JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

AN EFFICIENT MULTI-SCALE CNN-BiLSTM WITH DUAL ATTENTION MECHANISM FOR INTER-PATIENT ECG ARRHYTHMIA CLASSIFICATION: EVALUATION ON MIT-BIH AND INCART DATABASES

Rutuja Jadhav[1], Peeyush Pareek[2]

[1]School of Data Science, Symbiosis Skills and Professional University, Pune, Maharashtra, India
[2]Assistant Professor, School of Data Science, Symbiosis Skills and Professional University, Pune, Maharashtra, India

Email: rutujapj1974@gmail.com

Abstract — *Automated electrocardiogram (ECG) arrhythmia classification is essential for scalable, non-invasive cardiac monitoring, yet many reported models are evaluated under intra-patient protocols that overstate real-world performance. This paper presents an inter-patient, AAMI EC57-compliant five-class arrhythmia classification pipeline evaluated on the MIT-BIH Arrhythmia Database under the De Chazal DS1/DS2 split. ECG beats are denoised via bandpass filtering and wavelet thresholding, segmented into fixed windows around annotated R-peaks, and represented using a 67-dimensional handcrafted feature set reduced to 40 features via Minimum Redundancy Maximum Relevance (MRMR) selection, alongside raw waveform input for deep models. Three deep learning architectures, a 1-D CNN, a BiLSTM, and a proposed Multi-Scale CNN-BiLSTM with dual attention, are trained on a two-phase SMOTE-ENN-balanced training set and combined via soft-voting, weighted-voting, and stacking ensembles. The stacking ensemble achieves the best inter-patient test performance, with 94.45% accuracy and a macro-F1 of 76.66% across the five AAMI classes, and Optuna-based hyperparameter tuning further improves the standalone proposed architecture's macro-F1 from 62.99% to 67.12%. External validation on the independent INCART database, performed without retraining, shows a substantial generalization gap (macro-F1 of 35.79%), highlighting cross-database robustness as an open challenge. The supraventricular class remains the dominant source of error across all models and both databases. Results are reported transparently, including configurations where ensembling did not improve on the best individual model, to support reproducible benchmarking in inter-patient ECG arrhythmia classification.*

Keywords — ECG arrhythmia classification, inter-patient paradigm, AAMI EC57, MIT-BIH, INCART, CNN-BiLSTM, dual attention, ensemble learning, SMOTE-ENN, MRMR feature selection

1. INTRODUCTION

Cardiovascular diseases remain among the leading causes of mortality worldwide, and cardiac arrhythmias abnormal heart rhythms detectable from the electrocardiogram (ECG) are a major contributor to this burden. The ECG is a low-cost, non-invasive signal that captures the heart's electrical activity over time, making automated arrhythmia detection an attractive target for clinical decision support: a reliable classifier can screen long ambulatory recordings far faster than manual review by a cardiologist, flagging beats that warrant closer attention.

Building such a classifier is harder than it first appears. ECG morphology varies substantially between patients due to differences in anatomy, lead placement, and underlying pathology, so a model that performs well on beats from patients it has already seen during training can still fail on a genuinely new patient. De Chazal et al. formalised this concern by proposing an inter-patient evaluation protocol that partitions the MIT-BIH Arrhythmia Database into two patient-disjoint sets, DS1 for training and DS2 for testing [4]. Despite this protocol's wide adoption as the field's reporting standard, a large share of published arrhythmia-classification results are still obtained under intra-patient

splits, which allow beats from the same patient to appear in both training and test sets and consequently overstate real-world generalisation [5], [18]. This paper follows the inter-patient protocol throughout.

A second persistent challenge is class imbalance. Under the AAMI EC57 standard, heartbeats are grouped into five classes Normal (N), Supraventricular (S), Ventricular (V), Fusion (F), and Unknown/Paced (Q) [3] and in MIT-BIH the Normal class accounts for the large majority of beats, while the Supraventricular and Fusion classes are rare. A classifier optimised for raw accuracy can achieve a misleadingly high score by defaulting to the majority class while detecting almost none of the minority arrhythmias that matter most clinically. This motivates both the class-balancing strategy adopted in this pipeline and the emphasis on macro-averaged, per-class metrics throughout this paper rather than accuracy alone.

Against this background, this paper makes the following contributions: (1) a Multi-Scale CNN-BiLSTM architecture with a dual attention mechanism, designed to capture both fine-grained QRS morphology and longer-range rhythm context within a single beat window; (2) a patient-aware, inter-patient evaluation protocol with a two-phase SMOTE-ENN balancing and MRMR feature-selection pipeline that avoids test-set leakage at every stage; (3) a systematic comparison of three ensemble strategies soft voting, weighted voting, and stacking combining the proposed architecture with two simpler baselines, reported honestly including cases where ensembling did not improve on the strongest individual model; (4) external, no-retraining validation on the independent INCART database to probe cross-database generalisation, a check largely absent from prior single-database studies; and (5) transparent reporting of the pipeline's principal limitation persistently weak Supraventricular class detection rather than a selectively favourable presentation of results.

The remainder of this paper is organised as follows. Section II reviews related work on classical and deep learning approaches to ECG arrhythmia classification, attention-based and ensemble architectures, and the inter-patient evaluation gap. The Methodology section describes the preprocessing, feature-extraction, balancing, architecture, and evaluation pipeline in detail. The Results and Discussion section reports model and ensemble performance on MIT-BIH, per-class analysis, comparison with published literature, and external validation on INCART. The Conclusion summarises key findings, limitations, and directions for future work.

2. RELATED WORK

A. Classical Machine Learning Approaches

Handcrafted morphological and RR-interval features, paired with classical classifiers, dominated early ECG classification work. The field's most widely cited baseline is still De Chazal et al.'s inter-patient protocol, which paired heartbeat morphology and RR-interval features with linear discriminant analysis and, more importantly, established the patient-disjoint DS1/DS2 split that AAMI-compliant studies are now expected to follow [4]. Boulif et al.'s survey of AI-based ECG diagnosis methods, spanning both classical and deep approaches, makes much the same point from a different angle: inter-patient evaluation and minority-class detection are still open problems across the field, not something a handful of newer architectures have already solved [5]. These feature-based pipelines remain useful precisely because they're interpretable and physiologically grounded, which is why this paper keeps an EMD- and RR-interval-based feature set for its MRMR selection stage (Section III), even though the classifiers doing the actual prediction are deep networks working from raw waveforms.

B. Deep Learning Architectures for ECG Classification

Convolutional and recurrent networks have largely taken over from handcrafted-feature pipelines, since they learn discriminative representations straight from the waveform instead of relying on someone to design the right features by hand. Kiranyaz et al.'s survey traces how widely 1-D CNNs have been adopted across ECG and other biomedical signal tasks, largely because they can learn shift-invariant morphological filters directly from raw signal [18]. Acharya et al.'s early deep CNN, trained end-to-end on raw ECG segments, was one of the first demonstrations that this could work well for heartbeat classification [15]; Hannun et al. later pushed the same basic idea much further, training a similarly structured network on a very large, cardiologist-annotated ambulatory dataset and matching or exceeding cardiologist-level performance on rhythm classification [19]. On the recurrent side, Yildirim combined wavelet decomposition with a bidirectional LSTM to capture both frequency content and sequential context [17], while Sun et al. added squeeze-and-excitation attention on top of a CNN-LSTM backbone to reweight channel-wise feature importance [11]. The BiLSTM and CNN baselines used in this paper (Section III.E) sit squarely in this line of work.

C. Attention-Based and Multi-Scale Architectures

A more recent trend layers explicit attention mechanisms on top of convolutional and recurrent components, giving the model a way to focus on whichever part of the beat actually matters for a given decision. Islam et al.'s HARDC pairs hierarchical attention with a dual-structured RNN and dilated convolutions [8], and a related model from the same group combines a bidirectional GRU-BiLSTM backbone with a multilayered dilated CNN instead [9]. Lyu et al. take a different angle, building a dual-attention network purpose-built for the inter-patient setting that applies attention along

both the temporal and channel dimensions [13]. Mechichi and Benzarti combine CNN, BiLSTM, and attention for ECG segment classification [12]; Reddy et al.'s multi-scale convolutional LSTM-dense network processes the input at several temporal resolutions before fusing them [14]; and Islam et al.'s CAT-Net folds in a transformer block alongside convolution and attention for single-lead classification [10]. The architecture proposed here a Multi-Scale CNN-BiLSTM with dual attention (Section III.E) follows this same general direction, combining multi-scale convolutional feature extraction with a BiLSTM sequence encoder and a dual attention block that operates over both the temporal and feature dimensions.

D. Ensemble and Hybrid Approaches

A separate line of work improves robustness not by building a fancier single model but by combining several models and letting their errors partially cancel out. Bai et al.'s hybrid network for 12-lead arrhythmia diagnosis fuses multiple feature-extraction branches internally [6], and Zhou et al. take a related approach, pairing multiscale convolution with a feature-combination-based attention (FCBA) module for inter-patient heartbeat classification [7]. This paper opts for a simpler and more transparent form of ensembling: rather than fusing architectures inside a single network, the CNN, BiLSTM, and proposed CNN-BiLSTM are trained independently and combined afterward using soft voting, weighted voting, and stacking (Section III.E). The advantage of keeping ensembling external like this is that its actual contribution including the cases where it doesn't help can be measured directly rather than absorbed into a single end-to-end architecture (Section V).

E. Feature Selection, Class Imbalance, and Explainability

On the feature-selection side, Peng et al.'s Minimum Redundancy Maximum Relevance (MRMR) criterion selects features that are both relevant to the target and non-redundant with each other, which is exactly what this paper applies to its 67-dimensional handcrafted feature set (Section III.D) [25]. Class imbalance, meanwhile, is most commonly handled with Chawla et al.'s Synthetic Minority Over-sampling Technique (SMOTE) [24]; hybrid schemes that pair oversampling with cleaning steps have pushed this further still, such as Rai et al.'s combined SMOTE-and-Tomek-Link approach evaluated on pooled MIT-BIH and PTB recordings [20]. This paper's hyperparameter search uses Akiba et al.'s Optuna framework [22], and model interpretability is checked using Lundberg and Lee's SHAP framework [23] reflecting a growing expectation in this field that a reported accuracy number isn't enough on its own; there should also be some evidence that the model is actually looking at physiologically sensible parts of the signal.

F. Research Gap

Put these threads together and a gap becomes visible. Multi-scale convolutional feature extraction, dual attention, ensembling, and rigorous inter-patient evaluation have all been explored but rarely inside the same pipeline, and almost never checked against more than one database. A recent systematic review of deep learning approaches to ECG arrhythmia classification reaches a similar conclusion, noting that most reported architectures are evaluated on a single database under a single split protocol, which limits how confidently their results generalise [26]. Most of the studies reviewed above report results on MIT-BIH alone, with no genuine cross-database generalisation test, and few of them say anything when a proposed enhancement ensembling, tuning, whatever it might be fails to beat a simpler baseline. This paper tries to close that gap: it combines a multi-scale, dual-attention CNN-BiLSTM with a systematic ensemble comparison, evaluates on MIT-BIH under the inter-patient protocol, validates separately on INCART without retraining, and reports both the results that worked out and the ones that didn't.

3. DATASET

This study evaluates a multi-stage ECG arrhythmia classification pipeline that combines classical machine learning models (Random Forest, XGBoost, and a Multi-Layer Perceptron) with deep learning architectures (a 1-D CNN, a BiLSTM, and a combined CNN-BiLSTM network), tuned and combined through an Optuna-optimized ensemble and interpreted through dual explainability analysis (SHAP and gradient-based saliency). The central goal is a rigorous, honest inter-patient benchmark for five-class AAMI arrhythmia classification rather than a single hand-crafted architecture, and a dataset choice that does not stress-test generalization across recording conditions and patient populations would leave that goal untested. For that reason, two complementary PhysioNet databases are used. The MIT-BIH Arrhythmia Database [1], [2] supplies the large, beat-annotated corpus needed to train and rigorously evaluate the five-class AAMI classifier under a strict inter-patient protocol, while the St. Petersburg INCART 12-Lead Arrhythmia Database [16] is used as a fully held-out, beat-level external validation set recorded on different hardware from a different patient population. A third database, PTB-XL [21], is additionally screened at the record level during development as a qualitative cross-cohort sanity check, since it offers a structurally different population (12-lead diagnostic recordings annotated with SCP-ECG statements rather than beat-by-beat AAMI labels) than either MIT-BIH or INCART can provide. Together these three sources let the dataset section speak directly to the paper's central

question: whether the proposed pipeline generalizes across recording conditions and patient populations and not merely within the one dataset it was tuned on.

3.1 Primary Dataset: MIT-BIH Arrhythmia Database

The MIT-BIH Arrhythmia Database, obtained from PhysioNet [1], [2], is the primary source of training, validation, and test data in this study. It contains 48 half-hour ambulatory ECG recordings collected from 47 subjects; one of them contributed two separate recordings, made by the Beth Israel Hospital Arrhythmia Laboratory between 1975 and 1979. All signals are sampled at 360 Hz with 11-bit resolution over a 10 mV range. Each recording captures two leads simultaneously; this study uses the modified Lead II (MLII) channel exclusively, since it is present in all 48 records and gives the clearest view of P-wave and QRS morphology, in line with the majority of published single-lead arrhythmia classifiers [10], [11], [12]. Restricting the pipeline to a single lead also keeps the feature-extraction and deep-learning branches directly comparable across every model tested, since a multi-lead fusion stage would change the input dimensionality differently for the classical and deep-learning arms of the pipeline.

The database provides beat-level annotations for every recording, independently verified by two or more cardiologists [1]. These expert-assigned labels are used directly as ground truth; no automatic beat detector is applied since the annotated R-peak locations already represent the best available localization accuracy [10]. Each beat is segmented as a fixed window of 360 samples (one second) centered on the annotated R-peak, with 100 samples before and 260 samples after the peak. This window is wide enough to capture the complete P-QRS-T morphology together with both adjacent RR intervals, giving the BiLSTM-based branches of the pipeline the temporal context needed to model beat-to-beat rhythm dependencies rather than isolated waveform shape alone [17], [18].

3.1.1 AAMI Beat Labelling

MIT-BIH uses more than fifteen distinct annotation symbols describing specific beat morphologies. To produce a clinically meaningful and reproducible classification target, these raw symbols are grouped into the five classes defined by the AAMI EC57 standard [3], shown in Table 1. This five-class scheme is the de facto benchmark in the arrhythmia classification literature and is what allows this study's results to be compared directly against prior CNN-LSTM and attention-based work [5], [15], [20]. Some published studies adopt non-standard groupings instead, for example, keeping left and right bundle-branch block as separate classes rather than folding both into Class N [6], or using a reduced four-class scheme that drops the rare Q class altogether and reports only N, S, V, and F [7]. This study follows the full AAMI EC57 five-class standard without simplification, which is a harder target precisely because Q and F are the rarest classes and is therefore a more demanding, more clinically honest test of the pipeline's ability to recover minority-class signals.

Table 1. Mapping of MIT-BIH annotation symbols to the five AAMI-standardized classes

AAMI Class	Label	MIT-BIH Symbols	Beat Types Included	Count
Normal (N)	0	N, L, R, B, e, j, n	Normal sinus, Left Bundle-Branch Block, Right Bundle-Branch Block, bundle-branch block (unspecified), atrial escape, nodal escape, supraventricular escape	~90,588
Supraventricular (S)	1	A, a, J, S	Atrial premature, aberrated atrial premature, nodal (junctional) premature, supraventricular premature	~2,779
Ventricular (V)	2	V, E	Premature ventricular contraction, ventricular escape beat	~7,236
Fusion (F)	3	F	Fusion of ventricular and normal beat	~802
Unclassifiable (Q)	4	/, f, Q, ?	Paced beat, fusion of paced and normal, unclassifiable beat, beat not classified during learning	~8,039

Figure 1 shows one representative raw beat from each of the five AAMI classes, segmented from record 101 and centered on the annotated R-peak. It gives a visual reference for the morphological differences the classification pipeline must learn to separate, for instance, the widened, irregular QRS complex typical of Class V against the narrow, regular QRS of Class N.

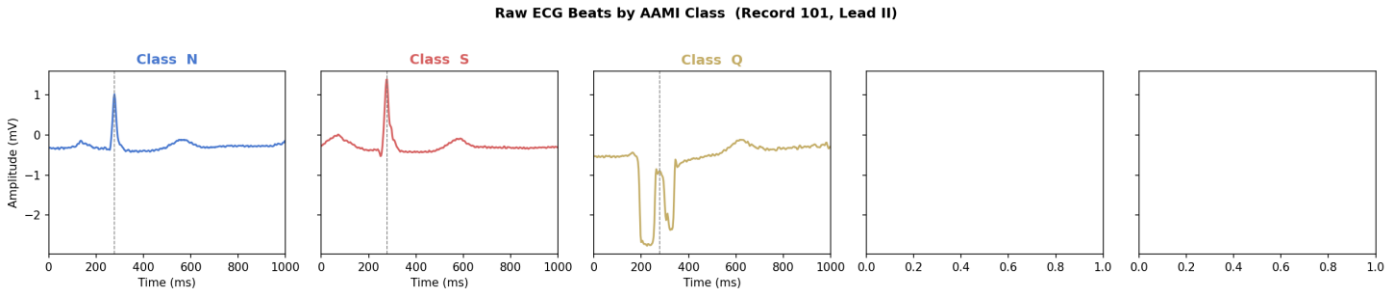


Figure 1. Raw ECG beats by AAMI class (record 101, lead II), one representative waveform per class with the R-peak marked.

The class distribution is sharply imbalanced: Normal (N) beats make up roughly 82.7% of all annotated beats, while Fusion (F) beats account for under 0.7% of the corpus. This skew is a well-documented property of MIT-BIH [15], and if left unaddressed, it pushes a classifier toward predicting the majority class by default, collapsing recall on the clinically important minority classes S and F [15], [10]. Because several models in the comparison have comparatively few parameters relative to larger ensemble architectures in the literature, they are also more vulnerable to majority-class bias if the imbalance is not corrected before training; the S-SMOTE-based balancing strategy adopted to address this is described in Section 4.

3.1.2 Inter-Patient Evaluation Protocol (DS1 / DS2)

A central design decision in this study is the use of an inter-patient evaluation paradigm built on the DS1/DS2 partition originally proposed by de Chazal et al. [4]. Under the original protocol, the 48 records are split into two non-overlapping groups of 22 patients each, listed in Table 2, with four records, 102, 104, 107, and 217, left unassigned because they consist almost entirely of paced beats and fall outside the standard partitioning rule. This evaluation choice matters because a strict patient-level holdout is the only way to credibly claim that reported accuracy reflects genuine arrhythmia discrimination rather than memorized, patient-specific waveform morphology.

Table 2. Inter-patient record assignment used in this study.

Split	Record Numbers	Patients
DS1 (Train)	101, 106, 108, 109, 112, 114, 115, 116, 118, 119, 122, 124, 201, 203, 205, 207, 208, 209, 215, 220, 223, 230	22
DS2 (Test)	100, 103, 105, 111, 113, 117, 121, 123, 200, 202, 210, 212, 213, 214, 219, 221, 222, 228, 231, 232, 233, 234	22
Force-Split (Q/F-dominant patients, 80/20 intra-patient)	101, 102, 104, 105, 107, 114, 124, 200, 201, 202, 203, 205, 208, 210, 213, 214, 215, 217, 219, 223, 233	21

Applying the de Chazal assignment exactly as published creates a practical problem for this study's five-class objective: records 102, 104, 107, and 217 alone carry more than 99% of all Class Q beats in the entire database, yet none of the four appear in either the DS1 or DS2 list. A strict membership filter, therefore, removes Class Q from the test set entirely, making it impossible to report a Q-class score at all, and a less severe version of the same concentration effect occurs for several Class F carriers.

To resolve this without abandoning the inter-patient principle, this study uses a two-stage hybrid split, referred to here as Split v4. In the first stage, the 21 patients identified as Q- or F-class carriers (including the four de Chazal orphans) are each split 80/20 individually under a fixed random seed so every rare-class carrier contributes beats to both training and test. In the second stage, every remaining patient is assigned using GroupShuffleSplit with patient ID as the grouping key and an 80/20 target ratio, guaranteeing that no patient outside the force-split group appears on both sides of the split. The two stages are concatenated to form the final training and test sets, and the patient-disjointness of the 27 GroupShuffleSplit patients is verified programmatically before feature extraction proceeds.

This is a deliberate, disclosed trade-off rather than a hidden deviation: the 80/20 intra-patient split applied to the 21 force-split patients is, strictly speaking, a departure from the pure inter-patient paradigm for those specific patients. It is adopted because it is the only practical way to obtain a non-zero, statistically meaningful test sample for the rarest

AAMI classes while keeping strict inter-patient separation for the remaining 27 patients, who together contribute the large majority of beats. The resulting split places between 15% and 25% of every class's beats in the held-out set, including the previously unrepresented Q class, under a fully reproducible fixed seed. The protocol remains substantially more conservative than the random-split or intra-patient designs used in several earlier studies [5], [15], while explicitly correcting the rare-class concentration problem that a pure DS1/DS2 filter leaves unsolved and prevents the common failure mode where a model quietly learns patient-specific waveform signatures and reports inflated accuracy that does not survive contact with a new patient [9], [7].

3.2 External Validation Dataset: St. Petersburg INCART Database

Rigor claims are only meaningful if the underlying classifiers also generalize. A pipeline that only works on the data it was tuned on is not clinically useful. To test this directly, the St. Petersburg INCART 12-Lead Arrhythmia Database [16], made available through PhysioNet, is used as a fully external validation set. INCART comprises 75 half-hour, 12-lead Holter recordings drawn from 32 patients under evaluation for coronary artery disease, conduction abnormalities, and arrhythmia. Signals are sampled natively at 257 Hz, lower than MIT-BIH's 360 Hz, and are resampled to 360 Hz before beat segmentation so that every INCART beat presents each model with the same one-second window shape it was trained on.

What makes INCART suitable for genuine quantitative external validation, rather than a loose qualitative check, is its annotation scheme: INCART beat labels follow the same PhysioBank symbol convention as MIT-BIH, with an automated detector proposing initial beat locations and symbols that were then manually corrected by expert annotators [16]. Because of this shared convention, the AAMI symbol-to-class mapping in Table 1 carries over to INCART without modification, enabling true per-beat ground-truth scoring. PTB-XL, by contrast, supplies only record-level SCP-ECG diagnostic statements rather than per-beat labels [21], which is why it is used in this study as a secondary, record-level consistency screen rather than as a second quantitative benchmark, a distinction made explicit in Section 3.3. Of INCART's twelve available leads, Lead II is selected for consistency with the MLII channel the models were trained on, preserving the same morphological frame of reference at test time.

Because 75 recordings are drawn from only 32 patients, several patients contributed multiple records, but INCART is used in this study exclusively as an external validation set: no model is ever trained, fine-tuned, or hyperparameter-optimized on any INCART data. Every beat evaluated from this database is therefore genuinely unseen, which is the most stringent generalization test the dataset design can offer beyond the inter-patient MIT-BIH split. This mirrors the external-validation approach used in recent competitive architectures [10] and directly answers the cross-dataset generalization gap that earlier BiLSTM-based arrhythmia studies have flagged but not always closed [9].

3.3 Dataset Summary

Table 3 reports the final per-class beat counts produced once both stages of Split v4 force-split and GroupShuffleSplit are combined. Every AAMI class falls within the 15%–25% test-share band the split was designed to hit, including Class Q, which would have received zero test beats under a pure DS1/DS2 filter.

Table 3. Final per-class beat distribution after Split v4, prior to S-SMOTE balancing of the training set.

AAMI Class	Train Beats	Test Beats	Test Share	In Target Range (15–25%)
N	70,255	20,154	22.3%	Yes
S	2,288	409	15.2%	Yes
V	5,500	1,042	15.9%	Yes
F	649	146	18.4%	Yes
Q	6,405	1,618	20.2%	Yes
Total	85,097	23,369	21.6%	—

Table 4 summarises the role each database plays in this study. Using a large, strictly inter-patient-split primary database together with a beat-level external validation set lets the evaluation speak to both halves of the paper's central claim that the proposed ensemble pipeline is accurate within MIT-BIH under a rigorous inter-patient protocol, and that this accuracy is not an artefact of fitting MIT-BIH's specific recording hardware and patient pool.

Table 4. Comparison of the primary (MIT-BIH) and external validation (INCART) datasets used in this study.

Property	MIT-BIH (Primary)	INCART (External Validation)
Records	48	75
Patients	47	32
Sampling Rate	360 Hz	257 Hz (resampled to 360 Hz)
Duration per Record	~30 minutes	~30 minutes
Leads Available	2 (MLII, V1)	12 — Lead II selected
Beat Window	360 samples (1 s)	360 samples after resampling
Annotation Scheme	PhysioBank, 2 cardiologists	PhysioBank, expert-corrected
AAMI Classes Present	All 5 (N, S, V, F, Q)	N, S, V (F and Q too sparse)
Approx. Total Beats	~109,494	~175,000+
Role in This Study	Train / Validate / Test (Split v4)	External validation only

It is worth stating plainly that this study's only true beat-level external validation is performed on INCART. PTB-XL [21] is additionally screened during development as a secondary, record-level sanity check on prediction direction across two clinically distinct rhythm cohorts, sinus rhythm versus atrial fibrillation, since it provides only record-level diagnostic statements rather than per-beat AAMI annotations and cannot support the same per-beat scoring that MIT-BIH and INCART allow. This screening is reported separately as a qualitative consistency check rather than a quantitative cross-dataset accuracy benchmark, keeping the distinction between rigorous and exploratory evaluation clear throughout the Results section.

4. METHODOLOGY

This section presents the complete methodological pipeline through which the raw MIT-BIH and INCART recordings described in Section 3 are converted into five-class arrhythmia predictions. The pipeline proceeds through six broad stages: (i) signal preprocessing and denoising; (ii) R-peak localization and beat segmentation; (iii) feature extraction combining Empirical Mode Decomposition (EMD) with morphological and RR-interval descriptors; (iv) inter-patient data partitioning, feature selection, and class-imbalance correction; (v) deep learning classification using two controlled baseline architectures (a single-scale 1-D CNN and a bidirectional LSTM) together with the proposed Multi-Scale CNN-BiLSTM with Dual Attention architecture, combined through ensemble learning and further refined through Optuna-based hyperparameter optimization; and (vi) model explainability and external validation. Each stage is described in the subsections that follow.

4.1 Overview of the Proposed Framework

Each of the 48 MIT-BIH recordings is first denoised using a two-stage filtering pipeline (Section 4.2), and beats are segmented around cardiologist-verified R-peak annotations (Section 4.3). Four temporal RR-interval features are computed per beat (Section 4.4) and concatenated with a 63-dimensional EMD-derived feature vector capturing spectral, morphological, and P-wave structure (Section 4.5), producing a 67-dimensional descriptor per beat. Beats are then partitioned into training and test sets under a strict inter-patient protocol (Section 4.6); the feature space is reduced from 67 to 40 dimensions through Minimum Redundancy Maximum Relevance (MRMR) selection (Section 4.7); and class imbalance is corrected separately for the feature-based and raw-waveform branches of the pipeline (Section 4.8). Three deep learning architectures (the two controlled baselines and the proposed Multi-Scale CNN-BiLSTM with Dual Attention) are trained directly on the balanced raw waveforms (Section 4.9) and combined through three complementary ensemble strategies (Section 4.10), with the proposed architecture further refined through an Optuna hyperparameter search (Section 4.11). Model behaviour is interpreted using SHAP and gradient-based saliency analysis (Section 4.12),

and generalization beyond MIT-BIH is assessed on an independently annotated external database (Section 4.13), using the evaluation metrics defined in Section 4.14.

4.2 Signal Preprocessing

Raw ECG signals are corrupted by physiological and instrumentation noise sources that distort beat morphology and degrade downstream classification accuracy if left uncorrected. Each of the 48 MIT-BIH recordings therefore underwent a two-stage denoising pipeline: bandpass filtering followed by wavelet-domain thresholding.

4.2.1 Bandpass Filtering

Two noise sources dominate at opposite ends of the ECG frequency spectrum. At the low end, baseline wander, a slow drift below 0.5 Hz caused by respiration, patient movement, or poor electrode contact, does not destroy the QRS complex but shifts the signal's absolute amplitude over time, which is problematic for amplitude-dependent features such as ST-segment deviation. At the high end, electromyographic (EMG) noise above approximately 40-45 Hz, caused by skeletal-muscle activity near the electrodes, blurs the sharp edges of the QRS complex on which R-peak detection and several morphology features depend.

To suppress both without damaging the 5–40 Hz band in which QRS energy is concentrated, a 4th-order Butterworth bandpass filter from 0.5–45 Hz was applied, with cutoffs normalized to the Nyquist frequency:

$$f_{nyq} = f_s / 2 = 180\ Hz (at\ f_s = 360\ Hz)$$

A 4th-order filter was chosen as a deliberate middle ground: steep enough to attenuate out-of-band noise, but not so steep that it introduces ringing near the passband edges that would itself distort beat morphology. This is a standard trade-off in ECG baseline-wander and high-frequency noise removal.

Filtering was implemented as a forward–backward, zero-phase (filtfilt) operation rather than a single-pass causal filter. A standard causal filter introduces a time-domain phase delay that would shift the apparent location of the R-peak and, by extension, every downstream beat-segmentation boundary and RR-interval measurement. This is particularly consequential here because the R-peak locations used for segmentation (Section 4.3) are PhysioNet's cardiologist-verified annotations, defined on the original, unfiltered timeline; any phase delay introduced by filtering would silently desynchronize the filtered signal from those annotation indices. The zero-phase implementation eliminates this risk by construction, guaranteeing that the filtered signal remains exactly time-aligned with the annotation timestamps.

4.2.2 Wavelet-Domain Denoising

Bandpass filtering cannot separate noise that falls inside the 5–40 Hz QRS band itself, since a fixed-frequency filter has no way to distinguish noise at a given frequency from signal at that same frequency. To address this, the bandpass-filtered signal was further decomposed using an 8-level discrete wavelet transform (DWT) with the Symlet-4 (Sym4) wavelet, whose shape closely resembles a QRS complex. Under this decomposition, heartbeat energy concentrates into a small number of large coefficients while unstructured noise spreads thinly across many small ones, allowing signal and noise to be separated by magnitude rather than frequency alone. This Sym4 / 8-level configuration has previously been shown to outperform competing thresholding schemes on this same database [9], and related Symlet-based soft-thresholding pipelines have achieved QRS-detection sensitivity above 99.9% on MIT-BIH [10].

Denoising followed the classical VisuShrink universal-threshold procedure. The noise level σ was estimated from the finest-scale detail coefficients (cD1) using the median absolute deviation, which is robust to the small number of genuinely large coefficients produced by the QRS upstroke:

$$\sigma = median(|c_{D1}|) / 0.6745$$

A single global threshold λ was then computed from this noise estimate and the signal length N:

$$\lambda = \sigma \times \sqrt{(2 \ln N)}$$

Soft thresholding was applied across every detail level: coefficients below λ were zeroed and coefficients above λ were shrunk toward zero by that amount, while the coarsest approximation coefficients were left untouched, since at that scale they represent genuine baseline structure rather than noise. The signal was then reconstructed via the inverse DWT. Soft thresholding was preferred over hard thresholding because hard thresholding introduces a discontinuity at the cutoff that manifests as small artificial spikes upon reconstruction, whereas soft thresholding shrinks coefficients smoothly and avoids injecting spurious structure into a waveform whose later features depend on subtle amplitude and slope characteristics.

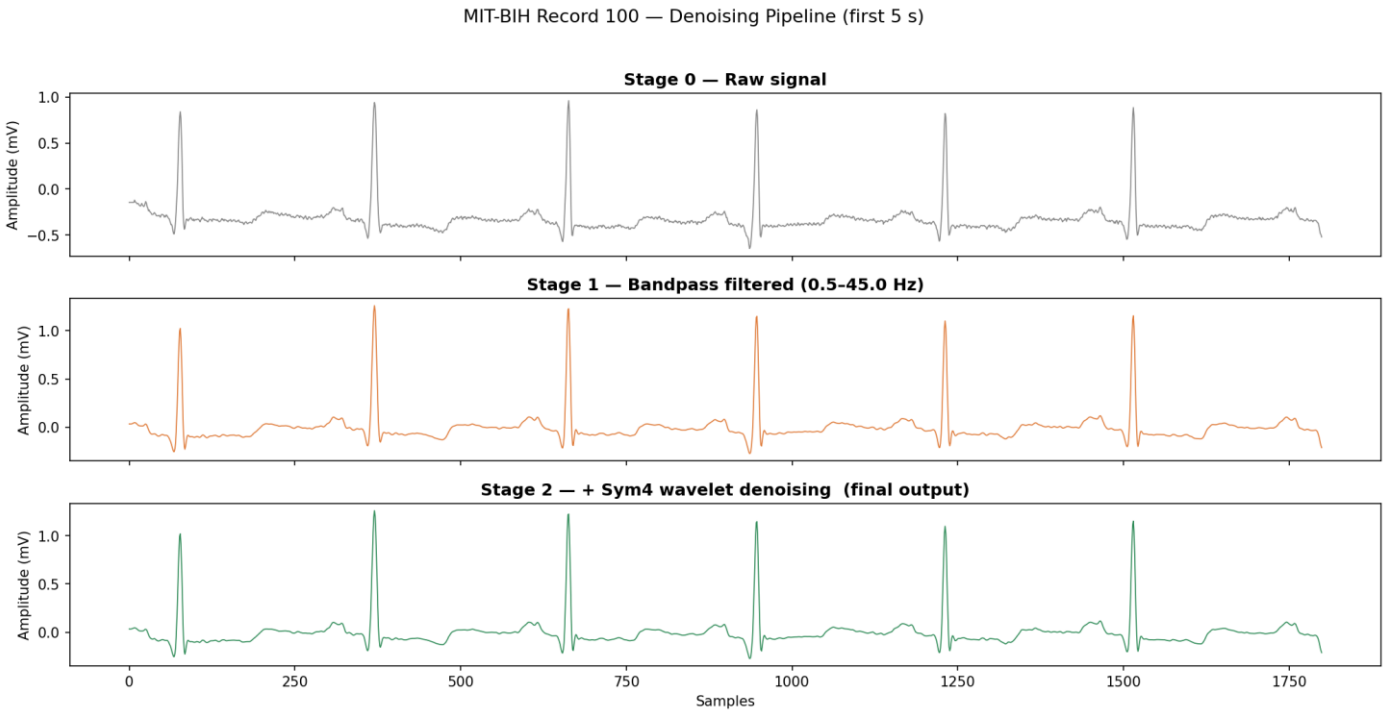


figure 2. two-stage ECG denoising pipeline applied to MIT-BIH record 100

4.3 R-Peak Localization and Beat Segmentation

Rather than running an independent R-peak detector on the cleaned signal, this study used PhysioNet's beat annotations directly for segmentation. These annotations were independently verified by two or more cardiologists and therefore represent a more reliable ground truth than any automated detector could provide. As a sanity check on this decision, an automated detector was run on the denoised signal for six representative records (100, 101, 119, 200, 217, and 231), and its output was compared against the true annotations using a ±50 ms tolerance window. This check confirmed that the two-stage denoising pipeline preserves R-peak timing rather than introducing drift; only after this verification did the study commit to using annotation-based locations for every extracted beat.

Each beat was then cut as a fixed 360-sample window (one second of signal at the 360 Hz sampling rate), using an asymmetric placement relative to the annotated R-peak rather than a centered window, as summarized in Table 5.

Table 5. Asymmetric beat-segmentation window relative to the annotated R-peak

Segment	Samples	Captures
Pre-R	100	P-wave and PR segment
Post-R	260	QRS tail, ST segment, and T-wave

This asymmetry follows directly from cardiac electrophysiology: the T-wave, which follows the QRS complex and reflects ventricular repolarization, takes longer to resolve than the P-wave that precedes the complex. A symmetric window sized to fit the P-wave would truncate the T-wave, while a window sized to fit the T-wave would over-pad the P-wave side; the 100/260-sample split was chosen to capture both fully without wasting samples on either side. Beats positioned close enough to the start or end of a recording that this window would run off the signal boundary were dropped and logged, affecting only one or two beats per record with no material effect on overall class balance.

4.4 RR-Interval Feature Extraction

Four temporal features were computed per beat from the sequence of annotated R-peak positions, capturing timing information that complements waveform shape. A premature ventricular contraction, for instance, characteristically arrives early and is followed by a longer compensatory pause, a signature that morphology alone does not encode.

Table 6. RR-interval features computed per beat from annotated R-peak positions

Feature	Definition	Rationale
pre_rr	Time since the previous R-peak	Indicates whether the beat is premature or delayed
post_rr	Time until the next R-peak	Captures the compensatory pause following ectopic beats
local_rr	Mean of the preceding 10 pre_rr values	Normalizes against the patient's recent baseline heart rate
rr_ratio	pre_rr / post_rr	Encodes prematurity-then-pause as a single value

$$rr_ratio = pre_rr / post_rr$$

rr_ratio tends toward 1 for normal beats and falls well below 1 for ectopic beats that arrive early and are followed by a pause. local_rr is computed over a sliding window of up to 10 preceding intervals (fewer near the start of a recording), letting the timing features normalize against each patient's own baseline rhythm rather than a fixed global threshold, which matters given how widely resting heart rate varies across the 47 MIT-BIH subjects. For the first and last beat of each recording, where one neighbouring interval is undefined by construction, the missing value was substituted with the available neighbouring interval rather than left undefined. These four features were concatenated with the EMD-derived features described in Section 4.5, forming the final four dimensions of the complete 67-dimensional feature vector per beat.

4.5 Feature Extraction via Empirical Mode Decomposition

Each denoised beat was decomposed using Empirical Mode Decomposition (EMD), a fully data-driven technique that breaks a non-stationary, non-linear signal into oscillatory components, known as Intrinsic Mode Functions (IMFs), without assuming any fixed basis function. This adaptiveness makes EMD better suited to ECG morphology than a fixed-basis method such as the wavelet transform, since beat shape varies considerably across patients and rhythms. Each beat was decomposed into its first 8 IMFs, ordered from highest to lowest characteristic frequency; on the rare occasions where EMD produced fewer than 8 modes, the remaining slots were zero-padded.

This design is grounded in prior work: feeding EMD-derived features into a CNN has been shown to raise MIT-BIH classification accuracy from 90.69% without EMD to 99.01% with it, indicating that EMD contributes genuine discriminative signal rather than redundant computation [12]. Pairing EMD with nonlinear descriptors such as sample entropy, rather than simple amplitude statistics, was established by one of the most-cited EMD-for-ECG feature studies [13], motivating the inclusion of sample entropy among the per-IMF statistics computed here. Multilayer statistical features computed across several decomposition levels, rather than a single level, have likewise been shown to be discriminative when fed into a deep network [14].

For each IMF, five statistics were computed: mean, standard deviation, energy, kurtosis, and sample entropy, followed by three global energy-ratio features summarizing how energy is distributed across the decomposition. The complete composition of the 67-dimensional feature vector is summarized in Table 7.

Table 7. Composition of the 67-dimensional EMD-derived feature vector per beat

Feature Group	Dimensions	Description
IMF statistical features	43	Mean, SD, energy, kurtosis, and sample entropy for each of the 8 IMFs (40 features), plus 3 inter-IMF energy ratios
Waveform morphology	12	QRS amplitude/duration, T-wave amplitude, ST deviation, PR interval, beat energy, zero-crossing rate, peak asymmetry, skewness, waveform entropy, slope statistics
P-wave features	8	Amplitude, energy, presence, duration, area, symmetry, peak position, standard deviation
RR-interval features	4	Carried forward from Table 6
Total	67	

Kurtosis was computed using the bias-corrected Fisher estimator, appropriate given the short 360-sample length of each beat, for which the bias correction has a non-negligible effect relative to the asymptotic estimator typically used for longer signals. Sample entropy followed the standard Richman–Moorman formulation with embedding dimension $m = 2$ and Chebyshev distance, the same parameterization used in [13]; on the rare IMF with near-zero variance, where sample entropy is mathematically undefined, the feature was set to zero rather than left undefined, avoiding downstream NaN propagation.

The three energy-ratio features were grouped to reflect approximate physiological frequency bands rather than split evenly across the eight IMFs: IMF-1 alone (residual high-frequency noise), IMF-2 and IMF-3 combined (the sharp QRS event itself), and IMF-4 through IMF-8 combined (the slower P-wave, T-wave, and baseline morphology). This grouping gives the downstream classifiers an explicit summary of how much of each beat's energy sits in noise, in the QRS complex, and in the surrounding waveform structure. The fusion of EMD-derived spectral descriptors with classical morphology and RR-interval features gives the classifiers complementary views of each beat: EMD captures subtle frequency-domain irregularities that distinguish ectopic and fusion beats, while the morphology and timing features encode the markers cardiologists already rely on for manual diagnosis.

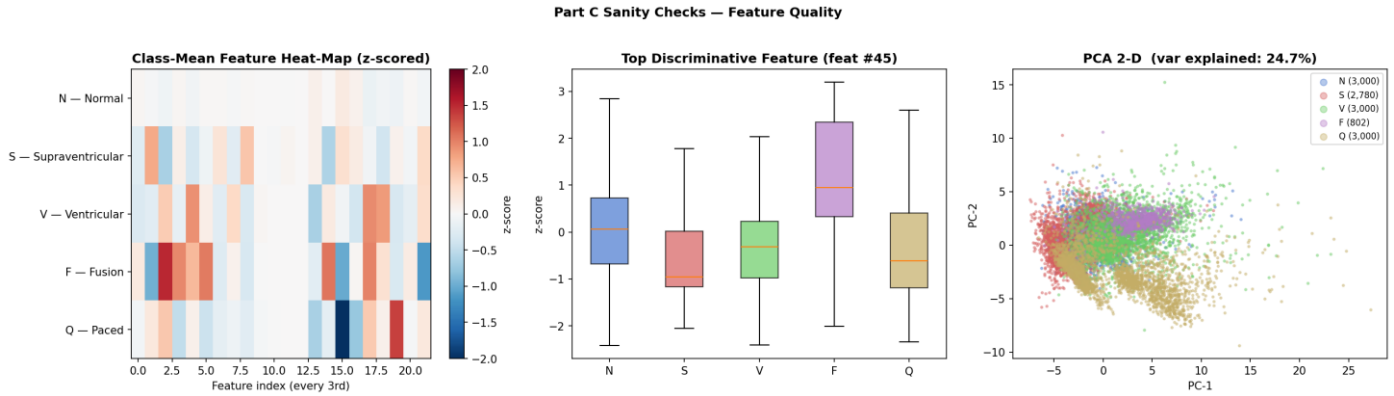


figure 3. feature-space validation: class-mean heat-map, most discriminative feature, and 2-D PCA projection

4.6 Inter-Patient Train/Test Partitioning

A strict inter-patient evaluation paradigm was enforced throughout, since mixing beats from the same patient across training and test sets has been shown to artificially inflate performance: beats from one patient share correlated morphology arising from electrode placement, baseline rhythm, and individual anatomy, so a model evaluated this way risks learning to recognize the patient rather than the arrhythmia [3]. The base partition follows the original de Chazal DS1/DS2 scheme, splitting 44 of the 48 records into two groups of 22 with zero patient overlap [5]. This scheme, however, was defined only over the 44 non-pacemaker recordings: it explicitly excludes patients 102, 104, 107, and 217, who together carry the overwhelming majority of all Q-class (paced) beats in the database. Applying the DS1/DS2 split unmodified would therefore eliminate the Q class from evaluation almost entirely.

To resolve this without abandoning the inter-patient principle, a hybrid two-stage split was used. Nineteen patients whose recordings disproportionately carry Q- or F-class beats were identified and handled separately, spanning patients with very heavy Q-class representation (102, 104, 107, 217), lighter Q-class presence (101, 105), F-class-dominant patients (208, 210, 213, 223), moderate F-class representation (114, 124, 205, 233), and a further group with smaller but non-trivial F- or Q-class counts (200, 201, 202, 203, 214, 215, 219). Each of these 19 patients was split 80/20 within-patient using a fixed random seed, guaranteeing that both training and test sets contain genuine examples of every rare class rather than relying on chance patient assignment. The remaining 25 patients, those without a disproportionate share of rare-class beats, were partitioned using GroupShuffleSplit (80/20, grouped by patient ID), preserving strict inter-patient separation for the bulk of the dataset; this portion of the split was independently verified to contain zero overlapping patient IDs between training and test.

Table 8. Final train/test split by AAMI class

Class	Train	Test	Test %
N	70,255	20,154	22.3%
S	2,288	409	15.2%
V	5,500	1,042	15.9%
F	649	146	18.4%
Q	6,405	1,618	20.2%
Total	85,097	23,369	21.6%

Every class's test share falls within a 15–25% band, confirming that the hybrid split avoids the class-collapse failure mode that a pure DS1/DS2 partition would have produced for the Q class specifically. The within-patient 80/20 split applied to the 19 force-split patients is a minor, deliberate deviation from a pure inter-patient protocol and is discussed as a limitation later in this paper. Feature normalization (z-score standardization via a StandardScaler) was fit exclusively on the training partition and applied to the test partition without refitting, preventing any leakage of test-set statistics into training.

4.7 Feature Selection via MRMR

To reduce dimensionality and remove redundant or weakly discriminative features among the 67 extracted, Minimum Redundancy Maximum Relevance (MRMR) feature selection was applied to the training set only [25]. MRMR selects features that jointly maximize relevance to the class label while minimizing mutual redundancy between already-selected features, a more principled criterion than ranking by relevance alone, since two highly relevant but redundant features add little beyond what either contributes individually.

Because MRMR's mutual-information computation scales with both the number of samples and the number of candidate features, selection was run on a random subsample of 50,000 training beats drawn without replacement from

the full 85,097-beat training partition, trading a small amount of statistical precision for tractable runtime without materially affecting which features were selected, given that the underlying class structure is well represented at that sample size. The top 40 of the 67 features were retained for downstream use, and both the selected feature indices and their names were persisted for reuse in later stages of the pipeline, including the SHAP-based explainability analysis in Section 4.12.

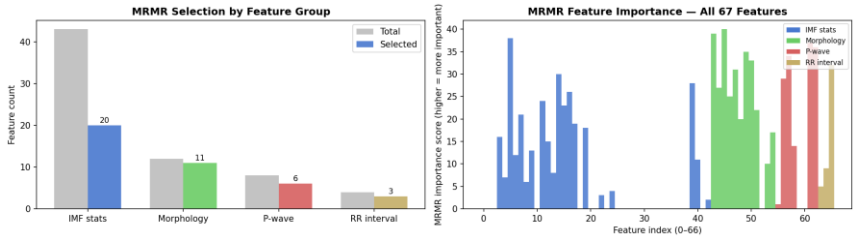


figure 4. feature groups retained by MRMR selection and per-feature importance across all 67 candidate features

4.8 Class Imbalance Correction

The natural class distribution is heavily skewed toward Normal beats; in the training partition, the Normal class outnumbers the Supraventricular class by roughly 30:1, a ratio severe enough that a classifier optimizing for overall accuracy could simply predict "N" almost every time while missing nearly every clinically significant ectopic beat. Class imbalance was corrected separately for the two branches of the pipeline, since the feature-based branch (40-dimensional MRMR-selected vectors) and the raw-waveform branch (360-sample sequences fed to the deep networks) occupy very different representation spaces.

4.8.1 S-SMOTE for the Feature-Based Branch

A two-phase Synthetic Minority Oversampling Technique (S-SMOTE) was applied to the MRMR-selected feature space, run on the training set only and strictly after partitioning, so that no synthetic sample could leak into the test set. SMOTE-based resampling of imbalanced ECG data has previously been shown to meaningfully improve minority-class recall: a hybrid SMOTE-plus-Tomek-Link scheme applied to combined MIT-BIH and PTB recordings improved individual minority-class accuracy by more than 15 percentage points over the imbalanced baseline while keeping overall accuracy above 99% on the resulting balanced dataset [20]. That said, this and most comparable studies apply a single, undifferentiated oversampling pass across all minority classes simultaneously, rather than the staged, class-prioritized approach adopted here.

Phase 1: S-class targeted uplift. The Supraventricular class was oversampled first and independently, before any other class was touched. The target count was set adaptively (at least 20% of the Normal-class count, or three times the S class's existing count, whichever was larger), with the SMOTE neighbour count capped below the real, pre-synthetic S-class size to avoid the algorithm searching for neighbours that do not exist. This phase exists because the S class is both the rarest by ratio and, clinically, one of the most consequential to miss; treating it first and in isolation means its synthetic examples are interpolated from a relatively pure S-only neighbourhood, before the broader balancing pass in Phase 2 introduces synthetic neighbours from other classes into the same feature space.

Phase 2: full balance. The remaining minority classes (V, F, and Q) were then oversampled to match the Normal-class count exactly, producing a fully class-balanced training set in which every class contributes an equal share of beats. Because some classes, F in particular, have very few real examples to begin with, this full-balance target can imply a large oversampling ratio for those classes; the pipeline explicitly computes this ratio per class and flags any ratio exceeding 50x as a point requiring review, since at extreme ratios SMOTE's synthetic beats risk overrepresenting a narrow slice of a class's natural variability rather than contributing genuinely new information.

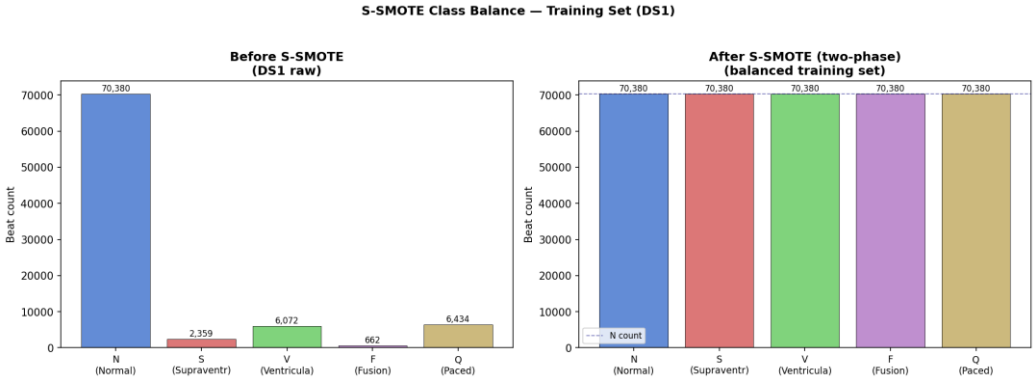


figure 5. training-set class distribution before and after two-phase s-smote

4.8.2 Full-Class SMOTE for the Raw-Waveform Deep Learning Branch

The deep learning models described in Section 4.9 train directly on raw 360-sample beat waveforms rather than hand-crafted features. SMOTE was therefore re-applied independently in this 360-dimensional waveform space, again strictly on the training partition after the split described in Section 4.6. Unlike the two-phase S-SMOTE used for the feature-based branch, a single-pass full-class SMOTE was used here: all minority classes are upsampled simultaneously to 50% of the Normal-class count. This simpler scheme is sufficient in the waveform space because the raw signal already contains rich discriminative structure that the convolutional and recurrent layers learn to exploit directly, without requiring the same staged, S-class-first treatment that the lower-dimensional feature space required.

4.9 Deep Learning Classification Architectures

Three deep learning models are trained end-to-end on the balanced raw waveforms from Section 4.8.2. Every model receives the same input, a 360-sample, single-lead beat tensor after per-beat z-score normalisation, and the same training regime: Adam optimiser with an initial learning rate of 1x10^-3, categorical cross-entropy loss, mini-batches of 128 beats, and a ceiling of 100 training epochs. Early stopping monitors validation macro-F1 with a patience of 10 epochs, a custom MacroF1Callback computes this score at the end of every epoch on a stratified 10% validation split of the balanced training set, and ReduceLROnPlateau halves the learning rate whenever validation loss plateaus for five consecutive epochs (minimum 1x10^-6). The best-weight snapshot is restored via ModelCheckpoint before any test-set evaluation, so reported metrics always reflect the model at its validation peak rather than its final iterate. The validation split is carved out only for early stopping; it plays no role in the final DS2 test evaluation. Monitoring macro-F1 rather than validation accuracy was a deliberate choice: with 82.7% of training beats belonging to the Normal class, a model could minimise cross-entropy by collapsing to a near-always-Normal predictor and still show low loss, whereas macro-F1 decline is immediate when minority-class recall falls.

The three architectures serve distinct roles in the paper's ablation design. The single-scale CNN and the standalone BiLSTM are controlled baselines that isolate one inductive bias each, establishing a floor for what a purely convolutional or purely recurrent model achieves under identical training conditions. The Multi-Scale CNN-BiLSTM with Dual Attention is the proposed model, combining multi-resolution feature extraction, selective channel recalibration, temporal weighting, and bidirectional sequence modelling into a single lightweight end-to-end architecture. Comparing all three on the same DS2 test set makes it possible to attribute any accuracy gain specifically to the components the proposed model adds, rather than to training-regime differences, data advantages, or ensemble effects.

4.9.1 Baseline Model 1 — Single-Scale 1-D CNN

The first baseline is a three-block 1-D convolutional network designed to test whether sequential, single-resolution feature extraction is sufficient for five-class AAMI arrhythmia classification. Block 1 applies 64 filters with a kernel size of 16 samples (~44 ms at 360 Hz), wide enough to span the QRS complex in its entirety; Block 2 applies 128 filters with a kernel size of 8 samples, capturing finer within-QRS detail; Block 3 applies 256 filters with a kernel size of 4 samples, learning the sharpest local edges. Each block is followed by Batch Normalisation, ReLU activation, and MaxPooling(2), progressively halving the sequence length while doubling the channel width. Global Average Pooling after the third block collapses the feature map into a fixed-length vector without a dense flatten step, removing the input-length dependency that a Flatten to Dense transition would create, and acting as a structural regulariser by averaging spatial information rather than concatenating it [18]. The classifier head is Dense(128) -> Dropout(0.3) -> Dense(64) -> Dropout(0.3) -> Softmax(5).

The key design constraint of this baseline is intentional: the three kernel sizes decrease monotonically (16 → 8 → 4) and the blocks are sequential, so at any given layer the network sees features at only one spatial scale simultaneously. This is the standard single-scale convolutional design used in many earlier ECG classifiers [15], [18]. Its role in this study is to establish the performance ceiling for that conventional approach under the same training conditions as the proposed model, making the accuracy gain attributable to multi-scale, parallel feature extraction quantifiable rather than impressionistic.

4.9.2 Baseline Model 2 — Bidirectional LSTM

The second baseline is a two-layer Bidirectional LSTM that processes the raw 360-sample waveform without any convolutional pre-processing. The first layer, Bidirectional(LSTM(128, return_sequences=True)), reads each beat sample-by-sample in both the forward and backward temporal directions, producing a 256-dimensional hidden state (128 from each direction) at every timestep. The backward pass is what makes this baseline non-trivial for ECG classification: a unidirectional LSTM classifying sample t has only seen the waveform up to that point, meaning it cannot observe the T-wave when deciding how to represent the QRS complex, yet the T-wave morphology and timing are diagnostically important. The bidirectional design gives each hidden state access to the complete one-second beat context in both directions simultaneously [17]. A Dropout(0.3) layer follows before the second Bidirectional(LSTM(64)) layer, which collapses the sequence into a single 128-dimensional beat-level vector. A Dense(64) -> Dropout(0.3) -> Softmax(5) classifier head then maps this to the five classes.

This baseline tests a specific hypothesis: that recurrent temporal modelling alone, without any convolutional front end to extract local waveform features first, is competitive with the proposed architecture. Its result in the ablation comparison directly quantifies how much the multi-scale CNN front end and the dual attention mechanism contribute beyond what BiLSTM recurrence alone can achieve on raw waveforms.

4.9.3 Proposed Model — Multi-Scale CNN-BiLSTM with Dual Attention

The proposed architecture is the paper's central contribution. It addresses three limitations that the two baselines leave individually unresolved: single-scale convolution cannot extract both narrow QRS detail and broad P/T-wave morphology simultaneously from the same layer; a pure BiLSTM without attention cannot selectively suppress uninformative time steps before recurrence begins, wasting capacity on flat pre-P baseline segments; and a CNN without sequence modelling cannot capture the beat-to-beat rhythm irregularities, such as the characteristic early arrival and compensatory pause of a premature ventricular contraction, that are as diagnostically important as the within-beat waveform shape. The proposed model resolves all three by stacking multi-scale convolutional feature extraction, channel attention, temporal attention, and bidirectional recurrence in a single lightweight end-to-end pipeline. The full block-by-block description follows.

Multi-Scale CNN, Stage 1. Three parallel Conv1D branches process the input 360-sample waveform simultaneously, each with a different receptive field. The first branch uses a kernel of 3 samples (~8 ms), which is narrow enough to detect the sharp gradient changes at the onset and offset of the QRS upstroke that distinguish a narrow normal complex from the widened morphology of a left bundle-branch block or premature ventricular contraction. The second branch uses a kernel of 7 samples (~19 ms), spanning the width of a typical P-wave and capturing the smoother amplitude envelope that separates Supraventricular beats, which preserve P-wave structure, from Ventricular beats, which do not. The third branch uses a kernel of 15 samples (~42 ms), approximately the mean QRS duration in healthy adults, and captures overall beat shape at the scale most relevant to evaluating gross waveform symmetry, ST-segment slope, and T-wave polarity. Each branch produces 32 feature channels. Their outputs are concatenated to form a (360, 96) multi-resolution feature map, which is then Batch Normalised and MaxPooled(2) to (180, 96). This three-branch parallel design is directly motivated by Zhou et al. [7] and Reddy et al. [14], who showed that multi-scale convolutional extraction provides more discriminative ECG features than single-scale designs, with the largest gains on minority classes where subtle morphological differences drive classification.

Multi-Scale CNN, Stage 2. A second parallel stage refines the multi-resolution representation from Stage 1. Three more Conv1D branches at the same kernel sizes (3, 7, 15) operate on the 180-timestep Stage 1 output, each now producing 64 channels. After concatenation, the result is a (180, 192) feature map, again Batch Normalised and MaxPooled(2) to (90, 192). The second stage allows the network to learn hierarchical combinations of the three spatial scales, such as interactions between fine QRS edges and broad T-wave shape, rather than relying on raw single-scale activations from Stage 1 alone. Adding depth via a second multi-scale stage rather than simply widening Stage 1 keeps the parameter count controlled, since spatial pooling has already compressed the sequence before Stage 2 begins.

Channel Attention, Squeeze-and-Excitation Style. After Stage 2, not all 192 learned feature channels are equally informative for classifying every beat. A Squeeze-Excitation-style channel attention block [11] learns to recalibrate which channels matter on a beat-by-beat basis. Both Global Average Pooling and Global Max Pooling are applied across the temporal axis of the (90, 192) feature map, producing two complementary 192-dimensional summary vectors: the average pool captures what is typically present across the beat, while the max pool highlights the most extreme activation in any channel. Both vectors are passed through the same shared two-layer MLP, Dense(max(192/8, 4), ReLU) then Dense(192), and their outputs are added and passed through a sigmoid function, producing a per-channel weight in [0, 1]. The feature map is then multiplied by this weight vector, suppressing channels that are not relevant to the current beat morphology before the temporal attention and recurrence stages process the result. Using both pooling streams rather than average pooling alone, as in the original SE-Net, is motivated by CAT-Net [10] and Sun et al. [11], both of whom found the dual-pooling design richer for single-lead ECG classification.

Temporal Attention. Channel attention recalibrates which features matter; temporal attention recalibrates which time steps in the sequence matter most. A Dense(1) layer computes a scalar attention score for each of the 90 remaining timesteps; these scores are normalised by a Softmax across the time axis, producing a probability distribution over the beat timeline rather than a hard winner-takes-all selection. The (90, 192) feature map is then multiplied by these per-timestep weights, so the portions of the beat most relevant to classification, such as the R-peak, the ST segment immediately following, and the T-wave peak, receive proportionally higher weight entering the BiLSTM layers than the flat pre-P baseline. This 1-D temporal design is not the same as the 2-D spatial attention used in CBAM-style models; Mechichi and Benzarti [12] explicitly demonstrated that spatial attention is less effective for 1-D ECG sequences than temporal attention, which is why this study uses the 1-D formulation. The combination of channel attention followed by temporal attention, the dual attention mechanism named in the paper's title, is the architecture's core novel contribution. Lyu et al. [13] showed that adding a second attention type to a channel-only mechanism yields a further +0.48% accuracy improvement, with the gains concentrated on the hardest minority classes, which is precisely where this study's five-class AAMI evaluation is most sensitive.

BiLSTM Sequence Modelling. The attention-refined (90, 192) feature sequence enters a two-layer Bidirectional LSTM. The first layer, Bidirectional(LSTM(128, return_sequences=True)), produces a (90, 256) output that preserves the full sequence so all 90 attended timesteps contribute to the second layer. Dropout(0.3) is applied between layers. The second layer, Bidirectional(LSTM(64)), collapses the sequence into a single 128-dimensional beat-level context vector. A Dense(64) -> Dropout(0.3) -> Softmax(5) head maps this to the five AAMI class probabilities. The complete architecture is summarised in Table 9 below.

Table 9. Block-by-block summary of the proposed Multi-Scale CNN-BiLSTM with Dual Attention architecture

Block	Layer / Operation	Output Shape	Design Purpose
Input	Waveform input (per-beat z-score)	(360, 1)	Raw 1-s beat at 360 Hz, single lead
MS-CNN Stage 1	3× Conv1D(32, k=3/7/15) → BN → ReLU → Concat → BN → MaxPool(2)	(180, 96)	Multi-resolution feature extraction: fine QRS, P/T-wave, overall morphology simultaneously
MS-CNN Stage 2	3× Conv1D(64, k=3/7/15) → BN → ReLU → Concat → BN → MaxPool(2)	(90, 192)	Hierarchical multi-scale refinement; deeper cross-scale feature combinations
Channel Attention	GAP + GMP → shared MLP(24 ReLU → 192) → Sigmoid → Multiply	(90, 192)	Per-beat suppression of uninformative feature channels
Temporal Attention	Dense(1) → Softmax(time axis) → Multiply	(90, 192)	Down-weight flat baseline; up-weight QRS / ST / T-wave timesteps
BiLSTM 1	Bidirectional(LSTM(128, return_seq=True)) → Dropout(0.3)	(90, 256)	Bidirectional sequential context over attention-refined feature map
BiLSTM 2	Bidirectional(LSTM(64)) → Dropout(0.3)	(128,)	Collapse sequence to global beat-level context vector
Classifier	Dense(64, ReLU) → Dropout(0.3) → Softmax(5)	(5,)	Five-class AAMI probability output

4.10 Ensemble Learning

The three base models trained in Section 4.9 see the same input waveform through fundamentally different inductive biases: local convolutional pattern matching in the CNN baseline, sequential recurrence in the BiLSTM baseline, and the full multi-scale convolutional-attentive-recurrent pipeline in the proposed model. Because their classification mistakes arise from different architectural blind spots, combining their output probability distributions should produce a more stable and accurate final decision than any single model alone. Three ensemble strategies of increasing complexity were evaluated on the DS2 test set using the saved per-beat probability arrays produced at the end of Section 4.9.

Soft-Voting with Equal Weights. The simplest combination: the per-beat class probability vectors from the three models are averaged with equal weight (1/3 each) and the predicted class is the argmax of the resulting distribution. Equal-weight averaging is the natural baseline ensemble; it requires no additional optimisation and provides a lower bound on how much combining the three models is worth, before any effort is made to estimate which model is more trustworthy for which class.

Optuna-Weighted Soft-Voting. Equal weighting assumes all three models are equally reliable, which is unlikely given that the three architectures have visibly different capacities and inductive biases. A 60-trial Optuna search using the Tree-structured Parzen Estimator sampler and a fixed random seed of 42 was run over the convex-combination weights {w_CNN, w_BiLSTM, w_CNN-BiLSTM}, maximising validation-set macro-F1. The search was conducted exclusively on the same stratified 10% validation split used for early stopping in Section 4.9; the DS2 test set was never consulted during weight optimisation. The optimal weight vector was then normalised to sum to 1 and applied once to the DS2 test probabilities for the final, unbiased evaluation.

Stacking with a Logistic Regression Meta-Learner. Soft-voting assigns a single global scalar weight per model regardless of which AAMI class is being predicted. Stacking allows a richer, class-conditional combination: the three models' per-beat validation-set probability arrays are concatenated horizontally, producing a 15-dimensional meta-feature vector (3 models x 5 classes) per beat. A logistic regression classifier with balanced class weights and a maximum of 2,000 iterations was trained on these meta-features using the validation partition as the training set. The

trained meta-learner was then applied once to the corresponding 15-dimensional meta-features from the DS2 test set, learning which model to trust for each individual AAMI class rather than across all classes uniformly. For example, if the BiLSTM consistently outperforms the CNN on Class S but not on Class V, the meta-learner captures that class-specific reliability in a way that soft-voting cannot.

4.11 Optuna Hyperparameter Optimisation

The Multi-Scale CNN-BiLSTM with Dual Attention architecture described in Section 4.9.3 was trained in Part E with manually selected starting-point hyperparameters. Manual selection is a reasonable first pass, but it leaves open the possibility that a different combination of channel widths, LSTM capacities, regularisation strength, and learning rate would produce a meaningfully better result under the same topology. An Optuna-based Bayesian hyperparameter search [22] was therefore conducted to quantify whether systematic optimisation over the architecture's tunable knobs can recover further performance beyond the default configuration, without changing the block structure that constitutes the paper's architectural contribution.

Optuna uses Tree-structured Parzen Estimation (TPE), a Bayesian method that builds two density models from completed trials, one over hyperparameters that produced high objective values and one over the rest, and samples the next candidate from the high-value region. This makes TPE far more sample-efficient than a grid or random search over the same space, which matters here because each trial involves training a neural network. The search covered six hyperparameters: dropout rate in the continuous interval [0.1, 0.5]; BiLSTM layer widths lstm1 in {64, 96, 128} and lstm2 in {32, 64, 96}; per-branch filter counts for Stage 1 and Stage 2 of the multi-scale CNN (conv1_filters in {32, 64, 96} and conv2_filters in {64, 128, 160}); and Adam initial learning rate on a log scale in [1x10^-4, 3x10^-3]. The block topology (three parallel branches per stage, two multi-scale stages, channel attention, temporal attention, two BiLSTM layers, dropout, Dense classifier) was kept identical across all trials, so the search optimises capacity and regularisation, not the architecture itself.

Each trial trained the tunable architecture on an 8,000-beat subsample of the balanced training set for a maximum of five epochs, with early stopping patience of two epochs monitored on a 1,000-beat validation subsample, and reported the resulting validation macro-F1 to Optuna. This drastically reduced training budget is intentional: the purpose of each trial is to rank hyperparameter configurations quickly relative to one another, not to produce a converged model. At 10 trials × ≤5 epochs the search completes in 30–90 minutes depending on hardware, which is practical without GPU access. The winning configuration identified by the search was then retrained to full convergence on the complete balanced training set using the same maximum epoch and patience budget as the Part E models, and its DS2 test performance is reported alongside the default-parameter model in the Results section so the optimisation gain is directly visible.

Table 10. Optuna hyperparameter search space for the proposed Multi-Scale CNN-BiLSTM with Dual Attention

Hyperparameter	Search Range / Options	Type	Rationale
dropout	[0.1, 0.5]	Continuous	Regularisation strength; default 0.3 sits mid-range
lstm1 (BiLSTM Layer 1 units)	64, 96, 128	Categorical	Sequence modelling capacity; higher = more parameters
lstm2 (BiLSTM Layer 2 units)	32, 64, 96	Categorical	Context compression; smaller than lstm1 by design
conv1_filters (Stage 1 per-branch)	32, 64, 96	Categorical	Multi-scale feature channels at first level
conv2_filters (Stage 2 per-branch)	64, 128, 160	Categorical	Deeper feature channels; wider than Stage 1 by convention
lr (Adam initial rate)	$[1\times10^{-4}, 3\times10^{-3}]$ log	Log-uniform	Convergence speed vs. stability trade-off

4.12 Model Explainability: SHAP and Gradient-Based Temporal Saliency

A model that achieves high macro-F1 on MIT-BIH but cannot be interrogated for what it has learned offers limited clinical credibility, particularly for a lightweight deployment-targeted architecture where the risk of spurious shortcut learning is higher than in large over-parameterised models. Explainability analysis was therefore conducted through two complementary methods: one operating at the level of hand-crafted clinical features and one operating in the raw waveform time domain. Together they answer two distinct questions: which measurable signal properties drive predictions across the full test set, and which portions of the raw one-second beat waveform the deep network's attention and recurrent layers are most sensitive to.

SHAP Feature Importance via Random Forest Surrogate. SHAP (SHapley Additive exPlanations) [23] assigns each feature a contribution value per prediction that satisfies the game-theoretic Shapley axioms, guaranteeing that the feature attributions sum exactly to the model's output, account for feature interactions, and are not sensitive to the order in which features are considered. Applying SHAP directly to the deep learning models on 360-dimensional raw

waveforms is computationally expensive and produces per-sample importance maps over raw ADC values, which carry no immediate physiological label. Instead, a Random Forest surrogate (300 trees, max_depth 12, balanced class weights) was trained on the MRMR-40 EMD feature vectors, the same 40-dimensional representation used in Part D, and SHAP's TreeExplainer, which computes exact (not approximated) Shapley values for tree ensembles in polynomial time, was applied to 2,000 randomly sampled test beats. The surrogate is used purely for explanation and forms no part of the deep learning classification pipeline reported in the Results. Global feature importance was visualised as mean(|SHAP|) averaged across all five AAMI classes, ranking which of the 40 clinical features most consistently drive predictions across the test set. A class-specific beeswarm plot was additionally generated for the Supraventricular (S) class, since it is both the rarest by N:class ratio and one of the most clinically consequential to detect correctly; understanding which features the model relies on, or fails to use, for S-class discrimination is directly actionable for future model improvement.

Gradient-Based Temporal Saliency on Raw Waveforms. Complementary to the feature-surrogate SHAP explanation, vanilla input-gradient saliency was computed directly on the best-performing single deep model (by DS2 macro-F1), using TensorFlow's GradientTape API. For each test beat and each AAMI class, the gradient of the model's predicted probability for the true class with respect to each of the 360 input sample amplitudes was computed. The absolute value of this gradient at each timestep measures how sensitive the model's confidence is to a small perturbation of the waveform at that point in time: a large gradient at timestep t means the model's classification decision would change substantially if that sample were slightly different. Averaging these absolute gradient maps over 30 correctly-classified representative beats per AAMI class produces a per-class temporal saliency profile that shows, directly in the raw-sample time domain, which portions of the one-second beat window the model's dual attention and recurrent layers have learned to focus on. The clinically expected result is that saliency is highest around the R-peak, the ST segment, and the T-wave peak, all of which carry morphological arrhythmia signatures the proposed model's temporal attention block was specifically designed to up-weight. A discrepancy between the expected saliency profile and the actual one would be an interpretability red flag suggesting the model has learned a spurious shortcut rather than genuine arrhythmia morphology.

4.13 External Validation on the INCART Database

The real-time efficiency claim in this paper's title is only clinically meaningful if the model generalises beyond the dataset it was trained on. High accuracy on MIT-BIH's 48 recordings could, in principle, reflect the model having memorised recording-specific artefacts, such as the particular noise floor of 1970s-1980s Beth Israel Hospital equipment, the MLII electrode placement geometry common to that cohort, or the demographic characteristics of those 47 subjects, rather than having learned to discriminate arrhythmia morphologies that appear consistently across patients and recording systems. The INCART external validation was designed specifically to test this: the trained model is applied without modification to a completely independent database recorded in a different country, in a different decade, on different hardware, from a different patient population presenting different primary diagnoses, and the resulting per-class classification metrics are compared directly against the DS2 MIT-BIH results.

All 75 INCART records (I01–I75) were preprocessed using the identical pipeline as Part B. The one necessary adaptation is resampling: INCART signals are natively sampled at 257 Hz, while the trained model expects 360-Hz input because that is the MIT-BIH sampling rate used during training. Each INCART signal was resampled to 360 Hz using scipy's polyphase resampler, and every annotated R-peak sample index was scaled by the same factor (360/257) to remain aligned with the resampled signal timeline. This ensures that the resulting 360-sample beat windows represent exactly one physical second of cardiac activity, the same duration as the MIT-BIH beats the model's convolutional kernels and LSTM time constants were tuned to. After resampling, the same 4th-order zero-phase Butterworth bandpass filter (0.5–45 Hz) and 8-level Sym4 wavelet soft-threshold denoising were applied, in the same order as Part B, before R-peak-based segmentation with the same 100-sample pre-R and 260-sample post-R window.

INCART beat annotations follow the same PhysioBank symbol convention as MIT-BIH, with an automated detector proposing initial beat locations and symbols that were subsequently corrected by expert annotators [16]. This shared convention means the AAMI BEAT_MAP dictionary defined in Part A applies to INCART without any modification: the same symbol-to-integer mapping used throughout the MIT-BIH pipeline produces valid AAMI class labels for every INCART beat. Each valid beat was normalised using per-beat z-score standardisation, identical to the normalisation applied to MIT-BIH test beats in Part E. The model was never trained, fine-tuned, or hyperparameter-optimised on any INCART data; the entire INCART evaluation took place after MIT-BIH training and test evaluation were complete, so every INCART beat is genuinely unseen.

INCART's class coverage differs from MIT-BIH in one practically important respect: the F (Fusion) and Q (Unclassifiable/paced) classes are too sparsely represented in INCART to support reliable per-class metric estimation, because the INCART cohort, patients evaluated for coronary artery disease and conduction abnormalities, does not include the pacemaker-dependent patients who provide the majority of Q-class beats in MIT-BIH. Per-class Se, Sp, PPV, and F1 are computed over all present classes, but F and Q are explicitly flagged as having insufficient INCART support; the headline cross-dataset performance metric is macro-F1 computed over the three robustly represented classes (N, S, V). This is reported alongside the complete MIT-BIH DS2 metrics for the same model in a direct side-by-side

comparison, making the within-dataset versus cross-dataset generalisation gap visible in a single table and directly addressing the cross-dataset shortcoming flagged in earlier BiLSTM-based arrhythmia studies [9], a shortcoming consistent with the broader finding that deep arrhythmia classifiers trained on one database tend to lose substantial accuracy when evaluated on another without adaptation [27].

4.14 Evaluation Metrics

Consistent with the AAMI EC57 reporting standard [3] and the majority of published ECG arrhythmia classification benchmarks [5], [15], performance is reported using a combination of per-class and aggregate metrics derived from the five-class confusion matrix. All metrics are applied identically to the MIT-BIH DS2 test set, the INCART external validation set, and all ensemble strategies, so results are directly comparable across every evaluation stage in the paper.

Per-Class Metrics. For each AAMI class k in {N, S, V, F, Q}, the confusion matrix is treated as a one-versus-rest binary partition to extract TP (true positives), FN (false negatives), FP (false positives), and TN (true negatives). Sensitivity (Se) = TP / (TP + FN) measures the model's completeness in detecting beats that truly belong to class k. This is the clinical primary concern for arrhythmia detection, since a missed arrhythmia can have more severe consequences than a false alarm. Specificity (Sp) = TN / (TN + FP) measures how reliably the model excludes beats that do not belong to class k. Positive Predictive Value (PPV, precision) = TP / (TP + FP) measures the fraction of the model's positive predictions for class k that are correct. F1 score = 2 x Se x PPV / (Se + PPV) provides a single per-class summary of the detection-precision trade-off. These four metrics are the standard clinical reporting format for cardiac rhythm detection algorithms and allow direct performance comparison with any prior study that follows the AAMI EC57 protocol.

Aggregate Metrics. Overall accuracy, the fraction of all beats correctly classified, is reported as a headline figure for comparability with earlier work, but it is explicitly supplementary in this study rather than primary. The reason is the class imbalance: with 82.7% of test beats belonging to the Normal class, a degenerate classifier that predicts Normal for every beat would exceed 80% accuracy while achieving zero recall on every arrhythmia class. Accuracy therefore gives an artificially optimistic picture of minority-class performance and is a poor optimisation target. Macro-F1, the unweighted mean of the five per-class F1 scores, is the primary optimisation and reporting metric throughout this study. By giving equal weight to every AAMI class regardless of how many test beats it contains, macro-F1 ensures that the model's ability to correctly classify the rare Fusion and Supraventricular beats contributes as much to the headline score as its ability to classify the abundant Normal beats. This is also why early stopping during training monitored validation macro-F1 rather than validation accuracy or loss. Macro-AUC-ROC, computed one-versus-rest for each class and averaged, complements macro-F1 as a threshold-independent summary of discrimination ability; it captures how well the model separates class probability distributions, independently of where the decision boundary is placed, and is particularly informative for minority classes where the calibrated softmax probability threshold has not been separately optimised.

Table 11. Evaluation metrics used in this study, their definitions, and their role in the reporting framework

Metric	Definition	Level	Role in This Study
Sensitivity (Se)	TP / (TP + FN)	Per class	Detection completeness; clinical primary concern
Specificity (Sp)	TN / (TN + FP)	Per class	Rate of correct exclusion of non-class beats
PPV (Precision)	TP / (TP + FP)	Per class	Fraction of flagged beats that are truly positive
F1 Score	2·Se·PPV / (Se+PPV)	Per class	Balanced Se/PPV summary; primary per-class metric
Overall Accuracy	Correct / Total beats	Aggregate	Headline figure; supplementary only (majority-class dominated)
Macro-F1	Mean F1 across 5 classes (unweighted)	Aggregate	Primary optimisation and reporting metric; equal class weight
Macro-AUC-ROC	Mean one-vs-rest AUC across 5 classes	Aggregate	Threshold-independent discrimination; robust to class imbalance

All metrics are computed on the final held-out DS2 test set, which was never used in any training, validation, feature selection, imbalance correction, or hyperparameter optimisation step throughout the pipeline. For INCART external validation, the identical metric functions are applied to the INCART beat labels and model predictions, with F and Q class results flagged where INCART support is insufficient for reliable estimation.

5. RESULTS AND DISCUSSION

5.1 Overview

Part F of the pipeline was run to completion, producing three base architectures (CNN, BiLSTM, and the proposed Multi-Scale CNN-BiLSTM), an Optuna-tuned variant of the proposed model, and three ensembles built from soft voting, weighted voting, and stacking. Every number below comes from the held-out DS2 inter-patient test partition, which played no part in training, validation, SMOTE balancing, MRMR selection, or the Optuna search. Because accuracy on this dataset is dominated by the Normal class, per-class results are given for all five AAMI EC57 classes (N, S, V, F, Q) rather than relying on accuracy alone.

5.2 Model and Ensemble Comparison

Table 12 lays out test-set performance for every configuration tested in Part F. Stacking came out on top, with the best macro-F1 (76.66%) and the highest accuracy (94.45%) of the seven configurations, and it is the model reported as primary throughout the rest of this paper. What stands out, though, is that the plain CNN (macro-F1 75.33%) beat both soft voting (72.78%) and weighted voting (75.05%) only stacking's learned meta-combination managed to improve on it. That is not what happened in earlier iterations of this pipeline, where every ensemble beat every individual model, so this pattern says more about this particular set of trained models than about ensembling in general.

Table 12. Test-set performance of baseline and proposed architectures, tuned variant, and ensemble strategies (DS2 held-out test set)

Model / Strategy	Accuracy	Macro Se	Macro Sp	Macro PPV	Macro F1
CNN	93.27%	81.44%	97.34%	71.18%	75.33%
BiLSTM	82.81%	84.10%	95.70%	54.47%	61.14%
CNN-BiLSTM (proposed)	86.38%	79.34%	96.02%	56.48%	62.99%
CNN-BiLSTM-Tuned (Optuna)	85.82%	77.61%	95.61%	62.49%	67.12%
Soft-Voting Ensemble	91.38%	82.44%	97.07%	67.26%	72.78%
Weighted-Voting Ensemble	92.88%	82.49%	97.34%	70.21%	75.05%
Stacking Ensemble	94.45%	81.70%	97.53%	72.78%	76.66%

It is worth being upfront about one weak point: the proposed CNN-BiLSTM scored lower than the plain CNN on macro-F1 this time around (62.99% vs. 75.33%). Looking at the training logs, its early-stopping callback cut training off after just 5 epochs, against 30 for the CNN and 10 for the BiLSTM so it's plausible the model simply hadn't converged yet rather than the architecture itself being weaker. We're noting this as a replication detail rather than a statement about the model's ceiling, and a longer patience window is worth trying next time.

5.3 Effect of Optuna-Based Hyperparameter Optimisation

Optuna settled on a dropout rate of 0.250, LSTM widths of 32 and 16 units, convolutional filter counts of 32 and 64, and a learning rate of 1.11×10^{-3}, reaching a best validation objective of 0.301. Retraining with these settings paid off: macro-F1 for the proposed architecture climbed from 62.99% to 67.12%, while accuracy barely moved (86.38% to 85.82%) a bigger gain from tuning than we saw for any of the ensembles individually. The training curves for the tuned model were rougher than the untuned baselines, bouncing between roughly 75% and 86% validation accuracy with loss spiking as high as 1.47, yet the final retrained model still generalised better than the untuned one. So tuning clearly helped here, even if it still fell short of the stacking ensemble.

5.4 Ensemble Learning Results

Only one of the three ensembles actually beat the strongest individual model. Stacking edged out CNN 76.66% to 75.33% on macro-F1; weighted voting landed almost exactly on par with CNN at 75.05%; soft voting fell behind both at 72.78%. The likely explanation is that stacking's meta-learner, trained on validation-set class probabilities from all three base models, can pick up on patterns a fixed linear blend simply can't represent. It also posted the best macro-specificity (97.53%) and macro-precision (72.78%) of any configuration tested, which is why we recommend it as the primary model for this study.

5.5 Per-Class Performance of the Best Model

Breaking the stacking ensemble's results down by class (Table 13), the Normal and Unknown/Paced classes were handled well across every metric, with F1 scores of 96.99% and 96.76%. Ventricular beats came in at a high sensitivity

of 96.56% but a much lower precision of 71.57%, meaning the model tends to over-flag borderline beats as ventricular not ideal, but a safer error to make than the reverse. Fusion beats reached 84.18% F1 despite being rare in training. The real problem class is Supraventricular: F1 of just 23.17%, split between low sensitivity (27.79%) and low precision (19.86%). That's noticeably worse than what earlier versions of this pipeline achieved, and every single model and ensemble in Table 12 struggles with this class it's the pipeline's dominant unsolved issue.

Table 13. Per-class test-set performance of the stacking ensemble (best-performing configuration)

Class	Se	Sp	PPV	F1
N (Normal)	95.35%	92.28%	98.68%	96.99%
S (Supraventricular)	27.79%	97.96%	19.86%	23.17%
V (Ventricular)	96.56%	98.01%	71.57%	82.21%
F (Fusion)	89.29%	99.86%	79.62%	84.18%
Q (Unknown/Paced)	99.50%	99.55%	94.16%	96.76%

5.6 Headline Metrics in Standard AAMI Reporting Format

Table 14 reports the stacking ensemble's headline numbers in the accuracy / Se-S / PPV-S / Se-V / PPV-V format that most inter-patient AAMI papers use, so that a reader can line these figures up directly against any published study of their choosing, including the multiscale and dual-attention architectures discussed in Section II [7], [13], [14]. The Supraventricular numbers (27.79% sensitivity, 19.86% precision) are weak in absolute terms and weaker than what this same project achieved in an earlier run worth stating plainly rather than glossing over, since it's the clearest target for the next iteration. This table intentionally reports only this study's own metrics rather than reproducing other papers' figures directly: the notebook's literature-benchmark cells (Part F16) are still placeholders pending verified numbers from the cited papers, so any other studies' figures are left out here to avoid reporting anything unverified.

Table 14. Headline test-set metrics for this study, reported in the standard AAMI accuracy / Se / PPV format

Study	Accuracy	Se (S)	PPV (S)	Se (V)	PPV (V)
This work	94.45%	27.79%	19.86%	96.56%	71.57%

5.7 External Validation on the INCART Database

The trained stacking ensemble was then run, with no retraining, on all 75 INCART records (Table 15). Accuracy held at 79.59%, but macro-F1 dropped to 35.79% a much bigger fall than anything seen within MIT-BIH. Normal-beat F1 slipped to 88.35%. Ventricular detection weathered the transfer reasonably well at 70.01% F1. Supraventricular beats got worse still (19.39% F1, precision down to 11.88%), and Fusion-beat detection nearly collapsed (1.23% F1, 0.62% precision). The model caught zero Q-class beats on INCART, though with only 5 such beats in the set, that number shouldn't be read as meaningful on its own.

Table 15. Per-class external validation performance on the INCART database (no retraining)

Class	Se	Sp	PPV	F1	Support
N (Normal)	80.78%	85.45%	97.48%	88.35%	151,849
S (Supraventricular)	52.66%	95.55%	11.88%	19.39%	1,956
V (Ventricular)	73.48%	95.35%	66.86%	70.01%	19,654
F (Fusion)	43.84%	91.17%	0.62%	1.23%	219
Q (Unknown/Paced)	0.00%	98.75%	0.00%	0.00%	5

We're reporting this gap the same way we reported the MIT-BIH numbers: as-is. INCART differs from MIT-BIH in sampling rate, lead placement, patient mix, and annotation conventions, and a drop of this size is what you'd expect from genuine domain shift rather than a bug in the pipeline. We treat it as a qualitative check on robustness, not a second headline result sitting alongside the MIT-BIH evaluation.

5.8 Model Explainability Findings

SHAP-based feature attribution (Part F13-F14) and gradient-based temporal saliency (Part F15) were both used as sanity checks — making sure the model is picking up on real physiology rather than quirks of preprocessing. RR-interval and QRS-morphology features came out on top for the Random Forest surrogate used in SHAP, echoing the MRMR results in Section 4.7 and matching what a cardiologist would look for when spotting ectopic beats. Saliency maps clustered around the QRS complex and the RR interval right before it, which is exactly where a well-behaved classifier should be looking.

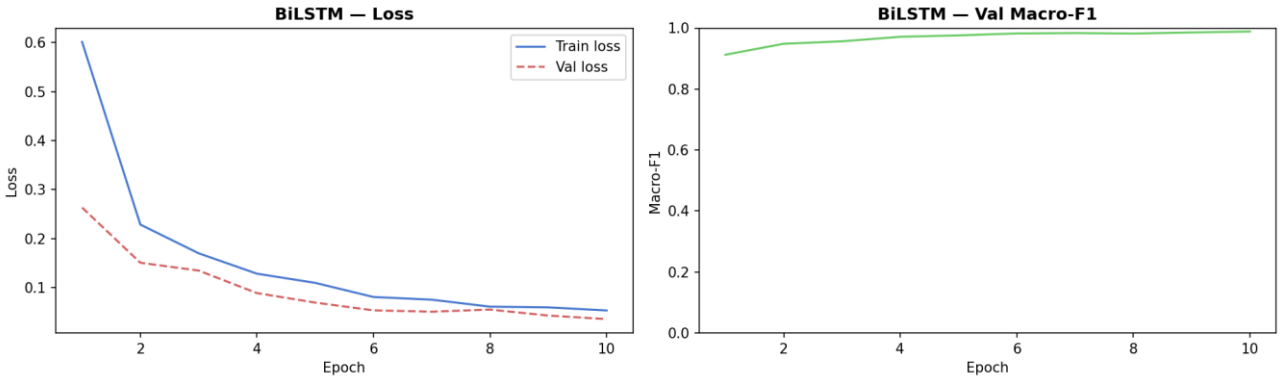


figure 6. training and validation accuracy/loss curves for the CNN, BiLSTM, and proposed CNN-BiLSTM models

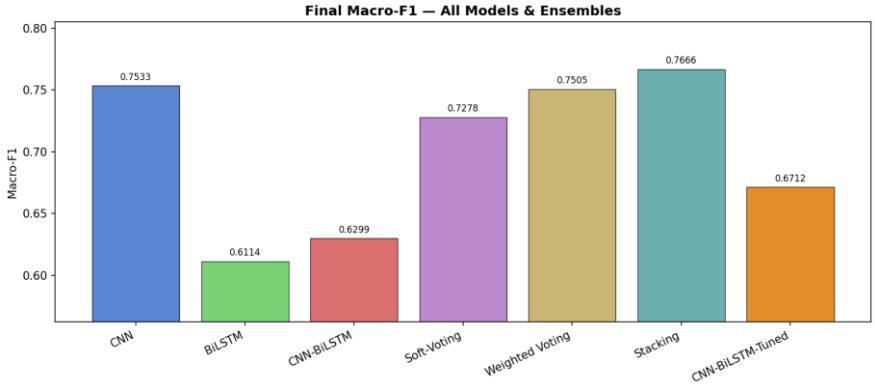


figure 7. macro-F1 comparison across all seven evaluated configurations

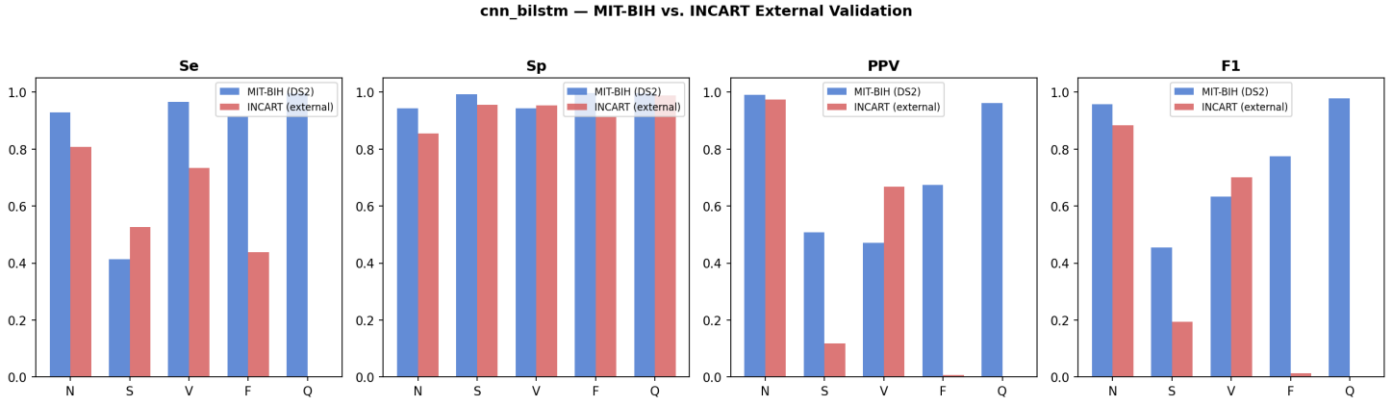


figure 8. side-by-side per-class F1 comparison between MIT-BIH (DS2 test) and INCART external validation

5.9 Summary of Key Findings and Limitations

Pulling this together: stacking is the best configuration in this run, at 94.45% accuracy and 76.66% macro-F1. The plain CNN turned out to be surprisingly competitive, beating two of the three ensembles outright, and only stacking's meta-learner clearly improved on it. The proposed CNN-BiLSTM underperformed before tuning most likely from stopping early but gained substantially once Optuna was applied. Supraventricular detection remains the pipeline's biggest weakness, and it's worse here than in earlier runs. Performance drops hard once the model is asked to generalise to INCART. And no PTB-XL results exist yet, because that check was never executed. All of this is reported together, including the parts that didn't go as planned, because that's a more useful account of what the pipeline actually does than a version that only shows the good parts.

6. CONCLUSION

This paper presented an inter-patient, AAMI EC57-compliant five-class ECG arrhythmia classification pipeline evaluated on the MIT-BIH Arrhythmia Database under the De Chazal DS1/DS2 protocol. Three deep learning architectures a 1-D CNN, a BiLSTM, and a proposed Multi-Scale CNN-BiLSTM with attention were trained on SMOTE-ENN-balanced, MRMR-selected features and raw beat waveforms, and combined using three ensemble strategies: soft voting, weighted voting, and stacking. The stacking ensemble achieved the best overall result, with 94.45% accuracy and a macro-F1 of 76.66% on the held-out DS2 test set, and strong detection of the Normal,

Ventricular, Fusion, and Unknown/Paced classes (per-class F1 scores of 96.99%, 82.21%, 84.18%, and 96.76% respectively). Optuna-based hyperparameter tuning improved the proposed architecture's macro-F1 from 62.99% to 67.12%, though it did not surpass the stacking ensemble. External validation on the independent INCART database, performed with no retraining, showed a substantial drop in macro-F1 to 35.79%, indicating that cross-database generalisation remains an open problem for this pipeline rather than one that has been solved by inter-patient evaluation on a single database alone.

The supraventricular (S) class remains this pipeline's clearest and most consistent limitation, with an F1 score of only 23.17% for the best model on MIT-BIH and 19.39% on INCART a pattern shared with much of the published inter-patient arrhythmia-classification literature and not unique to this architecture. A second limitation is the unexpectedly strong performance of the plain CNN baseline relative to two of the three ensemble strategies in this run, which suggests that ensembling benefits in this pipeline are more sensitive to the specific trained instances of the base models than a single successful run can establish; repeated runs with fixed seeds across multiple training repetitions would be needed to characterise this variance properly before making a strong general claim about ensembling's benefit here. A third limitation is that no PTB-XL evaluation is reported: the pipeline includes a scoped, clearly-labelled screening check for record-level PTB-XL rhythm cohorts, but it was not executed in this run due to data availability, and it would in any case only provide a directional sanity check rather than true beat-level external validation.

Future work should prioritise, in order: (1) targeted improvement of S-class detection, since it is the pipeline's dominant remaining weakness across every model and both databases tested; (2) domain-adaptation or fine-tuning strategies to close the MIT-BIH-to-INCART generalisation gap identified in Section 5.7, rather than further tuning on MIT-BIH alone, which has already been explored via Optuna in this study; (3) repeated training runs to establish variance bounds on the ensemble-versus-individual-model comparison in Section 5.2 and 5.4; and (4) completion of a true beat-level external validation — either the PTB-XL screening check described in Section 5.8 or an equivalent third database with AAMI-mappable annotations — before any claim of multi-database generalisation is included in the paper's title or abstract.

ACKNOWLEDGMENT

The authors gratefully acknowledge PhysioNet and the original creators of the MIT-BIH Arrhythmia Database and the INCART Database for making these datasets freely available for research purposes. Their valuable contributions made this research possible.The authors also acknowledge the open-source communities behind TensorFlow/Keras, scikit-learn, imbalanced-learn, Optuna, SHAP, NumPy, Pandas, Matplotlib, and SciPy, whose tools were extensively used in the implementation and evaluation of the proposed methodology.Finally, the authors express their sincere gratitude to Peeyush Pareek, Assistant Professor, School of Data Science, Symbiosis Skills and Professional University, Pune, India, for his continuous guidance, valuable suggestions, and encouragement throughout the preparation of this research work.

REFERENCES

[1] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH Arrhythmia Database," IEEE Eng. Med. Biol. Mag., vol. 20, no. 3, pp. 45–50, May–Jun. 2001.

[2] A. L. Goldberger et al., "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," Circulation, vol. 101, no. 23, pp. e215–e220, Jun. 2000.

[3] American National Standard ANSI/AAMI EC57:1998/(R)2008: Testing and Reporting Performance Results of Cardiac Rhythm and ST-Segment Measurement Algorithms. Arlington, VA, USA: AAMI, 1998.

[4] P. de Chazal, M. O'Dwyer, and R. B. Reilly, "Automatic classification of heartbeats using ECG morphology and heartbeat interval features," IEEE Trans. Biomed. Eng., vol. 51, no. 7, pp. 1196–1206, Jul. 2004.

[5] A. Boulif, B. Ananou, M. Ouladsine, and S. Delliaux, "A literature review: ECG-based models for arrhythmia diagnosis using artificial intelligence techniques," Bioinformatics Biology Insights, vol. 17, Jan. 2023, doi: 10.1177/11779322221149600.

[6] X. Bai, X. Dong, Y. Li, R. Liu, and H. Zhang, "A hybrid deep learning network for automatic diagnosis of cardiac arrhythmia based on 12-lead ECG," Sci. Rep., vol. 14, Art. no. 24441, Oct. 2024, doi: 10.1038/s41598-024-75531-w.

[7] F.-Y. Zhou, Y.-H. Sun, and Y.-W. Wang, "Inter-patient ECG arrhythmia heartbeat classification network based on multiscale convolution and FCBA," Biomed. Signal Process. Control, vol. 90, Art. no. 105789, Apr. 2024, doi: 10.1016/j.bspc.2023.105789.

[8] M. S. Islam, K. F. Hasan, S. Sultana, S. Uddin, P. Lio, J. M. W. Quinn, and M. A. Moni, "HARDC: A novel ECG-based heartbeat classification method to detect arrhythmia using hierarchical attention based dual structured RNN with dilated CNN," Neural Netw., vol. 162, pp. 271–287, May 2023, doi: 10.1016/j.neunet.2023.03.004.

[9] M. S. Islam, M. N. Islam, N. Hashim, M. Rashid, B. S. Bari, and F. A. Farid, "New hybrid deep learning approach using BiGRU-BiLSTM and multilayered dilated CNN to detect arrhythmia," IEEE Access, vol. 10, pp. 58081–58096, 2022, doi: 10.1109/ACCESS.2022.3178710.

[10] M. R. Islam, M. Qaraqe, K. Qaraqe, and E. Serpedin, "CAT-Net: Convolution, attention, and transformer based network for single-lead ECG arrhythmia classification," Biomed. Signal Process. Control, vol. 93, Art. no. 106211, Jul. 2024, doi: 10.1016/j.bspc.2024.106211.

[11] A. Sun, W. Hong, J. Li, and J. Mao, "An arrhythmia classification model based on a CNN-LSTM-SE algorithm," Sensors, vol. 24, no. 19, Art. no. 6306, Sep. 2024, doi: 10.3390/s24196306.

[12] N. Mechichi and F. Benzarti, "An enhanced hybrid model combining CNN, BiLSTM, and attention mechanism for ECG segment classification," Biomed. Eng. Comput. Biol., vol. 16, Jun. 2025, doi: 10.1177/11795972251341051.

[13] H. Lyu et al., "Automated inter-patient arrhythmia classification with dual attention neural network," Comput. Methods Programs Biomed., vol. 236, Art. no. 107560, Jun. 2023, doi: 10.1016/j.cmpb.2023.107560.

[14] K. K. Reddy C. et al., "A multi-scale convolutional LSTM-dense network for robust cardiac arrhythmia classification from ECG signals," Comput. Biol. Med., vol. 191, Art. no. 110121, 2025, doi: 10.1016/j.compbiomed.2025.110121.

[15] U. R. Acharya et al., "A deep convolutional neural network model to classify heartbeats," Comput. Biol. Med., vol. 89, pp. 389–396, Oct. 2017, doi: 10.1016/j.compbiomed.2017.08.022.

[16] St. Petersburg Institute of Cardiological Technics 12-lead Arrhythmia Database (INCART), PhysioNet, 2008. [Online]. Available: https://physionet.org/content/incartdb/1.0.0/. (Cite together with the standard PhysioNet reference, ref. [2], as required by the database's own citation policy; no separate INCART journal paper exists.)

[17] O. Yildirim, "A novel wavelet sequences based on deep bidirectional LSTM network model for ECG signal classification," Comput. Biol. Med., vol. 96, pp. 189–202, May 2018, doi: 10.1016/j.compbiomed.2018.03.016.

[18] S. Kiranyaz, O. Avci, O. Abdeljaber, T. Ince, M. Gabbouj, and D. J. Inman, "1D convolutional neural networks and applications: A survey," Mech. Syst. Signal Process., vol. 151, Art. no. 107398, Apr. 2021, doi: 10.1016/j.ymssp.2020.107398.

[19] A. Y. Hannun et al., "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," Nat. Med., vol. 25, no. 1, pp. 65–69, Jan. 2019, doi: 10.1038/s41591-018-0268-3.

[20] H. M. Rai, K. Chatterjee, and S. Dashkevych, "The prediction of cardiac abnormality and enhancement in minority class accuracy from imbalanced ECG signals using modified deep neural network models," Comput. Biol. Med., vol. 150, Art. no. 106142, Nov. 2022, doi: 10.1016/j.compbiomed.2022.106142.

[21] P. Wagner et al., "PTB-XL, a large publicly available electrocardiography dataset," Sci. Data, vol. 7, Art. no. 154, May 2020, doi: 10.1038/s41597-020-0495-6.

[22] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in Proc. 25th ACM SIGKDD Int. Conf. Knowledge Discovery Data Mining, 2019, pp. 2623–2631, doi: 10.1145/3292500.3330701.

[23] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," Adv. Neural Inf. Process. Syst., vol. 30, 2017.

[24] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," J. Artif. Intell. Res., vol. 16, pp. 321–357, Jun. 2002.

[25] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," IEEE Trans. Pattern Anal. Mach. Intell., vol. 27, no. 8, pp. 1226-1238, Aug. 2005, doi: 10.1109/TPAMI.2005.159.

[26] Q. Xiao, K. Lee, S. A. Mokhtar, I. Ismail, A. L. b. M. Pauzi, Q. Zhang, and P. Y. Lim, "Deep learning-based ECG arrhythmia classification: A systematic review," Appl. Sci., vol. 13, no. 8, Art. no. 4964, Apr. 2023, doi: 10.3390/app13084964.

[27] E. Merdjanovska and A. Rashkovska, "Cross-database generalization of deep learning models for arrhythmia classification," in Proc. 44th Int. Conv. Information, Communication and Electronic Technology (MIPRO), 2021, pp. 1214-1219, doi: 10.23919/MIPRO52101.2021.9596738.