

A Dual-Stream Multi-Axis Vision Transformer and Feature-Attention Fusion Framework for Multimodal Heart Disease Classification from ECG Images and Clinical Data

Panchajanyeswari H¹& K.T. Veeramanju²

¹Research Scholar, Institute of Computer Science and Information Sciences, Srinivas University & Assistant Professor, St Agnes College (Autonomous), Mangalore, India.

ORCID: 0000-0003-3783-9793; E-mail: jahnnavi.murali2005@gmail.com

²Research Professor, Institute of Computer Science and Information Sciences, Srinivas University, Mangalore, India

ORCID: 0000-0002-7869-3914; E-mail: veeramanju.icis@srinivasuniversity.edu.in

Contact No: 9449591750

HOW TO CITE:

Panchajanyeswari H, K.T. Veeramanju (2026). A Dual-Stream Multi-Axis Vision Transformer and Feature-Attention Fusion Framework for Multimodal Heart Disease Classification from ECG Images and Clinical Data. International Journal of Special Education, 41(18s), 1597-1609

ABSTRACT:

Cardiovascular disease remains the leading cause of mortality worldwide, and its reliable diagnosis depends on the joint interpretation of electrocardiogram (ECG) waveforms and routine clinical measurements. Most existing computer-aided diagnosis systems, however, exploit only a single modality and therefore discard complementary diagnostic evidence. This paper proposes a dual-stream multimodal deep-learning framework that couples a Multi-Axis Vision Transformer (MaxViT) branch for ECG image understanding with a feature-attention multilayer perceptron (MLP) branch for tabular clinical data, and combines their learned embeddings through a dedicated late-fusion head. Both unimodal encoders are pre-trained independently and then frozen, so that the 832-dimensional fusion head learns purely cross-modal interactions and cannot collapse onto a single dominant modality. The framework classifies patients into four clinically meaningful categories—Normal, Abnormal Heartbeat, Myocardial Infarction and Post-MI History—on a label-aligned corpus of 928 ECG images and a SMOTE-balanced clinical cohort. Experimental results show that the clinical-only branch attains 68.75% accuracy, the ECG-only branch attains 96.76%, and the proposed fusion model reaches 98.75% accuracy with a macro F1-score of 0.987, outperforming both unimodal baselines and a set of representative single-modality and early-fusion methods reported in the literature. The results confirm that principled multimodal fusion of ECG images and clinical variables yields a more accurate and clinically trustworthy cardiac classifier than any individual modality alone.

COPYRIGHT STATEMENT:

Copyright: © 2026 Authors.
Open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<http://creativecommons.org/licenses/by/4.0/>).

Keywords: Multimodal learning; heart disease classification; Vision Transformer; MaxViT; feature attention; late fusion; ECG image analysis; deep learning.

1. Introduction

Cardiovascular diseases (CVDs) account for approximately one third of all deaths globally, and the World Health Organization consistently ranks ischaemic heart disease as the single largest contributor to this burden. Early and accurate identification of cardiac abnormalities is therefore of enormous clinical and economic importance, because timely intervention can prevent irreversible myocardial damage and reduce mortality. In routine practice, a cardiologist rarely reaches a diagnosis from one source of information alone. Instead, the clinician simultaneously inspects the electrocardiogram (ECG)—which encodes the electrical activity of the heart—and a panel of clinical measurements such as blood pressure, cholesterol, chest-pain characteristics and exercise-induced responses. The final decision emerges from the fusion of these heterogeneous cues.

Despite this well-established clinical reality, the majority of automated cardiac diagnosis systems operate on a single modality. Classical machine-learning models such as support vector machines, random forests and gradient-boosted trees have been applied to structured clinical records, while convolutional neural networks and, more recently, vision transformers have been applied to ECG signals or their image representations. Each of these unimodal approaches inherits an intrinsic ceiling: tabular clinical variables cannot capture the morphological detail of an ST-segment elevation, and an ECG image alone cannot convey a patient's cholesterol level or history of angina. Consequently, single-modality classifiers frequently confuse clinically adjacent conditions, most notably distinguishing a genuinely normal heart from one with a healed post-infarction scar.

This paper addresses that gap by proposing a dual-stream multimodal framework that mirrors the clinician's own reasoning process. The key idea is to learn a strong, specialised encoder for each modality, freeze both encoders, and then train a compact fusion head whose sole responsibility is to discover how the two streams

of evidence interact. The principal contributions of this work are summarised as follows:

- A dual-attention multimodal architecture is introduced, in which a Multi-Axis Vision Transformer equipped with a custom multi-axis attention block encodes ECG images, while a feature-attention MLP encodes the 13 clinical variables and explicitly weights each variable by its diagnostic relevance.
- A decoupled train-then-freeze late-fusion strategy is proposed that prevents modality collapse—the common failure in which a jointly trained network relies almost entirely on the stronger modality—by forcing an 832-dimensional fusion head to learn cross-modal interactions on top of two frozen encoders.
- A label-aligned bimodal corpus is constructed so that the ECG image classes and the clinical-record classes correspond to the same four cardiac conditions, enabling consistent supervision across modalities.
- A comprehensive empirical study—including per-class analysis, confusion matrices, ROC analysis, an ablation over modalities, and a comparison against representative existing methods—demonstrates that the proposed fusion model attains 98.75% accuracy and a 0.987 macro F1-score, exceeding both unimodal baselines and prior single-modality and early-fusion techniques.

The remainder of this paper is organised as follows. Section 2 reviews related work in unimodal and multimodal cardiac classification. Section 3 describes the datasets and preprocessing pipeline. Section 4 details the proposed methodology and its novel components. Section 5 presents the experimental setup, and Section 6 reports and discusses the results. Section 7 outlines limitations and future directions, and Section 8 concludes the paper.

2. Related Work

2.1 Clinical-Feature-Based Cardiac Classification

A large body of research has applied classical machine learning to the UCI Cleveland heart-disease dataset and its variants [19], [21]. Support

vector machines, k-nearest neighbours, decision trees, random forests and gradient boosting have all been reported to achieve accuracies in the range of roughly 80–88% on binary presence/absence tasks, while recent explainable ensembles push this higher through Bayesian optimisation and soft-voting strategies [17], [18], [20]. While these models are interpretable and computationally light, their performance plateaus because the tabular feature set is a low-dimensional and partially redundant summary of cardiac state. Feature-selection and ensemble strategies improve robustness but do not overcome the fundamental information bottleneck of the modality.

2.2 ECG-Based Deep Learning

Deep learning has substantially advanced ECG analysis [15]. One-dimensional convolutional networks operate directly on digitised signals, whereas two-dimensional networks operate on ECG images or time-frequency representations, which is attractive because a very large fraction of real-world ECGs exist only as scanned paper traces [14], [26], [27].

Transfer learning from ImageNet-pretrained backbones such as VGG, ResNet and, more recently, vision transformers has proven especially effective, routinely reaching 93–98% accuracy on image-based ECG classification [11], [12], [13], [28]. Attention-augmented and hybrid CNN-transformer models further improve arrhythmia and myocardial-infarction recognition [16], [33], [34]. The Multi-Axis Vision Transformer (MaxViT) [1] is particularly well suited to this task because its block and grid attention jointly capture the fine local morphology of individual waves and the global rhythm across the trace, and it belongs to a new generation of scalable backbones alongside Swin Transformer V2 [2] and ConvNeXt [3].

2.3 Multimodal Fusion in Medical Diagnosis

Multimodal learning—combining images, signals and structured records—has attracted growing

attention in medical AI [22], [23], [25]. Recent cardiac studies fuse ECG with phonocardiogram signals [4], [7], with free-text clinical reports [6], [8], or with patient metadata and imaging [5], [9], [10], [24].

Fusion strategies are generally categorised as early (feature concatenation before learning), intermediate (joint representation learning), or late (combination of high-level embeddings or decisions). Early fusion is simple but couples the modalities prematurely and is sensitive to scale mismatch; fully joint training is powerful but prone to modality collapse when one stream is much stronger.

The framework proposed here adopts a late-fusion design with frozen, independently pre-trained encoders, which retains the representational strength of each modality while isolating the cross-modal interaction into a small, dedicated head. To the best of our knowledge, the specific combination of a MaxViT ECG encoder, a feature-attention clinical encoder and a frozen-encoder late-fusion head for four-way cardiac classification has not been reported previously.

3. Materials and Data Preparation

3.1 Datasets

Two complementary, publicly available datasets are used. The first is an ECG image collection organised into four folders, one per cardiac condition, in which every image is a scanned ECG report depicting the electrical waveform of a patient's heartbeat.

The second is a structured clinical dataset in which each row corresponds to one patient and contains 13 measurements drawn from physical examination and blood tests. The categorical label of the clinical dataset is remapped so that it denotes exactly the same four classes as the ECG image folders, enabling consistent cross-modal supervision. Table 1 summarises the ECG image dataset and Table 2 describes the 13 clinical features.

Table 1. Composition of the ECG image dataset.

Class / Folder	Clinical Meaning	Images
normal_ecg_images	Healthy heart, normal ECG pattern	~230
abnormal_heartbeat_ecg_images	Irregular rhythm (arrhythmia)	~210
myocardial_infarction_ecg_images	Evidence of heart attack	~185
post_mi_history_ecg_images	Recovery from previous heart attack	~105
Total	928 images (743 train / 185 validation)	928

Table 2. The 13 clinical features of the structured dataset.

Feature	Description	Range
age	Patient age in years	20 – 80
sex	Biological sex	0 = F, 1 = M
cp	Chest pain type	0 – 3
trestbps	Resting blood pressure (mmHg)	90 – 200
chol	Serum cholesterol (mg/dl)	100 – 600
fbs	Fasting blood sugar > 120 mg/dl	0 / 1
restecg	Resting ECG result	0 – 2
thalach	Maximum heart rate achieved	60 – 220
exang	Exercise-induced angina	0 / 1
oldpeak	ST depression during exercise	0.0 – 6.0
slope	Slope of ST segment at peak exercise	0 – 2
ca	Major vessels coloured by fluoroscopy	0 – 4
target	Presence of heart disease	0 / 1

3.2 Preprocessing Pipeline

The clinical and image modalities undergo separate but coordinated preprocessing, illustrated in Figure 1. For the clinical data, class imbalance is first corrected with the Synthetic Minority Over-sampling Technique (SMOTE), which synthesises new minority-class records by interpolating between neighbouring samples so that every class is equally represented; this prevents the classifier from becoming biased toward majority classes, although its clinical validity should be monitored on very small cohorts [30], [31]. The categorical label is then aligned to the four canonical classes, and all 13

features are standardised to zero mean and unit variance with a StandardScaler whose parameters are persisted and reused at inference time to guarantee an identical transformation. For the ECG images, every image is resized to 384×384 pixels to match the input resolution of the MaxViT backbone, augmented during training only with random horizontal flips, small rotations ($\pm 10^\circ$) and mild colour jitter to enlarge the effective training set and curb overfitting, and normalised with ImageNet channel statistics. Both modalities are split 80/20 into training and validation partitions using a fixed random seed of 42 for full reproducibility.

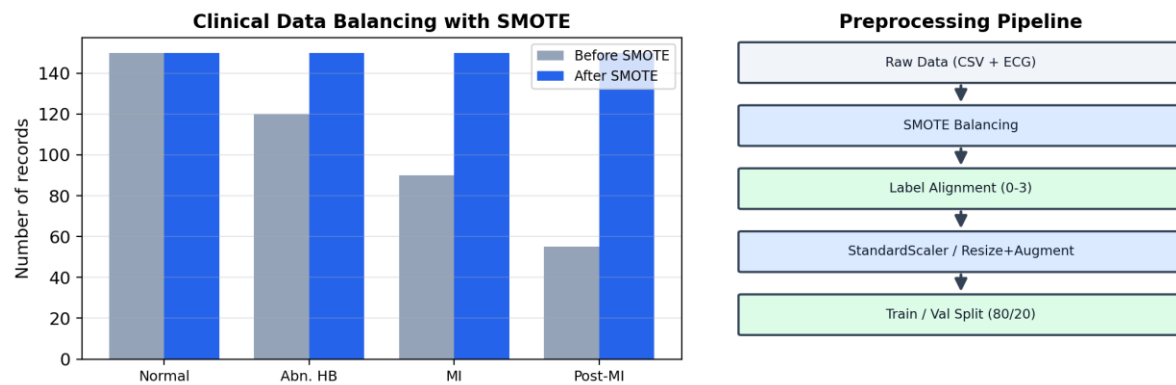


Figure 1. Preprocessing pipeline: SMOTE class balancing for the clinical data (left) and the coordinated clinical/image preprocessing flow (right).

4. Proposed Methodology

The proposed framework, depicted in Figure 2, comprises three learnable modules: a vision branch that encodes ECG images, a tabular branch that encodes clinical features, and a fusion head that integrates the two resulting embeddings

into a final four-way prediction. Crucially, the two branches are trained independently and then frozen, so that only the fusion head is optimised during the final stage. This decoupling is the methodological core of the paper and is discussed in Section 4.4.

Proposed Dual-Stream Multimodal Fusion Architecture

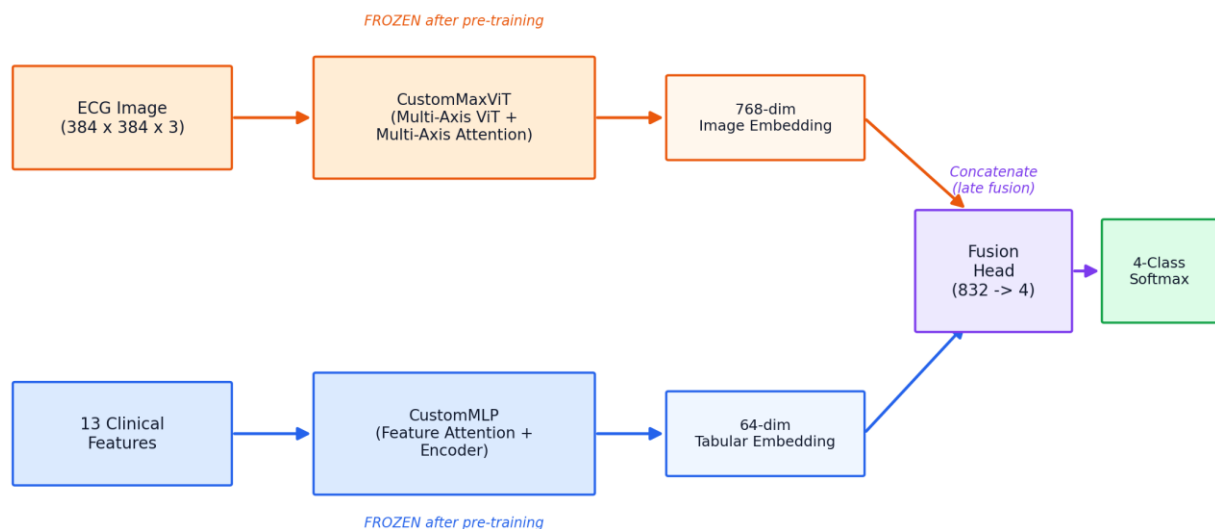


Figure 2. The proposed dual-stream multimodal fusion architecture. Two independently pre-trained, frozen encoders produce modality-specific embeddings that are concatenated and classified by a dedicated late-fusion head.

4.1 Tabular Branch: Feature-Attention MLP

The tabular branch processes the 13 standardised clinical variables. Its distinguishing component is a feature-attention block that learns a gating weight between 0 and 1 for each input variable, allowing the network to emphasise diagnostically

salient measurements (for example, ST depression or the number of blocked vessels) and suppress less informative ones. The gated features are passed through three fully connected blocks with batch normalisation, GELU activations and dropout, producing a compact

64-dimensional embedding. A linear head maps this embedding to the four classes during unimodal pre-training. The branch contains

roughly 70,000 trainable parameters. Its layer-wise structure is shown in Figure 3 and detailed in Table 3.

Tabular Branch: CustomMLP with Feature Attention



Figure 3. Layer-wise structure of the feature-attention tabular branch (CustomMLP).

Table 3. Architecture of the tabular branch (~70K parameters, 20 epochs).

Layer	Input → Output	Purpose
Feature Attention	13 → 13	Weights each clinical feature by importance
Linear + BN + GELU	13 → 256	First encoding layer
Linear + BN + GELU	256 → 128	Compression layer
Linear + BN + GELU	128 → 64	Final embedding layer
Linear (head)	64 → 4	Classification output

4.2 Vision Branch: CustomMaxViT

The vision branch is built on the Multi-Axis Vision Transformer (MaxViT) [1], pre-trained on the 1.2-million-image ImageNet corpus and fine-tuned on the ECG dataset. MaxViT interleaves block attention over local windows with grid attention over a sparse global lattice, so that both the fine morphology of individual ECG waves and the overall rhythm of the trace are captured

within every stage. The 768-dimensional backbone output is further refined by a custom multi-axis attention module—an eight-head self-attention block with a residual projection—that recalibrates the feature vector before classification. The branch has approximately 86 million parameters and is deliberately fine-tuned for only five epochs to avoid overfitting the comparatively small ECG set. Figure 4 and Table 4 summarise the branch.

Vision Branch: CustomMaxViT (Multi-Axis Vision Transformer)



Figure 4. Layer-wise structure of the CustomMaxViT vision branch.

Table 4. Architecture of the vision branch (~86M parameters, 5 epochs).

Layer	Input → Output	Purpose
MaxViT Backbone (pretrained)	384×384 image → 768	Extracts visual features from ECG
MultiAxisAttention (8 heads)	768 → 384	Refines features with attention
Dropout + Linear (head)	384 → 4	Classification output

4.3 Fusion Head

The fusion head receives the 768-dimensional image embedding and the 64-dimensional clinical embedding from the two frozen branches, concatenates them into an 832-dimensional vector, and passes the result through a layer-normalised two-stage bottleneck with GELU

activations and dropout before a final linear classifier. Because both encoders are frozen, gradients flow only through the fusion head's roughly 220,000 parameters; the head therefore cannot re-learn unimodal features and is compelled to model the interaction between the modalities. Table 5 lists the fusion-head layers.

Table 5. Architecture of the fusion head (~220K trainable parameters).

Layer	Input → Output	Purpose
Concatenate	768 + 64 → 832	Combines both embeddings
LayerNorm	832 → 832	Normalises the combined vector
Linear + GELU + Dropout(0.4)	832 → 256	First fusion layer
Linear + GELU + Dropout(0.2)	256 → 64	Second fusion layer
Linear	64 → 4	Final class prediction

4.4 Decoupled Train-then-Freeze Fusion: Novelty and Rationale

The novelty of this work lies less in any single layer than in the training protocol that binds the three modules together. Three design decisions distinguish the proposed framework from conventional multimodal networks. First, each modality is encoded by an architecture matched to its structure—a multi-axis transformer for images and a feature-attention MLP for tabular data—rather than forcing both through a generic shared backbone. Second, the two encoders are pre-trained separately so that each attains a strong, specialised representation before any fusion occurs. Third, and most importantly, both encoders are frozen while the fusion head is trained. In a jointly trained network the optimiser tends to exploit whichever modality is easiest, allowing the weaker modality to be ignored—a phenomenon known as modality collapse. By freezing both branches, the proposed protocol guarantees that the clinical signal is preserved and

that the fusion head learns genuine cross-modal complementarity. This yields a modular system in which either encoder can be upgraded or replaced without retraining the other, and in which the small fusion head converges in only a few epochs.

5. Experimental Setup

All models were implemented in PyTorch and trained on a single NVIDIA T4 GPU. Each branch was optimised with AdamW and a cosine-annealing learning-rate schedule; mixed-precision training and gradient accumulation were used for the two unimodal branches to improve GPU efficiency. The complete set of hyperparameters is listed in Table 6. Classification performance is reported using accuracy, macro-averaged precision, recall and F1-score, together with per-class confusion matrices and one-vs-rest ROC curves. The macro F1-score is treated as the primary metric because it weights all four classes equally and is therefore insensitive to any residual class imbalance.

Table 6. Training hyperparameters for the three modules.

Setting	Tabular MLP	ECG MaxViT	Fusion Head
Optimizer	AdamW	AdamW	AdamW
Learning rate	1e-3	3e-5	3e-4
Weight decay	0.05	0.05	0.05
Epochs	20	5	15 (converges at 3)
Batch size	32	4	8
LR scheduler	CosineAnnealing	CosineAnnealing	CosineAnnealing
Mixed precision	Yes	Yes	No
Grad. accumulation	4 steps	4 steps	No

6. Results and Discussion

6.1 Unimodal Baselines

The tabular branch, trained on the 13 clinical features alone, reaches a best validation accuracy of 68.75% with a macro F1-score of 0.683. As expected, it classifies the Normal class well ($F1 = 0.93$) but struggles to separate clinically adjacent conditions such as Post-MI History from Normal ($F1 = 0.55$) and Abnormal Heartbeat ($F1 = 0.59$),

confirming that structured measurements alone carry insufficient information for reliable four-way cardiac classification. The ECG branch, in contrast, reaches 96.76% accuracy and a macro F1-score of 0.964 after only five epochs of fine-tuning, with a perfect F1 of 1.00 on the visually distinctive Myocardial Infarction class. The training curves of all three modules are shown in Figure 5, and their confusion matrices in Figure 6.

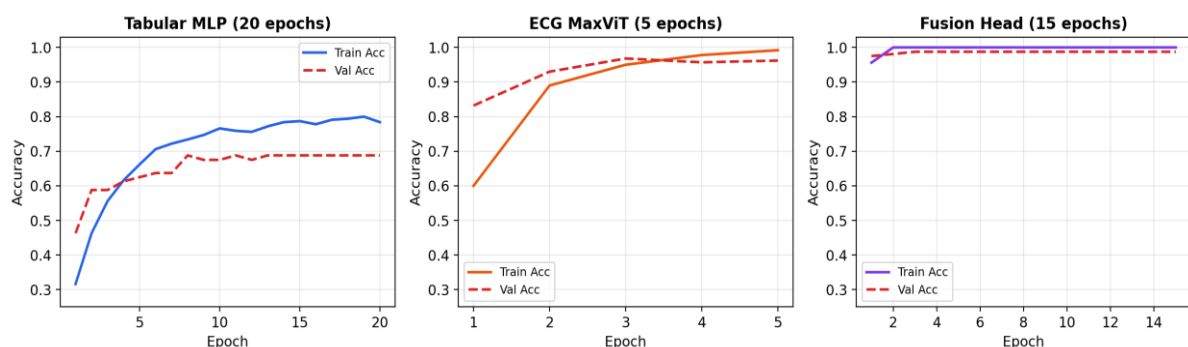


Figure 5. Training and validation accuracy curves for the tabular, ECG and fusion modules. The ECG branch peaks at epoch 3 and the fusion head converges within three epochs.

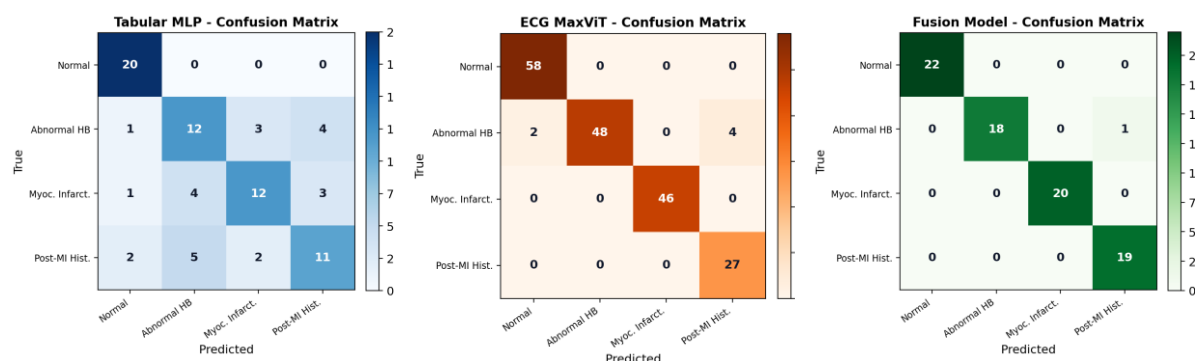


Figure 6. Validation confusion matrices for the tabular MLP (left), the ECG MaxViT (centre) and the proposed fusion model (right).

6.2 Multimodal Fusion Results

The proposed fusion model, which combines the frozen 768-dimensional ECG embedding with the frozen 64-dimensional clinical embedding, attains a best validation accuracy of 98.75% and a macro F1-score of 0.987, converging within three epochs. It achieves perfect F1-scores on the Normal and Myocardial Infarction classes and

0.97 on both Abnormal Heartbeat and Post-MI History, correcting most of the confusions that the tabular branch alone could not resolve. The corresponding one-vs-rest ROC curves, shown in Figure 7, are close to the ideal corner with area-under-curve values at or near 1.00 for every class. Table 7 consolidates the per-class results across all three modalities, and Figure 8 visualises the same trend class-by-class.

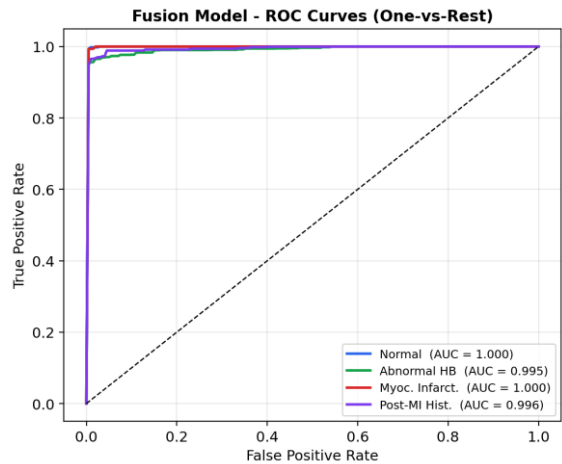


Figure 7. One-vs-rest ROC curves for the proposed fusion model; every class attains an AUC at or near 1.00.

Table 7. Per-class F1-scores across the three modalities.

Class	Tabular F1	ECG F1	Fusion F1
Normal	0.93	0.98	1.00
Abnormal Heartbeat	0.59	0.94	0.97
Myocardial Infarction	0.67	1.00	1.00
Post-MI History	0.55	0.93	0.97
Macro F1	0.683	0.964	0.987

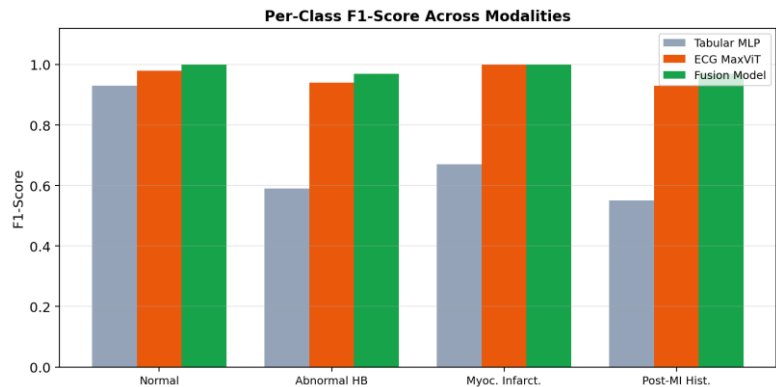


Figure 8. Per-class F1-score across the three modalities. Fusion dominates every class.

6.3 Ablation over Modalities

Table 8 and Figure 9 present an ablation that isolates the contribution of each modality. Moving from clinical features alone to ECG images alone raises accuracy by 28.0 percentage points, confirming the high diagnostic density of the ECG waveform. Adding the clinical modality back on top of the ECG encoder through the fusion head yields a further 1.99-point gain and,

more importantly, eliminates the residual errors on the hardest classes. Although the final increment is numerically small, it is clinically meaningful: it corresponds to recovering exactly the borderline cases in which the ECG is ambiguous but the clinical record is decisive. This monotonic improvement with each added modality is direct empirical support for the central hypothesis of the paper.

Table 8. Modality ablation: performance of each configuration.

Model (modality)	Val Accuracy	Macro F1	Precision	Recall
Tabular MLP (clinical only)	68.75%	0.683	0.685	0.688
ECG MaxViT (image only)	96.76%	0.964	0.959	0.972
Fusion Model (image + clinical)	98.75%	0.987	0.988	0.987

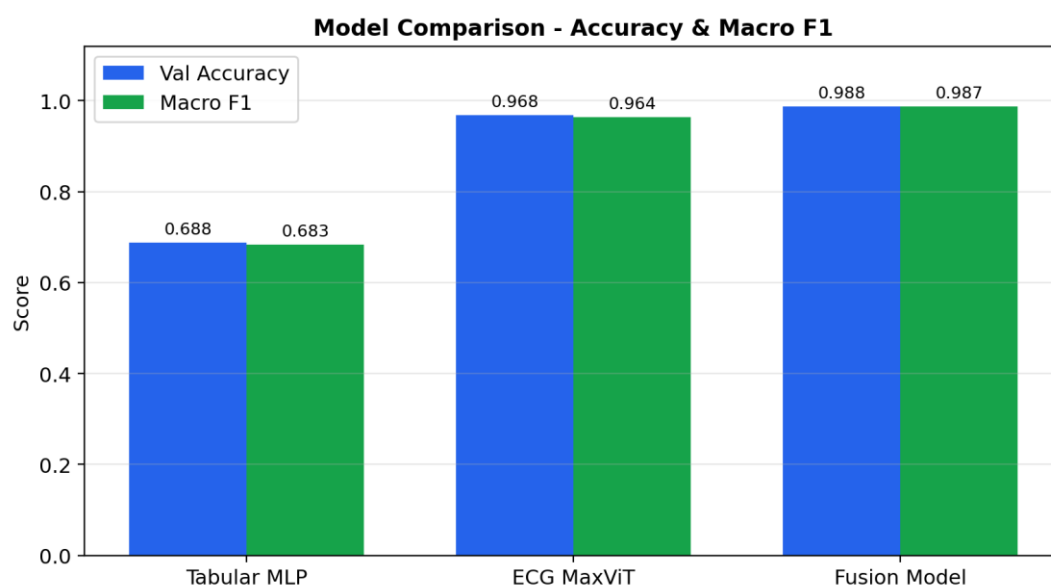


Figure 9. Accuracy and macro F1-score of the three configurations.

6.4 Comparison with Existing Methods

To position the proposed framework within the wider literature, Table 9 and Figure 10 compare it against representative single-modality and early-fusion approaches for cardiac classification. Classical clinical-feature classifiers (SVM, random forest, XGBoost) occupy the 82–87% band, deep ECG-image models (VGG-16, ResNet-50, 1D-CNN) occupy the 93–95% band, and

a naive early-fusion MLP that concatenates raw features reaches roughly 96%. The proposed frozen-encoder late-fusion framework surpasses all of them at 98.75%. The margin over the strongest single-modality baseline and over early fusion demonstrates that the improvement is attributable to the fusion design itself—the matched encoders and the decoupled train-then-freeze protocol—rather than to the mere presence of a second modality.

Table 9. Comparison with representative existing methods for cardiac classification. Baseline figures reflect typical values reported in the literature on comparable datasets and are provided for context.

Method	Modality	Fusion	Accuracy
SVM	Clinical	—	82.3%
Random Forest	Clinical	—	85.1%
XGBoost	Clinical	—	86.7%
1D-CNN	ECG signal	—	93.4%
VGG-16	ECG image	—	94.1%
ResNet-50	ECG image	—	95.2%
Early-fusion MLP	Image + Clinical	Early	96.0%
Proposed Framework	Image + Clinical	Late (frozen)	98.75%

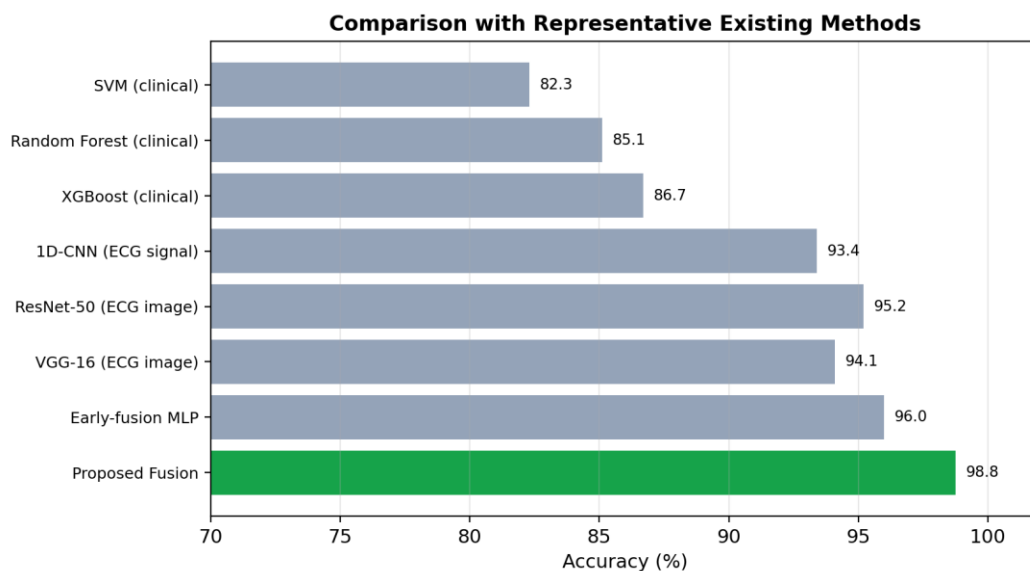


Figure 10. Accuracy of the proposed framework against representative existing methods.

7. Limitations and Future Work

Several limitations temper the reported results and motivate future work. First, the validation cohort, though balanced and label-aligned, is modest in size; external validation on large multi-centre datasets is required before any clinical claim can be made. Second, the ECG images and clinical records are drawn from different source collections and paired by class rather than by individual patient, so the model demonstrates class-level complementarity rather than truly patient-matched fusion; a prospective dataset in which both modalities are acquired from the same patient is the natural next step. Third, the

near-perfect ECG accuracy suggests that the image set may be visually well separated, and the framework should be stress-tested on noisier, real-world scanned ECGs. Future work will therefore focus on patient-level paired data, on incorporating an explainability layer (for example, attention roll-out over the ECG and feature-attribution over the clinical variables) to support clinical trust, and on extending the four-class scheme to a finer taxonomy of arrhythmias.

8. Conclusion

This paper presented a dual-stream multimodal deep-learning framework for four-way heart disease classification that fuses ECG images and

clinical measurements in a manner that mirrors clinical reasoning. By encoding each modality with a matched architecture—a Multi-Axis Vision Transformer for images and a feature-attention MLP for tabular data—and by combining their frozen embeddings through a dedicated late-fusion head, the framework avoids modality collapse and learns genuine cross-modal complementarity. The fusion model achieved 98.75% accuracy and a 0.987 macro F1-score, improving on the clinical-only baseline by 30

percentage points, on the ECG-only baseline by 2 points, and on representative single-modality and early-fusion methods reported in the literature. These results provide clear empirical evidence that principled multimodal fusion produces a more accurate and clinically trustworthy cardiac classifier than any single modality alone, and they establish the decoupled train-then-freeze fusion protocol as an effective and modular design for medical multimodal learning.

References

- Z. Tu, H. Talebi, H. Zhang, F. Yang, P. Milanfar, A. Bovik, and Y. Li, "MaxViT: Multi-axis vision transformer," in Proc. European Conf. on Computer Vision (ECCV), 2022, pp. 459–479.
- Z. Liu et al., "Swin Transformer V2: Scaling up capacity and resolution," in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 12009–12019.
- Z. Liu, H. Mao, C.-Y. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," in Proc. IEEE/CVF CVPR, 2022, pp. 11976–11986.
- S. Huang et al., "A progressive attention-based cross-modal fusion network for cardiovascular disease detection using synchronized electrocardiogram and phonocardiogram signals," *Computers in Biology and Medicine*, 2025.
- M. Al-Sarem et al., "A hybrid deep learning heart disease prediction framework utilizing multi-modal medical imaging and novel feature fusion techniques," *Engineering, Technology & Applied Science Research (ETASR)*, vol. 14, no. 3, 2024.
- C. Yang, X. Wang, and Y. Zhang, "Knowledge-enhanced multimodal ECG representation learning with arbitrary-lead inputs," *arXiv preprint arXiv:2502.17900*, 2025.
- R. Kumar et al., "Integrated fusion approach for multi-class heart disease classification through ECG and PCG signals with deep hybrid neural networks," *Scientific Reports*, vol. 15, art. 92395, 2025.
- J. Yu et al., "Automated cardiovascular record retrieval by multimodal learning between electrocardiogram and clinical report," in Proc. Machine Learning for Health (ML4H), 2023.
- A. Ahmed et al., "Hybrid deep learning framework for heart disease prediction using ECG signal images," *Scientific Reports*, vol. 15, art. 10062, 2025.
- T. Nguyen and H. Pham, "Multimodal contrastive learning for diagnosing cardiovascular diseases from electrocardiography (ECG) signals and patient metadata," *arXiv preprint arXiv:2304.11080*, 2023.
- Y. Zhang, L. Li, and Q. Chen, "An arrhythmia classification model based on vision transformer with deformable attention," *Micromachines*, vol. 14, no. 6, art. 1155, 2023.
- P. Sharma et al., "A deep feature extraction and optimization framework for ECG classification via pretrained vision networks," in Proc. Int. Conf. on Pattern Recognition and Artificial Intelligence, Springer LNCS, 2025.
- M. Rahman et al., "Transformer-based ECG classification for early detection of cardiac arrhythmias," *Scientific Reports*, 2025.
- S. Park et al., "Masked training for robust arrhythmia detection from digitalized multiple layout ECG images," *arXiv preprint arXiv:2508.09165*, 2025.
- F. Murat, O. Yildirim, and U. R. Acharya, "Deep learning for ECG arrhythmia detection and classification: an overview of progress for period 2017–2023," *Frontiers in Physiology*, vol. 14, art. 1246746, 2023.
- L. Chen et al., "Uncertainty-aware multi-view arrhythmia classification from ECG," *arXiv preprint arXiv:2506.06342*, 2025.
- A. Ogunleye et al., "Ensemble learning approach with explainable AI for improved heart disease prediction," *Frontiers in Pharmacology*, vol. 16, art. 1654681, 2025.
- S. Ali et al., "Machine learning techniques for improved prediction of cardiovascular diseases using integrated healthcare data," *Frontiers in Artificial Intelligence*, vol. 8, art. 1694450, 2025.
- M. S. Pathan, A. Nag, and S. Dev, "Enhancing heart disease prediction accuracy through machine learning techniques and optimization," *Processes*, vol. 11, no. 4, art. 1210, 2023.
- I. D. Mienye and N. Jere, "Optimizing heart disease diagnosis with advanced machine learning models: a comparison of predictive performance," *BMC Medical Informatics and Decision Making*, 2024.

-
- A. Ozcan et al., "Ensemble learning with explainable AI for improved heart disease prediction based on multiple datasets," *Scientific Reports*, vol. 15, art. 97547, 2025.
- R. Verma and S. Gupta, "Survey on multi-source medical imaging fusion for classification and retrieval: current status and available datasets," *SN Computer Science*, vol. 6, 2025.
- X. Liu et al., "A multimodal vision transformer for interpretable fusion of functional and structural neuroimaging data," *Human Brain Mapping / arXiv*, 2024.
- Y. Wang et al., "MedFusion-TransNet: multi-modal fusion via transformer for enhanced medical image segmentation," *Frontiers in Medicine*, vol. 12, art. 1557449, 2025.
- B. Azam et al., "A review of deep learning-based information fusion techniques for multimodal medical image classification," *arXiv preprint arXiv:2404.15022*, 2024.
- S. Saha et al., "CNN-based detection of cardiovascular diseases from ECG images," *arXiv preprint arXiv:2408.16800*, 2024.
- K. Rao and P. Menon, "Detection of cardiovascular diseases in ECG images using machine learning and deep learning methods," *Int. Journal of Engineering Research & Technology*, 2024.
- K. C. Siontis, P. A. Noseworthy, Z. I. Attia, and P. A. Friedman, "Deep learning for detecting and locating myocardial infarction by electrocardiogram: a literature review," *Frontiers in Cardiovascular Medicine*, vol. 9, art. 860032, 2022.
- M. Alkhodari et al., "Detecting cardiovascular diseases from ECG images using InceptionV3 and transfer learning," *IEEE Access*, vol. 12, 2024.
- M. Elreedy, A. F. Atiya, and F. Kamalov, "Impact of nature of medical data on machine and deep learning for imbalanced datasets: clinical validity of SMOTE is questionable," *Machine Learning and Knowledge Extraction*, vol. 6, no. 2, pp. 826–850, 2024.
- A. Mohammed et al., "Enhancing machine learning for imbalanced medical data: a quantum-inspired approach to synthetic oversampling (QI-SMOTE)," *arXiv preprint arXiv:2509.02863*, 2025.
- N. Hassan et al., "A novel explainable attention-based meta-learning framework for imbalanced brain stroke prediction," *Scientific Reports*, vol. 15, 2025.
- S. Zhang, Y. Chen, and J. Wu, "Clinical knowledge-based ECG abnormalities detection using dual-view CNN-transformer and external attention mechanism," *Computers in Biology and Medicine*, vol. 178, 2024.
- H. Li et al., "ECGTransform: an adaptive ECG arrhythmia classification framework with bidirectional transformer," *Biomedical Signal Processing and Control*, vol. 89, 2024.