A Systematic Comparison of Data Representations for Transformer-Based ECG Arrhythmia Classification

Mona Aman, Godbright Uiso, Carine Mukamakuza, Vijayakumar Bhagavatula
[1] Carnegie Mellon University Africa
Regional ICT Center of Excellence Bldg
Kigali, Rwanda
{amona, guisso, cmukamakuza}@andrew.cmu.edu, kumar@ece.cmu.edu

Abstract

Automated electrocardiogram (ECG) classification plays a key role in detecting cardiac arrhythmias efficiently and objectively. Despite major advances in deep learning, there remains no consensus on whether one-dimensional (1D) temporal or two-dimensional (2D) time–frequency representations yield superior diagnostic accuracy. This study presents a controlled comparison between Vision Transformer (ViT) architectures trained on raw 1D ECG sequences and Short-Time Fourier Transform (STFT)-based 2D spectrograms using the CPSC2018 dataset. Both models share comparable architectures and parameter counts to isolate the effect of signal representation. The 1D-ViT achieved the highest overall accuracy (96.5%) and F1-score (96.5%), confirming that preserving temporal continuity is critical for arrhythmia discrimination. The 2D-ViT achieved lower accuracy (92.6%) due to temporal information loss, though it maintained competitive calibration (AUC 98.6%) and generalization. A bidirectional fusion model combining both encoders through cross-attention exhibited complementary behavior but did not surpass the 1D baseline. These findings indicate that while spectro-temporal information can enhance interpretability and stability, temporal-domain fidelity remains the dominant factor for reliable ECG classification.

Introduction

Automated analysis of the electrocardiogram (ECG) is a critical clinical imperative for addressing cardiovascular diseases (CVDs), the leading cause of global mortality (Rifat 2022). Deep learning (DL) has emerged as the state-of-the-art solution, superseding traditional labor-intensive interpretation(Wu and Guo 2025). While early architectures like Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) showed promise, they struggled to capture long-range temporal dependencies(Wu and Guo 2025).The advent of the Transformer architecture (Vaswani et al. 2023), adapted as the Vision Transformer (ViT) (Dosovitskiy et al. 2021), has overcome these limitations. By employing self-attention, Transformers effectively model global dependencies across entire ECG recordings, establishing them as the new standard for robust analysis(Vaid et al. 2023).

Despite the consensus on the power of Transformer-based models, a fundamental dilemma persists in computational cardiology: how should the ECG signal be represented for optimal input? This splits the field into two competing approaches: 1) the **1D time-series approach**, which processes the raw signal directly, preserving perfect temporal fidelity, and 2) the **2D image-based approach**, which converts the signal into a time-frequency representation (e.g., scalogram via CWT , or Mel-Spectrogram) to leverage image-optimized ViT models (Dosovitskiy et al. 2021). Previous comparative studies have failed to resolve this dichotomy conclusively, resulting in a "Missing Fair Comparison". These failures stem from pervasive methodological inconsistencies, including the use of mismatched model architectures with non-equivalent parameter counts, non-standardized datasets, and variable, non-uniform 1D-to-2D conversion techniques that may obscure true performance differences (Vaid et al. 2023) (Song and Lee 2024).

Resolving this methodological dilemma is critical for advancing clinically robust cardiac AI. The choice of representation dictates the model's feature sensitivity: conditions like Atrial Fibrillation (AFIB) are defined by millisecond-precision R-R interval irregularity, demanding maximal temporal fidelity(Duan et al. 2022). Conversely, complex or rare arrhythmias, such as General Supraventricular Tachycardia (GSVT), may manifest subtle features better captured by explicit frequency-domain analysis (Zhang et al. 2023). If current representations systematically discard crucial information required for specific diagnoses, deployment risks diagnostic failure in critical clinical scenarios. A controlled evaluation is therefore essential to quantify these trade-offs and guide the next generation of model design. Furthermore, the small percentage point gains in accuracy identified through rigorous comparison translate into a significant increase in correct diagnoses when deployed across large patient populations.

The objective of this study is to definitively **isolate the impact of data representation** by conducting a rigorously controlled head-to-head comparison between 1D,2D and Hybrid Vision Transformer models. We address the lack of standardization by ensuring experimental parity:

1. **Equivalent Architectures:** We utilized 1D-ViT (11.2M parameters) and 2D-ViT (11.0M parameters) with identical training configurations whilst the Hybrid approach

used encoders from prior mentioned model architectures.

2. **Standardized Data:** All models were trained, validated, and tested using the same standardized CPSC2018 12-lead ECG dataset, employing a strict patient-level splitting paradigm to prevent data leakage.

Related Work

The 1D vs. 2D Dichotomy in Transformer-Based ECG Analysis

The automated analysis of ECG signals has rapidly evolved from traditional 1D-CNNs (Mhetre, Mhetre, and Lokhande 2025) and RNNs (Zacarias et al. 2024) to Transformer-based architectures. This shift, sparked by the "Attention is All You Need" paradigm (Vaswani et al. 2023), has established Transformers as the state-of-the-art due to their ability to capture global dependencies. However, this has created a fundamental dichotomy in data representation:

1. **1D Time-Series:** This approach processes the raw ECG signal directly, often with 1D-Transformer architectures. Its primary strength is preserving perfect temporal fidelity, which is critical for rhythm-based diagnostics.
2. **2D Image-Based:** This approach converts the 1D signal into a 2D time-frequency representation (e.g., CWT, STFT, or Mel-spectrograms (Song and Lee 2024)) and applies powerful Vision Transformers (ViT) (Dosovitskiy et al. 2021). This method excels at spectral analysis and leverages models pre-trained on massive image datasets, as demonstrated by models like HeartBEiT (Vaid et al. 2023).

The Missing Fair Comparison and Rise of Hybrid Models

The critical question of which representation is superior remains unanswered. Previous comparative studies have been inconclusive due to a pervasive "Missing Fair Comparison." For instance, a comprehensive study by Narotamo et al. (Narotamo et al. 2024) using GRU,LSTM and 1D-CNN found that 2D ViTs generally outperformed 2D models on the PTB-XL dataset. However, they also noted that 2D representations showed strengths for specific cardiac conditions and that hybrid fusion models were promising.

This highlights the central gap: these studies, including (Vaid et al. 2023) and (Song and Lee 2024), fail to provide a definitive answer because they suffer from **methodological inconsistencies**, such as mismatched model architectures, non-equivalent parameter counts, and different datasets. It is therefore impossible to isolate the true impact of the **data representation** from architectural choices.

Recognizing that both 1D (temporal) and 2D (spectral) views offer complementary information, many have explored hybrid fusion approaches (Narotamo et al. 2024). These range from simple fusion to more complex bidirectional and cross-attention mechanisms (Dong et al. 2023) that allow information to flow between the two modalities. A summary of these approaches is presented in Table 1.

However, the design of an optimal fusion architecture is premature without a fundamental understanding of each modality's independent trade-offs. This study addresses this critical gap by conducting the first rigorously controlled head-to-head comparison, using architecturally-equivalent 1D and 2D Vision Transformers under identical experimental conditions to definitively isolate the impact of data representation.

Approach Category	Key Architectures	Strengths	Limitations
1D Time-Series Models	1D-CNN, LSTM, GRU, 1D-Transformer	Temporal fidelity, direct processing	Limited spectral analysis, local receptive fields
2D Image-Based Models	2D-CNN, Vision Transformer	Frequency analysis, global attention	Temporal resolution loss, artifacts
Multimodal Fusion	Early/Late/Intermediate fusion	Complementary information	Suboptimal integration
Bidirectional Transformers	Cross-attention	Dynamic exchange	Computational cost

Table 1: Summary of Key Approaches in ECG Classification Literature

Methodology

This study used the 12-lead large-scale electrocardiogram database for the study of rheumatic disorders to ensure the use of a comprehensive and publicly accessible clinical dataset, thereby facilitating robust model generalization and reproducibility (Zheng et al. 2020). The dataset comprises 45,152 12-lead ECG recordings representing the standard clinical configuration. Each recording captures 10 seconds of cardiac activity at a sampling rate of 500 Hz.For the multiclass classification task, the signals were categorized into four merged superclasses: Atrial Fibrillation (AFIB), General Supraventricular Tachycardia (GSVT), Sinus Bradycardia (SB), and Sinus Rhythm (SR) (Zheng et al. 2020). The complete class distribution and merging scheme are detailed in Table 2.

A comprehensive preprocessing pipeline was applied to all ECG signals to ensure signal quality and consistency, as illustrated in Figure 1. The pipeline consisted of sequential filtering, resampling, normalization, and quality control steps.

First, bandpass filtering was performed using a fourth-order Butterworth filter with cutoff frequencies of 0.5 to 50 Hz to remove baseline wander and high-frequency noise while preserving clinically relevant waveform components(Zheng et al. 2020). Zero-phase forward-backward filtering (`filtfilt`) prevented phase distortion. A 50 Hz notch filter with quality factor 30 was then applied to eliminate powerline interference (Payal Kohli 2023).

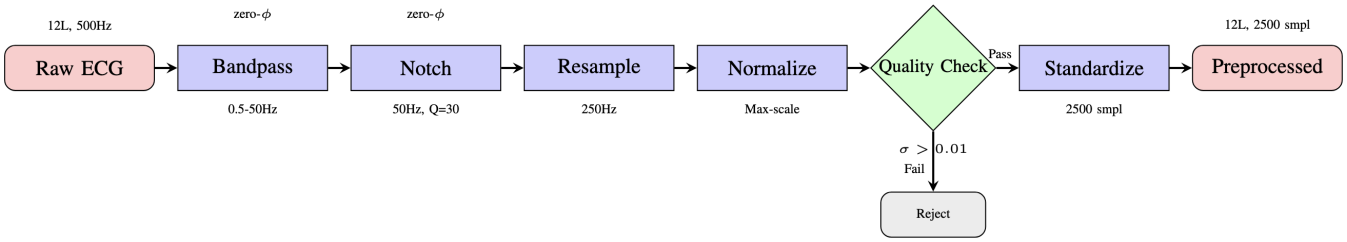

Figure 1: ECG signal preprocessing pipeline. The workflow includes bandpass filtering (0.5-50 Hz) for noise removal, notch filtering for powerline interference, resampling to 250 Hz, max-rescaling normalization, quality control check, and length standardization to 2,500 samples.

Signals were resampled to 250 Hz using linear interpolation, balancing computational efficiency with sufficient temporal resolution for accurate rhythm classification. Each recording was standardized to a fixed length of 2,500 samples (10 seconds at 250 Hz) through truncation of longer recordings or zero-padding of shorter ones.

Max-rescaling normalization was applied per lead, dividing each signal by its maximum absolute value to standardize amplitudes to the range $[-1, 1]$ while preserving relative amplitude relationships between leads. The normalization was computed as $x_{normalized} = x / \max(|x|)$.

A rigorous quality control procedure excluded corrupted or non-informative recordings. Signals exhibiting minimal variance (standard deviation below 0.01 per lead) or maximum normalized amplitudes exceeding 1.5 were discarded. Additional checks verified the absence of non-finite values throughout processing.

The preprocessing pipeline was implemented using SciPy for signal processing functions and NumPy for numerical computations. All filtering operations used zero-phase filtering to prevent temporal distortion of ECG waveforms, which is critical for accurate feature extraction. The bandpass filter range of 0.5–50 Hz corresponds to the diagnostic bandwidth recommended by the American Heart Association for computerized ECG analysis, ensuring preservation of relevant cardiac features including P-waves, QRS complexes, and T-waves while removing artifacts.

Model Architectures

This study implements two architecturally equivalent Vision Transformer (ViT) models that differ only in their input data representation: 1D temporal signals and 2D time-frequency spectrograms. Both architectures share identical transformer configurations to enable fair comparison and subsequent multi-modal fusion.

1D Temporal Transformer Architecture The 1D-ViT processes raw ECG time series directly, preserving crucial temporal precision for rhythm-based arrhythmia detection. As shown in Fig. 3, the architecture begins with input ECG signals of dimension $R^{12\times2500}$ representing 12 leads and 2,500 samples (10 seconds at 250 Hz). The signal is partitioned into non-overlapping patches of 25 samples each via a 1D convolutional layer with stride 25, resulting in 100 temporal patches per lead. These patches are linearly projected into a 384-dimensional embedding space through learned weight matrices. A learnable classification token (CLS) is added to the patch sequence, and learnable positional embeddings are added to preserve temporal ordering information, yielding a sequence of length 101.

The transformer encoder backbone consists of 6 identical layers, each comprising multi-head self-attention (MHSA) with 6 attention heads and a feed-forward network (FFN) with expansion ratio of 4 (hidden dimension: 1,536). Layer normalization precedes each sub-layer, and residual connections facilitate gradient flow. Stochastic depth regularization with linearly increasing drop path rate (0 to 0.1 across layers) is applied during training to improve generalization. The attention mechanism computes scaled dot-product attention as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (1)$$

where $d_k = 64$ is the dimension per head. Following the final transformer layer, the CLS token representation undergoes layer normalization and is projected through a linear classification head to produce logits for the four arrhythmia classes.

2D Spectro-Temporal Transformer Architecture The 2D-ViT processes time-frequency representations of ECG signals, using spectral information that captures the frequency modulations characteristic of complex arrhythmias. As illustrated in Fig. 4, raw ECG signals are first transformed into 2D representations using a Mel spectrogram or continuous wavelet transform (CWT). For Mel spectrograms, Short-Time Fourier Transform (STFT) is computed with 256-point FFT, 64-sample hop length, and Hann windowing, followed by projection onto 128 mel-scale frequency bins spanning [0.5 - 40 Hz]. The resulting power spectrograms are converted to decibel scale and normalized per lead, yielding representations of shape $R^{12\times128\times T}$ where T denotes the number of time frames. Alternatively, CWT-based scalograms are computed using Morlet wavelets across 128 logarithmically spaced scales, providing multi-resolution time-frequency decomposition with enhanced temporal resolution at higher frequencies corresponding to QRS complexes (Qiu et al. 2024).

These 2D representations (128×40 after temporal downsampling for computational efficiency) are partitioned into non-overlapping patches of size 16×4, creating 80 spatial

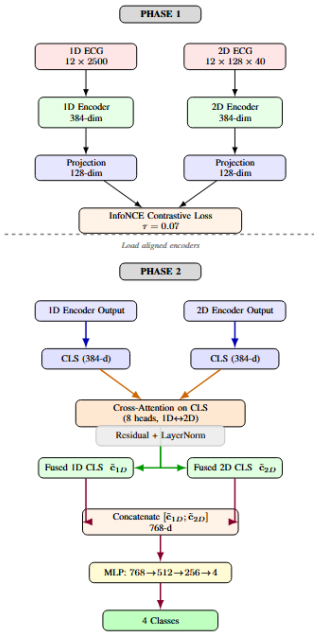

Figure 2: Hybrid Architecture

patches per lead. Each patch is flattened and linearly embedded in the same 384-dimensional space as the 2D ViT, with 2D sinusoidal positional encodings added to preserve spatial relationships in both frequency and time dimensions. A learnable CLS token is prepended, forming a sequence of length 81.

The transformer architecture mirrors the 1D variant exactly: 6 encoder layers with 6 attention heads each (head dimension: 64), FFN expansion ratio of 4, layer normalization, residual connections, and identical stochastic depth schedule. This architectural equivalence ensures that performance differences arise solely from input representation rather than model capacity disparities. The final CLS token representation is processed by the same linear classification head architecture, projecting to four arrhythmia class logits. Dropout regularization with probability 0.1 is applied in the attention and FFN modules during training, while an additional dropout layer with probability 0.2 precedes the classification head to prevent overfitting.

Two-Phase 1D–2D Fusion Architecture The proposed fusion framework integrates temporal and spectro-temporal ECG representations through a two-phase training strategy Fig. 2, designed to align and combine heterogeneous modalities in a shared latent space while preserving modality-specific information relevant to arrhythmia detection.

Phase 1: Cross-Modal Representation Alignment. In Phase 1, the 1D and 2D encoders are trained jointly using a symmetric InfoNCE objective to align their latent spaces. The 1D encoder processes raw ECG waveforms $\mathbf{x}^{(1D)} \in R^{12\times T}$, while the 2D encoder operates on time–frequency representations $\mathbf{x}^{(2D)} \in R^{12\times 128\times 40}$ derived via STFT–mel or CWT transforms. Each encoder produces a CLS token embedding $\mathbf{z}_i^{(1D)}, \mathbf{z}_i^{(2D)} \in R^{384}$, which are projected into a shared 128-dimensional space and L_2-normalized.

The contrastive alignment loss encourages paired 1D–2D embeddings of the same patient to be close while pushing apart mismatched pairs:

$$\mathrm{L}_{align} = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{\exp\left(\mathrm{sim}(\mathbf{p}_i^{(1D)},\mathbf{p}_i^{(2D)})/\tau\right)}{\sum_j \exp\left(\mathrm{sim}(\mathbf{p}_i^{(1D)},\mathbf{p}_j^{(2D)})/\tau\right)} + \log\frac{\exp\left(\mathrm{sim}(\mathbf{p}_i^{(2D)},\mathbf{p}_i^{(1D)})/\tau\right)}{\sum_j \exp\left(\mathrm{sim}(\mathbf{p}_i^{(2D)},\mathbf{p}_j^{(1D)})/\tau\right)}\right],$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes cosine similarity and τ=0.07 is the temperature parameter. This stage produces modality-specific encoders whose latent features are geometrically aligned, enabling seamless fusion in the subsequent phase.

Phase 2: Bidirectional Cross-Attention Fusion and Classification. In Phase 2, the aligned encoders are frozen or lightly fine-tuned, and their CLS embeddings $\mathbf{c}^{(1D)}, \mathbf{c}^{(2D)} \in R^{384}$ are combined through bidirectional cross-attention layers that learn dynamic inter-modality interactions. Each CLS token attends to the other modality:

$$\tilde{\mathbf{c}}^{(1D)} = \mathrm{CA}(\mathbf{q}{=}\mathbf{c}^{(1D)}, \mathbf{k}{=}\mathbf{c}^{(2D)}, \mathbf{v}{=}\mathbf{c}^{(2D)}),$$
$$\tilde{\mathbf{c}}^{(2D)} = \mathrm{CA}(\mathbf{q}{=}\mathbf{c}^{(2D)}, \mathbf{k}{=}\mathbf{c}^{(1D)}, \mathbf{v}{=}\mathbf{c}^{(1D)}),$$

where CA denotes multi-head cross-attention with eight heads (head dimension: 48). Residual connections and layer normalization produce the refined embeddings $\hat{\mathbf{c}}^{(1D)}$ and $\hat{\mathbf{c}}^{(2D)}$, which are concatenated to form a fused vector of size 768.

This fused representation is passed through a lightweight multilayer perceptron (MLP) classifier (768→512→256→4) with dropout (0.2 before the classifier) to output four arrhythmia class logits. Dropout with probability 0.1 is also applied in attention and feed-forward modules to improve generalization. The architectural equivalence across both encoders ensures that observed gains originate from fusion synergy rather than model capacity differences.

Saliency Visualization for GSVT. For interpretability, we generate saliency maps for the *GSVT* class to highlight regions contributing most to classification. For a test sample with true label GSVT, input–gradient saliency is computed as $\left|\partial f_{\mathrm{GSVT}}/\partial \mathbf{x}^{(1D)}\right|$ and $\left|\partial f_{\mathrm{GSVT}}/\partial \mathbf{x}^{(2D)}\right|$, aggregated across channels. The 1D saliency is overlaid on the ECG waveform, emphasizing high-importance temporal intervals, while the 2D saliency is visualized as a heatmap over the spectro-temporal plane, revealing discriminative frequency–time patterns associated with supraventricular tachycardia events. Additionally, the fusion module's learned modality weights α_{1D} and α_{2D} quantify the relative contribution of temporal versus spectro-temporal cues in the final decision, offering an interpretable view of model reasoning.

Statistical Significance Testing To rigorously assess the statistical significance of performance differences between models, we employed two complementary approaches:

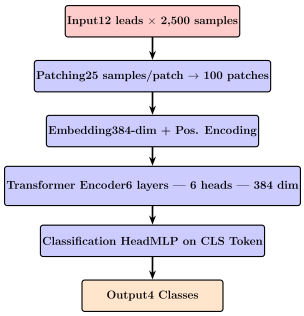

Figure 3: 1D Transformer Architecture

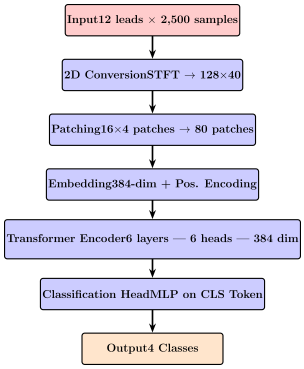

Figure 4: 2D Transformer Architecture

McNemar's Test : A paired non-parametric test that compares model disagreements on the same test samples. For each model pair (1D-ViT vs 2D-ViT, 1D-ViT vs Fusion, Fusion vs 2D-ViT), we constructed a 2×2 contingency table of correct/incorrect predictions and computed the chi-square statistic with continuity correction. The null hypothesis assumes no significant difference in classification performance.

Bootstrap Confidence Intervals : We generated 10,000 bootstrap samples from the test set (sampling with replacement) to compute 95% confidence intervals for all evaluation metrics. This approach accounts for sampling variability and provides robust estimates of performance bounds.

All statistical tests used $\alpha = 0.05$ significance level. P-values below 0.05 indicate statistically significant differences between models.

Implementation Details. Both encoders share identical depth (6 layers, 6 attention heads) and hidden size (d=384). Phase 1 uses projection heads of size 128 for contrastive alignment, while Phase 2 employs cross-attention fusion and a compact MLP classifier. The two-phase setup enforces robust modality alignment before supervised fusion, improving multi-view interpretability and stability across arrhythmia classes.

Merged Rhythms	Merged to	Sample
AFIB, AF	AFIB	9735
SVT, AT, SAAWR, ST, AVNRT, AVRT	GSVT	9772
SB	SB	16539
SR, SI	SR	8156

Table 2: Rhythms categories sample size

Metric	1D ViT	2D ViT	95% CI (1D)
Accuracy	96.52%	92.63%	[95.3, 97.6]
F1-Score	96.52%	92.60%	[95.3, 97.6]
Precision	96.52%	92.67%	[95.3, 97.6]
Recall	96.52%	92.63%	[95.3, 97.6]
AUC	99.16%	98.62%	[98.9, 99.4]
Specificity	98.90%	97.91%	[98.5, 99.2]
2D 95% CI: Acc [91.0, 94.2], AUC [98.2, 99.0]			
McNemar's: $\chi^2 = 13.50$, $p < 0.001$***			

Table 3: Performance comparison with 95% confidence intervals. Non-overlapping CIs and McNemar's test ($p < 0.001$) confirm statistically significant superiority of 1D ViT.

Results and Analysis

Comparative Performance of 1D and 2D Architectures

Our experiments compared two deep learning architectures for ECG classification:a 1D Vision Transformer (1D-ViT) and a 2D Vision Transformer (2D-ViT). Both models were trained for 50 epochs on the preprocessed ECG dataset comprising four cardiac rhythm classes. The 1D ViT demonstrated superior performance across all evaluation metrics.

As shown in Table 3, the 1D ViT achieved an overall accuracy of 96.52%, significantly outperforming the 2D model's 92.63% accuracy. Notably, the area under the ROC curve (AUC) for the 1D ViT reached 99.16%, compared to 98.62% for the 2D model, indicating excellent discriminative ability.

McNemar's Statistical test confirmed that the 1D-ViT's superiority over the 2D-ViT is highly statistically significant ($\chi^2 = 13.50$, $p < 0.001$). The 95% confidence interval for the accuracy difference is [+3.2%, +4.6%], and notably, the confidence intervals for 1D-ViT and 2D-ViT do not overlap, providing strong evidence of performance difference. Bootstrap analysis with 10,000 iterations confirms the robustness of these estimates to sampling variability.

Analysis of per-class performance (Table 4) revealed that the 1D ViT maintained consistently high performance across all four classes, with F1-scores ranging from 93.63% to 98.50%. In contrast, the 2D model showed more variability, particularly struggling with Class 1 (F1-score: 85.04%) compared to the 1D ViT's 95.97% for the same class.

Training dynamics revealed that both models converged effectively, though the 1D ViT demonstrated faster initial learning and better generalization (Figures 5 and 6). The training loss curves show smooth convergence with minimal overfitting, as evidenced by the close alignment between training and validation losses.

Class	1D F1	2D F1	Diff.	Rhythm
0	98.50%	98.56%	-0.06%	SB
1	95.97%	85.04%	+10.93%	SR
2	93.63%	88.07%	+5.56%	AFIB
3	95.87%	93.52%	+2.35%	GSVT

Table 4: Per-class F1-score comparison (SB: Sinus Bradycardia, SR: Sinus Rhythm, AFIB: Atrial Fibrillation, GSVT: Supraventricular Tachycardia). The 1D ViT shows major advantages for SR and AFIB detection.

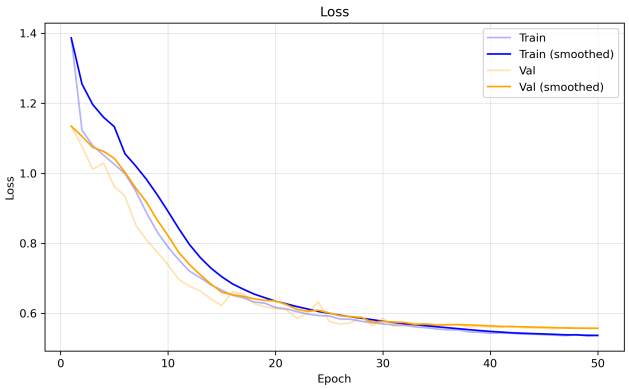

Figure 5: Training and validation loss curves for 2D model over 50 epoch.

The confusion matrices (Figures 7 and 8) further illustrate this difference: the 1D ViT exhibited fewer misclassifications, with the majority of errors concentrated between Classes 1 and 2 (48 misclassifications). The 2D model showed notably higher confusion between Classes 1 and 2 (168 misclassifications), suggesting difficulty in distinguishing between these similar rhythm patterns.

Multi-Modal Fusion Approach

Two-Phase Training Strategy The proposed multi-modal fusion framework integrates complementary information from 1D temporal ECG waveforms and 2D time–frequency representations through a two-phase training strategy. Phase 1 establishes cross-modal alignment between encoders via contrastive learning, while Phase 2 fine-tunes the aligned representations for supervised arrhythmia classification.

In Phase 1, the 1D and 2D encoders are trained with a symmetric InfoNCE objective to align their latent embeddings in a shared feature space, enforcing consistency between temporal and spectro-temporal views of the same cardiac activity. In Phase 2, the pretrained encoders are loaded and fused through bidirectional cross-attention modules, where each modality adaptively attends to the other, learning joint representations that capture both fine-grained waveform morphology and broader frequency structure. The concatenated embeddings are passed to an MLP classifier for final prediction.

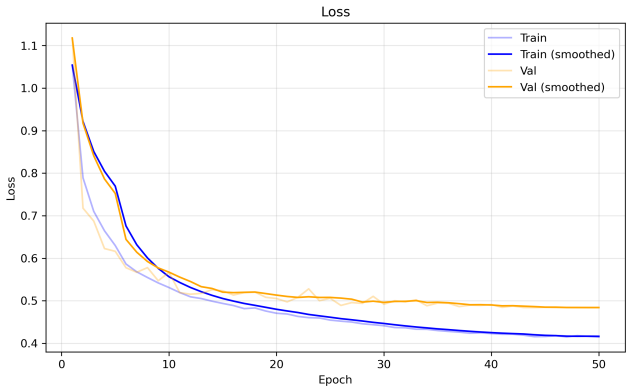

Figure 6: Training and validation loss curves for 1D ViT over 50 epochs

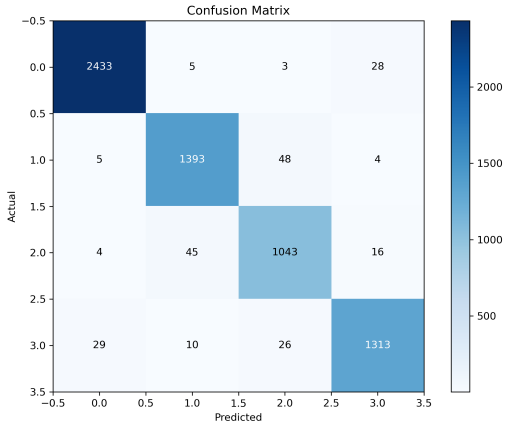

Figure 7: Confusion matrix for the 1D ViT showing concentrated misclassifications between Classes 1 and 2 (48 instances), indicating difficulty in distinguishing between similar rhythm patterns.

Model Performance Table 5 summarizes the quantitative results for the Fusion configuration. The model achieved a weighted F1-score of 94.03%, accuracy of 94.02%, and an overall AUC of 99.26%, demonstrating robust generalization across all arrhythmia classes. Per-class AUC values ranged from 0.987 for SR to 0.998 for SB, indicating high discriminative performance across distinct rhythm types.

Interpretability and Saliency Analysis To investigate how the models handle ambiguity, we generated input-gradient saliency maps for a challenging GSVT case (Fig. 9) where both single-modality models failed. The visualizations illustrate how each model's reasoning process differs. The 1D-ViT, focusing purely on temporal continuity, misclassified the segment as Sinus Rhythm (SR). Conversely, the 2D-ViT, analyzing the signal's spectral properties, misclassified the same segment as Sinus Bradycardia (SB). This example pinpoints a specific ambiguity that confuses both encoders when they are isolated. The Fusion model, however, correctly identified the GSVT rhythm. Its

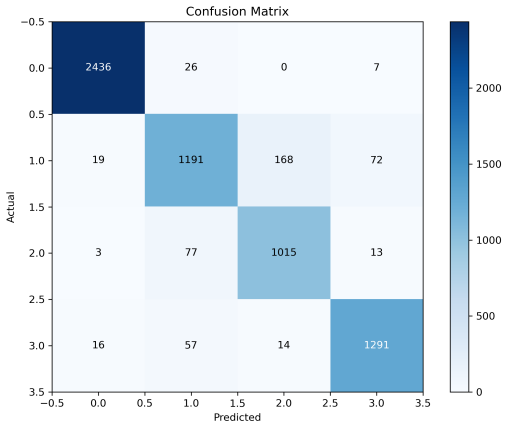

Figure 8: Confusion matrix for the 2D model showing notably higher confusion between Classes 1 and 2 (168 misclassifications), suggesting difficulty in distinguishing between these similar rhythm patterns.

Metric	Fusion
Accuracy	94.35%
Precision	94.15%
Recall (Sensitivity)	94.02%
Specificity	97.96%
F1-Score	94.03%
AUC	99.26%
Class 0 (SB) F1	97.73%
Class 1 (SR) F1	89.63%
Class 2 (AFIB) F1	91.49%
Class 3 (GSVT) F1	94.09%

Table 5: Performance of the Fusion model across all arrhythmia classes. The model maintains high discriminative ability with balanced sensitivity and specificity.

cross-attention mechanism assigned a balanced 57% weight to the 1D encoder and 43% to the 2D encoder. This demonstrates that the model integrated both temporal morphology and spectral cues, using their combined evidence to resolve the ambiguity that neither could overcome alone. This analysis provides the practical motivation for fusion: it is not intended to surpass the 1D-ViT in overall accuracy, but to enhance robustness by correctly resolving ambiguous edge cases where both single modalities fail. While such specific failures are rare—as evidenced by the 1D-ViT's high aggregate F1-score on this class (Table 4) —the fusion model's ability to correct them demonstrates its potential for improving clinical reliability

Overall Performance A comparative analysis across all evaluated architectures (Table 6) confirms that the **1D-ViT** achieved the highest overall classification performance. Its direct access to raw temporal information enables precise modeling of rhythm morphology and interval variability, yielding an accuracy and weighted F1-score of 96.52% and an AUC of 99.16%. Statistical validation via McNemar's test confirms 1D-ViT's significant superiority over both 2D-

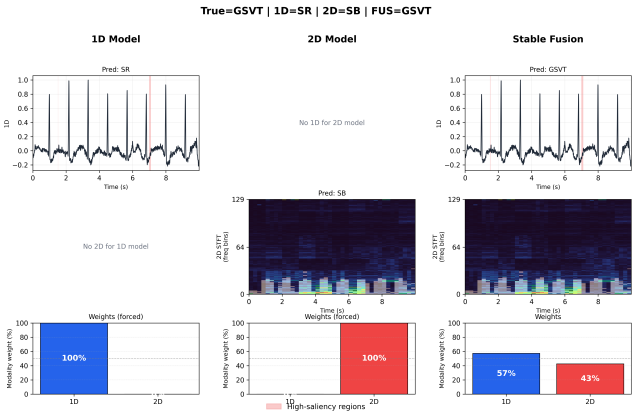

Figure 9: Saliency visualization for the **GSVT** class using the Fusion model. Up: 1D ViT highlighting temporal waveform peaks. Middle: 2D model emphasizing frequency-localized activations. Down: fused model combining both cues with adaptive modality weighting (1D: 57%, 2D: 43%).

ViT ($\chi^2 = 13.50$, $p < 0.001$) and Fusion ($\chi^2 = 4.90$, $p = 0.027$), with non-overlapping 95% confidence intervals [95.3%, 97.6%] vs [91.0%, 94.2%] and [92.8%, 95.7%] respectively.

The **2D-ViT** (92.63% accuracy, 98.62% AUC) suffered from partial loss of temporal resolution and phase information during STFT conversion, limiting its ability to differentiate rhythm-specific transitions.

The **Fusion model** achieved intermediate performance (94.35% accuracy, 99.26% AUC highest). Critically, while Fusion significantly underperformed 1D-ViT ($p = 0.027$), it showed *no significant improvement* over the simpler 2D-ViT ($p = 0.139$, $\chi^2 = 2.19$), indicating that added architectural complexity lacks statistical justification. Qualitative analysis revealed that fusion can help in edge cases—such as distinguishing GSVT from morphologically similar rhythms—by integrating spectral context.

Overall, statistical testing validates that 1D Transformers remain the most efficient and accurate choice for ECG arrhythmia classification. Multimodal fusion provides interpretive flexibility but no significant performance gain over simpler baselines, suggesting the 1D-ViT architecture for deployment scenarios prioritizing accuracy and efficiency.

Computational Considerations

From an efficiency standpoint, the 1D-ViT demonstrates the lowest computational cost, serving as the baseline (1.0×). The 2D-ViT requires slightly more processing (1.1×) due to spectrogram computation and increased input dimensionality. The Fusion model, which doubles encoder complexity and introduces cross-attention, demands approximately twice the computational load (2.0×). Given its higher inference cost without significant accuracy benefit, the Fusion architecture is best reserved for exploratory or interpretability-driven tasks rather than deployment in real-time systems.

Model	Accuracy	Precision	Recall	F1	AUC	Cost	p-value vs 1D
1D-ViT	**96.52%**	**96.52%**	**96.52%**	**96.52%**	99.16%	1.0×	–
Fusion	94.35%	94.15%	94.02%	94.03%	**99.26%**	2.0×	0.027*
2D-ViT	92.63%	92.67%	92.63%	92.60%	98.62%	1.1×	<0.001***

***$p < 0.001$, *$p < 0.05$ (McNemar's test). 95% CI: 1D [95.3, 97.6], Fusion [92.8, 95.7], 2D [91.0, 94.2] Fusion vs 2D: $p = 0.139$ (not significant). Bootstrap: 10,000 samples.

Table 6: Comprehensive model comparison with statistical validation. 1D-ViT significantly outperforms both alternatives.

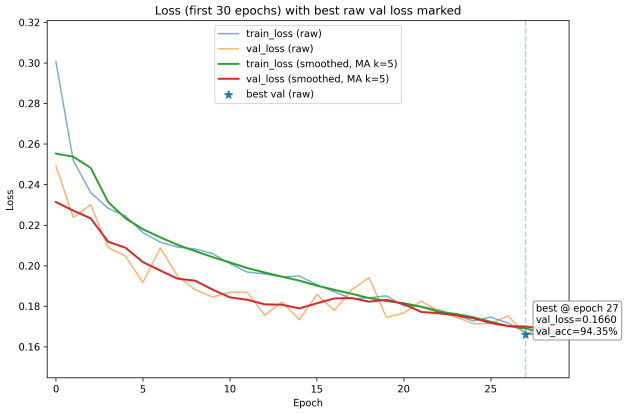

Figure 10: Training and validation loss for the Fusion model (first 30 epochs). The validation loss reaches its *raw* minimum at epoch 27; with an early-stopping, training halts at epoch 30 and the epoch 27 checkpoint is used for reporting. Smoothed curves are overlaid for readability.

Toward Clinical Translation: Human-AI Co-Design

Real-World Deployment Challenges: While our controlled evaluation on CPSC2018 demonstrates clear performance advantages for temporal representations, clinical deployment introduces additional complexity. Real-world ECG data exhibits greater variability in signal quality due to patient motion, electrode impedance drift, and diverse acquisition devices. Patient populations in clinical settings span broader age ranges, include comorbidities affecting cardiac electrophysiology, and present with arrhythmia distributions that may differ substantially from research datasets. The 1D-ViT's superior performance on CPSC2018 suggests robustness to these challenges, as temporal fidelity preservation should be advantageous regardless of signal perturbations. However, prospective validation on clinical data streams is essential before deployment.

Performance-Interpretability Trade-offs for Clinical Adoption: Our findings reveal a critical design decision for clinical AI systems: the 1D-ViT achieves highest accuracy (96.52%) but provides single-modality reasoning, while the Fusion model sacrifices performance (94.35%) for enhanced interpretability through joint temporal-spectral saliency analysis. This trade-off has direct implications for human-in-the-loop workflows. In high-stakes clinical scenarios requiring physician oversight, the Fusion model's ability to highlight both temporal waveform features and frequency-domain anomalies may facilitate faster expert verification, even if overall accuracy is lower. Conversely, for automated screening workflows with lower clinical risk, the simpler 1D-ViT's superior accuracy and computational efficiency (1.0× cost vs. 2.0× for Fusion) makes it the preferred choice.

Co-Design with Clinicians: Future work should involve cardiologists in defining deployment requirements through participatory design. Key considerations include: (1) acceptable false positive/negative rates for specific arrhythmias (e.g., life-threatening AFIB vs. benign Sinus Bradycardia), (2) latency constraints for real-time monitoring vs. retrospective analysis, (3) interpretability needs varying by clinical context (emergency department vs. outpatient screening), and (4) integration with existing electronic health record systems. Such co-design ensures AI models align with clinical workflows rather than imposing technically optimal but clinically impractical solutions.

Deployment Requirements: Clinical systems must balance multiple constraints. Latency requirements vary: ambulatory monitors require <1s inference for real-time alerts, while retrospective analysis tolerates higher latency for improved accuracy. Our models' computational costs (1D: 1.0× baseline, Fusion: 2.0×) are acceptable for cloud-based analysis but may challenge edge deployment on wearable devices. Power and memory constraints on such devices favor the simpler 1D architecture, suggesting differentiated deployment: lightweight 1D models on wearables for continuous monitoring, with cloud-based Fusion models for detailed analysis when connectivity permits.

This work demonstrates that the choice between 1D and multimodal architectures is not purely technical—it reflects a fundamental design decision about the role of AI in clinical care. By providing rigorous performance comparisons alongside interpretability analysis, we enable informed decisions that balance accuracy, explainability, and clinical workflow integration.

Conclusion

This study conducted a controlled comparison of three Transformer-based architectures—1D-ViT, 2D-ViT, and a cross-attention Fusion model—for ECG arrhythmia classification on the CPSC2018 dataset. All models were trained under identical conditions to isolate the impact of signal representation. Results confirm that preserving temporal fidelity yields the highest classification performance. The 1D-ViT achieved 96.5% accuracy and 99.16% AUC, outperforming both 2D and Fusion counterparts. Spectro-temporal

approaches offer complementary interpretive cues but not predictive advantage. Future work should investigate hybrid approaches that embed frequency-aware features within temporal Transformers to preserve interpretability without compromising accuracy or efficiency.

Limitations and Future Work

Alternative Time-Frequency Representations: This study employed Short-Time Fourier Transform (STFT) as the primary 2D representation. We acknowledge that 2D model performance is sensitive to the choice of transform and its parameters. Alternative representations such as Continuous Wavelet Transform (CWT) or Mel-spectrograms may yield different results, as they offer distinct time-frequency tradeoffs. We initially explored CWT-based scalograms, which provide superior temporal resolution at higher frequencies relevant to QRS complex detection. However, computational constraints limited comprehensive evaluation of CWT architectures. Future work should systematically compare multiple 2D representations (STFT, CWT, Mel-spectrograms) under identical experimental conditions to determine if spectro-temporal approaches can match or exceed 1D temporal performance with optimized transform selection.

Dataset Generalization: Our conclusions are drawn from the CPSC2018 dataset, a large publicly available repository of 12-lead ECG recordings. While this dataset provides robust evaluation, performance may vary on clinical datasets with different arrhythmia distributions, patient demographics (age, comorbidities), acquisition conditions (ambulatory vs. hospital), and higher levels of signal noise (motion artifacts, electrode placement variability). The controlled nature of CPSC2018 may not fully represent real-world clinical complexity. Future validation on diverse clinical datasets, including those with rare arrhythmias, pediatric populations, and challenging signal quality conditions, is essential for clinical deployment.

References

Dong, Y.; Zhang, M.; Qiu, L.; Wang, L.; and Yu, Y. 2023. An Arrhythmia Classification Model Based on Vision Transformer with Deformable Attention. *Micromachines*, 14(6): 1155.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. ArXiv:2010.11929 [cs].

Duan, J.; Wang, Q.; Zhang, B.; Liu, C.; Li, C.; and Wang, L. 2022. Accurate detection of atrial fibrillation events with R-R intervals from ECG signals. *PLoS ONE*, 17(8): e0271596.

Mhetre, P. S.; Mhetre, S. L.; and Lokhande, S. S. 2025. Automated ECG Signal Analysis with 1D CNN: A Deep Learning Approach.

Narotamo, H.; Dias, M.; Santos, R.; Carreiro, A. V.; Gamboa, H.; and Silveira, M. 2024. Deep learning for ECG classification: A comparative study of 1D and 2D representations and multimodal fusion approaches. *Biomedical Signal Processing and Control*, 93: 106141.

Payal Kohli. 2023. ECG Signal Quality: A Practical Guide for ECG Readings.

Qiu, C.; Li, H.; Qi, C.; and Li, B. 2024. Enhancing ECG classification with continuous wavelet transform and multi-branch transformer. *Heliyon*, 10(5): e26147.

Rifat, A. 2022. The State of Cardiovascular Disease in G20+ Countries. *HPHR Journal*, HSIL(2022).

Song, M.-S.; and Lee, S.-B. 2024. Comparative study of time-frequency transformation methods for ECG signal classification. *Frontiers in Signal Processing*, 4: 1322334.

Vaid, A.; Jiang, J.; Sawant, A.; Lerakis, S.; Argulian, E.; Ahuja, Y.; Lampert, J.; Charney, A.; Greenspan, H.; Narula, J.; Glicksberg, B.; and Nadkarni, G. N. 2023. A foundational vision transformer improves diagnostic performance for electrocardiograms. *npj Digital Medicine*, 6(1): 108.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2023. Attention Is All You Need. ArXiv:1706.03762 [cs].

Wu, Z.; and Guo, C. 2025. Deep learning and electrocardiography: systematic review of current techniques in cardiovascular disease diagnosis and management. *BioMedical Engineering OnLine*, 24: 23.

Zacarias, H.; Marques, J. A. L.; Felizardo, V.; Pourvahab, M.; and Garcia, N. M. 2024. ECG Forecasting System Based on Long Short-Term Memory. *Bioengineering*, 11(1): 89.

Zhang, Y.; Yi, J.; Chen, A.; and Cheng, L. 2023. Cardiac arrhythmia classification by time–frequency features inputted to the designed convolutional neural networks. *Biomedical Signal Processing and Control*, 79: 104224.

Zheng, J.; Zhang, J.; Danioko, S.; Yao, H.; Guo, H.; and Rakovski, C. 2020. A 12-lead electrocardiogram database for arrhythmia research covering more than 10,000 patients. *Scientific Data*, 7(1): 48.