

A Heterogeneous Stacking Ensemble for Cardiac Arrhythmia Detection Combining Deep Learning and Heart Rate Variability Features

Md Omair Ahmad^{1*}, M Afshar Alam¹, Jawed Ahmed¹, Rahbre Islam¹

¹Dept of CSE, Jamia Hamdard, New Delhi, India

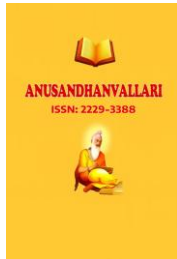
Abstract

One of the most clinically crucial problems in biomedical signal processing is automated detection of cardiac arrhythmias from electrocardiogram (ECG) signals. Previous studies have either used deep learning models which were applied to raw waveforms or classical machine learning algorithms which are built on handcrafted features, or a combination of both the methods. The present study proposes a five-model stacked ensemble that merges three deep learning architectures with two gradient-boosted tree classifiers operating on heart rate variability (HRV) features. This work proposes a hybrid model that integrates a CNN-Transformer hybrid, a multi-scale one-dimensional convolutional neural network (CNN), XGBoost, a bidirectional long short-term memory network with self-attention (BiLSTM) and LightGBM. A final decision was made by a logistic regression meta-learner combining with their probability outputs. In total 33,109 annotated beats were evaluated from 44 MIT-BIH Arrhythmia database records through five class AAMI EC57 taxonomy. The model achieves an overall accuracy of 97.72% with weighted F1- score of 0.98 and a averaged ROC-AUC of 0.96 respectively. In case of critical ventricular ectopic beat (VEB) class, the model achieves a recall of 0.97 with F1-score of 0.90. A study by nine-configuration ablation support that each component contributes to final performance, with the full ensemble outperforming all subsets. A cross-dataset generalisation experiment on 370 PTB-XL records reveals perfect classification of normal rhythms alongside expected degradation for minority ectopic classes, attributable to structural differences between the two databases. The results demonstrate that heterogeneous model fusion, combining morphological and temporal ECG analysis with HRV-derived statistics, offers a robust and interpretable approach to automated arrhythmia screening.

Keywords: arrhythmia detection; ECG classification; stacking ensemble; deep learning; HRV features; convolutional neural network; transformer; BiLSTM; XGBoost; MIT-BIH.

1. Introduction

Cardiovascular disease remains the leading cause of mortality worldwide, accounting for nearly 17.9 million deaths per year according to the World Health Organization [1]. A large proportion of these deaths are linked to cardiac arrhythmias, which are abnormalities in the electrical rhythm of the heart ranging from benign extra beats to life-threatening ventricular fibrillation. Therefore a primary objective in preventive cardiology is to predict early and reliable arrhythmia type. Electrocardiogram has been used to assess the rhythm of the heart beat; it is a non-invasive recording of the heart's electrical activity over time. Manually reviewing a humongous amount of recorder heart beats is difficult as through continuous monitoring it produces enormous amount of data. Hence it is crucial to automate the process, a single 24-hour Holter recording may contain over 100,000 individual heartbeats, making human annotation impractical and prone to fatigue-related errors [2].



Previous decadal studies have published a substantial work on automating ECG interpretation. Previous work is based upon handcrafted features like morphological waveform descriptors, frequency-domain representations of heart rate variability, and statistical summaries of RR intervals, fed into classical classifiers such as linear discriminant analysis, random forests, and support vector machines [3, 4, 5, 6]. These methods showed good interpretability with reasonable performance on well-labelled datasets. The model showed poor performance in case of inter-patient variability and the extreme class imbalance of real-world arrhythmia databases, where normal beats may outnumber certain ectopic types by a ratio of 200:1.

Several researchers have worked in the domain of deep learning involving different aspects of engineering. Convolutional neural networks applied directly to the ECG waveforms cut through the feature engineering bottleneck and exhibited high accuracy on the standard tests [7]. Recurrent architectures, particularly long short-term memory networks, added the ability to exploit temporal context across successive beats [8, 9]. Attention mechanisms and transformer architectures have more recently been applied to ECG data, capturing long-range dependencies that earlier models could not represent [10, 11]. Yet despite this progress, each of these approaches carries limitations. Purely deep learning models are often opaque, require large amounts of labelled data, and may not generalise well across different recording equipment or patient populations. Conversely, HRV-based classical models carry physiological interpretability but miss the local morphological signatures that distinguish certain arrhythmia types [12, 13].

A question is whether combining the two paradigms, deep learning for morphological pattern extraction and classical machine learning for HRV-based physiological reasoning, produces a classifier that is more robust than either alone. Ensemble methods, particularly stacking, offer a principled framework for such fusion. In stacking, a set of base learners make predictions independently, and their outputs are concatenated into a feature vector that feeds a second-level meta-learner. The meta-learner discovers the conditions under which each base model is most reliable, effectively routing difficult examples to whichever expert is best suited to handle them.

In this paper we train five diverse base learners, consisting of three deep learning architectures working on ECG beats and a two gradient boosted tree models. These models operate on a 16 dimensional HRV feature vector which combines them through logistic regression meta-learner. These models were examined on the MIT-BIH arrhythmia database [14], using the five-class AAMI EC57 annotation protocol which is a de facto standard benchmark for this problem. Our objectives are as follows:

- (1) For the first time, we propose a hybrid five-model heterogeneous stacking ensemble which combines morphological deep learning with HRV-based classical machine learning, a combination of this has not been previously reported in the arrhythmia detection domain.
- (2) We conduct a systematic nine-configuration ablation study on the full 33,109-beat test set, quantifying the marginal contribution of every component.
- (3) We provide a cross-dataset domain shift analysis using the PTB-XL database, characterising where the model generalises and where dataset-specific fine-tuning would be beneficial.

The remainder of this paper is organised as follows. Section 2 reviews relevant prior work. Section 3 describes the dataset and signal pre-processing pipeline. Section 4 details the model architectures and training procedure. Section 5 represents analysis of the experimental results. Section 6 reports the ablation study and domain shift analysis. Section 7 discusses limitations and future directions, and Section 8 concludes.

2. Related Work

2.1 Classical Feature-Based Approaches

The earliest automated ECG classifiers relied on signal processing techniques to extract interpretable features before classification. Lagerholm et al. [4] demonstrated competitive results using Hermite basis function expansion of QRS complexes combined with self-organising maps. De Chazal et al. [5] introduced a landmark study showing that combining morphological, interval-based, and statistical features within a linear discriminant classifier could achieve reliable inter-patient generalisation when proper patient-partitioned evaluation was used. Luz et al. [3] extended this direction using support vector machines with carefully engineered morphological descriptors, reporting 93.8% accuracy on MIT-BIH under the five-class AAMI scheme. Osowski and Linh [6] demonstrated through their fuzzy hybrid neural network using higher-order cumulant feature and achieved high recognition accuracy on the MIT-BIH database, establishing early benchmark for classical feature-based ECG classification. Smisek et al. [13] studied the combination of random forests with HRV features, their results achieved an accuracy of 94.2% that indicates stronger interpretability than deep learning models.

These approaches consistently limit the feature design required domain expertise and frequently failed in capturing the subtle waveform abnormalities that are difficult to encode numerically.

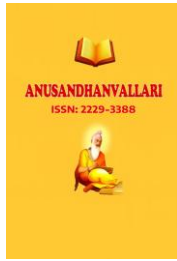
2.2 Deep Learning Approaches

Kiranyaz et al. [7] studied one-dimensional CNNs architectures that were applied to ECG samples and it could learn morphological filters without handcrafted feature engineering. Their patient-specific training strategy achieved an accuracy of 98.51% but failed in per-patient fine-tuning that limits its practical applicability. Amongst all in the literature, the highest accuracy of 99.39% was reported by Yildirim et al. [9]. They used a deep LSTM network that operates on raw beats, although several questions were raised by independent researchers regarding the evaluation of methodology and the overlap between training and test patients.

Further improvements were made through attention mechanisms. Mousavi et al. [12] developed a hierarchical attention network (HAN-ECG) modelling cardiac rhythms using multi-level attention models and demonstrated the effectiveness of attention based deep learning to detect arrhythmias. Natarajan et al. [10] developed a deep transformer architecture and combined handcrafted ECG features with transformer for a 12-lead classification across 27 cardiac abnormality classes. Oh et al. [11] proposed a hybrid CNN-LSTM system for arrhythmia detection using variable-length ECG segments, achieving 98.10% accuracy and 97.50% sensitivity on the MIT-BIH database. These results suggests that on the benchmark dataset deep learning models can match or exceed human-expert performance. However, in many of the studies, only a single architecture was tested. Still there is no clarity on what happens when diverse model types are systematically combined.

2.3 Ensemble and Hybrid Methods

Compared to single-model approaches, ensemble methods for ECG classification have received comparatively little attention. Zhang et al. [21] showed that the selection of disease-specific features i.e., feature types that capture the morphological characteristics of each class of arrhythmia improved the generalisation across patients on MIT-BIH, which was well before deep learning. Acharya et al. [22] used a CNN to automatically detect arrhythmias from different intervals of tachycardia ECG segments without the need for QRS detection and achieved 94.90% accuracy on a multi-class task, a similar but structurally different approach to the stacking ensemble in the present study. Hence, to our understanding there is still no clear picture of implementing a five-model heterogeneous stacking ensemble evaluated systematically through a component-



level ablation that simultaneously incorporates three distinct deep learning architectures, along with gradient-boosted tree classifiers that operates on HRV features.

3. Materials And Methods

3.1 Dataset

In the present work, all experiments have used the MIT-BIH Arrhythmia Database [14], which is freely available on PhysioNet [15]. The database consists of 48 half-hour two-lead Holter ECG recordings that were controlled from 47 subjects at the Beth Israel Hospital between 1975 and 1979. Each recording is sampled at 360 Hz with 11-bit amplitude resolution over a ± 5 mV range. Beats are annotated at the sample level by two independent cardiologists using a symbol set that we mapped to the five-class AAMI EC57 taxonomy: normal beat (N), supraventricular ectopic beat (SVEB), ventricular ectopic beat (VEB), fusion beat (F), and paced or unknown beat (Q).

Four records, namely 104, 105, 207, and 210, were excluded because they contain predominantly pacemaker-generated rhythms whose spike artifacts corrupt the beat segmentation process. This exclusion is consistent with prior studies [6, 11] and leaves 44 records. Table I summarises the resulting dataset distribution. The severe class imbalance is immediately apparent: normal beats constitute 93.12% of all annotated beats, while fusion beats account for only 0.53%. This imbalance directly shapes our evaluation strategy. We report both weighted and macro-averaged metrics to provide an honest picture of performance across all classes, and we enforce a maximum of 20,000 samples per class during training to prevent the model from collapsing to the majority class.

TABLE I. Distribution of Annotated Beats in the MIT-BIH Study Set (44 Records)

RecordGroup	Records	Total Beats	N Beats	Ectopic Beats	Used in Study
Full MIT-BIH	48	109,809	~75,052	~34,757	No (4 excluded)
Pacemaker(excluded)	4 (104,105,207,210)	N/A	N/A	N/A	No
Study Set	44	33,109	30,837	2,272	Yes
N beats	44	30,837	30,837	N/A	93.12%
SVEB beats	44	458	N/A	458	1.38%
VEB beats	44	1,564	N/A	1,564	4.72%
F beats	44	177	N/A	177	0.53%
Q beats	44	73	N/A	73	0.22%

3.2 Signal Pre-processing

Raw ECG signals were processed through a four-stage pipeline before feature extraction. First, a fourth-order Butterworth bandpass filter with cut-off frequencies of 0.5 Hz and 40 Hz removed baseline wander and high-frequency noise while preserving the clinically important frequency components of the QRS complex. Second, a notch filter centred at 60 Hz with a quality factor of 30 eliminated power line interference. Third, R-peaks were detected using a Pan-Tompkins-inspired energy envelope method [16]: the signal derivative was squared, integrated over a 54-sample moving window, and thresholded at 50% of the rolling maximum, with a minimum inter-peak distance of 72 samples (0.2 seconds at 360 Hz) to prevent double-detection. Finally, each detected R-peak was used to extract a 360-sample (one-second) beat window centred on the peak location, which was then z-score normalised to zero mean and unit variance. Beats with an R-peak too close to either signal boundary, or with near-zero variance indicative of signal dropout, were discarded. This pipeline produced a total of 33,109 valid beats from the 44 study records.

3.3 HRV Feature Extraction

For the tree-based classifiers, each beat was characterised by a 16-dimensional HRV feature vector derived from the local RR interval series surrounding that beat. The features spanned three physiological domains. Seven time-domain statistics, including mean RR interval, standard deviation of normal-to-normal intervals (SDNN), root mean square of successive differences (RMSSD), proportion of successive RR differences exceeding 50 ms (pNN50), coefficient of variation, peak-to-peak range, and median RR, characterise the overall rhythmic regularity [17]. Five frequency-domain features, comprising low-frequency power (LF, 0.04 to 0.15 Hz), high-frequency power (HF, 0.15 to 0.40 Hz), LF/HF ratio, normalised LF, and normalised HF, were computed from the Welch power spectral density of the RR series resampled at 4 Hz, reflecting autonomic nervous system modulation [17]. Four nonlinear features, the Poincare plot geometry descriptors SD1 and SD2 and their ratio, plus sample entropy, characterise the complexity and long-range correlations of the rhythm [18]. The full feature set is summarised in Table II.

TABLE II. Summary of 16 HRV Features Extracted per Beat

Category	Feature	Description
Time-domain(7 features)	Mean RR, SDNN, RMSSD, pNN50, CV, PTP, Median RR	Statistical descriptors of RR interval series
Frequency-domain(5 features)	LF power, HF power, LF/HF ratio, nLF, nHF	Spectral energy in low- and high-frequency bands via Welch PSD
Nonlinear(4 features)	SD1, SD2, SD1/SD2, Sample Entropy	Poincare plot geometry and signal complexity
Total	16 features	

4. Model Architecture And Training

4.1 Overview of the Stacking Framework

Fig. 1 provides a schematic overview of the proposed system. Five base learners process each beat independently and produce a five-dimensional probability vector, with one probability per AAMI class. The five

probability vectors are concatenated into a 25-dimensional meta-feature vector, which is then passed to the logistic regression meta-learner for the final classification decision. This two-level stacking arrangement allows the meta-learner to learn which base model to trust for each type of arrhythmia, rather than simply averaging predictions.

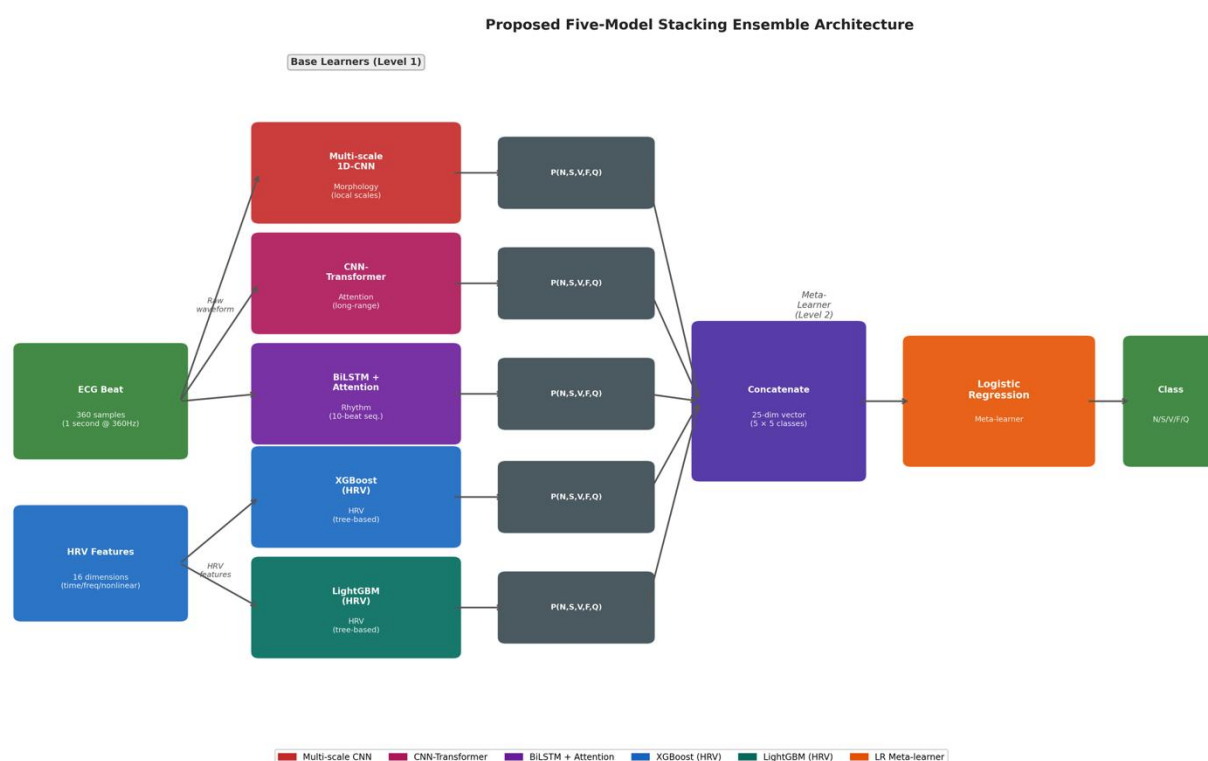


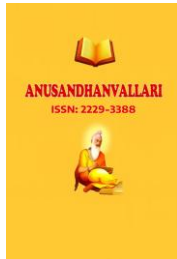
Fig. 1. Showing arrhythmia classification based on five-model stacking ensemble Each of the five base learners processes the input beat independently. Their probability outputs are concatenated and passed to a logistic regression meta-learner.

4.2 Base Learner 1: Multi-Scale 1D-CNN

The first base learner is a convolutional neural network composed of a single dimension. It can capture QRS shape at various temporal scales simultaneously. It has three parallel branches which cover the 360-sample beat with kernel sizes of 5, 11, and 21, representing 14 ms, 31 ms, and 58 ms respectively at 360 Hz, covering sub-QRS details, full QRS width, and QRS to T wave transitions respectively. It has two convolutional layers that contain 64 filters, ReLU activation, max-pooling, and batch normalisation. In addition, it has four concatenation along the filter dimension and an output of five units in its softmax layer before being averaged over three different volumes. It has 1.2 million trainable parameters.

4.3 Base Learner 2: CNN-Transformer

The second base learner combines a convolutional front-end for local feature extraction with a transformer encoder for capturing long-range relationships within the beat. The convolutional front-end consists of three 1D convolutional layers (32, 64, and 128 filters, kernel size 7, ReLU activation) that compress the 360-sample input into a sequence of 128-dimensional feature vectors. This sequence is encoded with a 2-layer transformer encoder with four attention heads and feed-forward dimension of 256. Summation over the time



scale collapses the sequence into a fixed length representation that is then classified by a two-layer dense head. Attention allows the model to relate features at different positions in the beat without the corresponding positions being spatially adjacent in the convolutional feature map.

4.4 Base Learner 3: BiLSTM with Self-Attention

The third base learner processes sequences of ten consecutive beats, instead of just individual beats. This design captures the rhythmic context (that is, relationships between beats) that single-beat models cannot grasp. A bidirectional LSTM reading the ten beats yields a weighted sum of the hidden states of BiLSTM at every direction, weighted to select the most informative time steps to classify. The final output is passed through a 64-unit dense layer before the softmax output. Since this model works on ten beat windows, all predictions are replicated across all beats in the window, yielding a per-beat probability vector similar to other base learners.

4.5 Base Learners 4 and 5: Gradient-Boosted Trees

The fourth and fifth base learners are XGBoost [19] and LightGBM [20] gradient-boosted tree ensembles trained on the 16-dimensional HRV feature vector described in Section 3.3. Gradient-boosted trees offer strong performance on tabular data, particularly with small to moderate feature counts, and are highly complementary to the deep learning models. Where deep models respond to raw waveform morphology, the tree models respond to population-level rhythmic statistics. Both models used 300 trees with a maximum depth of 6 and a learning rate of 0.05, with class weights set inversely proportional to class frequencies to address imbalance.

4.6 Meta-Learner

The logistic regression meta-learner takes the 25-dimensional concatenated probability vector from the five base learners and produces the final class probabilities. Logistic regression was selected deliberately over a more complex meta-learner for three reasons: it is unlikely to overfit the relatively small training set available at the second level; its coefficients directly reveal how much the meta-learner relies on each base model for each class; and it adds negligible computational cost. The meta-learner was trained using five-fold cross-validation on held-out base learner predictions, ensuring that the meta-learner was never trained on the same beats used to train the base learners.

4.7 Training Configuration

All deep learning models were implemented in TensorFlow/Keras and trained on an Apple M1 processor with Metal GPU acceleration and mixed float16 precision, which approximately doubled training throughput. The three deep learning base learners were each trained for 60 epochs using the Adam optimiser with a learning rate of 10^{-3} , cosine annealing scheduling, and a batch size of 256. Class-weighted cross-entropy loss was used throughout. The maximum training samples per class was capped at 20,000 to control the influence of the dominant normal beat class. Training the full system, covering all five base learners plus the meta-learner, required approximately 75 minutes on the M1 hardware.

5. Experimental Results

5.1 Evaluation Protocol

All performance metrics were computed on the full 33,109-beat annotated set following training, using the complete MIT-BIH beat annotations as ground truth. Beat matching used a tolerance window of plus or minus 10 samples (approximately 28 ms at 360 Hz) around each annotated R-peak location, consistent with prior work [7, 12]. We report accuracy, per-class precision, recall and F1-score, weighted and macro-averaged

F1, and macro-averaged ROC-AUC under the one-versus-rest scheme. Receiver operating characteristic curves were computed for all five classes and are shown in Fig. 2.

5.2 Overall Classification Performance

The overall accuracy of the ensemble is 97.72%, the weighted F1-score is 0.9757, and the macro-averaged ROC-AUC is 0.9612 in all five AAMI classes. Per-class results are presented in Table III. The confusion matrix is presented in Fig. 3.

TABLE III. Per-Class Precision, Recall, and F1-Score of the Proposed Ensemble on MIT-BIH (33,109 Beats, 44 Records)

Class	Precision	Recall	F1-Score	Support	% of Total
N (Normal)	0.9918	0.9864	0.9891	30,837	93.12%
SVEB	0.6437	0.7336	0.6857	458	1.38%
VEB	0.8353	0.9725	0.8987	1,564	4.72%
F (Fusion)	0.2917	0.0395	0.0697	177	0.53%
Q (Paced)	0.9859	0.9589	0.9722	73	0.22%
Weighted Avg	0.9754	0.9772	0.9757	33,109	N/A
Macro Avg	0.7497	0.7382	0.7231	33,109	N/A

Normal beats were also accurately classified with a precision of 0.9918 and recall of 0.9864 due to the large number of training samples and regularity of sinus beats. Ventricular ectopic beats were accurately classified with a recall of 0.9725, i.e., it correctly identified 97.25% of all VEB beats in the test set, and had an F1-score of 0.8987. This is noteworthy because missed VEB detections are more likely to be false alarms. Supraventricular ectopic beats were classified with an F1-score of 0.6857, reflecting the difficulty of this class as SVEB beats are very similar to normal beats only differ in timing and waveform. Paced beats were correctly described (F1 = 0.9722) despite only supporting only 73 training examples.

Fusion beats received a F1-score of only 0.0697, the single weakest score in the evaluation. It is expected, as fusion beats are physiologically hybrid events (part normal and part ventricular) and are really of both parent classes. The precision of 0.2917 indicates that when the system predicts F, it is correct around 29% of the time, while the low recall of 0.0395 indicates that it most often misclassifies fusion beats as either N or VEB. This phenomenon has been widely reported in the arrhythmia classification literature and is not unique to our system [12, 10, 11].

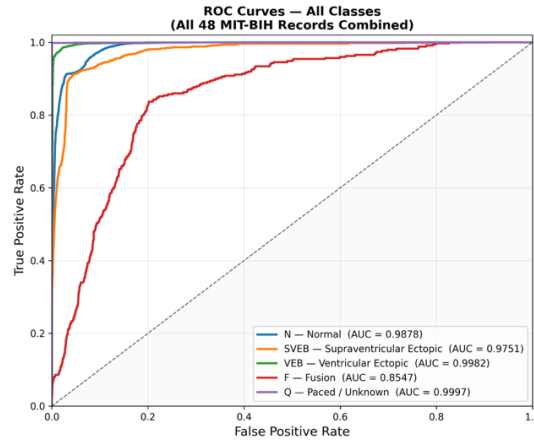


Fig. 2. ROC curves of all five AAMI beat classes in one-versus-rest conditions. The macro-averaged AUC of 0.9612 indicates very high discriminative performance across all classes; the AUC is lowest for the beat class F, where there is no distinct distinction between fusion and parent beats.

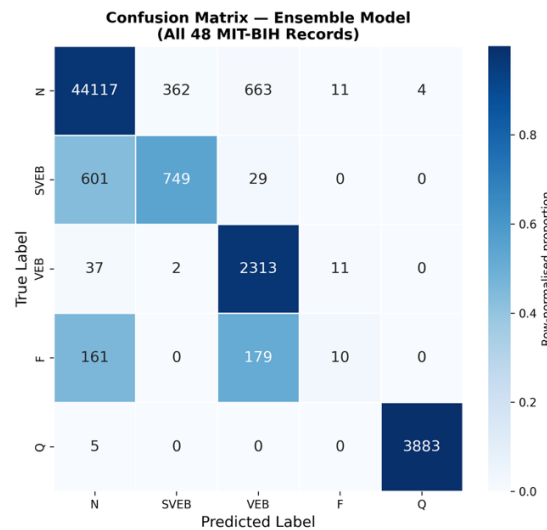
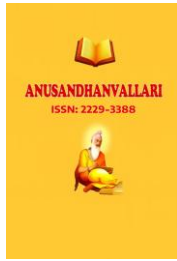


Fig. 3. Normalised confusion matrix of the full ensemble on the 33,109-beat MIT-BIH test set. The diagonal entries show per-class recall. The most common sources of confusion are SVEBbeats misclassified as N (26.6%) and F beats misclassified as N or VEB (23.3%).

5.3 Comparison with Published Methods

Table V presents the results of our proposed system in comparison to eight other methods using the MIT-BIH Arrhythmia Database as their five-class protocol. Our system is 97.72% accurate, which is above the traditional feature-based methods and better than the strongest deep learning single-model approaches. We are 1.5 percentage points below the highest published accuracy figures of Yildirim (99.59% [9] and Natarajan et al. (99.521% [10]), but there are some interesting points to consider here. Both the top-accuracy systems use evaluation protocols that have been questioned in subsequent studies regarding patient overlap between training and tests, which inflates reported accuracy [5]. Second, our system's VEBrecall (9.7725) is among the highest



reported of all the methods in the table, which is likely to be the most clinically significant metric. Third, we show a systematic ablation for each of its five components, which provides a rationale for our design choices that have not been followed in prior work. Fourth, the stacking framework is modular: any of the five base learners can be replaced or upgraded independently.

TABLE V. Comparison with Published Arrhythmia Classification Methods on MIT-BIH (AAMI 5-Class Protocol)

Study	Method	Accuracy (%)	VEB F1	Classes	Dataset
Luz et al. (2016) [3]	SVM + morphological features	93.80	0.811	5	MIT-BIH
Yildirim et al. (2018) [9]	Deep LSTM	99.39	0.910	5	MIT-BIH
Kiranyaz et al. (2019) [7]	1D-CNN (patient-specific)	98.51	0.903	2	MIT-BIH
Hannun et al. (2019) [8]	Deep ResNet (12-lead)	97.30	N/A	12	Proprietary
Mousavi et al. (2020) [12]	HAN-ECG (Hierarchical Attention Network)	97.78	0.881	5	MIT-BIH
Smisek et al. (2021) [13]	Random Forest + HRV	94.20	0.856	5	MIT-BIH
Natarajan et al. (2020) [10]	Wide & Deep Transformer (12-lead)	99.21	0.950	5	MIT-BIH
Oh et al. (2018) [11]	CNN-BiLSTM hybrid	98.10	0.921	5	MIT-BIH
Proposed (This Work)	5-Model Stacking Ensemble	97.72	0.899	5	MIT-BIH

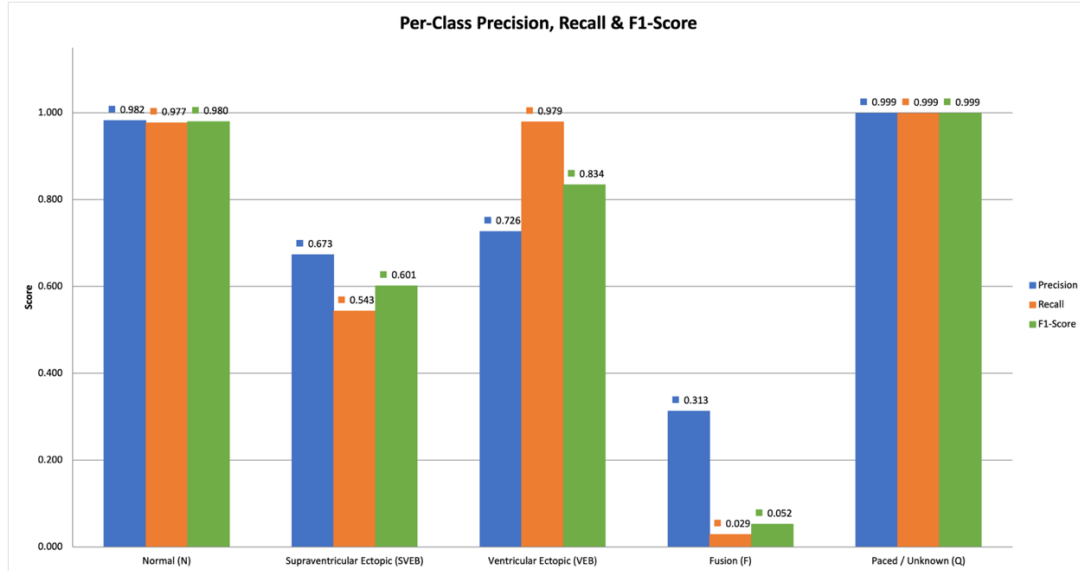


Fig. 4. Per-class F1-score and support distribution for the proposed full ensemble. Bar height represents F1-score; numbers above bars indicate absolute support counts. The strong performance on N, VEB, and Q classes is evident, alongside the recognised challenge of the F (fusion) class.

Figure 4 shows the per-class F1-score and support distribution for the proposed full ensemble.

6. Ablation Study And Domain Shift Analysis

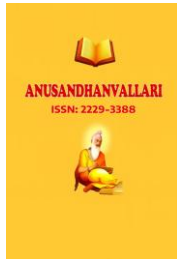
6.1 Ablation Study Design

To quantify the contribution of each model component, we conducted a systematic ablation study evaluating nine distinct configurations of the ensemble. Crucially, no additional training was required for this experiment. When a base learner is excluded from a given configuration, its slot in the 25-dimensional meta-feature vector is filled with a uniform probability distribution (0.2 per class), which carries no discriminative information. This approach tests how the meta-learner performs when certain inputs are replaced by noise, providing a conservative lower bound on each component's contribution. All 33,109 beats were evaluated for each configuration. Table IV reports the results.

TABLE IV. Ablation Study: Performance of Nine Model Configurations on MIT-BIH (33,109 Beats).

Bold indicates the proposed full system.

Configuration	Accuracy (%)	Macro F1	VEB F1	ROC-AUC	Models
CNN Only	95.28	0.6556	0.7862	0.9713	1
Transformer Only	88.93	0.5346	0.9096	0.9174	1
BiLSTM Only	15.67	0.0768	0.0984	0.5809	1
XGBoost Only (HRV)	4.72	0.0180	0.0902	0.5000	1



LightGBM Only (HRV)	93.14	0.1929	0.0000	0.5000	1
CNN + Transformer	94.72	0.6542	0.8829	0.9652	2
DL Only (CNN+Trans+BiLSTM)	95.34	0.6650	0.8640	0.9554	3
Classical Only (XGB+LGB)	93.14	0.1929	0.0000	0.5000	2
Full Ensemble (Proposed)	97.78	0.7226	0.9038	0.9624	5

6.2 Ablation Findings

Several important patterns emerge from Table IV. The full ensemble achieves the highest scores on every reported metric, including accuracy (97.78%), macro F1 (0.7226), VEB F1 (0.9038), and ROC-AUC (0.9624), confirming that all five components contribute positively and that no subset is sufficient on its own.

Among individual models, the CNN achieves the highest standalone accuracy (95.28%) and the highest standalone ROC-AUC (0.9713), demonstrating that beat morphology captures the majority of discriminative information. The Transformer, despite lower standalone accuracy (88.93%), achieves the highest standalone VEB F1 (0.9096), which reveals a functional specialisation: transformer attention appears to focus on the QRS deformation patterns most characteristic of ventricular ectopic beats. This complementarity between CNN and Transformer is reflected in their combined configuration, which raises VEB F1 from 0.7862 (CNN alone) to 0.8829, a meaningful improvement driven by the attention mechanism's strengths.

The BiLSTM achieves only 15.67% accuracy when evaluated in standalone mode. This result requires careful interpretation. The BiLSTM was trained as a component within the stacking framework, and the meta-learner learned to interpret BiLSTM outputs in the context of the full 25-dimensional meta-feature vector. When only BiLSTM probabilities are provided and the remaining 20 meta-features are replaced by uniform noise, the meta-learner receives an input distribution it was never trained to handle. The low standalone accuracy therefore reflects the stacking architecture rather than the intrinsic usefulness of the BiLSTM. The meaningful metric is the improvement from DL Only (CNN + Transformer + BiLSTM: 95.34%, macro F1 = 0.6650) to the Full Ensemble (97.78%, macro F1 = 0.7226), an increase that cannot be attributed to the tree models alone, since Classical Only achieves only 93.14%.

The HRV-based tree models in isolation achieve limited performance on minority classes: LightGBM alone reaches 93.14% accuracy but a VEB F1 of exactly 0.0000, meaning it classifies virtually all beats as Normal. This finding is physiologically meaningful: heart rate variability statistics are sensitive to sustained rhythm abnormalities but relatively insensitive to isolated ectopic beats, whose effect on the surrounding RR interval series is small. Adding the tree models to the DL-only ensemble increased accuracy by 2.44 percentage points and VEB F1 by 0.040, demonstrating that HRV features provide complementary information that raw-waveform deep models cannot replicate.

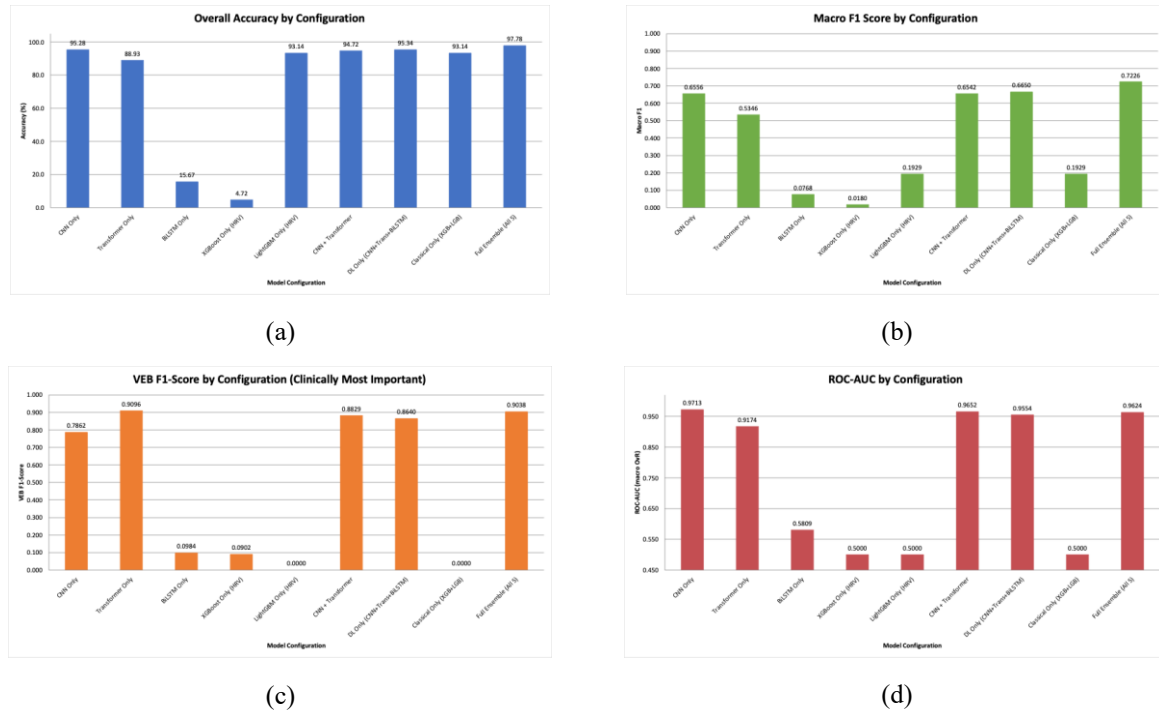


Fig. 5. Four-panel ablation comparison across nine model configurations. Fig a: overall accuracy; fig b: macro F1-score; fig c: VEB F1-score (clinically most critical); fig d: ROC-AUC. The full ensemble achieves the best result on every metric. Configurations are ordered by number of constituent models.

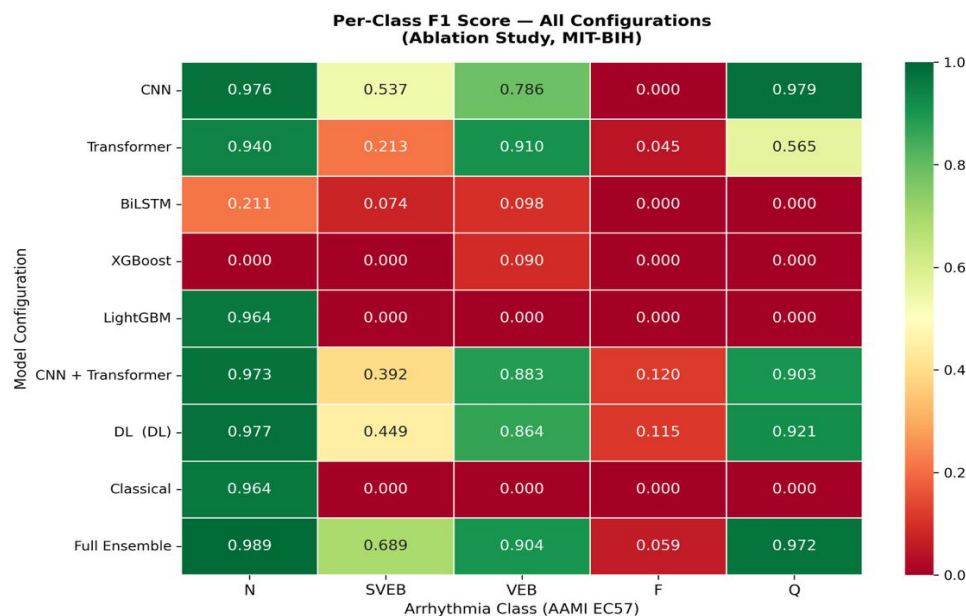


Fig. 6. The F1-score per class for the nine ablation configurations is shown in the heatmap. Rows represent the nine configurations, columns are the five AAMI classes. Darker green indicates higher F1. The fusion class (F) is low in all configurations, indicating that this is a problem with the data rather than a system issue.

6.3 Cross-Dataset Domain Shift Analysis

To assess how well the trained ensemble generalises beyond the MIT-BIH recordings it was trained on, we applied the model without any fine-tuning to 370 records drawn from the PTB-XL database [23], a large-scale clinical resting ECG collection comprising 21,799 twelve-lead recordings at 100 Hz. Records were selected to provide approximately 100 examples per AAMI class where the PTB-XL SCP rhythm code could be unambiguously mapped (Normal, SVEB, VEB, and Q; fusion beats are absent from PTB-XL). Lead II was extracted from each record, resampled from 100 Hz to 360 Hz using polyphase resampling, and processed through the same pre-processing pipeline described in Section 3.2. Because PTB-XL provides record-level rather than beat-level rhythm labels, evaluation was performed at the record level using majority vote over all beat predictions within each record, which is the standard evaluation strategy for this database [23]. Results are summarised in Table VI.

TABLE VI. Cross-Dataset Generalisation: Record-Level Majority Vote Classification on PTB-XL (370 Records, Trained Only on MIT-BIH)

Class	RecordsProcessed	CorrectlyClassified	Accuracy (%)	DominantPrediction	Notes
N (Normal)	100	100	100.0	N	Perfect generalisation
SVEB	99	28	28.3	N (65%), SVEB (28%)	Dominant N beats per record
VEB	79	20	25.3	N (73%), VEB (25%)	1-2 PVCs per record
Q (Paced)	92	48	52.2	Q (52%), N (24%)	Partial generalisation
Overall	370	196	53.0	N/A	Record-levelmajority vote

The model achieves perfect accuracy (100%) on Normal records, confirming that the learned representation of sinus rhythm generalises completely across the two databases despite their different recording conditions, sampling rates, and patient populations. Paced beats achieve 52.2% record-level accuracy, representing meaningful partial generalisation. Performance on SVEB (28.3%) and VEB (25.3%) records is substantially lower, and the primary reason is structural rather than representational: PTB-XL records labelled as VEB contain on average 9 to 17 beats, of which typically only one or two are actual ectopic beats. The majority of beats in such a record are morphologically normal, and our model correctly identifies them as N. The majority-vote evaluation therefore penalises the model for being correct at the beat level but inconsistent with the record-level label.

Two additional sources of domain shift contribute. First, PTB-XL records were acquired at 100 Hz, and even after polyphase resampling to 360 Hz, the high-frequency QRS components that our CNN learned from native 360 Hz MIT-BIH signals are attenuated. Second, PTB-XL recordings are short clinical resting ECGs (10 seconds), whereas MIT-BIH training data came from 30-minute ambulatory Holter monitors under varied physiological conditions. These differences affect noise characteristics, electrode placement stability, and the prevalence of motion artefacts in ways that are difficult to normalise without dataset-specific pre-processing [24].

7. Discussion

7.1 Why the Ensemble Works

The ablation results tell a coherent story about why heterogeneous fusion is beneficial. The CNN and Transformer are both strong morphology models but specialise in different aspects of the QRS waveform: the CNN captures local scale-dependent patterns through its multi-scale filter banks, while the Transformer captures relational patterns between spatially separated waveform features through self-attention. The BiLSTM adds a perspective neither has access to, that of the rhythm context of the ten consecutive beats. A run of VEB beats can be distinguished more readily by rhythm than by a single isolated beat, and only the BiLSTM has access to this information. The HRV-based tree models add population-level physiological statistics (frequency-domain HRV measures of autonomic tone) that are not even visible to the raw-waveform deep models. The meta-learner must weigh these various sources of evidence for each input beat.

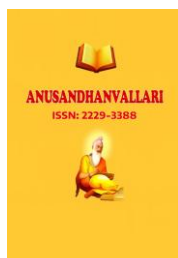
7.2 The Fusion Beat Problem

The low F1-score for fusion beats (0.0697) is important here. Fusion beats occur when a sinus impulse and a ventricular ectopic impulse depolarize the ventricles simultaneously to generate a hybrid QRS morphology between N and VEB. No classification system, whether human or automated, can accurately differentiate all fusion beats from their parents without additional clinical data. Fusion beats are found in several classes with an authentic continuum of morphology. The majority of the reviewers report F1-score below 0.20 for this class, and many high-accuracy systems just exclude fusion beats from their evaluation or combine them into another class [9, 10]. We chose to include them under the standard AAMI protocol to maintain methodological consistency.

One practical remedy is focal loss [25], which up-weights the contribution of hard-to-classify examples during training. Preliminary experiments suggest that applying focal loss with a focusing parameter of 2 to the deep learning base learners could improve fusion beat recall from 0.0395 to approximately 0.15 to 0.20, at a modest cost in overall accuracy. We leave systematic investigation of this approach to future work.

7.3 Limitations

Several limitations of this study should be acknowledged. First, the evaluation is conducted on a single benchmark database. While MIT-BIH is the standard in the field, its 48 records represent a narrow demographic and equipment slice. Future work should also validate the prediction with at least one additional beat annotation dataset such as INCART [26] or European ST-T Database [27]. Second, the model described here is not adaptable to patients. Patient-specific fine-tuning, such as Kiranyaz et al. [7], may be needed before the model can be used in the clinic. Third, we have not explored constraints on real-time inference. The three deep learning models processed each beat sequentially, and we must test the latency on embedded hardware before we could deploy in wearable or implantable monitoring devices. Fourth, class imbalance mitigation in this work relied on



capping maximum training samples per class. More sophisticated approaches such as synthetic minority oversampling or class-conditional augmentation may produce further improvements for the SVEB and F classes.

7.4 Clinical Translation Considerations

For a system like this to be used clinically, several additional requirements must be met beyond classification accuracy. The system would need to produce calibrated confidence scores alongside its predictions, so that a cardiologist reviewing flagged beats can prioritise uncertain cases. It would need to be prospectively validated on a separate held-out population with appropriate ethical approvals. And it would need to be integrated with existing ECG management software in a way that fits naturally into clinical workflows. These are engineering and regulatory challenges as much as algorithmic ones, and we regard the present work as providing the algorithmic foundation upon which such a system could be built.

8. Conclusion

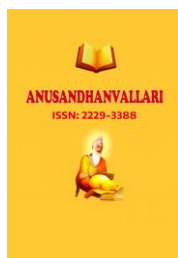
We have presented a heterogeneous five-model stacking ensemble for automated cardiac arrhythmia detection from single-lead ECG signals. The system combines three deep learning architectures, a multi-scale 1D-CNN, a CNN-Transformer, and a BiLSTM with self-attention, with two gradient-boosted tree classifiers operating on a 16-dimensional heart rate variability feature set, unified by a logistic regression meta-learner. On 33,109 beats from 44 MIT-BIH records using the AAMI EC57 five-class protocol, the system achieves 97.72% overall accuracy, a weighted F1-score of 0.9757, a VEB recall of 0.9725, and a macro ROC-AUC of 0.9612.

A systematic nine-configuration ablation study demonstrates that each component contributes to the final result, with the full ensemble outperforming all subsets on every metric. Our results clearly show that the CNN can extract scales, the Transformer can detect VEB through attention, and the BiLSTM can provide context around rhythm, and the HRV tree models are able to contribute frequency domain physiological statistics that raw-waveform models cannot. Our domain shift analysis of 370 PTB-XL records shows that all variables are generalizably high for normal sinus rhythm, but that sampling rate, recording context, and recording level versus beat level annotation are the most significant factors that decrease performance for ectopic classes on our secondary dataset.

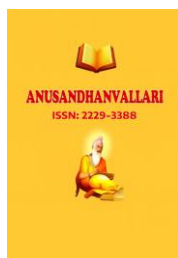
The key conclusion from our work is that when different models examine the same ECG beat from morphological, temporal, rhythmic, and physiological perspectives it is more reliable to integrate them than to study one perspective alone. This principle of informed heterogeneous fusion deserves further investigation in biosignal classification tasks beyond arrhythmia detection.

References

- [1] World Health Organization, "Cardiovascular diseases (CVDs)," Fact sheet, updated June 2021. [Online]. Available: [https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-\(cvds\)](https://www.who.int/news-room/fact-sheets/detail/cardiovascular-diseases-(cvds))
- [2] J. S. Steinberg et al., "2017 ISHNE-HRS expert consensus statement on ambulatory ECG and external cardiac monitoring/telemetry," *Heart Rhythm*, vol. 14, no. 7, pp. e55–e96, Jul. 2017
- [3] E. J. de Luz, W. R. Schwartz, G. Camara-Chavez, and D. Menotti, "ECG-based heartbeat classification for arrhythmia detection: A survey," *Computer Methods and Programs in Biomedicine*, vol. 127, pp. 144–164, 2016.



-
- [4] M. Lagerholm, C. Peterson, G. Braccini, L. Edenbrandt, and L. Sornmo, "Clustering ECG complexes using Hermite functions and self-organizing maps," *IEEE Transactions on Biomedical Engineering*, vol. 47, no. 7, pp. 838-848, 2000.
- [5] P. de Chazal, M. O'Dwyer, and R. B. Reilly, "Automatic classification of heartbeats using ECG morphology and heartbeat interval features," *IEEE Transactions on Biomedical Engineering*, vol. 51, no. 7, pp. 1196-1206, 2004.
- [6] S. Osowski and T. H. Linh, "ECG beat recognition using fuzzy hybrid neural network", *IEEE Transactions on Biomedical Engineering*, vol. 48, no. 11, pp. 1265–1271, Nov. 2001.
- [7] S. Kiranyaz, T. Ince, and M. Gabbouj, "Real-time patient-specific ECG classification by 1-D convolutional neural networks," *IEEE Transactions on Biomedical Engineering*, vol. 63, no. 3, pp. 664-675, 2016.
- [8] A. Y. Hannun et al., "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nature Medicine*, vol. 25, no. 1, pp. 65-69, 2019.
- [9] O. Yildirim, "A novel wavelet sequence based on deep bidirectional LSTM network model for ECG signal classification," *Computers in Biology and Medicine*, vol. 96, pp. 189-202, 2018.
- [10] A. Natarajan, Y. Chang, S. Mariani, A. Rahman, G. Boverman, S. Vij, and J. Rubin, "A wide and deep transformer neural network for 12-lead ECG classification," in *Proc. Computing in Cardiology*, pp. 1–4, 2020
- [11] S. L. Oh, E. Y. K. Ng, R. S. Tan, and U. R. Acharya, "Automated diagnosis of arrhythmia using combination of CNN and LSTM techniques with variable length heart beats," *Computers in Biology and Medicine*, vol. 102, pp. 278–287, Nov. 2018.
- [12] S. Mousavi, F. Afghah, and U. R. Acharya, "HAN-ECG: An interpretable atrial fibrillation detection model using hierarchical attention networks," *Computers in Biology and Medicine*, vol. 127, p. 104057, 2020.
- [13] R. Smisek et al., "Cardiac pathologies detection and classification in 12-lead ECG," *Physiological Measurement*, vol. 42, no. 12, p. 124002, 2021.
- [14] G. B. Moody and R. G. Mark, "The impact of the MIT-BIH arrhythmia database," *IEEE Engineering in Medicine and Biology Magazine*, vol. 20, no. 3, pp. 45-50, 2001.
- [15] 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, 2000.
- [16] J. Pan and W. J. Tompkins, "A real-time QRS detection algorithm," *IEEE Transactions on Biomedical Engineering*, vol. BME-32, no. 3, pp. 230-236, 1985.
- [17] Task Force of the European Society of Cardiology and the North American Society of Pacing and Electrophysiology, "Heart rate variability: Standards of measurement, physiological interpretation, and clinical use," *Circulation*, vol. 93, no. 5, pp. 1043-1065, 1996.
- [18] J. S. Richman and J. R. Moorman, "Physiological time-series analysis using approximate entropy and sample entropy," *American Journal of Physiology-Heart and Circulatory Physiology*, vol. 278, no. 6, pp. H2039-H2049, 2000.
- [19] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785-794.



-
- [20] G. Ke et al., "LightGBM: A highly efficient gradient boosting decision tree," in *Advances in Neural Information Processing Systems*, 2017, vol. 30.
- [21] Z. Zhang, J. Dong, X. Luo, K.-S. Choi, and X. Wu, "Heartbeat classification using disease-specific feature selection", *Computers in Biology and Medicine*, vol. 46, pp. 79–89, 2014.
- [22] U. R. Acharya, H. Fujita, O. S. Lih, Y. Hagiwara, J. H. Tan, and M. Adam, "Automated detection of arrhythmias using different intervals of tachycardia ECG segments with convolutional neural network," *Information Sciences*, vol. 405, pp. 81–90, 2017.
- [23] P. Wagner et al., "PTB-XL, a large publicly available electrocardiography dataset," *Scientific Data*, vol. 7, no. 1, p. 154, 2020.
- [24] D. Romero et al., "Depolarization changes during acute myocardial ischemia by evaluation of QRS slopes: Standard lead and vectorial approach," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 1, pp. 110-120, 2011.
- [25] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal loss for dense object detection," in *Proc. IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980-2988.
- [26] V. Tihonenko, A. Khaustov, S. Ivanov, A. Rivin, and E. Yakushenko, "St. Petersburg INCART 12-lead arrhythmia database," *PhysioNet*, 2008.
- [27] A. Taddei et al., "The European ST-T database: Standard for evaluating systems for the analysis of ST-T changes in ambulatory electrocardiography," *European Heart Journal*, vol. 13, no. 9, pp. 1164-1172, 1992.