# DeepArrhythmia: Segment-Contextualized ECG Arrhythmia Classification via Selective Evidence Acquisition

**Jiahui Li, Ruili Fang, Zishuai Liu, WenZhan Song, Jin Lu, Fei Dou**
University of Georgia
{jl57095,ruili.fang,zishuai.liu,wsong,Jin.Lu,Fei.Dou}@uga.edu

## Abstract

Beat-level Electrocardiography (ECG) arrhythmia detection aims to assign an arrhythmia class to each beat in a recording, yet many existing systems treat beats as isolated local instances. This is limiting because beat labels often depend on multi-beat rhythm context, including timing, compensatory pauses, and beat-to-beat morphological consistency. We present **DeepArrhythmia**, a tool-grounded multimodal framework for segment-contextualized beat-level ECG arrhythmia classification. Given a multi-beat ECG segment, DeepArrhythmia combines the raw ECG signal and a rendered waveform image, localizes R peaks to identify beat instances, and produces structured beat-level predictions. The framework decouples physiological measurement from evidence integration using specialized tools for beat localization, numerical rhythm–morphology extraction, and morphology-focused textual analysis. DeepArrhythmia uses segment-level confidence to route between minimal and rich evidence states, since richer physiological evidence is not uniformly useful. This agentic design integrates rhythm context, explicit physiological grounding, and selective evidence acquisition for decision making.

## 1 Introduction

Cardiac arrhythmias are common rhythm disorders associated with cardiovascular disease [Koppad, 2021] and are routinely detected and assessed using electrocardiography (ECG). As a non-invasive recording of cardiac electrical activity, ECG remains central to clinical arrhythmia evaluation, where recordings have traditionally been reviewed by experts to identify and label abnormal beats [Hong et al., 2020]. This clinical workflow has motivated extensive research on automated beat-level ECG arrhythmia classification, whose basic goal is to assign an arrhythmia class to each beat in an ECG recording [Ebrahimi et al., 2020]. Despite substantial progress, many existing beat-level systems still treat each beat primarily as a local classification instance, using isolated beat crops or short beat-centered windows as input.

This local treatment is limited because beat-level arrhythmia classification is beat-instance prediction, but not purely beat-local. Each prediction corresponds to an individual beat instance, typically anchored by its R-peak location. The evidence needed to classify a beat often extends beyond the target complex to the surrounding rhythm context. As illustrated in Figure 1, ventricular ectopic beats (VEBs) and supraventricular ectopic beats (SVEBs) can both occur prematurely and produce a post-ectopic pause, so timing alone may be ambiguous. Their distinction often depends on contextual morphology: VEBs typically show a wide and aberrant QRS complex, whereas SVEBs tend to preserve a QRS morphology closer to neighboring normal beats. These examples motivate segment-contextualized beat-level classification, where each beat receives its own label while the decision can draw on morphology, rhythm timing, and beat-to-beat consistency within a multi-beat ECG segment.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

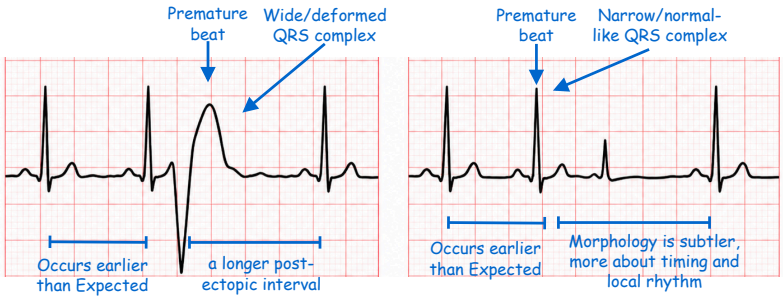



Figure 1: VEB/SVEB discrimination requires morphology in context.

This context dependence exposes two limitations in existing automated beat-level systems. First, isolated-beat and short-window methods preserve local morphology but underuse rhythm context, such as preceding and following R–R intervals, compensatory pauses, and cross-beat morphological consistency. Second, longer-context end-to-end models may process more signal context, but their predictions are often not explicitly grounded in physiological anchors or intermediate measurements. As a result, it is difficult to verify whether a decision is supported by recognizable ECG cues such as R–R dynamics, QRS morphology, P-wave behavior, or compensatory pauses. A natural way to address this limitation is to decouple physiological measurement from evidence integration: specialized ECG tools can provide precise beat localization and structured rhythm–morphology measurements, while a multimodal model can integrate heterogeneous evidence for beat-level decision making. However, richer physiological evidence should not be acquired indiscriminately. For easy segments, additional measurements may add computation without improving accuracy; for noisy or weakly informative signals, auxiliary evidence can even mislead the model. Thus, the key question is not only how to incorporate physiological evidence, but when such evidence should be acquired.

We propose **DeepArrhythmia**, a tool-grounded multimodal framework for segment-contextualized beat-level ECG arrhythmia classification. Given a multi-beat ECG segment, DeepArrhythmia uses both the raw ECG signal and a rendered ECG waveform image as complementary inputs. A Peak Detector first localizes R peaks to identify beat instances and provide temporal anchors for evidence alignment and structured output. The central multimodal agent then produces beat-level predictions under a minimal evidence state using the signal, image, and detected beat anchors. When the segment-level confidence indicates that the current evidence may be insufficient, DeepArrhythmia selectively invokes a Feature Extractor and a Morphology Analyzer to obtain explicit numerical rhythm–morphology measurements and textual morphology evidence before producing refined beat-level predictions. This design combines multi-beat rhythm context, structured physiological grounding, and confidence-guided selective evidence acquisition within a unified agentic framework. Building on the above, we make the following contributions:

1. We reformulate beat-level arrhythmia classification as segment-contextualized, peak-aligned prediction. Rather than classifying isolated beat crops, DeepArrhythmia predicts labels for R-peak-aligned beats within multi-beat ECG segments, allowing each decision to combine local morphology with surrounding rhythm context.
2. We introduce a tool-grounded multimodal ECG agent that separates physiological measurement from evidence integration. Specialized tools expose detected peaks, numerical rhythm–morphology measurements, and morphology-focused textual descriptions as structured intermediate evidence that is used both for beat-level prediction and for interpretation and error analysis.
3. We propose confidence-guided selective evidence acquisition. Motivated by the non-uniform marginal value of physiological evidence, DeepArrhythmia uses confidence as an evidence gate to balance under-evidenced decisions on difficult rhythm-dependent segments against over-evidenced decisions where weak auxiliary measurements add cost or noise when applied indiscriminately. The acquisition policy is learned through threshold-induced routed supervised fine-tuning, so the agent learns *when* richer evidence is worth acquiring.

Together, these elements form a single framework that couples segment-level rhythm context, tool-grounded physiological evidence, and confidence-guided evidence acquisition for beat-level ECG classification.

## 2 Related Work

**Beat-Level Arrhythmia Classification and Rhythm Context** Automated beat-level ECG arrhythmia classification has been studied extensively using both feature-based and end-to-end learning approaches. Feature-based methods typically extract handcrafted rhythm and morphology descriptors, such as RR intervals, R-peak amplitudes, and beat-shape features, and then apply conventional machine learning classifiers [Limam and Precioso, 2017, Mondéjar-Guerra et al., 2019]. In contrast, end-to-end approaches directly process raw ECG signals or transformed representations, such as time–frequency images, using deep neural networks [Huang et al., 2019, Rajkumar et al., 2019, Hannun et al., 2019, Zhao et al., 2019]. These approaches achieve strong predictive performance, but two limitations remain common. Many operate on isolated beat crops or short windows, underusing surrounding rhythm context. Some performance may rely on patient-specific or intra-patient evaluation, limiting evidence of subject-disjoint generalization. In contrast, we study beat-level arrhythmia classification under subject-disjoint splits using 10-second segments that preserve rhythm.

**Tool-Grounded Multimodal ECG Reasoning and Structured Evidence** Recent ECG-specific multimodal large language models (MLLMs) integrate ECG signals, rendered ECG images, and textual reasoning to support ECG understanding and improve interpretability. GEM [Lan et al., 2025] aligns time-series and image representations for grounded ECG interpretation, while PULSE [Liu et al., 2024] and ECG-R1 [Jin et al., 2026] pursue ECG-oriented vision–language reasoning. Although MLLMs are effective at integrating heterogeneous evidence, they remain limited in fine-grained numerical tasks, including precise event localization and explicit physiological measurement [Arbuzov et al., 2025, Fons et al., 2024, Li et al., 2025]. ECG-Agent explores tool calling for multi-turn ECG interaction [Chung et al., 2026]; however, it primarily presents tool outputs to improve interpretability rather than incorporating them as structured intermediate evidence for downstream decision-making. In contrast, DeepArrhythmia separates physiological measurement from evidence integration, enabling arrhythmia classification that is both more accurate and more interpretable.

**Selective Evidence Acquisition and Adaptive Tool Use** Recent agentic frameworks Yao et al. [2023], Jin et al. [2025], Chung et al. [2026] enable models to selectively invoke the tools most appropriate for downstream tasks. However, these approaches often depend heavily on the model's prior knowledge and on carefully designed prompts. In expert domains such as arrhythmia classification, such reliance can limit reliable tool use, particularly when the triggering conditions are implicit, as in variations in the complexity of ECG time series. DeepArrhythmia instead explicitly leverages the relationship between model confidence and predictive accuracy, using confidence as an evidence gate for selective tool invocation and thereby balancing predictive performance against computational cost.

## 3 Method

**Problem Formulation** Let $x = (x_{\mathrm{ts}}, x_{\mathrm{img}})$ denote a time-normalized ECG context segment of fixed duration $T = 10$ seconds. For a dataset $d$ with sampling frequency $f_d$, the signal view is represented on the native sampling grid as $x_{\mathrm{ts}} \in \mathbb{R}^{C \times L_d}$, with $L_d = \lfloor T f_d \rceil$, where $C$ is the number of ECG input channels or leads. Thus, $L_d$ may vary across datasets, but every segment spans the same physical $T$-second context. The rendered waveform view $x_{\mathrm{img}} \in \mathbb{R}^{H \times W \times 3}$ is generated from the same $T$ interval. Each segment $x$ contains a variable number of beat instances $B_x = \{b_{x,k}\}_{k=1}^{K_x}$, where $K_x$ depends on the subject's heart rate and rhythm within the $T$ window. The goal is to assign an arrhythmia label $y_{x,k} \in \mathcal{Y}$ to each beat instance $b_{x,k}$, where $\mathcal{Y} = \{N, S, V, F\}$ denotes normal, supraventricular ectopic, ventricular ectopic, and fusion beats under the AAMI mapping. The model outputs a structured beat-level sequence $\hat{Y}(x) = \{[a_{x,k} : \hat{y}_{x,k}]\}_{k=1}^{K_x}$, where $a_{x,k}$ is the temporal anchor used to identify beat $b_{x,k}$ in the serialized output. In practice, these anchors are instantiated as R-peak sample positions by the Peak Detector described below. The task is therefore beat-level classification conditioned on segment-level context, rather than sample-wise labeling or segment-level classification. The overview of the DeepArrhythmia framework is shown in Figure 2.

### 3.1 Tool-Grounded Evidence Interfaces and Multimodal Integration

DeepArrhythmia separates physiological measurement from multimodal evidence integration. Specialized ECG tools expose structured physiological evidence, while the central multimodal agent consumes this evidence together with signal and image tokens to produce peak-aligned beat predic-

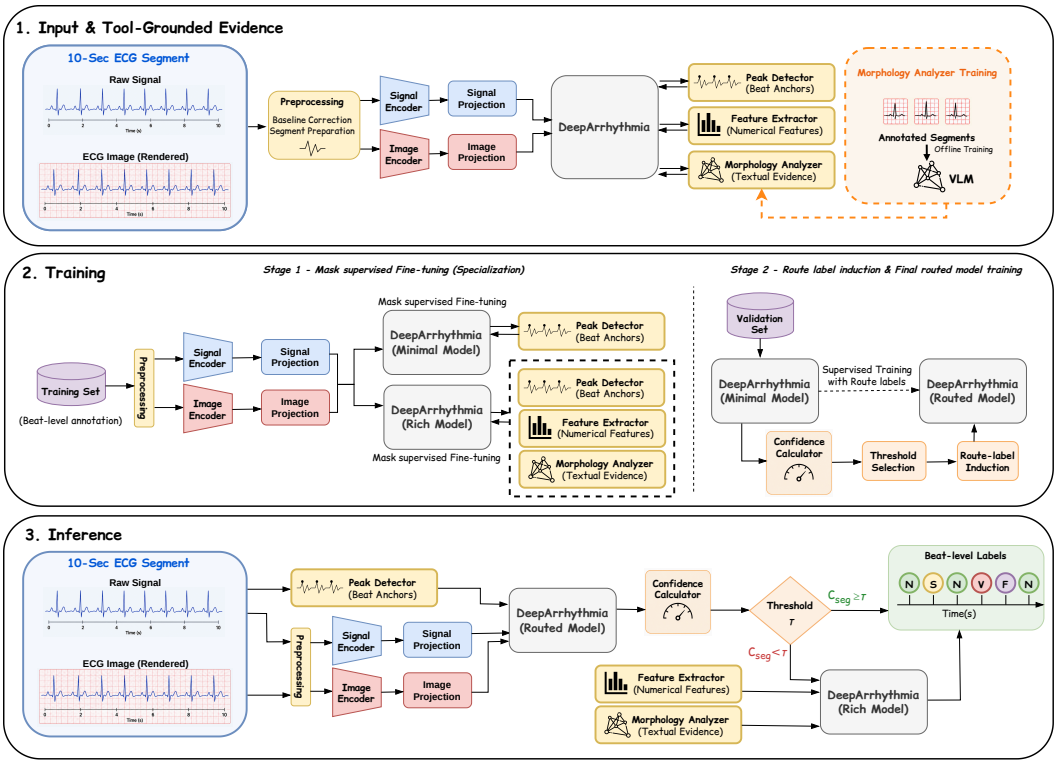



Figure 2: **Overview of the DeepArrhythmia framework. (1)** Multimodal ECG inputs with three tool-grounded evidence sources. **(2)** Two-stage mask-SFT yields Minimal, Rich, and Routed specialists. **(3)** At inference, confidence $C_{\text{seg}}$ vs. threshold $\tau$ routes between minimal and rich evidence.

tions. The central agent receives ECG signal tokens, rendered ECG image tokens, and serialized tool outputs. Signal and image views are encoded separately and projected into the language-model embedding space. Tool interactions are represented with explicit `<Call>` and `<Output>` spans. The final answer is emitted as structured beat-level text, e.g., $[a_1 : N]\ [a_2 : S]\ \cdots\ [a_K : V]$.

**Beat-anchor interface.** The *Peak Detector* instantiates the abstract beat anchors $A_x$ by localizing candidate R peaks from $x_{\text{ts}}$. These anchors determine where beat-level labels are emitted, and they also define the alignment for numerical feature extraction and morphology analysis.

**Numerical physiology interface.** Given $x_{\text{ts}}$ and $A_x$, the *Feature Extractor* produces per-beat rhythm and morphology measurements, $Z_{\text{feat}}(x) = T_{\text{feat}}(x_{\text{ts}}, A_x)$, including RR intervals, normalized RR intervals, R-peak amplitude, higher-order statistics, and numerical morphology descriptors. These measurements provide explicit numerical grounding for rhythm and morphology cues.

**Textual morphology interface.** The *Morphology Analyzer* produces morphology-aware textual evidence, $Z_{\text{morph}}(x) = T_{\text{morph}}(x_{\text{img}}, A_x, Z_{\text{feat}})$, from beat-centered ECG image crops and auxiliary context. It is trained by split-safe teacher distillation on training subjects only and receives no ground-truth label at inference time. Its output is used as supportive morphology-aware evidence rather than as a fully independent explanation. The same intermediate outputs are both consumed by the central agent for prediction and retained as auditable evidence for interpretation and error analysis. Thus, the central agent acts as an evidence integrator and decision maker, rather than a standalone ECG signal processor.

### 3.2 Confidence-Guided Selective Evidence Acquisition

DeepArrhythmia operationalizes selective evidence acquisition as a routing decision between two physiological evidence states. After beat anchors have been obtained, the model first predicts under a minimal peak-guided evidence state and then uses the *Confidence Calculator* to decide whether the current evidence appears sufficient or whether optional richer evidence should be acquired. The

Confidence Calculator is the fourth tool interface in DeepArrhythmia, but unlike the Peak Detector, Feature Extractor, and Morphology Analyzer, it does not produce new physiological measurements. Instead, it acts as an evidence-gating tool that controls the transition between evidence states.

We define the minimal evidence state contains the synchronized signal view, rendered waveform view, and beat anchors: $E_{\min}(x) = \{x_{\text{ts}}, x_{\text{img}}, A_x\}$, where $A_x = \{a_{x,k}\}_{k=1}^{K_x}$ denotes the beat anchors returned by the Peak Detector. The rich evidence state augments this minimal state with structured numerical rhythm–morphology measurements and textual morphology evidence: $E_{\text{rich}}(x) = E_{\min}(x) \cup \{Z_{\text{feat}}(x), Z_{\text{morph}}(x)\}$. Thus, the routing problem is not whether to classify the segment, but whether to classify it using only $E_{\min}$ or to acquire $E_{\text{rich}}$ before the final prediction.

The model first generates an initial structured prediction under $E_{\min}$. For each beat anchor $a_{x,k}$, we obtain a class posterior $p_{x,k} \in [0,1]^{|\mathcal{Y}|}$ from the language-model logits at the corresponding label-generation position in the structured output. Concretely, we apply a softmax over the canonical class tokens $\mathcal{Y} = \{N, S, V, F\}$, and the initial predicted label is $\hat{y}_{x,k}^{\min} = \arg\max_{y \in \mathcal{Y}} p_{x,k,y}$. The Confidence Calculator converts these beat-level posteriors into a segment-level evidence-sufficiency score. The beat-wise confidence is $c_{x,k} = \max_{y \in \mathcal{Y}} p_{x,k,y}$, and the segment-level confidence is computed by averaging over the $K_x$ beats in the segment: $C_{\text{seg}}(x) = \frac{1}{K_x} \sum_{k=1}^{K_x} c_{x,k}$. We route once per segment rather than once per beat, because the optional Feature Extractor and Morphology Analyzer operate on segment-level rhythm and morphology context. Appendix E compares alternative segment-level confidence statistics, and Appendix F analyzes threshold sensitivity, calibration, and routing rates.

Given a dataset-specific threshold $\tau_d$, DeepArrhythmia follows the policy

$$\pi(x) = \begin{cases} E_{\min}(x), & C_{\text{seg}}(x) \geq \tau_d, \\ E_{\text{rich}}(x), & C_{\text{seg}}(x) < \tau_d. \end{cases}$$

If $C_{\text{seg}}(x) \geq \tau_d$, the model returns the initial minimal-evidence prediction $\hat{Y}_{\min}(x) = \{[a_{x,k} : \hat{y}_{x,k}^{\min}]\}_{k=1}^{K_x}$. If $C_{\text{seg}}(x) < \tau_d$, the model invokes the optional evidence-producing tools, obtains $Z_{\text{feat}}$ and $Z_{\text{morph}}$, and produces a refined prediction under $E_{\text{rich}}$: $\hat{Y}_{\text{rich}}(x) = \{[a_{x,k} : \hat{y}_{x,k}^{\text{rich}}]\}_{k=1}^{K_x}$.

This policy treats confidence as a practical proxy for evidence sufficiency, not as a direct estimator of the true utility of additional tools. The design is intended to balance two failure modes: under-evidence, where difficult rhythm-dependent segments are classified from insufficient physiological context, and over-evidence, where optional measurements are applied uniformly even when they may be redundant, costly, or noisy. Appendix L provides a complete example of this inference pipeline.

### 3.3 Two-stage Training and Implementation

We train DeepArrhythmia with two stages of masked supervised fine-tuning over serialized tool-use trajectories. Each trajectory contains generated tool-call tokens, externally returned tool-output spans, and a final structured beat-level answer. For each dataset, we first construct ***subject-disjoint training and test partitions***. The test subjects are never used for tool distillation, threshold selection, or supervised fine-tuning. We further split the training subjects into a Stage 1 optimization split $D_1$ and a held-out Stage 2 route-induction split $D_2$ using a 9:1 ratio. The split $D_2$ is held out from Stage 1 training and is used only to select the confidence threshold and construct routed SFT trajectories. The selected threshold is fixed before test-time evaluation. We apply shift-based segment re-anchoring on the training split to mitigate class imbalance. This augmentation preserves the fixed 10-second context horizon while increasing exposure to minority-class beats at different positions within the segment. Details are provided in Appendix B.1.

**Masked SFT objective.** Let $\tau^* = (\tau_1^*, \ldots, \tau_N^*)$ denote a target token trajectory and let $m_n \in \{0,1\}$ indicate whether token $\tau_n^*$ is supervised. The masked SFT loss is $\mathcal{L}_{\text{SFT}}(\theta) = -\sum_{(x,\tau^*)} \sum_{n=1}^{N} m_n \log p_\theta(\tau_n^* \mid x, \tau_{<n}^*)$. We set $m_n = 1$ for generated tool-call tokens and final-answer tokens, and $m_n = 0$ for externally returned tool-output spans and other context tokens. Thus, the model is trained to decide which tools to call and to generate the final structured beat-level prediction, while deterministic tool outputs are consumed rather than imitated.

**Stage 1: evidence-state specialization.** Stage 1 trains two branch-specific agents corresponding to the two evidence states. The minimal-evidence agent $M_{\min}$ is trained on short trajectories that

call the Peak Detector, receive beat anchors $A_x$, and then emit the final structured prediction under $E_{\min}(x)$ The rich-evidence agent $M_{\text{rich}}$ is trained on trajectories that call the Peak Detector, Feature Extractor, and Morphology Analyzer before producing the final prediction under $E_{\text{rich}}(x)$ This stage teaches the model the output syntax, the alignment between beat anchors and class labels, and how to predict under both minimal and rich evidence budgets before adaptive routing is introduced.

**Stage 2: confidence-gated routed SFT.** Stage 2 trains the routed agent $M_{\text{route}}$ to follow the confidence-gated acquisition policy. We first run $M_{\min}$ on the held-out route-induction split $D_2$. For each segment, the Confidence Calculator computes the segment-level score from the minimal-evidence class posteriors. We also run $M_{\text{rich}}$ on the same split to obtain the prediction that would result from acquiring full rich evidence. For dataset $d$, we sweep candidate thresholds and select $\tau_d = \arg\max_\tau \text{MicroF1}\left(\hat{Y}_\tau; D_2\right)$, where $\hat{Y}_\tau(x)$ is either $\hat{Y}_{\min}(x)$ or $\hat{Y}_{\text{rich}}(x)$. The selected threshold induces deterministic route labels on $D_2$: segments above the threshold stop after minimal evidence, whereas segments below the threshold acquire the Feature Extractor and Morphology Analyzer outputs before final prediction. These route labels are threshold-induced evidence-acquisition decisions, not normal-versus-abnormal labels.

We then construct routed trajectories for $M_{\text{route}}$. Each trajectory calls the Peak Detector, produces an initial minimal-evidence prediction, queries the Confidence Calculator, and receives the resulting confidence score. If $C_{\text{seg}}(x) \geq \tau_d$, the trajectory terminates with the minimal prediction. If $C_{\text{seg}}(x) < \tau_d$, the trajectory continues by invoking the Feature Extractor and Morphology Analyzer and then emits the rich-evidence prediction. We initialize $M_{\text{route}}$ from $M_{\min}$ to preserve the minimal-evidence prediction behavior and then fine-tune it on these routed trajectories. At inference time, the same threshold $\tau_d$ is fixed and applied without access to test labels.

**Implementation.** We instantiate Qwen3.5-4B as Multimodal LLM backbone and use pretrained ECG-CoCa encoder for signal modality [Zhao et al., 2025]. The ECG-CoCa encoder is kept frozen during DeepArrhythmia training, so optimization focuses on multimodal evidence integration, tool-use generation, and the evidence-acquisition policy. Details are in Appendix A.

## 4 Experiments

**Datasets and baselines.** We evaluate model performance on four electrocardiogram (ECG) datasets: MIT-BIH Arrhythmia [Moody and Mark, 2001], MIT-BIH Supraventricular [Greenwald et al., 1990], INCART [Tihonenko et al., 2008], and VitalDB Arrhythmia [Eun et al., 2026]. We compare DeepArrhythmia against four baseline families. The first is a conventional machine-learning baseline, SVM [Mondéjar-Guerra et al., 2019]. The second is a set of deep-learning classifiers: 2D CNN [Limam and Precioso, 2017], STFT 2D CNN [Huang et al., 2019], 1D ResNet [Hannun et al., 2019], Transformer [Hu et al., 2022], xECG [Lunelli et al., 2025], and TCN [Ingolfsson et al., 2021]. The third is a set of general-purpose multimodal language models: Gemini 3.1 [Team et al., 2023], GPT-5.4 [Singh et al., 2025], Qwen3.5-4B [Bai et al., 2023], and Gemma-3-27B [Gemma Team, 2025]. The fourth is a set of ECG-oriented vision-language models: PULSE [Liu et al., 2024], GEM [Lan et al., 2025], and ECG-R1 [Jin et al., 2026]. For language-model-based baselines, we append the outputs of the Peak Detector and Feature Extractor to the input prompt to ensure a fair comparison under matched tool-derived evidence. The Morphology Analyzer is excluded because it is trained with dataset-specific supervision, which would compromise the fairness of this comparison.

**Non-agentic baseline ablation.** We evaluate matched non-agentic fusion baselines to isolate the contribution of each modality. Specifically, we replace the decoder-only backbone with a fusion classifier that consumes the same input ingredients but performs no tool calling or sequential decision making. The model encodes each 10-second ECG segment with the ECG-CoCa tower and uses pre-computed Qwen3.5-4B embeddings for the rendered ECG image and text evidence. For each detected R peak, it pools a beat-local ECG representation, concatenates it with the global ECG representation, image embedding, structured peak-derived numerical features, and optional morphology/prompt text embedding, and then applies a supervised fusion head to predict a label for each beat.

**Experiment setting.** For data partitioning, we adopt the DS1/DS2 protocol [Mondéjar-Guerra et al., 2019], which separates training and test sets by subject identity with balanced numbers of IDs. This split prevents subject overlap between training and testing and reduces information leakage. Following prior ECG annotation practice, we exclude class $Q$, which denotes unknown or unclassifiable beats.

Table 1: Performance comparison across datasets using Macro-F1 and Micro-F1. Best results are shown in bold, and second-best results are underlined. For machine learning and deep learning models, we report the mean over three random seeds. For VLMs, we use a decoding temperature of 0. For datasets with absent classes in the test split, Macro-F1 is averaged over classes present in the ground truth.

| | MIT-BIH | | MIT-BIH-SUP | | INCART | | VitalDB | |
|---|---|---|---|---|---|---|---|---|
| **Model** | **Ma** | **Mi** | **Ma** | **Mi** | **Ma** | **Mi** | **Ma** | **Mi** |
| *DeepArrhythmia* | | | | | | | | |
| Rich-specialization model | **0.593** | 0.950 | 0.614 | 0.952 | **0.710** | 0.981 | 0.858 | 0.911 |
| Simple routing baseline | 0.566 | 0.943 | 0.613 | **0.953** | 0.692 | 0.979 | 0.861 | 0.914 |
| Threshold-induced routed model | 0.588 | **0.951** | **0.615** | **0.953** | 0.689 | 0.979 | **0.885** | 0.922 |
| Minimal-specialization model | 0.534 | 0.917 | 0.562 | 0.938 | 0.620 | 0.963 | 0.877 | **0.923** |
| *Matched non-agentic fusion baseline* | | | | | | | | |
| Signal + image + peaks | 0.456 | 0.856 | 0.361 | 0.739 | 0.536 | 0.910 | 0.631 | 0.792 |
| Signal + image + peaks + features | **0.593** | 0.928 | 0.516 | 0.912 | 0.637 | 0.966 | 0.770 | 0.873 |
| Signal + image + peaks + features + morphology | 0.567 | 0.937 | 0.536 | 0.926 | 0.641 | 0.967 | 0.821 | 0.900 |
| *Machine learning model* | | | | | | | | |
| SVM [Mondéjar-Guerra et al., 2019] | 0.523 | 0.934 | 0.575 | 0.938 | 0.681 | **0.985** | 0.600 | 0.747 |
| *Deep learning models* | | | | | | | | |
| 2D CNN [Limam and Precioso, 2017] | 0.423 | 0.839 | 0.486 | 0.829 | 0.587 | 0.938 | 0.618 | 0.780 |
| STFT CNN [Huang et al., 2019] | 0.366 | 0.853 | 0.395 | 0.832 | 0.462 | 0.930 | 0.373 | 0.682 |
| 1D ResNet [Hannun et al., 2019] | 0.362 | 0.760 | 0.467 | 0.814 | 0.501 | 0.869 | 0.654 | 0.786 |
| Transformer [Hu et al., 2022] | 0.363 | 0.907 | 0.427 | 0.913 | 0.513 | 0.962 | 0.746 | 0.850 |
| xECG [Lunelli et al., 2025] | 0.440 | 0.893 | 0.512 | 0.901 | 0.523 | 0.934 | 0.698 | 0.820 |
| TCN [Ingolfsson et al., 2021] | 0.444 | 0.704 | 0.451 | 0.783 | 0.444 | 0.831 | 0.676 | 0.811 |
| *Closed-source VLMs* | | | | | | | | |
| Gemini 3.1 [Team et al., 2023] | 0.410 | 0.878 | 0.522 | 0.849 | 0.653 | 0.924 | 0.389 | 0.693 |
| GPT-5.4 [Singh et al., 2025] | 0.352 | 0.835 | 0.307 | 0.727 | 0.428 | 0.920 | 0.252 | 0.664 |
| *Open-source VLMs* | | | | | | | | |
| Qwen3.5-4B [Bai et al., 2023] | 0.414 | 0.785 | 0.443 | 0.865 | 0.517 | 0.861 | 0.373 | 0.709 |
| Gemma-3-27B [Gemma Team, 2025] | 0.378 | 0.880 | 0.277 | 0.817 | 0.367 | 0.878 | 0.209 | 0.608 |
| *ECG VLMs* | | | | | | | | |
| PULSE [Liu et al., 2024] | 0.211 | 0.646 | 0.226 | 0.660 | 0.216 | 0.670 | 0.242 | 0.568 |
| GEM [Lan et al., 2025] | 0.205 | 0.459 | 0.194 | 0.468 | 0.273 | 0.465 | 0.285 | 0.398 |
| ECG-R1 [Jin et al., 2026] | 0.199 | 0.446 | 0.116 | 0.237 | 0.270 | 0.532 | 0.243 | 0.515 |

We therefore report results on the clinically meaningful beat-label set $\mathcal{Y} = \{N, S, V, F\}$, where $N$, $S$, $V$, and $F$ denote normal, supraventricular ectopic, ventricular ectopic, and fusion beats, respectively. All four datasets are highly imbalanced under the coarse AAMI mapping; full beat-level class statistics are provided in Appendix G. To assess whether the observed gains can be attributed to data augmentation alone, Appendix C reports a complementary evaluation of the augmentation strategy on the TCN deep-learning baseline. A repository containing the implementation, configurations, and reproduction instructions is available at `https://github.com/jimmylihui/DeepArrhythmia_repo`.

**DeepArrhythmia variants** We compare the minimal- and rich-specialization models with the threshold-induced routed model. For additional analysis, we also include a simple routing baseline. This baseline is trained on trajectories derived from the Stage 2 split, where each sample is assigned to the rich-evidence trajectory if the rich-specialization model yields a more accurate prediction than the minimal-specialization model, and to the minimal-evidence trajectory otherwise.

### 4.1 Main Results: Context, Evidence, and Acquisition Policy

Table 1 shows that DeepArrhythmia variants achieve the best or tied-best performance across datasets and metrics. These comparisons highlight three main advantages of the framework.

**Contextual information improves performance.** DeepArrhythmia explicitly leverages segment-level context and consistently outperforms the conventional machine-learning, deep-learning, and VLM baselines. The value of contextual information is further supported by Appendix O, where increasing the available context length leads to consistent improvements in performance. This

interpretation is also reinforced by the morphology-analyzer statistics and qualitative examples in Appendix S and Appendix T, which show that rhythm-related cues and cross-beat morphological consistency frequently contribute to model decisions.

**Context must be processed appropriately.** These results indicate that contextual information is beneficial only when it is represented and modeled appropriately. On the representation side, the matched non-agentic fusion baselines improve substantially when structured features are added, and the morphology modality yields further gains on three of the four datasets, even though both are ultimately derived from the same raw ECG signal. On the modeling side, however, access to additional evidence alone is insufficient: SVM and VLM baselines can consume contextual information, yet they remain well below the strongest DeepArrhythmia variants. Even under closely matched inputs, DeepArrhythmia outperforms the fusion baselines (on MIT-BIH, Macro/Micro-F1 0.593/0.950 vs. 0.567/0.937 for the non-agentic fusion baseline with the full evidence set including morphology), and even without the Morphology Analyzer it exceeds the SVM and Gemini 3.1 (Macro-F1 0.5694 vs. 0.523 and 0.410; Table 2), suggesting that model architecture plays a central role in translating heterogeneous evidence into accurate beat-level predictions. This interpretation is further supported by Appendix N, where perturbing the evidence for a target beat changes the prediction for that beat while leaving neighboring beats largely unaffected, indicating a targeted link between evidence and classification. Model capacity is likewise important: Appendix R shows that performance improves with backbone size (0.8B to 4B).

**Not all additional evidence is equally useful.** Additional evidence is most helpful for difficult segments, but it does not provide uniform benefit for every sample. Appendix I shows that richer evidence improves performance primarily on low-confidence segments, whereas high-confidence segments benefit much less. Moreover, low-quality evidence can be detrimental. As discussed in Appendix D, the weaker utility of VitalDB features suggests lower signal quality or noisier labels, and this is consistent with the observation that the rich-specialization model underperforms the minimal-specialization model on VitalDB. This dataset-dependent behavior motivates selective evidence acquisition rather than always invoking every tool. By combining confidence thresholds with routing, the threshold-induced routed model consistently matches or exceeds the simple routing baseline and achieves the best or tied-best performance among DeepArrhythmia variants while reducing reliance on uninformative evidence.

Beyond the main within-dataset benchmarks, Appendix R shows that performance improves with backbone size. The gains are statistically reliable: 95% bootstrap confidence intervals on MIT-BIH Macro-F1 do not overlap (DeepArrhythmia 0.593 [0.583, 0.603] vs. SVM 0.523 [0.516, 0.530]), and the ordering is preserved under balanced accuracy (Appendix M). The framework is also robust to the routing threshold, whose confidence is well calibrated (ECE 0.024–0.058; Appendix F), to peak-detector errors (Appendix P), and to signal noise (Appendix Q). We also identify two important limitations. Appendix J demonstrates limited cross-dataset generalization, although DeepArrhythmia degrades less than the baselines (mean off-diagonal Micro-F1 0.698 vs. 0.457 for the 1D ResNet), and Appendix M shows that performance on extremely rare classes remains weak in absolute terms.

### 4.2 Ablation and Adaptive Evidence Acquisition

**Ablation Study.** Table 2 summarizes the ablation results on MIT-BIH. Shift augmentation is a critical component of DeepArrhythmia: removing it reduces Micro-F1 from 0.9501 to 0.9120, indicating that the model benefits substantially from improved data balance. The ablations further show that richer evidence sources—including the ECG image, Feature Extractor, and Morphology Analyzer—each contribute positively to performance, with the Feature Extractor having a particularly important role. Finally, the Peak Detector is indispensable because it provides the beat anchors required for downstream alignment and beat-level classification; without it, the model cannot localize candidate beats reliably enough to perform classification.

Table 2: Ablation on the MIT-BIH test split.

| Configuration | Macro-F1 | Micro-F1 |
|---|---|---|
| Full DeepArrhythmia | 0.5929 | 0.9501 |
| w/o shift augmentation | 0.4259 | 0.9120 |
| *Modality and subsystem ablations* | | |
| w/o ECG image | 0.5394 | 0.9426 |
| w/o Feature Extractor | 0.3527 | 0.9066 |
| w/o Morphology Analyzer | 0.5694 | 0.9433 |
| w/o Peak Detector | – | – |

**Adaptive Evidence Acquisition.** Table 3 compares the computational profiles of the rich-specialization model, minimal-specialization model, simple routing baseline, threshold-induced

Table 3: Computational cost and latency comparison of DeepArrhythmia variants across datasets.

| | MIT-BIH | | | MIT-BIH-SUP | | | INCART | | | VitalDB | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Policy | A. Cal | T/s | Lat. | A. Cal | T/s | Lat. | A. Cal | T/s | Lat. | A. Cal | T/s | Lat. |
| Rich-Spec. (Agentic) | 3 | 648 | 3.89 | 3 | 646 | 3.85 | 3 | 652 | 3.94 | 3 | 666 | 4.02 |
| Rich-Spec. (Non-agentic) | 3 | 573 | 3.40 | 3 | 570 | 3.39 | 3 | 580 | 3.50 | 3 | 595 | 3.59 |
| Simple Routing | 1.78 | 365 | 3.24 | 1.99 | 409 | 3.28 | 2.24 | 474 | 3.53 | 1.92 | 413 | 3.48 |
| Threshold Routing[a] | 3.41 | 510 | 3.58 | 3.20 | 459 | 3.40 | 3.07 | 436 | 3.45 | 2.52 | 319 | 3.27 |
| Minimal-Spec. | 1 | 183 | 2.83 | 1 | 178 | 2.73 | 1 | 186 | 2.88 | 1 | 197 | 3.01 |

*A. Cal* = average tool calls per sample; *T/s* = tokens per sample; *Lat.* = seconds per sample; [a] adds one tool call.

routed model, and rich-specialization non-agentic baseline. The minimal-specialization model is the most efficient variant, with an average latency of 2.86 seconds per sample across datasets. By contrast, the rich-specialization model is the most expensive because it always acquires the full set of evidence. The rich-specialization non-agentic baseline removes the decoder-only backbone, yet its computational cost remains high, with an average latency of 3.47 seconds per sample, indicating that inference cost is dominated largely by evidence acquisition rather than by autoregressive decoding alone. The simple routing baseline occupies an intermediate operating point between the rich- and minimal-specialization models; however, its latency varies only modestly across datasets, ranging from 3.24 to 3.48 seconds per sample, which suggests comparatively limited adaptation to dataset-specific evidence utility. In contrast, the threshold-induced routed model exhibits greater variation in computational cost, with latency ranging from 3.27 to 3.58 seconds per sample. This pattern is consistent with more selective routing: on datasets such as MIT-BIH, where richer evidence is more beneficial, the model allocates a larger proportion of samples to the rich-evidence trajectory, whereas on datasets such as VitalDB, where richer evidence is less informative, it routes more samples to the minimal-evidence trajectory to reduce computational cost. Appendix A summarizes this device setup. All latency values are measured with three independent single-GPU A6000 workers, batch size 1, bfloat16 weights, and greedy decoding with up to 8192 new tokens.

**Expert evaluation of generated morphology analyses.** We conducted a blinded expert evaluation of the generated morphology analyses to assess explanation quality and clinical plausibility; the full evaluation protocol is provided in Appendix W. Across the three systems, 200 samples per system (600 total) were independently scored by three experts on seven criteria. We compared analyses generated by Gemini 3.1 Pro with beat-label conditioning, Gemini 3.1 Pro without beat-label conditioning, and the DeepArrhythmia Morphology Analyzer. As shown in Tables 4, DeepArrhythmia received strong ratings for physiological correctness (4.92/5), use of segment context (4.39/5), clinical utility (4.22/5), and overall usefulness (4.19/5). Relative to Gemini 3.1 Pro without label conditioning, DeepArrhythmia is more competitive on physiological correctness and contextual grounding, but it remains below the label-conditioned Gemini 3.1 Pro teacher across all reported metrics. This pattern suggests that access to beat-label information enables the teacher model to produce more diagnostically explicit and evidentially complete analyses. Additional interpretation statistics and qualitative examples are provided in Appendix S and T, and Appendix U examines the effect of alternative teacher models. Notably, the Morphology Analyzer is trained with label-conditioned teacher distillation but is used without label input at inference time.

Table 4: Human expert evaluation of morphology analyses. Mean (STD) on a 5-point scale.

| | Clinical Metrics | | | | Utility & Safety Metrics | | | |
|---|---|---|---|---|---|---|---|---|
| Model | Diag. Supp. | ECG Evid. | Physiol. Corr. | Seg. Context | Evid. Compl. | Uncert. & Safety | Clin. Utility | Overall |
| Gemini (w/ labels) | 4.98 (0.14) | 5.00 (0.00) | 5.00 (0.00) | 5.00 (0.00) | 4.92 (0.31) | 4.88 (0.33) | 4.99 (0.10) | 4.97 (0.19) |
| Gemini (wo labels) | 4.60 (1.01) | 4.68 (0.84) | 4.84 (0.58) | 4.37 (0.84) | 4.38 (0.94) | 4.74 (0.73) | 4.67 (0.89) | 4.61 (0.86) |
| DeepArrhythmia | 3.88 (0.62) | 4.25 (0.89) | 4.92 (0.27) | 4.39 (0.75) | 3.51 (0.90) | 4.14 (0.78) | 4.22 (0.91) | 4.19 (0.86) |

## 5 Conclusion and Limitations

We presented **DeepArrhythmia**, a tool-grounded multimodal framework for segment-contextualized beat-level ECG arrhythmia classification from multi-beat ECG segments. The central idea is to treat each prediction as a beat-instance decision while allowing the evidence for that decision to come from

surrounding rhythm context, structured physiological measurements, and morphology-focused textual evidence. DeepArrhythmia decouples physiological measurement from evidence integration through specialized tools for beat localization, numerical rhythm–morphology extraction, and morphology analysis, while using confidence-guided routing to decide when richer evidence should be acquired. Across four public arrhythmia datasets, the results show that multi-beat context and structured physiological grounding improve beat-level classification, and that adaptive evidence acquisition provides a practical alternative to always invoking all tools. Beyond predictive performance, the retained intermediate evidence offers a basis for interpretation, error analysis, and expert review. DeepArrhythmia targets offline and retrospective analysis and clinical decision support, such as Holter review and second reading of flagged segments, rather than hard real-time on-device alarming.

Several limitations remain. The confidence-guided routing policy may be vulnerable to domain shift, because confidence calibration can vary across devices, sampling rates, patient populations, and noise conditions. Future work should therefore replace fixed thresholds with more robust uncertainty-aware routing, such as label-free or conformal threshold selection, and validate the framework on multi-center cohorts. Performance on extremely rare classes remains limited, especially for fusion beats, and improving recognition of these classes, for example through generative synthesis of realistic fusion morphology, remains an important open challenge.

## Acknowledgments and Disclosure of Funding

This work received no external funding. The authors declare no competing interests.

## References

Mikhail L Arbuzov, Alexey A Shvets, and Sisong Beir. Beyond exponential decay: Rethinking error accumulation in large language models. *arXiv preprint arXiv:2505.24187*, 2025.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wen Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Hyunseung Chung, Jungwoo Oh, Daeun Kyung, Jiho Kim, Yeonsu Kwon, Min-Gyu Kim, and Edward Choi. Ecg-agent: On-device tool-calling agent for ecg multi-turn dialogue. *arXiv preprint arXiv:2601.20323*, 2026.

Dhaval S Desai and Said Hajouli. Arrhythmias. In *StatPearls [Internet]*. StatPearls Publishing, 2023.

Zahra Ebrahimi, Mohammad Loni, Masoud Daneshtalab, and Arash Gharehbaghi. A review on deep learning methods for ecg arrhythmia classification. *Expert Systems with Applications: X*, 7:100033, 2020.

Da-In Eun, Kayoung Shim, Hyunsoo Lee, Yeji Lim, Hanbyeol Lim, Hyeonhoon Lee, Jiwon Lee, and Hyung-Chul Lee. Vitaldb arrhythmia database: An anesthesiologist-validated large-scale intraoperative arrhythmia dataset with beat and rhythm labels. *Scientific Data*, 2026.

Elizabeth Fons, Rachneet Kaur, Soham Palande, Zhen Zeng, Tucker Balch, Manuela Veloso, and Svitlana Vyetrenko. Evaluating large language models on time series feature understanding: A comprehensive taxonomy and benchmark. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21598–21634, 2024.

Gemma Team. Gemma 3 technical report, 2025.

Ary L Goldberger, Zachary D Goldberger, and Alexei Shvilkin. *Goldberger's Clinical Electrocardiography-E-Book: A Simplified Approach*. Elsevier Health Sciences, 2023.

Scott David Greenwald, Ramesh S Patil, and Roger G Mark. Improved detection and classification of arrhythmias in noise-corrupted electrocardiograms using contextual information. In *Proceedings of Computers in Cardiology*, pages 461–464. IEEE, 1990.

Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine*, 25(1):65–69, 2019.

Shenda Hong, Yuxi Zhou, Junyuan Shang, Cao Xiao, and Jimeng Sun. Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review. *Computers in biology and medicine*, 122:103801, 2020.

Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, 2023.

Rui Hu, Jie Chen, and Li Zhou. A transformer-based deep neural network for arrhythmia detection using continuous ecg signals. *Computers in biology and medicine*, 144:105325, 2022.

Jingshan Huang, Binqiang Chen, Bin Yao, and Wangpeng He. Ecg arrhythmia classification using stft-based spectrogram and convolutional neural network. *IEEE access*, 7:92871–92880, 2019.

Thorir Mar Ingolfsson, Xiaying Wang, Michael Hersche, Alessio Burrello, Lukas Cavigelli, and Luca Benini. Ecg-tcn: Wearable cardiac arrhythmia detection with a temporal convolutional network. In *2021 IEEE 3rd International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, pages 1–4. IEEE, 2021.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*, 2025.

Jiarui Jin, Haoyu Wang, Xingliang Wu, Xiaocheng Fang, Xiang Lan, Zihan Wang, Deyun Zhang, Bo Liu, Yingying Zhang, Xian Wu, et al. Ecg-r1: Protocol-guided and modality-agnostic mllm for reliable ecg interpretation. *arXiv preprint arXiv:2602.04279*, 2026.

Deepali Koppad. Arrhythmia classification using deep learning: A review. *WSEAS Trans. Biol. Biomed*, 18:96–105, 2021.

Xiang Lan, Feng Wu, Kai He, Qinghao Zhao, Shenda Hong, and Mengling Feng. Gem: Empowering mllm for grounded ecg understanding with time series and images. *arXiv preprint arXiv:2503.06073*, 2025.

Jiahui Li, Yida Zhang, Zixuan Zeng, Jiayu Chen, Xiang Zhang, Jin Lu, WenZhan Song, and Fei Dou. Peak-r1: Instruction-tuned large language models for robust j-peak detection in cardiomechanical signals. In *NeurIPS 2025 Workshop on Learning from Time Series for Health*, 2025.

Mohamed Limam and Frederic Precioso. Atrial fibrillation detection and ecg classification based on convolutional recurrent neural network. In *2017 computing in cardiology (CinC)*, pages 1–4. IEEE, 2017.

Ruoqi Liu, Yuelin Bai, Xiang Yue, and Ping Zhang. Teach multimodal llms to comprehend electrocardiographic images. *arXiv preprint arXiv:2410.19008*, 2024.

Riccardo Lunelli, Angus Nicolson, Samuel Martin Pröll, Sebastian Johannes Reinstadler, Axel Bauer, and Clemens Dlaska. Benchecg and xecg: a benchmark and baseline for ecg foundation models. *arXiv preprint arXiv:2509.10151*, 2025.

Lucie Maršánová, Andrea Němcová, Radovan Smíšek, Martin Vítek, and Lukáš Smital. Advanced p wave detection in ecg signals during pathology: evaluation in different arrhythmia contexts. *Scientific reports*, 9(1):19053, 2019.

Víctor Mondéjar-Guerra, Jorge Novo, José Rouco, Manuel G Penedo, and Marcos Ortega. Heartbeat classification fusing temporal and morphological information of ecgs via ensemble of classifiers. *Biomedical Signal Processing and Control*, 47:41–48, 2019.

George B Moody and Roger G Mark. The impact of the mit-bih arrhythmia database. *IEEE engineering in medicine and biology magazine*, 20(3):45–50, 2001.

Jiapu Pan and Willis J Tompkins. A real-time QRS detection algorithm. *IEEE Transactions on Biomedical Engineering*, BME-32(3):230–236, 1985.

A Rajkumar, M Ganesan, and R Lavanya. Arrhythmia classification on ecg using deep learning. In *2019 5th international conference on advanced computing & communication systems (ICACCS)*, pages 365–369. IEEE, 2019.

Erendira Rendon, Roberto Alejo, Carlos Castorena, Frank J Isidro-Ortega, and Everardo E Granda-Gutierrez. Data sampling methods to deal with the big data multi-class imbalance problem. *Applied Sciences*, 10(4):1276, 2020.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, et al. Openai gpt-5 system card. *arXiv preprint arXiv:2601.03267*, 2025.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Viktor Tihonenko, Alexander Khaustov, Sergey Ivanov, and Alexei Rivin. St Petersburg INCART 12-lead arrhythmia database. PhysioNet, 2008. URL `https://physionet.org/content/incartdb/1.0.0/`. Version 1.0.0.

Ziyu Xu, Chen Dan, Justin Khim, and Pradeep Ravikumar. Class-weighted classification: Trade-offs and robust approaches. In *International conference on machine learning*, pages 10544–10554. PMLR, 2020.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=WE_vluYUL-X`.

Wei Zhao, Jing Hu, Dongya Jia, Hongmei Wang, Zhenqi Li, Cong Yan, and Tianyuan You. Deep learning based patient-specific classification of arrhythmia on ecg signal. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1500–1503. IEEE, 2019.

Yubao Zhao, Jiaju Kang, Tian Zhang, Puyu Han, and Tong Chen. Ecg-chat: A large ecg-language model for cardiac disease diagnosis. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2025. doi: 10.1109/ICME59968.2025.11209476.

Tengda Zhou, Shaoyang Men, Jingxian Liang, Baoxian Yu, Han Zhang, and Xiaomu Luo. 1d u-net++: an effective method for ballistocardiogram j-peak detection. *Journal of Mechanics in Medicine and Biology*, 21(10):2140058, 2021.

# DeepArrhythmia: Segment-Contextualized ECG Arrhythmia Classification via Selective Evidence Acquisition


A Implementation Details 14
B Additional Method Details 14
C Baseline Augmentation 17
D VitalDB Result Analysis 17
E Routing Scheme Analysis 18
F Sensitivity to the Routing Threshold and Calibration 18
G Dataset Class Distribution 20
H Impact of Data Splitting 20
I Accuracy Distribution by Prediction Confidence 20
J Cross-Dataset Performance 21
K Knowledge Distillation for the Morphology Analyzer 25
L Example of the Complete Pipeline for DeepArrhythmia 26
M Detailed Result Analysis for DeepArrhythmia 28
N Faithfulness Evaluation 28
O Impact of Context Length 31
P Impact of the Peak Detector 32
Q Impact of Noise 33
R Impact of Backbone Scale 34
S Interpretation Analysis 35
T Interpretation Examples 35
U Effect of the Teacher Model in the Morphology Analyzer 36
V Broader Impacts and Limitations 37
W Human Evaluation Protocol 39

# A Implementation Details

Table 5: Training configuration used for supervised fine-tuning of DeepArrhythmia.

| Parameter | Configuration / Value |
|---|---|
| Backbone model | Qwen3.5-4B |
| ECG encoder | Pretrained ECG-CoCa checkpoint |
| GPUs | 3 × NVIDIA A6000 GPUs |
| Memory | 49 GB |
| Precision | bfloat16 |
| Attention | FlashAttention-2 |
| Distributed training | DeepSpeed ZeRO-3 |
| Per-device batch size | 1 |
| Gradient accumulation | 4 |
| Effective global batch size | 12 |
| Learning rate | $2 \times 10^{-5}$ |
| Max sequence length | 8000 |

Table 5 summarizes the training configuration used for DeepArrhythmia. The total GPU hours required to train the model on MIT-BIH, MIT-BIH-SUP, INCART, and VitalDB were 61.30, 127.05, 57.72, and 179.98, respectively.

# B Additional Method Details

## B.1 Shift-based augmentation

To mitigate class imbalance, we apply shift-based augmentation under the AAMI coarse class mapping as shown in Fig 3. We first compute class frequencies and define target relative abundances for non-normal classes with respect to the normal (N) class. For each non-N beat, we generate additional fixed-length (10 s) segments by re-anchoring the window so that the beat's R-peak appears at predefined fractional offsets along the segment axis. Rarer classes are assigned more distinct offsets to approach the target ratio, whereas classes already near the target are not oversampled. Candidate windows are clipped to valid record boundaries and deduplicated against the standard non-overlapping grid. During data loading, only augmentation files present on disk are ingested and merged with the base training split.

## B.2 Detailed tool definitions

**Peak Detector.** To handle variable numbers of beats per segment, the first step is robust R-peak localization. We implement the peak detector using a 1D-UNet++ architecture [Zhou et al., 2021]. Given an input segment $x_{ts}$, the detector outputs a point-wise probability sequence:

$$p_t = f_\theta(x_{ts})_t, \qquad p_t \in [0,1], \quad t = 1, \ldots, L, \tag{1}$$

where $f_\theta$ denotes the detector parameterized by $\theta$.

Training uses point-wise supervision at each time index. Let $y_t \in [0,1]$ denote the smoothed target label, obtained by applying a Gaussian kernel centered at each annotated peak to improve robustness. The optimization objective is point-wise binary cross-entropy:

$$\mathcal{L}_{\text{peak}} = -\frac{1}{L} \sum_{t=1}^{L} \left[ y_t \log p_t + (1 - y_t) \log(1 - p_t) \right]. \tag{2}$$

At inference time, candidate peaks are obtained from the probability sequence using a classical peak-finding procedure [Pan and Tompkins, 1985] with thresholding and non-maximum selection, yielding the final R-peak set $\mathcal{P} = \{\hat{t}_k\}_{k=1}^{K}$ for each segment.

**Feature Extractor.** To complement learned representations with explicit physiological cues, we extract a set of handcrafted numerical features from each detected beat Hannun et al. [2019]. Specifically, five feature groups are used: (i) R–R interval, (ii) normalized R–R interval, (iii) R-peak amplitude, (iv) higher-order statistics (HOS), and (v) numerical morphology descriptors.

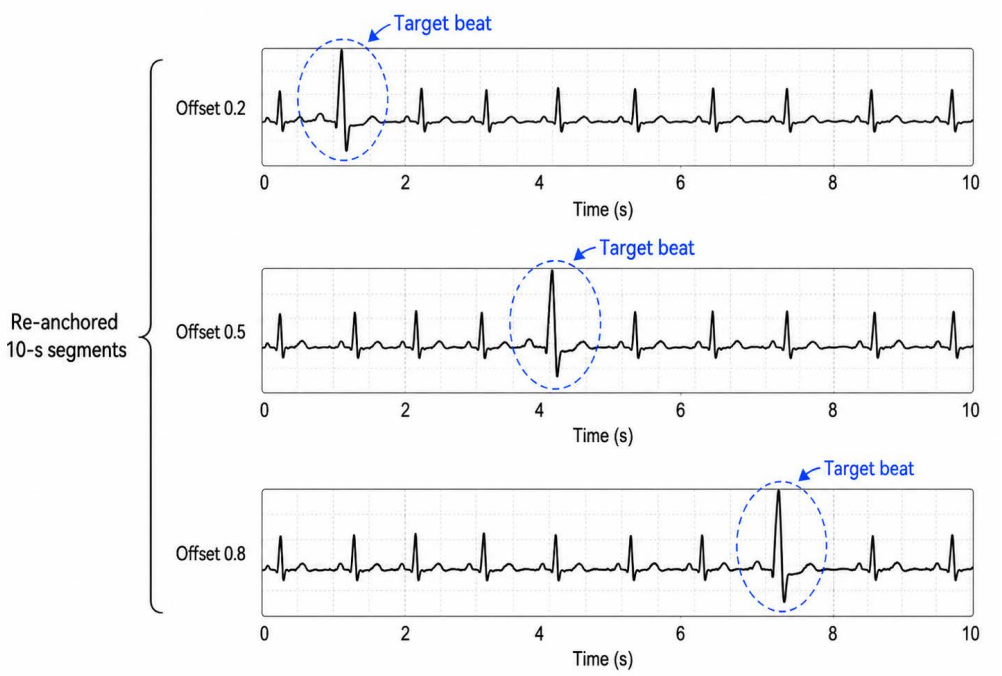



Figure 3: Illustration of the shift-based augmentation strategy used to alleviate class imbalance. Rare classes are oversampled by placing their R-peaks at different positions within the segment, thereby increasing data diversity.

For HOS, each beat is partitioned into five temporal sub-intervals, and both skewness and kurtosis are computed for each sub-interval, resulting in a 10-dimensional feature vector. Morphology descriptors are derived from Euclidean distances in the (sample, amplitude) space between the R-peak and four characteristic points selected from predefined local windows. These points are identified as follows: $\max(\text{beat}[0, 40])$, $\min(\text{beat}[75, 85])$, $\min(\text{beat}[95, 105])$, and $\max(\text{beat}[150, 180])$.

**Morphology Analyzer.** To model beat morphology in an interpretable form, we employ a vision-language morphology analyzer that takes as input the segment ECG image together with beat-local auxiliary context. For a queried beat within the current segment, the student analyzer generates a textual morphology rationale:

$$t_m = g_\phi(x_{\text{img}}, c)\,, \tag{3}$$

where $t_m$ denotes the generated morphology text, $x_{\text{img}}$ denotes the ECG image of the current segment, and $c$ denotes the auxiliary context provided by the Feature Extractor for the queried beat.

We train $g_\phi$ with a knowledge-distillation framework using an online teacher LLM (e.g., Gemini 3.1). Given image input, ground-truth class label, and auxiliary context features (e.g., RR intervals and peak amplitude), the teacher produces reference explanations:

$$\tilde{t}_m = g_\psi(x_{\text{img}}, y, c)\,, \tag{4}$$

where $g_\psi$ denotes the teacher model, $y$ is the ground-truth label of the queried beat, and $c$ is the corresponding auxiliary context. The impact of teacher-model choice is analyzed in Appendix U.

To prevent data leakage, teacher-generated rationales are produced exclusively for beats in the training partition; at inference time, the Morphology Analyzer operates without access to the ground-truth label.

The student is optimized with a token-level morphology distillation objective:

$$\mathcal{L}_{\text{morph}} = -\sum_{n=1}^{N} \log p_\phi\big(\tilde{t}_{m,n} \mid \tilde{t}_{m,<n}, x_{\text{img}}, c\big)\,, \tag{5}$$

where $N$ is the target explanation length. The generated morphology text is passed to the central agent as structured evidence for final arrhythmia classification. In our experiments, Qwen3.5-4B serves as the base model for the student Morphology Analyzer.

**Confidence Calculator.** The Confidence Calculator maps the current beat-level class posteriors to a single segment-level routing signal. For each detected beat, it takes the largest predicted class probability as that beat's confidence, and it then averages those beat-wise confidences across the segment to obtain $C_{\text{seg}}$. This scalar $C_{\text{seg}}$ is the confidence statistic used to decide whether the additional evidence-producing tools should be invoked. Accordingly, routing is performed once per segment in the present study; per-beat routing was not explored here. In the final agentic system, the Confidence Calculator is queried once at most per segment, immediately after peak detection and before any optional evidence call.

### B.3 Detailed training objectives

We introduce two modality placeholders, `<Signal>` and `<Image>`, to indicate insertion points for modality-specific tokens. Given time-series input $x_T$ and image input $x_I$, modality features are extracted and projected into the language-model embedding space using independent encoders and projectors:

$$e_T = \text{Encoder}_T(x_T; \theta_{E_T}), \qquad z_T = \text{Proj}_T(e_T; \theta_{P_T}), \tag{6}$$

$$e_I = \text{Encoder}_I(x_I; \theta_{E_I}), \qquad z_I = \text{Proj}_I(e_I; \theta_{P_I}), \tag{7}$$

where $z_T \in \mathbb{R}^{N_T \times d}$ and $z_I \in \mathbb{R}^{N_I \times d}$ denote signal and image token sequences in a shared embedding space, and $\theta_M = \{\theta_{P_T}, \theta_{P_I}\}$ denotes projector parameters. The language model is conditioned on textual context $x_{\text{text}}$ together with $(z_T, z_I)$ by injecting $z_T$ at `<Signal>` and $z_I$ at `<Image>`.

We serialize each trajectory as a token sequence $\tau = (\tau_1, \ldots, \tau_T)$, containing optional `<Call>` spans, externally returned `<Output>` spans, and the final peak-aligned textual answer. The final answer is generated as ordinary language-model text, e.g., `[77:N] [370:N] [663:N] [947:N].`

Let $X = (x_{\text{text}}, z_T, z_I)$ denote the multimodal conditioning context, and let $h_i$ denote the autoregressive language-model state before generating token $\tau_i$. The agent defines the standard next-token probability $p_\theta(\tau_i \mid X, \tau_{<i})$ over the language-model vocabulary.

Training uses masked next-token supervision over valid tool-action tokens and final-answer tokens. The objective is written as the sum of tool-action and class-answer token losses:

$$\mathcal{L}_{\text{agent}} = \mathcal{L}_{\text{tool}} + \mathcal{L}_{\text{cls}}, \tag{8}$$

with

$$\mathcal{L}_{\text{tool}} = -\sum_{i=1}^{T} m_i^a \log p_\theta(\tau_i^* \mid X, \tau_{<i}^*), \quad \mathcal{L}_{\text{cls}} = -\sum_{i=1}^{T} m_i^y \log p_\theta(\tau_i^* \mid X, \tau_{<i}^*), \tag{9}$$

where $\tau^* = (\tau_1^*, \ldots, \tau_T^*)$ is the serialized trajectory containing tool calls, tool outputs, and the final textual prediction. The masks $m_i^a$ and $m_i^y$ select supervised tool-action tokens and final-answer tokens, respectively. Externally provided tool-output tokens are included as context and masked out from the loss.

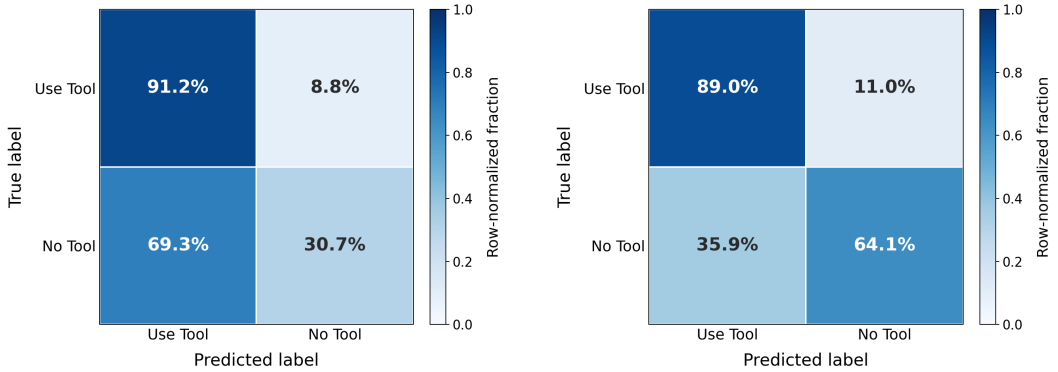
### B.4 Threshold Selection

We compute a segment-level confidence score for each example using the Confidence Calculator and choose a dataset-specific threshold that maximizes the micro F1 score for beat classification for Stage 2 SFT set. This threshold is fixed at inference time. The resulting dataset-specific thresholds are reported in Table 6.

Table 6: Micro-F1 confidence thresholds for tool-use decision

| Dataset | MIT-BIH Arrhythmia | MIT-BIH Supraventricular | INCART | VitalDB |
|---|---|---|---|---|
| Threshold | 0.990529 | 0.980933 | 0.993532 | 0.98519 |

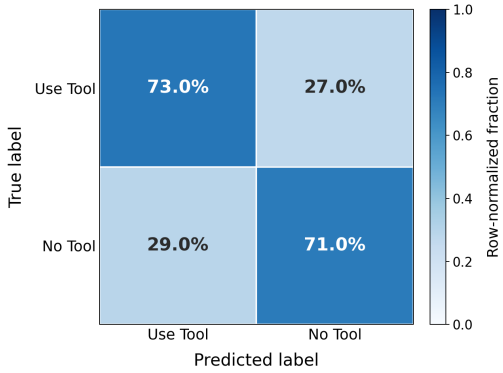
Figure 4 presents the row-normalized confusion matrices for the tool-use decision across datasets. On MIT-BIH Arrhythmia, MIT-BIH superventricular, and Incart, the classifier exhibits a strong tendency



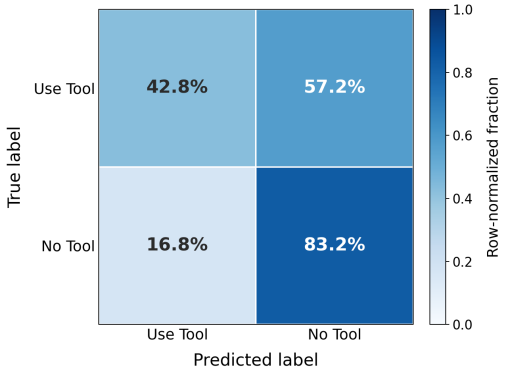

(a) Confusion matrix on MIT-BIH Arrhythmia.




(b) Confusion matrix on MIT-BIH Supraventricular.




(c) Confusion matrix on INCART.




(d) Confusion matrix on VitalDB.

Figure 4: Confusion-matrix panels for tool use across four datasets.

to invoke tools. In contrast, VitalDB exhibits a clear shift toward the no-tool prediction: 83.2% of true no-tool cases are correctly classified, whereas only 42.8% of true tool-use cases are identified. This pattern suggests a two-part limitation on VitalDB: first, the short-branch confidence signal used for routing is less effective at separating samples that should remain on the no-tool branch from those that should escalate to the richer branch; second, even after invocation, the optional evidence appears less useful on this dataset, as further discussed in Section D.

## C Baseline Augmentation

To assess whether standard imbalance-mitigation techniques can strengthen a conventional baseline, we evaluate the TCN model under three training settings: the original non-augmented baseline, class-weighted loss Xu et al. [2020], and class-balanced sampling augmentation Rendon et al. [2020]. Table 7 summarizes the resulting Macro-F1 and Micro-F1 scores across the four datasets. Class-weighted loss yields the most consistent improvement, outperforming the non-augmented baseline on both metrics for all datasets. By contrast, class-balanced sampling produces mixed effects: it improves Macro-F1 on INCART and MIT-BIH-SUP, but reduces Micro-F1 on three of the four datasets. Despite these modest gains, all augmented TCN variants remain substantially below DeepArrhythmia. Overall, these results indicate that standard imbalance-aware training offers only limited improvements for the TCN baseline.

## D VitalDB Result Analysis

The VitalDB performance becomes worse after adding the auxiliary ECG-classification evidence. To analyze whether this evidence is intrinsically useful, we evaluate the discriminative power of the evidence features alone.

Table 7: TCN baseline performance under alternative imbalance-mitigation strategies. We report Macro-F1 and Micro-F1 on the standard test split of each dataset.

| | MIT-BIH | | INCART | | MIT-BIH-SUP | | VitalDB | |
|---|---|---|---|---|---|---|---|---|
| **Training setting** | **Ma** | **Mi** | **Ma** | **Mi** | **Ma** | **Mi** | **Ma** | **Mi** |
| Non-augmented baseline | 0.4679 | 0.7048 | 0.4403 | 0.8249 | 0.4336 | 0.7796 | 0.5069 | 0.8114 |
| Class-weighted loss | **0.4796** | **0.7370** | 0.4705 | **0.8500** | 0.4597 | **0.8257** | **0.5223** | **0.8149** |
| Class-balanced sampling augmentation | 0.4446 | 0.6708 | **0.4712** | 0.8285 | **0.4680** | 0.8208 | 0.4895 | 0.7850 |

For each beat, we parse the auxiliary feature vector

$$\mathbf{x}_i = [\mathrm{RR}_i, \mathrm{normRR}_i, \mathrm{amp}_i, \mathrm{HOS}_i, \mathrm{myMorph}_i] \in \mathbb{R}^{23}, \tag{10}$$

from the evidence text. We then train a balanced logistic regression classifier using only $\mathbf{x}_i$, without ECG images or language-model predictions.

Table 8: Feature-only classification performance.

| Dataset | Accuracy | Macro-F1 | Weighted-F1 |
|---|---|---|---|
| MIT-BIH | 0.7080 | 0.5209 | 0.7895 |
| VitalDB | 0.4879 | 0.3851 | 0.5315 |

The results in Table 8 show a clear dataset-level gap in feature-only separability. On MIT-BIH, the handcrafted evidence features alone achieve 0.7080 accuracy, 0.5209 Macro-F1, and 0.7895 Weighted-F1, indicating that they retain meaningful class information even without ECG images or language-model predictions. On VitalDB, the same feature-only classifier drops to 0.4879 accuracy, 0.3851 Macro-F1, and 0.5315 Weighted-F1, showing that these auxiliary features are substantially less discriminative on that dataset.

This gap helps explain the downstream behavior. When the auxiliary evidence is relatively informative, as on MIT-BIH, conditioning on it can support the final prediction. When the evidence is weak, as on VitalDB, adding it can inject noise rather than useful information, which may degrade final performance.

## E Routing Scheme Analysis

We further analyze the routing mechanism on the MIT-BIH Arrhythmia test set by comparing two segment-level confidence statistics for optional tool invocation: the mean beat-wise confidence within a segment and the minimum beat-wise confidence within a segment. The mean-based statistic reflects the overall certainty of the segment, whereas the minimum-based statistic emphasizes the most uncertain beat. Table 9 reports the confidence threshold for each routing rule, the resulting Macro-F1 and Micro-F1, and the number of segments assigned to the no-tool and tool-enabled branches.

Both routing rules achieve the same Macro-F1 of 0.588, indicating comparable class-balanced performance. However, the mean-confidence rule attains a slightly higher Micro-F1 (0.951 vs. 0.948) while routing slightly fewer segments to the tool-enabled branch (2,954 vs. 2,982) and slightly more segments to the no-tool branch (1,244 vs. 1,216). By contrast, the minimum-confidence rule is more aggressive, because a single low-confidence beat can send the entire segment to the tool-enabled branch.

## F Sensitivity to the Routing Threshold

Because the per-dataset threshold $\tau_d$ is selected on a held-out split, a reasonable concern is that the routed model is brittle to the precise value of $\tau_d$ and may not transfer across acquisition conditions. We evaluate this directly by characterizing performance as a continuous function of the threshold. For a given $\tau$, the policy of Section 3 emits the minimal-evidence prediction on segments whose confidence satisfies $C_{\text{seg}} \geq \tau$ and the rich-evidence prediction otherwise; sweeping $\tau$ thus interpolates

Table 9: Comparison of segment-level routing rules on the MIT-BIH Arrhythmia test set. Confidence denotes the tuned threshold used by each routing rule. The No tool and Use tool columns report the number of evaluation segments assigned to the no-tool and tool-enabled branches, respectively.

| Routing signal | Confidence | Macro-F1 | Micro-F1 | No tool | Use tool |
|---|---|---|---|---|---|
| Mean confidence | 0.990529 | 0.588 | 0.951 | 1,244 | 2,954 |
| Minimum confidence | 0.990214 | 0.588 | 0.948 | 1,216 | 2,982 |

the entire family of policies between the pure minimal-evidence model ($\tau \to 0$) and the pure rich-evidence model ($\tau \to 1$). As $C_{\text{seg}}$ concentrates near unity—the minimal-evidence model is confidently correct on the dominant normal class—all non-trivial routing occurs within $\tau \in [0.98, 1.0]$, which we adopt as the operative range in Figure 5.

Performance is stable across this range on all four datasets. Micro-F1 spans only $0.9498$–$0.9524$ (MIT-BIH), $0.9509$–$0.9530$ (MIT-BIH Supraventricular), $0.9746$–$0.9807$ (INCART), and $0.9116$–$0.9255$ (VitalDB), a maximum variation of $1.4$ points with no discontinuity. A mis-specified threshold therefore degrades accuracy by at most a few tenths of a point rather than inducing a collapse, indicating that $\tau_d$ need not be finely tuned to remain effective.

The value of acquiring rich evidence nonetheless varies markedly across datasets, and the optimal operating point is not consistently at either extreme. On INCART, rich evidence is monotonically beneficial and performance is maximized as $\tau \to 1$. On VitalDB the trend is reversed: both Micro-F1 and Macro-F1 peak near the minimal-evidence end ($0.9255$ and $0.8927$ at $\tau \approx 0.98$) and decline monotonically toward $\tau = 1$, so indiscriminate acquisition is actively harmful, consistent with the noisier VitalDB features analyzed in Appendix D. On MIT-BIH and MIT-BIH Supraventricular the Micro-F1 curve is non-monotonic and attains an interior optimum ($0.9524$ at $\tau \approx 0.997$ and $0.9530$ at $\tau \approx 0.995$, respectively) that marginally exceeds the always-rich endpoint ($0.9504$ and $0.9522$); withholding rich evidence on the most confident segments is thus not merely cheaper but slightly more accurate than acquiring it everywhere.

Taken together, these results show that the utility of physiological evidence is heterogeneous—beneficial on some datasets, detrimental on others, and best applied selectively on the rest—so no fixed acquisition policy is uniformly optimal. Confidence-gated routing accommodates this heterogeneity automatically, keeping the model on the favorable side of the accuracy–cost trade-off on every dataset while remaining robust to the exact choice of threshold.

**Calibration of the routing score.** The confidence of a beat is $c = \max_y p_y$ over the posterior on $\{\mathrm{N}, \mathrm{S}, \mathrm{V}, \mathrm{F}\}$, and the routing score $C_{\text{seg}}$ is the mean beat confidence within a segment. On the minimal-evidence predictions of each subject-disjoint test split we bin all beats into $M = 10$ equal-width confidence bins and report $\mathrm{ECE} = \sum_m \frac{|B_m|}{N} |\mathrm{acc}(B_m) - \mathrm{conf}(B_m)|$ and $\mathrm{MCE} = \max_{m:|B_m|>0} |\mathrm{acc}(B_m) - \mathrm{conf}(B_m)|$ (Table 10). ECE is low throughout ($0.024$–$0.058$): the reliability curve tracks the diagonal at the high-confidence end, where the large majority of beats lie. The larger MCE ($0.117$–$0.164$) is driven by sparsely populated mid-confidence bins with mild over-confidence, which is benign for routing because the policy only requires the score to *order* segments by difficulty. Under cross-dataset shift calibration is not guaranteed to transfer, which is why $\tau_d$ is re-selected per dataset. The selected thresholds all fall in the narrow band $\tau_d \in [0.9809, 0.9935]$ (Table 10), inside the swept range, so substituting another dataset's threshold changes Micro-F1 by at most the span of the sweep ($0.3$, $0.2$, $0.6$, and $1.4$ points on MIT-BIH, MIT-BIH Supraventricular, INCART, and VitalDB). A genuinely new site could shift the confidence distribution itself; label-free alternatives such as fixing $\tau$ at a confidence percentile of the incoming data are left to future work.

**Alternative routers.** We compare the threshold router with a stochastic entropy router and a learned (logistic-regression) router. The simple threshold matches or exceeds both (Macro-F1 on MIT-BIH $0.593$ vs. $0.592$/$0.586$; INCART $0.629$ vs. $0.628$/$0.622$; VitalDB $0.892$ vs. $0.886$/$0.894$ for threshold vs. entropy/learned).

**Routing rates.** Table 11 reports how often the rich-evidence path is selected under each dataset's threshold ($C_{\text{seg}} < \tau_d$). Arrhythmia-containing segments trigger the rich path $2.2$–$5.3\times$ more often than normal-only segments. Rich routing concentrates in the lowest-confidence quartile, where the

Table 10: Calibration of the minimal-evidence confidence (10 equal-width bins) and the selected thresholds $\tau_d$.

| Dataset | ECE | MCE | $\tau_d$ |
|---|---|---|---|
| MIT-BIH | 0.058 | 0.121 | 0.9905 |
| MIT-BIH Supraventricular | 0.029 | 0.117 | 0.9809 |
| INCART | 0.024 | 0.155 | 0.9935 |
| VitalDB | 0.042 | 0.164 | 0.9852 |

Table 11: Fraction of segments routed to the rich-evidence path, overall and by segment type.

| Dataset | Rich (all) | Normal-only | Abnormal |
|---|---|---|---|
| MIT-BIH | 47.4% | 32.2% | 72.0% |
| MIT-BIH Supraventricular | 43.4% | 17.3% | 70.0% |
| INCART | 31.3% | 9.1% | 48.1% |
| VitalDB | 17.1% | 8.2% | 27.4% |

minimal-evidence model is least accurate (accuracy $0.81$–$0.91$ vs. $0.97$–$1.00$ elsewhere), and it is lowest on VitalDB, where rich evidence is least helpful. Because noise lowers confidence, degraded segments are routed to rich evidence more often, so the mechanism fails safe (Appendix Q).

## G Dataset Class Distribution

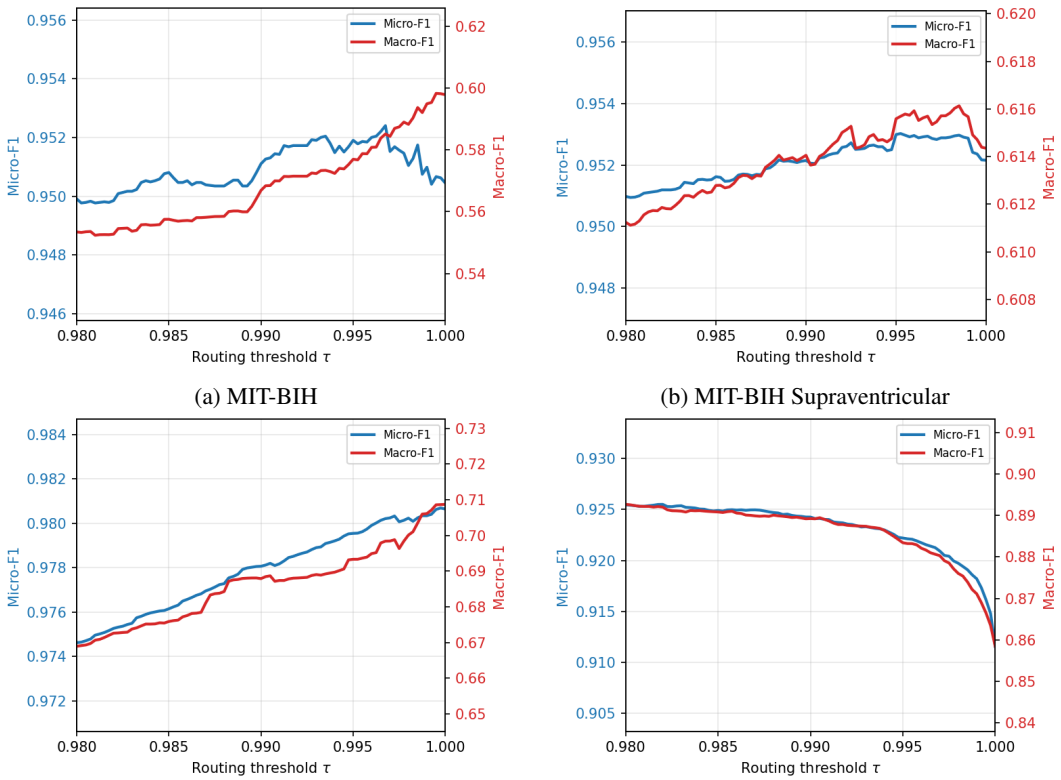
Table 12 shows that all four datasets are strongly imbalanced, but the degree and structure of imbalance vary substantially across datasets. Normal beats dominate every dataset, accounting for 68.19% to 87.95% of all annotations, which means that aggregate metrics can be heavily influenced by the majority class. MIT-BIH and INCART are both dominated by normal beats, but INCART contains a noticeably larger ventricular proportion (11.38%), whereas MIT-BIH includes a relatively larger fraction of unclassifiable beats (Q, 7.35%). MIT-BIH Superventricular remains normal-heavy overall, yet contains the highest proportion of supraventricular beats among the three conventional ECG benchmarks (6.61%), making it useful for evaluating supraventricular discrimination. VitalDB exhibits the most distinct distribution: compared with the other datasets, it contains a much lower normal-beat proportion and a substantially higher supraventricular proportion (28.60%), while ventricular beats remain comparatively rare. Fusion beats are extremely scarce in every dataset and entirely absent in VitalDB, which helps explain the instability of class-F evaluation.

## H Impact of data splitting

Prior ECG studies often report strong performance on MIT-BIH [Limam and Precioso, 2017, Huang et al., 2019]; however, reported results are highly sensitive to the data-partition protocol. In this work, we adopt the subject-disjoint DS1/DS2 split, which prevents patient overlap between training and test sets and therefore mitigates identity leakage. By contrast, random beat-level splitting can assign beats from the same subject to both sets, yielding an easier but less clinically realistic evaluation scenario. Figure 6 presents the confusion matrices obtained under the two evaluation protocols for the 2D CNN baseline Limam and Precioso [2017]. The random split produces markedly higher apparent performance, increasing SVEB recall from 9.7% to 90.0% and fusion-beat (F) recall from 0.5% to 87.5%. This large discrepancy indicates substantial leakage risk under random splitting and supports the use of subject-disjoint evaluation for reliable clinical assessment.

## I Accuracy Distribution by Prediction Confidence

Figure 7 illustrates the relationship between segment-level prediction confidence and Micro-F1 on the MIT-BIH dataset. Most segments cluster at very high confidence values (above 0.99), indicating that a substantial portion of the test set is relatively easy for the lightweight baseline. By contrast, the gains from routing are concentrated in the low-confidence regime, and the magnitude of improvement increases as confidence decreases. This pattern suggests that low-confidence segments correspond to




(a) MIT-BIH

(b) MIT-BIH Supraventricular

(c) INCART

(d) VitalDB

Figure 5: Beat-level Micro-F1 (left axis, blue) and Macro-F1 (right axis, red) versus the routing threshold $\tau$ over the operative range $[0.98, 1.0]$. Performance is flat in $\tau$ on all datasets (Micro-F1 varies by $\leq 1.4$ points), and the effect of rich evidence is dataset-dependent: monotonically beneficial on INCART, detrimental on VitalDB, and maximized at an interior threshold on MIT-BIH and MIT-BIH Supraventricular, where selective routing slightly surpasses always acquiring rich evidence.

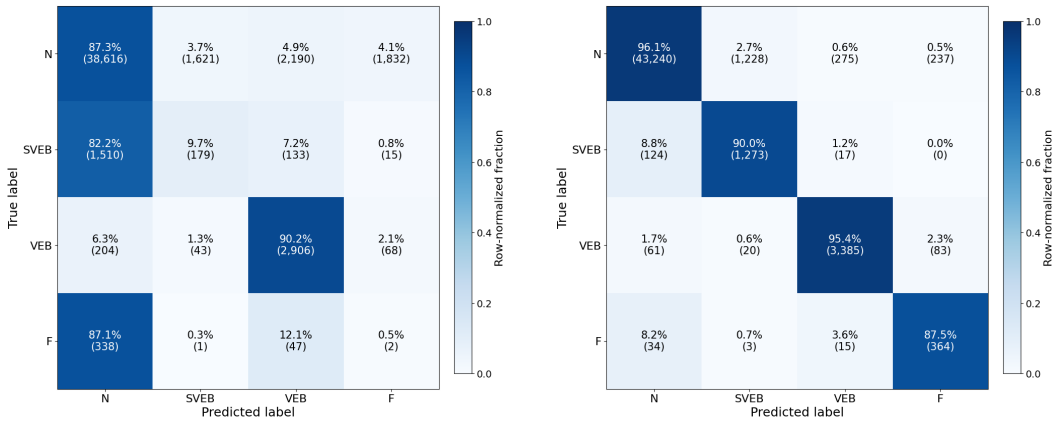
Table 12: Beat-level class distribution across datasets (datasets shown as columns).

| Metric | MIT-BIH | INCART | MIT-BIH SV | VitalDB |
|---|---|---|---|---|
| Records | 48 | 75 | 78 | 482 |
| Total beats | 109,494 | 175,874 | 184,582 | 640,107 |
| N (%) | 90,631 (82.77) | 153,676 (87.38) | 162,339 (87.95) | 436,503 (68.19) |
| S (%) | 2,781 (2.54) | 1,960 (1.11) | 12,198 (6.61) | 183,052 (28.60) |
| V (%) | 7,236 (6.61) | 20,013 (11.38) | 9,943 (5.39) | 16,097 (2.51) |
| F (%) | 803 (0.73) | 219 (0.12) | 23 (0.01) | 0 (0.00) |
| Q (%) | 8,043 (7.35) | 6 ($<$ 0.01) | 79 (0.04) | 4,455 (0.70) |

genuinely difficult cases and that richer evidence improves accuracy precisely where the baseline is least reliable. Overall, these results support the premise that a confidence-gated routing strategy achieves a favorable accuracy–efficiency trade-off by selectively invoking additional evidence only for challenging inputs.

## J Cross Dataset Performance

Table 13 reports cross-dataset generalization, with rows denoting the training dataset and columns denoting the evaluation dataset. For computational efficiency, we evaluate each source–target pair on a 10% sample (fixed seed) of the corresponding target test set for DeepArrhythmia with threshold-induced routing. Nothing from the target dataset is used for training, tuning, normalization, or




(a) Subject-disjoint DS1/DS2 split on MIT-BIH.

(b) Random split on MIT-BIH.

Figure 6: Comparison of MIT-BIH evaluation outcomes under DS1/DS2 and random split protocols.

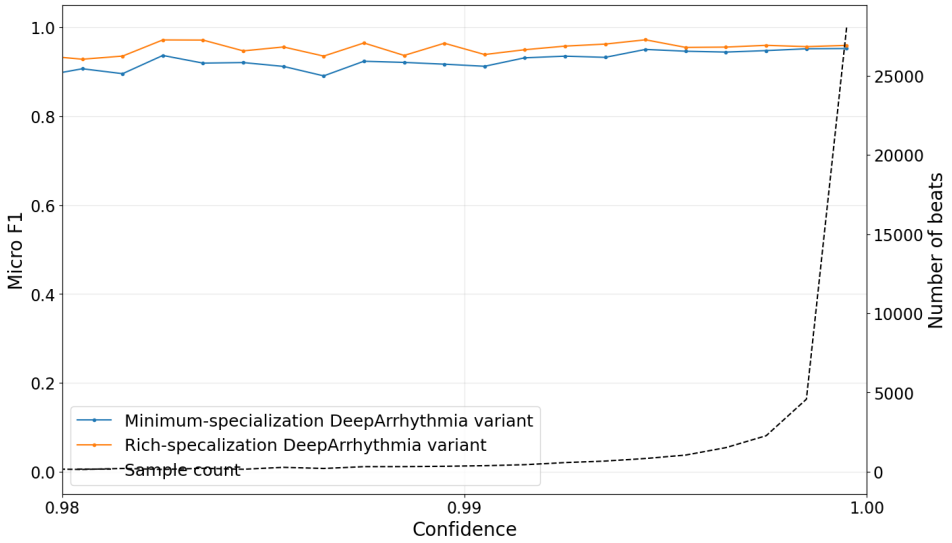



Figure 7: Relationship between prediction confidence and segment-level Micro-F1 on the MIT-BIH dataset. The x-axis shows the confidence score assigned to each segment by the minimal-specialization DeepArrhythmia variant, whereas the y-axis reports segment-level Micro-F1 for both the minimal- and rich-specialization DeepArrhythmia variants. We additionally visualize the number of samples at each confidence level to characterize the distribution of evaluation segments.

threshold selection: DeepArrhythmia keeps its per-dataset checkpoint and routing policy, the 1D ResNet uses the source's normalization statistics, and the SVM uses its source-fitted scaler. In-domain performance is highest for all four target datasets, with matched train–test Micro-F1 values of 0.9521 on MITBIH, 0.9527 on Superventricular, 0.9791 on INCART, and 0.9023 on VitalDB. Cross-dataset transfer is nevertheless highly uneven. Models trained on Superventricular and INCART generalize comparatively well to MITBIH and to each other, whereas the VitalDB-trained model transfers poorly overall, most notably to Superventricular (0.0935). MITBIH-trained models transfer reasonably to Superventricular (0.9272) but degrade substantially on INCART and VitalDB (0.5540 and 0.5514, respectively). Overall, these results indicate substantial dataset shift and pronounced source–target asymmetry despite strong in-domain accuracy.

To assess whether this behavior is specific to our framework or an intrinsic property of the task, we run the identical cross-dataset protocol on two representative baselines: the raw-signal 1D ResNet (Table 14) and the feature-based SVM (Table 15). Both baselines show the same source–target asymmetry, confirming that the shift is driven by dataset factors (sampling rate, lead configuration,

Table 13: Cross-dataset generalization performance measured by Micro-F1.

| Train \ Test | MITBIH | Superventricular | INCART | VitalDB |
|---|---|---|---|---|
| MITBIH | 0.9521 | 0.9272 | 0.5540 | 0.5514 |
| Superventricular | 0.9351 | 0.9527 | 0.8711 | 0.6482 |
| INCART | 0.9243 | 0.8942 | 0.9791 | 0.6735 |
| VitalDB | 0.5835 | 0.0935 | 0.7214 | 0.9023 |

Table 14: Cross-dataset generalization performance of the 1D ResNet baseline measured by Micro-F1.

| Train \ Test | MITBIH | Superventricular | INCART | VitalDB |
|---|---|---|---|---|
| MITBIH | 0.7596 | 0.8911 | 0.8635 | 0.6714 |
| Superventricular | 0.0743 | 0.8135 | 0.4786 | 0.5827 |
| INCART | 0.4199 | 0.6787 | 0.8693 | 0.5204 |
| VitalDB | 0.1012 | 0.0759 | 0.1299 | 0.7859 |

acquisition device, and class priors) rather than by any single model. Crucially, when trained on informative source datasets such as Superventricular and INCART, DeepArrhythmia transfers to unseen target datasets more effectively than either baseline. For the Superventricular-trained model, DeepArrhythmia attains a Micro-F1 of 0.9351 on MITBIH and 0.8711 on INCART, compared with 0.0743 and 0.4786 for the 1D ResNet and 0.8827 and 0.8624 for the SVM. The INCART-trained model shows the same advantage, reaching 0.9243 on MITBIH and 0.8942 on Superventricular, versus 0.4199 and 0.6787 for the ResNet and 0.8825 and 0.8886 for the SVM. Averaged over all cross-dataset transfers originating from these two sources, DeepArrhythmia achieves a mean Micro-F1 of 0.824, ahead of the SVM (0.811) and well ahead of the 1D ResNet (0.459). The SVM's figure is inflated by off-diagonal collapse to majority-class prediction (Macro-F1 $0.649 \rightarrow 0.294$), whereas DeepArrhythmia keeps detecting abnormal beats. Over all twelve off-diagonal transfers, DeepArrhythmia retains a mean Micro-F1 of 0.698 versus 0.457 for the 1D ResNet. DeepArrhythmia retains abnormal-beat detection on most transfers; its worst case is confined to the VitalDB-trained direction, where it over-predicts a single abnormal class. These results indicate that, given a sufficiently informative training distribution, DeepArrhythmia yields stronger cross-dataset generalization than both a raw-signal deep network and a feature-based classifier under matched conditions.

Table 15: Cross-dataset generalization performance of the SVM baseline measured by Micro-F1.

| **Train \ Test** | **MITBIH** | **Superventricular** | **INCART** | **VitalDB** |
|---|---|---|---|---|
| MITBIH | 0.9366 | 0.8886 | 0.9384 | 0.6735 |
| Superventricular | 0.8827 | 0.9374 | 0.8624 | 0.6735 |
| INCART | 0.8825 | 0.8886 | 0.9845 | 0.6735 |
| VitalDB | 0.8827 | 0.4505 | 0.8624 | 0.7501 |

## K Knowledge Distillation for Morphology Analyzer

To obtain a lightweight yet effective morphology analyzer, we adopt a knowledge distillation framework [Hsieh et al., 2023]. Specifically, a teacher model is provided with the ground-truth beat label, ECG images, beat position, and related ECG features, and is prompted to generate morphology-focused explanations for each beat. During the inference,the student Morphology Analyzer is label-free. In our setup, Gemini 3.1 serves as the teacher model and Qwen3.5-4B serves as the student model. For reproducibility, we provide the prompt format below.

```
# This is a 10-second ECG segment from the MIT-BIH Arrhythmia Database
# (Record 100, Segment 0). The signal uses detailed AAMI beat labels.
# The following abnormal beats are annotated (time measured from segment start):

- t = 5.68 s -> label: A, peak amplitude: 1.255 mV,
  RR interval to previous beat: 0.656 s

## Normal/reference beats in the same segment:
- t = 0.21 s -> label: N, peak amplitude: 1.160 mV,
  RR interval to previous beat: N/A
- t = 1.03 s -> label: N, peak amplitude: 1.315 mV,
  RR interval to previous beat: 0.814 s
- t = 1.84 s -> label: N, peak amplitude: 1.350 mV,
  RR interval to previous beat: 0.814 s
- t = 2.63 s -> label: N, peak amplitude: 1.235 mV,
  RR interval to previous beat: 0.789 s
- t = 3.42 s -> label: N, peak amplitude: 1.210 mV,
  RR interval to previous beat: 0.789 s
- t = 4.21 s -> label: N, peak amplitude: 1.275 mV,
  RR interval to previous beat: 0.789 s
- t = 5.03 s -> label: N, peak amplitude: 1.310 mV,
  RR interval to previous beat: 0.817 s
- t = 6.67 s -> label: N, peak amplitude: 1.305 mV,
  RR interval to previous beat: 0.994 s
- t = 7.52 s -> label: N, peak amplitude: 1.285 mV,
  RR interval to previous beat: 0.842 s
- t = 8.33 s -> label: N, peak amplitude: 1.280 mV,
  RR interval to previous beat: 0.811 s
- t = 9.12 s -> label: N, peak amplitude: 1.200 mV,
  RR interval to previous beat: 0.792 s
- t = 9.89 s -> label: N, peak amplitude: 1.205 mV,
  RR interval to previous beat: 0.769 s

## Task Instructions:
For each annotated beat listed above, inspect the waveform at the given time
position and explain the specific morphological features that justify its label
(e.g., wide/narrow QRS, premature timing, abnormal P-wave, T-wave inversion,
compensatory pause, amplitude differences, etc.). Use the provided peak
amplitudes and RR intervals for both abnormal and normal beats as quantitative
support, especially when comparing premature or wide-complex beats against
nearby normal reference beats. Be concise, refer to exact time positions,
and avoid disclosing information that directly reveals the class label.
```

# L Example of Complete Pipeline for DeepArrhythmia

To illustrate the end-to-end behavior of DeepArrhythmia, we provide a concise example interaction pipeline below.

```
human: Given the ECG segment. Identify the arrhythmia class for each R peak.
Use the format [position:class] for each beat, separated by spaces.

LLM: Call_tools[Peak Detector]

Tool_output: R peaks are located at the following sample positions:
77 370 663 947 1231 1515 1809 2045 2403 2706 2998 3283 3560.

LLM: [77:N] [370:N] [663:N] [947:N] [1231:N] [1515:N] [1809:N] [2045:N] [2403:N] [2706:N] [2998:N]
[3283:N] [3560:N]

LLM: Call_tools[Confidence Calculator]

Tool_output: The confidence is 0.989392. which is lower than Threshold 0.990529, Call Feature Extractor
and Morphology Analyzer.

LLM: Call_tools[Feature Extractor]

Tool_output: Best ECG-classification features from config RR+norm_RR+HOS+myMorph (C=10): RR [pre, next,
local, global] in samples describes beat timing, normalized RR [pre, next, local, global] shows timing
relative to the record trend, peak amplitude (mV) captures R-peak height, HOS summarizes higher-order
statistics of beat-shape intervals, and myMorph measures distances to representative morphology landmarks:

[77:RR=77,293,77,77;norm_RR=0.2693,1.0247,0.2694,0.2688;amp=1.1600;HOS=[1.1053,-0.3433,1.2196,0.0950,-0.2
560,4.6480,2.4145,2.9714,2.7759,2.2308];myMorph=[0.9410,1.0038,0.9457,1.0186]]

[370:RR=293,293,185,77;norm_RR=1.0246,1.0247,0.6472,0.2688;amp=1.3150;HOS=[-0.4416,0.3670,0.9802,-0.1661,
0.1701,2.0221,2.4089,2.4058,2.0685,2.0092];myMorph=[0.9454,1.0031,0.9396,1.0078]]

[663:RR=293,284,221,185;norm_RR=1.0246,0.9932,0.7731,0.6458;amp=1.3500;HOS=[-0.7455,1.1639,1.2957,-0.0901
,-0.3184,2.8724,3.3559,3.1715,2.4122,1.8911];myMorph=[0.9140,1.0022,0.9523,1.0351]]

[947:RR=284,284,237,221;norm_RR=0.9931,0.9932,0.8282,0.7715;amp=1.2350;HOS=[-0.4047,-1.4507,1.2064,0.0547
,0.0933,2.7067,6.1521,2.9994,2.0822,2.1874];myMorph=[0.9015,1.0039,0.9182,0.9532]]

[1231:RR=284,284,246,237;norm_RR=0.9931,0.9932,0.8613,0.8265;amp=1.2100;HOS=[-0.5737,-0.0519,1.3420,-0.47
98,-0.7527,3.0223,2.1574,3.3234,2.4888,3.3500];myMorph=[0.9323,1.0023,0.9439,1.0004]]

[1515:RR=284,294,252,246;norm_RR=0.9931,1.0282,0.8833,0.8595;amp=1.2750;HOS=[0.2622,1.3791,1.2894,0.2687,
0.2932,2.5440,5.3323,3.1648,1.9164,2.1889];myMorph=[0.9039,1.0038,0.9501,1.0355]]

[1809:RR=294,236,258,252;norm_RR=1.0281,0.8254,0.9041,0.8815;amp=1.3100;HOS=[-0.3015,0.7845,1.1879,0.2678
,-0.0181,2.2311,3.6662,2.8801,2.3006,1.7732];myMorph=[0.9045,1.0019,0.9446,1.0376]]

[2045:RR=236,358,256,258;norm_RR=0.8253,1.2520,0.8943,0.9022;amp=1.2550;HOS=[-0.7270,-0.4906,1.1749,-0.51
42,-0.3069,2.6356,2.8848,2.8712,3.0628,2.4644];myMorph=[0.8997,1.0033,0.9322,1.0002]]

[2403:RR=358,303,267,256;norm_RR=1.2519,1.0597,0.9341,0.8924;amp=1.3050;HOS=[-0.2710,0.8416,1.0584,0.2052
,-0.3983,1.9386,3.6821,2.5817,2.2463,3.0717];myMorph=[0.9381,1.0025,0.9712,1.0192]]

[2706:RR=303,292,271,267;norm_RR=1.0596,1.0212,0.9466,0.9321;amp=1.2850;HOS=[-0.1878,0.8982,1.2062,-0.490
5,-0.1132,1.8725,4.3606,2.9380,2.6907,2.3904];myMorph=[0.9376,1.0024,0.9540,1.0004]]

[2998:RR=292,285,292,271;norm_RR=1.0211,0.9967,1.0219,0.9446;amp=1.2800;HOS=[-0.9765,1.4327,1.3017,-0.458
0,-0.2072,3.3586,4.3886,3.1866,2.7695,2.8285];myMorph=[0.9363,1.0026,0.9572,1.0056]]

[3283:RR=285,277,291,273;norm_RR=0.9966,0.9688,1.0191,0.9514;amp=1.2000;HOS=[-0.9414,0.8672,1.2883,-0.102
3,0.1537,3.7850,3.5034,3.1642,1.8658,2.7679];myMorph=[0.9032,1.0026,0.9443,0.9939]]

[3560:RR=277,303,290,274;norm_RR=0.9687,1.0597,1.0135,0.9551;amp=1.2050;HOS=[-0.7466,1.2535,1.2960,0.1509
,0.5796,3.2926,4.4371,3.1819,2.4824,2.8179];myMorph=[0.9191,1.0023,0.9366,1.0154]]

LLM: Call_tools[Morphology Analyzer]

Tool_output: Abnormal Beat: t = 5.68 s

Premature Timing: The RR interval of 0.656 s is significantly shorter than the preceding and subsequent
baseline RR intervals, approximately 0.79 to 0.82 s. This beat occurs earlier than expected in the
underlying rhythm.

Morphology: The QRS complex is narrow and similar in width to the surrounding beats. The preceding P wave
is likely abnormal or hidden within the preceding T wave. The peak amplitude is 1.255 mV, which is within
the range of the surrounding beats in this segment.
```

Reference Beats: t = 0.21 s, 1.03 s, 1.84 s, 2.63 s, 3.42 s, 4.21 s, 5.03 s

Consistent Morphology: These beats form a stable run. They exhibit narrow QRS complexes with consistent upright P waves preceding each QRS. Their RR intervals are relatively regular, clustering around 0.789 s to 0.817 s. Their amplitudes are also consistent, ranging from 1.160 mV to 1.350 mV.

t = 6.67 s

Compensatory Pause: This beat follows the premature beat at 5.68 s. The RR interval from the premature beat to this beat is 0.994 s, which is notably longer than the baseline rhythm. This represents a pause before the rhythm resumes its regular pattern. Morphology returns to the baseline pattern.

t = 7.52 s, 8.33 s, 9.12 s, 9.89 s

Return to Baseline: After the pause, the rhythm returns to the stable baseline pattern seen earlier. RR intervals normalize, ranging from 0.842 s to 0.769 s, and QRS morphology remains narrow and consistent, with stable amplitudes of approximately 1.200 mV to 1.285 mV.

Summary: The beat at 5.68 s is identified as abnormal because it occurs earlier than expected, has a narrow QRS complex, and is followed by a longer RR interval. The remaining beats show consistent narrow QRS morphology, associated P waves, and regular timing.

LLM: [77:N] [370:N] [663:N] [947:N] [1231:N] [1515:N] [1809:N] [2045:SVEB] [2403:N] [2706:N] [2998:N] [3283:N] [3560:N]

## M Detail Result Analysis for DeepArrhythmia

Table 16 and Figure 8 jointly characterize the class-wise behavior of the DeepArrhythmia threshold-induced routed variant across the four evaluation datasets, with the feature-based SVM and the raw-signal 1D ResNet included in Table 16 as reference baselines. Table 16 reports precision, recall, and F1 for each class, whereas Figure 8 visualizes the corresponding confusion patterns for DeepArrhythmia. Across all datasets, DeepArrhythmia recognizes normal beats reliably, with F1 values ranging from 0.9396 to 0.9908, indicating that the model preserves high accuracy on the dominant class; the 1D ResNet is lower on this class (0.8499–0.9234), whereas the SVM is comparable (0.8224–0.9928). Ventricular beats are also classified robustly by DeepArrhythmia, with F1 scores above 0.78 on all four datasets, whereas the 1D ResNet is markedly weaker on this class (F1 between 0.4027 and 0.8019), underscoring that access to raw signal alone does not guarantee reliable ventricular discrimination. Performance on supraventricular beats is more variable: for DeepArrhythmia it is weakest on MIT-BIH Arrhythmia (F1 = 0.3922) but improves substantially on MIT-BIH Supraventricular, INCART, and VitalDB, where the corresponding F1 scores reach 0.6726, 0.8117, and 0.8532, respectively. On this class DeepArrhythmia is the strongest of the three methods on every dataset; the 1D ResNet is consistently the weakest (e.g., F1 = 0.0728 on MIT-BIH and 0.2174 on INCART), and the SVM lies in between.

The principal limitation concerns fusion beats. As shown in Table 16, absolute performance for class F remains poor for all three methods on the datasets where meaningful evaluation is possible. This weakness is expected given the extreme rarity and heterogeneity of fusion morphologies, but the comparison with the baselines is still informative: on MIT-BIH, DeepArrhythmia attains the highest F1 for class F (0.0659), ahead of both the 1D ResNet (0.0404) and the SVM (0.0253), and on INCART it reaches 0.0992 while the 1D ResNet attains only 0.0603 and the SVM remains at 0.0000. We do not report F for MIT-BIH Supraventricular or VitalDB because the corresponding evaluation splits contain fewer than five fusion-beat examples, making class-wise estimates too unstable for meaningful comparison. On MIT-BIH, all three methods are evaluated on the identical DS2 test split ($\approx$50k beats: $\approx$44k N, $\approx$2k S, $\approx$3.2k V, and only 388 F). Taken together, these results show that the difficulty of the minority classes (S and F) is shared across a feature-based classifier, a deep raw-signal network, and DeepArrhythmia, and is therefore intrinsic to the extreme class imbalance of beat-level ECG classification rather than specific to our framework; among the three methods, DeepArrhythmia consistently attains the strongest minority-class performance. The confusion matrices in Figure 8 are consistent with these quantitative findings: residual errors are predominantly misses toward the normal class rather than confusion among abnormal classes. On MIT-BIH, 1,173 of 1,833 S beats (64%) are predicted as N (32% correct), because S beats preserve a near-normal QRS and are distinguished mainly by timing and P-wave cues; 354 of 388 F beats (91%) are predicted as N (5% as V), since a fusion beat is a blend of a normal and a ventricular complex. The diagonal corresponding to normal beats remains dominant across datasets.

To quantify statistical uncertainty, we compute 95% bootstrap confidence intervals for MIT-BIH Macro-F1: DeepArrhythmia 0.593 [0.583, 0.603], SVM 0.523 [0.516, 0.530], and 1D ResNet 0.362 [0.356, 0.367]; the intervals do not overlap. Under class-balanced evaluation (balanced accuracy, i.e., mean per-class recall) on MIT-BIH, the ordering is preserved: DeepArrhythmia 0.568, SVM 0.516, and 1D ResNet 0.449.

The rendered waveform image is derived from the raw signal and therefore carries no independent information, yet removing it lowers MIT-BIH Macro-F1 from 0.5929 to 0.5394 (Table 2). A per-beat analysis shows that the beats recovered by the image are overwhelmingly abnormal beats that the model without the image labels as normal, chiefly S beats: 138 of the 143 recovered S beats were labelled N without the image. These are premature beats (normalized pre-RR $\approx$0.71) with a near-normal QRS, for which RR timing alone is insufficient; the image provides a complementary spatial representation of P-wave presence and subtle QRS shape that the scalar features compress away.

## N Faithfulness Evaluation

To assess whether the generated explanations reflect evidence used by the model, we perform an evidence-deletion test on the MIT-BIH test set using 160 beats stratified by cue (QRS 53, P-wave 54, RR 53). For numerical and textual morphology analyses that cite QRS width, P-wave evidence, or RR-interval regularity, we apply targeted perturbations that selectively disrupt the cue explicitly

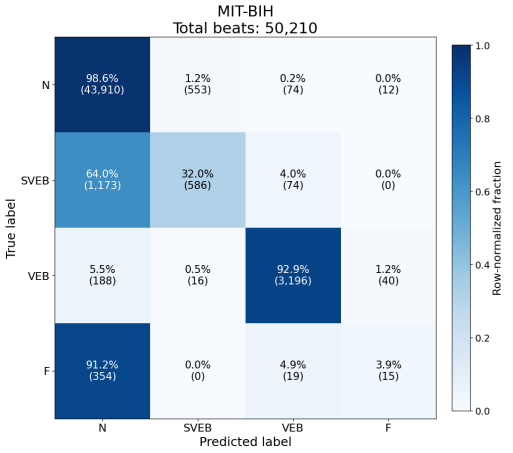
Table 16: Class-wise precision, recall, and F1 for DeepArrhythmia, SVM, and 1D ResNet across the four evaluation datasets. '–' indicates that the corresponding class is not reported because the evaluation split contains fewer than five beats.

| Model | Class | MIT-BIH | | | MIT-BIH-SUP | | | INCART | | | VitalDB | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | Prec. | Rec. | F1 | Prec. | Rec. | F1 | Prec. | Rec. | F1 | Prec. | Rec. | F1 |
| DeepArrhythmia | N | 0.9624 | 0.9857 | 0.9739 | 0.9872 | 0.9745 | 0.9808 | 0.9869 | 0.9947 | 0.9908 | 0.9121 | 0.9689 | 0.9396 |
| | S | 0.5074 | 0.3197 | 0.3922 | 0.6244 | 0.7289 | 0.6726 | 0.7849 | 0.8405 | 0.8117 | 0.9159 | 0.7985 | 0.8532 |
| | V | 0.9503 | 0.9291 | 0.9396 | 0.7907 | 0.8176 | 0.8040 | 0.9602 | 0.9154 | 0.9373 | 0.8328 | 0.7379 | 0.7825 |
| | F | 0.2239 | 0.0387 | 0.0659 | – | – | – | 0.3333 | 0.0583 | 0.0992 | – | – | – |
| SVM | N | 0.9533 | 0.9802 | 0.9665 | 0.9735 | 0.9714 | 0.9724 | 0.9908 | 0.9949 | 0.9928 | 0.8340 | 0.8111 | 0.8224 |
| | S | 0.3791 | 0.1415 | 0.2060 | 0.6096 | 0.5855 | 0.5973 | 0.8068 | 0.7410 | 0.7725 | 0.5965 | 0.6359 | 0.6155 |
| | V | 0.8696 | 0.9199 | 0.8938 | 0.7017 | 0.7616 | 0.7304 | 0.9596 | 0.9562 | 0.9579 | 0.3794 | 0.3470 | 0.3625 |
| | F | 0.0328 | 0.0206 | 0.0253 | – | – | – | 0.0000 | 0.0000 | 0.0000 | – | – | – |
| 1D ResNet | N | 0.9366 | 0.7919 | 0.8582 | 0.9872 | 0.8244 | 0.8985 | 0.9880 | 0.8667 | 0.9234 | 0.9177 | 0.7914 | 0.8499 |
| | S | 0.0591 | 0.0946 | 0.0728 | 0.2319 | 0.7614 | 0.3555 | 0.1313 | 0.6334 | 0.2174 | 0.6534 | 0.7791 | 0.7107 |
| | V | 0.3342 | 0.8230 | 0.4753 | 0.5542 | 0.6829 | 0.6119 | 0.7123 | 0.9174 | 0.8019 | 0.2809 | 0.7111 | 0.4027 |
| | F | 0.0265 | 0.0851 | 0.0404 | – | – | – | 0.0324 | 0.4320 | 0.0603 | – | – | – |

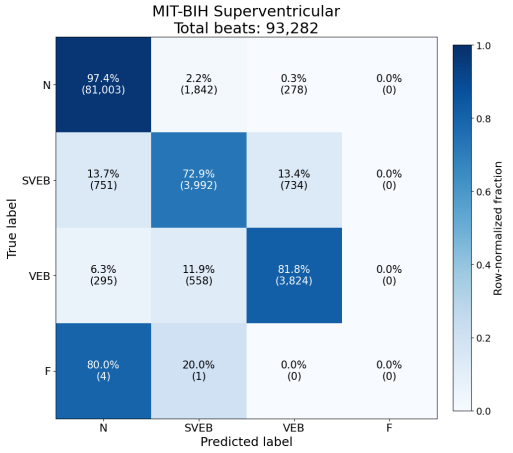
Table 17: Faithfulness evaluation via evidence deletion. We report classification accuracy after targeted perturbations and matched control perturbations. Targeted perturbations disrupt the cue explicitly referenced in the explanation, while control perturbations are applied to unrelated regions. Results are shown overall and by cue type. Lower accuracy under targeted perturbations indicates that the model relies on the cited evidence.

| Cue type | Targeted accuracy | Control accuracy | Sample size |
|---|---|---|---|
| Overall | 3 / 160 (1.88%) | 148 / 160 (92.50%) | 160 |
| QRS | 1 / 53 (1.89%) | 51 / 53 (96.23%) | 53 |
| P_WAVE | 1 / 54 (1.85%) | 49 / 54 (90.74%) | 54 |
| RR | 1 / 53 (1.89%) | 48 / 53 (90.57%) | 53 |

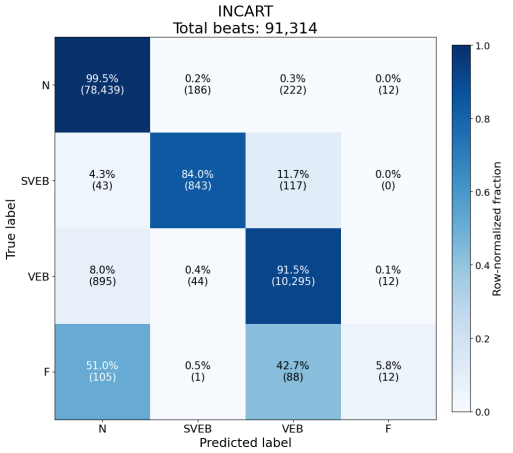
referenced in the explanation and compare the resulting classification accuracy against matched control perturbations applied to unrelated regions. As shown in Table 17, targeted perturbations cause a dramatic drop in accuracy relative to control perturbations, both overall (1.88% vs. 92.50%) and within each cue category. These results suggest that the generated explanations are at least partially faithful, in the sense that the model's predictions depend strongly on the evidence explicitly cited in the explanation.



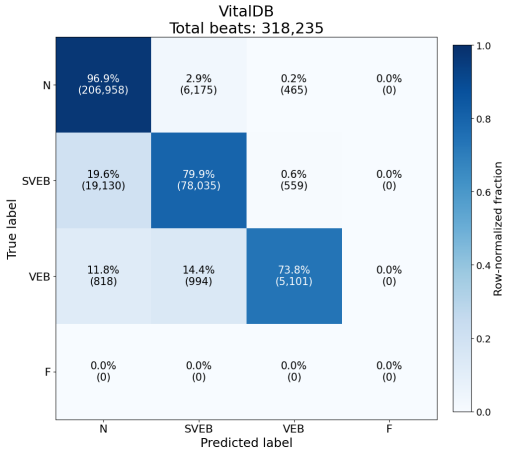

(a) Confusion matrix on MIT-BIH Arrhythmia.




(b) Confusion matrix on MIT-BIH Supraventricular.

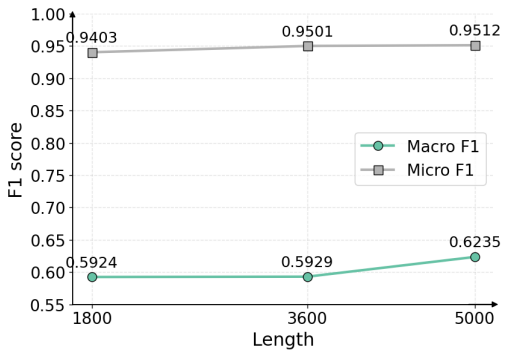



(c) Confusion matrix on INCART.

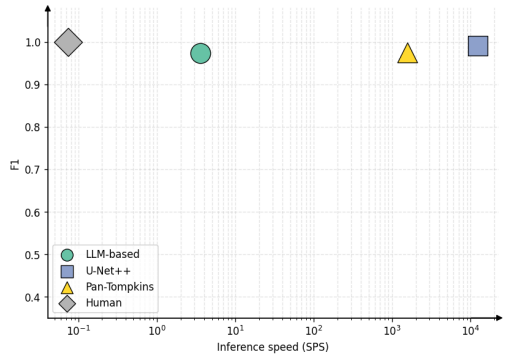



(d) Confusion matrix on VitalDB.

Figure 8: Confusion-matrix panels for DeepArrhythmia across four datasets.




(a) Impact of context length on DeepArrhythmia performance.




(b) Peak detector trade-off between throughput (samples per second) and F1 score.

Figure 9: Sensitivity analyses for context length and peak-detector selection.

Table 18: Boundary versus interior beat classification on the MIT-BIH test set. Boundary beats (the first and last beat of each 10-second window) are only marginally harder overall, and the residual gap is concentrated on the context-dependent abnormal classes, while the normal class is essentially unaffected.

| **Beat group** | **#Beats** | **Micro-F1** | **Macro-F1** | **N** | **S** | **V** | **F** |
|---|---|---|---|---|---|---|---|
| Boundary (first/last) | 8,066 | 0.9435 | 0.5673 | 0.970 | 0.360 | 0.901 | 0.038 |
| Interior | 41,981 | 0.9521 | 0.5987 | 0.975 | 0.400 | 0.950 | 0.070 |

Table 19: Effect of beat position within the fixed context window on abnormal (S/V/F) beat classification, MIT-BIH test set. Boundary beats are the weakest; interior beats are stronger; and re-classifying each boundary beat in a window re-centered on it (overlapping/sliding-window inference) recovers the gap and matches interior-beat performance. "Boundary" and "Re-centered" are the same 867 beats.

| **Beat position** | **#Beats** | **Acc.** | **Macro-F1** | **S** | **V** | **F** |
|---|---|---|---|---|---|---|
| Boundary (original) | 867 | 0.606 | 0.466 | 0.445 | 0.915 | 0.038 |
| Interior (reference) | 4,680 | 0.683 | 0.507 | 0.490 | 0.961 | 0.072 |
| Re-centered (sliding) | 867 | **0.694** | **0.584** | **0.531** | **0.966** | **0.254** |

# O Impact of Context Window Length and Boundary Truncation

Figure 9a evaluates the sensitivity of DeepArrhythmia with rich-evidence to input context length. Increasing the context window from 1,800 samples (5 s) to 3,600 samples (10 s) yields consistent gains, with Macro-F1 improving from 0.5924 to 0.5929 and Micro-F1 improving from 0.9403 to 0.9501. Extending the context further to 5,000 samples provides an additional improvement, increasing Macro-F1 to 0.6235 and Micro-F1 to 0.9512. These results indicate that longer temporal context strengthens rhythm-level reasoning and particularly benefits class-balanced performance.

To assess the effect of fixed-window boundary truncation, we compare beats at the segment boundary (the first and last beat of each 10-second window) against interior beats on the MIT-BIH test set (Table 18). Boundary beats are only marginally harder overall—Micro-F1 is $0.9435$ versus $0.9521$ for interior beats (a $0.9$-point gap), and the dominant normal class is essentially unaffected ($0.970$ vs. $0.975$). The residual gap is instead concentrated on the abnormal classes that most depend on cross-beat context: V-class F1 drops from $0.950$ to $0.901$ and the S and F classes decline similarly, because a boundary beat lacks a complete preceding/following RR interval and neighbouring morphology. This robustness is largely attributable to the shift-based re-anchoring augmentation, which exposes each beat at many within-window positions during training.

Because this degradation is confined to the abnormal classes—the normal class is already unaffected by boundary position—we test the overlapping-window remedy precisely where it could help: the abnormal (S/V/F) boundary beats. We therefore select every S, V, or F beat that occurs first or last in a window and can be re-centered within the record, yielding $867$ target beats (of which 3 lie too close to a record edge to center and are excluded). For each such beat we re-classify it inside a $10$-second window *re-centered* on it, emulating an overlapping (sliding-window) inference scheme in which every beat is evaluated where it has full surrounding context; the $N$ class and interior beats are excluded because they show no boundary deficit to recover and would only dilute the measurement. As shown in Table 19, re-centering these boundary beats raises their accuracy from $0.606$ to $0.694$ and Macro-F1 from $0.466$ to $0.584$, recovering the boundary gap and matching interior-beat performance, with the largest gain on the context-dependent fusion class (F1 $0.038 \rightarrow 0.254$); 131 of the $867$ beats are corrected while only $54$ regress. This confirms that boundary degradation stems from missing context rather than a model deficiency, and that a lightweight overlapping-window scheme resolves it. We use non-overlapping windows in the main results for training and inference efficiency, since overlap multiplies the compute per beat.

Table 20: Class-wise decrease in F1 score ($\Delta$F1 = $F1_{clean}$ - $F1_{stress}$) under controlled peak-detection stress conditions: masking (a true R-peak is removed), mislocalization (a peak is shifted by 6–30 samples), and extra (a spurious R-peak is inserted). "–" indicates the interfered beat has no ground-truth label (masked or spurious). Positive values denote degradation; negative values indicate improvement. Class S is the most sensitive to mislocalization and spurious peaks, whereas under masking the collateral effect is largest for the minority F and V classes.

| Label | Mask Interfered | Mask Non-interfered | Mislocalize Interfered | Mislocalize Non-interfered | Extra Interfered | Extra Non-interfered |
|---|---|---|---|---|---|---|
| N | – | 0.004 | 0.020 | 0.014 | – | 0.028 |
| S | – | 0.019 | 0.256 | 0.136 | – | **0.302** |
| V | – | 0.043 | 0.137 | 0.025 | – | 0.072 |
| F | – | 0.056 | 0.039 | 0.033 | – | 0.050 |

# P Impact of Peak Detector

Peak detection is a critical upstream stage in beat-centered ECG analysis because errors in R-peak localization propagate directly to subsequent feature extraction, morphology reasoning, and beat-label alignment. To quantify this dependency, we compare representative detectors on 10-second MIT-BIH segments, including Pan–Tompkins [Pan and Tompkins, 1985], U-Net++ [Zhou et al., 2021], an LLM-based detector [Li et al., 2025], and human annotations as a reference.

Figure 9b summarizes the trade-off between detection quality, measured by F1 score using a 30 ms peak-matching tolerance [Li et al., 2025], and runtime throughput, measured in samples per second. Throughput is measured on the same single-GPU A6000 inference setup described in Appendix A. Among the automated methods, U-Net++ achieves the highest F1 score (0.9913) while also providing the fastest inference speed (12,379 segments per second); only the human-annotation reference performs better in detection quality. This favorable accuracy–efficiency profile motivates our use of U-Net++ as the default peak detector in DeepArrhythmia.

To further characterize the downstream consequences of detector errors, we evaluate three controlled perturbation settings on the rich-evidence variant of DeepArrhythmia, using a class-enriched 400-segment MIT-BIH subset (100 all-normal segments and 100 segments each containing S, V, or F beats): *masking*, which removes a true R peak; *mislocalization*, which shifts a detected peak to the left or right by 6 to 30 samples; and *extra*, which inserts a spurious R peak into a wide gap between two true beats. These perturbations isolate three common failure modes of practical peak detectors—missed detections, temporal offsets, and false detections—and allow us to examine how each propagates to beat-level classification.

Table 20 reveals three main patterns. First, for non-interfered collateral beats, mislocalization causes larger degradation than masking; directly masked beats are omitted from prediction and are therefore not directly comparable. The largest degradation occurs for supraventricular beats ($\Delta$F1 = 0.256), followed by ventricular beats ($\Delta$F1 = 0.137), whereas the effect is smaller for normal and fusion beats ($\Delta$F1 values of 0.020 and 0.039, respectively). This suggests that precise beat alignment is especially important for classes whose discrimination depends on localized morphology and rhythm context, particularly class S. Second, the collateral effect on non-interfered beats remains modest under masking ($\Delta$F1 between 0.004 and 0.056), but becomes larger under mislocalization, again most notably for class S ($\Delta$F1 = 0.136). A plausible interpretation is that a missed beat primarily removes one decision anchor, whereas a mislocalized beat can actively distort downstream feature extraction and morphology reasoning. Third, inserting a spurious peak produces the same qualitative signature but the strongest collateral effect on class S of all three settings ($\Delta$F1 = 0.302), while remaining modest overall (Micro-F1 $\Delta$ = 0.043); a false anchor injects an inconsistent RR pattern and morphology descriptor that most affects the timing- and context-dependent S class. Examining the spurious anchors themselves, only 34.6% are labelled normal, whereas 63.4% are assigned an arrhythmic class (52.0% V, 11.4% S). Because prediction is anchored to detected peaks, spurious detections are thus not robustly rejected—a limitation shared with detector-dependent beat-level pipelines that motivates tighter detector–classifier coupling in future work.

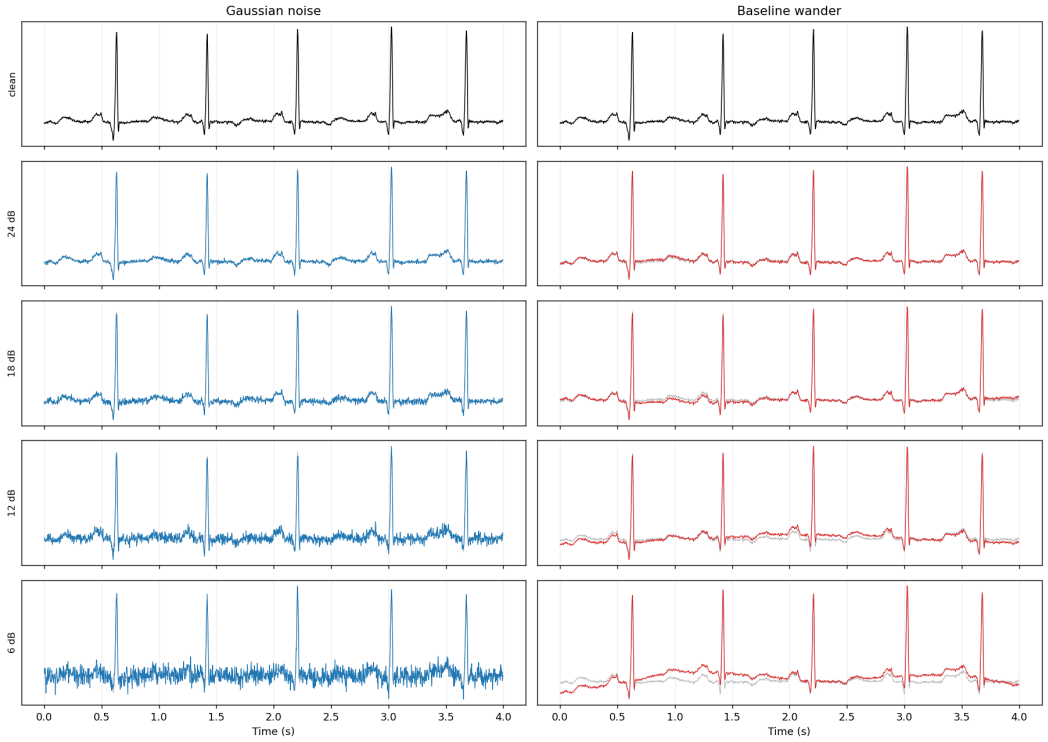


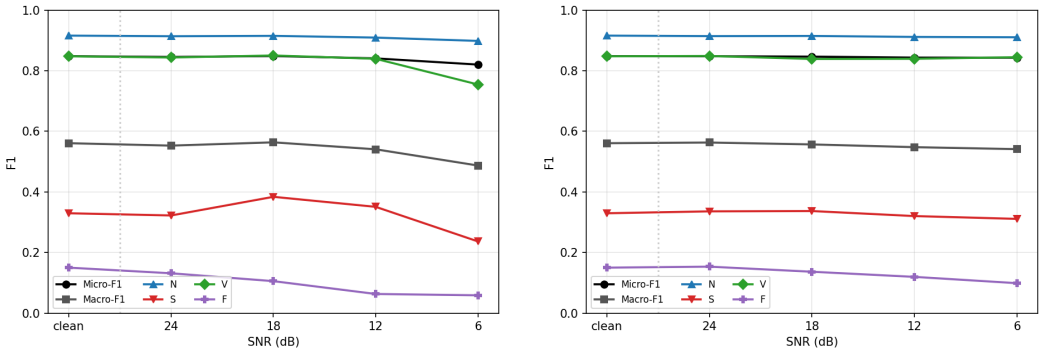

Figure 10: Illustration of the two noise types at each SNR on a representative 4-second MIT-BIH segment (lead MLII). The clean trace (top) is overlaid in grey on each corrupted signal for reference. Gaussian noise (left, blue) adds broadband high-frequency fluctuations, while baseline wander (right, red) adds low-frequency drift.

## Q Impact of Noise

A natural concern for any morphology-aware classifier is its resilience to signal corruption. We therefore quantify how beat-level accuracy responds to two canonical ECG disturbances—*additive Gaussian noise* and *baseline wander*—applied to each 10-second segment at signal-to-noise ratios (SNR) of 24, 18, 12, and 6 dB. Figure 10 illustrates the two disturbances at each level: at 24 dB the corruption is barely perceptible, whereas at 6 dB Gaussian noise obscures the P and T waves while leaving the QRS complex identifiable, and baseline wander induces a pronounced low-frequency drift of the isoelectric line.

Crucially, each perturbation is propagated through the entire input pipeline rather than a single modality: the corrupted waveform is re-encoded by the ECG-CoCa encoder, the ECG image is re-rendered from it, and the numerical morphology descriptors (peak amplitude, HOS, and myMorph) are re-extracted from it. We hold the R-peak anchors and RR intervals at their clean values so as to isolate the effect of *signal quality* from that of *peak-detection error*, which is examined separately in Appendix P. All conditions use the rich-evidence variant and are evaluated on the same 400-segment MIT-BIH subset as the peak-perturbation study (5,374 beats), ensuring that differences reflect noise sensitivity alone rather than sampling variation. Because the subset is class-enriched, the clean reference (Micro-F1 0.8482) sits below the 0.9501 of Table 1 for the same model, while Macro-F1 (0.5609 vs. 0.5929) is close. SNR is defined as $10 \log_{10}$(signal power / noise power) per segment and lead. Figure 11 visualizes the corresponding degradation curves.

Across every tested condition, DeepArrhythmia degrades gracefully rather than collapsing. Baseline wander is almost entirely absorbed: from the clean reference to the most severe 6 dB level, Micro-F1 declines by only 0.5 points ($0.848 \rightarrow 0.843$) and Macro-F1 by 2.0 points ($0.561 \rightarrow 0.541$). This is expected, as the median-filter baseline removal in preprocessing suppresses low-frequency drift while preserving beat morphology. Additive Gaussian noise is more disruptive yet remains well tolerated over the practical range: accuracy is statistically flat down to 12 dB, staying within one




(a) Additive Gaussian noise

(b) Baseline wander

Figure 11: Beat-level F1 (Micro, Macro, and per-class N/S/V/F) as a function of SNR under (a) additive Gaussian noise and (b) baseline wander, with the clean condition as the leftmost reference. Performance degrades gracefully and never collapses: baseline wander is almost fully absorbed, while Gaussian noise leaves accuracy essentially unchanged down to 12 dB and induces only a localized decline at 6 dB, concentrated on the morphology-dependent V class and the minority S and F classes.

Table 21: Beat-level Micro-F1 and Macro-F1 under additive Gaussian noise and baseline wander (rich-evidence variant, 400-segment MIT-BIH subset, 5,374 beats).

| | Gaussian noise | | Baseline wander | |
|---|---|---|---|---|
| SNR | Micro-F1 | Macro-F1 | Micro-F1 | Macro-F1 |
| Clean | 0.8482 | 0.5609 | 0.8482 | 0.5609 |
| 24 dB | 0.8456 | 0.5529 | 0.8480 | 0.5632 |
| 18 dB | 0.8487 | 0.5638 | 0.8467 | 0.5569 |
| 12 dB | 0.8409 | 0.5408 | 0.8435 | 0.5478 |
| 6 dB | 0.8204 | 0.4872 | 0.8431 | 0.5413 |

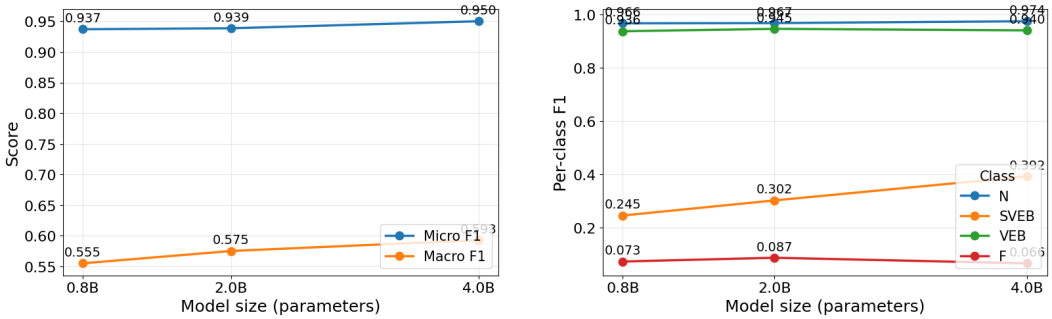
Micro-F1 point of the clean baseline, and a pronounced decline emerges only at the extreme 6 dB setting (Micro-F1 $-2.8$, Macro-F1 $-7.4$ points).

The structure of this degradation is consistent with the design of the framework. As Figure 11 shows, the loss is concentrated on classes whose discrimination depends on fine morphology: under Gaussian noise at 6 dB the V-class F1 falls from $0.848$ to $0.755$ and the minority S and F classes deteriorate, whereas the rhythm-anchored N class remains essentially stable ($0.916 \rightarrow 0.899$). Intuitively, noise corrupts the morphological evidence—the rendered waveform image and the HOS/myMorph descriptors—but not the timing-based RR intervals and peak anchors, so decisions grounded in rhythm context stay reliable. Taken together, these results demonstrate that DeepArrhythmia is robust to realistic levels of signal noise, exhibiting only graceful, morphology-localized degradation even at severe SNR (Table 21). Noise also lowers confidence, so a fixed threshold routes degraded segments to rich evidence more often; the routing mechanism therefore fails safe.

## R Impact of Backbone Scale

We assess backbone scale by training three DeepArrhythmia variants with rich-evidence that differ only in the Qwen3.5 backbone (0.8B, 2B, and 4B), while holding the remaining architecture and training settings fixed. All models are evaluated on the same held-out MIT-BIH split using beat-level Micro-F1, Macro-F1, and per-class F1.

Figure 12 indicates a clear positive scaling trend with increasing backbone size. At the aggregate level, beat-level Micro-F1 improves from 0.937 (0.8B) to 0.950 (4B), while Macro-F1 increases from 0.555 to 0.593. At the class level, the gains for N and VEB are relatively modest, rising from 0.966 to 0.974 and from 0.936 to 0.940, respectively, which is expected given that both classes already operate near ceiling performance. In contrast, SVEB benefits most substantially from increased




(a) Aggregate beat-level Micro-F1 and Macro-F1.

(b) Per-class beat-level F1.

Figure 12: Effect of Qwen3.5 backbone scale on DeepArrhythmia beat classification on MIT-BIH Arrhythmia. Performance increases with backbone size, primarily through improved *S*-class recognition.

model scale, with F1 improving from 0.245 to 0.392, suggesting that larger backbones primarily enhance recognition of the minority class. Fusion beats remain challenging across all model sizes, with F1 remaining close to zero throughout.

## S Interpretation Analysis

DeepArrhythmia provides an additional advantage beyond classification accuracy: improved interpretability. The morphology analyzer generates clinically grounded textual rationales that align model predictions with recognizable ECG patterns, thereby supporting transparent decision-making. Figure 13 shows the top bigram distributions of generated interpretations for each class on the MIT-BIH dataset. Three recurring diagnostic cues emerge.

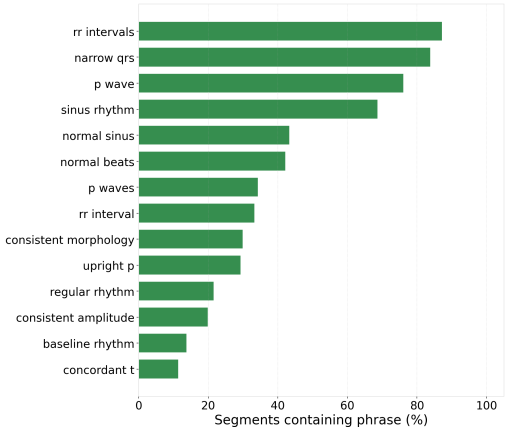
**QRS range** QRS-width descriptors are among the most frequent bigrams across all four classes. In particular, "narrow QRS" is more prevalent in normal and supraventricular beats, whereas "wide QRS" appears more frequently in ventricular beats. For fusion beats, narrow- and wide-QRS descriptors appear at comparable rates, which is consistent with their mixed morphology. This pattern is clinically plausible: wide-complex tachycardia (QRS $\geq$ 120 ms) is commonly associated with ventricular-origin rhythms, including monomorphic ventricular tachycardia, polymorphic ventricular tachycardia, and ventricular fibrillation [Desai and Hajouli, 2023].

**P wave** P-wave-related bigrams are highly ranked for normal and supraventricular classes, but are less prominent for ventricular beats. This trend is also physiologically meaningful: ventricular depolarization can dominate surface ECG morphology and obscure atrial activity, making P-wave visibility less reliable in ventricular rhythms [Goldberger et al., 2023]. Because P-wave detection remains challenging in noisy clinical recordings [Maršánová et al., 2019], multimodal visual-language reasoning may provide complementary support for identifying subtle atrial signatures.

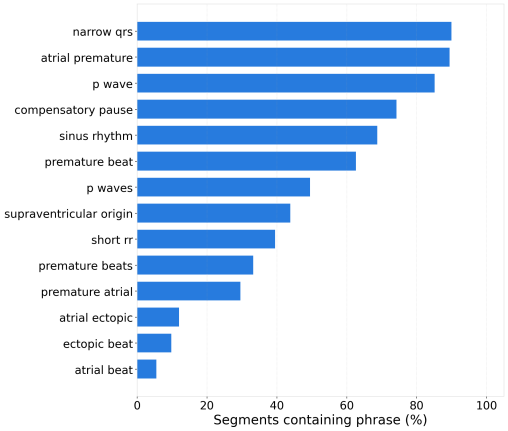
**RR interval** RR-interval regularity remains a central cue for rhythm discrimination. Normal beats are associated with regular intervals (e.g., Figure 13a), while premature atrial and ventricular events are more often linked to irregular interval patterns and compensatory pauses. Importantly, RR-interval evidence is inherently contextual and cannot be inferred reliably from isolated beats. This observation supports segment-level modeling, where temporal context is explicitly incorporated into classification.

## T Interpretation Examples

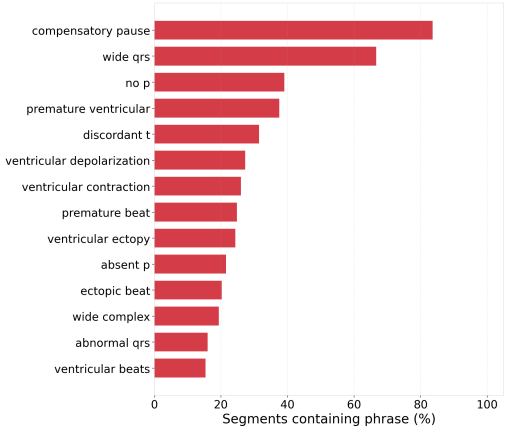
We present representative ECG segments with analyses generated by the morphology analyzer to analyze the interpretability of DeepArrhythmia predictions. In Figure 14, the model identifies a supraventricular beat using clinically relevant evidence, including premature timing, a narrow QRS complex, an abnormal or partially merged P wave, and a compensatory pause. The explanation further contrasts the target beat with neighboring normal beats through RR-interval and P-wave comparisons, providing transparent justification for the assigned class.



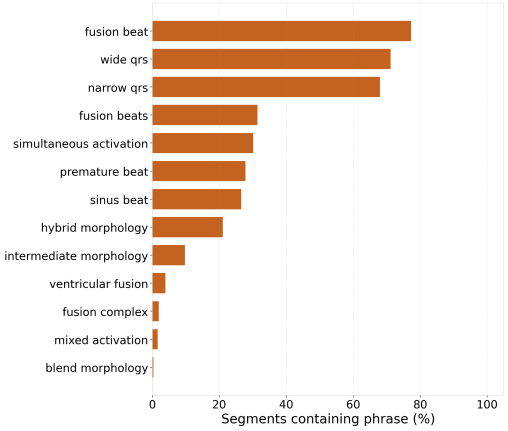

(a) Bigram distribution for class N.




(b) Bigram distribution for class S.




(c) Bigram distribution for class V.




(d) Bigram distribution for class F.

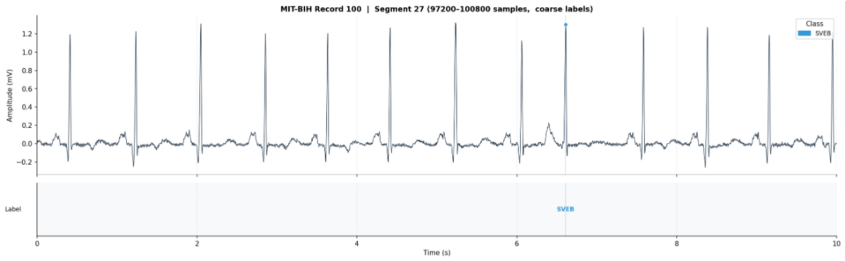
Figure 13: Class-wise bigram statistics extracted from interpretation captions across beat classes N, S, V, and F.

Figure 15 provides a second positive example with a similar reasoning pattern. In this case, the model emphasizes altered atrial activity (e.g., absent or poorly resolved P-wave morphology) together with elevated peak amplitude relative to nearby normal beats, supporting its class decision with both morphological and quantitative cues.

Figure 16 illustrates a failure case that clarifies current limitations. The first error occurs for an early beat where incomplete contextual information (e.g., limited visibility of the preceding RR interval and P-wave morphology) leads to a ventricular prediction based on partial evidence. The second error occurs in a noisy region, where signal corruption distorts morphology and RR-interval estimates, resulting in an incorrect classification. These examples indicate that performance remains sensitive to context truncation and local noise artifacts.

## U Effect of the Teacher Model in the Morphology Analyzer

Because the current teacher used for morphology-rationale distillation is Gemini 3.1 [Team et al., 2023], which is closed-source, we additionally compare it with an open-source alternative, Gemma 4-31B [Team et al., 2024]. This ablation isolates teacher choice while keeping the student architecture and downstream classification pipeline fixed. We report the Macro-F1 and Micro-F1 performance of the rich evidence variant of DeepArrhythmia on the MIT-BIH dataset in Table 22. The student distilled from Gemini 3.1 achieves higher Macro-F1 and Micro-F1 than the student distilled from Gemma 4 (0.593 vs. 0.572 and 0.950 vs. 0.944, respectively), indicating that teacher quality has




Abnormal Beat at t = 6.61 s
Premature Timing: The RR interval to the previous beat is 0.550 s, which is significantly shorter than the preceding normal RR intervals, ranging from 0.778 s to 0.828 s. This confirms that the beat occurs prematurely.

Narrow QRS: The QRS complex at this time is not widened. Its morphology and duration are consistent with the surrounding normal beats in the segment, suggesting that the abnormality is mainly reflected in timing and atrial activity rather than ventricular conduction.

Abnormal P wave: The P wave preceding this QRS complex is morphologically different from the uniform P waves observed before surrounding beats. It appears partially merged with the preceding T wave and has a slightly sharper, more pointed contour.

Compensatory Pause:The RR interval following this beat, from 6.61 s to 7.58 s, is 0.972 s, which is longer than the baseline RR interval of approximately 0.8 s. This longer pause indicates that the premature beat disrupts the otherwise regular rhythm.

Normal Beats for Contrast: The surrounding beats at approximately 0.41 s, 1.24 s, and 2.05 s show a regular rhythm and consistent morphology. Their RR intervals remain between 0.772 s and 0.828 s. Each beat is preceded by a uniform upright P wave, followed by a narrow QRS complex and a consistent T wave. The peak amplitudes of the surrounding beats range from 1.125 mV to 1.320 mV. The abnormal beat at 6.61 s has an amplitude of 1.300 mV, which falls within this range. Therefore, amplitude alone is not the main distinguishing feature.

Figure 14: Interpretation example from MIT-BIH Record 100 (segment 27). Morphology analyzer identifies a supraventricular beat using premature timing, narrow QRS morphology, and a compensatory pause relative to surrounding normal beats.

Table 22: Effect of teacher choice on downstream beat-classification performance of the morphology-analyzer student. Higher values are better.

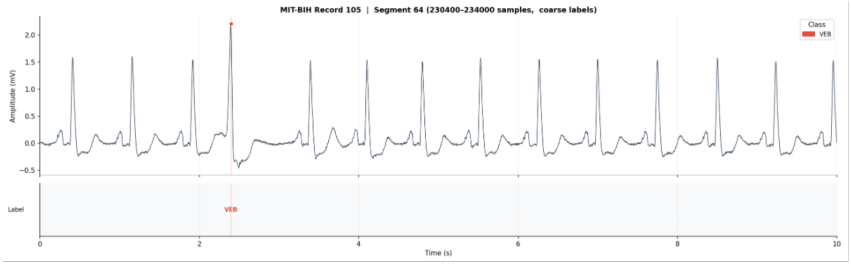
| Teacher model | Macro-F1 | Micro-F1 |
|---|---|---|
| Gemini 3.1 | 0.593 | 0.950 |
| Gemma 4 | 0.572 | 0.944 |

a measurable effect on the usefulness of the learned morphology evidence. Figure 17 provides an illustrative output from the Gemma-4-distilled student on MIT-BIH Record 100; the rationale is relatively ambiguous and less clinically specific, for example describing the P wave only as absent or merged.

# V Broader Impacts and Limitations

DeepArrhythmia illustrates the potential of tool-grounded agentic systems to improve both predictive performance and interpretability in ECG analysis. By integrating peak detection, feature extraction, and morphology-aware reasoning, the framework produces evidence-linked predictions rather than purely opaque outputs. This design may support clinical workflows by enabling faster beat-level screening, prioritizing suspicious segments for expert review, and providing traceable rationales aligned with established electrophysiological cues, including RR-interval dynamics, QRS morphology, and P-wave characteristics. Such support may be particularly valuable in resource-constrained settings where specialist availability is limited.

Beyond near-term clinical assistance, our results highlight a broader methodological opportunity: using a VLM as a central agent to orchestrate specialized physiological domain models. Rather than replacing decades of progress in physiological signal processing, this paradigm operationalizes that historical advancement by integrating established detectors, feature extractors, and morphology analyzers into a unified reasoning loop. The resulting architecture is inherently extensible: new domain tools can be added as modules while the central agent maintains cross-modal coordination and decision synthesis. In this sense, DeepArrhythmia represents an expandable AI-centered framework that is strengthened by domain knowledge, with potential applicability beyond ECG to other physio-




Abnormal beat At t = 2.40 s, this beat occurs earlier than expected. The RR interval to the previous beat is 0.481 s, which is significantly shorter than the surrounding baseline RR intervals ranging from 0.694 s to 0.761 s.

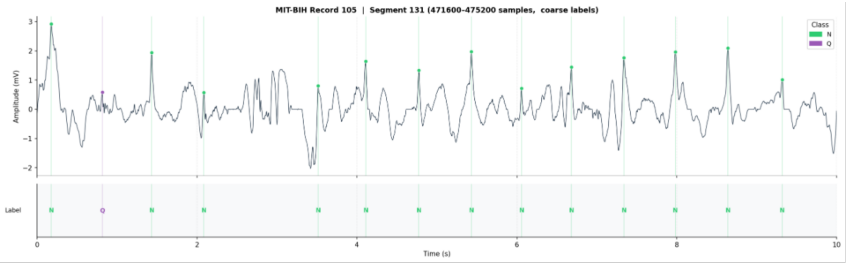
The QRS complex is visibly wider and more slurred than the sharp, narrow complexes of the surrounding beats. Its peak amplitude is 2.210 mV, which is much higher than the surrounding amplitude range of approximately 1.5 to 1.6 mV.

No clear P wave is visible before this QRS complex. The following beat at t = 3.39 s occurs after a prolonged RR interval of 0.997 s, indicating a pause after the early abnormal beat. Reference beats The surrounding reference beats show consistent morphology. They have sharp, narrow, and uniform QRS complexes. Their peak amplitudes are tightly clustered between 1.510 mV and 1.600 mV.

Their rhythm is relatively regular, with RR intervals ranging from 0.694 s to 0.761 s, excluding the longer pause after the abnormal beat. Each surrounding reference beat is preceded by a small and consistent P wave.

Selected observations The beat at t = 1.92 s is the last reference beat before the abnormal beat. Its RR interval of 0.761 s helps establish the baseline rhythm. The beat at t = 3.39 s is the first reference beat after the abnormal beat. It follows the prolonged RR interval of 0.997 s, while its morphology returns to the surrounding reference pattern. The beats at t = 0.41 s and t = 1.16 s represent the steady underlying rhythm. Their regular timing and consistent shape contrast clearly with the single abnormal beat at t = 2.40 s.

Figure 15: Second interpretation example from MIT-BIH Record 105 (segment 64). The model highlights altered P-wave morphology and elevated amplitude as evidence for its prediction.




Abnormal Beat: t = 0.18 s: This beat occurs very early in the segment. While the full preceding RR interval is not visible, the QRS complex is notably wide and abnormal in shape compared to the subsequent beats in the tracing. It has a significantly higher positive amplitude, reaching nearly 3 mV, and lacks a preceding P-wave.

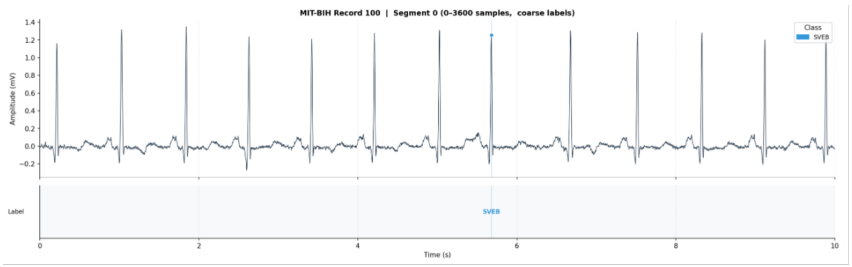
Abnormal Beat: t = 2.09 s: This beat appears slightly premature compared to the regular rhythm established by the subsequent beats. Its morphology is intermediate; the QRS complex is somewhat wider and more irregular than the later baseline beats at approximately 3.5 s, 4.5 s, and 5.5 s, but not as wide or abnormal as the early beat at 0.18 s.

Reference Beats: The remaining beats exhibit a consistent baseline morphology. They feature distinct upright P-waves, sharp and narrow QRS complexes, and regular T-waves. Based on the provided ECG segment, these beats provide a stable reference pattern for comparing the abnormal beats.

Figure 16: Failure-case example from MIT-BIH Record 105 (segment 131). Misclassifications are associated with incomplete temporal context for an early beat and morphology distortion under localized noise.

logical monitoring tasks. The peak detector, feature extractor, and morphology analyzer operate on fiducial points and their local context, so they transfer to other quasi-periodic signals (e.g., systolic peaks in PPG, J-peaks in BCG, pulse onsets in arterial blood pressure); only the ECG-CoCa encoder is modality-specific.

Several limitations, however, remain. First, performance is still heterogeneous across classes, and fusion beats remain difficult to detect because of extreme class imbalance and substantial morphological variability. Second, as shown in the interpretation examples, prediction quality degrades under incomplete temporal context and low signal quality, indicating sensitivity to context truncation and noise contamination. Third, although results are strong across four benchmark datasets, the present study is retrospective; therefore, prospective utility across institutions, devices, and patient populations is not yet established. Fourth, LLM-based components may reflect biases in training data and can generate plausible but imperfect explanations, so outputs should be used to assist—not




Reference Beats t = 0.21, 1.03, 1.84, 2.63, 3.42, 4.21, 5.03, 6.67, 7.52, 8.33, 9.12, 9.89 s

P wave: Present with consistent morphology and regularly preceding the QRS complex.

QRS complex: Narrow, less than 120 ms, with typical morphology. Peak amplitudes range from 1.160 mV to 1.350 mV. Timing: RR intervals are relatively regular, averaging approximately 0.80 s.

T wave: Positive polarity, with no inversion noted.

Overall interpretation: These beats show regular P QRS T sequences, stable morphology, and consistent timing, providing the baseline reference pattern for comparison.

Abnormal Beat t = 5.68 s

P wave: Absent or merged. There is no distinct baseline P wave preceding this QRS complex.

QRS complex: Narrow, less than 120 ms, with morphology very similar to the reference beats. The peak amplitude is 1.255 mV, which is consistent with the baseline range.

Timing: Premature. The RR interval to the previous beat is 0.656 s, which is significantly shorter than the established baseline average of approximately 0.80 s.

T wave: Positive polarity, with no inversion.

Overall interpretation: The beat at 5.68 s is identified as abnormal because it occurs earlier than expected and lacks a distinct preceding baseline P wave, while maintaining a narrow QRS complex similar to the surrounding reference beats.

Figure 17: Illustrative interpretation produced by the morphology-analyzer student distilled from Gemma 4 on MIT-BIH Record 100 (segment 0).

replace—clinical judgment. Finally, real-world deployment requires additional safeguards, including uncertainty calibration, robust out-of-distribution detection, privacy-preserving data governance, and human-in-the-loop oversight to reduce automation bias.

Overall, DeepArrhythmia should be regarded as an assistive decision-support framework rather than an autonomous diagnostic system. It targets offline and retrospective analysis, such as Holter review and second reading of flagged segments, rather than hard real-time on-device alarming. We evaluate on beat-annotated datasets because the task requires per-beat AAMI labels; large resources such as PTB-XL and MIMIC-IV-ECG provide record-level 12-lead diagnostic statements and are therefore outside the beat-level scope of this paper. Future work should emphasize multicenter external validation, improved robustness for rare classes and noisy recordings (e.g., generative synthesis of realistic fusion morphology), label-free or conformal selection of the routing threshold, and prospective workflow-level evaluation of effects on diagnostic quality, patient safety, and clinician workload.

## W Human Evaluation Protocol

We evaluate generated morphology-analysis outputs from three sources: Gemini 3.1 Pro with label conditioning, Gemini 3.1 Pro without label conditioning, and DeepArrhythmia. For each source, 200 samples are randomly selected, resulting in 600 total evaluated samples. All generated outputs are anonymized before review and presented in blinded, randomized order to three experts so that the evaluators assess the text itself rather than model identity. Each sample is graded independently according to the seven criteria summarized in Table 23.

The evaluator instructions were as follows: for each anonymized generated output, read the morphology analysis and assign a score from 1 to 5 for each criterion in Table 23; score each sample independently; and base the judgment on the clinical quality of the generated output, including whether it provides class-consistent diagnostic support, cites observable ECG evidence, remains

Table 23: Description of the seven criteria used in the expert evaluation of morphology analyses.

| Criterion | Description and scoring guidance |
|---|---|
| **Diagnostic Support** | Does the explanation provide convincing evidence for the predicted arrhythmia label (*N*, *S*, *V*, or *F*) rather than merely restating the answer? A score of 5 indicates that the explanation strongly supports the predicted class with class-consistent ECG evidence; a score of 1 indicates that the explanation does not support the prediction or supports the wrong class. |
| **ECG Evidence Grounding** | Are the stated findings explicitly grounded in observable ECG evidence or tool-derived measurements, such as RR-interval pattern, QRS width/shape, P-wave morphology, compensatory pause, or amplitude differences? A score of 5 requires specific, verifiable evidence; a score of 1 indicates vague, generic, or unsupported statements. |
| **Physiological Correctness** | Are the described ECG findings medically plausible and internally consistent? A score of 5 indicates no clear physiological errors; a score of 1 indicates major factual errors, contradictions, or clinically implausible reasoning. |
| **Use of Segment Context** | Does the explanation appropriately use surrounding beats and temporal context when needed, especially for rhythm-related discrimination? A score of 5 indicates effective use of neighboring beats and context; a score of 1 indicates isolated-beat reasoning when context is essential. |
| **Evidence Completeness** | Does the explanation cover the main evidence needed for this case, including both morphology and rhythm cues when relevant? A score of 5 indicates that the important cues are covered with no major omissions; a score of 1 indicates that most key evidence is missing. |
| **Uncertainty and Safety** | Does the output avoid overclaiming in noisy, ambiguous, or low-context cases, and does it appropriately acknowledge limited evidence? A score of 5 indicates well-calibrated and safe phrasing; a score of 1 indicates unjustified certainty or unsafe overinterpretation. |
| **Clinical Review Utility** | Would this output help an expert review the case more efficiently, verify the prediction, or understand possible failure modes? A score of 5 indicates strong practical usefulness for expert review; a score of 1 indicates little or no value for clinical assessment. |

physiologically plausible, uses segment context when needed, covers the key evidence, expresses uncertainty appropriately, and would be useful in expert review.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the paper's core claims about the tool-augmented architecture, multi-dataset evaluation, ablations, and interpretability. These claims are supported by the Method and Experiment sections and the appendix analyses.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The appendix includes a dedicated "Broader Impacts and Limitations" section. It discusses class imbalance, noise sensitivity, retrospective evaluation, possible LLM explanation errors, and deployment constraints.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not present formal theorems or proof-based theoretical results. Its contributions are methodological and empirical.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The paper describes the datasets, DS1/DS2 split, preprocessing, augmentation, model structure, training stages, ablations, and evaluation setup in the Method and Experiment sections. The appendix adds concrete implementation settings such as hardware configuration, batch size, learning rate, and sequence length, and it documents the split-safe curation protocol and prompt format used for morphology-teacher distillation.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: The codebase and reproduction instructions are publicly available at `https://github.com/jimmylihui/DeepArrhythmia_repo`. The ECG datasets used in the experiments are public. In particular, the morphology-distillation prompt format is disclosed in the appendix, whereas the raw teacher outputs are not treated as an open artifact because they were generated through a proprietary API.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: The paper specifies the subject-disjoint data split, 10-second segmentation, baseline correction, augmentation scheme, Stage 1/Stage 2 training setup, dataset-specific confidence-threshold selection, evaluation setup, and appendix implementation details such as GPU configuration, batch size, learning rate, and sequence length. These details are sufficient to understand how the reported results were obtained.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The main performance and ablation tables report mean for the trainable methods, with variability tied to three random seeds. Deterministic VLM baselines are evaluated with decoding temperature 0. Appendix M additionally reports 95% bootstrap confidence intervals and balanced accuracy on MIT-BIH.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The appendix reports the GPU type/count, memory, precision mode, attention kernel,training runtime, and distributed training setup, and the paper reports inference latency.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: To the best of our knowledge, the work uses retrospective public ECG datasets and does not involve new human-subject data collection. The paper also discusses risks, limitations, and responsible clinical positioning in the appendix.

   Guidelines:

- The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The appendix "Broader Impacts and Limitations" section discusses positive impacts such as assistive clinical screening and interpretability. It also discusses negative risks including bias, imperfect explanations, privacy/governance concerns, automation bias, and deployment limitations.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper does not release a new high-risk model or scraped dataset as part of the submission. It evaluates on existing public ECG datasets and does not describe a separate controlled-release artifact.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.

- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We use existing public ECG datasets, pretrained encoders, open-weight models, closed-source API models, and software components, all of which are cited in the paper. The PhysioNet datasets (MIT-BIH Arrhythmia, MIT-BIH Supraventricular, and INCART) are distributed under the Open Data Commons Attribution License v1.0, and VitalDB is distributed under the Creative Commons Attribution 4.0 International License (CC BY 4.0). Open-weight models are used under their respective model licenses, and closed-source models are accessed through their providers' API terms.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: The public repository (`https://github.com/jimmylihui/DeepArrhythmia_repo`) contains the implementation, configurations, and reproduction instructions. This released code asset is documented alongside the artifact through the repository materials referenced in the paper.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: The work does not involve crowdsourcing or a prospective patient-facing human-subject study. However, we include an expert evaluation of generated ECG morphology analyses by experts. We provide the evaluation instructions, scoring rubric, and criteria in Appendix W. Screenshots are not applicable because the evaluation was conducted using a structured rubric rather than a public crowdsourcing interface.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: The study does not conduct prospective human-subject research and does not collect new patient data. All experiments use retrospective public de-identified ECG datasets. The expert evaluation only asks experts to rate anonymized model-generated text (Appendix W) and involves no patient contact.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: LLM/VLM components are integral to the proposed method, including the central agent, the Morphology Analyzer, confidence-guided tool orchestration, teacher-distilled morphology explanations, and multiple LLM/VLM baselines. Their role is described in the abstract, the Method and Experiment sections, and the appendix.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.