Final Report: An Interpretable Deep Learning Framework for Wearable AFib Detection

Student Name: Yifei Tao

Date: May 5th, 2026

1. Introduction

1.1 What problem are you going to solve?

The proliferation of wearable devices, such as the Apple Watch and Fitbit, has democratized cardiac monitoring, enabling millions to screen for Atrial Fibrillation (AFib). While these devices demonstrate high sensitivity, they often suffer from lower specificity in real-world settings, leading to a critical "false positive" problem [1]. Many of these false alerts are triggered by benign rhythm variants such as Sinus Tachycardia, a regular but rapid rhythm often caused by anxiety, exercise, or caffeine. A major pain point lies in the "Inconclusive" or "Possible AFib" alerts provided by current consumer-grade ECGs. When a user receives such a notification without any explanation of why, it can trigger a cycle of health anxiety or cyberchondria [2]. This anxiety further elevates the heart rate, leading to more false alerts and creating a debilitating feedback loop. Current deep learning models behave as black boxes, providing only a binary or categorical label without showing the morphological evidence (such as the presence or absence of a P-wave) that the model relied on. This project aims to break that cycle by developing an interpretable framework that classifies single-lead ECG recordings into Normal, AFib, and Other rhythms, and produces a faithful patient-facing natural-language explanation grounded in the model's actual evidence.

1.2 Why is this problem important?

The clinical and psychological burden of false positives is substantial. A recent case study described a patient who performed 916 smartwatch ECGs in one year due to anxiety over "inconclusive" alerts, leading to 12 unnecessary emergency department visits [2]. Such over-utilization strains healthcare systems and subjects patients to unnecessary stress and testing [3]. A larger follow-up study in 2024 confirmed that 20% of wearable AF users reported intense anxiety in response to rhythm notifications, and this group showed twice the rate of diagnostic tests and 4.5 times the ablation rate compared to non-users [19]. From a psychological perspective, the inability of current devices to explain why a rhythm was flagged is itself the threat: uncertainty without context drives the anxiety. Providing a transparent output that describes which morphological features the model focused on, even when the classification is uncertain, can serve as a digital "second opinion" and empower users with actionable, interpretable health data [4].

1.3 Why is machine learning promising?

Traditional signal processing relies on handcrafted features such as R-R interval irregularity (e.g., RMSSD). However, Sinus Tachycardia with minor respiratory arrhythmia can mimic the irregularity of

AFib in short segments, leading to misclassification by rule-based systems [5]. Deep learning, particularly Convolutional Neural Networks (CNNs), shows strong promise here because it can learn hierarchical representations directly from raw waveforms. A CNN can extract fine-grained morphological features such as the subtle shape of a P-wave often buried in noise [6]. More importantly, recent advances in Explainable AI (XAI) such as Grad-CAM [7] make it possible to visualize where the model focused its attention, and recent work in language modeling has shown that grounding generated explanations in structured evidence reduces hallucination [20]. Combining these tools allows the project to move beyond accuracy alone and target a more clinically meaningful goal: an AI output that is both predictive and faithfully explained.

1.4 Major Contributions

The contributions of this project are: (1) a 3-class CNN classifier on the PhysioNet 2017 single-lead ECG dataset using the original expert-annotated labels; (2) a Grad-CAM implementation with a physiological windowing step that summarizes saliency maps into structured features over the P-wave, QRS, and T-wave regions; (3) an LLM-based explanation pipeline that takes these features as input under explicit grounding rules; and (4) a manual faithfulness evaluation of 40 generated explanations following Xie et al. [20], including a prompt-design ablation that raised the overall faithful rate from 72.5% to 95%.

2. Related Work

2.1 Deep Learning for ECG-Based AFib Detection

Automated AFib detection from single-lead ECG recordings has advanced substantially since the PhysioNet/Computing in Cardiology Challenge 2017 established a standardized four-class benchmark using consumer-grade AliveCor recordings [5]. Early approaches combined handcrafted rhythm features with ensemble classifiers, achieving macro-averaged F1 scores up to 0.868 [5]. Subsequent deep learning work improved on these results. Fan et al. [8] showed that a multi-scale CNN applied directly to raw waveforms could capture morphological features without manual engineering, achieving 98.13% accuracy on 20-second segments. Hybrid CNN-LSTM architectures further improved inter-beat temporal modeling [9, 10], and more recent work using data augmentation on PhysioNet 2017 reached F1_AVG = 0.833 on the full four-class task [11].

Despite these advances, most existing models report performance on the standard four-class benchmark (Normal, AFib, Other, Noisy) and do not explicitly examine which non-AFib rhythms (e.g., Sinus Tachycardia) drive the false-positive alerts that contribute to health anxiety in wearable users [2]. The present work uses the same three rhythm labels (Normal, AFib, Other) as in the original dataset for training.

2.2 Explainable AI (XAI) and Interpretability in Healthcare

Even highly accurate models are unlikely to be trusted without some explanation of their reasoning [12]. Grad-CAM [7] is the most widely used XAI method for ECG analysis. It generates a temporal heatmap by weighting convolutional feature maps according to class-specific gradients. Hicks et al. [13] applied Grad-CAM to 12-lead ECGs and confirmed that the resulting maps highlighted clinically meaningful landmarks such as the QRS complex and T-wave. Beyond arrhythmia classification, Grad-CAM has been

applied to myocardial infarction detection [16] and extended to multimodal 1D and 2D ECG inputs for broader cardiac arrhythmia classification [17], demonstrating its versatility across cardiac diagnostic tasks. Jones et al. [14] further showed that saliency maps can meaningfully improve ECG classification interpretability when aligned with known physiological features. Baig et al. [18] found that combining Grad-CAM with SHAP was complementary: Grad-CAM identified broad regions of interest while SHAP provided finer, time-step-level importance scores. When both methods agreed on the same morphological feature, this convergence strengthened the clinical justification for the model's output.

That said, saliency maps should not be accepted uncritically. Arun et al. [15] evaluated eight popular saliency methods and found that every one failed at least one trustworthiness criterion, including localization accuracy and reproducibility. A visually compelling heatmap does not automatically constitute reliable clinical evidence, and this concern directly motivates the faithfulness constraints discussed in Section 2.4.

2.3 Health Anxiety and Psychological Well-Being in Wearable Device Users

Consumer wearables have proven effective for cardiac monitoring [1, 3], but evidence is mounting that they can also cause unintended psychological harm. Rosman et al. [2] described a patient with paroxysmal AF who performed 916 smartwatch ECGs in a single year, ultimately developing illness anxiety disorder that required cognitive behavioral therapy. Importantly, anxiety escalations were triggered not only by AF alerts but equally by "inconclusive" notifications, ambiguous outputs that the patient had no framework to interpret. The authors explain this through the Uncertainty and Anticipation Model of Anxiety: when a device flags a possible abnormality but provides no morphological rationale, uncertainty itself becomes the threat.

A larger follow-up study by Rosman et al. [19] confirmed this pattern in a propensity-matched cohort of 172 AF patients. Wearable users showed significantly higher cardiac symptom preoccupation ($P = 0.03$) and treatment concerns ($P = 0.02$) than non-users. Twenty percent reported intense anxiety in response to irregular rhythm notifications, and this group drove substantially higher healthcare consumption, including twice the rate of diagnostic tests and 4.5 times the ablation rate relative to matched non-users ($P = 0.04$). These findings make clear that the core problem is not false classification itself, but the absence of an explanation that helps non-expert users distinguish a benign state from a real arrhythmia. This is the design gap the current project directly targets.

2.4 Faithfulness of AI-Generated Explanations in Medical Contexts

Since the central contribution of this project is a natural-language explanation grounded in ECG morphology, the key technical risk is ensuring that generated text faithfully reflects the model's evidence rather than introducing unsupported claims. Xie et al. [20] identify two failure modes in generative medical AI: intrinsic errors, where output contradicts the source evidence, and extrinsic errors, where output introduces claims that cannot be verified at all. Their review found that nearly 30% of AI-generated radiology summaries contained factual errors, and that standard metrics like ROUGE and BERTScore cannot detect them. A fluent, confident explanation can therefore still be clinically wrong.

The proposed pipeline addresses this by conditioning the LLM's output on structured Grad-CAM evidence rather than allowing free generation from the model's general knowledge. Restricting the input

to what the saliency map has explicitly identified, for example "high activation in the P-wave region; low rhythm interval variance," limits the space for extrinsic errors and turns generation into a translation task rather than an inference task.

Evidence from the ECG-language modeling literature supports this approach. Li et al. [21] showed that a frozen ClinicalBERT model can encode meaningful cardiac semantics without any task-specific fine-tuning, confirming that clinical language models already understand the vocabulary of ECG descriptions. Yu et al. [22] found that structuring LLM prompts around explicit, feature-level ECG information outperforms unconstrained prompting, with the removal of the domain guidance component alone causing a 0.076 drop in macro-F1. Both results support the same design principle: grounding the LLM's context in structured, evidence-based inputs produces more faithful and reliable outputs, which is the primary evaluation criterion for the proposed explanation pipeline.

3. Methods

3.1 Dataset

This project uses the PhysioNet/Computing in Cardiology Challenge 2017 dataset [5], which contains 8,528 single-lead ECG recordings collected via AliveCor consumer devices, sampled at 300 Hz, and ranging from 9 to 61 seconds in duration. Each recording carries one of four expert-annotated labels: Normal sinus rhythm (N = 5,076), Other rhythm (O = 2,415), Atrial Fibrillation (A = 758), and Too Noisy to classify (~ = 279). Noisy recordings are excluded from all experiments, leaving a final set of 8,249 recordings across three rhythm classes. This dataset closely mirrors the signal quality and lead configuration of current consumer smartwatches, making it directly applicable to the wearable health monitoring context this project addresses.

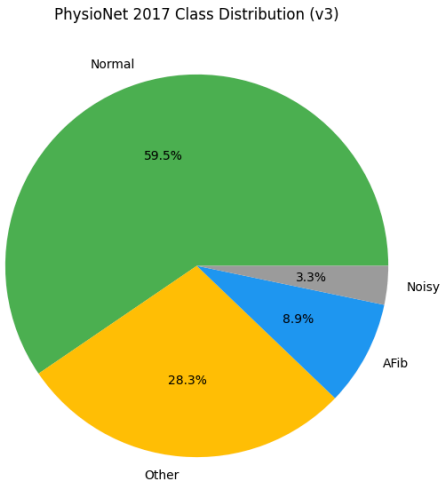

Figure 1: Class distribution of the PhysioNet 2017 dataset. Normal rhythms dominate the training set, while AF accounts for only 8.9% of recordings, reflecting a real-world

screening population with significant class imbalance.

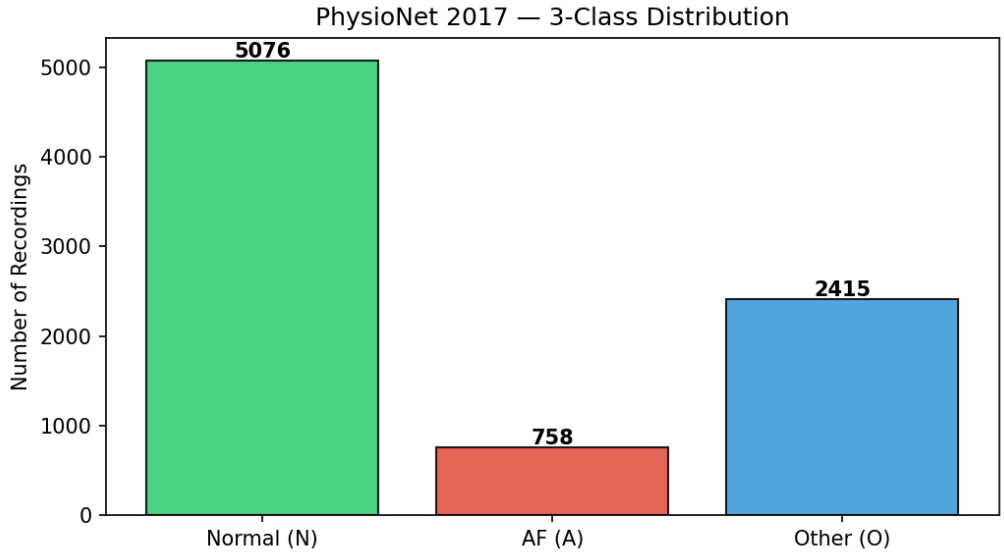

Figure 2: Class distribution of the three rhythm classes used in this project (Normal, Other, AFib).

3.2 Data Preprocessing and Experimental Setup

All recordings are truncated or zero-padded to a fixed length of 2,700 samples (9 seconds at 300 Hz) and z-score normalized per recording. The 8,249 recordings are split into training (80%), validation (10%), and test (10%) sets using stratified sampling to preserve class proportions, yielding 6,599 training, 825 validation, and 825 test samples. To address the substantial class imbalance (Normal:Other:AF ≈ 6.7:3.2:1), weighted cross-entropy loss is applied during training, with class weights computed using sklearn's "balanced" strategy.

3.3 Classification Backbone

A 1D Convolutional Neural Network (CNN) serves as the classification backbone. The architecture consists of three convolutional blocks, each containing a Conv1d layer with increasing filter counts (32, 64, 128), followed by batch normalization, ReLU activation, and max pooling. A global average pooling (GAP) layer then collapses the temporal dimension, and a fully connected layer produces 3-class logits for Normal, AFib, and Other. GAP is used instead of a flatten operation specifically because it preserves the temporal feature maps from the last convolutional block, which is what makes Grad-CAM possible in the next stage. The model is trained using the Adam optimizer with an initial learning rate of 1e-3 and a ReduceLROnPlateau scheduler over 30 epochs, with the best checkpoint selected by validation macro F1.

3.4 Grad-CAM Implementation

Grad-CAM [7] is applied to the trained CNN to generate a temporal saliency map for each test recording, indicating which regions of the input ECG the model focused on when making its prediction. The implementation hooks into the last convolutional layer (block 3) and captures both forward activations and the gradients of the predicted class score with respect to those activations. The channel-wise gradients

are globally averaged over time to produce per-channel weights, which are then used to compute a weighted sum of the feature maps. A ReLU is applied to keep only positive contributions, and the resulting map is upsampled back to the original signal length of 2,700 samples and normalized to the range [0, 1]. For all subsequent analyses, Grad-CAM is computed with respect to the model's predicted class, so the saliency map reflects the evidence the model used for its own decision rather than for any reference label.

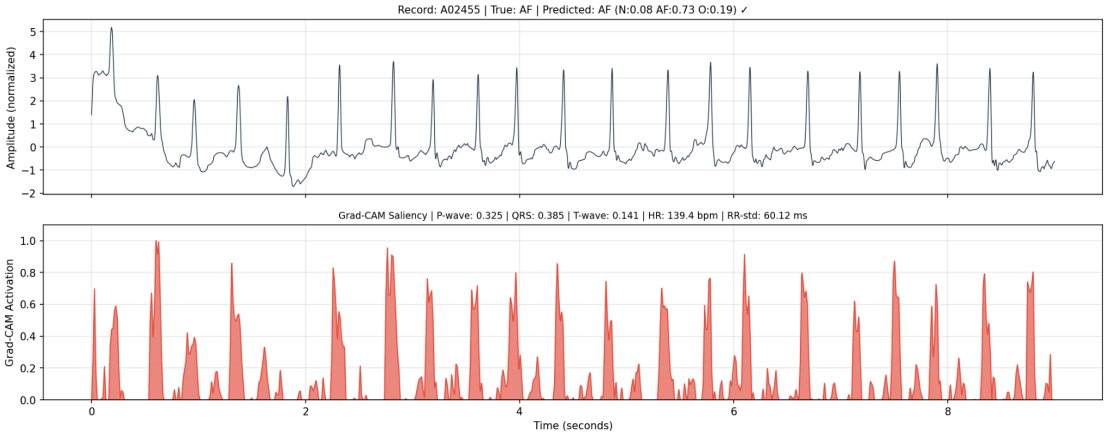

Figure 3: Example Grad-CAM saliency map for a correctly classified AFib sample. Top: raw ECG signal. Bottom: Grad-CAM activation along the same time axis, with peaks indicating regions the model focused on.

3.5 Physiological Window Analysis

To convert raw saliency values into clinically interpretable features, the project introduces a windowing step that aggregates Grad-CAM activations within three physiologically defined regions around each detected R-peak. R-peaks are detected using the NeuroKit2 library. For each R-peak at sample index r, three windows are extracted: the P-wave region (samples r-30 to r-12, corresponding to roughly 100 ms to 40 ms before the R-peak), the QRS complex (r-15 to r+15, ±50 ms around the R-peak), and the T-wave region (r+45 to r+120, roughly 150 ms to 400 ms after the R-peak). The mean Grad-CAM activation within each window is averaged across all R-peaks in the recording, producing three scalar features per sample: P-wave activation, QRS activation, and T-wave activation. Two rhythm-level features are also computed from the R-peak series: mean heart rate (in BPM) and the standard deviation of R-R intervals (RR-std, in ms), where the latter serves as an indicator of rhythm regularity.

3.6 LLM Explanation Pipeline

The five structured features from Section 3.5 (three regional activations plus heart rate and RR-std), together with the model's predicted class and confidence, form the input to a language model that generates a patient-facing explanation. The LLM used is Claude Sonnet 4 (claude-sonnet-4-20250514) accessed through the Anthropic API. Activation values are first discretized into qualitative levels ("low" if < 0.3, "moderate" if 0.3–0.6, "high" if ≥ 0.6) and RR-std is mapped to either "regular" (≤ 50 ms) or "irregular" (> 50 ms), so that the LLM does not see raw numeric values that could leak into the user-facing explanation. The prompt template enforces four explicit rules: only describe features that

appear in the structured evidence; do not introduce unsupported information; do not assert that the user has or does not have any condition; and always recommend consulting a healthcare professional. The instruction is class-specific, so the prompt for an AFib prediction differs from the prompt for a Normal or Other prediction in tone and emphasis. Generation is constrained to 300 tokens. Following the design principle in Xie et al. [20], this setup turns the generation task into translation from structured evidence to natural language rather than open-ended inference, which limits the surface area for hallucination [21, 22].

3.7 Faithfulness Evaluation Protocol

To evaluate whether the generated explanations actually reflect the Grad-CAM evidence, 40 explanations are sampled from the test set, balanced across predicted classes and across correct and incorrect predictions. Each explanation is manually annotated against its corresponding structured evidence following the taxonomy of Xie et al. [20]. An intrinsic error is recorded when the explanation contradicts the evidence, for example claiming P-wave clarity when the P-wave activation is low. An extrinsic error is recorded when the explanation introduces clinical claims that are not present in the evidence, for example asserting the rhythm is regular when no rhythm information is supplied. An explanation is labeled as faithful if and only if it contains neither type of error. Faithful rate is reported overall and broken down by predicted class.

In addition to this initial round of annotation, a small prompt-design ablation is conducted: the prompt is revised based on the failure mode observed in the first round, and the affected explanations are regenerated and re-annotated under the same protocol, with all other components of the pipeline held fixed. This isolates the effect of prompt structure on faithfulness. The specific failure mode that motivated the revision, and the resulting change in faithful rate, are reported in Section 4.

3.8 Novelty

Most prior deep learning work on PhysioNet 2017 reports performance on the standard four-class benchmark and does not pair the classifier with a downstream natural-language explanation [8, 9, 10, 11]. Visual saliency methods such as Grad-CAM have been applied to ECG analysis [13, 14, 16, 17, 18], but typically for visual inspection by domain experts rather than as the input to a patient-facing language model. This project combines the two by routing Grad-CAM evidence through structured physiological windows into an LLM, and adds a manual faithfulness evaluation that quantifies whether the resulting explanations are grounded in what the model actually focused on. The prompt-design ablation in Section 3.7 also provides a small, controlled measurement of how prompt structure affects faithfulness, a question the existing faithfulness literature [20] frames but does not directly test in the consumer single-lead ECG setting.

4. Results

4.1 Classification Performance

The CNN backbone achieved an overall test set accuracy of 64% and a macro F1 score of 0.53 on the held-out test set, with the best checkpoint selected at the epoch with peak validation macro F1 of 0.5148. Per-class results show strong Normal detection (precision = 0.74, recall = 0.78, F1 = 0.76), but weaker

AFib detection (precision = 0.35, recall = 0.46, F1 = 0.40) and weaker Other detection (precision = 0.48, recall = 0.38, F1 = 0.43), reflecting the substantial class imbalance and the inherent heterogeneity of the Other class.

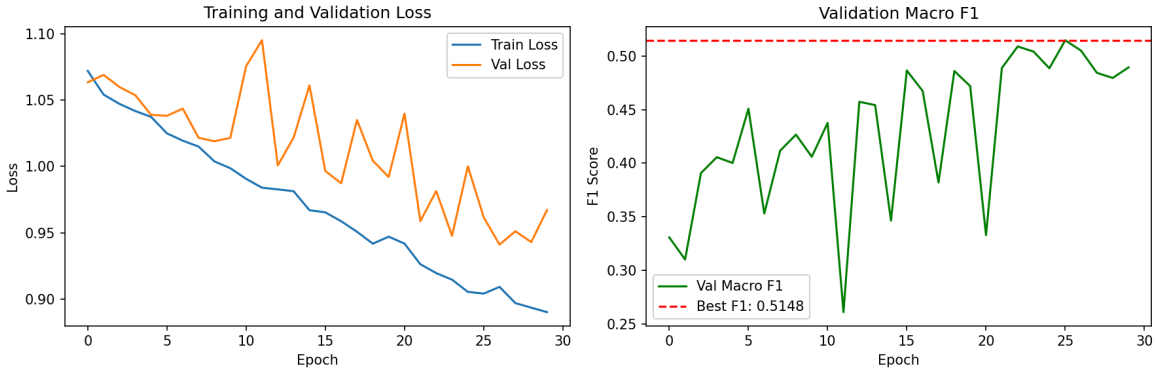

Figure 4: Training and validation loss curves (left) and validation macro F1 score over 30 epochs (right). The persistent gap between training and validation loss indicates mild overfitting, attributable to the small effective dataset size for the minority classes.

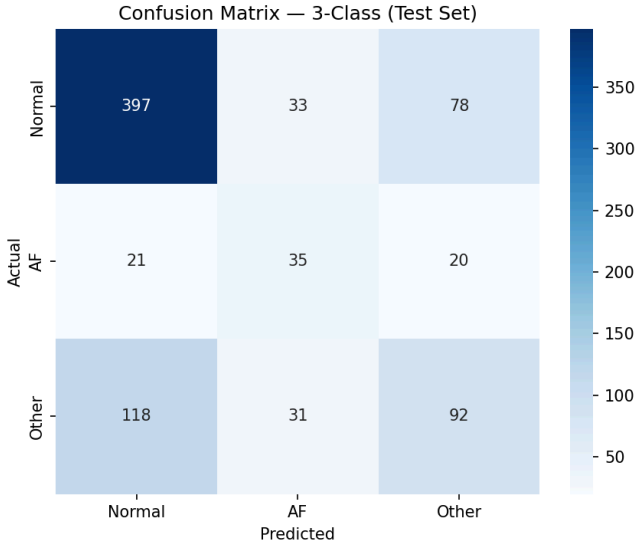

Figure 5: Confusion matrix on the held-out test set (n = 825). Of 76 AFib samples, 35 were correctly identified, 21 were misclassified as Normal, and 20 as Other. Of 508 Normal samples, 397 were correctly identified.

The confusion matrix reveals two clinically meaningful error patterns. The 21 AFib recordings misclassified as Normal represent a safety concern, as genuine arrhythmia cases would receive a falsely

reassuring output. The 33 Normal recordings misclassified as AFib represent the false-positive scenario that drives health anxiety in wearable users [2, 19]. Both error directions motivate the explanation pipeline: even when the classification is wrong, a faithful Grad-CAM-grounded explanation gives the user morphological context rather than only a categorical label.

4.2 Grad-CAM Activation Patterns by Class

This table summarizes the average Grad-CAM activation within each physiological window, computed over all test samples for which R-peak detection succeeded.

Region	Normal	AFib	Other
P-wave activation	0.147	0.176	0.172
QRS activation	0.106	**0.216**	0.170
T-wave activation	0.245	0.167	0.226
Mean HR (BPM)	71.5	90.1	78.5
RR-std (ms)	82.1	**164.3**	117.1

The most informative pattern is in the AFib row. AFib samples show the highest QRS activation (0.216 vs 0.106 for Normal) and the highest beat-to-beat variability (RR-std = 164.3 ms vs 82.1 ms for Normal). This is consistent with the clinical definition of AFib, where rhythm irregularity at the QRS level is the defining feature. P-wave activation is comparable across classes, which suggests that the model relies more on rhythm-level cues than on P-wave morphology in this single-lead, low signal-to-noise setting.

4.3 LLM Explanation Examples

Two representative explanations generated by the pipeline are shown below to illustrate intended and unintended behavior.

Correctly classified AFib. Prediction: AFib (73% confidence). Evidence: HR = 139 BPM, irregular rhythm, P-wave activation moderate, QRS activation moderate.

> "The ECG analysis system detected signs consistent with atrial fibrillation. The system identified an irregular heartbeat pattern and focused its attention on the P-wave region and QRS complex. Your heart rate was elevated. It is important to consult with a healthcare professional promptly to discuss these findings and determine the appropriate next steps for your care."

Correctly classified Normal (V1, unfaithful). Prediction: Normal (49% confidence). Evidence: HR = 77 BPM, regular rhythm, all activations low.

> "The system analyzed your ECG and detected a normal heart rhythm. The P-waves, QRS complexes, and T-waves all appear as expected in a normal rhythm."

The AFib explanation accurately reflects the evidence. The Normal explanation illustrates the failure mode described in Section 4.4: all activations were low, yet the LLM described wave features it had no evidence for.

4.4 Faithfulness Evaluation

Manual annotation of 40 generated explanations following the protocol in Section 3.7 yielded an overall faithful rate of 72.5% (29 of 40). Three explanations contained intrinsic errors (7.5%) and ten contained extrinsic errors (25.0%).

The class-level breakdown is more informative than the overall number. AFib explanations were faithful in 17 of 18 cases (94%), and Other explanations were faithful in 8 of 8 cases (100%). Normal-class explanations were faithful in only 4 of 14 cases (29%).

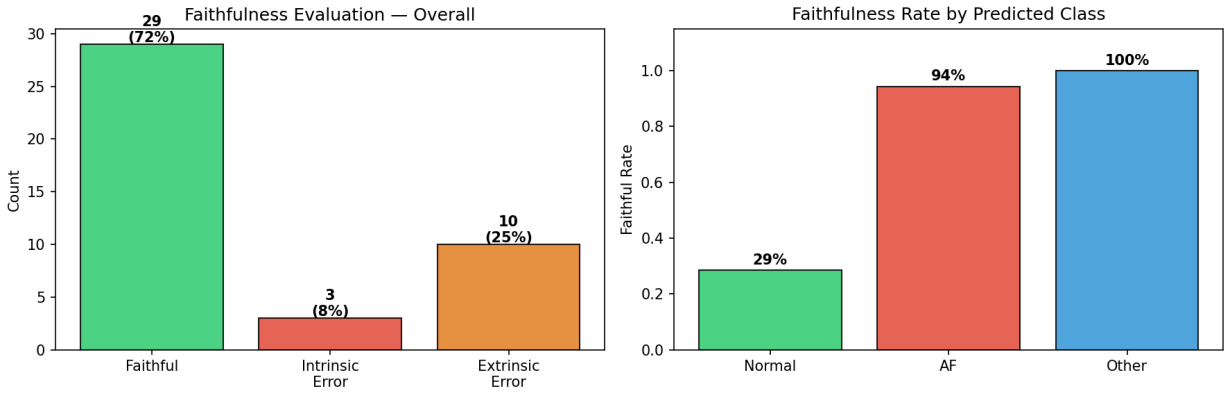

Figure 6: Faithfulness evaluation results. Left: overall counts of Faithful, Intrinsic Error, and Extrinsic Error across the 40 annotated explanations. Right: faithful rate broken down by predicted class.

Inspection of the failed Normal-class explanations revealed a consistent pattern. When all three Grad-CAM activations were low (i.e., the model did not strongly focus on any specific morphological region), the LLM filled the gap by describing wave features it had no evidence for, for example claiming that "the P-waves, QRS complexes, and T-waves all appear as expected in a normal rhythm." This is a textbook extrinsic error: the explanation adds clinical content that is not supported by the structured evidence. AFib and Other explanations did not exhibit this pattern, because their Grad-CAM evidence typically contained non-zero directional signal (high QRS activation, irregular rhythm) that the LLM could anchor on.

4.5 Prompt Improvement Ablation

To address the Normal-class failure mode, the prompt was revised with one targeted addition: when all three activation values are below 0.3, the LLM is explicitly instructed to describe only heart rate and rhythm regularity, and not to describe any wave-level feature. All other components of the pipeline were

held fixed, and the explanations for the 14 Normal-class samples were regenerated and re-annotated under the same protocol.

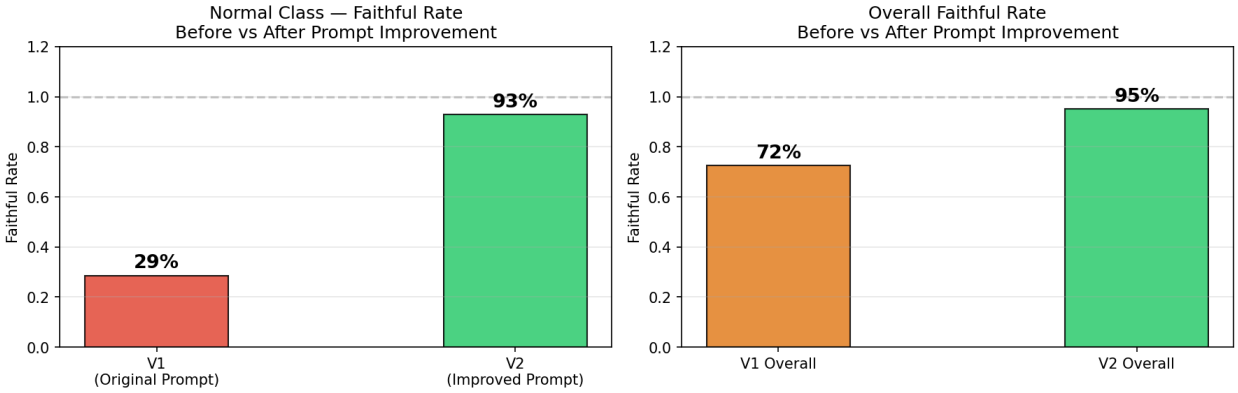

Figure 7: Faithful rate before and after the prompt revision. Left: Normal class only (29% → 93%). Right: overall across all 40 samples (72.5% → 95%).

After the revision, the Normal-class faithful rate rose from 29% (4 of 14) to 93% (13 of 14), and the overall faithful rate across all 40 samples rose from 72.5% to 95% (38 of 40). This ablation isolates prompt structure as a variable and shows it has a substantial effect on faithfulness. The overall faithful rate of 95% is not sufficient for clinical deployment, but the result demonstrates that targeted prompt design can close much of the gap between what the model focused on and what the LLM describes.

5. Discussion

5.1 Summary of Contributions

This project built an end-to-end pipeline for interpretable wearable ECG classification, covering three stages: a 3-class CNN classifier, a Grad-CAM saliency analysis with physiological windowing, and an LLM-based explanation generator with a faithfulness evaluation. The primary finding is not in the classification performance itself, which is modest at macro F1 = 0.53, but in the faithfulness evaluation. The initial prompt achieved 94% and 100% faithful rates for AFib and Other predictions respectively, but only 29% for Normal predictions. A single targeted prompt revision raised the Normal faithful rate to 93% and the overall rate to 95%. This demonstrates that the structure of the prompt has a measurable and large effect on whether the LLM stays within the bounds of the evidence it is given.

5.2 What Worked and What Did Not

The Grad-CAM analysis produced clinically interpretable patterns. AFib samples showed the highest QRS activation and the highest RR variability across all three classes, which aligns with the clinical definition of AFib as an irregularly irregular rhythm. This suggests the model is picking up on the right signal rather than spurious correlations in the data.

The classification performance, however, has clear limitations. The 10-to-1 class imbalance between Normal and AFib recordings meant that even with weighted cross-entropy loss, the model still leaned

toward predicting Normal, resulting in an AFib recall of 0.46. The Other class is inherently difficult because it groups together many different rhythm types with no shared morphological signature. These classification limitations are worth acknowledging, though they do not undermine the main contribution of the project, which concerns the explanation layer rather than the classifier itself.

The LLM explanation pipeline worked well when the Grad-CAM evidence was directional. It failed when all activations were low, which occurred most often for Normal predictions. In retrospect this is predictable: a language model given evidence that says nothing has nothing to anchor on, and defaults to fabricated clinical description.

5.3 Implications for Researchers and Clinicians

For technology researchers, the main takeaway is that faithfulness evaluation of LLM-generated medical explanations requires manual annotation. Standard metrics such as ROUGE and BERTScore do not detect the types of errors identified here [20]. The failure mode found in the Normal class would have been invisible without comparing the generated text against the structured evidence line by line.

For clinicians, the pipeline demonstrates that a saliency-based approach can produce patient-facing explanations that are grounded in the model's actual behavior rather than post-hoc rationalization. However, the current system is not ready for clinical use. A faithful rate of 95% means roughly one in twenty explanations still contains an unsupported claim, which is not an acceptable error rate in a medical context.

5.4 Ethics and Privacy

The dataset used in this project consists of de-identified ECG recordings and does not contain personally identifiable information. However, ECG signals can in principle be used to infer sensitive health information beyond cardiac rhythm, and any real deployment of this pipeline would need to address data storage and access controls carefully.

A more immediate ethical concern is the risk of false reassurance. If a user with genuine AFib receives a Normal classification paired with a confident-sounding explanation, they may delay seeking care. This risk is not eliminated by faithfulness evaluation alone: an explanation can be faithful to the model's evidence and still be clinically wrong if the underlying classification is incorrect. Any clinical application of this type of system would require prospective validation and regulatory approval before deployment.

5.5 Future Directions

Several directions follow naturally from this work. The most limiting factor for classification is the absence of fine-grained labels within the Other class. A dataset with expert-annotated sub-labels for rhythms such as Sinus Tachycardia, atrial flutter, and SVT would allow a more targeted classification setup and a more meaningful case study. On the explanation side, the prompt-design ablation here is small and manual. A more systematic investigation of how different prompt structures affect faithfulness across a larger sample would strengthen the conclusions. Finally, the assumption that faithful explanations reduce user anxiety has not been tested empirically. A user study with real smartwatch users would be a meaningful next step.

References

[1] Marco V. Perez, Kenneth W. Mahaffey, Haley Hedlin, et al. 2019. Large-Scale Assessment of a Smartwatch to Identify Atrial Fibrillation. *New England Journal of Medicine*. 381, 20 (2019), 1909–1917. DOI:https://doi.org/10.1056/NEJMoa1901183

[2] Lindsey Rosman, Anil Gehi, and Rachel Lampert. 2020. When smartwatches contribute to health anxiety in patients with atrial fibrillation. *Cardiovascular Digital Health Journal.* 1, 1 (2020), 9–10. DOI: https://doi.org/10.1016/j.cvdhj.2020.06.004

[3] Kirk J. Wyatt, Michael J. Maniaci, Paul R. Nobrega, et al. 2020. User-initiated photoplethysmogram alert for atrial fibrillation: initial surveillance experience of a large healthcare system. *Journal of the American Medical Informatics Association*. 27, 9 (2020), 1432–1437. DOI:https://doi.org/10.1093/jamia/ocaa137

[4] Gašper Štiglic, Primož Kocbek, Nino Ficco, et al. 2020. Interpretability of machine learning-based prediction models in healthcare. *WIREs Data Mining and Knowledge Discovery*. 10, 5 (2020), e1379. DOI:https://doi.org/10.1002/widm.1379

[5] Gari D. Clifford, Chengyu Liu, Benjamin Moody, et al. 2017. AF Classification from a Short Single Lead ECG Recording: The PhysioNet/Computing in Cardiology Challenge 2017. In *Proceedings of Computing in Cardiology* (CinC '17), Vol. 44. IEEE, 1–4.

[6] Yuguang Ye, Kavimbi Chipusu, Muhammad Awais Ashraf, Bijiao Ding, Yifeng Huang, and Jianlong Huang. 2025. Hybrid CNN-BLSTM architecture for classification and detection of arrhythmia in ECG signals. *Scientific Reports*. 15, 1 (2025), 34510. DOI:https://doi.org/10.1038/s41598-025-17671-1

[7] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, et al. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV '17)*. IEEE, 618–626.

[8] Xiaomao Fan, Qihang Yao, Yunpeng Cai, Fen Miao, Fangmin Sun, and Ye Li. 2018. Multiscaled Fusion of Deep Convolutional Neural Networks for Screening Atrial Fibrillation From Single Lead Short ECG Recordings. IEEE Journal of Biomedical and Health Informatics. 22, 6 (2018), 1744–1753. DOI: https://doi.org/10.1109/JBHI.2018.2858789

[9] Georgios Petmezas, Kostas Haris, Leandros Stefanopoulos, et al. 2021. Automated Atrial Fibrillation Detection using a Hybrid CNN-LSTM Network on Imbalanced ECG Datasets. Biomedical Signal Processing and Control. 63 (2021), 102194. DOI: https://doi.org/10.1016/j.bspc.2020.102194

[10] Chen Chen, Zhengchun Hua, Ruiqi Zhang, Guangyuan Liu, and Wanhui Wen. 2020. Automated arrhythmia classification based on a combination network of CNN and LSTM. Biomedical Signal Processing and Control. 57 (2020), 101819. DOI: https://doi.org/10.1016/j.bspc.2019.101819

[11] Zhaocheng Yu, Junxin Chen, Yu Liu, et al. 2022. DDCNN: A Deep Learning Model for AF Detection From a Single-Lead Short ECG Signal. IEEE Journal of Biomedical and Health Informatics. 26, 10 (2022), 4987–4995. DOI: https://doi.org/10.1109/JBHI.2022.3191754

[12] Muhammad Aurangzeb Ahmad, Carly Eckert, and Ankur Teredesai. 2018. Interpretable Machine Learning in Healthcare. In *Proceedings of the 9th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics* (ACM-BCB '18). ACM, 559–560. DOI: https://doi.org/10.1145/3233547.3233667

[13] Steven A. Hicks, Jonas L. Isaksen, Vajira Thambawita, et al. 2021. Explaining Deep Neural Networks for Knowledge Discovery in Electrocardiogram Analysis. *Scientific Reports*. 11 (2021), 10949. DOI: https://doi.org/10.1038/s41598-021-90285-5

[14] Yola Jones, Fani Deligianni, and Jeff Dalton. 2020. Improving ECG Classification Interpretability using Saliency Maps. In *Proceedings of the IEEE 20th International Conference on Bioinformatics and Bioengineering* (BIBE '20). IEEE.

[15] Nishanth Arun, Nathan Gaw, Praveer Singh, et al. 2021. Assessing the Trustworthiness of Saliency Maps for Localizing Abnormalities in Medical Imaging. *Radiology: Artificial Intelligence*. 3, 6 (2021), e200267. DOI: https://doi.org/10.1148/ryai.2021200267

[16] V. Jahmunah, E.Y.K. Ng, Ru-San Tan, Shu Lih Oh, and U. Rajendra Acharya. 2022. Explainable Detection of Myocardial Infarction using Deep Learning Models with Grad-CAM Technique on ECG Signals. *Computers in Biology and Medicine*. 146 (2022), 105550. DOI: https://doi.org/10.1016/j.compbiomed.2022.105550

[17] Fatma Murat Duranay, Ender Murat, Özal Yıldırım, Yakup Demir, Ru-San Tan, Niranjana Sampathila, and U. Rajendra Acharya. 2026. MM-GradCAM: An Improved Multimodal GradCAM Method with 1D and 2D ECG Data for Detection of Cardiac Arrhythmia. *Scientific Reports*. 16 (2026), 7919. DOI: https://doi.org/10.1038/s41598-026-38654-w

[18] Zuraiz Baig, Sidra Nasir, Rizwan Ahmed Khan, and Muhammad Zeeshan Ul Haque. 2025. ArrhythmiaVision: Resource-Conscious Deep Learning Models with Visual Explanations for ECG Arrhythmia Classification. *arXiv preprint* arXiv:2505.03787v1.

[19] Lindsey Rosman, Rachel Lampert, Songcheng Zhuo, et al. 2024. Wearable Devices, Health Care Use, and Psychological Well-Being in Patients With Atrial Fibrillation. *Journal of the American Heart Association*. 13 (2024), e033750. DOI: https://doi.org/10.1161/JAHA.123.033750

[20] Qianqian Xie, Edward J. Schenck, He S. Yang, Yong Chen, Yifan Peng, and Fei Wang. 2023. Faithful AI in Medicine: A Systematic Review with Large Language Models and Beyond. *Research Square* (preprint). DOI: https://doi.org/10.21203/rs.3.rs-3661764/v1

[21] Jun Li, Che Liu, Sibo Cheng, Rossella Arcucci, and Shenda Hong. 2023. Frozen Language Model Helps ECG Zero-Shot Learning. *Proceedings of Machine Learning Research*. 227 (2023), 402–415.

[22] Han Yu, Peikun Guo, and Akane Sano. 2023. Zero-Shot ECG Diagnosis with Large Language Models and Retrieval-Augmented Generation. *Proceedings of Machine Learning Research*. 225 (2023), 650–663.