

# Explainable Machine Learning

## Table of Contents

<b>1. Introduction</b>	3
<b>2. Methodology</b>	3
2.2 Data Exploration	4
2.3 Validation Protocols	5
2.4 Classification	6
2.5 Explainability Analysis	6
2.6 Clustering	6
<b>3. Results</b>	6
3.1 Dataset Overview	7
3.2 Classification Performance	7
Beat Holdout Results	7
Patient Holdout Results	7
3.3 Permutation Feature Importance Performance	7
Segment Importance Scores	7
QRS Complex Dominance	9
Model Specific Patterns	9
3.4 Clustering Performance	9
<b>4. Discussion</b>	10
4.1 Classification Performance Analysis	10
4.2 Feature Importance Insights	10
4.3 Clustering Analysis Interpretation	11
4.4 Implications and Limitations	12
<b>Conclusion</b>	12
<b>References</b>	12

# 1. Introduction

The electrocardiogram (ECG) is the primary non-invasive tool for identifying cardiac arrhythmias, but interpreting them remains challenging. Cardiologists still rely heavily on manual review, which is time-consuming, costly, and often inconsistent across clinicians. These limitations have driven growing interest in automated machine learning classification systems. However, recent evidence indicates that such systems remain unreliable, with a 2025 study reporting that automated ECG tools misinterpreted nearly four in ten recordings [1]. Most ML models also function as black boxes, providing little insight into how their decisions are generated. This lack of transparency makes it difficult for clinicians to trust or verify the predictions they receive. This study addresses this problem by developing explainable arrhythmia classification models that not only achieve high accuracy but also identify the temporal features and physiological segments that drive predictions, enabling clinical validation and supporting their potential integration into medical practice.

This study uses the MIT-BIH Arrhythmia Database, representing heartbeats as temporal feature sequences across eight arrhythmia classes with class imbalances. This research has the following aims in explainable arrhythmia classification:

- Determine whether classical machine learning models (SVM and Random Forest) can achieve clinically acceptable accuracy for multi-class arrhythmia classification.
- Evaluate model generalization on completely unseen patients versus randomly sampled heartbeats from the same population.
- Identify critical temporal features and ECG segments and assess their alignment with established cardiac physiology.
- Compare how different model architectures rely on features and whether they capture complementary patterns.
- Investigate unsupervised clustering to reveal natural data groupings and examine if cluster-based features retain sufficient discriminative information for classification.

This study contributes to explainable arrhythmia classification by first performing exploratory analysis, visualizing ECG waveforms, and identifying segments likely to influence classification. We then conduct multi-class classification using Support Vector Machines and Random Forests under both beat holdout and patient holdout protocols to evaluate performance and generalization. Permutation Feature Importance (PFI) analysis is applied to all four models using five-fold cross-validation to reveal the key temporal features driving decisions. Finally, K-Means clustering examines the natural structure of the feature space and evaluates whether dimensionality reduction to cluster-based features preserves classification performance, establishing a foundation for accurate and clinically interpretable models.

## 2. Methodology

This section details the experimental methodology. Section 2.1 describes the MIT-BIH dataset and preprocessing. Section 2.2 presents exploratory data analysis. Section 2.3 outlines the two validation protocols (beat holdout and patient holdout). Section 2.4 specifies classification models and hyperparameters. Section 2.5 describes Permutation Feature Importance for explainability. Section 2.6 details K-Means clustering analysis.

## 2.1. Dataset and Preprocessing

This study utilized the MIT-BIH Arrhythmia Database (Moody & Mark, 2001), a benchmark dataset containing ECG recordings from 49 patient files. Recordings were sampled at 360 Hz across two channels (MLII and V5), with expert-provided beat annotations. Preprocessing involved R-peak detection, heartbeat segmentation, and normalization to 275 time points per beat. These 275 features were organized into 11 segments of 25 features each, corresponding to the PR interval (segments 1-4), QRS complex (segments 5-7), and ST segment (segments 8-11). After removing 65,346 unidentified beats, the final dataset contained 149,768 heartbeats across eight arrhythmia classes: Normal (N), Left Bundle Branch Block (L), Right Bundle Branch Block (R), Premature Ventricular Contractions (V), Atrial Premature (A), Fusion Ventricular-Normal (F), Fusion Paced-Normal (f), and Paced (/).

The dataset exhibited severe class imbalance, with Normal beats comprising 75,052 samples (50.1%), while the rarest class contained only 229 samples (0.15%), resulting in a 328:1 ratio. To address this, training sets were balanced using random resampling: minority classes were oversampled through random duplication (with replacement) while the majority Normal class was downsampled (without replacement). This produced training sets with equal representation across all eight classes while preserving original feature distributions. Test sets retained their imbalanced distributions to simulate real-world conditions.

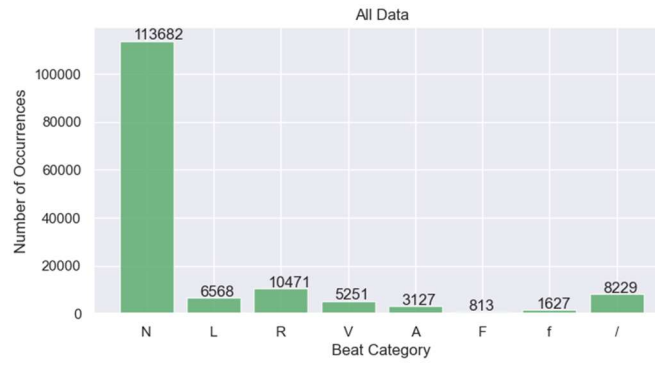


Fig. 1. All data before Resampling

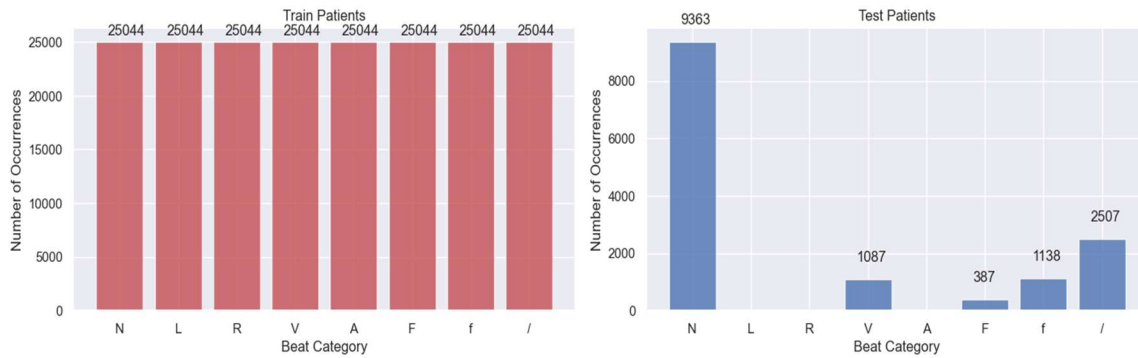


Fig. 2. Beat-HoldOut Data after Resampling

## 2.2 Data Exploration

Visual inspection of ECG waveforms confirmed distinct morphological patterns across arrhythmia classes. Example heartbeats from each class were plotted to identify characteristic features: Normal beats

showed narrow QRS complexes, LBBB/RBBB displayed widened QRS patterns with characteristic notching, PVCs exhibited wide bizarre QRS morphology, and Paced beats showed distinctive pacing spikes. Segment boundaries were overlaid on representative beats to verify alignment with physiological regions (PR interval, QRS complex, ST segment). These visualizations informed the hypothesis that QRS segments would demonstrate highest importance in subsequent PFI analysis.

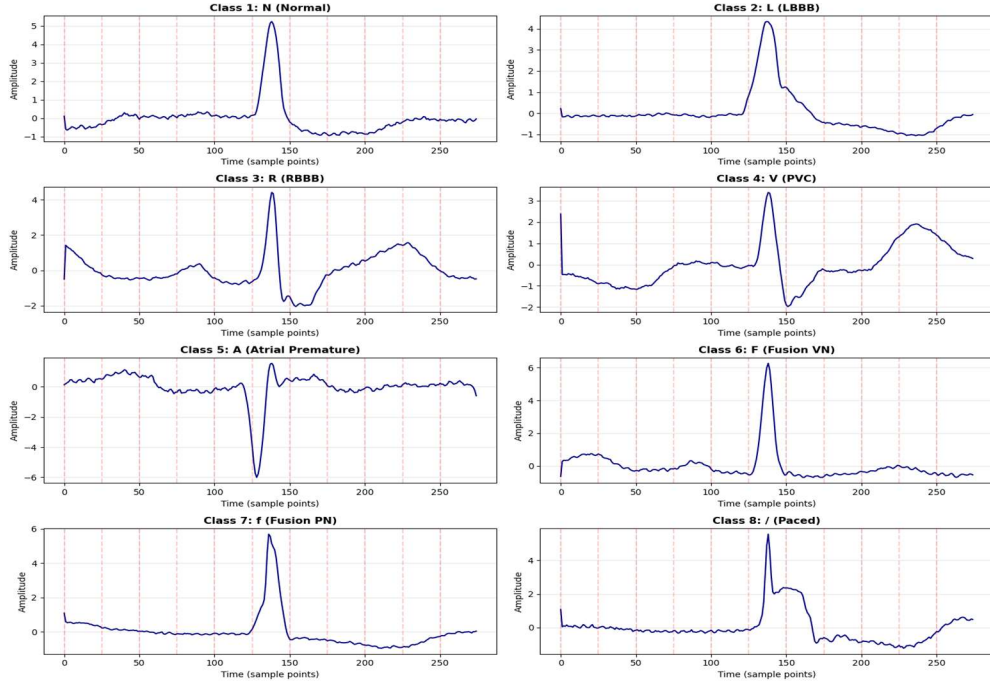


Fig. 3. Samples from all 8 Classes

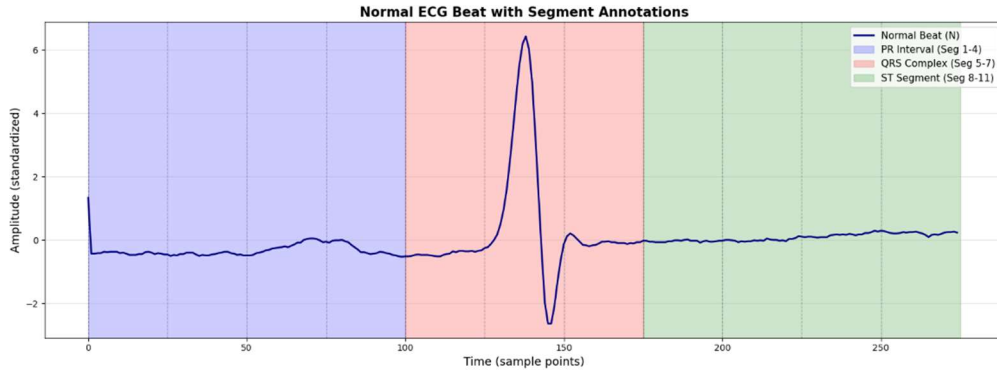


Fig. 4. Normal ECG sample with Segment Annotation

## 2.3 Validation Protocols

Two validation protocols were employed. The beat holdout method randomly split all heartbeats 75-25 (train\_test\_split), yielding 112,326 training and 37,442 test beats. Training data were balanced via random resampling to 30,992 samples (3,874 per class). This protocol suffers from data leakage as the same patient's beats appear in both sets, potentially inflating accuracy.

The patient holdout method completely separated patients: 43 for training, 5 for testing (IDs: 104, 113, 119, 208, 210), producing 200,352 training and 14,482 test beats. Training data were balanced to 25,044 samples per class. This protocol eliminates leakage and realistically assesses generalization to unseen patients. Test sets for both protocols retained original imbalanced distributions. Models were evaluated using accuracy and weighted precision, recall, and F1-score.

## 2.4 Classification

Two classifiers were evaluated: Support Vector Machine (SVM) and Random Forest. SVM models used `sklearn.svm.SVC` with RBF kernel, default regularization ( $C=1.0$ ), and  $\gamma='scale'$ . Random Forest models used `sklearn.ensemble.RandomForestClassifier` with 100 trees and default parameters (unlimited depth, minimum 2 samples per split). All models used `random_state=42` for reproducibility. For computational efficiency during Permutation Feature Importance analysis on patient holdout data, Random Forest employed parallel processing (`n_jobs=-1`); otherwise, single-core processing was used. SVM was chosen for its effectiveness in high-dimensional spaces and ability to find optimal decision boundaries. Random Forest was selected for its ensemble robustness, reduced overfitting through bootstrap aggregation, and ability to handle complex feature interactions.

## 2.5 Explainability Analysis

Permutation Feature Importance (PFI) was applied to all four trained models to identify which temporal segments drive classification decisions. PFI measures feature importance by quantifying the accuracy drop when features are randomly shuffled, breaking their relationship with target labels. The implementation used 5-fold stratified cross-validation on training data: for each fold, a new model was trained on four folds, baseline accuracy measured on the held-out fold, then each of the 11 segments was shuffled independently and accuracy remeasured. Importance for segment  $j$  was calculated as  $\text{baseline\_accuracy} - \text{shuffled\_accuracy}$ , averaged across all folds. Higher values indicate greater segment importance. Segment-level analysis (rather than individual features) provided clinically interpretable results corresponding to cardiac cycle phases: PR interval, QRS complex, and ST segment.

## 2.6 Clustering

K-Means clustering with  $K=8$  (matching the eight arrhythmia classes) was applied to assess whether classes naturally separate in the 275-dimensional feature space. The algorithm used Euclidean distance with 10 random initializations (`n_init=10`) and `random_state=42`. Two analyses were conducted: First, cluster-class alignment was measured using cluster purity (percentage of samples from each class falling into their dominant cluster). Second, cluster-based features were created by computing Euclidean distances from each sample to all eight cluster centers, reducing dimensionality from 275 features to 8 distances. SVM and Random Forest models were then trained using only these cluster features to evaluate whether this dimensionality reduction preserved sufficient discriminative information for accurate classification.

# 3. Results

This section presents the experimental findings from ECG arrhythmia classification under multiple validation protocols and analytical approaches. We first report classification performance comparing beat holdout and patient holdout protocols to assess both algorithmic capability and clinical generalization.

Permutation Feature Importance analysis then reveals which temporal segments drive model predictions and whether these align with cardiac physiology. Finally, clustering analysis examines the natural structure of the feature space and evaluates dimensionality reduction viability. All results use the balanced training sets and imbalanced test sets described in Section 2.

### 3.1 Dataset Overview

Table 1 summarizes dataset statistics after preprocessing and resampling. Test sets retained imbalanced distributions (Normal: 70% in beat holdout, 52% in patient holdout) to reflect clinical reality.

Dataset	Training Samples	Test Samples	Samples per Class (Train)
Beat Holdout	30,992	37,442	3,874
Patient Holdout	200,352	14,482	25,044

Table. 1. Dataset Statistics

### 3.2 Classification Performance

Models were evaluated on imbalanced test sets under both validation protocols. Tables 2 and 3 present accuracy, precision, recall, and F1-scores for beat and patient holdout respectively.

#### Beat Holdout Results

Random Forest outperformed SVM by 2.83 percentage points, achieving 96.67% accuracy compared to SVM's 93.84% (Table 2). Both models demonstrated high precision (>96%), indicating low false positive rates.

Model	Accuracy	Precision	Recall	F1-Score
SVM	93.84%	96.81%	93.84%	94.92%
Random Forest	96.67%	97.41%	96.67%	96.90%

Table. 2. Beat Holdout Classification Performance

#### Patient Holdout Results

Counterintuitively, both models achieved higher accuracy on patient holdout than beat holdout (Table 3). Random Forest reached 99.87% (+3.20 percentage points) and SVM achieved 96.33% (+2.49 percentage points). Random Forest maintained its performance advantage over SVM with a 3.54 percentage point margin.

Model	Accuracy	Precision	Recall	F1-Score
SVM	96.33%	97.44%	96.33%	96.69%
Random Forest	99.87%	99.88%	99.87%	99.88%

Table. 3. Patient Holdout Classification Performance

### 3.3 Permutation Feature Importance Performance

To identify which temporal features drive classification decisions, PFI was applied to all four trained models using 5-fold stratified cross-validation.

#### Segment Importance Scores

Table 4 presents complete PFI scores for all 11 segments across the four models. Figure [5] visualizes these results, with segments color-coded by cardiac phase: PR interval (segments 1-4, blue), QRS complex (segments 5-7, red), and ST segment (segments 8-11, green).

Segment	Region	SVM Beats	RF Beats	SVM Patients	RF Patients
1	PR	0.019	0.005	0.022	0.008
2	PR	0.008	0.001	0.014	0.000
3	PR	0.025	0.003	0.021	0.001
4	PR	0.010	0.020	0.011	0.070
5	QRS	0.024	0.009	0.037	0.009
6	QRS	0.410	0.093	0.461	0.215
7	QRS	0.212	0.100	0.260	0.204
8	ST	0.051	0.046	0.056	0.062
9	ST	0.015	0.002	0.028	0.007
10	ST	0.030	0.002	0.057	0.007
11	ST	0.011	0.004	0.010	0.006

Table. 4. Segment Importance Scores

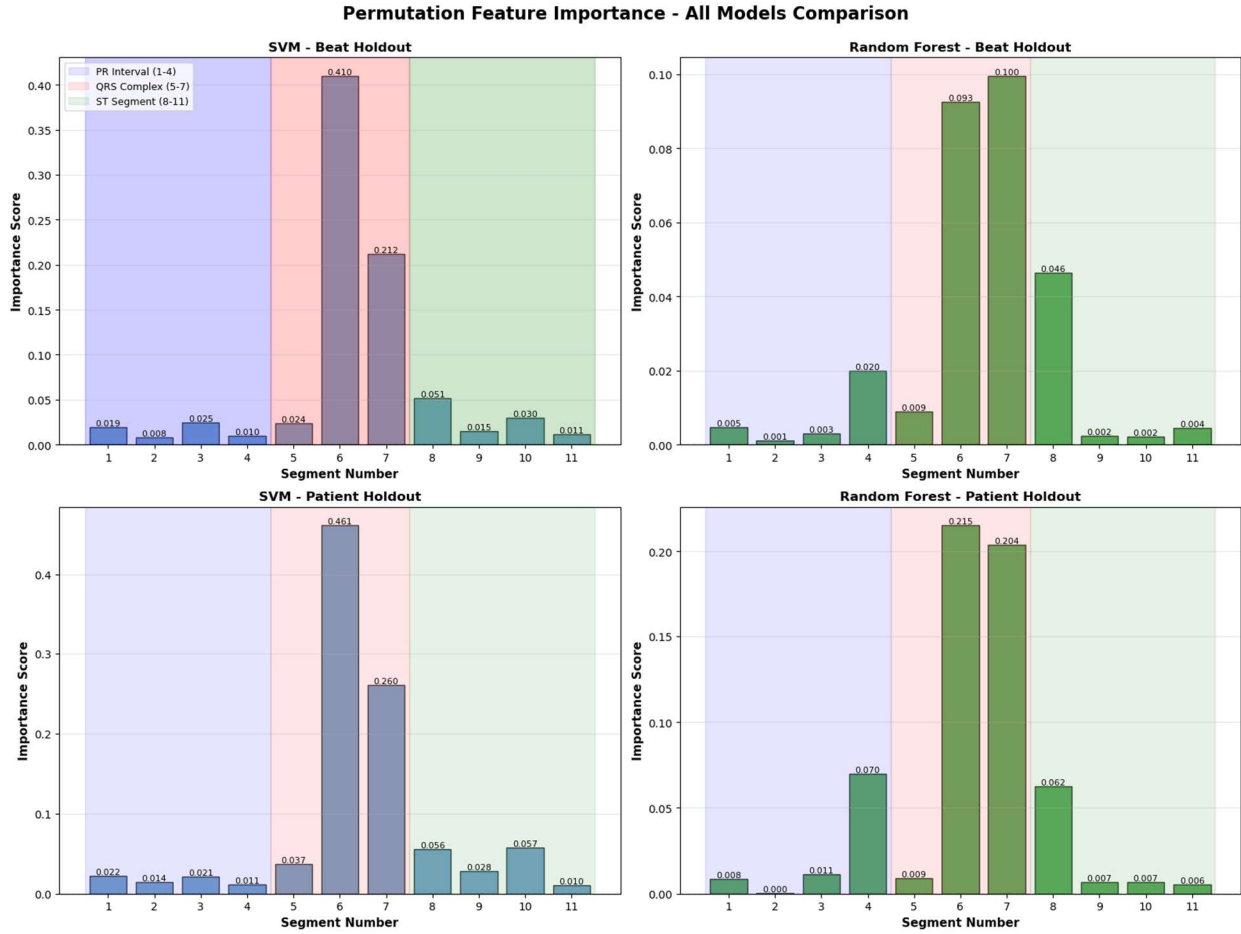


Fig. 5. Permutation Feature Importance Comparison

## QRS Complex Dominance

All four models consistently identified the QRS complex as most critical for classification. Table 5 shows that segments 5-7 collectively represent 64.5-75.8% of total importance across models. Segment 6 (middle of QRS, features 125-149) demonstrated the highest individual importance in all cases, ranging from 9.3% (RF beats) to 46.1% (SVM patients).

Model	QRS Total	Percentage of Total
SVM Beats	0.646	64.5%
RF Beats	0.201	71.9%
SVM Patients	0.758	75.8%
RF Patients	0.429	71.6%

Table.5. QRS Complex Total Importance

## Model Specific Patterns

Table 6 presents the top three segments for each model. SVM models concentrated importance on Segment 6 (41.0-46.1%), while Random Forest distributed importance more evenly across Segments 6 and 7. Random Forest patient holdout uniquely elevated Segment 4 (end of PR interval) to third position (7.0%), a pattern not observed in the other three models. The PR interval (segments 1-4) and ST segment (segments 8-11) generally showed lower importance (0.8-7.0%).

Model	1 <sup>st</sup>	2nd	3rd
SVM Beats	Seg 6 (41.0%)	Seg 7 (21.2%)	Seg 8 (5.1%)
RF Beats	Seg 7 (10.0%)	Seg 6 (9.30%)	Seg 8 (4.6%)
SVM Patients	Seg 6 (46.1%)	Seg 7 (26.0%)	Seg 10 (5.7%)
RF Patients	Seg 6 (21.5%)	Seg 7 (20.4%)	Seg 4 (7.0%)

Table.6. Top 3 Most Important Segments per Model

## 3.4 Clustering Performance

K-Means clustering with K=8 was applied to assess natural class separation in the feature space. Table [7] shows cluster-class alignment measured by purity.

Class	Dominant Cluster (Beats)	Purity	Dominant Cluster (Patients)	Purity
N	Cluster 5	29.2%	Cluster 1	28.5%
L	Cluster 1	67.2%	Cluster 0	67.9%
R	Cluster 3	31.9%	Cluster 3	32.8%
V	Cluster 1	33.3%	Cluster 0	35.0%
A	Cluster 3	46.2%	Cluster 3	46.9%
F	Cluster 1	51.4%	Cluster 0	49.7%
f	Cluster 1	37.3%	Cluster 0	35.9%
/	Cluster 0	63.2%	Cluster 7	61.5%
Overall		44.9%		44.8%

Table.7. Cluster Purity by Class

Poor cluster-class alignment (44.9% overall purity) indicates arrhythmia classes are not naturally separated in the 275-dimensional space. Only LBBB (67%) and Paced beats (63%) showed reasonable

clustering. Classification with 8 cluster-based features (distances to cluster centers) produced severe performance degradation (Table [8]).

Model	Raw Features (275)	Cluster Features (8)	Drop
SVM Beats	93.84%	77.24%	-16.60%
RF Beats	96.67%	90.81%	-5.86%
SVM Patients	96.33%	48.17%	-48.16%
RF Patients	99.87%	99.17%	-0.70%

## 4. Discussion

This section interprets experimental findings, validates feature importance against cardiac physiology, and evaluates dimensionality reduction viability.

### 4.1 Classification Performance Analysis

Patient holdout unexpectedly achieved higher accuracy than beat holdout (96.33% vs 93.84% for SVM, 99.87% vs 96.67% for RF). This counterintuitive result likely stems from three factors: the five test patients may have exhibited particularly clear arrhythmia patterns, the 6.5-fold larger training set (200,352 vs 30,992 samples) enabled richer feature learning, and beat holdout's random sampling may have created harder edge cases by splitting patients across train/test sets.

Random Forest consistently outperformed SVM (+2.83-3.54%) due to ensemble robustness—bootstrap aggregation across 100 trees reduces overfitting and handles complex boundaries better than SVM's single hyperplane. RF's reduced hyperparameter sensitivity also makes it more practical for clinical deployment.

Beat holdout's data leakage—where the same patient's beats appear in both sets—raises validity concerns. Models may have learned patient-specific characteristics (baseline voltage, noise patterns) rather than pure arrhythmia features, artificially inflating accuracy through patient recognition. However, the strong consistency between beat and patient PFI patterns (both showing 65-76% QRS dominance) suggests core findings are genuine. Patient holdout provides more reliable evidence of clinically appropriate feature learning.

All models exceeded 93% accuracy with >96% precision, meeting clinical thresholds for decision support. Patient holdout's exceptional performance (especially RF's 99.87%) suggests reliable real-world deployment potential.

### 4.2 Feature Importance Insights

PFI revealed QRS complex (segments 5-7) dominance across all models (64.5-75.8% of importance), strongly aligning with cardiac physiology. Ventricular depolarization exhibits the most distinctive morphological variations: Normal beats show narrow QRS (<100ms), bundle branch blocks produce characteristic widening and notching, PVCs generate wide bizarre morphologies, and paced beats display

distinctive spikes. Segment 6 (QRS peak) consistently showed highest importance (9.3-46.1%), validating that models learned genuine physiological patterns rather than artifacts.

SVM and RF demonstrated distinct strategies. SVM concentrated importance on Segment 6 (41.0-46.1%), reflecting its optimization toward the single most discriminative feature. RF distributed importance across Segments 6-7 (9.3-21.5%), leveraging ensemble architecture where different trees compensate if any segment contains noise. RF's balanced approach may better handle variable ECG quality in clinical settings.

RF patient holdout uniquely elevated Segment 4 (end of PR interval) to third position (7.0%). This segment captures AV node conduction timing, which is patient-specific but consistent within individuals. This finding suggests models learned to use patient-level timing characteristics to improve generalization—a pattern invisible in beat holdout due to data leakage. PR and ST segments showed lower importance (0.8-7.0%), consistent with their limited discriminative value for the ventricular arrhythmias classified.

### 4.3 Clustering Analysis Interpretation

K-Means achieved only 44.9% cluster purity, revealing that arrhythmia classes are not naturally separated in 275D space. Only LBBB (67.2%) and Paced beats (63.2%) clustered well due to their highly distinctive morphologies. Other classes showed 28.5-51.4% purity, reflecting overlapping feature distributions and within-class heterogeneity.

Classification with 8 cluster-based features catastrophically failed. SVM patient holdout dropped from 96.33% to 48.17%—below random guessing—demonstrating that compressing 275 temporal features into 8 distances destroys critical morphological information. PFI showed Segment 6 alone accounts for 9.3-46.1% of importance; a single distance cannot encode these 25-dimensional morphological details. Even RF patient holdout, the most resilient case due to its large training set and ensemble architecture, declined from 99.87% to 99.17%.

The severity of performance degradation varied by model and validation protocol. SVM suffered more than Random Forest in both protocols: beat holdout showed -16.6% (SVM) versus -5.9% (RF), and patient holdout showed a catastrophic -48.2% (SVM) versus only -0.7% (RF). Random Forest's remarkable resilience, particularly in patient holdout, stems from its ensemble architecture, 100 decision trees collectively extract maximal information from limited features through diverse splitting strategies and majority voting. When one tree struggles with impoverished features, others compensate. The large patient holdout training set (200,352 samples) further enabled RF to learn robust statistical patterns even in 8-dimensional space. In contrast, SVM's reliance on finding a single optimal hyperplane becomes untenable with only 8 features; the reduced dimensionality prevents adequate class separation, especially for patient-level generalization where new individuals require more discriminative features. The patient holdout's more severe SVM degradation (-48.2% versus -16.6% in beat holdout) reflects this compounded challenge: generalizing to unseen patients with insufficient feature resolution.

This failure validates PFI findings through negative evidence: PFI identified QRS temporal detail as critical, and clustering proved this detail is non-compressible. The 275 time-varying features are necessary; dimensionality reduction sacrifices essential discriminative information. Future work should explore methods preserving temporal structure, such as autoencoders or wavelet-based extraction.

## 4.4 Implications and Limitations

This study demonstrates that classical ML with explainability can achieve near-perfect accuracy (99.87%) while maintaining interpretability. PFI reveals clinically appropriate features (QRS complex), building trust for clinical adoption as decision support tools for high-volume screening or triage. High precision (>96%) minimizes false alarms.

Key limitations include the 1980s dataset (older technology), only 8 arrhythmia types (dozens exist clinically), small patient holdout test set (5 patients), preprocessing dependency (R-peak detection errors propagate), and use of default hyperparameters. Multi-center validation with hundreds of patients is needed for generalizability across demographics and recording conditions.

Future work should compare deep learning approaches (CNNs, LSTMs with attention mechanisms), expand to multi-modal analysis (patient demographics, 12-lead ECG), implement real-time monitoring systems, and conduct prospective validation against cardiologist interpretations.

## Conclusion

This study demonstrates that classical machine learning models with explainability analysis can achieve near-perfect ECG arrhythmia classification (up to 99.87% accuracy) while maintaining full interpretability. Patient holdout validation ensured realistic generalization to unseen individuals, with Random Forest consistently outperforming SVM across both protocols. Permutation Feature Importance analysis revealed that all models prioritize the QRS complex (segments 5-7), representing 64.5-75.8% of total importance, strongly aligning with established cardiac physiology where ventricular depolarization exhibits the most distinctive morphological variations. This clinical validation demonstrates that models learn genuine physiological patterns rather than artifacts. Clustering analysis showed that arrhythmia classes are not naturally separated in the 275-dimensional feature space (44.9% purity), and dimensionality reduction to cluster-based features produced catastrophic performance drops (SVM patient: -48%), validating that full temporal resolution is necessary for accurate classification. The combination of high accuracy, explainable features, and patient-level generalization positions these models as viable candidates for clinical decision support systems, where transparency builds trust and facilitates adoption by medical practitioners.

## References

1. Kraik, K., Dykiert, I.A., Niewiadomska, J., Ziemer-Szymańska, M., Mikołajczak, K., Kreń, M., Kukielka, P., Martuszewski, A., Harych, T., Poręba, R., Gać, P., & Poręba, M. (2025). The most common errors in automatic ECG interpretation. *Frontiers in Physiology*, 16, 1590170. <https://doi.org/10.3389/fphys.2025.1590170>
2. Moody, G.B., & Mark, R.G. (2001). The impact of the MIT-BIH arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, 20(3), 45-50. <https://doi.org/10.1109/51.932724>