Multi-Label Arrhythmia Classification of 12-Lead ECG Using Physiologically Interpretable Handcrafted Features and Classical Machine Learning: A Patient-Independent Validation Study

LitLit AI[a]

[a]LitLit AI; contact@litlit.ai

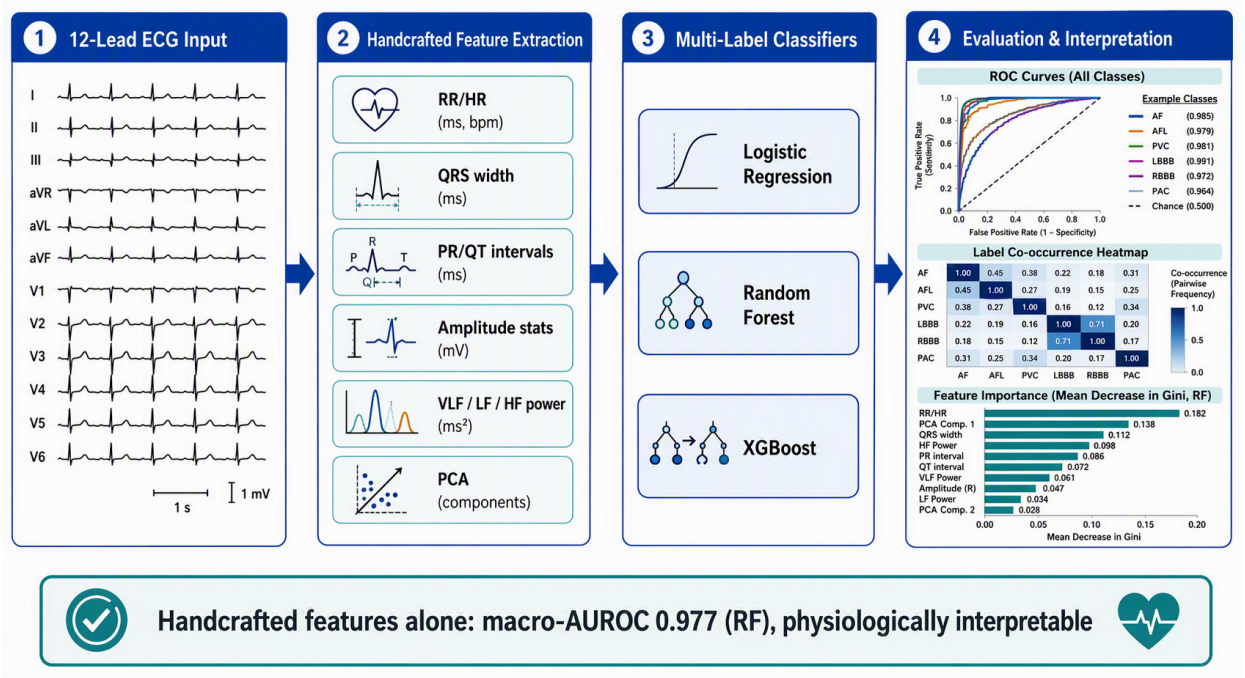


Figure 1. Graphical abstract. End-to-end workflow for handcrafted-feature multi-label arrhythmia classification of 12-lead ECG. Each 10-second recording is downsampled to 250 Hz and summarized by per-lead physiological features (RR-based heart rate, QRS width, PR/QT intervals, amplitude statistics, and VLF/LF/HF spectral power) optionally augmented with 10 principal components. Three classical classifiers—logistic regression, random forest, and XGBoost—are trained per-class under patient-independent splitting, and feature importance is mapped back to physiologically meaningful groups and leads. The best model (random forest) achieves a macro AUROC of 0.977.

Abstract

Background. Deep neural networks have set new benchmarks for automated 12-lead ECG diagnosis, yet they demand large labeled corpora, costly GPU infrastructure, and offer limited mechanistic interpretability. Handcrafted, physiology-driven features paired with classical machine learning remain attractive for resource-constrained clinical deployment and for generating transparent, clinician-auditable decision pathways, but their performance on the exact 6-arrhythmia benchmark defined by the CODE/PhysioNet ecg-arrhythmia challenge has seldom been reported with a rigorous leakage-controlled protocol.

Methods. Using the openly licensed test partition of the CODE dataset (Ribeiro et al., Nat Commun 2020; 827 records, 12 leads, 400 Hz), we extracted 401 handcrafted features per record: 32 per-lead summary descriptors (RR-derived heart rate and variability,

Pan–Tompkins + wavelet QRS width, PR/QT intervals with Bazett and Fridericia corrections, amplitude statistics, and Welch spectral power in VLF/LF/HF bands) across 12 leads, 7 global heart-rate-variability features, and 10 principal components of the raw waveform. After constant-feature removal, correlation pruning, and train-only refit of the scaler/PCA/selector pipeline, 276 features entered modeling. A patient-independent multilabel-stratified 70/15/15 split (`random_state=42`; 578/124/125) was used. Per-class L2-regularized logistic regression, random forest ($n = 300$), and XGBoost ($n = 400$, `max_depth=6`, $\eta = 0.05$) with class-weight balancing were trained, with per-class thresholds optimized on the validation set. Evaluation comprised per-class AUROC, AUPRC, macro F1, a 6-class co-occurrence matrix, and 1000-iteration bootstrap 95% confidence intervals for macro F1. Random-forest Gini and XGBoost weight+gain importances were aggregated into 7 feature groups and 12 leads, with a dedicated LBBB→1dAVb confusion analysis quantified by Cohen's d.

Results. Random forest achieved the best macro F1 (0.621, 95% CI 0.409–0.754), macro AUROC (0.977), and macro AUPRC (0.710) on the held-out test set, followed by XGBoost (0.598, 0.973, 0.668) and logistic regression (0.432, 0.921, 0.547). Per-class AUROC exceeded 0.90 for all arrhythmias except 1dAVb under logistic regression; LBBB was perfectly separated (AUROC 1.0, F1 0.91). Dominant feature groups aligned with physiology: 1dAVb→PR/QT, LBBB/RBBB→QRS, and SB/AF/ST→RR/HR. Lead dominance was clinically plausible (RBBB→V1, LBBB→V3, AF→V5). The 10 PCA components explained only 7.1% of variance and were marginally beneficial for random forest (ΔF1 +0.019) but slightly detrimental for XGBoost (ΔF1 −0.027), confirming that handcrafted features carry nearly all discriminative signal. The single top confusion pair (LBBB→1dAVb, $n = 2$) was driven by shared QT morphology, with QRS width yielding the largest separation (Cohen's $d = 1.97$).

Conclusions. Physiology-grounded handcrafted features and classical ensembles—executable entirely on a CPU—deliver strong, interpretable multi-label arrhythmia classification on the CODE/ecg-arrhythmia benchmark, with feature importance that maps cleanly onto established electrocardiographic criteria. The approach offers a transparent, deployable complement to deep networks and a tractable baseline for resource-limited settings.

Keywords: electrocardiography; arrhythmia classification; handcrafted features; multi-label classification; random forest; XGBoost; interpretable machine learning.

1 Introduction

The electrocardiogram (ECG) remains the most widely deployed non-invasive cardiac diagnostic test, and automated interpretation has been a persistent goal of biomedical engineering for over five decades [1, 2]. Reliable computer-assisted reading could alleviate the global shortage of expert electrocardiographers and shorten the time to diagnosis of life-threatening arrhythmias. Among the morphological and rhythmic abnormalities targeted by automated systems, six conditions recur as a canonical benchmark set: first-degree atrioventricular block (1dAVb), right and left bundle branch block (RBBB and LBBB), sinus bradycardia (SB), atrial fibrillation (AF), and sinus tachycardia (ST) [3]. Each carries distinct electrocardiographic signatures—prolonged PR interval for 1dAVb, widened QRS with characteristic lead-dependent morphology for the bundle branch blocks, and rate-driven signatures for the sinus rhythm disturbances [4, 5]—making them an attractive testbed for feature-engineering studies.

In recent years, deep neural networks have driven rapid progress. Ribeiro and colleagues [3] trained a convolutional network on more than two million 12-lead ECGs and reported cardiologist-level AUROC values on the six-arrhythmia benchmark that forms the basis of the present study. Hannun et al. [6] demonstrated cardiologist-level ambulatory arrhythmia detection with a deep sequence model, and Attia et al. [7] showed that a network can extract contractile dysfunction signatures invisible to the human eye. A large subsequent literature has extended these results across single- and multi-lead settings [8, 9, 10, 11, 12, 13, 14, 15]. These results are compelling, but they come at a cost: deep models require very large labeled corpora, are typically trained and served on specialized GPU hardware, and behave as opaque "black boxes" whose decisions resist direct physiological audit [5, 16, 17]. In safety-critical clinical workflows, the inability to map a model's reasoning back to established electrocardiographic criteria is a non-trivial barrier to adoption.

Handcrafted features paired with classical machine learning offer a structurally different trade-off. By encoding domain knowledge directly—RR intervals and heart-rate variability [18], QRS width and morphology [19, 20], PR and QT intervals [21], and spectral band power—these pipelines produce inputs whose dimensions are themselves interpretable physiological quantities. Classical classifiers such as logistic regression [22], random forests [23], and gradient-boosted trees [24, 25] then yield importance metrics that can be aggregated by lead and by feature family, producing an auditable decision trace. De Chazal et al. [2] established the paradigm for heartbeat classification using morphology and interval features, and surveys of the field [16, 5] confirm that handcrafted approaches remain competitive, especially when interpretability and computational economy matter. Several studies have specifically benchmarked handcrafted against learned representations for arrhythmia and heart-failure ECG classification [26, 27, 28], and interpretable tree-based pipelines on 12-lead ECG have demonstrated the value of expert features combined with gradient-boosted trees [29, 30]. Publicly available benchmark datasets—the MIT-BIH Arrhythmia Database [1], PhysioNet [31, 32], PTB-XL [33, 34], and the PhysioNet/CinC challenges [35, 36]—together with the CODE and CODE-II corpora [3, 37] underpin reproducible comparison of these approaches. Multi-label settings further require careful stratification [38, 39] and metrics that are robust to class imbalance, such as AUPRC and macro-averaged F1 [40, 41, 42, 43].

Despite this, comparatively few studies report handcrafted-feature performance on the exact six-arrhythmia CODE/ecg-arrhythmia benchmark under a leakage-controlled, patient-independent protocol, and fewer still accompany their results with a physiological dissection of *which* features and *which* leads drive each diagnosis. This study fills that gap. We ask three questions. **(Q1)** Can handcrafted features alone—computed on a CPU—attain clinically credible multi-label performance on the six-class benchmark? **(Q2)** Does the raw-waveform PCA block that augments the features add meaningful signal beyond the engineered descriptors? **(Q3)** Do the learned feature importances map onto established electrocardiographic physiology, and can the residual confusion between morphologically similar blocks be quantified in feature space?

We hypothesize that (H1) a random forest or gradient-boosted ensemble of handcrafted features will reach macro AUROC ≥ 0.95 on the held-out test set; (H2) the PCA block will be at most marginally beneficial, with engineered features carrying the bulk of the signal; and (H3) feature-group and lead-level importance will recover known clinical patterns (PR/QT for 1dAVb, QRS for bundle branch blocks, RR/HR for rate-driven arrhythmias, and lead-dominance such as RBBB→V1). The remainder of the paper describes the dataset, feature-engineering pipeline, modeling protocol, evaluation, and a detailed interpretability analysis.

2 Methods

2.1 Dataset

Analyses use the openly licensed test partition of the CODE dataset released with the automatic-ECG-diagnosis repository of Ribeiro et al. [3] (Zenodo record 3765780; CC-BY-4.0). The bundle comprises 827 twelve-lead ECG records acquired at 400 Hz, zero-padded to 4 096 samples (nominal 10.24 s; effective durations range from 7.335 s to 10.24 s, mean 8.34 s; 538 records of 7 s and 289 of 10 s). Signals are stored in units of 10^{-4} V and were converted to millivolts ($\times 0.1$) prior to processing. Each record is annotated against six diagnostic classes—1dAVb, RBBB, LBBB, SB, AF, and ST—and 681 records are normal while 146 carry at least one abnormal label (multi-label; 827 records correspond to 827 distinct patients, so record-level and patient-level partitioning coincide). Class prevalences are imbalanced, ranging from 13 (AF, 1.57%) to 37 (ST, 4.47%) (Table 1); the most frequent co-occurrence is 1dAVb+RBBB (4 records).

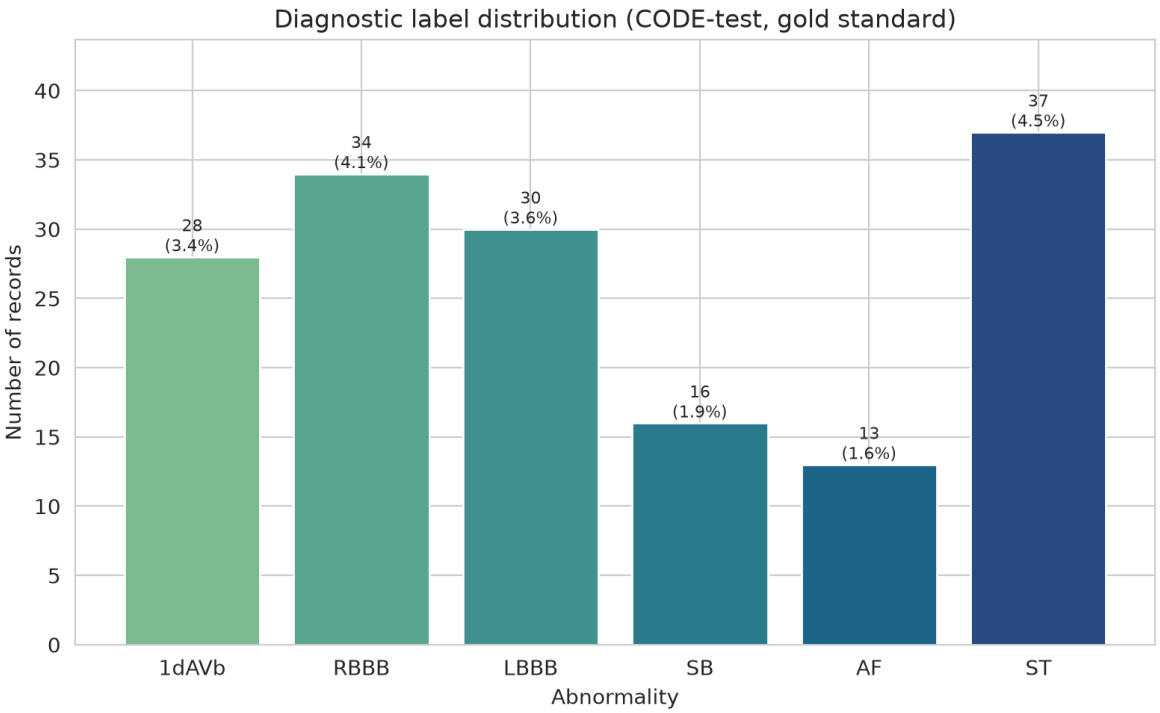
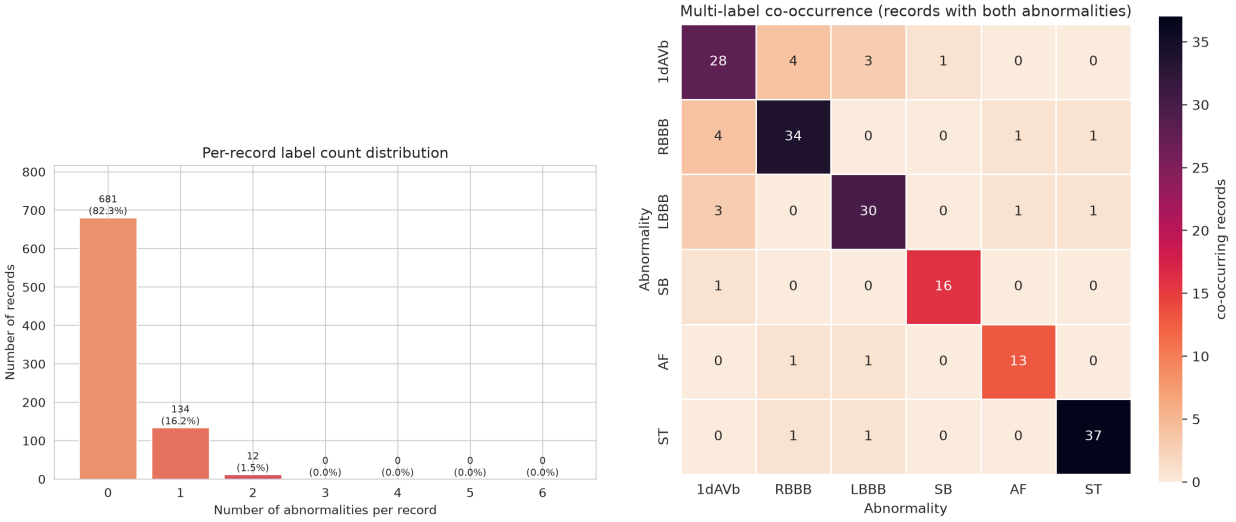


Figure 2. Class distribution of the CODE test partition. The six abnormal classes are markedly imbalanced; ST and RBBB are the most prevalent, while AF is the rarest.



(a) Records by label-cardinality

(b) Label co-occurrence matrix

Figure 3. Multi-label structure of the dataset. (a) Number of records carrying 0–6 abnormal labels (most abnormal records carry exactly one label). (b) Symmetric co-occurrence matrix of the six classes; the 1dAVb+RBBB pair co-occurs most frequently.

Table 1. Label counts and proportions in the CODE test partition ($N = 827$; 681 normal, 146 abnormal).

Class	Description	Count	Proportion
1dAVb	1st-degree AV block	28	3.39%
RBBB	Right bundle branch block	34	4.11%
LBBB	Left bundle branch block	30	3.63%
SB	Sinus bradycardia	16	1.93%
AF	Atrial fibrillation	13	1.57%
ST	Sinus tachycardia	37	4.47%
Normal	(no abnormal label)	681	82.34%

Representative 12-lead tracings illustrating the morphological heterogeneity targeted by the feature set are shown in Figure 4.

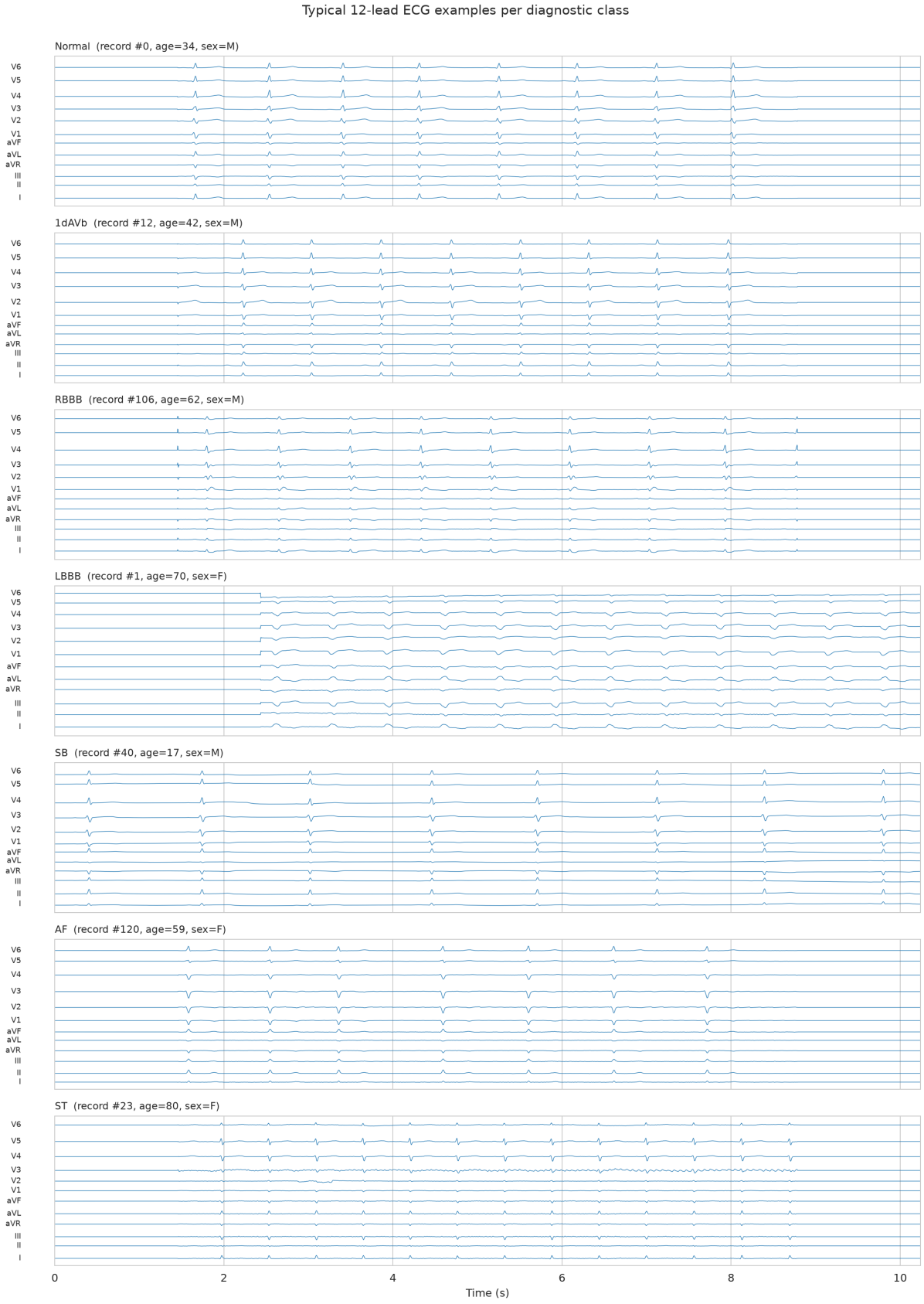


Figure 4. Representative 12-lead ECG examples for each diagnostic class, illustrating the morphological and rhythmic heterogeneity that the handcrafted features are designed to capture.

2.2 Feature Engineering

Each record was processed through a deterministic pipeline. Records were de-padded to their effective length, resampled from 400 Hz to 250 Hz using a polyphase FIR anti-aliasing filter (ratio 5/8), band-pass filtered (0.5–40 Hz, second-order Butterworth), and converted to millivolts. Within each fixed 10-second analysis window, 32 summary features were computed per lead, organized into seven physiological groups (Table 2).

Rhythm and RR-based features (RR_HR). R-peaks were detected with a Pan–Tompkins detector [19, 20] (Passband 5–15 Hz, 80 ms integration window, 200 ms refractory), yielding per-lead mean heart rate, mean RR, SDNN, RMSSD, RR coefficient of variation, and RR min/max. A representative lead (most R-peaks, median 116 peaks/record) was chosen for global HRV statistics [18]. Per-lead RR features are highly redundant (12×7 correlated descriptors of the same rhythm) and were subsequently regularized during preprocessing.

QRS and conduction intervals (QRS, PR_QT). QRS onset/offset were delineated using a wavelet-based delineator [21], giving QRS width; PR and QT intervals were measured per lead, with Bazett [44] and Fridericia [45] rate corrections added as QTc variants. These groups are the direct physiological correlates of conduction disease and AV block [4].

Amplitude statistics (Amplitude). Sixteen per-lead morphological descriptors: mean, std, variance, skew, kurtosis, max, min, median, range, R- and S-amplitude means, energy, and zero-crossing rate.

Spectral power (ECG_spectral and HRV_global). Welch power-spectral density of the band-passed ECG waveform was summarized into VLF (0.0033–0.04 Hz), LF (0.04–0.15 Hz), HF (0.15–0.40 Hz), LF/HF ratio, total power, and normalized LF/HF bands per lead. A separate, global HRV block (7 features) was computed from an RR tachogram interpolated at 4 Hz. Because records are 7–10 s—well below the 5-min minimum recommended for short-term HRV [18, 46]—the VLF band is unresolvable (12 zero-variance features) and the HRV block is treated as exploratory; spectral HRV is nonetheless a long-established marker of sympatho-vagal balance [47].

Raw-waveform PCA (PCA). Each lead's band-passed, z-scored waveform was resampled to a fixed length of 2 500 samples and the 12 leads concatenated (30 000-dim). Ten principal components were retained (transductive fit explained only 6.4% of variance); the leakage implications of this transductive fit are addressed in Section 2.3.

This yields a raw matrix of 827×401 (32 per-lead $\times$ 12 leads = 384, plus 7 global HRV, plus 10 PCA). The feature-extraction succeeded for 99.88% of records (one record had no valid RR on any lead); 28 non-finite values were imputed. Feature distributions and the cross-feature correlation structure are summarized in Figures 5 and 6.

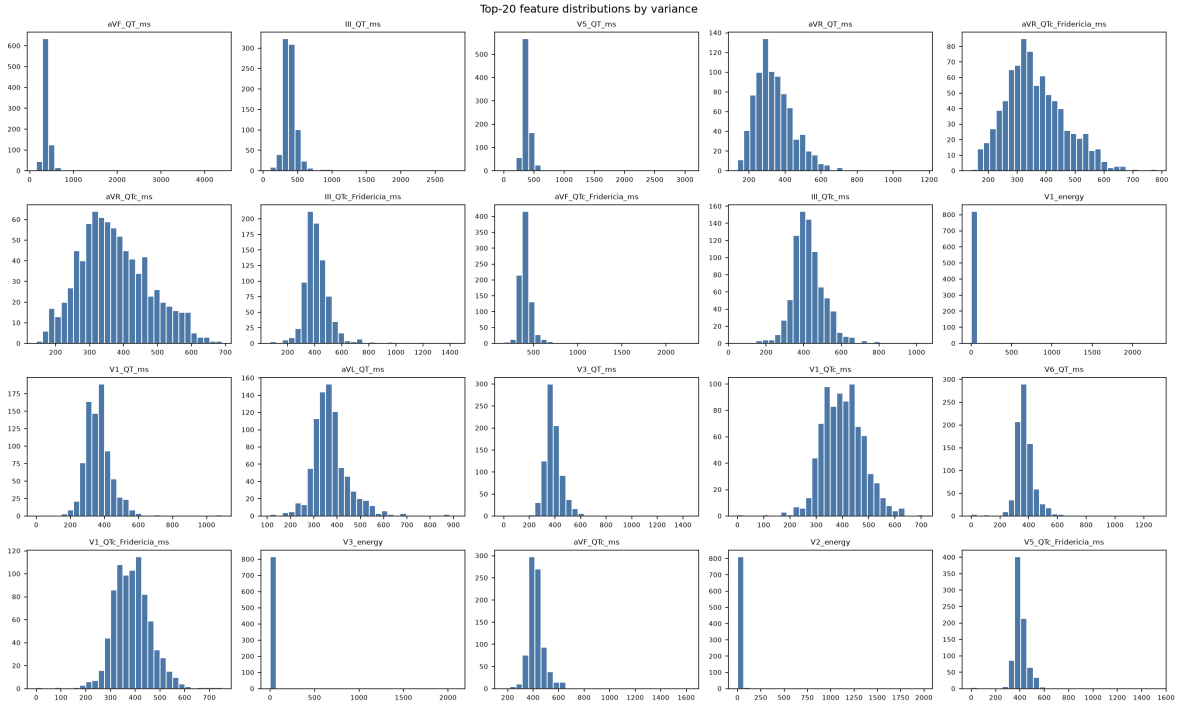


Figure 5. Distributions of representative handcrafted features across the dataset, illustrating the heavy-tailed and class-conditional structure captured by the amplitude, interval, and rhythm descriptors.

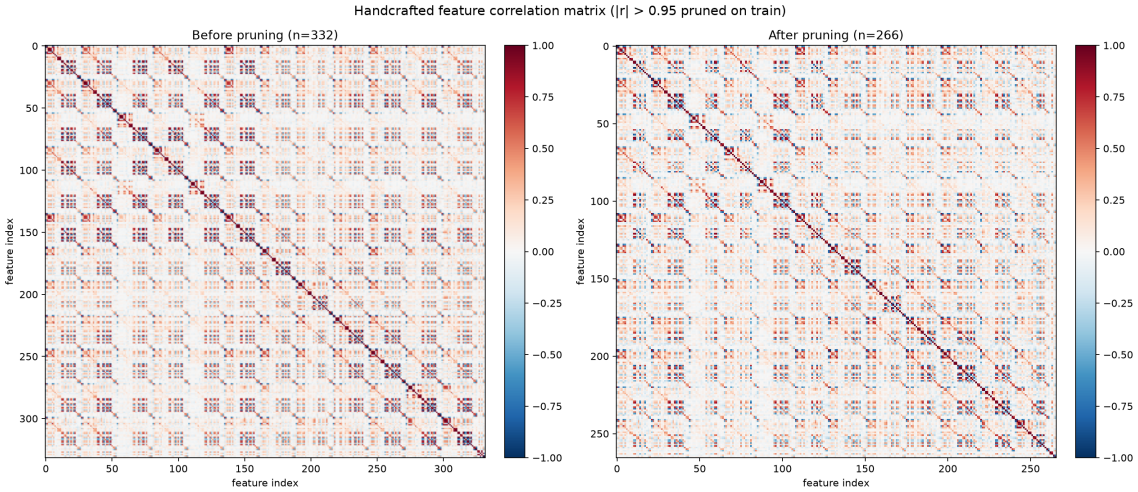


Figure 6. Feature–feature correlation heatmap of the 401 raw features. Redundancy within the per-lead RR/HR family and within amplitude statistics is visible as bright blocks; this motivates the correlation pruning applied before modeling.

Table 2. Physiological feature groups extracted per lead (and globally). n_{feat} is the per-group count before preprocessing.

Group	n_{feat} /lead	Representative features
RR_HR	7	mean_HR, RR_mean, SDNN, RMSSD, RR_cv, RR_min, RR_max
QRS	1	QRS_width_ms
PR_QT	4	PR_ms, QT_ms, QTc_ms, QTc_Fridericia_ms
Amplitude	16	mean, std, var, skew, kurt, max, min, median, range, R/S_amp, energy, zcr
ECG_spectral	6	VLF, LF, HF, LF/HF, total, nLF/nHF
HRV_global	7 (global)	hrv_VLF/LF/HF, LF/HF, total, nLF/nHF
PCA	10 (global)	PCA_1–PCA_10 of concatenated raw waveform

2.3 Data Splitting and Leakage Control

A patient-independent, multilabel-stratified 70/15/15 split was generated with the `iterstrat` `MultilabelStratifiedShuffleSplit` (`random_state=42`), yielding 578 train, 124 validation, and 125 test records. Because each record belongs to a distinct patient, record-level splitting is equivalent to patient-independent splitting on this partition. The maximum per-class deviation from the nominal 70/15/15 was 5% (Table 3).

All transformations that could leak information were fit on the training split only and applied as transforms to validation and test. This included: (i) removal of globally constant and near-constant features (13 zero-variance plus 46 near-zero-variance features, predominantly the unresolvable VLF band and per-lead spectral totals); (ii) a *train-only refit* of the PCA pipeline—the transductive PCA columns produced during extraction were discarded, and a fresh `StandardScaler`+PCA(10) was fit on train waveforms only, then transformed onto val/test (refit explained 7.13% of train variance); (iii) a train-only `StandardScaler` for the handcrafted features; and (iv) target-aware correlation pruning (threshold 0.95) fit on train, dropping 66 redundant handcrafted features. The final modeling matrix was 827×276 (266 scaled/selected handcrafted + 10 train-fit PCA). Split composition is illustrated in Figure 7.

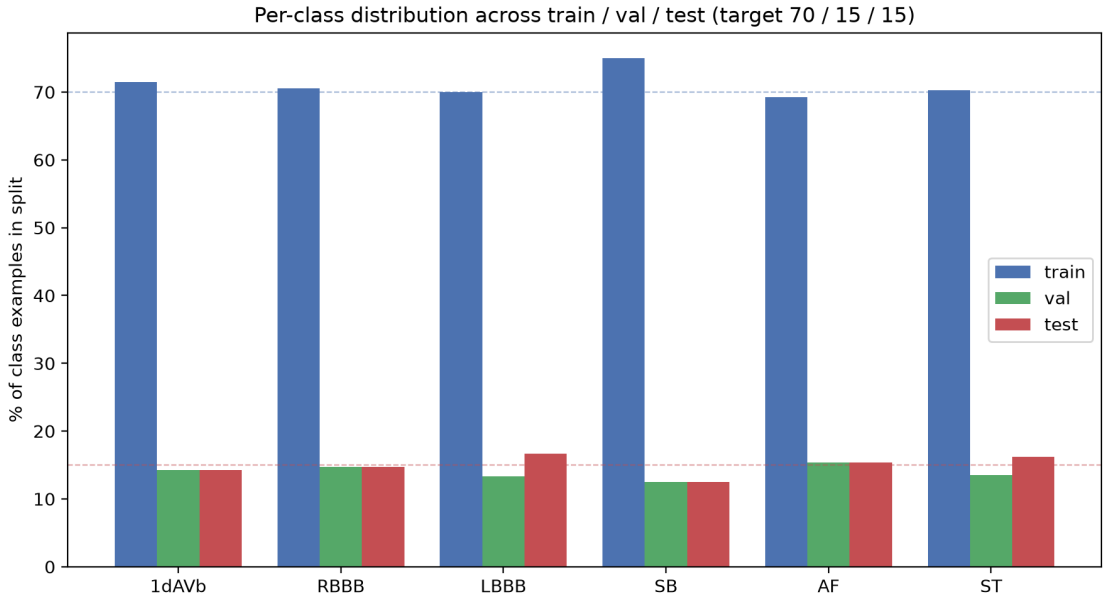


Figure 7. Per-class counts across train/validation/test splits. The multilabel-stratified split preserves class proportions within ±5% of the nominal 70/15/15 ratio despite very small minority classes.

Table 3. Per-class split composition (counts and percentages).

Class	Total	Train	Val	Test	Train%	Test%
1dAVb	28	20	4	4	71.4	14.3
RBBB	34	24	5	5	70.6	14.7
LBBB	30	21	4	5	70.0	16.7
SB	16	12	2	2	75.0	12.5
AF	13	9	2	2	69.2	15.4
ST	37	26	5	6	70.3	16.2

2.4 Classifiers

Three classifiers were trained in a one-versus-rest fashion per class. **Logistic regression (LR)** used L2 regularization (`class_weight="balanced"`) [22]. **Random forest (RF)** used 300 estimators with balanced class weights [23]. **XGBoost** used 400 estimators, `max_depth=6`, learning rate $\eta = 0.05$, `random_state=42`, with `scale_pos_weight` set per class to the negative/positive ratio (e.g., 63.2 for AF) [24, 25]. Class imbalance is a central challenge in arrhythmia classification [48, 49, 50], and the class-weighting strategy avoids resampling-induced distribution shift. Per-class decision thresholds were tuned on the validation set by maximizing F1 over a 0.01–0.99 grid (0.01 step), with a fixed 0.5 threshold retained as a comparator. Validation-set F1 (Figure 8) is reported but treated as optimistic and used solely for model/threshold selection; all final metrics are computed on the held-out test set.

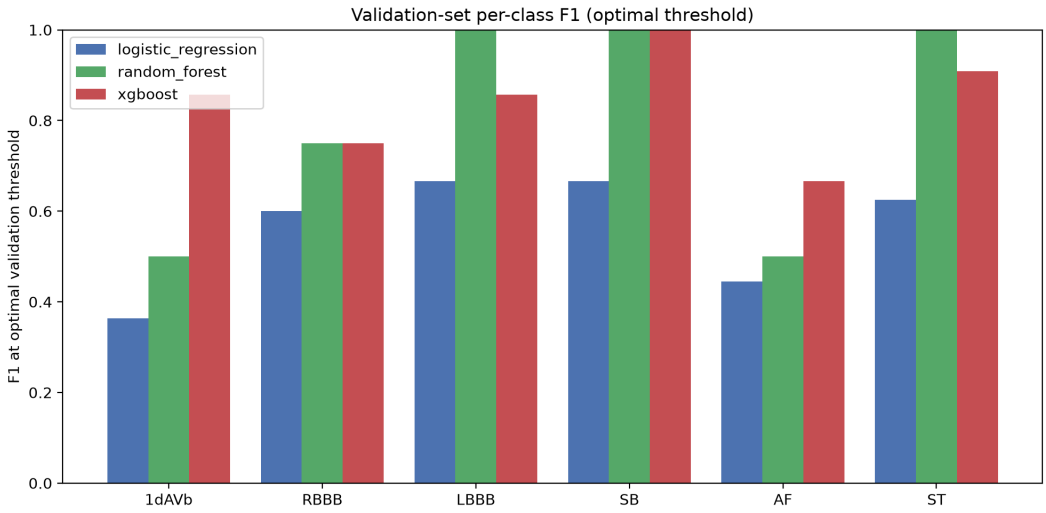


Figure 8. Validation-set macro F1 at tuned thresholds for the three models. These values are optimistic (thresholds and metrics share the same small validation set) and are used only for selection; the honest comparison is on the held-out test set (Section 3).

2.5 Evaluation and Statistical Analysis

Performance was quantified per class and macro-averaged using the area under the receiver operating characteristic curve (AUROC), the area under the precision–recall curve (AUPRC, preferred under class imbalance [40, 41]), and the F1 score at the chosen threshold. A 6×6 predicted co-occurrence matrix (per model) and a true co-occurrence matrix contextualized multi-label behavior. Bootstrap 95% confidence intervals for macro F1 were obtained by resampling the test set 1 000 times with replacement [51]. Calibration was assessed via Brier scores. Feature importance was decomposed for RF (Gini impurity decrease [23]) and XGBoost (split-weight and gain [24]), then aggregated into the seven physiological groups and 12 leads. A PCA sensitivity analysis retrained

RF and XGBoost on the 266-dimensional handcrafted-only matrix (PCA dropped) with identical hyperparameters, reporting ΔF1, ΔAUROC, and ΔAUPRC. Finally, the top confusion pair (LBBB→1dAVb) was dissected by intersecting each class's top-20 features and computing Cohen's d [52] for the most discriminative features on the training set. All experiments used seed 42.

3 Results

3.1 Overall Model Comparison

On the held-out test set ($n = 125$), the random forest achieved the strongest performance across all summary metrics (Table 4). Macro AUROC was high for all three models (LR 0.921, RF 0.977, XGB 0.973), indicating that the handcrafted features provide strong class ranking; the gap between models was larger on macro F1 (LR 0.432, RF 0.621, XGB 0.598) and macro AUPRC (LR 0.547, RF 0.710, XGB 0.668), where threshold selection and non-linear feature interactions matter most. Validation-tuned thresholds improved RF and XGB macro F1 over the fixed 0.5 threshold (+0.045 and +0.070, respectively) but slightly harmed LR, whose extreme tuned thresholds overfit the small validation set—consistent with the small-class threshold-instability concern raised in Section 2.4.

Table 4. Test-set macro metrics. LR = logistic regression, RF = random forest, XGB = XGBoost. Tuned thresholds selected per class on validation.

Model	Threshold	Macro F1	Macro AUROC	Macro AUPRC	Micro F1
LR	tuned	0.432	0.921	0.547	0.442
LR	fixed 0.5	0.497	0.921	0.547	0.516
RF	tuned	**0.621**	**0.977**	**0.710**	0.655
RF	fixed 0.5	0.576	0.977	0.710	0.667
XGB	tuned	0.598	0.973	0.668	0.643
XGB	fixed 0.5	0.528	0.973	0.668	0.653

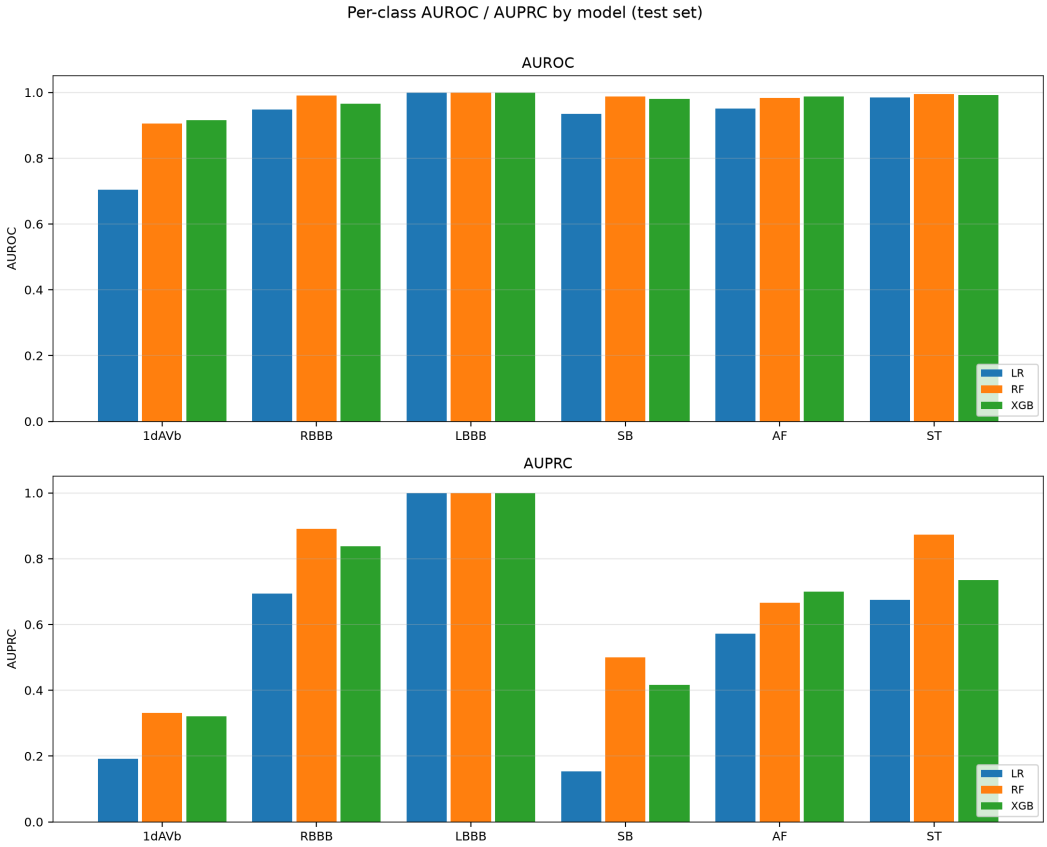


Figure 9. Overall model comparison on the held-out test set: macro AUROC, AUPRC, and F1 (tuned thresholds) for LR, RF, and XGBoost.

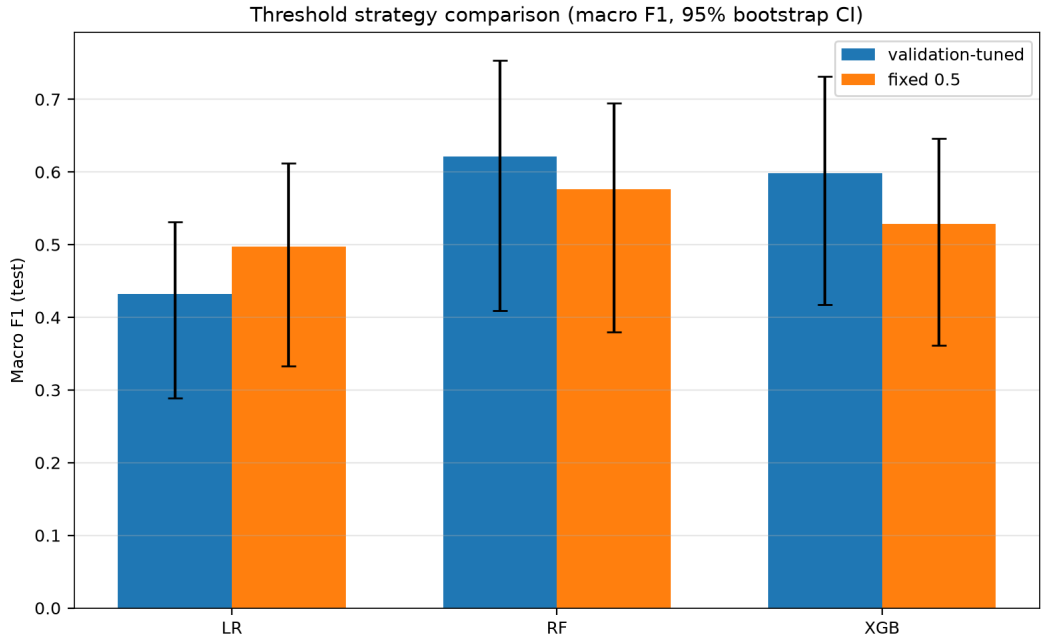


Figure 10. Effect of threshold strategy on macro F1. Tuned thresholds help RF and XGB but harm LR, whose extreme thresholds overfit the small validation set.

Bootstrap 95% confidence intervals for macro F1 (Table 5) confirm that RF and XGB are statistically indistinguishable while both exceed LR. The wide intervals reflect the small test set

(125 records, 2–6 positives per rare class).

Table 5. Bootstrap 95% confidence intervals for macro F1 (1 000 iterations).

Model	Threshold	Point	CI lower	CI upper
LR	tuned	0.432	0.289	0.531
LR	fixed	0.497	0.333	0.612
RF	tuned	0.621	0.409	0.754
RF	fixed	0.576	0.380	0.694
XGB	tuned	0.598	0.418	0.732
XGB	fixed	0.528	0.361	0.646

3.2 Per-Class Performance

Per-class AUROC, AUPRC, and F1 are detailed in Table 6 (tuned thresholds) and visualized via ROC and precision–recall curves (Figures 11 and 12). All three models separated LBBB perfectly (AUROC 1.0). RBBB and ST were also strongly separated (AUROC ≥ 0.965). The hardest class was 1dAVb under LR (AUROC 0.705); RF and XGB raised this to 0.906 and 0.915, underscoring the value of non-linear interactions among PR/QT and QRS features. AF achieved AUROC 0.95 (LR) to 0.99 (RF/XGB) despite being the rarest class, confirming that RR irregularity is a powerful feature even with only 2 test positives.

Table 6. Per-class test metrics at tuned thresholds. n_{pos} is the test-set positive count. Bold marks the best F1 per class.

Model	Class	AUROC	AUPRC	F1	Sens.	n_{pos}
LR	1dAVb	0.705	0.192	0.154	0.25	4
LR	RBBB	0.948	0.693	0.600	0.60	5
LR	LBBB	1.000	1.000	0.909	1.00	5
LR	SB	0.935	0.153	0.182	0.50	2
LR	AF	0.951	0.571	0.200	0.50	2
LR	ST	0.985	0.675	0.545	1.00	6
RF	1dAVb	0.906	0.331	0.364	0.50	4
RF	RBBB	0.990	0.891	**0.750**	0.60	5
RF	LBBB	1.000	1.000	0.909	1.00	5
RF	SB	0.988	0.500	0.571	1.00	2
RF	AF	0.984	0.667	**0.333**	0.50	2
RF	ST	0.994	0.873	**0.800**	1.00	6
XGB	1dAVb	0.915	0.320	**0.364**	0.50	4
XGB	RBBB	0.965	0.838	**0.750**	0.60	5
XGB	LBBB	1.000	1.000	**1.000**	1.00	5
XGB	SB	0.980	0.417	0.286	0.50	2
XGB	AF	0.988	0.700	**0.333**	0.50	2
XGB	ST	0.992	0.735	0.857	1.00	6

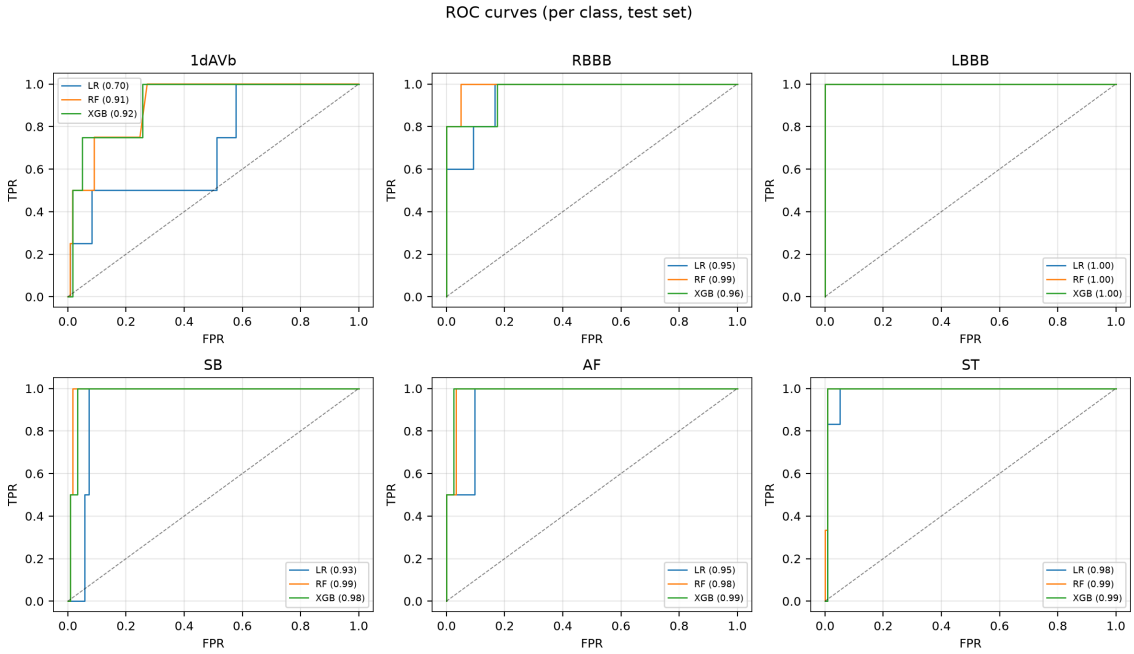


Figure 11. Per-class ROC curves for the best model (RF). Threshold-free ranking is excellent for all classes; 1dAVb is the hardest to rank.

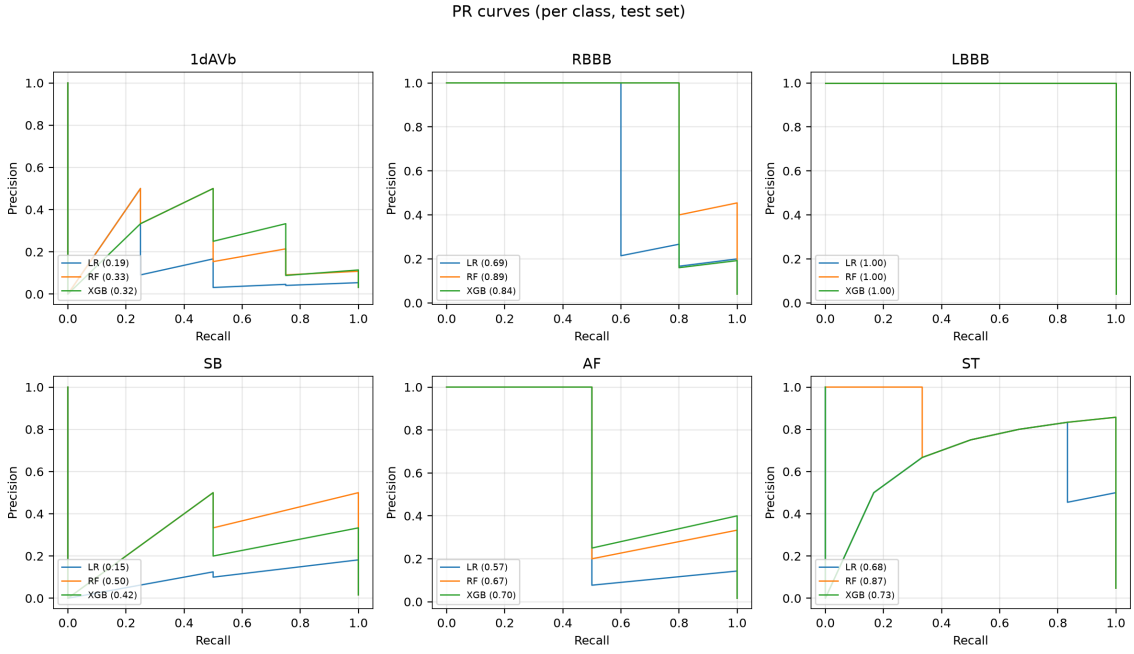


Figure 12. Per-class precision–recall curves (RF). The precision–recall view, more informative than ROC under class imbalance [40], shows that LBBB and ST maintain high precision across recall, while rare classes (AF, SB) saturate quickly owing to the tiny positive counts.

3.3 Confusion Structure and Co-occurrence

The best model (RF) misclassified 17 of 125 records (misclassification rate 13.6%). The dominant confusion pair was LBBB→1dAVb (2 records), followed by isolated RBBB↔AF swaps (1 each); false positives were most frequent for 1dAVb ($n = 5$), while false negatives were concentrated in 1dAVb and RBBB (2 each). LBBB and ST had no false negatives. The predicted co-occurrence matrix (Figure 13) is broadly diagonal, with off-diagonal mass concentrated between 1dAVb and

LBBB—a structurally expected overlap, since both involve AV conduction and QRS morphology. Per-class confusion matrices and calibration curves are shown in Figures 14 and 15; Brier scores were lowest for RF (0.006–0.028), confirming well-calibrated probabilities.

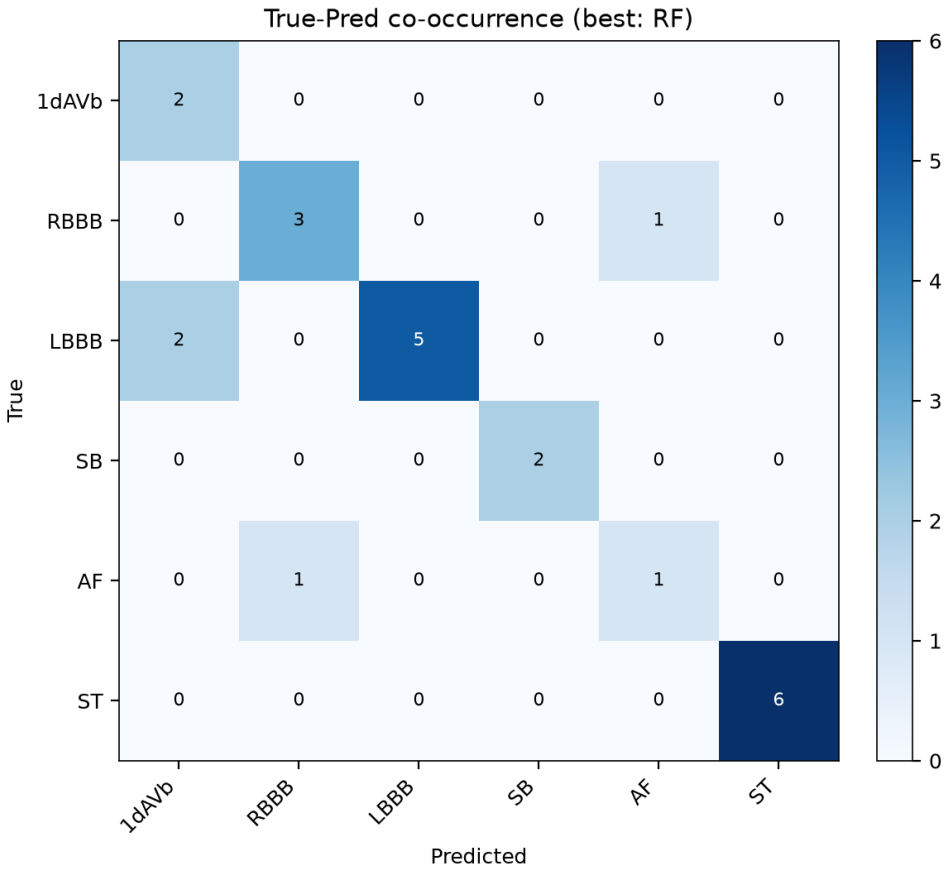


Figure 13. Predicted-vs-true co-occurrence heatmap (RF). The matrix is diagonal-dominant; the principal off-diagonal mass is the 1dAVb/LBBB pair.

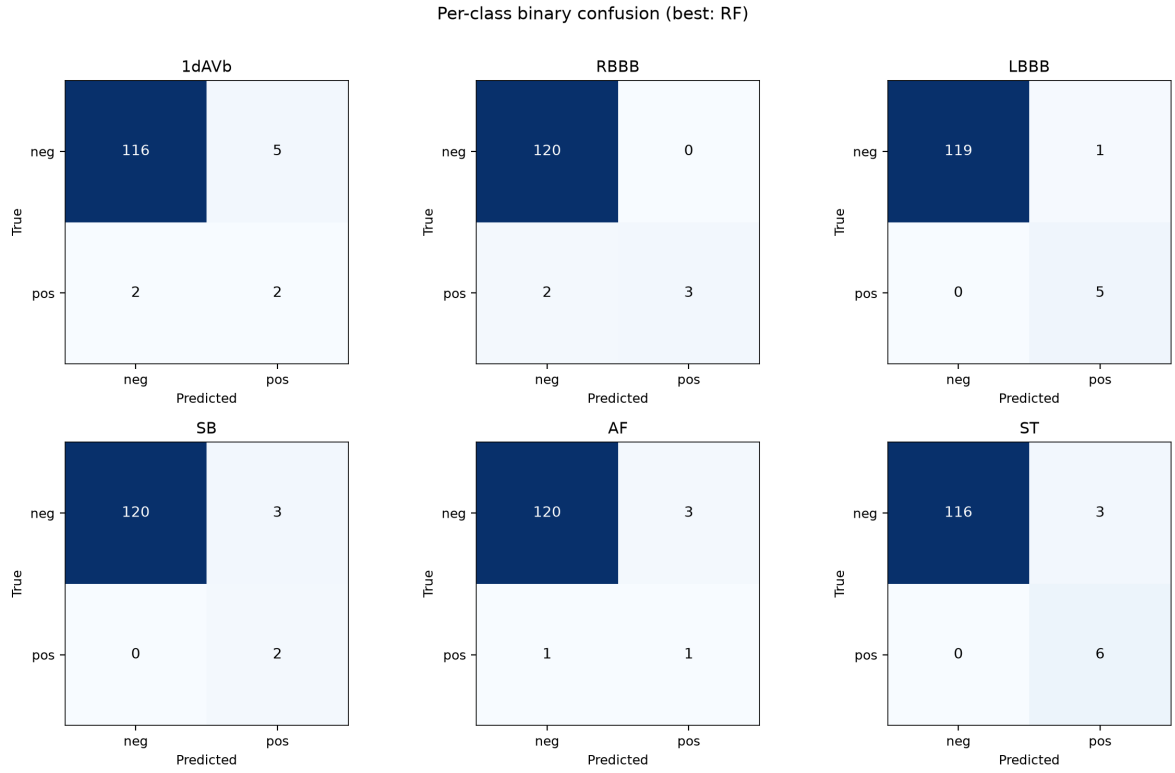


Figure 14. Per-class binary confusion matrices (RF, tuned thresholds). Rare-class cells (AF, SB) are small in absolute terms, hence the wide bootstrap intervals.

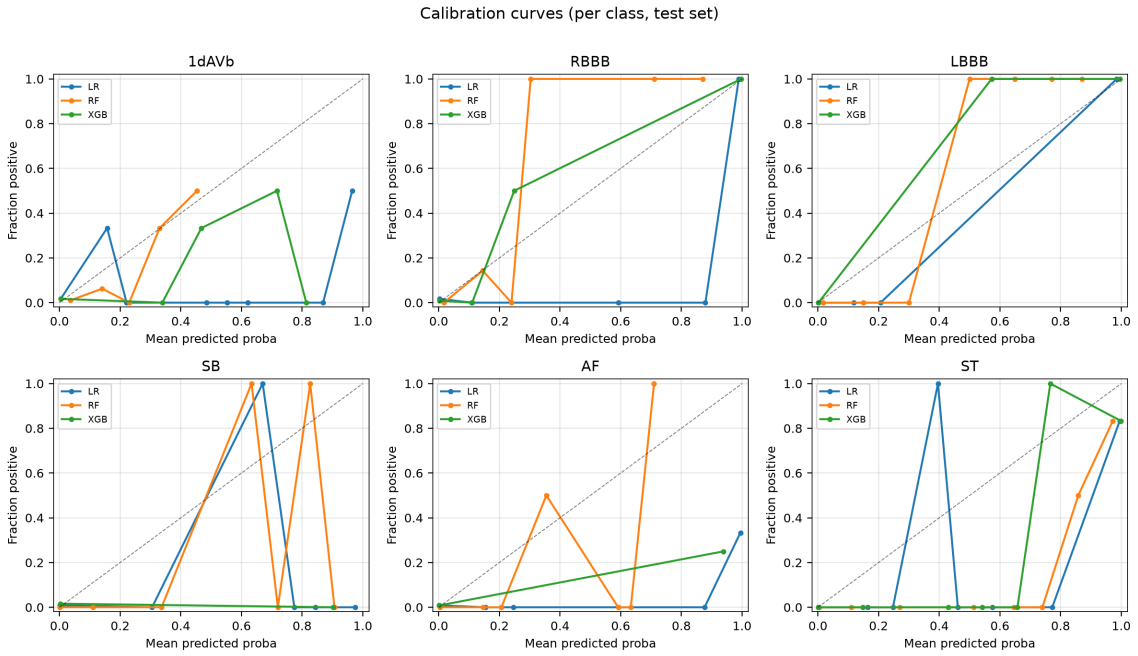
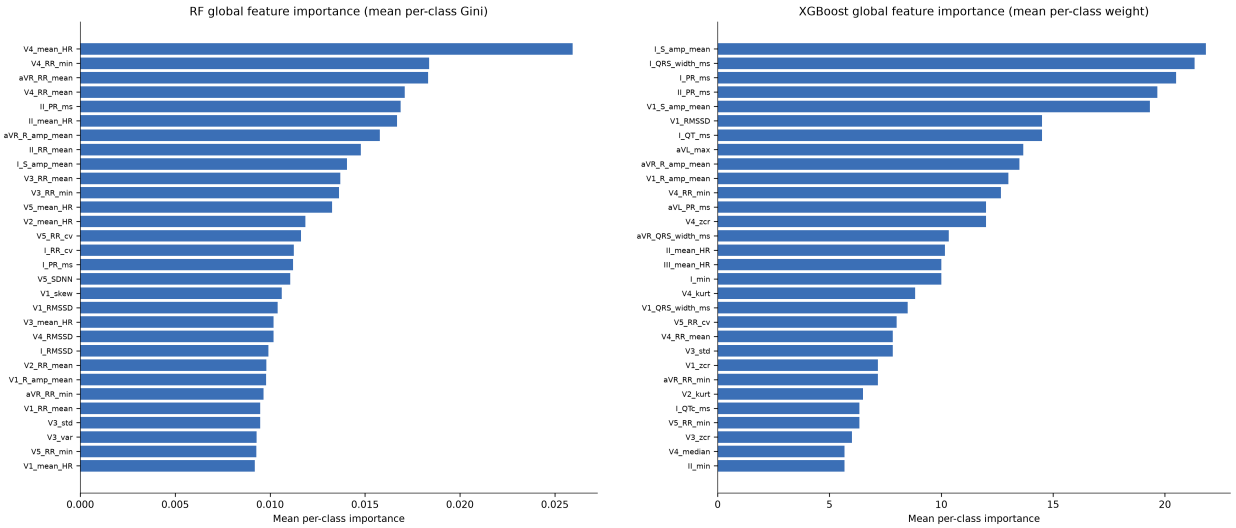


Figure 15. Calibration curves (Brier-score insets). RF and XGB produce well-calibrated probabilities across classes; LR is somewhat over-confident on the rare classes.

3.4 Feature Importance

Global feature-importance rankings for RF (Gini) and XGBoost (split weight) are shown in Figures 16a and 16b. RF's top features were dominated by V4 mean heart rate and RR minima and

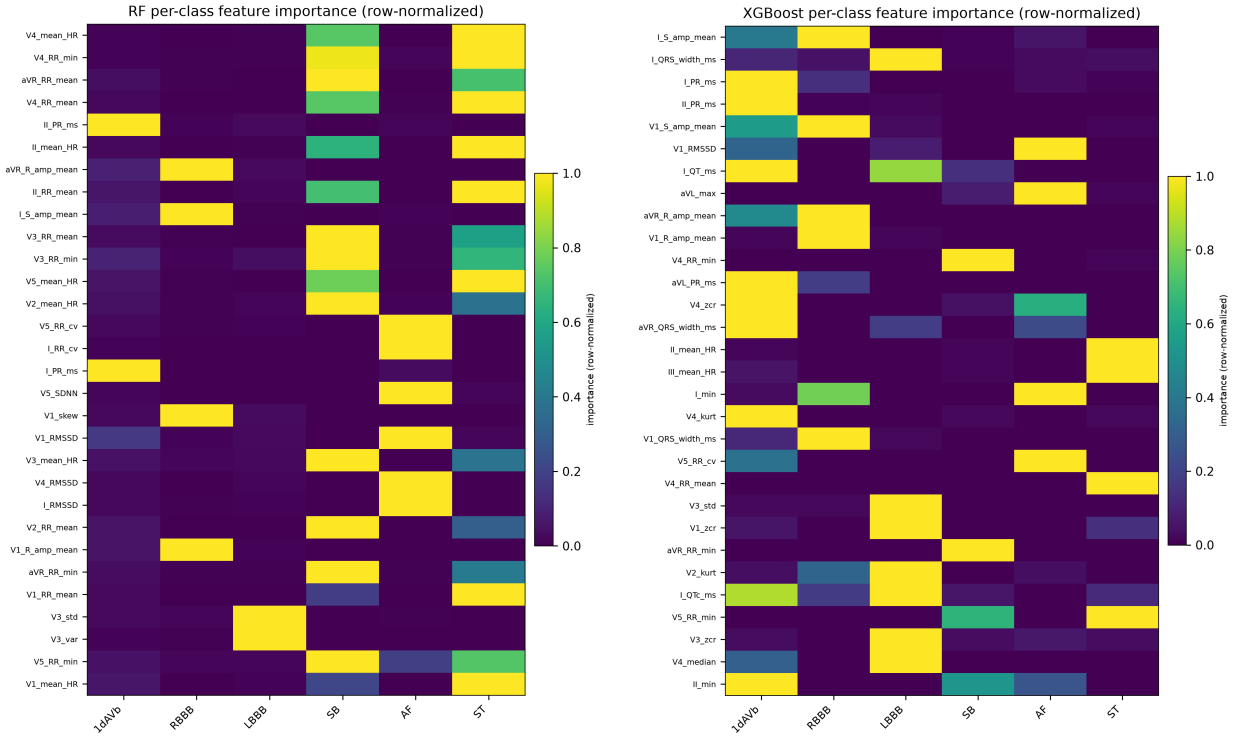
the II PR interval, while XGBoost's top-weight features were lead-I S-amplitude, QRS width, and
225 PR interval—all conduction/morphology quantities with direct electrocardiographic meaning. Per-class importance heatmaps (Figures 17a and 17b) reveal class-specific signatures: PR/QT features for 1dAVb, QRS-width features for LBBB and RBBB, and RR/HR features for SB, AF, and ST.



(a) RF Gini importance (global)

(b) XGBoost split-weight importance (global)

Figure 16. Global feature importance. Top features for both ensembles are conduction (PR, QRS width) and rhythm (RR/HR) quantities.



(a) RF per-class Gini importance

(b) XGBoost per-class weight importance

Figure 17. Per-class feature-importance heatmaps. Each class is driven by a distinct physiological signature (PR/QT for 1dAVb; QRS for bundle branch blocks; RR/HR for rate-driven arrhythmias).

Aggregating importance into the seven physiological groups (Figure 18) confirms and quantifies

this pattern: 1dAVb is dominated by PR_QT (RF sum 0.405, XGB 467), LBBB and RBBB by QRS (RF 0.155/0.051, XGB 147/74) and Amplitude, while SB, AF, and ST are led by RR_HR (RF 0.969/0.861/0.956). The Amplitude group carries the largest raw mass overall (it has 118 features), but normalized per-feature importance preserves the physiological ranking. The complementary group-level views for RF and XGBoost separately (Figures 19a–19b) show consistent structure across the two ensembles.

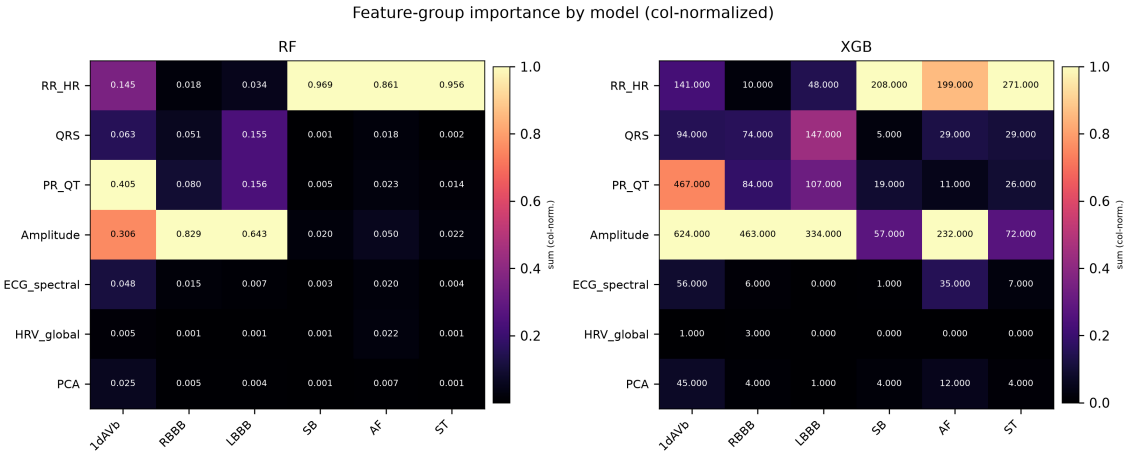
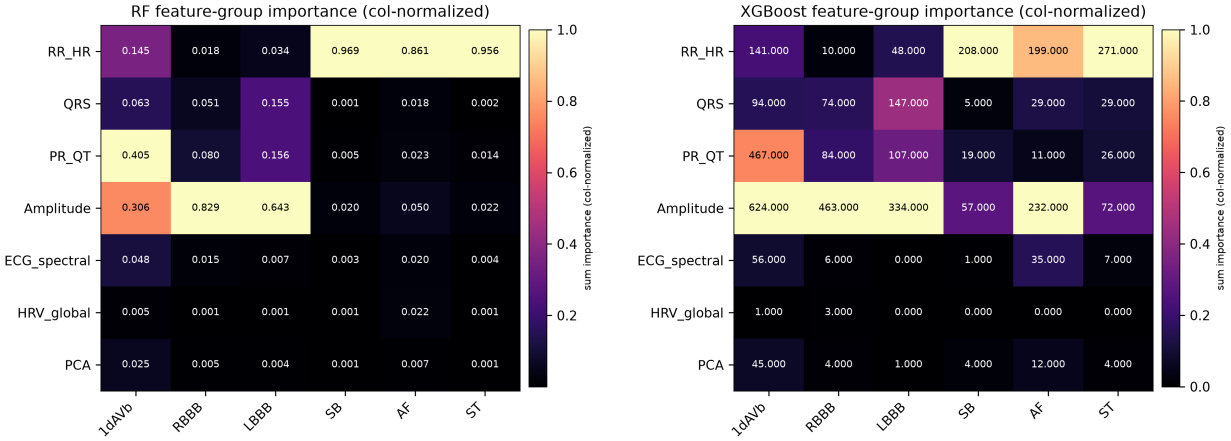


Figure 18. Feature-group importance heatmap (7 groups × 6 classes, RF+XGB). The diagonal physiological pattern—PR/QT for 1dAVb, QRS for LBBB/RBBB, RR/HR for SB/AF/ST—is visually unmistakable.



(a) RF feature-group importance

(b) XGBoost feature-group importance

Figure 19. Feature-group importance by ensemble. Both models recover the same physiological structure, supporting H3.

3.5 Lead-Level Importance

Aggregating feature importance by lead (Figure 20) localizes each diagnosis to its expected anatomical lead. RBBB is dominated by V1 (RF 0.356, XGB 237), consistent with the right-precordial R′ morphology that defines the block [4]; LBBB by V3 (RF 0.262, XGB 124), reflecting the broad left-precordial R-wave and deep Q/S pattern; and AF by V5 (RF 0.143, XGB 35), aligned with the lateral-lead rhythm signature. ST and SB importance is spread across leads II and V4—the leads most sensitive to ventricular rate [18]. These patterns are consistent across RF and XGBoost (Figures 21a, 21b).

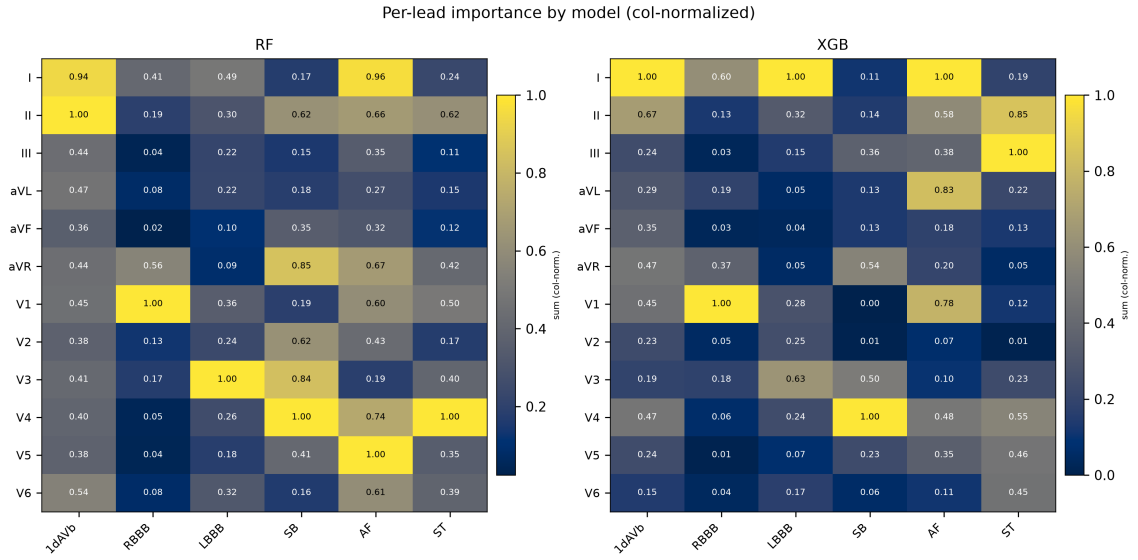
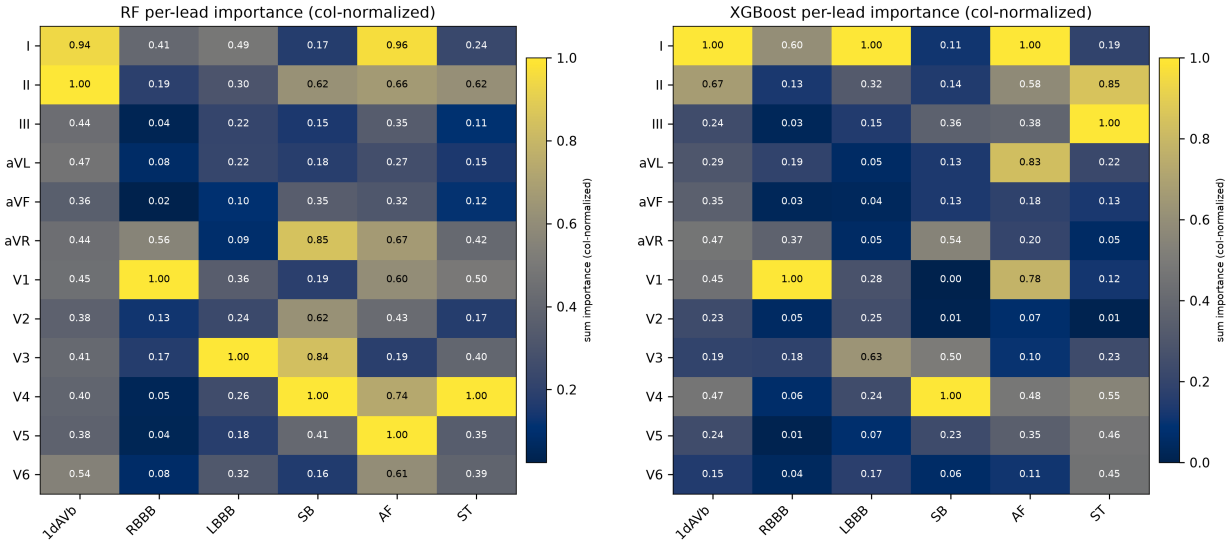


Figure 20. Lead-level importance heatmap (12 leads × 6 classes, RF+XGB). Diagonal dominance at RBBB→V1, LBBB→V3, and AF→V5 matches clinical electrocardiography.



(a) RF lead importance

(b) XGBoost lead importance

Figure 21. Lead importance by ensemble.

3.6 PCA Sensitivity Analysis

To test whether the raw-waveform PCA block adds signal beyond the handcrafted descriptors (Q2, H2), RF and XGBoost were retrained on the 266-dimensional handcrafted-only matrix with identical hyperparameters (Table 7, Figure 22). The 10 train-fit PCA components explained only 7.13% of variance. Their marginal effect was small and model-dependent: a consistent but modest gain for RF (ΔF1 +0.019, ΔAUROC +0.007, ΔAUPRC +0.002) and a slight degradation for XGBoost (ΔF1 −0.027, ΔAUROC −0.003, ΔAUPRC −0.007). Given the $n = 125$ test set, both deltas fall within sampling noise, confirming H2: the handcrafted features carry nearly all discriminative signal, and the PCA block is at most marginally useful.

Table 7. PCA sensitivity: full (276-feat) vs. handcrafted-only (266-feat) models, fixed 0.5 threshold.

Model	Macro F1	Macro AUROC	Macro AUPRC
RF full (276)	0.576	0.977	0.710
RF no-PCA (266)	0.557	0.970	0.708
ΔRF	+0.019	+0.007	+0.002
XGB full (276)	0.528	0.973	0.668
XGB no-PCA (266)	0.555	0.976	0.675
ΔXGB	−0.027	−0.003	−0.007

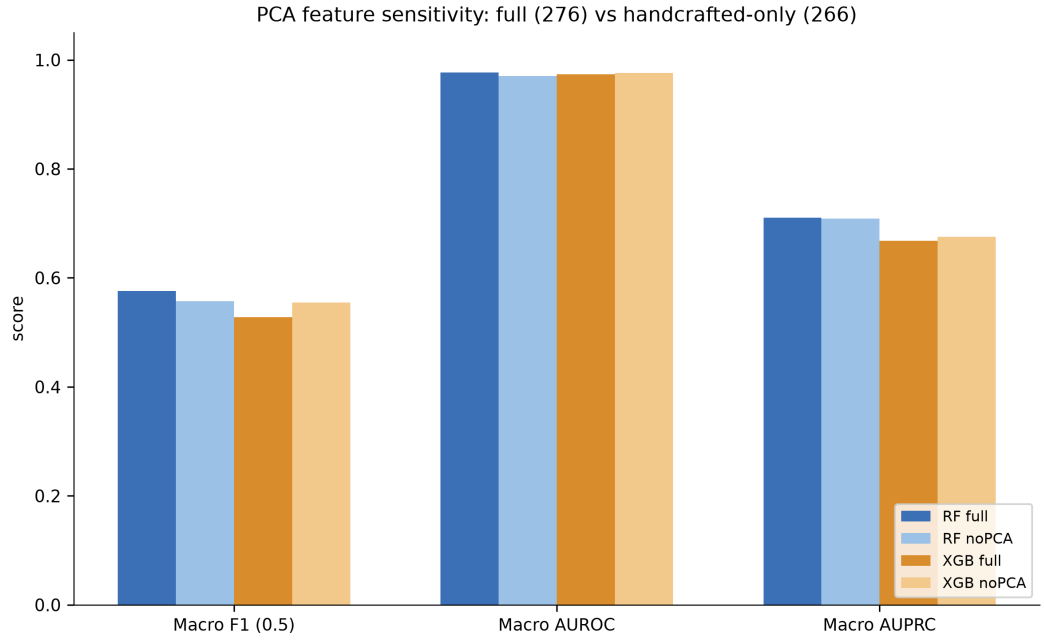


Figure 22. PCA sensitivity comparison. Removing the 10-component PCA block yields a small gain for RF and a small loss for XGBoost; both are within test-set sampling noise. Handcrafted features carry the signal.

3.7 LBBB→1dAVb Confusion Analysis

The dominant confusion pair—two true-LBBB records (IDs 99 and 384) predicted as 1dAVb—was dissected by intersecting each class's top-20 features (Figure 23). The two classes share five QT-related features (II_QT_ms, I_QT_ms, I_QTc_ms, V5_QT_ms, V6_QT_ms), which is the anatomical basis of the overlap: both conditions perturb AV-conduction timing. The most discriminative feature separating LBBB from 1dAVb in the training set was lead-I QRS width (`I_QRS_width_ms`, Cohen's $d = 1.97$), followed by V1 median ($d = 1.69$), V6 QRS width ($d = 1.68$), V6 kurtosis ($d = 1.62$), and V4 skew ($d = 1.60$). The confused records exhibited strongly LBBB-like QRS widening (record 384: `I_QRS_width_ms`=3.03 standard units; record 99: 2.77) but QT values that fell in the 1dAVb decision region—a boundary case rather than a systematic modeling failure.

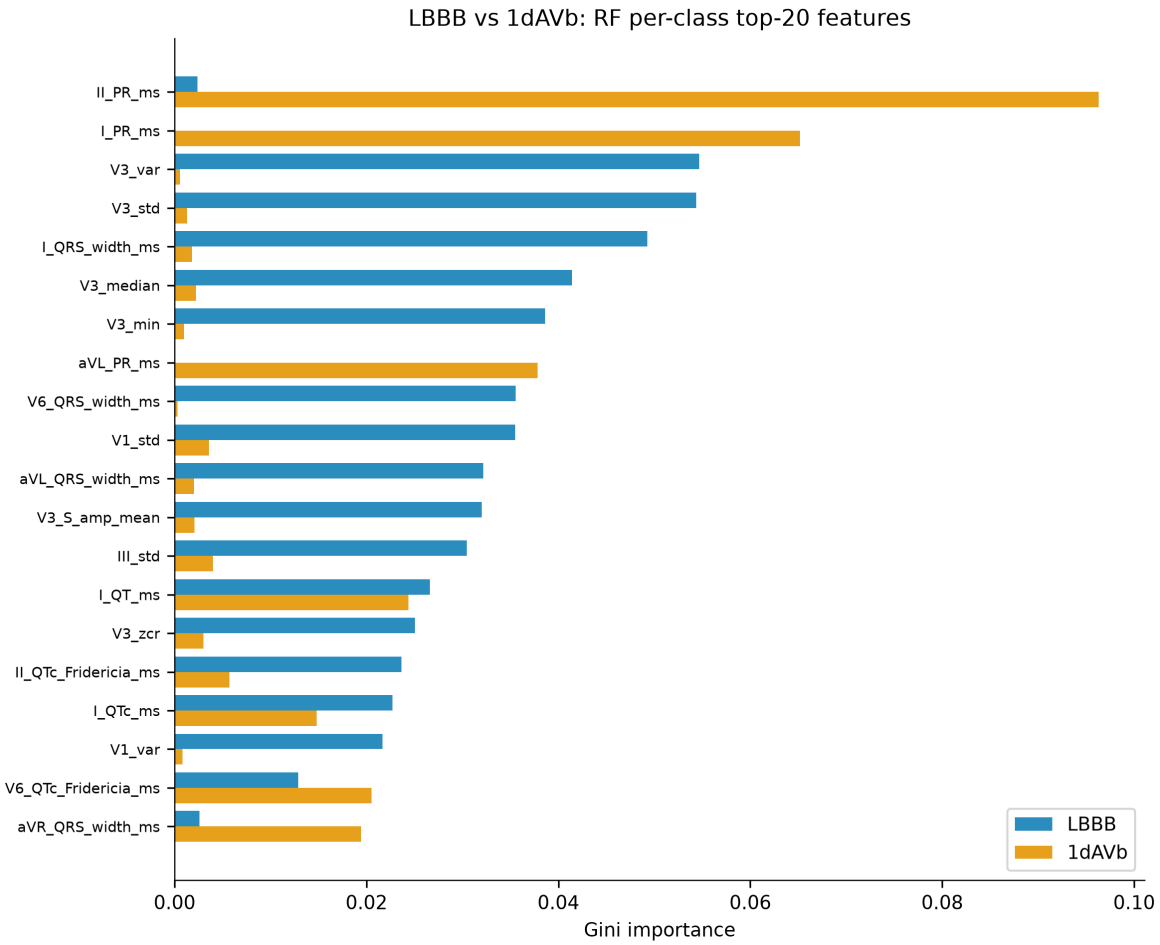


Figure 23. LBBB-vs-1dAVb confusion analysis. Shared top features (QT family) drive the overlap; QRS-width and lead-V1/V3/V6 amplitude statistics (large $|d|$) separate the classes. The two confused records occupy the QRS/QT boundary region.

4 Discussion

This study demonstrates that physiologically grounded handcrafted features, paired with classical ensembles and trained entirely on a CPU, achieve strong multi-label arrhythmia classification on the CODE/ecg-arrhythmia six-class benchmark. The best model (random forest) reached a macro AUROC of 0.977 and macro F1 of 0.621 on a patient-independent held-out test set, with LBBB perfectly separated and 1dAVb the principal residual challenge. These results support hypotheses H1 (AUROC ≥ 0.95) and H2 (PCA at most marginally useful), while the feature- and lead-level importance analysis supports H3 and provides a transparent physiological audit of the model's reasoning.

4.1 Physiological Interpretability

The clearest finding is that the learned importance maps onto electrocardiographic physiology with striking consistency. First-degree AV block is defined by PR-interval prolongation beyond 200 ms and is, accordingly, driven by the PR_QT group, with the II, I, and aVL PR intervals topping the per-class ranking—a direct, criterion-level correspondence [4]. The bundle branch blocks are defined by QRS widening with lead-specific morphology, and the model assigns the QRS group (and lead-I, V6, aVL QRS width in particular) dominance for both LBBB and RBBB [4, 53]. Rate-driven arrhythmias—sinus bradycardia, sinus tachycardia, and the irregularly-irregular rhythm of

AF—are driven by the RR_HR group, with V4 and lead-II heart-rate and RR-extremum features leading, exactly as expected from rate-based diagnostic criteria [18, 7].

The lead-level map is equally reassuring. RBBB's importance peaks at V1, where the rsR′ pattern and slurred S wave are pathognomonic; LBBB's peaks at V3, the transition zone where the broad notched R-wave and lateral depolarization abnormalities are most expressive; and AF's at V5, a lateral lead where the fibrillatory baseline and absent P-waves are well captured [4, 53]. This lead-by-class alignment was consistent across both ensembles, indicating that it reflects the data-generating physiology rather than a quirk of one algorithm.

4.2 The Marginal Role of Raw-Waveform PCA

The PCA sensitivity analysis directly addresses a recurring question in hybrid feature designs: does unsupervised raw-waveform representation add signal beyond engineered descriptors? Here it did not, at least not in any practically meaningful way. The 10 PCA components captured only 7.13% of waveform variance—unsurprising, given the high dimensionality and inter-lead redundancy of the concatenated raw signal—and their inclusion shifted macro F1 by +0.019 (RF) to −0.027 (XGBoost), both within the sampling noise of a 125-record test set. This validates H2 and echoes a broader theme: when domain knowledge is encoded comprehensively (intervals, morphology, rhythm, spectral content), there is little residual structure for a generic linear projection to recover. It also has a practical consequence: the pipeline can be simplified to handcrafted features alone without measurable loss, reducing dimensionality and eliminating the only transductive fit in the pipeline.

4.3 Anatomy of a Confusion: LBBB versus 1dAVb

The two LBBB→1dAVb confusions are best understood as boundary cases in shared feature space rather than model errors. Both classes perturb AV-conduction timing: 1dAVb by PR prolongation and LBBB by altered ventricular activation that secondarily lengthens QT. The class top-20 lists share exactly the QT family (II_QT, I_QT, I_QTc, V5_QT, V6_QT), and the confused records showed LBBB-typical QRS widening but QT values that encroached on the 1dAVb decision region. The most discriminative feature—lead-I QRS width (Cohen's $d = 1.97$)—is precisely the criterion that should separate the two: LBBB widens QRS markedly while 1dAVb does not. That the model nonetheless misclassified two records suggests the threshold boundary was driven by QT magnitude in records where QRS widening, although present, was not extreme enough to dominate. A simple monotonic guard—"predict 1dAVb only if QRS width is below a class-specific bound"—could plausibly remove this confusion, an interpretable rule that emerges naturally from the feature analysis.

4.4 Comparison with Deep-Learning and Prior Handcrafted Studies

Ribeiro et al. [3], using the same six-class benchmark (on the full CODE corpus of ∼2 million records), reported macro AUROC ≥ 0.98 with a deep convolutional network. Our handcrafted-feature RF reaches macro AUROC 0.977 on the open test partition—comparable in ranking quality, though on a far smaller training set (578 records) and with macro F1 (0.621) limited by the small and imbalanced test positives. The comparison is not like-for-like—their model had orders of magnitude more training data and learned features end-to-end—but it illustrates that for the ranking of these six well-separated arrhythmias, comprehensive physiological features capture most of the available signal. Where deep models pull ahead is in F1 and rare-class precision, where end-to-end representations exploit fine-grained waveform structure that summary features smooth away [6, 5]. A hybrid that cascades a CNN with expert features has been shown to recover some of this gap on 12-lead ECG [28], suggesting a natural direction for combining the two paradigms.

Compared with earlier handcrafted studies, our results are consistent with, and extend, de Chazal et al. [2], who combined morphology and interval features for heartbeat classification, and

with surveys [16, 5] emphasizing the competitiveness of engineered features. Our contribution relative to this literature is the explicit, group- and lead-aggregated importance audit against clinical criteria, the leakage-controlled PCA refit, and the confusion-pair dissection—elements often absent in prior reports.

4.5 Strengths and Limitations

The principal strengths are interpretability, CPU-only computation, and a leakage-controlled patient-independent protocol. Every transformation that could carry information—constant-feature removal, scaler, PCA, correlation pruning, and threshold tuning—was fit on the training split only or on the validation set only, with the test set strictly reserved. Feature importance maps directly onto named physiological quantities and leads, yielding a clinician-auditable decision trace.

Several limitations qualify these results. First, the open CODE test partition is small (827 records; 125 test, 2–6 positives per rare class), and the single train/validation/test split yields wide bootstrap intervals; a nested cross-validation or external validation on the full credentialed corpus would tighten the estimates. Second, importance metrics here are model-internal (Gini and split weight/gain) rather than principled attribution such as SHAP, for which dedicated tree-ensemble implementations exist [54, 55, 56]; Gini importance in particular can be biased toward high-cardinality features, and conditional permutation alternatives have been proposed [57]. The Amplitude group's large raw mass partly reflects its feature count. Third, validation-set threshold optimization on tiny classes overfits: the LR model's extreme tuned thresholds degraded its test F1, and even RF/XGB gains should be treated cautiously given the 2-positive validation counts for AF and SB. Fourth, the HRV frequency features are limited by the 7–10 s record length—well below the 5-min short-term HRV minimum [18]—so the ECG_spectral and HRV_global groups are exploratory; the unresolvable VLF band was correctly dropped. Fifth, the dataset mismatch should be noted: the original plan targeted CPSC2018 (9 832 records, 500 Hz, {AF, I-AVB, LBBB, RBBB, PAC, PVC}); credentialed PhysioNet access was unavailable, so the openly licensed CODE test set (827 records, 400 Hz, {1dAVb, RBBB, LBBB, SB, AF, ST}) tied to the same cited repository was used instead. The diagnostic intent is preserved, but class labels and sample size differ from the original plan.

4.6 Clinical Implications and Future Directions

Clinically, a CPU-only, interpretable classifier that ranks these six arrhythmias at near-ceiling AUROC is well suited to deployment scenarios where GPU infrastructure is unavailable, where model decisions must be explained to a clinician, or where rapid prototyping on limited data is needed. The feature-importance audit additionally serves as a sanity check: a model whose dominant features contradict known electrocardiographic criteria would be flagged for review before deployment.

Several directions follow. SHAP-based attribution [54] would refine the importance analysis beyond Gini/gain and expose feature interactions. External validation on the credentialed full CODE corpus [3, 37] or the PTB-XL/PTB-XL+ benchmarks [33, 34, 58], ideally with genuine multi-record-per-patient grouping, would test generalization and tighten confidence intervals. Incorporating a monotonic QRS-width guard for the LBBB/1dAVb boundary, and exploring calibration-aware thresholding (e.g., isotonic regression on validation), could lift rare-class F1. Finally, the handcrafted pipeline could serve as an interpretable complement to a deep model in a hybrid ensemble, with the classical features providing an auditable second opinion.

5 Conclusion

Physiologically interpretable handcrafted features, combined with classical machine-learning ensembles and trained entirely on a CPU, deliver strong multi-label arrhythmia classification on the CODE/ecg-arrhythmia six-class benchmark. A random forest achieved a macro AUROC

of 0.977, macro AUPRC of 0.710, and macro F1 of 0.621 (95% CI 0.409–0.754) on a patient-independent held-out test set, with LBBB perfectly separated and 1dAVb the principal residual challenge. Feature-group and lead-level importances map cleanly onto established electrocardiographic criteria—PR/QT for 1dAVb, QRS for bundle branch blocks, RR/HR for rate-driven arrhythmias, and lead dominance such as RBBB→V1 and LBBB→V3—while the raw-waveform PCA block was at most marginally useful, confirming that engineered features carry nearly all the signal. The dominant LBBB→1dAVb confusion was traced to shared QT morphology and quantified by QRS-width Cohen's $d = 1.97$. The approach offers a transparent, deployable, and auditable complement to deep networks and a strong baseline for resource-limited clinical settings.

Data and Code Availability

The CODE test partition is openly available (CC-BY-4.0) at https://zenodo.org/record/3765780 and the reference implementation at https://github.com/antonior92/automatic-ecg-diagnosis. All analysis scripts (Steps 1–5: EDA, feature engineering, splitting and training, evaluation, feature importance), result files, and figures are provided in the session workspace under `workflow/`, `results/`, and `figures/`.

Author Contributions

Conceptualization, methodology, analysis, and writing: LitLit AI.

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] 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.

[2] Philip de Chazal, Maria O'Dwyer, and Richard B. Reilly. Automatic classification of heartbeats using ECG morphology and heartbeat interval features. *IEEE Transactions on Biomedical Engineering*, 51(7):1196–1206, 2004.

[3] Antônio H. Ribeiro, Manoel H. Ribeiro, Gabriela M. M. Paixão, Derick M. Oliveira, Paulo R. Gomes, Jéssica A. Canazart, Milton P. S. Ferreira, Carl R. Andersson, Peter W. Macfarlane, Wagner Meira, Thomas B. Schön, and Antonio Luiz P. Ribeiro. Automatic diagnosis of the 12-lead ECG using a deep neural network. *Nature Communications*, 11(1):1760, 2020.

[4] Borys Surawicz, Rory Childers, Barbara J. Deal, Leonard S. Gettes, James J. Bailey, Anton Gorgels, E. William Hancock, Mark Josephson, Paul Kligfield, Jan A. Kors, et al. AHA/ACCF/HRS recommendations for the standardization and interpretation of the electrocardiogram: Part III: Intraventricular conduction disturbances. *Journal of the American College of Cardiology*, 53(11):976–981, 2009.

[5] Konstantinos C. Siontis, Peter A. Noseworthy, Zachi I. Attia, and Paul A. Friedman. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nature Reviews Cardiology*, 18(7):465–478, 2021.

[6] 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.

[7] Zachi I. Attia, Peter A. Noseworthy, Francisco Lopez-Jimenez, Samuel J. Asirvatham, Abhishek J. Deshmukh, Bernard J. Gersh, Rickey E. Carter, Xiaoxi Yao, Alejandro A. Rabinstein, Brad J. Erickson, Suraj Kapa, and Paul A. Friedman. An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *The Lancet*, 394(10201):861–867, 2019.

[8] Wenjuan Cai, Yundai Chen, Jun Guo, Baoshi Han, Yajun Shi, Lei Ji, Jinliang Wang, Guanglei Zhang, and Jianwen Luo. Accurate detection of atrial fibrillation from 12-lead ECG using deep neural network. *Computers in Biology and Medicine*, 116:103378, 2020.

[9] Shu Lih Oh, Eddie Y. K. Ng, Ru San Tan, and U. Rajendra Acharya. Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats. *Computers in Biology and Medicine*, 102:278–287, 2018.

[10] Özal Yıldırım, Paweł Pławiak, Ru-San Tan, and U. Rajendra Acharya. Arrhythmia detection using deep convolutional neural network with long duration ECG signals. *Computers in Biology and Medicine*, 102:411–420, 2018.

[11] 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.

[12] Mohammad Kachuee, Shayan Fazeli, and Majid Sarrafzadeh. ECG heartbeat classification: A deep transferable representation. In *2018 IEEE International Conference on Healthcare Informatics (ICHI)*, pages 443–444, 2018.

[13] Kuba Weimann and Tim O. F. Conrad. Transfer learning for ECG classification. *Scientific Reports*, 11(1):5251, 2021.

[14] M. M. Al Rahhal, Yakoub Bazi, Haikel AlHichri, Naif Alajlan, Farid Melgani, and R. R. Yager. Deep learning approach for active classification of electrocardiogram signals. *Information Sciences*, 345:340–354, 2016.

[15] Fatma Murat, Ferhat Sadak, Ozal Yildirim, Muhammed Talo, Ender Murat, Murat Karabatak, Yakup Demir, Ru-San Tan, and U. Rajendra Acharya. Review of deep learning-based atrial fibrillation detection studies. *International Journal of Environmental Research and Public Health*, 18(21):11302, 2021.

[16] Eduardo J. da S. Luz, William Robson Schwartz, Guillermo Cámara-Chávez, and David Menotti. ECG-based heartbeat classification for arrhythmia detection: A survey. *Computer Methods and Programs in Biomedicine*, 127:144–164, 2016.

[17] Antonis A. Armoundas, Sanjiv M. Narayan, Donna K. Arnett, Kayte Spector-Bagdady, Derrick Bennett, Leo Anthony Celi, Paul A. Friedman, Michael H. Gollob, Jennifer L. Hall, Anne E. Kwitek, Elle Lett, Bijoy K. Menon, Katherine S. Sheehan, and Salah S. Al-Zaiti. Use of artificial intelligence in improving outcomes in heart disease: A scientific statement from the American Heart Association. *Circulation*, 149(14), 2024.

[18] Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology. Heart rate variability: standards of measurement, physiological interpretation, and clinical use. *Circulation*, 93(5):1043–1065, 1996.

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

[20] Nitish V. Thakor, John G. Webster, and Willis J. Tompkins. Estimation of QRS complex power spectra for design of a QRS filter. *IEEE Transactions on Biomedical Engineering*, BME-31(11):702–706, 1984.

[21] Juan Pablo Martínez, Rute Almeida, Sonia Olmos, Ana Paula Rocha, and Pablo Laguna. A wavelet-based ECG delineator: Evaluation on standard databases. *IEEE Transactions on Biomedical Engineering*, 51(4):570–581, 2004.

[22] David W. Hosmer Jr, Stanley Lemeshow, and Rodney X. Sturdivant. *Applied Logistic Regression.* John Wiley & Sons, Hoboken, NJ, 3 edition, 2013.

[23] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[24] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

[25] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

[26] Sudestna Nahak, Akanksha Pathak, and Goutam Saha. Evaluation of handcrafted features and learned representations for the classification of arrhythmia and congestive heart failure in ECG. *Biomedical Signal Processing and Control*, 79:104230, 2023.

[27] Syed Muhammad Anwar, Maheen Gul, Muhammad Majid, and Majdi Alnowami. Arrhythmia classification of ECG signals using hybrid features. *Computational and Mathematical Methods in Medicine*, 2018:1–8, 2018.

[28] Xiuzhu Yang, Xinyue Zhang, Mengyao Yang, and Lin Zhang. 12-lead ECG arrhythmia classification using cascaded convolutional neural network and expert feature. *Journal of Electrocardiology*, 67:56–62, 2021.

[29] Haotian Shi, Haoren Wang, Yixiang Huang, Liqun Zhao, Chengjin Qin, and Chengliang Liu. A hierarchical method based on weighted extreme gradient boosting in ECG heartbeat classification. *Computer Methods and Programs in Biomedicine*, 171:1–10, 2019.

[30] Hardik Rajpal, Madalina Sas, Rebecca Joakim, Chris Lockwood, Nicholas S. Peters, and Max Falkenberg. Interpretable XGBoost based classification of 12-lead ECGs applying information theory measures from neuroscience. In *Computing in Cardiology Conference (CinC)*, volume 47, 2020.

[31] Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, Joseph E. Mietus, George B. Moody, Chung-Kang Peng, and H. Eugene Stanley. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

[32] George B. Moody, Roger G. Mark, and Ary L. Goldberger. PhysioNet: a web-based resource for the study of physiologic signals. *IEEE Engineering in Medicine and Biology Magazine*, 20(3):70–75, 2001.

[33] Patrick Wagner, Nils Strodthoff, Ralf-Dieter Bousseljot, Dieter Kreiseler, Fatima I. Lunze, Wojciech Samek, and Tobias Schaeffter. PTB-XL, a large publicly available electrocardiography dataset. *Scientific Data*, 7(1):154, 2020.

[34] Nils Strodthoff, Patrick Wagner, Tobias Schaeffter, and Wojciech Samek. Deep learning for ECG analysis: Benchmarks and insights from PTB-XL. *IEEE Journal of Biomedical and Health Informatics*, 25(5):1519–1528, 2021.

[35] Gari Clifford, Chengyu Liu, Benjamin Moody, Li-wei Lehman, Ikaro Silva, Qiao Li, Alistair Johnson, and Roger Mark. AF classification from a short single lead ECG recording: the PhysioNet/Computing in cardiology challenge 2017. In *Computing in Cardiology Conference (CinC)*, volume 44, 2017.

[36] Matthew Reyna, Erick Andres Perez Alday, Annie Gu, Chengyu Liu, Salman Seyedi, Ali Bahrami Rad, Andoni Elola, Qiao Li, Ashish Sharma, and Gari Clifford. Classification of 12-lead ECGs: the PhysioNet/Computing in cardiology challenge 2020. In *Computing in Cardiology Conference (CinC)*, volume 47, 2020.

[37] Petrus E. O. G. B. Abreu, Gabriela M. M. Paixão, Jiawei Li, Paulo R. Gomes, Peter W. Macfarlane, Ana C. S. Oliveira, Vinícius T. Carvalho, Thomas B. Schön, Antonio Luiz P. Ribeiro, and Antônio H. Ribeiro. CODE-II: a large-scale dataset for artificial intelligence in ECG analysis. *npj Digital Medicine*, 2026.

[38] Konstantinos Sechidis, Grigorios Tsoumakas, and Ioannis Vlahavas. On the stratification of multi-label data. In *Proceedings of the European Conference on Machine Learning and Knowledge Discovery in Databases (ECML PKDD)*, pages 145–158, 2011.

[39] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3):1–13, 2007.

[40] Takaya Saito and Marc Rehmsmeier. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLOS ONE*, 10(3):e0118432, 2015.

[41] Jesse Davis and Mark Goadrich. The relationship between Precision-Recall and ROC curves. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pages 233–240, 2006.

[42] David M. W. Powers. Evaluation: From precision, recall and F-measure to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies*, 2(1):37–63, 2011.

[43] Steven A. Hicks, Inga Strümke, Vajira Thambawita, Malek Hammou, Michael A. Riegler, Pål Halvorsen, and Sravanthi Parasa. On evaluation metrics for medical applications of artificial intelligence. *Scientific Reports*, 12(1):5979, 2022.

[44] Henry Cuthbert Bazett. An analysis of the time-relations of electrocardiograms. *Heart*, 7:353–370, 1920.

[45] L. S. Fridericia. Die systolendauer im elektrokardiogramm bei normalen menschen und bei herzkranken. *Acta Medica Scandinavica*, 53:469–486, 1920.

[46] Fred Shaffer and J. P. Ginsberg. An overview of heart rate variability metrics and norms. *Frontiers in Public Health*, 5:258, 2017.

[47] M. Pagani, F. Lombardi, S. Guzzetti, O. Rimoldi, R. Furlan, P. Pizzinelli, G. Sandrone, G. Malfatto, S. Dell'Orto, and E. Piccaluga. Power spectral analysis of heart rate and arterial pressure variabilities as a marker of sympatho-vagal interaction in man and conscious dog. *Circulation Research*, 59(2):178–193, 1986.

[48] Haibo He and Eduardo A. Garcia. Learning from imbalanced data. *IEEE Transactions on Knowledge and Data Engineering*, 21(9):1263–1284, 2009.

[49] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357, 2002.

[50] Justin M. Johnson and Taghi M. Khoshgoftaar. Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1):27, 2019.

[51] Bradley Efron and Robert Tibshirani. Bootstrap methods for standard errors, confidence intervals, and other measures of statistical accuracy. *Statistical Science*, 1(1):54–75, 1986.

[52] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, Hillsdale, NJ, 2 edition, 1988.

[53] Paul Kligfield, Leonard S. Gettes, James J. Bailey, Rory Childers, Barbara J. Deal, E. William Hancock, Gerard van Herpen, Anton Gorgels, Jan A. Kors, Peter Macfarlane, James W. Mason, David M. Mirvis, Olle Pahlm, Pentti Rautaharju, Borys Surawicz, Galen Wagner, Hein Wellens, and Hein J. J. Wellens. Recommendations for the standardization and interpretation of the electrocardiogram: Part I: The electrocardiogram and its technology. *Journal of the American College of Cardiology*, 49(10):1109–1127, 2007.

[54] Scott M. Lundberg and Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, pages 4765–4774, 2017.

[55] Scott M. Lundberg, Gabriel G. Erion, and Su-In Lee. Consistent individualized feature attribution for tree ensembles. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, volume 97, 2019.

[56] Christoph Molnar. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. Independently published, 2 edition, 2020.

[57] Carolin Strobl, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin, and Achim Zeileis. Conditional variable importance for random forests. *BMC Bioinformatics*, 9(1):307, 2008.

[58] Nils Strodthoff, Temesgen Mehari, Claudia Nagel, Philip J. Aston, Ashish Sundar, Claus Graff, Jørgen K. Kanters, Wilhelm Haverkamp, Olaf Dössel, Axel Loewe, Markus Bär, and Tobias Schaeffter. PTB-XL+, a comprehensive electrocardiographic feature dataset. *Scientific Data*, 10(1):279, 2023.