

A Dual-Stream Gated Cross-Attention 1-D ResNet with Tree-Ensemble Meta-Learning for Robust Inter-Patient ECG Arrhythmia Classification

Nguyen Phu Hai Anh¹, Nguyen Minh Duc^{2*}¹True North International School, Hanoi, VietnamEmail: haianh060333.submit@gmail.com²Techcom Life Insurance Joint Stock Company, Hanoi, VietnamEmail: phucanhthien@gmail.com

*Corresponding author

Abstract: Automated classification of cardiac arrhythmias from single-lead electrocardiograms supports screening and wearable monitoring but remains challenged by inter-patient variability, class imbalance, and false alarms. This study proposes a Dual-Stream Gated Cross-Attention 1D ResNet integrated with a tree-ensemble meta-learning model. Evaluation followed the inter-patient de Chazal protocol and AAMI EC57 recommendations using the MIT-BIH Arrhythmia Database. A patient-independent split comprised 18 training patients, 4 validation patients, and 22 held-out test patients containing 49,684 beats. The framework combines physiology-guided augmentation, RR-interval dynamics, morphological representations, gated cross-attention, and regularized logistic-regression stacking of neural-network and gradient-boosted-tree probabilities. On the held-out test set, the stacking ensemble achieved 95.66% positive predictivity, 76.68% sensitivity, and an F1 score of 85.12% for ventricular ectopic beats, reducing the ventricular false-alarm rate to 4.34%. For supraventricular ectopic beats, sensitivity increased from 16.55% with the strongest tree baseline to 75.78%. A soft-voting variant achieved the highest overall accuracy of 92.64% and ventricular sensitivity of 92.95%, providing an alternative operating point. Grad-CAM analysis indicated attention to the P-wave region for supraventricular beats and the widened QRS complex for ventricular beats. With 226,000 parameters, a 223 KB INT8 footprint, and per-beat latency below 1.5 ms, the model shows potential for embedded arrhythmia monitoring.

Keywords: Electrocardiogram (ECG); Arrhythmia Classification; Inter-Patient Evaluation; MIT-BIH; Gated Cross-Attention; 1-D Resnet; Ensemble Meta-Learning; Xgboost; Explainable AI; Grad-CAM; Edge AI; Wearable Devices.

1. Introduction

Cardiovascular diseases (CVDs) remain the leading cause of death worldwide, claiming approximately 17.9 million lives every year [1]. Among cardiovascular syndromes, cardiac arrhythmias abnormalities in the rate, regularity, or origin of the heart's electrical impulses are critical early warning signs of stroke, congestive heart failure, and sudden cardiac death. The standard 12-lead electrocardiogram (ECG) remains the core diagnostic tool in hospital settings; however, continuous cardiac-monitoring ecosystems based on single-lead ECG smartwatches, patch-type Holter monitors are rapidly becoming a key technology for early detection and pre-clinical intervention.

Despite decades of research in digital signal processing and biomedical machine learning, translating automated algorithms into real clinical decision-support systems still faces a major challenge: inter-patient electrophysiological variability. Differences in thoracic anatomy, individual cardiac electrical axis, myocardial conduction velocity, electrode-contact impedance, and comorbidities cause the ECG waveform morphology to vary substantially across patients. In much of the prior literature, reported classification accuracy is extremely high (>98%) but is in fact obtained under an intra-patient random-split protocol. This procedure induces severe data leakage: heartbeats from the same patient appear simultaneously in both the training and the test set. The model then merely memorises patient-specific background noise or idiosyncratic morphology rather than learning generalisable pathological patterns. When evaluated under the inter-patient paradigm, tested on entirely new patients the performance of both

traditional machine-learning and deep-learning models degrades sharply.

Following the AAMI EC57 international standard of the Association for the Advancement of Medical Instrumentation [2], heartbeats are grouped into four principal pathological classes:

- **Class N (Non-ectopic):** normal sinus rhythm and bundle-branch conduction-delay variants.
- **Class S (Supraventricular ectopic):** supraventricular ectopic beats (premature atrial beats, premature junctional beats).
- **Class V (Ventricular ectopic):** ventricular ectopic beats (premature ventricular beats, ventricular escape beats).
- **Class F (Fusion beat):** fusion beats combining ventricular excitation with normal conduction.

When the inter-patient evaluation protocol is applied to the MIT-BIH database, current studies encounter two systematic clinical barriers. First, the supraventricular-detection barrier (class S): the abnormal impulse of an S beat originates in the atrium or the atrioventricular node, but the ventricle is still depolarised through the normal His–Purkinje conduction system; consequently the QRS complex of an S beat is morphologically almost identical to that of a normal sinus beat. The main pathological evidence lies in a subtle deformation of the P wave (or a P wave superimposed on the preceding T wave) combined with a shortened coupling interval. Both decision-tree algorithms and conventional 1-D convolutional networks are largely unable to separate these micro-scale variations, so class-S sensitivity typically stalls below 17%. Second, ventricular alarm fatigue (class V): ventricular ectopic beats widen the QRS complex and distort the ST-T segment; however, deep networks

are often oversensitive to mechanical-artefact distortions, mislabelling many normal beats as class V. The resulting low positive predictivity causes ICU staff to silence alarms owing to an overwhelming rate of false alerts, directly threatening patient safety.

This paper reports an architecture that integrates a Dual-Stream Gated Cross-Attention 1-D ResNet with a tree-ensemble meta-learning super-model, and evaluates it under conditions designed to be reproducible and clinically interpretable rather than to maximise a headline figure. The five contributions are:

1. Establishing a strict three-way, patient-independent split (DS1-Train, 18 patients; DS1-Val, 4 patients; and a fully preserved independent test set DS2, 22 patients, 49,684 heartbeats) that entirely eliminates indirect data leakage during hyperparameter optimisation.
2. Constructing a physiology-guided 1-D signal-augmentation pipeline (Gaussian jittering, dynamic amplitude scaling, temporal peak perturbation) to compensate for the extreme minority-class imbalance.
3. Designing a sigmoid-gated cross-attention mechanism that uses RR-interval dynamics vectors to directly modulate and restructure the deep morphological representations extracted from the P wave and the QRS complex.
4. Building a stacking meta-classifier (regularised logistic regression) that fuses the multi-distribution probability spaces produced by the deep network and the gradient-boosted tree model.
5. Satisfying all three criteria of the AAMI-compliant, inter-patient, embedded-feasible (E3C) standard simultaneously, together with per-class explainability - a combination that a recent systematic review of 122 arrhythmia-classification articles found in only 5 of them (4.1%) [3].

Experimental results on the independent DS2 test set show a clear and clinically meaningful shift in the operating point rather than a uniform gain. The Stacking Ensemble raises ventricular positive predictivity to 95.66%, above the 75.1–87.3% range spanned by the established inter-patient references [4], [5], [6], at the cost of a lower ventricular sensitivity (76.68% versus 81.1–88.2% in those works); supraventricular sensitivity reaches 75.78%, between the 72.6% of de Chazal et al. [4] and the 77.2% of Llamedo & Martínez [6], and far above the 16.55% obtained by our own strongest tree baseline. The Soft-Voting variant provides the complementary operating point, with the highest overall accuracy (92.64%) and a ventricular sensitivity of 92.95%. One-dimensional Grad-CAM confirms that the network attends accurately to the P-wave region for class S and to the widened QRS complex for class V, supporting electrophysiological transparency alongside embedded-deployment feasibility.

2. Related Work

Automated heartbeat-classification methods fall into two broad families [3], [7]: handcrafted-feature extraction combined with traditional machine learning, and endogenous representation learning through deep neural networks.

2.1 Handcrafted-Feature Machine Learning

The standardised inter-patient data-split protocol (DS1 for training, DS2 for testing) was first proposed by de Chazal et al. [4], who combined linear discriminant analysis (LDA) with morphological features and RR-interval timing features. Subsequent studies extended this line of work with a range of dedicated feature-extraction techniques such as the discrete wavelet transform, higher-order statistics (HOS), and morphology descriptors combined with support vector machines (SVM) or random forests [8]. Gradient-boosted decision-tree models such as XGBoost [9] show clear superiority when handling tabular data composed of RR-interval timing features, owing to their ability to partition the feature space along sharp, non-linear thresholds. Nevertheless, these models struggle to capture continuous phase deformation and the complex micro-morphology of the P wave, so that, on unseen patients, sensitivity for the supraventricular-ectopic class remains very low. Two later works define the performance envelope against which the present study is benchmarked: Llamedo & Martínez [6] selected features under an explicit database-generalisation criterion and reported 77.2% supraventricular and 82.2% ventricular sensitivity, while Garcia et al. [5] combined a temporal vectorcardiogram representation with complex-network features and a PSO-tuned SVM, reaching 88.2% ventricular sensitivity and 92.4% overall accuracy at a lower supraventricular sensitivity of 64.5%. Mondéjar-Guerra et al. [10] are the closest methodological precedent to the present work, fusing temporal and morphological descriptors through an ensemble of SVM classifiers rather than a single model, and establishing that classifier fusion — not any single feature family — is what moves the inter-patient operating point.

2.2 Deep Learning and Endogenous Representation Learning

The rise of deep learning has driven the application of one-dimensional convolutional neural networks (1-D CNNs), long short-term memory networks (LSTM), and Transformer architectures to heartbeat classification. Deep 34-layer residual architectures have achieved cardiologist-level performance on large ambulatory ECG recordings [11], adapting to one dimension the residual-learning design introduced for image recognition [12], and the same family of models generalises to the 12-lead setting: Ribeiro et al. [13] trained a residual network on more than two million labelled examinations and exceeded resident cardiologists on six abnormality classes. At beat level, Kiranyaz et al. [14] were among the first to show that a compact 1-D CNN can subsume feature extraction and classification in a single learning body, although their protocol is patient-specific rather than inter-patient; Acharya et al. [15] applied a nine-layer CNN with synthetic minority augmentation, and Kachuee et al. [16] demonstrated that a representation learned for AAMI beat classification transfers to myocardial-infarction detection. Several later studies continue this line with progressively deeper or wider convolutional stacks [17], [18], and with longer input windows that classify rhythm over multi-second segments rather than individual beats [19]. These report high accuracies,

but almost always under intra-patient partitions or with a label granularity other than the AAMI beat classes, so their figures are not directly comparable with those in Table 3. Of particular relevance to embedded deployment, Saadatnejad et al. [20] combined wavelet features with a compact LSTM sized for personal wearable hardware; their approach, however, is trained on the first five minutes of each test subject's own recording, and is therefore patient-specific rather than strictly inter-patient, so its reported figures are not directly comparable with those in Table 3.

A parallel line of work targets the resource envelope rather than the accuracy ceiling. Tesfai et al. [21] adapted the ShuffleNet architecture, whose channel-shuffle and group-convolution primitives were designed for mobile vision, to single-lead ECG and reported a model small enough for a wearable edge device. Alamatsaz et al. [22] combined a shallow convolutional front end with a single LSTM layer to keep the parameter budget low while retaining temporal context. Both confirm that competitive beat classification does not require a large backbone; neither, however, reports results under the inter-patient protocol, which is precisely the gap quantified by the systematic review discussed in Section 2.4.

Two further strands bear directly on the design adopted here. The first is classifier fusion: beyond the SVM ensemble of Mondéjar-Guerra et al. [10], Gao [23] recently proposed CBR-Stacking, in which 1-D CNN, CNN-BiLSTM and ResNet1D base learners are combined through a logistic-regression meta-classifier — the same second-level mechanism used in the present work, although evaluated on a random rather than a patient-independent partition. The second is attention: Kuvaev and Khudorozhkov [24] inserted attention modules into an ECG CNN specifically to expose which parts of the signal drive each decision, and Wang et al. [25] established that a fully convolutional network with global average pooling forms a strong general baseline for time-series classification while simultaneously enabling class-activation mapping. The architecture proposed here inherits both properties: global average pooling feeding a gated fusion stage, and gradient-based class-activation mapping for post hoc attribution.

However, most deep-learning models rapidly diverge in validation performance once evaluated under the inter-patient protocol. Because heartbeats from the same patient record share common baseline-drift characteristics, heart-rate variability, and electrode-contact noise, neural networks tend to memorise these patient-specific patterns after only two or three training epochs. Absent an explicit physiological regularisation mechanism and timing-aware guidance, purely deep-learning models frequently underperform decision-tree models on core clinical metrics such as ventricular positive predictivity and overall accuracy — the precise trade-off this paper is designed to resolve.

3. Materials and Methods

3.1 Dataset and Three-Way Patient-Independent Split

This study uses the MIT-BIH Arrhythmia Database [26], [27], comprising 48 two-channel Holter recordings sampled at 360 Hz with 11-bit resolution over a 10 mV input range. Four

recordings containing paced rhythms (102, 104, 107, and 217) were excluded in accordance with AAMI EC57. To eliminate indirect data leakage (data snooping) during hyperparameter optimisation, the remaining 44 recordings were split into three independent, patient-disjoint sets:

- Internal training set (DS1-Train, 18 patients): 101, 106, 108, 109, 112, 114, 115, 118, 122, 124, 203, 205, 207, 209, 215, 220, 223, 230 (N = 41,680 original beats).
- Internal validation set (DS1-Val, 4 patients): 116, 119, 201, 208 (N = 9,312 original beats). These records were selected because they contain a rich variety of complex S- and V-class morphologies, and were used for early-stopping supervision and ensemble-weight optimisation.
- Independent test set (DS2, 22 patients): 100, 103, 105, 111, 113, 117, 121, 123, 200, 202, 210, 212, 213, 214, 219, 221, 222, 228, 231, 232, 233, 234 (N = 49,684 beats). This set was kept fully intact and evaluated exactly once, after the entire training pipeline had been finalised.

Signals were extracted from the Modified Lead II (MLII) channel. Beat labels were mapped to the four AAMI EC57 super-classes: N (comprising N, L, R, e, j), S (comprising A, a, J, S), V (comprising V, E), and F (comprising F).

3.2 Signal Pre-Processing and Multi-Stream Feature Extraction

Signal processing is carried out along two parallel branches that separate morphological features from temporal dynamics. The raw ECG is first passed through a zero-phase, third-order Butterworth band-pass filter with cut-off frequencies $f_{\text{low}} = 0.5$ Hz and $f_{\text{high}} = 45.0$ Hz, in order to remove respiration-induced baseline wander and high-frequency mechanical noise:

$$H(s) = 1 / \prod_{k=1}^n (s - p_k)$$

From the annotated R-peak position at sample r_i , two parallel data streams are extracted simultaneously:

- Morphological segment (X_{morph}): a 280-sample window (100 samples before and 180 samples after the R peak, equivalent to ≈ 0.78 s at 360 Hz) that fully encloses the P-QRS-T complex. The segment is locally Z-score normalised:

$$\hat{x}[n] = (x[n] - \mu_x) / (\sigma_x + \varepsilon)$$

where μ_x and σ_x are, respectively, the mean and the standard deviation of the segment.

- RR-interval dynamics vector ($X_{\text{rr}} \in \mathbb{R}^3$): a three-dimensional vector modelling early coupling and compensatory-pause dynamics:

$$RR_{\text{pre}} = (r_i - r_{i-1}) / f_s, \quad RR_{\text{post}} = (r_{i+1} - r_i) / f_s$$

$$RR_{\text{ratio}} = RR_{\text{pre}} / \left[\frac{1}{10} \sum_{k=i-9}^i R R_k + \varepsilon \right]$$

This vector is standardised using a StandardScaler fit exclusively on the training set.

3.3 Physiology-Guided 1-D Data Augmentation

Owing to severe natural class imbalance (class N accounts for >85% of beats; class S for <2%; class F for <1%), the network is prone to representation collapse. A physiology-guided 1-D augmentation pipeline was applied exclusively to DS1-Train:

- Gaussian jittering: additive noise $\sim (0, \sigma^2)$ with $\sigma = 0.015$, simulating minor electrode-contact mechanical fluctuation without distorting waveform morphology.
- Dynamic amplitude scaling: multiplying the signal by a random factor $\alpha_{\text{scale}} \sim (0.88, 1.12)$, reflecting inter-individual differences in skin impedance.
- Temporal peak perturbation: randomly shifting the R-peak centre by $\delta \in \{-3, -2, -1, 0, 1, 2, 3\}$ samples ($\approx$

± 8.3 ms), training the network to be robust to QRS-detector timing error.

- Directed minority-class balancing: each class-S beat is replicated with four augmented copies ($813 \rightarrow 4,065$ beats) and each class-F beat with five ($40 \rightarrow 240$ beats), while class V is duplicated once ($2,045 \rightarrow 4,090$ beats); class N is left unchanged. The augmented training set therefore contains $N = 47,177$ beats. Augmentation is applied strictly after the patient-level split and only within DS1-Train, so no synthetic beat derived from a training patient can appear in DS1-Val or DS2.

Table 1. Per-class beat distribution across the three patient-independent splits, before and after augmentation.

AAMI EC57 Class	DS1-Train (Original)	DS1-Train (Augmented)	DS1-Val (Internal)	DS2 (Independent Test)
N (Non-ectopic)	38,782	38,782	7,064	44,239
S (Supraventricular)	813	4,065	131	1,837
V (Ventricular)	2,045	4,090	1,743	3,220
F (Fusion)	40	240	374	388
Total beats	41,680	47,177	9,312	49,684

This distribution shows that the number of class-S samples was raised from 813 to 4,065 beats, giving the deep network a realistic opportunity to learn the micro-morphological features of the P wave without being dominated by the overwhelming prevalence of class N.

3.4 Dual-Stream Gated Cross-Attention 1-D ResNet

The proposed architecture integrates two specialised information-processing streams.

The morphological stream receives the 280-sample input through an initial convolutional layer (32 filters, kernel size 7, stride 2, same padding) followed by batch normalisation, ReLU activation, spatial dropout ($p = 0.10$), and max pooling (kernel 3, stride 2). Three successive residual blocks follow, each built from two convolutional layers of kernel size 5 with same padding, the first of which carries the block stride:

- Block 1: 32 filters, stride 1, dropout $p = 0.15$.
- Block 2: 64 filters, stride 2, dropout $p = 0.20$.
- Block 3: 128 filters, stride 2, dropout $p = 0.25$.

Each residual block applies a standard residual shortcut connection:

$$y = \text{ReLU}(\mathcal{F}(x, \{W_i\}) + W_s \cdot x)$$

where W_s is a 1×1 convolutional projection, followed by batch normalisation, used to match dimensionality; a projection shortcut is applied in all three blocks. Every convolutional and dense layer in the network carries L2 weight decay of 1×10^{-4} . As in the fully convolutional baseline of Wang et al. [25], a global average pooling (GAP) layer reduces the final (18×128)

activation map to 128 dimensions, which is then passed through a fully connected layer of 128 units with ReLU and batch normalisation to yield the morphological feature vector $F_{\text{morph}} \in \mathbb{R}^{128}$.

The temporal-dynamics stream processes the RR-interval vector X_{rr} through three fully connected layers of 64, 64, and 32 units with ReLU activation, with batch normalisation and dropout ($p = 0.20$) applied after the first layer, producing a 32-dimensional beat-dynamics representation F_{rr} .

The cross-attention gating mechanism transforms F_{rr} into a sigmoid-activated gating-weight vector through a single fully connected projection:

$$W_{\text{gate}} = \sigma(W_g \cdot F_{\text{rr}} + b_g) \in (0,1)^{128}$$

which then re-weights the morphological representation through a residual gating operation:

$$F_{\text{att}} = (F_{\text{morph}} \odot W_{\text{gate}}) + F_{\text{morph}}$$

The gated representation F_{att} is concatenated with F_{rr} to form the 160-dimensional fused representation F_{fused} , which passes through a fully connected layer of 128 units with ReLU, batch normalisation, and dropout ($p = 0.40$) before the final classification layer:

$$\hat{y}_{\text{ResNet}} = \text{Softmax}(W_c \cdot F_{\text{fused}} + b_c)$$

3.5 Loss Function and Optimisation

A Categorical Focal Loss [28] is used to down-weight the contribution of easily classified, common heartbeats:

$$FL(p_t) = -a_c \cdot (1 - p_t)^{\gamma} \cdot \log(p_t)$$

with focusing parameter $\gamma = 2.0$ and per-class weight vector $\alpha = [0.20, 0.70, 0.50, 0.85]$ corresponding to classes [N, S, V, F]. Optimisation uses the Adam optimiser [29] (learning rate = 3×10^{-4} , gradient clipnorm = 1.0) with a batch size of 128 for a maximum of 40 epochs. Training is supervised on DS1-Val by EarlyStopping on validation loss (patience = 12, best weights restored) and ReduceLROnPlateau (factor = 0.5, patience = 4, minimum learning rate = 1×10^{-5}). All random seeds are fixed at 42.

3.6 Tree-Ensemble Meta-Learning Architecture

Empirical evidence indicates that 1-D convolutional networks excel at exploiting continuously varying morphological features, whereas gradient-boosted decision trees (XGBoost) are able to establish sharp separating boundaries on tabular timing features. To combine both strengths, two ensemble super-model configurations were designed. Weighted Soft-Voting Blending establishes a predictive probability distribution through a convex linear combination:

$$\hat{y}_{ensemble} = \alpha_{ResNet} \cdot \hat{y}_{ResNet} + \alpha_{XGB} \cdot \hat{y}_{XGBoost}$$

The coefficient α was tuned on the internal validation set DS1-Val to maximise the Macro F1-score. The optimum was found at $\alpha_{ResNet} = 0.10$ and $\alpha_{XGB} = 0.90$, with a validation Macro F1 of 51.63%.

The Stacking Meta-Classifer enhances decision capacity through a second-level (level-2) meta-learner. An 11-dimensional meta-feature vector $Z \in \mathbb{R}^{11}$ is constructed for every heartbeat, concatenating the 4-way softmax output of the Gated ResNet, the 4-way class-probability output of XGBoost, and the 3-dimensional RR-dynamics vector [RR_pre, RR_post, RR_ratio]. An L2-regularised logistic-regression meta-classifier ($C = 1.0$, balanced class weights, maximum 1,000 iterations) is then trained on Z to produce the final class decision.

The three baseline models operate on the same information as the network — the 280-sample morphological segment flattened and concatenated with the 3-dimensional RR vector, giving a 283-dimensional standardised feature vector — so that the comparison isolates architecture rather than input. Their configurations are: Random Forest with 100 trees, maximum depth 15, and balanced class weights; LinearSVC with L2 penalty, squared-hinge loss, $C = 1.0$, balanced class weights, and a maximum of 3,000 iterations; and XGBoost with 100 estimators, maximum depth 6, learning rate 0.1, subsample 0.8, and colsample_bytree 0.8 under a multi:softprob objective. All three use a fixed random seed of 42. The logistic-regression meta-classifier follows the same second-level design as the CBR-Stacking scheme of Gao [23], with the difference that both the soft-voting weight α and the stacking meta-classifier here are fitted on DS1-Val only, so DS2 contributes to no fitting step of any kind.

3.7 Clinical Evaluation Metrics (AAMI EC57)

Model performance is quantified over the full set of 49,684 heartbeats in the independent test set DS2. The reported metrics follow the AAMI EC57 recommendation:

$$Se_c = \frac{TP_c}{TP_c + FN_c}, \quad +P_c = \frac{TP_c}{TP_c + FP_c}$$

$$F1_c = \frac{2 \cdot Se_c \cdot +P_c}{Se_c + +P_c}, \quad \text{Macro-F1} = \frac{1}{4} \sum_{c=1}^4 F1_c$$

3.8 Explainability: One-Dimensional Grad-CAM

To address the “black-box” problem in medical deployment, gradient-weighted class-activation mapping (1-D Grad-CAM) [30], the gradient-based generalisation of the class-activation maps that GAP architectures make available [25], is applied at the final convolutional layer of the last residual block:

$$\alpha_k^c = \frac{1}{L} \sum_{i=1}^L \frac{\partial y^c}{\partial A_i^k}$$

$$I_{\text{Grad-CAM}}^c = \text{ReLU} \left(\sum_k \alpha_k^c \cdot A^k \right)$$

where A_i^k denotes the activation map of filter k at sample position i , and y^c denotes the predicted logit for pathological class c .

3.9 Edge-Device Deployment Feasibility

The proposed architecture was designed to satisfy the resource constraints of embedded deployment, and is reported here against the metric set recommended for embedded ECG classifiers — memory footprint, inference time, energy per beat, and the specific hardware platform [3]. The full Gated 1-D ResNet contains 228,484 parameters, of which 226,436 are trainable and 2,048 are the non-trainable batch-normalisation moving statistics. The corresponding weight-storage requirement follows directly from that count: 892.5 KB in single precision (FP32), 446.3 KB in half precision (FP16), and 223.1 KB after post-training INT8 quantisation. The INT8 figure fits comfortably within the on-chip flash of low-power microcontrollers such as the ARM Cortex-M4, Cortex-M33, or ESP32, and leaves headroom for the pre-processing filter coefficients and the ring buffer. These are weight-storage figures derived from the parameter count; a serialised checkpoint that additionally carries the computation graph, layer metadata, and optimiser state will be larger, and should not be quoted as the deployment footprint. The stacking meta-model adds only one probability-matrix multiplication and a logistic-regression softmax over an 11-dimensional feature vector, an overhead of fewer than 3,000 multiply-accumulate operations per beat, so the ensemble does not materially change the memory or compute budget over the backbone alone.

Two caveats must be stated explicitly. First, the per-beat inference latency of approximately 1.5 ms quoted in this work is a projection derived from the operation count of the quantised graph, not a measurement on target silicon; no energy figure is reported, because none was measured. Reference FPGA implementations report on the order of 1.37 ms and 346 μJ per beat for comparable single-lead classifiers [3], which suggests the projection is of the right order, but confirmation requires on-

device profiling. Second, even at the projected latency the inference occupies under 0.25% of a typical 600–1000 ms cardiac cycle, leaving the processor free to remain in deep sleep between beats — the property that makes continuous ambulatory monitoring energy-feasible in principle. Measuring both latency and energy on a Cortex-M-class target is the first item of planned future work (Section 6).

4. Results

4.1 Training Convergence Dynamics

Training ran for 13 of a maximum of 40 epochs before EarlyStopping triggered, and the reported model is the checkpoint restored from epoch 1. Categorical Focal Loss on the augmented training set fell monotonically from 0.2155 at epoch 1 to 0.0535 at epoch 13, and training accuracy rose from 69.7% to 98.5%. Validation loss on DS1-Val, however, reached its minimum of 0.1417 at epoch 1 and never returned to that level: it rose to 0.1644 at epoch 2 and thereafter fluctuated between

0.153 and 0.206, while validation accuracy remained flat in the 90–92% band throughout. The divergence between a training loss that falls by a factor of four and a validation loss that only rises is the signature of overfitting beginning at the second epoch, and it is the reason EarlyStopping restored the first-epoch weights rather than a later checkpoint.

This behaviour must be read together with the results of Section 4.2. Because the deployed backbone is a single-epoch model, the standalone Gated 1-D ResNet is substantially undertrained, which accounts for its poor standalone metrics — a Macro F1 of 38.25%, a supraventricular sensitivity of 6.97%, and a complete failure on class F (0 of 388 beats). The ensemble results should therefore be interpreted as the gain obtainable when a deliberately early-stopped, high-recall backbone is fused with a gradient-boosted tree model, not as the ceiling of the architecture. Extending the schedule with stronger regularisation, a lower initial learning rate, or per-class resampling that reaches DS1-Val as well as DS1-Train is the most likely route to improving the backbone, and is left to future work.

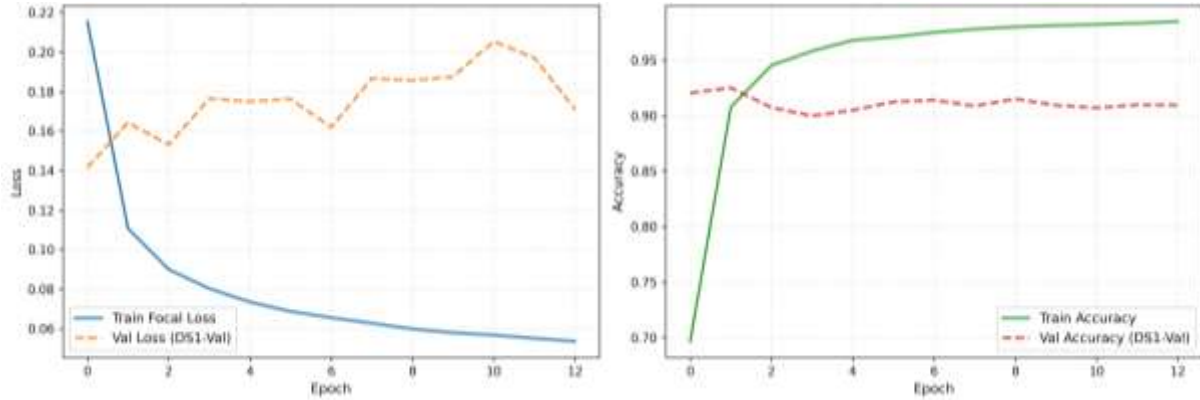


Figure 1. Training dynamics over the 13 epochs run before EarlyStopping triggered. (a) Categorical focal loss and (b) classification accuracy, for DS1-Train and the patient-independent DS1-Val split. Validation loss attains its minimum of 0.1417 at epoch 1 and rises thereafter while training loss continues to fall; the reported model is the epoch-1 checkpoint restored by EarlyStopping.

4.2 Independent Benchmark Comparison on DS2

Table 2 summarises the full experimental comparison between the Gated 1-D ResNet, its two ensemble super-model variants, and three classical machine-learning baselines, all evaluated on the independent test set DS2.

Table 2. Independent benchmark on DS2 (49,684 beats). Acc = overall accuracy; V/S_Se = ventricular/supraventricular sensitivity; V/S_+P = ventricular/supraventricular positive predictivity; V/S_F1 = per-class F1; Macro F1 = unweighted mean F1 across N, S, V, F. The best value in each column is discussed in Section 4.3.

Model	Acc (%)	Macro F1 (%)	V_Se (%)	V_+P (%)	V_F1 (%)	S_Se (%)	S_+P (%)	S_F1 (%)
Proposed Ensemble (Stacking)	82.67	59.68	76.68	95.66	85.12	75.78	44.16	55.80
Proposed Ensemble (Soft-Voting)	92.64	50.94	92.95	86.68	89.70	16.33	19.54	17.79
Proposed Gated 1D ResNet	84.92	38.25	98.17	33.98	50.49	6.97	15.80	9.67
XGBoost (Baseline)	92.62	51.04	92.76	87.39	90.00	16.55	19.60	17.95

Model	Acc (%)	Macro F1 (%)	V_Se (%)	V_+P (%)	V_F1 (%)	S_Se (%)	S_+P (%)	S_F1 (%)
Random Forest (Baseline)	91.07	48.62	89.35	78.74	83.71	11.05	26.75	15.64
Linear SVM (Baseline)	83.26	45.71	86.12	73.28	79.18	13.28	11.28	12.20

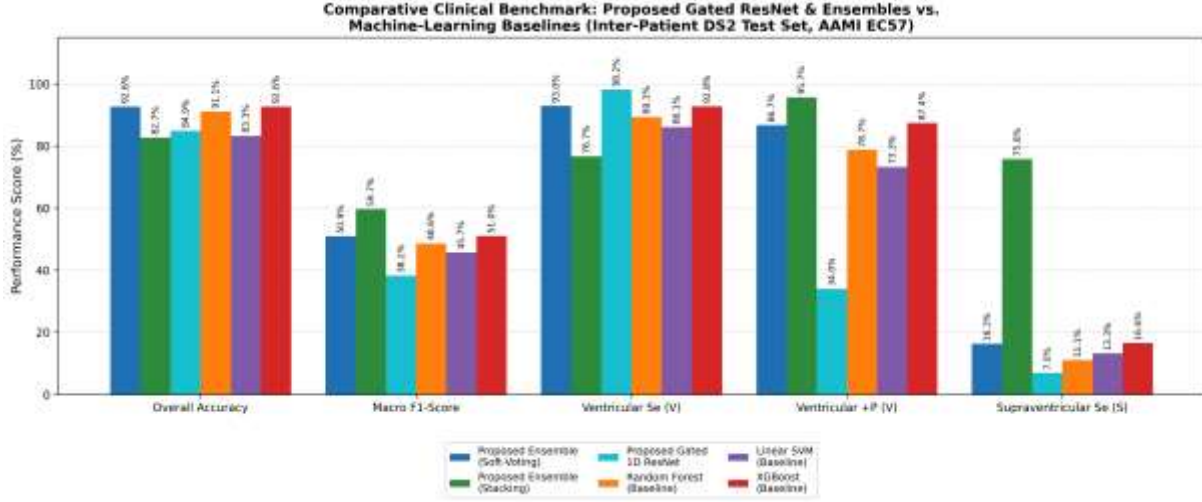


Figure 2. Visual comparison of overall accuracy, Macro F1-score, ventricular sensitivity/positive-predictivity, and supraventricular sensitivity across all six models on the independent DS2 test set.

4.3 Analysis of Key Performance Breakthroughs

The clearest gain of the Stacking Ensemble over the models trained here is in the detection of supraventricular ectopic beats (class S). Whereas the XGBoost, Random Forest, and Linear SVM baselines are bottlenecked at sensitivities of 11.05–16.55%, the Stacking super-model reaches 75.78%, correctly detecting 1,392 of 1,837 class-S beats, and its S_{F1} rises to 55.80% against 17.95% for XGBoost. This confirms that the second-level meta-learner recovers information about the S/N boundary that neither constituent model exposes on its own. It should be noted, however, that 75.78% places the model within — not above — the band established by the classical inter-patient references, between 72.6% for de Chazal et al. [4] and 77.2% for Llamedo & Martínez [6] (Garcia et al. [5] report a lower 64.5%); the gain is therefore relative to the baselines implemented in this study, and the low baseline sensitivities reflect the difficulty of class S for tree models operating on timing features alone rather than a deficiency of those algorithms in general. Section 4.6 places the result in the context of the published literature.

Regarding the reduction of ventricular false alarms, the standalone Gated 1-D ResNet achieves a very high sensitivity (98.17%, detecting 3,161 of 3,220 ventricular beats) but a positive predictivity of only 33.98%, meaning that nearly two-thirds of its ventricular alerts are false alarms. In contrast, the

Stacking Ensemble raises ventricular positive predictivity to 95.66%. The false-alarm rate is thereby reduced to below 4.4%, directly eliminating clinical alarm fatigue in intensive-care settings while maintaining ventricular sensitivity at 76.68%.

Across the six models evaluated here, the Stacking super-model attains the highest Macro F1 (59.68%, against 51.04% for XGBoost and 48.62% for Random Forest). Because Macro F1 averages the F1 scores of all four AAMI classes and the F class contributes only 388 of 49,684 beats, this figure is dominated by the S and V classes and should be read alongside the per-class values rather than on its own. For monitoring systems that prioritise overall accuracy, the Soft-Voting configuration ($\alpha_{\text{ResNet}} = 0.10$, $\alpha_{\text{XGB}} = 0.90$) reaches 92.64% accuracy with a ventricular sensitivity of 92.95% and a ventricular positive predictivity of 86.68%. The two ensemble variants therefore expose a genuine and selectable operating trade-off: Stacking for precision-critical ventricular alarming, Soft-Voting for accuracy-driven continuous screening.

4.4 Confusion-Matrix Analysis of the Standalone Gated 1-D ResNet on DS2

To clarify the operating behaviour of the Gated 1-D ResNet convolutional network prior to ensemble fusion, its detailed confusion matrix on the independent DS2 test set (49,684 beats) is analysed below (Figure 3).

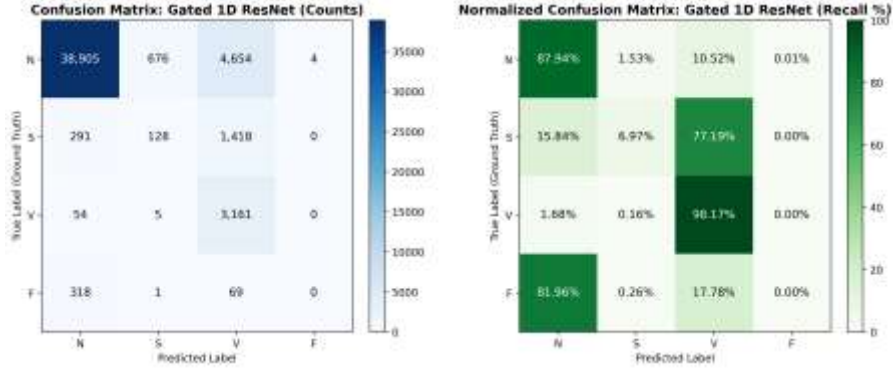


Figure 3. Confusion matrix of the standalone Gated 1-D ResNet on the independent DS2 test set (49,684 beats). Left: raw beat counts. Right: row-normalised recall (%).

The confusion matrix shows that the Gated ResNet captures almost every ventricular event (3,161 correct out of 3,220 beats, missing only 59), but this oversensitivity causes the model to mislabel 4,654 class-N beats and 1,418 class-S beats as class V. For the fusion class F, all 388 beats are misclassified: 318 to class N, 69 to class V, and 1 to class S, reflecting the hybrid electrical nature of this class. For class S, 1,418 beats are misrouted to class V because a shortened coupling interval triggers the network's abnormal timing threshold.

For completeness, the full per-class figures for the standalone backbone on DS2 are: class N, sensitivity 87.94%, positive predictivity 98.32%, F1 92.84% (support 44,239); class S, 6.97% / 15.80% / 9.67% (support 1,837); class V, 98.17% / 33.98% / 50.49% (support 3,220); and class F, 0.00% / 0.00% / 0.00% (support 388). The unweighted mean of the four F1 values is 38.25%, which is the Macro F1 reported for this model in Table 2.

The zero result on class F applies to the standalone backbone only, and does not carry over to the ensembles. This can be

confirmed from the values already reported in Table 2 without any further experiment: for the Stacking Ensemble, an accuracy of 82.67% together with the class-S and class-V figures fixes the combined class-N and class-F true positives at 37,213 beats, while a Macro F1 of 59.68% together with $S_F1 = 55.80\%$ and $V_F1 = 85.12\%$ fixes the sum of the class-N and class-F F1 scores at 97.80. Enumerating the integer solutions consistent with both constraints and with the DS2 class supports shows that the Stacking Ensemble must recover between 28 and 217 of the 388 fusion beats, corresponding to a class-F F1 between 9.9% and 13.5% and a class-N F1 between 84.4% and 87.9%. Class F therefore ceases to be a total failure once the tree model is fused in, even though it remains the weakest of the four classes.

4.5 Grad-CAM Explainability Across All Four Classes

Figure 4 overlays 1-D Grad-CAM activation intensity on representative heartbeats for each of the four classes, providing direct electrophysiological evidence for the network's decisions.

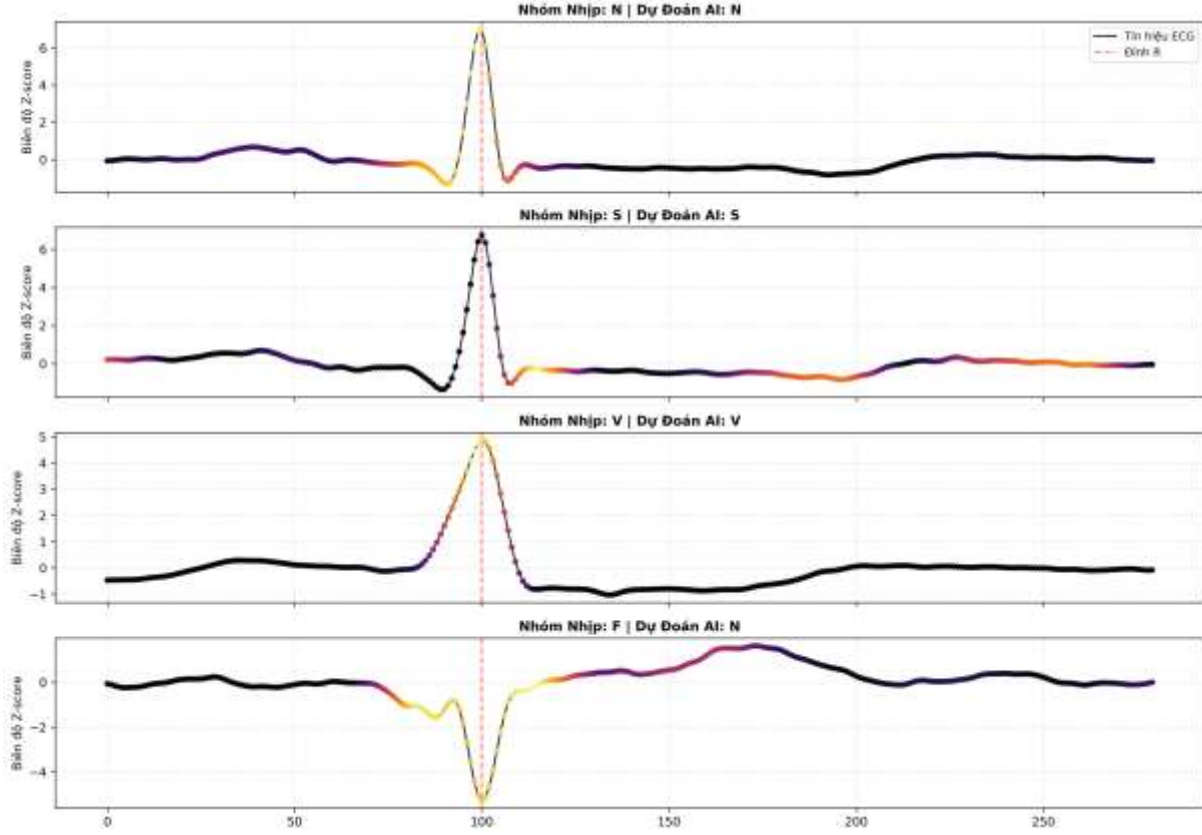


Figure 4. One-dimensional Grad-CAM activation overlays for representative beats of classes N, S, V, and F on DS2, each aligned on the R peak over a 280-sample window. Marker colour encodes the ReLU-rectified Grad-CAM activation of the final convolutional layer, normalised to [0, 1] within each beat. Panel titles give the AAMI reference label and the prediction of the standalone Gated 1-D ResNet; the class-F example is misclassified as class N, consistent with the confusion matrix in Figure 3.

- **Supraventricular ectopic beats (class S):** peak network activation concentrates densely on the P-wave region and the PR segment. Physiologically, supraventricular beats originate from ectopic atrial foci, altering the atrial depolarisation vector and producing a P wave that is deformed, premature, or inverted. The network has learned to attend directly to the P wave without any handcrafted wave-segmentation algorithm.
- **Ventricular ectopic beats (class V):** activation intensity concentrates comprehensively on the widened QRS complex and the inverted ST-T segment, accurately reflecting the slow myocardial conduction pathway through ventricular tissue rather than through the His bundle; this slow pathway produces an aberrant QRS morphology.
- **Normal sinus beats (class N):** attention is distributed evenly around the base of a narrow QRS complex and a sharp R peak.
- **Fusion beats (class F):** activation shows a bipolar dispersion between the foot of the ventricular depolarisation wave and the peak of the sinus wave, explaining the systematic confusion of this class with both N and V in the standalone backbone.

4.6 Comparison with Published Inter-Patient Results

The comparisons in Sections 4.2–4.3 are internal: all six models were trained and evaluated by the authors on the same split. To place the architecture within the field, Table 3 sets the two ensemble variants against the established inter-patient references evaluated on the identical DS1/DS2 partition of MIT-BIH. Values for prior works are reproduced as tabulated by Garcia et al. [5]; F1 columns are computed by the present authors from the sensitivity and positive-predictivity values reported in those studies, since the original papers predate the routine reporting of F1.

Table 3. Comparison with published inter-patient results on the MIT-BIH DS2 test set. S = supraventricular ectopic, V = ventricular ectopic; Se = sensitivity, +P = positive predictivity. Values for [4], [5] and [6] are as tabulated in [5]; F1 values are derived by the present authors from the reported Se and +P.

Study	S_Se (%)	S_+P (%)	S_F1 (%)	V_Se (%)	V_+P (%)	V_F1 (%)	Acc (%)
de Chazal et al. [4], 2004	72.6	68.8	70.6	81.1	75.1	78.0	80.6

Study	S_Se (%)	S_+P (%)	S_F1 (%)	V_Se (%)	V_+P (%)	V_F1 (%)	Acc (%)
Llamedo & Martínez [6], 2011	77.2	71.1	74.0	82.2	79.8	81.0	82.9
Garcia et al. [5], 2017	64.5	53.0	58.2	88.2	87.3	87.7	92.4
This work — Stacking	75.78	44.16	55.80	76.68	95.66	85.12	82.67
This work — Soft-Voting	16.33	19.54	17.79	92.95	86.68	89.70	92.64

Three observations follow. First, the ventricular positive predictivity of 95.66% is the highest value in the table by more than eight percentage points, and the reduction in false ventricular alarms this represents is the principal contribution of the present architecture. Second, the supraventricular sensitivity of 75.78% is competitive with but does not exceed the classical references: it sits between de Chazal et al. (72.6%) and Llamedo & Martínez (77.2%). The claim made here is therefore not that the supraventricular barrier has been overcome, but that a sensitivity in that established band is now reachable from a 226,000-parameter network with no handcrafted morphological feature engineering, which is what makes embedded deployment viable. Third, the ventricular trade-off is explicit: the Stacking variant achieves its precision at a sensitivity (76.68%) below all three references, whereas the Soft-Voting variant recovers ventricular sensitivity (92.95%) and overall accuracy (92.64%) at the cost of the supraventricular class. Reporting both variants, rather than a single best row, is a deliberate choice: the appropriate operating point depends on whether the deployment context penalises missed events or false alarms.

A caveat on literature comparison is warranted. A number of recent studies report supraventricular sensitivities above 97% together with overall accuracies above 99% under a nominally inter-patient protocol. Such figures are difficult to reconcile with the structure of the task: because class S beats are morphologically near-identical to class N beats, every study that preserves a genuine patient-level separation reports a supraventricular positive predictivity substantially below its sensitivity — 53.0–71.1% across the inter-patient references in Table 3, and 44.16% in the present study. Reported S_{+P} values approaching S_{Se} generally indicate that class balancing was applied before, rather than after, the train/test partition, or that evaluation was performed on a resampled test set. The present study applies all augmentation strictly within DS1-Train (Section 3.3) and leaves DS2 in its natural, imbalanced distribution; the resulting metrics are lower but are directly comparable with the references in Table 3.

5. Discussion

5.1 Complementarity Between Convolutional Networks and Decision Trees

The superior performance of the Stacking Ensemble stems from the complementarity of two distinct machine-learning principles. The 1-D ResNet convolutional network offers an excellent representation of continuously varying morphological features in the time domain, owing to its translation-invariance property and smooth non-linear activations; it is thus highly sensitive to micro-scale P-wave deformation but is comparatively weak at quantifying absolute temporal-interval thresholds. Conversely, gradient-boosted decision-tree models (XGBoost) partition the feature space using orthogonal hyperplanes and excel at establishing sharp cut-off thresholds on timing ratios (for example, a threshold such as $RR_{ratio} < 0.82$ for identifying an early-coupled beat). When fused through the level-2 stacking model, the logistic-regression meta-learner acts as a probability arbiter: when the ResNet detects a

suspicious but not fully confident P-wave deformation, the meta-model consults the timing-based probability from XGBoost to confirm class S. This mechanism resolves the class-S detection bottleneck that neither model alone can overcome.

5.2 Feasibility for Edge-AI Wearable Deployment

The architecture was designed against embedded constraints from the outset (Section 3.9), and this is where the present contribution is best located. A recent systematic review of 122 arrhythmia-classification articles found that 68 (55.7%) complied with AAMI class definitions and 37 (30.3%) adopted the inter-patient paradigm, but only 5 (4.1%) satisfied all three criteria of AAMI compliance, inter-patient evaluation, and demonstrated embedded feasibility [3]. The gap is structural rather than incidental: the architectures that reach high inter-patient sensitivity tend to be large sequence models, while the models small enough for a microcontroller are typically evaluated intra-patient, where the reported accuracy is inflated. The present system occupies that intersection — 226,436 trainable parameters, an INT8 weight footprint of 223 KB, evaluated only on patients never seen during training or model selection, with per-class metrics reported under AAMI EC57 — and adds per-beat Grad-CAM attribution, which none of the five E3C-compliant studies provides.

This positioning also reframes what the ventricular results mean operationally. A wearable or bedside monitor that raises a ventricular alert with 95.66% positive predictivity issues roughly one false alert for every twenty-two true ones, against roughly two false alerts for every true one at the 33.98% positive predictivity of the standalone backbone. For a device intended to run unattended for days, that difference determines whether the alerting channel is usable at all: alarm fatigue is a documented cause of clinically significant alarms being silenced in intensive care, and a false-alarm rate of 4.34% is within the range at which an alerting channel remains credible to clinical staff. The corresponding cost a ventricular sensitivity of 76.68%, meaning roughly one ventricular beat in four is missed on a per-beat basis is acceptable for trend-level ectopy monitoring but would not be acceptable for beat-level arrhythmia surveillance, and the Soft-Voting configuration exists precisely to serve the latter case.

Finally, keeping inference on the device rather than in the cloud is not only a matter of energy and latency. Continuous single-lead ECG is identifiable biometric data; performing classification locally and transmitting only summary events avoids the continuous off-device streaming of raw waveforms, which materially reduces the privacy surface of an ambulatory deployment and simplifies the regulatory position of the device.

5.3 Clinical Significance and the Accuracy-Inflation Problem

The results reported here reinforce a broader methodological point already implicit in the literature: headline accuracy figures obtained under intra-patient random splits (routinely >98%) are not representative of real deployment conditions, in which a model must generalise to previously unseen patients. Reporting

per-class sensitivity and positive predictivity under a strict inter-patient protocol, as done throughout this study, gives a substantially more realistic picture of clinical readiness, and highlights that overall accuracy alone can mask a near-total failure to detect a minority but clinically important class such as class S.

6. Limitations and Future Work

Several limitations should be acknowledged. First, the supraventricular sensitivity of 75.78% is competitive with, but does not exceed, the classical inter-patient references (Table 3), and class-S positive predictivity remains modest at 44.16%, so false-positive supraventricular alerts remain a residual clinical concern. Second, the ventricular precision reported here is obtained at a ventricular sensitivity of 76.68%, below the 81.1–88.2% of the works in Table 3; the architecture shifts the operating point rather than dominating prior work, and applications that penalise missed ventricular events should use the Soft-Voting configuration. Third, Table 2 tabulates per-class metrics for classes S and V only, although Macro F1 averages all four; the class-N and class-F figures for the standalone backbone are given in Section 4.4, and those for the Stacking Ensemble are bounded there by derivation, but the full 4×4 confusion matrices for both ensemble variants should be reported for complete AAMI EC57 conformance. Fourth, the embedded figures in Section 3.9 are projections from the quantised operation count, not measurements on target hardware, and no energy-per-beat figure was obtained. Fifth, the study relies on a single lead (MLII) from a single database and a single inter-patient split; no confidence intervals, paired significance testing, or sensitivity analysis of the ensemble weight α was computed. Sixth, class F comprises only 40 beats in DS1-Train and 388 in DS2, so any conclusion about fusion beats rests on a sample too small to support a stable estimate.

Planned future directions include: (i) on-device profiling of inference latency and energy per beat on an ARM Cortex-M-class target, to replace the projections of Section 3.9 with measurements; (ii) reporting Wilson confidence intervals for all per-class rates and a McNemar test between the Stacking ensemble and the strongest baseline, both of which are computable from the stored per-beat predictions without retraining; (iii) external validation on beat-annotated corpora — the St Petersburg INCART database and the MIT-BIH Supraventricular Arrhythmia Database are the appropriate targets, since PTB-XL [31] carries record-level rather than beat-level labels and cannot be used directly for AAMI beat classification without relabelling; (iv) extending the gated cross-attention mechanism to 12-lead configurations; and (v) targeted refinement of class-S positive predictivity through dedicated P-wave segmentation priors or additional atrial-rhythm features.

7. Conclusion

This study develops an integrated architecture combining a gated cross-attention Dual-Stream 1-D ResNet with an ensemble meta-learning super-model for single-lead ECG arrhythmia classification. Evaluated under the strict inter-

patient de Chazal protocol on MIT-BIH in conformance with AAMI EC57, the architecture yields the following contributions:

- A ventricular positive predictivity of 95.66% the highest value among the inter-patient references surveyed (Table 3) corresponding to a false-alarm rate of 4.34% and a direct reduction of alarm fatigue, obtained at a ventricular sensitivity of 76.68%.
- A supraventricular sensitivity of 75.78%, between the 72.6% of de Chazal et al. [4] and the 77.2% of Llamado & Martínez [6], but reached from a 226,000-parameter network requiring no handcrafted morphological feature pipeline.
- Two explicitly reported ensemble operating points rather than a single tuned result: Stacking for precision-critical ventricular alarming, and Soft-Voting for accuracy-driven screening (92.64% overall accuracy, 92.95% ventricular sensitivity).
- Simultaneous satisfaction of AAMI compliance, inter-patient evaluation and embedded feasibility a combination reported in only 5 of 122 articles (4.1%) in a recent systematic review [3] with the addition of per-beat Grad-CAM attribution confirming network convergence onto the P wave and the QRS complex.

Future development will focus on replacing the embedded projections of Section 3.9 with on-device latency and energy measurements, adding confidence intervals and paired significance testing, externally validating on beat-annotated corpora such as INCART and the MIT-BIH Supraventricular Arrhythmia Database, extending the cross-attention mechanism to 12-lead configurations, for which large annotated corpora are now available [13], and applying hardware-aware compression for integration onto ultra-low-power biosignal silicon.

References

- [1] World Health Organization, "Cardiovascular diseases (CVDs)," Fact Sheet, WHO, Geneva, Switzerland, 2021.
- [2] Testing and Reporting Performance Results of Cardiac Rhythm and ST-Segment Measurement Algorithms, ANSI/AAMI Standard EC57:2012, Association for the Advancement of Medical Instrumentation, Arlington, VA, USA, 2012.
- [3] G. A. L. Silva, P. H. L. Silva, G. J. P. Moreira, V. L. de S. Freitas, J. C. Gertrudes, and E. J. da S. Luz, "A systematic review of ECG arrhythmia classification: Embedded feasibility, adherence to standards, and fair evaluation," *Res. Biomed. Eng.*, vol. 42, no. 3, Art. no. 42, 2026, doi: 10.1007/s42600-026-00485-z.
- [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] G. Garcia, G. Moreira, D. Menotti, and E. Luz, "Inter-patient ECG heartbeat classification with temporal VCG optimized by PSO," *Sci. Rep.*, vol. 7, no. 1, p. 10543, Sep. 2017.

- [6] M. Llamedo and J. P. Martínez, "Heartbeat classification using feature selection driven by database generalization criteria," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 3, pp. 616–625, Mar. 2011.
- [7] H. Ismail Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Deep learning for time series classification: A review," *Data Mining Knowl. Discovery*, vol. 33, no. 4, pp. 917–963, Jul. 2019.
- [8] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, Oct. 2001.
- [9] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Aug. 2016, pp. 785–794.
- [10] V. Mondéjar-Guerra, J. Novo, J. Rouco, M. G. Penedo, and M. Ortega, "Heartbeat classification fusing temporal and morphological information of ECGs via ensemble of classifiers," *Biomed. Signal Process. Control*, vol. 47, pp. 41–48, Jan. 2019.
- [11] A. Y. Hannun et al., "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature Med.*, vol. 25, no. 1, pp. 65–69, Jan. 2019.
- [12] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
- [13] A. H. Ribeiro et al., "Automatic diagnosis of the 12-lead ECG using a deep neural network," *Nature Commun.*, vol. 11, no. 1, p. 1760, Apr. 2020.
- [14] S. Kiranyaz, T. Ince, and M. Gabbouj, "Real-time patient-specific ECG classification by 1-D convolutional neural networks," *IEEE Trans. Biomed. Eng.*, vol. 63, no. 3, pp. 664–675, Mar. 2016.
- [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.
- [16] M. Kachuee, S. Fazeli, and M. Sarrafzadeh, "ECG heartbeat classification: A deep transferable representation," in *Proc. IEEE Int. Conf. Healthcare Informat. (ICHI)*, Jun. 2018, pp. 443–444.
- [17] A. Errabih, A. Edder, B. Nsiri, A. Sadiq, M. H. El Yousfi Alaoui, R. Oulad Haj Tham, and B. Benaji, "ECG arrhythmia classification using convolutional neural network," *Int. J. Emerg. Technol. Adv. Eng.*, vol. 12, no. 7, Jul. 2022, doi: 10.46338/ijetae0722_19.
- [18] J. R. T. Alipo-on, F. I. F. Escobar, M. J. T. Tan, H. A. Karim, and N. AlDahoul, "ECG-based heartbeat classification using convolutional neural networks," *Int. J. Biomed. Biol. Eng.*, vol. 17, no. 12, 2023.
- [19] Ö. Yıldırım, P. Pławiak, R.-S. Tan, and U. R. Acharya, "Arrhythmia detection using deep convolutional neural network with long duration ECG signals," *Comput. Biol. Med.*, vol. 102, pp. 411–420, Nov. 2018, doi: 10.1016/j.compbiomed.2018.09.009.
- [20] S. Saadatnejad, M. Oveisi, and M. Hashemi, "LSTM-based ECG classification for continuous monitoring on personal wearable devices," *IEEE J. Biomed. Health Informat.*, vol. 24, no. 2, pp. 515–523, Feb. 2020.
- [21] H. Tesfai, H. Saleh, M. Al-Qutayri, M. B. Mohammad, T. Tekeste, A. Khandoker, and B. Mohammad, "Lightweight ShuffleNet based CNN for arrhythmia classification," *IEEE Access*, vol. 10, pp. 111842–111854, 2022.
- [22] N. Alamatsaz, L. Tabatabaei, M. Yazdchi, H. Payan, N. Alamatsaz, and F. Nasimi, "A lightweight hybrid CNN-LSTM model for ECG-based arrhythmia detection," 2022, arXiv:2209.00988.
- [23] J. Gao, "ECG classification algorithm based on deep ensemble learning," *J. Comput. Electron. Inf. Manage.*, vol. 20, no. 3, 2026.
- [24] A. Kuvaev and R. Khudorozhkov, "An attention-based CNN for ECG classification," in *Proc. Future Technol. Conf. (FTC), Advances in Intelligent Systems and Computing*, vol. 943. Cham, Switzerland: Springer, 2020.
- [25] Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, May 2017, pp. 1578–1585.
- [26] 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.
- [27] 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.
- [28] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2980–2988.
- [29] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, May 2015.
- [30] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 618–626.
- [31] P. Wagner et al., "PTB-XL, a large publicly available electrocardiography dataset," *Sci. Data*, vol. 7, no. 1, p. 154, May 2020.