Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

# A Lightweight, CPU-Deployable, and Interpretable ECG Arrhythmia Classification Pipeline Using the MIT-BIH Arrhythmia Database

**Weihao Cheng**

weihao_cheng@hdu.edu.cn

Hangzhou Dianzi University https://orcid.org/0009-0009-1891-4760

**Research Article**

**Keywords:** ECG, arrhythmia classification, MIT-BIH, interpretable machine learning, feature engineering, CPU deployment

**Posted Date:** December 23rd, 2025

**DOI:** https://doi.org/10.21203/rs.3.rs-8415145/v1

**License:** This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

**Additional Declarations:** The authors declare no competing interests.

# Abstract

Electrocardiogram (ECG) arrhythmia classification is a foundational task in medical artificial intelligence, yet many high-performing deep learning approaches require GPU resources and may be difficult to interpret in clinical settings. This study presents a lightweight, CPU-deployable, and interpretable pipeline for heartbeat-level arrhythmia classification using the MIT-BIH Arrhythmia Database. ECG beats were segmented around annotated R-peaks and mapped into five AAMI-style aggregated classes (N, S, V, F, Q). We extracted compact time-domain statistics and frequency-domain energy features, then trained and compared three classical machine learning models: Logistic Regression (LR), Random Forest (RF), and Support Vector Machine (SVM). Using a record-wise split to reduce data leakage risk, RF achieved the highest accuracy (0.864), while LR provided the strongest one-vs-rest ROC-AUC (0.806). Class-wise ROC curves and feature-importance analysis suggested that spectral energy and amplitude-related statistics contributed substantially to discrimination. Overall, the results demonstrate that interpretable, resource-efficient ECG classification remains feasible without deep networks, supporting practical deployment in CPU-only environments and rapid prototyping of medical AI systems.


# 1. Introduction

Arrhythmias are a major contributor to morbidity and mortality worldwide, and ECG remains the most accessible noninvasive signal for screening and diagnosis. Automated ECG interpretation has therefore become a central topic in medical AI, with applications ranging from bedside monitoring to ambulatory Holter analysis. Public physiological signal repositories have played an essential role in enabling reproducible research; among them, PhysioNet provides datasets and tooling that have shaped decades of signal-processing and machine-learning development in biomedical informatics [1].

The MIT-BIH Arrhythmia Database is one of the most influential standardized ECG benchmarks and has historically served as a common evaluation substrate for arrhythmia detection algorithms [2][3]. Despite the recent dominance of deep learning, classical machine learning pipelines remain attractive in scenarios where computational budgets are limited or interpretability is prioritized—e.g., embedded devices, CPU-only hospital systems, or rapid clinical decision support prototypes.

This work focuses on a “small but complete” engineering contribution: a fully reproducible, CPU-deployable pipeline that (i) segments beats from annotated R-peaks, (ii) uses lightweight handcrafted features, (iii) evaluates multiple classical models under a record-wise split, and (iv) provides interpretable feature-importance analysis. The goal is not to claim state-of-the-art performance, but to demonstrate that a practical medical AI pipeline can be built quickly, transparently, and reproducibly with a modest feature set—an approach that can serve as a strong baseline for subsequent deep-learning expansion.

# 2. Related Work

# 2.1 Public ECG resources and benchmarking

PhysioNet and its associated toolkits were introduced to support research on complex physiologic signals, making datasets like MIT-BIH widely accessible for benchmarking [1]. The MIT-BIH Arrhythmia Database, in particular, has had enduring impact on algorithm evaluation and comparative studies of arrhythmia detectors [2][3].

## 2.2 AAMI-style beat grouping for arrhythmia classification

Many ECG heartbeat-classification studies group beat annotations into higher-level categories aligned with AAMI testing and reporting guidance, commonly using the N/S/V/F/Q-style aggregation for algorithm evaluation [4]. This reduces label granularity while maintaining clinically meaningful groupings for supraventricular and ventricular ectopy assessment.

## 2.3 Classical ML vs. deep learning

Feature-based methods have historically delivered strong performance on arrhythmia classification and remain valuable for interpretability and efficiency. For example, mixture-of-features approaches have been used to classify AAMI-recommended classes on MIT-BIH-like settings [5]. Deep learning methods can further improve performance by learning representations directly from raw waveforms; for instance, convolutional architectures have been reported for heartbeat classification under AAMI-style grouping [6]. However, deep models may require more computation, careful tuning, and can be harder to interpret—motivating lightweight baselines such as the pipeline proposed here.

# 3. Methods

## 3.1 Dataset and access

We used the MIT-BIH Arrhythmia Database hosted on PhysioNet [3]. Signals and annotations were programmatically accessed using the WFDB Python package, which provides reading and processing utilities for physiological waveforms and annotation files [7].

#### Dataset rationale

MIT-BIH has long been used as a reference benchmark for arrhythmia detection and evaluation, making it suitable for a reproducible baseline study [2][3].

## 3.2 Beat segmentation

For each recording, we used annotated R-peak positions from the standard atr annotation stream and extracted a fixed-length beat-centered window. Each beat segment covered 0.2 s pre-R and 0.4 s post-R (total 0.6 s). This window provides a compact representation of local morphology and rhythm context while remaining computationally efficient.

## 3.3 Label mapping (five-class aggregation)

Each beat annotation symbol was mapped into one of five aggregated categories (N, S, V, F, Q), consistent with widely used AAMI-style evaluation conventions [4]. Briefly:

**N**

normal and bundle-branch block–type beats

**S**

supraventricular ectopic beats

**V**

ventricular ectopic beats

**F**

fusion beats

**Q**

paced/unknown/other beats (as applicable under aggregation)

# 3.4 Feature extraction

We extracted a lightweight feature set designed for CPU efficiency and interpretability:

**Time-domain statistics**

mean, standard deviation, min, max, peak-to-peak amplitude (ptp), RMS, and mean absolute value.

**Frequency-domain energy**

we computed a real FFT for each beat (after normalization) and summarized spectral energy in coarse bands (five bins) plus total energy.

These features are intentionally compact so that (i) training and inference remain fast on CPU, and (ii) feature importance can be directly interpreted and discussed.

# 3.5 Models and training protocol

We trained and compared three classical models:

**Logistic Regression (LR)** with standardized features and class balancing

**Random Forest (RF)** with balanced subsampling

**Support Vector Machine (SVM)** with RBF kernel, standardized features, and probability estimates

**Record-wise split**

To reduce overly optimistic performance caused by correlated beats from the same record appearing in both train and test, we used a record-level grouping split (GroupShuffleSplit). This is an important engineering step for fairer evaluation in beat-level tasks.

# 3.6 Evaluation metrics

Given the strong class imbalance typical of MIT-BIH-derived beat datasets, we reported:

**Accuracy**

**Macro-F1 (emphasizes minority-class performance)**

**One-vs-rest ROC-AUC (OVR AUC) for multi-class discrimination assessment**

# 4. Experiments

# 4.1 Dataset Summary, Class Distribution, and Experimental Setup

A total of 109,460 heartbeat segments were extracted from the MIT-BIH Arrhythmia Database using beat-centered windows (0.2 s pre-R and 0.4 s post-R). The resulting dataset exhibited a highly imbalanced class distribution, which is typical for long-term ambulatory ECG recordings: N constituted the majority class, whereas F and S were minor classes. This imbalance is clinically realistic but introduces an evaluation challenge: a model may achieve high accuracy by predominantly predicting the majority class, while still performing poorly for rare yet clinically important arrhythmias.

To mitigate overly optimistic evaluation caused by leakage across beats from the same record, we adopted a record-wise split (GroupShuffleSplit). Under this protocol, beats from the same record are assigned exclusively to either training or test sets, which better approximates generalization to unseen recordings. We report three complementary metrics: Accuracy, Macro-F1, and one-vs-rest ROC-AUC (OVR AUC). Accuracy reflects overall correctness; Macro-F1 emphasizes balanced performance across classes by averaging per-class F1 equally; OVR AUC captures threshold-independent separability for each class against the rest and is especially informative in imbalanced multi-class settings.

Table 1
Model comparison on MIT-BIH (record-wise split)

| Model | Accuracy | Macro-F1 | OVR AUC |
|---|---|---|---|
| Logistic Regression | **0.424** | **0.365** | **0.806** |
| Random Forest | **0.864** | **0.332** | **0.709** |
| SVM | **0.604** | **0.313** | **0.704** |

The results show a clear trade-off:

Random Forest (RF) achieved the highest accuracy (0.864), suggesting strong performance on the dominant class and robust rule-based partitioning of feature space. However, its Macro-F1 (0.332) is lower than LR's, implying that performance on minor classes is not proportionally improved.

Logistic Regression (LR) achieved the best OVR AUC (0.806) and highest Macro-F1 (0.365) among the three. This indicates better overall class separability and a more balanced performance distribution than RF, despite lower accuracy.

SVM achieved moderate accuracy (0.604) and Macro-F1 (0.313). In this pipeline, SVM may be limited by feature simplicity and class overlap, especially between N and S.

# 4.2 ROC Analysis: Class-wise Discrimination Performance

In multi-class arrhythmia classification, ROC curves provide a class-wise view of how well the model separates each category from the rest across all thresholds. Figure 1 presents one-vs-rest ROC curves for the five aggregated classes. Two patterns are especially noteworthy:

(1) Strong separability for Q and V classes.

The ROC curves indicate that Q and V are more easily separable under the current feature representation. This may arise because paced/unknown-like beats and ventricular ectopy often introduce distinct morphological or spectral energy characteristics. In practical terms, a high AUC implies that even if the final decision threshold is adjusted (e.g., to reduce false positives), the model can maintain relatively strong sensitivity-specificity trade-offs.

(2) Lower separability for S class.

The S class has the lowest ROC performance among the five, which is consistent with the clinical and algorithmic difficulty of distinguishing supraventricular ectopy from normal beats using short-window morphology alone. Supraventricular beats may differ subtly, and the discriminatory signal may require rhythm context (e.g., RR intervals, local variability) rather than purely local morphology. This observation sets up a concrete future-work direction: add RR-based features or longer context windows while still remaining CPU-friendly.

# 4.3 Confusion matrix (error patterns)

The normalized confusion matrix (Fig. 2) reveals the structure of misclassifications and provides insight beyond scalar metrics.

(1) Diagonal dominance and class stability

Several classes show strong diagonal values (e.g., Q and V), indicating stable recognition patterns. This supports the ROC findings and suggests that the chosen feature set contains adequate discriminative

cues for these classes.

(2) N–S confusions: the central bottleneck

A prominent error mode is confusion between N and S, which is common in heartbeat-level classification. This is an important point to discuss carefully in SCI writing:

**Engineering explanation**

the feature set emphasizes amplitude statistics and coarse spectral energy; these may not fully capture subtle supraventricular morphological differences.

**Physiological explanation**

supraventricular beats can appear similar to normal beats in short segments, and the discriminating cues often emerge in timing (prematurity) or broader context rather than morphology alone.

(3) Minor classes and imbalance sensitivity

For minor classes such as F, the matrix often shows dispersion into major classes. This is expected under imbalance and limited sample size. It is also the reason Macro-F1 is relatively low even when accuracy is high.

(4) Why this matters clinically

From a clinical screening perspective, false negatives in arrhythmia classes (e.g., predicting N for an arrhythmia beat) are typically more concerning than false positives. The confusion matrix gives you a ready-made narrative to justify future threshold calibration (e.g., increase sensitivity for S/V) and cost-sensitive training if the deployment scenario demands it.

# 4.4 Feature importance (interpretability)

Figure 3 shows the top-10 features ranked by absolute LR coefficients. Among them, amplitude-related statistics (e.g., absolute mean, RMS, standard deviation) and coarse spectral energy features (fft bands) were consistently influential. This supports the intuition that both waveform magnitude and frequency characteristics contribute to class separation in beat-level arrhythmia classification.

# 4.5 Representative spectral patterns (qualitative visualization)

Figure 4 provides a qualitative visualization of representative spectral-energy profiles for each class. Although the curves share a common shape (expected due to normalization and the inherent structure of ECG signals), subtle differences are visible across bands.

This plot supports two important points:

Coarse spectral summarization is informative. Even with only five energy bands, the model benefits from frequency-domain information.

Classes differ in nuanced ways, which may not be fully separable using only coarse bands, especially for borderline categories such as N vs S. This further motivates augmenting the feature set with rhythm context or more refined spectral descriptors if higher sensitivity is needed.

## 5. Discussion

This study highlights a practical and reproducible approach to ECG arrhythmia classification that is computationally efficient and interpretable. Importantly, the results illuminate the trade-offs between accuracy-driven performance and balanced multi-class recognition under severe imbalance.

# 5.1 Why classical ML remains valuable in medical AI

While deep learning can be highly accurate, real-world clinical and engineering settings often require rapid prototyping, transparency, and deployment feasibility on limited hardware. The proposed pipeline addresses these constraints by using compact features and CPU-friendly models, making it attractive for:

embedded or bedside deployments,

research baselines and ablation studies,

educational settings and rapid iteration cycles.

The interpretability of Logistic Regression is particularly valuable: coefficients provide a direct mapping from features to decisions, allowing error analysis and feature refinement without resorting to opaque latent representations.

# 5.2 Interpreting the metric trade-offs (Accuracy vs Macro-F1 vs AUC)

Your results show an instructive pattern:

RF has high accuracy but relatively low Macro-F1. Under imbalance, accuracy can be dominated by the majority class (N). RF may learn highly effective rules for N, improving accuracy substantially, but still struggle with minority classes.

LR has the best AUC and Macro-F1 among the three. This indicates better overall separability and a more balanced classification profile, even if the default thresholding yields lower accuracy.

# 5.3 Error analysis: why N and S are hard, and what to do next

The confusion matrix suggests that N–S discrimination is the main bottleneck. This is expected when using short beat windows and morphology-focused features. Two low-cost improvements are strongly justified:

RR interval features: pre-RR, post-RR, RR ratio, local mean RR, etc. These capture prematurity and rhythm changes that are crucial for supraventricular ectopy.

Morphological descriptors: approximate QRS width, slopes, or wavelet energies; these can be computed efficiently without GPUs.

You can present this as a targeted future direction: “improving features is likely more effective than simply changing the classifier.”

## 5.4 Clinical translation considerations

From a deployment perspective, misclassifications have asymmetric cost. In many screening contexts, sensitivity for arrhythmias (S/V) may be prioritized over specificity, suggesting:

probability calibration,

class-specific thresholds,

cost-sensitive training.

Your pipeline is well-suited to such adjustments because it outputs calibrated-like probabilities (for LR/SVM) and can be tuned without extensive retraining overhead.

# 6.Conclusion and Future Work

## 6.1 Conclusion

We developed a lightweight, interpretable, and CPU-deployable pipeline for heartbeat-level arrhythmia classification using beat-centered segmentation and compact handcrafted features. Across three classical models, Random Forest achieved the highest accuracy (0.864), while Logistic Regression achieved the best discrimination ability measured by OVR ROC-AUC (0.806) and the highest Macro-F1 among compared models (0.365). The ROC and confusion-matrix analyses indicate strong separability for Q and V classes and identify N–S confusion as the primary challenge, consistent with limited rhythm context in short beat windows. Feature-importance analysis further demonstrates that amplitude-related statistics and coarse spectral energy features drive model decisions, supporting transparent interpretation and efficient deployment.

## 6.2 Future Work

To strengthen both performance and clinical relevance while preserving CPU efficiency, future work will pursue:

### 6.2.1 Rhythm-aware feature augmentation:

Add pre-/post-RR intervals, RR ratios, and local HRV descriptors to improve discrimination between N and S, which is the dominant error mode observed.

### 6.2.2 Feature refinement with minimal overhead:

Explore wavelet energy features, QRS width approximations, and slope-based morphology descriptors. These remain computationally light but can capture subtler morphological cues.

### 6.2.3 Imbalance-aware learning and thresholding:

Investigate cost-sensitive optimization, re-sampling strategies, and class-specific decision thresholds. This is especially relevant for minority classes such as F.

### 6.2.4 External validation and robustness testing:

Validate the pipeline on additional ECG datasets and evaluate robustness to noise, baseline wander, and lead variations to better approximate real-world clinical conditions.

### 6.2.5 Deployment-oriented evaluation:

Benchmark inference latency and memory footprint on CPU-only devices and evaluate calibrated probabilities for safe integration into clinical decision support.

## References


1. Goldberger AL, Amaral LAN, Glass L, Hausdorff JM, Ivanov PC, Mark RG, Mietus JE, Moody GB, Peng C-K, Stanley HE (2000) PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. Circulation 101(23):e215–e220. https://doi.org/10.1161/01.CIR.101.23.e215
2. Moody GB, Mark RG (2001) The impact of the MIT-BIH arrhythmia database. IEEE Eng Med Biol Mag 20(3):45–50. https://doi.org/10.1109/51.932724
3. PhysioNet (n.d.). MIT-BIH Arrhythmia Database (mitdb), v1.0.0. Retrieved from https://physionet.org/content/mitdb/1.0.0/
4. Association for the Advancement of Medical Instrumentation (2012) /2020 ANSI/AAMI EC57: Testing and reporting performance results of cardiac rhythm and ST segment measurement algorithms (ANSI/AAMI EC57:2012 (R2020)). (Preview). Retrieved from https://webstore.ansi.org/preview-pages/AAMI/preview_ANSI%2BAAMI%2BEC57-2012%2B%28R2020%29.pdf
5. Das MK, Ari S (2014) & (others as listed in the article). (2014). ECG beats classification using mixture of features. Comput Math Methods Med, Article ID 178414. https://pmc.ncbi.nlm.nih.gov/articles/PMC4897569/

6. Kachuee M, Fazeli S, Sarrafzadeh M (2018) ECG heartbeat classification: A deep transferable representation. arXiv preprint, arXiv:1805.00794. https://arxiv.org/abs/1805.00794
7. PhysioNet (2023) Waveform Database Software Package (WFDB) for Python. Retrieved from https://physionet.org/content/wfdb-python/

# Figures

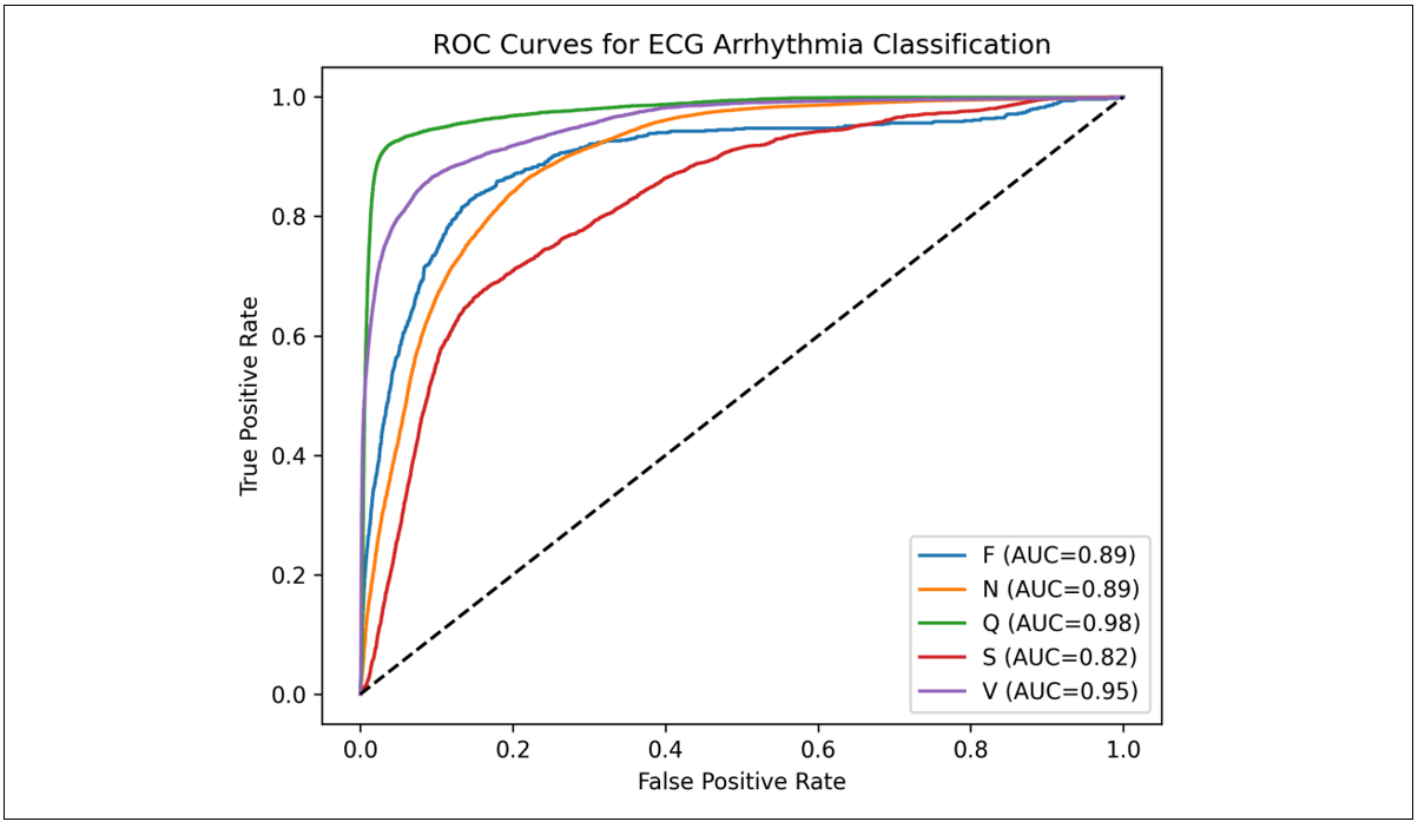



**Figure 1**

**Class-wise ROC curves for five-class ECG arrhythmia classification (one-vs-rest)**

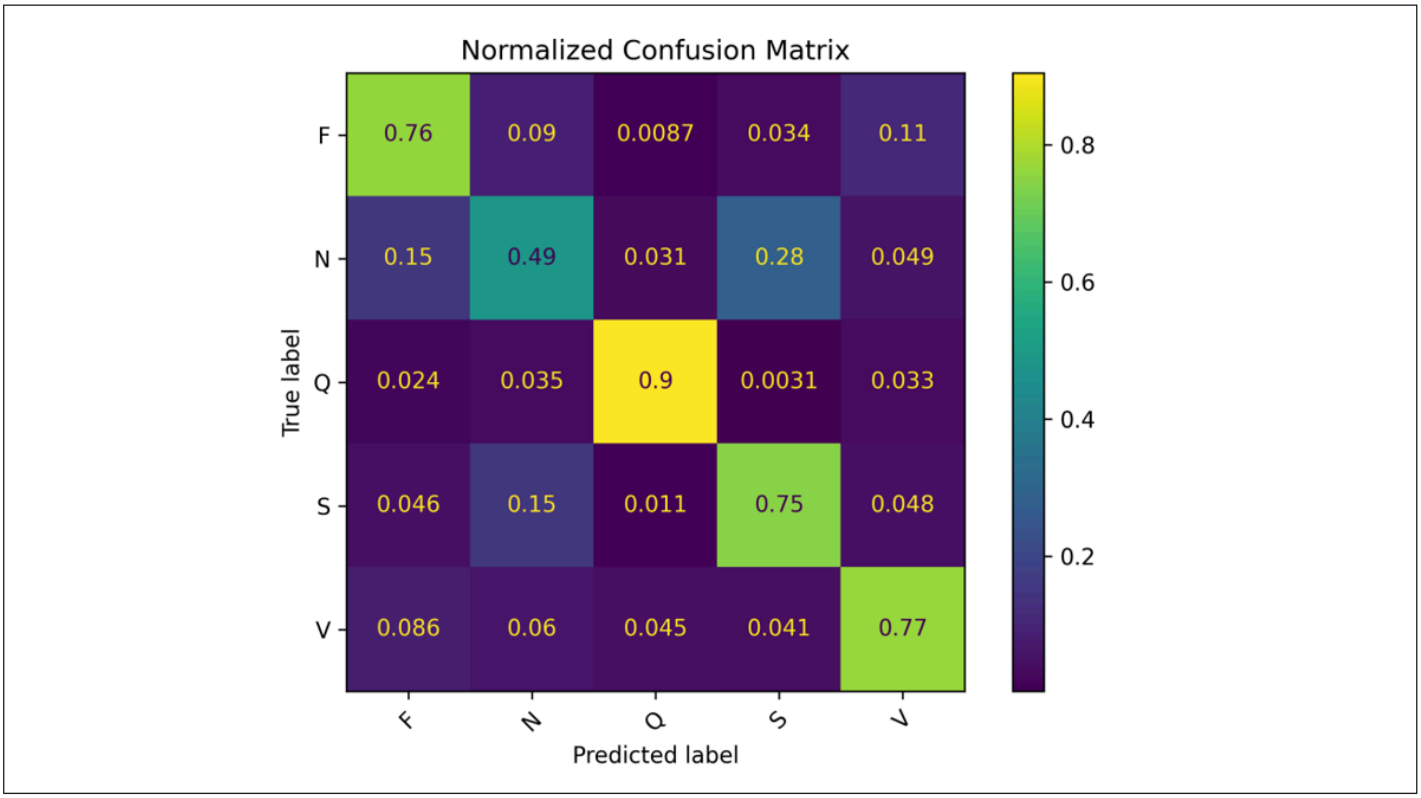



Figure 2

Normalized confusion matrix revealing dominant error modes across five arrhythmia categories

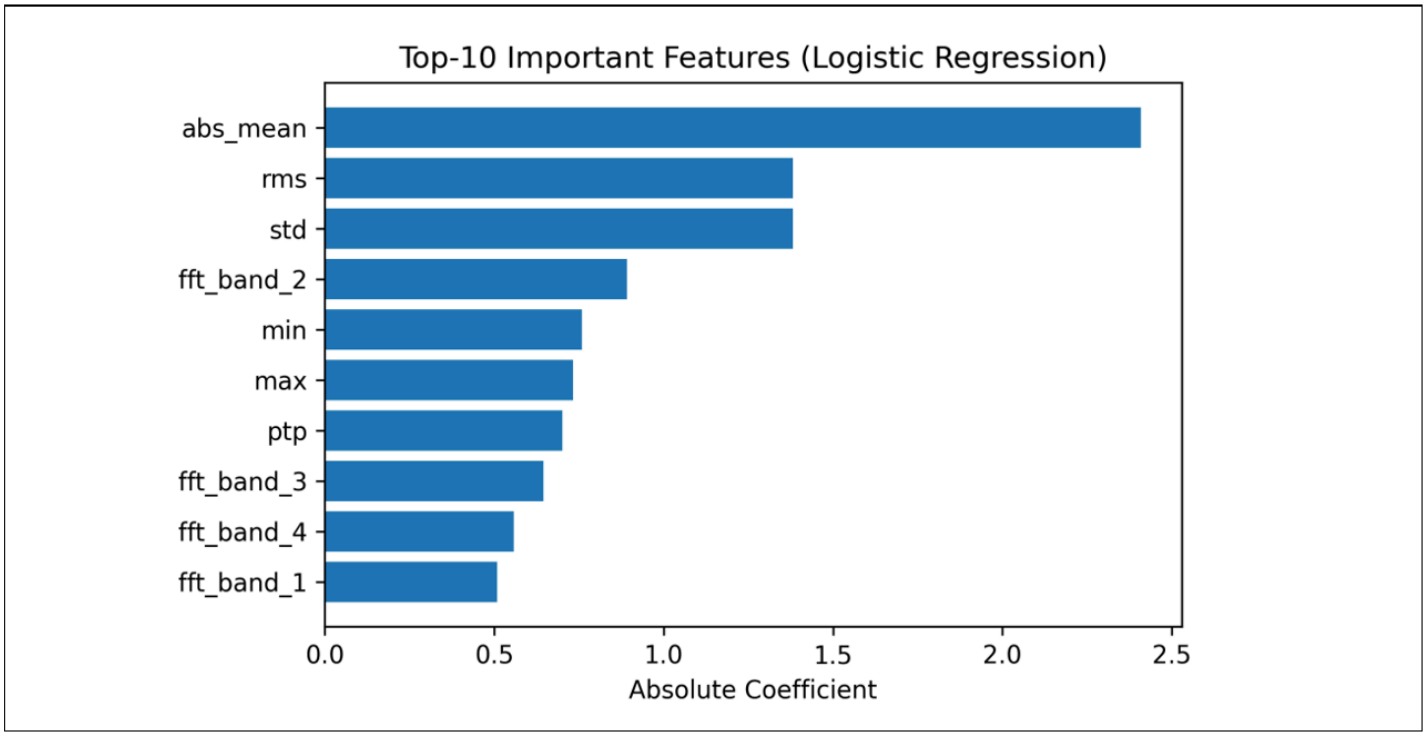



Figure 3

Top-10 most influential handcrafted features in Logistic Regression (absolute coefficient magnitude)

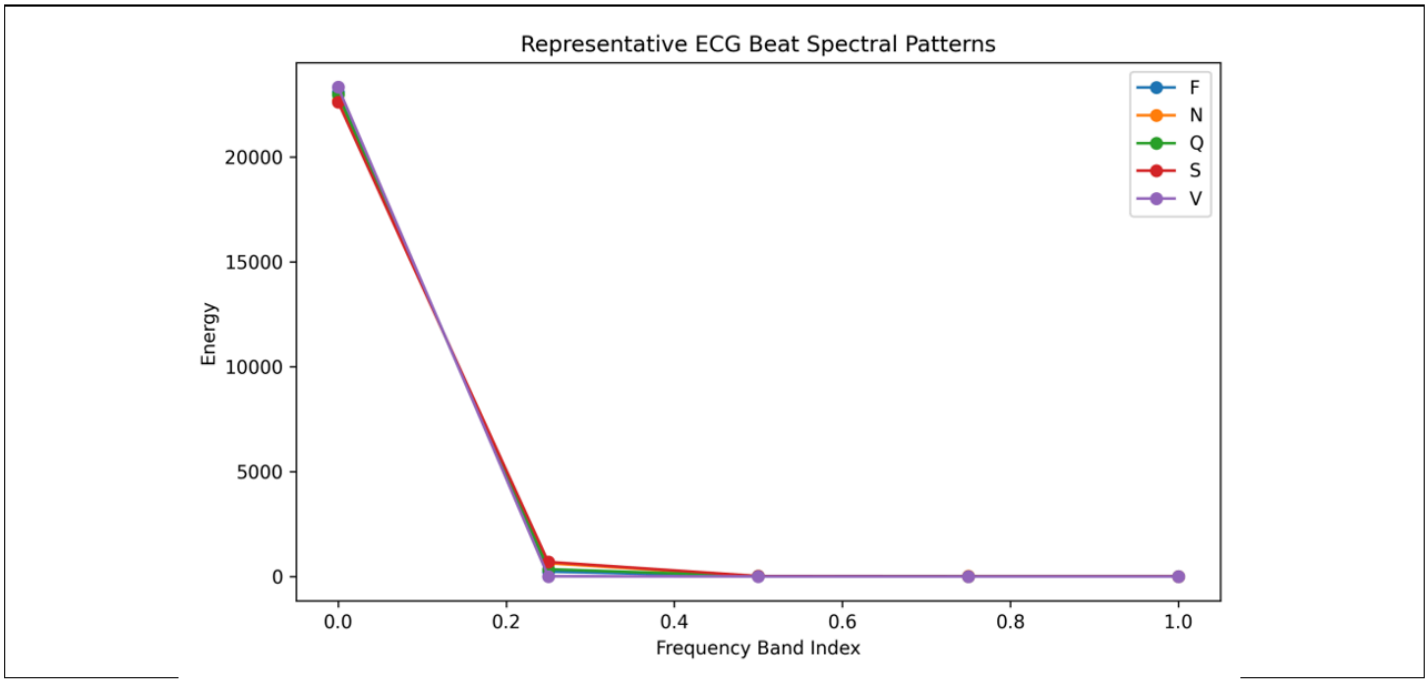



Figure 4

Representative spectral-energy signatures across arrhythmia classes using coarse FFT band summaries