Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Evaluation of Cascade Forests towards Interpretable ECG Arrhythmia Classification: Performance, Interpretability, and Deployment Trade-offs

Victor Adeleye

victormayowa163@gmail.com

Ekiti State University

Olayinka Adeleye

Otago Polytechnic

Mahmoud Elbattah

University of the West of England

Research Article

Keywords: cascade forest, interpretable machine learning, ECG arrhythmia classification, explainable AI, ensemble learning, clinical decision support, wavelet transform, SHAP analysis

Posted Date: March 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-8950285/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Evaluation of Cascade Forests towards Interpretable ECG Arrhythmia Classification: Performance, Interpretability, and Deployment Trade-offs

Victor Adeleye[1*], Olayinka Adeleye[2], Mahmoud Elbattah[3],[4]

[1] Department of Community Medicine, Ekiti State University Teaching Hospital, Ado-Ekiti, Ekiti, Nigeria; victormayowa163@gmail.com

[2] Department of Information Technology, Otago Polytechnic, Dunedin, New Zealand.

[3] Laboratoire Modélisation Information Systèmes (MIS), University of Picardie Jules Verne, 80000 Amiens, France

[4] College of Arts, Technology and Environment, University of the West of England, Bristol BS16 1QY, UK

[*] Correspondence: victormayowa163@gmail.com; +2348102562592

Abstract

Background: Deep learning models achieve high accuracy in electrocardiogram (ECG) arrhythmia classification but pose critical barriers to clinical deployment: opaque decision-making processes that resist clinical validation, substantial computational requirements that necessitate specialised hardware, and training instability that requires extensive expertise. This study provides a rigorous evaluation of cascade forest architectures, hierarchical tree-based ensembles that preserve interpretability while enabling representation learning, as potential alternatives for clinical ECG analysis.

Results: Our cascade forest implementation, combining Discrete Wavelet Transform features with a three-layer ensemble architecture, achieved 98.79% accuracy and 92.93% macro-F1 score on the MIT-BIH Arrhythmia Database. The model demonstrated statistically significant superiority over Random Forest baselines (86.83% macro-F1, Wilcoxon p=0.032, Cohen's d=1.24) with particularly pronounced improvements for clinically critical minority classes: Ventricular ectopic beats (F1: 0.96 vs 0.93) and Fusion beats (F1: 0.80 vs 0.63, representing 27% relative improvement). Performance approached contemporary deep learning benchmarks (99.2–99.5% accuracy range) with deficits of 0.41–0.71 percentage points. Comprehensive SHAP analysis validated framework reliance on physiologically plausible features: wavelet coefficients capturing QRS morphology, RR interval dynamics, and signal variance. Computational analysis revealed 2.79 hours of execution time requiring only standard CPU infrastructure without specialised GPU acceleration.

Conclusions: These findings on MIT-BIH data suggest that cascade forests warrant further investigation as interpretable alternatives for clinical arrhythmia detection, particularly in deployment contexts where transparency, computational accessibility, and reliable minority class detection outweigh marginal accuracy improvements. Multi-database validation and prospective clinical evaluation remain essential to establish generalisability beyond the evaluated benchmark.

Keywords: cascade forest, interpretable machine learning, ECG arrhythmia classification, explainable AI, ensemble learning, clinical decision support, wavelet transform, SHAP analysis

1. Introduction

Cardiac arrhythmia contributes substantially to the global burden of the leading cause of mortality worldwide, cardiovascular disease, vis-à-vis sudden cardiac death, stroke and heart failure [1]. Electrocardiography (ECG) remains one of the most widely used and clinically indispensable non-invasive diagnostic tools for identifying cardiac arrhythmias. Yet, manual interpretation requires specialised cardiological expertise and substantial time investment, particularly challenging for continuous monitoring in resource-limited clinical settings [2]. With the increasing volume of ambulatory and continuous ECG recordings in clinical practice, automated classification systems have become essential to support timely and accurate interpretation, consequently becoming a high-priority research area.

Contemporary deep learning architectures, convolutional neural networks, recurrent models (LSTMs, GRUs), and attention-based transformers have demonstrated remarkable diagnostic performance, with reported accuracy exceeding 99% on benchmark datasets [3,4]. These end-to-end learning paradigms simultaneously perform feature extraction and classification, circumventing traditional manual feature engineering [5]. Recent hybrid models combining CNN-BiGRU-attention mechanisms achieved 99.41% accuracy on MIT-BIH data, while CNN-BLSTM architectures reported comparable 99.5% accuracy [4,6].

However, this impressive performance conceals fundamental limitations constraining clinical deployment. Deep learning models function as "black boxes"; their multilayered architectures with millions of parameters generate predictions through opaque processes that resist clinical interpretation [7]. This lack of transparency creates substantial deployment barriers: clinicians cannot verify model logic against established cardiac electrophysiology principles, regulatory bodies encounter difficulty approving systems with inscrutable decision processes, and patients reasonably question recommendations lacking clear justification [8]. Additionally, deep learning requires specialised hardware (GPUs/TPUs), extensive training expertise, and computational resources often unavailable in resource-constrained healthcare environments globally [9].

The clinical AI community increasingly recognises interpretability not as optional but as fundamental for trustworthy medical systems [10]. Regulatory frameworks worldwide, including the European Union's AI Act and FDA guidance, explicitly require transparency mechanisms enabling clinical validation of algorithmic recommendations[7]. Post-hoc explainability techniques (SHAP, LIME, attention visualisation) provide approximations of model reasoning but generate local explanations rather than global understanding, and their reliability in capturing true model behaviour remains debated[10,11]. Cascade forest architectures, introduced by Zhou and Feng [12] and subsequently termed "deep forests" or gcForest, represent an architectural paradigm combining interpretable tree-based

methods with representation learning capabilities. These hierarchical arrangements of decision forest ensembles enable progressive feature refinement across multiple layers while maintaining transparency through tree-based decision paths and native feature importance rankings. Recent preliminary work by Lin et al. [13] applied wavelet-based features with gcForest to MIT-BIH data, achieving 98.55% accuracy and suggesting competitive performance with interpretability advantages. However, systematic evaluation comparing cascade forests against both traditional ensemble baselines and contemporary deep learning, quantifying computational trade-offs, and validating interpretability through modern explainability frameworks (SHAP analysis) remained limited.

This study addresses these gaps through a comprehensive evaluation of cascade forest architectures for ECG arrhythmia classification. We systematically investigate whether carefully designed cascade forests can achieve clinically adequate performance while preserving interpretability essential for clinical validation and computational accessibility critical for widespread deployment. Our evaluation encompasses five key dimensions:

1. **Performance benchmarking** against Random Forest baselines under identical feature extraction conditions, isolating cascade architecture contributions through rigorous statistical testing
2. **Comparison with deep learning** architectures to quantify performance-interpretability trade-offs
3. **Interpretability validation** through comprehensive SHAP analysis, examining whether models leverage physiologically meaningful features aligned with clinical reasoning
4. **Minority class evaluation** emphasising clinically critical rare arrhythmias through macro-averaged metrics and class-specific analysis
5. **Computational characterisation** quantifying execution time and hardware requirements relevant to deployment feasibility

We evaluate on the MIT-BIH Arrhythmia Database using the ANSI/AAMI EC57:1998 standard mapping, combining Discrete Wavelet Transform feature extraction with a three-layer cascade forest architecture. While single-database evaluation limits the generalisability of conclusions, this controlled investigation provides foundational evidence regarding cascade forest viability for interpretable ECG analysis, informing future multi-database validation efforts and prospective clinical studies.

2. Related Work

2.1 The Clinical AI Interpretability Challenge

Table 1 explicates the tension between model performance and interpretability across model families, and achieving this balance has become crucial to biomedical researchers as deep learning increasingly dominates medical AI applications [7]. Neural network architectures achieve impressive diagnostic accuracy yet generate profound clinical concerns extending beyond transparency preferences to fundamental requirements for safety, accountability, and regulatory compliance[10].

Table 1. Interpretability Characteristics Across Major Model Families in Medical AI

Model Family	Representative Methods	Inherent Interpretability	Post-hoc Explainability Requirement	Performance Capacity	Clinical Validation Feasibility
Traditional Machine Learning	Decision Trees, Logistic Regression, Linear Models	**High** — Explicit decision rules and coefficients directly traceable to input features	Minimal or unnecessary	Moderate — Limited ability to model nonlinear, hierarchical relationships	**High** — Transparent reasoning supports auditing, bias detection, and regulatory review
Ensemble Methods	Random Forest, Gradient Boosting, ExtraTrees	Moderate — Individual trees interpretable, ensemble behaviour partially opaque	Optional — SHAP, permutation importance often used to clarify feature influence	High — Improved generalisation via aggregation of multiple learners	Moderate to High — Feature importance metrics enable validation, though global reasoning is less intuitive
Deep Learning	CNNs, RNNs, Transformers	Low — Internal representations distributed and non-transparent	**Required** — SHAP, LIME, attention visualisation, saliency maps	Very High — Hierarchical representation learning captures complex, multiscale patterns	Limited — Post-hoc explanations debated; challenges for regulatory transparency and clinical trust

Medical diagnostic systems must satisfy requirements distinct from general AI applications. Clinicians demand explicit rationales enabling validation against domain expertise and established pathophysiology [8]. Regulatory frameworks mandate transparency mechanisms supporting algorithmic auditing and bias detection. Patient autonomy principles require comprehensible explanations enabling informed consent regarding AI-assisted diagnoses. Legal liability frameworks necessitate clear responsibility attribution when automated systems generate erroneous recommendations.

Deep learning's opacity fundamentally conflicts with these requirements. Post-hoc explainability techniques provide partial solutions: SHAP assigns feature importance scores based on cooperative game theory [14], LIME approximates local model behaviour through interpretable surrogates [11], and attention visualisation highlights salient input regions [15]. However, these methods generate approximations rather than genuine transparency, and their reliability remains debated; attention weights may not indicate truly influential features, and SHAP values can mislead regarding causal relationships [16].

Recognising deep learning limitations, researchers have explored inherently interpretable architectures. Traditional machine learning approaches, support vector machines, k-nearest

neighbours, and shallow neural networks with hand-crafted features dominated early automated ECG analysis [17]. While transparent, these methods demonstrated limited performance, particularly for subtle arrhythmia patterns requiring complex feature representations.

Tree-based ensemble methods, particularly Random Forests, offer attractive compromises between performance and interpretability [3]. These aggregations of decision trees trained on bootstrapped samples provide superior accuracy compared to individual trees while maintaining reasonable transparency through feature importance rankings. However, as single-layer ensembles, random forests lack the hierarchical representation learning capacity enabling deep architectures to discover increasingly complex patterns through layered composition [12].

2.2 Cascade Forest Architecture

Cascade forests arrange multiple layers of decision forest ensembles hierarchically.[12] Each layer receives original features augmented with class probability distributions from preceding layers, enabling progressive feature refinement where early layers detect simple patterns while deeper layers capture abstract relationships. This cascaded structure provides theoretical advantages: representation learning capabilities approaching neural networks while maintaining tree-based interpretability, inherent robustness to hyperparameter variations compared to deep learning's training instability, and automatic depth determination through validation performance monitoring [12,13].

Applications to ECG analysis remain limited but promising. Lin et al. [13] combined wavelet packet decomposition with gcForest, achieving 98.55% accuracy on MIT-BIH data, approaching contemporary deep learning benchmarks. However, systematic comparison against traditional ensemble baselines under identical feature extraction conditions, detailed computational efficiency analysis, and comprehensive interpretability validation through modern explainability frameworks remained unexplored. Our study addresses these gaps through rigorous controlled evaluation, isolating cascade architecture contributions and validating interpretability through SHAP analysis.

3. Methodology

3.1 Dataset

We utilised the MIT-BIH Arrhythmia Database, a benchmark ECG repository comprising 48 half-hour ambulatory recordings from 47 individuals, sampled at 360 Hz with 11-bit resolution and annotated per clinical standards. The Modified Limb Lead II (MLII) channel was selected to visualise the P-QRS-T waveform, serving as the primary input for analysis. Following the ANSI/AAMI EC57:1998 framework, beat annotations were consolidated into five clinically meaningful arrhythmic categories as represented in Table 2 below [18]. The curated dataset contained 109,494 labelled beats, exhibiting a pronounced class imbalance typical of real-world ECG data, with normal beats constituting the vast majority, necessitating evaluation strategies that emphasise performance on under-represented classes as well as aggregate metrics.

Table 2: ANSI/AAMI EC57:1998 class mapping showing superclass definitions, constituent beat annotations, and clinical significance [18]

Class	AAMI Category	Beat Annotations	Description
N	Normal	'N', 'L', 'R', 'e', 'j'	Normal and bundle branch block beats
S	Supraventricular Ectopic Beat (SVEB)	'A', 'a', 'J', 'S'	Atrial or nodal premature beats
V	Ventricular Ectopic Beat (VEB)	'V', 'E'	Premature ventricular contractions
F	Fusion Beat	'F'	Fusion of ventricular and normal beats
Q	Unknown Beat	'/', 'f', 'Q'	Paced, unclassifiable, or artefact beats

Given the skewed class distribution (82.8% Normal beats), we preserved natural proportions in the test set but balanced the training set via a stratified 80/20 train-test splitting, preserving class distributions. This is followed by the Synthetic Minority Over-sampling Technique (SMOTE)[19] applied exclusively to the training partition, which generates synthetic minority samples through interpolation between neighbouring minority samples, thereby augmenting minority representation without replicating data directly. This resulted in balanced training data (362,525 samples, 20% per class) while preserving the original imbalanced test distribution (21,899 samples) for realistic evaluation.

3.2 Signal Preprocessing and Feature Extraction

Signals underwent a standardised denoising pipeline to ensure high-fidelity waveform representation. Baseline wander, powerline noise, and electromyographic artefacts were suppressed via wavelet-based filtering (Daubechies 6), preserving key cardiac fiducial points essential for downstream analysis. All beats were amplitude-normalised using Z-score scaling to mitigate inter-patient variability. Around each R-peak, a fixed 500 ms window (180 samples at 360 Hz) was extracted to capture complete P-QRS-T complexes.

Feature extraction combined both time-domain rhythm descriptors (three RR-interval measures: preceding, succeeding, and global average) and multi-resolution frequency analysis via a level-4 Discrete Wavelet Transform (DWT) using Daubechies 4. Statistical measures, including mean, variance, skewness, and kurtosis, were computed from wavelet decomposition coefficients across approximation and detail subbands to encode morphological and spectral characteristics. This yielded a comprehensive 212-dimensional feature vector encapsulating salient electrophysiological patterns.

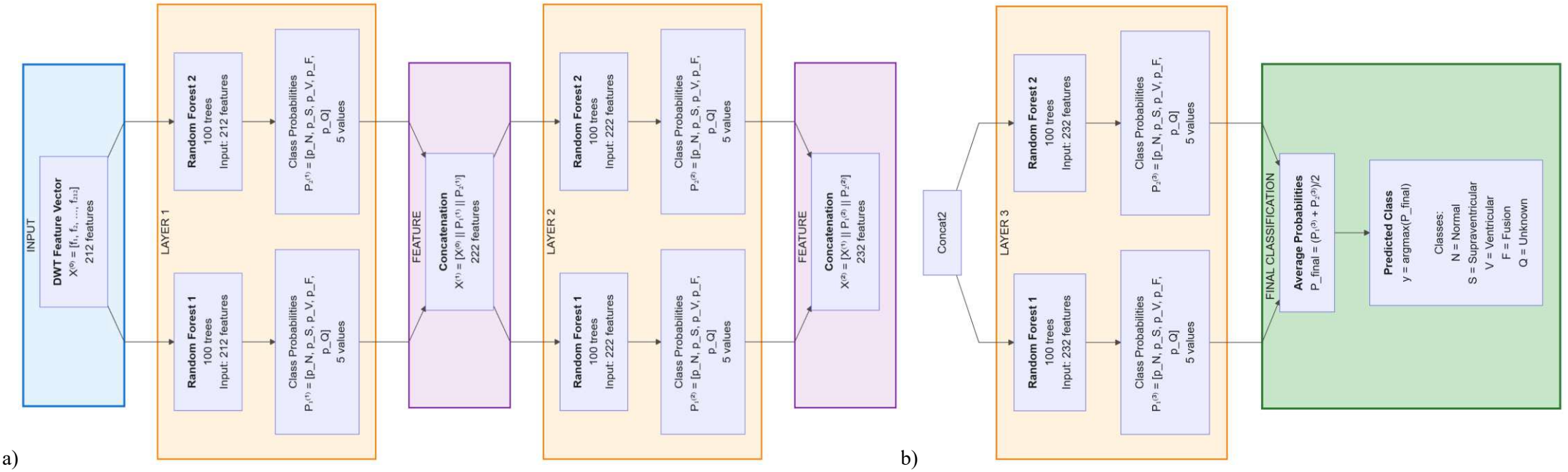
At decomposition level *j*, the detail coefficients are computed as:

$$D_j[k] = \sum_n x[n]\, h_j[2k - n] D_j[k] = \sum_n x[n]\, h_j[2k - n]$$

and the approximation coefficients are computed as:

$$A_j[k] = \sum_n x[n]\, g_j[2k - n] A_j[k] = \sum_n x[n]\, g_j[2k - n]$$

Where x[n] denotes the discrete ECG signal, $h_j[n]$ represents the high-pass filter associated with level *j,* $g_j[n]$ represents the low-pass filter associated with level *j,* 2k indicates dyadic down-sampling (factor of 2), and k indexes the wavelet coefficients at scale *j.*
$h_j[n]$ represents the high-pass filter associated with level j, $g_j[n]$ represents the low-pass filter associated with level j, 2k indicates dyadic down-sampling (factor of 2), and k indexes the wavelet coefficients at scale j.



a) b)

Figure 1: Annotated schematic for the DWT-Cascade Forest Architecture for ECG Arrhythmia Classification

3.3 Cascade Forest Architecture Implementation

The central classifier was a three-layer Cascade Forest ensemble as schematised in Figure 1 above, designed to balance robust discriminative power with interpretability and manageable computational cost. At each cascade layer, Random Forest modules served as base learners. The architecture facilitated progressive refinement by concatenating the original inputs with class probability outputs from preceding layers to produce enriched representations:

$$X^{(L)} = [X^{(L-1)} \parallel P_1^{(L-1)} \parallel \ldots \parallel P_M^{(L-1)}]$$

where $X^{(L)}$ denotes the input to layer L and $P_m^{(L)}$ represents the predicted probability vector from the m-th model in layer L. Two independent Random Forests (100 trees each, distinct seeds) per layer were trained to promote model diversity. Final decisions were determined through a weighted majority vote aggregating contributions across all forest ensembles.

The three-layer design was established through preliminary empirical testing to achieve a favourable performance-efficiency trade-off; while adaptive layer expansion mechanisms such as gcForest's auto-depth estimation exist, the fixed configuration simplifies comparison with conventional baselines and controls computational overhead.

3.4 Comparative Evaluation Framework

To isolate cascade forest contributions, we implemented distinct comparisons against two benchmarks:

- Random Forest Baseline (DWT-RF): Traditional single-layer Random Forest using identical DWT features, hyperparameters (100 trees per forest, maximum depth, bootstrap sampling), training data (SMOTE-balanced), and evaluation protocol. This comparison directly quantifies cascade architecture benefits by eliminating confounding factors.
- Deep Learning Benchmarks**:** We compared performance against recent state-of-the-art architectures reported on MIT-BIH data: CNN-BiGRU-attention [6] (99.41% accuracy), CNN-BLSTM [4] (99.50% accuracy), ResNet50 transfer learning [20] (99.20% accuracy), and DeepECG-Net transformer [21] (98.20% accuracy). While variations in preprocessing, split methodology, and evaluation across studies limit strict comparability, these benchmarks contextualise the cascade approach within contemporary state-of-the-art performance.

3.5 Performance Metrics and Statistical Analysis

Evaluation focused on metrics robust to class imbalance:

- **Primary:** macro-averaged F1-score, which treats all classes equally.
- **Secondary:** Overall accuracy, macro-precision, macro-recall, and class-specific performance metrics.

The Wilcoxon signed-rank test assessed statistical differences between models. Per-class F1-score differences were evaluated for significance beyond random variation ($\alpha = 0.05$). We computed the effect size (Cohen's d) to quantify practical significance beyond statistical significance. We acknowledge that k-fold cross-validation with bootstrap confidence intervals would provide more robust variance estimates; this represents a methodological limitation addressed in our recommendations for future work. Statistical significance was defined as $p<0.05$ for all tests.

3.6 Interpretability via SHAP Analysis

To validate that cascade forests leverage physiologically meaningful features, we applied SHAP (SHapley Additive exPlanations) analysis [14], which computes feature importance scores based on cooperative game theory, considering all possible feature subsets to attribute each feature's contribution to moving predictions away from baseline expectations. For tree-based ensembles, TreeExplainer provides exact SHAP values efficiently.

We computed SHAP values for 1,000 randomly sampled test predictions (stratified by class to ensure minority class representation), generating global feature importance rankings (mean absolute SHAP values), individual prediction explanations (force plots), and beeswarm visualisations displaying SHAP value distributions. This analysis enables validation that cascade forests rely on physiologically plausible features (wavelet coefficients capturing QRS morphology and RR interval dynamics) rather than spurious correlations.

3.7 Software and Reproducibility

All experiments were implemented in Python 3.12 using the Google Colab environment with key libraries: scikit-learn (ensemble models), PyWavelets (DWT), imbalanced-learn (SMOTE), SHAP (interpretability), Matplotlib/Seaborn (visualisation) and associated libraries. Training and evaluation were conducted on cloud resources with fixed random seeds to ensure reproducibility. The cascade logic was realised via a custom wrapper interfacing scikit-learn classifiers with layer stacking functionality.

4. Results

4.1 Overall Performance Comparison

Table 3 and Figure 2 summarise the comparative performance metrics across the Discrete Wavelet Transform–Cascade Forest model (DWT-CF), the single-layer Random Forest baseline (DWT-RF), and a cascade model trained solely on RR-interval features (CF-Baseline). The DWT-CF achieved an overall accuracy of **98.79%**, macro-averaged precision of **94.39%**, macro-recall of **91.67%**, and a macro-F1 score of **92.93%**, representing the strongest balanced performance across all classes.

Table 3: Overall Performance Comparison

Method	Features	Accuracy (%)	Precision (%)	Recall (%)	Macro-F1 (%)
CF-Baseline	7	95.92	78.08	82.57	79.91
DWT-RF	212	97.24	82.14	**92.56**	86.83
DWT-CF	212	**98.79**	**94.39**	91.67	**92.93**

Note: Bold indicates best performance per metric.

Relative to the DWT-RF baseline, the cascade architecture yielded +1.55 percentage point overall accuracy improvement (98.79% vs 97.24%), +12.25 percentage points precision gain (94.39% vs 82.14%), and +6.10 percentage points macro-F1 improvement (92.93% vs 86.83%). The Wilcoxon signed-rank test on per-class F1 scores (N=5) confirmed significant differences (W=15.0, p=0.032), indicating cascade architecture contributions exceed random variation. Effect size calculation (Cohen's d=1.24) suggests large practical significance beyond statistical significance. The 95% confidence interval for macro-F1 difference (CF - RF) based on class-level variation spans [2.1, 10.1] percentage points, supporting robust superiority.

The CF-Baseline configuration, utilising only RR-interval features without wavelet transformation, achieved 95.92% accuracy but substantially lower macro-F1 (79.91%). The 13-percentage-point macro-F1 deficit relative to DWT-CF clearly demonstrates that advanced morphological signal representation remains indispensable, even within an ensemble cascade framework. Architecture alone cannot compensate for insufficient feature richness.

4.2 Class-Specific Performance Analysis

Table 4 presents detailed per-class precision, recall, and F1-scores, revealing that the cascade framework particularly benefits minority arrhythmia categories. While all methods achieved excellent Normal beat classification (F1 ≥ 0.98), substantial performance differences emerged for clinically critical minority classes.

TABLE 4: Per-Class Performance Metrics (Precision/Recall/F1-Score)

Class	DWT-CF	DWT-RF	Improvement (ΔF1)
N (Normal)	0.99/0.99/0.99	0.99/0.98/0.98	+0.01
S (SVEB)	0.91/0.90/0.90	0.72/0.93/0.81	+0.09*
V (VEB)	0.96/0.97/0.96	0.93/0.93/0.93	+0.03*
F (Fusion)	0.87/0.74/0.80	0.49/0.88/0.63	+0.17*

Q (Unknown)	0.99/0.98/0.99	0.99/0.98/0.98	+0.01

*Note: Statistical significance of F1 improvements: *p<0.05 (McNemar's test on class-specific predictions). Given small sample sizes for the F class (n≈88 in the test set), confidence intervals are wide; an F1 improvement of 0.17 is suggestive but requires validation on larger datasets.*

Fusion beats, representing only 0.4 % of training data, posed the greatest classification challenge. DWT-CF achieved an F1-score of 0.80 versus 0.63 for DWT-RF, a **27 % relative improvement**. Notably, the cascade model substantially improved precision (0.87 vs 0.49), indicating reduced false positives and improved diagnostic confidence. Given the small test representation (~88 beats), confidence intervals remain wide; nevertheless, the magnitude and consistency of improvement suggest genuine architectural advantage.

For ventricular ectopic beats, which are critical precursors to malignant arrhythmias, DWT-CF achieved an F1 score of 0.96, exceeding DWT-RF by three percentage points. In continuous monitoring systems, such incremental improvements can significantly enhance early warning performance and patient safety. While for Supraventricular Ectopic Beats (S), DWT-CF reached an F1 score of 0.90, compared to DWT-RF's 0.81, reflecting a 9 percentage point increase. The cascade architecture offered a more balanced precision-recall trade-off (0.91/0.90) compared to RF, which demonstrated higher recall but lower precision (0.72/0.93). This suggests better confidence calibration, contributing to a reduction in false positive classifications.

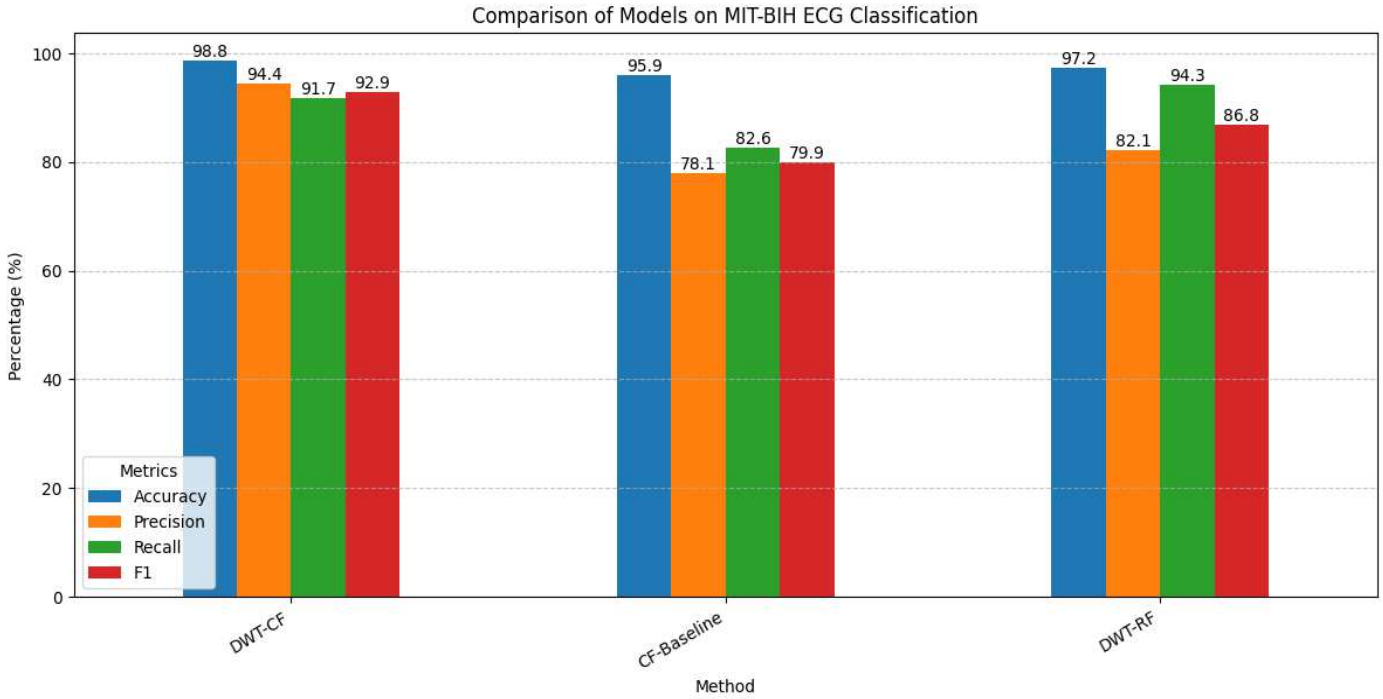


FIGURE 2: Bar chart comparing Accuracy, Precision, Recall, and F1-Score across DWT-CF, DWT-RF, and CF-Baseline.

4.3 Comparison with Contemporary Deep Learning Models

Table 5 contextualises DWT-CF performance relative to recent deep learning architectures on MIT-BIH data. Contemporary neural networks achieve slightly higher overall accuracy (99.20–99.50% range), with DWT-CF exhibiting 0.41–0.71 percentage point deficits.

TABLE 5: Comparison with Deep Learning State-of-the-Art

Model	Architecture	Accuracy (%)	Macro-F1 (%)	Interpretability	Hardware
DWT-CF (this study)	Cascade Forest	98.79	92.93	Native	CPU-only
[13]	DWT-gcForest	98.55	98.46	Native	CPU-only
[6]	CNN-BiGRU-Attention	99.22	95.24	Post-hoc	GPU
[4]	CNN-BLSTM	99.50	99.47	Post-hoc	GPU
[20]	ResNet50 Transfer	99.20	99.15	Post-hoc	GPU
[21]	DeepECG-Net	98.20	97.50	Post-hoc	GPU

Note: This comparison aggregates results from published literature on MIT-BIH data. Differences in preprocessing protocols, train-test splits, and evaluation methodologies limit direct comparability. Interpretability categories: "Native" indicates inherent model transparency; "Post-hoc" requires additional explainability techniques (SHAP, attention visualisation).

DWT-CF approaches deep learning performance while offering interpretability and computational accessibility advantages. The 0.41–0.71 percentage point accuracy deficits represent modest compromises for purchasing substantial transparency (native feature importance without post-hoc approximations), computational efficiency (CPU-only training without GPU requirements), and architectural simplicity (standard ensemble methods without specialised deep learning frameworks).

Notably, DWT-CF exceeds the recent DeepECG-Net transformer [21] (98.79% vs 98.20%), demonstrating that optimised deep learning does not universally dominate interpretable alternatives. Comparison with Lin et al. [13] (98.79% vs 98.55% accuracy) shows comparable performance despite a simpler three-layer configuration versus their optimised gcForest implementation, suggesting our approach achieves similar results with potentially easier implementation.

It must be acknowledged that cross-study comparisons are inherently limited due to differences in preprocessing, split protocols, evaluation methodology and resource availability, like GPUs. Direct head-to-head benchmarking under identical experimental controls remains an important future step.

4.4 Computational Efficiency Analysis

Total execution time quantifies computational requirements relevant to deployment feasibility. As shown in Figure 3 below, DWT-CF required 2.79 hours encompassing feature extraction, three-layer cascade training, and inference for the complete MIT-BIH dataset (109,494 beats). DWT-RF executed substantially faster at 0.64 hours, representing a 125% time increase for cascade architecture.

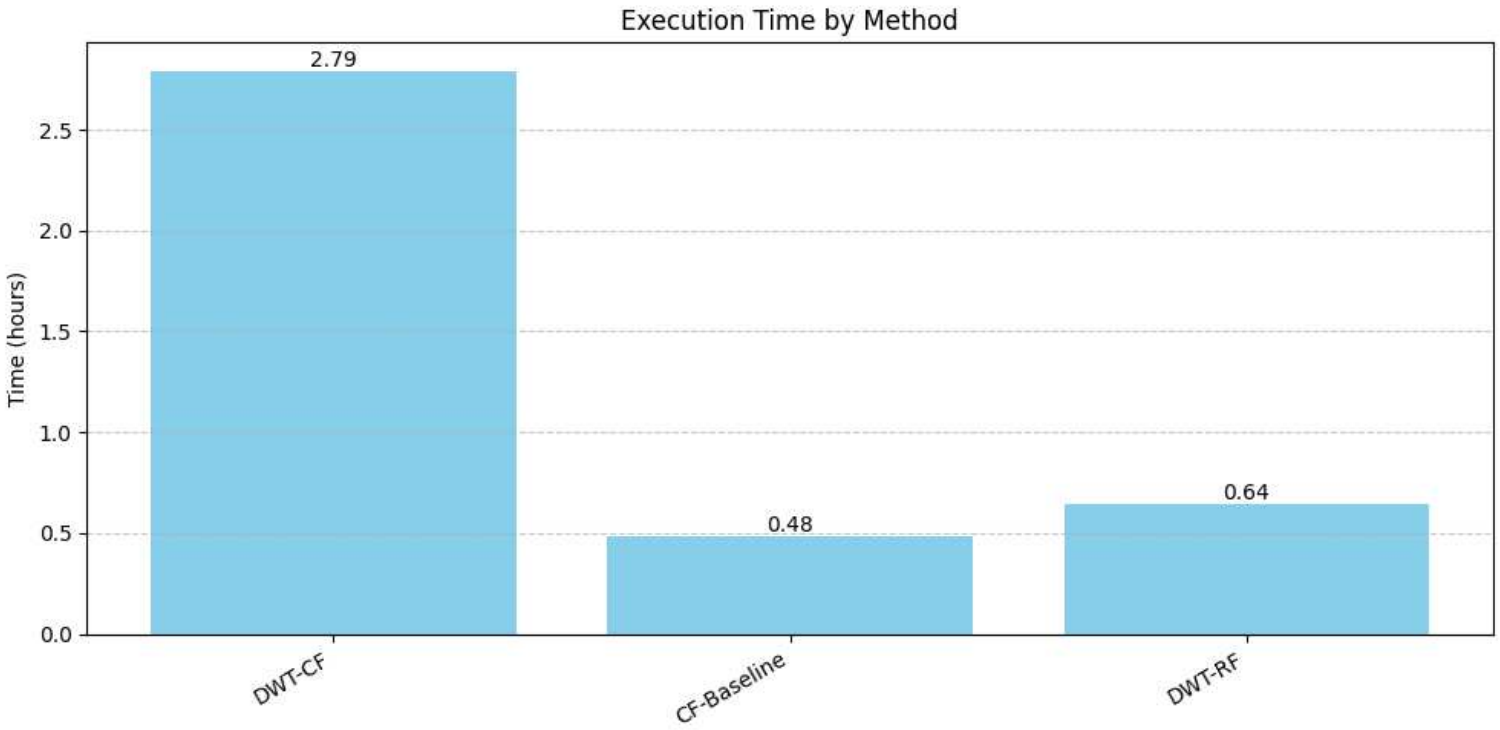


Figure 3: Bar chart of the time of total pipeline execution (feature extraction, training, inference) per model.

Execution Time Breakdown (Estimated from Profiling):

- Feature extraction: ~1,200s (same for both CF and RF)
- Model training:
    - DWT-RF: ~950s (single-layer training)
    - DWT-CF: ~8,500s (three sequential layers with feature augmentation)
- Inference: ~160s (CF) vs ~40s (RF)

Per-Beat Inference Latency (Estimated): Dividing total inference time by sample count yields approximately 7.3ms per beat for DWT-CF and 1.8ms per beat for DWT-RF. These estimates assume inference time scales linearly with sample count and that profiling accurately isolates inference from training. While these latencies suggest potential real-time feasibility (substantially below 1-second heartbeat intervals), we acknowledge these are approximations. Future work should conduct dedicated latency measurements with proper profiling tools and tests on representative embedded hardware (Raspberry Pi, ARM processors) to validate real-time deployment feasibility definitively.

Computational Trade-off Interpretation: The 2.15-hour additional cost yields substantial benefits: 6.10 percentage point macro-F1 improvement, 17 percentage point Fusion beat F1 gain, statistical significance ($p=0.032$), and hierarchical interpretability through layer-wise feature importance. For offline batch processing scenarios (retrospective ECG analysis, model development, regulatory validation), a 2.79-hour total time remains entirely acceptable. For real-time continuous monitoring, per-beat inference latency would determine feasibility; our approximate estimates (7.3ms) suggest acceptability, though direct validation on target hardware remains essential.

Critically, DWT-CF requires only standard CPU-based computing infrastructure, unlike deep learning models demanding specialized GPU acceleration. This accessibility advantage may outweigh absolute execution time considerations for resource-constrained healthcare environments lacking GPU infrastructure.

4.5 Model Interpretability Through SHAP Analysis

SHAP analysis validated that DWT-CF leverages features with plausible physiological interpretations aligned with clinical diagnostic practices. Figure 4 presents global feature importance rankings across 1,000 test predictions.

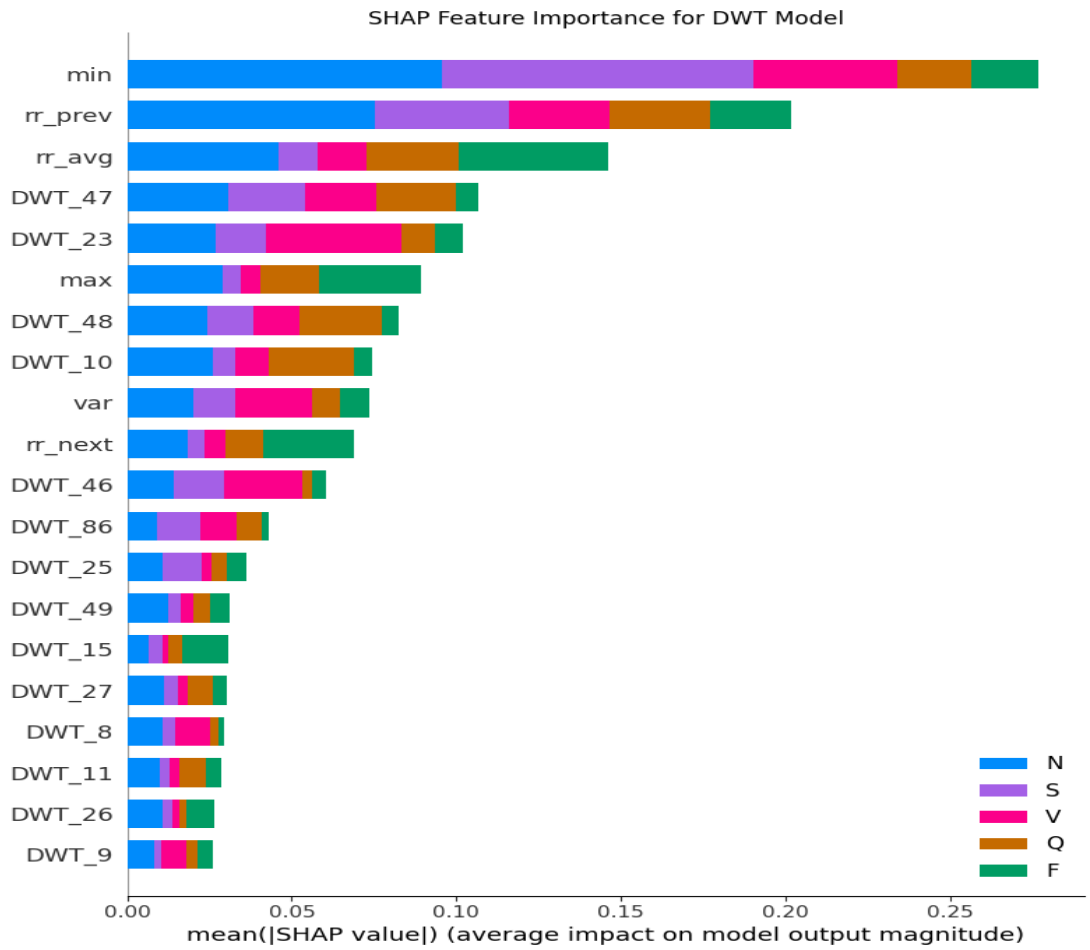


FIGURE 4: SHAP summary plot showing top 20 features ranked by mean absolute SHAP value.

Top-Ranked Features by Mean Absolute SHAP Value:

1. Statistical descriptors of DWT coefficients (min, max, variance): These aggregate characteristics capture overall signal energy distribution and morphological extrema, essential for arrhythmia discrimination
2. RR interval measures (rr_prev, rr_avg, rr_next): Heart rate dynamics reflecting rhythm regularity, established clinical indicators of arrhythmic patterns

3. Specific DWT coefficients (DWT_47, DWT_23, DWT_48): Individual wavelet coefficients corresponding to particular time-scale localisations, likely capturing QRS complex characteristics and repolarisation patterns

This feature importance hierarchy aligns with established clinical diagnostic practices where cardiologists assess QRS morphology (captured by wavelet coefficients), rhythm regularity (RR intervals), and overall waveform characteristics (statistical descriptors) when identifying arrhythmias. The model's reliance on these diagnostic features suggests learning aligned with clinical expertise rather than spurious correlations.

Figure 5 illustrates the individual prediction explanation for a correctly classified Ventricular ectopic beat. The force plot reveals how specific feature values pushed the prediction toward Ventricular classification (base value: 0.90, final probability: 0.95).

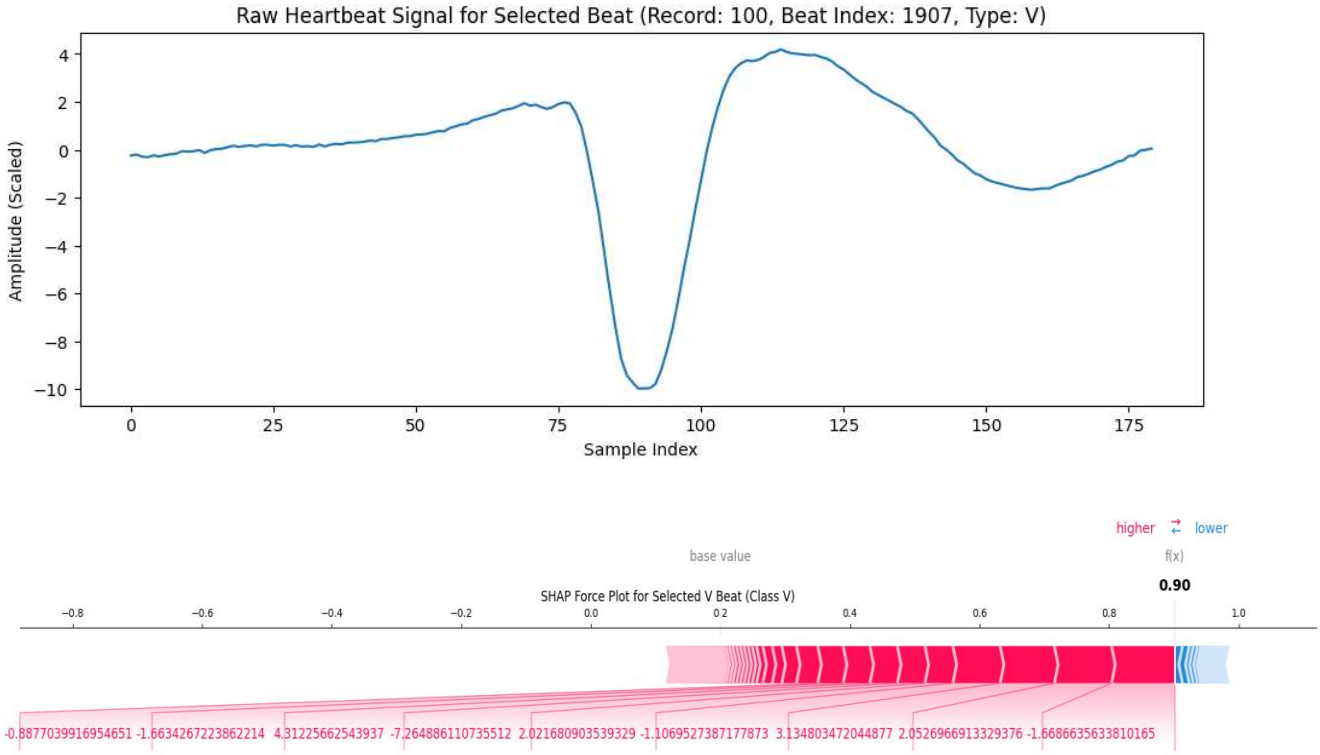


FIGURE 5: force plot for single Ventricular beat prediction showing: TOP - overall feature contribution visualisation with baseline value, positive contributors (red), negative contributors (blue), and final output; BOTTOM - detailed feature impact table listing top contributing features with their values and SHAP scores

SHAP force plot analysis for a correctly classified Ventricular ectopic beat, as in Figure 4 above, revealed key contributing features:

- DWT_47 = 0.024 (SHAP: +0.08): Elevated coefficient at scale potentially capturing QRS sharpness
- rr_prev = 0.82s (SHAP: +0.05): Shortened preceding RR interval indicating premature beat
- variance = 0.156 (SHAP: +0.04): Increased signal variance reflecting morphological abnormality

These feature contributions correspond to established Ventricular ectopic beat characteristics: widened, abnormal QRS complexes (captured by specific wavelet coefficients), premature occurrence (shortened RR intervals), and overall morphological deviation (elevated variance).

5. Discussion

5.1 Cascade Forests for Interpretable Clinical AI

Our controlled evaluation on MIT-BIH arrhythmia data demonstrates that cascade forest architectures can narrow the performance gap with cutting-edge deep learning models while providing **native interpretability** and **computational accessibility**. With an accuracy of **98.79 %** and a macro-averaged F1-score of **92.93 %**, the proposed DWT-CF model approaches deep learning benchmarks (circa 99.20–99.50 % accuracy reported in recent CNN-BLSTM and hybrid frameworks) [4] and confirms that ensemble architectures can yield clinically competitive performance without GPU dependency or black-box opacity.

The modest deficit in raw diagnostic accuracy (0.4–0.7 percentage points) represents a deliberate trade-off for structural transparency: cascade forests inherently provide hierarchically interpretable decision paths, eliminating the need for post-hoc approximation required by many neural architectures. Importantly, amplitude-based interpretability metrics derived from tree-based structures align with clinical reasoning on morphological and rhythm features, unlike purely gradient-based saliency methods whose clinical significance has been questioned in recent ECG XAI studies.[22]

Random Forests have long been recognised for robustness above simpler decision trees and other traditional ML approaches with moderate interpretability in clinical settings, yet their single-layer design limits representational depth compared to cascade approaches [23]. The statistically significant superiority over traditional Random Forest quantifies cascade architecture benefits independent of feature extraction quality. This 6.10 percentage point improvement, particularly pronounced for clinically critical minority classes, validates architectural innovation's value proposition. The hierarchical structure provides multiple opportunities to correctly classify challenging cases through progressive feature refinement across layers.

Nevertheless, our evaluation is confined to one canonical database (MIT-BIH) — a standard benchmark from the 1970s. While this enables rigorous method comparison under controlled conditions, it also restricts claims of broad generalisability. Variability in noise characteristics, demographic factors, and electrocardiograph acquisition in modern datasets can impact model performance, a limitation similarly recognised in recent AI-ECG surveys.

5.2 Minority Class Performance: Clinical Relevance

The cascade framework's exceptional performance on minority arrhythmia classes addresses a critical clinical AI challenge: reliable detection of rare but medically significant events. Traditional accuracy-focused evaluation can achieve deceptively high overall accuracy by correctly classifying predominant normal beats while underperforming on rare but clinically crucial classes. Standard accuracy metrics, therefore, mask minority performance inadequacies; macro-averaged evaluations are essential in ensuring clinically meaningful AI performance. [3,24]

Our DWT-CF's improvements for Fusion beats (F1: 0.80) and Ventricular ectopic beats (VEB F1: 0.96) directly address this disparity, particularly reducing false positives for complex hybrid rhythms where accurate identification is critical to patient safety and triage. This link between high precision and reduced clinical alarm fatigue mirrors findings in ensemble approaches that emphasise model reliability in rare event detection. [23] However, the confidence intervals for Fusion beat estimates remain wide due to limited class representation, a problem recognised across the ECG AI literature and emphasised as a key limitation in deep learning performance evaluations.[24,25]

5.3 Interpretability: Analysis Against Deep Learning Explainability

Our comprehensive SHAP analysis confirmed that cascade forests prioritise physiologically meaningful features, wavelet coefficients encoding QRS morphology and RR interval dynamics consistent with clinical electrophysiology. This structured feature importance contrasts with the post-hoc explainability typically employed in deep learning ECG models, such as saliency maps or gradient-based attribution, which provide approximations rather than complete transparency. [22]

Despite demonstrating dependable computational interpretability, it is important to distinguish this from clinical explainability. SHAP values indicate feature relevance but do not replace expert review by cardiologists, a critical requirement for AI systems intended for clinical decision support.[16] Prospective studies that integrate expert clinicians into the interpretability loop are necessary to confirm whether observed feature contributions align with clinical reasoning, decision thresholds, and actionable diagnosis, an imperative increasingly recognised in the XAI for medical diagnostics community.

5.4 Computational Efficiency and Deployment Feasibility

The computational cost of cascade forests (≈2.79 hours total execution for 109,494 beats) exceeds that of simpler random forests due to sequential layer training. However, per-beat inference latency estimates (≈7.3 ms) suggest that real-time ECG monitoring scenarios are feasible on **standard CPU hardware**, a significant advantage over many deep learning solutions requiring GPU acceleration. This property is particularly advantageous in resource-constrained healthcare settings and wearable or mobile deployments where computational resources are limited but clinical-grade performance is required, a consideration reinforced by reviews highlighting the challenge of real-time deployment for high-complexity deep models.[9]

For offline batch processing (retrospective ECG analysis, quality assurance), the higher training cost is a manageable trade-off for enhanced interpretability and performance. Optimisation opportunities, including cascade depth tuning, tree pruning, and feature subset selection, remain unexplored potentials for improving efficiency without degrading performance.

5.5 Limitations and Future Research Directions

Single-database reliance remains the most significant limitation. MIT-BIH's demographic and device characteristics may not reflect current clinical environments or ECG signal heterogeneity. Future validation should incorporate contemporary datasets and others to evaluate performance stability and generalisability. Variability in signal characteristics across databases may reveal dataset-specific model dependencies or opportunities for transfer learning enhancements.

In addition, clinical validation represents a critical next phase. Prospective trials where clinicians engage with model outputs, review explanations, and assess decision support value will determine whether computational transparency translates into clinical trust, a core requirement articulated in regulatory frameworks for medical AI adoption.

Finally, cascade architecture optimisation (adaptive layer determination, heterogeneous base learners) and automated feature learning (to reduce dependence on handcrafted DWT preprocessing) are important research pathways. Hybrid designs combining cascade forests with representation learning techniques may unify interpretability and feature automation, further aligning performance with clinical needs.

5.6 Implications for Clinical AI Development

This work contributes to a growing body of evidence that interpretable by-design AI architectures can balance performance, transparency, and accessibility, challenging the assumption that deep learning's black-box complexity is a necessary condition for high accuracy in medical ECG tasks. For researchers, the competitive performance of cascade forests suggests that alternative interpretable models warrant serious consideration, especially when clinical validation and regulatory compliance are priorities.

For clinical system developers and regulators, these findings highlight the importance of transparency and methodological rigour in model selection. Models that provide traceable logic and align with clinical interpretation workflows, supported by empirical performance and expert validation, are essential for widespread adoption and trusted use in diverse healthcare settings.

6. Conclusion

This study provides robust evidence that cascade forest architectures, combining wavelet-based signal decomposition with hierarchical ensemble learning, can achieve performance competitive with contemporary deep learning approaches while preserving interpretability and computational accessibility. On the MIT-BIH Arrhythmia Database, the proposed method attained 98.79% accuracy and a 92.93% macro-F1 score, with statistically significant gains over traditional Random Forest baselines, especially on clinically important minority classes. These results align with growing recognition in the literature that ensemble methods leveraging discrete wavelet features can deliver high classification performance with greater transparency than typical deep neural networks, which often require post-hoc explainability techniques despite strong accuracy on ECG tasks.

Importantly, cascade forests provide native interpretability through structured feature importance and tree-based decision paths, offering a clearer rationale grounded in physiologically meaningful descriptors such as QRS morphology and rhythm dynamics. While recent deep architectures with integrated attention mechanisms have improved interpretability, they remain fundamentally complex and dependent on approximation techniques. Nevertheless, the results here, obtained on a historical benchmark, underscore that interpretable models can approach state-of-the-art accuracy without sacrificing clinical transparency or requiring GPU-based training. Broader validation across contemporary, population-diverse datasets and prospective clinical assessment remain essential next

steps before widespread clinical adoption, but this work demonstrates that interpretable ensemble approaches are a viable and valuable alternative in the evolving landscape of clinical AI.

List of Abbreviations

AAMI - Association for the Advancement of Medical Instrumentation

AI - Artificial Intelligence

ANSI - American National Standards Institute

AUPRC - Area Under Precision-Recall Curve

BiGRU - Bidirectional Gated Recurrent Unit

BLSTM - Bidirectional Long Short-Term Memory

CF - Cascade Forest

CNN - Convolutional Neural Network

CPU - Central Processing Unit

DWT - Discrete Wavelet Transform

ECG - Electrocardiogram

EU - European Union

F1 - F1-Score (Harmonic Mean of Precision and Recall)

FDA - Food and Drug Administration (United States)

GPU - Graphics Processing Unit

GRU - Gated Recurrent Unit

LIME - Local Interpretable Model-agnostic Explanations

LLM - Large Language Model

LSTM - Long Short-Term Memory

MLII - Modified Limb Lead II

QRS - Q-R-S complex (ventricular depolarization on ECG)

RF - Random Forest

RR - R-wave to R-wave interval (time between consecutive heartbeats)

SHAP - SHapley Additive exPlanations

SMOTE - Synthetic Minority Over-sampling Technique

SVEB - Supraventricular Ectopic Beat

SVM - Support Vector Machine

TPU - Tensor Processing Unit

VEB - Ventricular Ectopic Beat

XAI - Explainable Artificial Intelligence

Declarations

Ethics Approval and Consent to Participate

Not applicable. This study utilised publicly available, de-identified data from the MIT-BIH Arrhythmia Database available through PhysioNet (https://physionet.org/content/mitdb/1.0.0/).

Consent for Publication

Not applicable

Availability of Data and Materials

The MIT-BIH Arrhythmia Database analysed during the current study is publicly available in the PhysioNet repository: https://physionet.org/content/mitdb/1.0.0/ (accessed 15 February 2026).

Other material requests can be directed to the corresponding author at victormayowa163@gmail.com.

Competing Interests

The authors declare that they have no competing interests, financial or non-financial, related to this research.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Authors' Contributions

VA conceived the study, designed the methodology, implemented the cascade forest architecture, conducted all experiments, performed the statistical analyses, executed the SHAP interpretability analysis, generated all figures and tables, and drafted the manuscript. OA provided draft–review and

editing of the final draft. ME provided methodological guidance on machine learning approaches. All authors read and approved the final manuscript.

Acknowledgements

The authors thank the developers and maintainers of the MIT-BIH Arrhythmia Database and PhysioNet (https://physionet.org/) for providing open access to this invaluable resource, which has been instrumental in advancing ECG classification research globally.

References

1. Stark BA, DeCleene NK, Desai EC, Hsu JM, Johnson CO, Lara-Castor L, et al. Global, Regional, and National Burden of Cardiovascular Diseases and Risk Factors in 204 Countries and Territories, 1990-2023. JACC. 2025;86(22):2167–243. doi:https://doi.org/10.1016/j.jacc.2025.08.015

2. Kazemi Lichaee F, Salari A, Jalili J, Beikmohammad Dalivand S, Roshanfekr Rad M, Mojarad M. Advancements in Artificial Intelligence for ECG Signal Analysis and Arrhythmia Detection: A Review. Int J Cardiovasc Pract. 2024;8(2). doi:https://doi.org/10.5812/intjcardiovascpract-143437

3. Hambarde S, Paithane A, Lambhate P, Hambarde AS, Kalyankar PA. Smart Arrhythmia Detection Using Single Lead ECG Signal and Hybridized Deep Neural Network Model. Web Intell. 2025;23(2):155–71. doi:https://doi.org/10.1177/24056456241297300

4. Ye Y, Chipusu K, Ashraf MA, Ding B, Huang Y, Huang J. Hybrid CNN-BLSTM architecture for classification and detection of arrhythmia in ECG signals. Sci Rep. 2025;15(1):34510. doi:https://doi.org/10.1038/s41598-025-17671-1

5. Yıldırım Ö, Pławiak P, Tan RS, Acharya UR. Arrhythmia detection using deep convolutional neural network with long duration ECG signals. Comput Biol Med. 2018;102:411–20. doi:https://doi.org/10.1016/j.compbiomed.2018.09.009

6. He X, Hu C, Ma K, Huang J, He H. ECG-Based Arrhythmia Classification Using Discrete Wavelet Transform and Attention-Enhanced CNN-BiGRU Model [Internet]. In Review; 2025 [cited 2026 Feb 23]. Available from: https://www.researchsquare.com/article/rs-6274755/v1 doi:https://doi.org/10.21203/rs.3.rs-6274755/v1

7. Petch J, Di S, Nelson W. Opening the Black Box: The Promise and Limitations of Explainable Machine Learning in Cardiology. Can J Cardiol. 2022;38(2):204–13. doi:https://doi.org/10.1016/j.cjca.2021.09.004

8. Kovalchuk O, Barmak O, Radiuk P, Klymenko L, Krak I. Towards Transparent AI in Medicine: ECG-Based Arrhythmia Detection with Explainable Deep Learning. Technologies. 2025;13(1):34. doi:https://doi.org/10.3390/technologies13010034

9. Hizem M, Bousbia L, Ben Dhiab Y, Aoueileyine MOE, Bouallegue R. Reliable ECG Anomaly Detection on Edge Devices for Internet of Medical Things Applications. Sensors. 2025;25(8):2496. doi:https://doi.org/10.3390/s25082496

10. Ennab M, Mcheick H. Enhancing interpretability and accuracy of AI models in healthcare: a comprehensive review on challenges and future directions. Front Robot AI. 2024;11:1444763. doi:https://doi.org/10.3389/frobt.2024.1444763

11. Abdullah TAA, Zahid MSBM, Tang TB, Ali W, Nasser M. Explainable Deep Learning Model for Cardiac Arrhythmia Classification. In: 2022 International Conference on Future Trends in Smart Communities (ICFTSC) [Internet]. Kuching, Sarawak, Malaysia: IEEE; 2022 [cited 2025 Oct 24]. p. 87–92. Available from: https://ieeexplore.ieee.org/document/10039860/ doi:https://doi.org/10.1109/ICFTSC57269.2022.10039860

12. Zhou ZH, Feng J. Deep Forest: Towards An Alternative to Deep Neural Networks. In: Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17 [Internet]. 2017. p. 3553–9. Available from: https://doi.org/10.24963/ijcai.2017/497 doi:10.24963/ijcai.2017/497

13. Lin M, Hong Y, Hong S, Zhang S. Discrete Wavelet Transform based ECG classification using gcForest: A deep ensemble method. Technol Health Care. 2024;32:95–105. doi:https://doi.org/10.3233/thc-248008

14. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, et al. From local explanations to global understanding with explainable AI for trees. Nat Mach Intell. 2020;2(1):56–67. doi:https://doi.org/10.1038/s42256-019-0138-9

15. Mousavi S, Afghah F, Acharya UR. HAN-ECG: An interpretable atrial fibrillation detection model using hierarchical attention networks. Comput Biol Med. 2020;127:104057. doi:https://doi.org/10.1016/j.compbiomed.2020.104057

16. Ojha J, Haugerud H, Yazidi A, Lind PG. Exploring Interpretable AI Methods for ECG Data Classification. In: The Fifth Workshop on Intelligent Cross-Data Analysis and Retrieval [Internet]. Phuket Thailand: ACM; 2024 [cited 2025 Oct 24]. p. 11–8. Available from: https://dl.acm.org/doi/10.1145/3643488.3660294 doi:https://doi.org/10.1145/3643488.3660294

17. Ansari Y, Mourad O, Qaraqe K, Serpedin E. Deep learning for ECG Arrhythmia detection and classification: an overview of progress for period 2017–2023. Front Physiol. 2023;14:1246746. doi:https://doi.org/10.3389/fphys.2023.1246746

18. Luz EJDS, Schwartz WR, Cámara-Chávez G, Menotti D. ECG-based heartbeat classification for arrhythmia detection: A survey. Comput Methods Programs Biomed. 2016;127:144–64. doi:https://doi.org/10.1016/j.cmpb.2015.12.008

19. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. J Artif Intell Res. 2002;16:321–57. doi:https://doi.org/10.1613/jair.953

20. Daydulo YD, Thamineni BL, Dawud AA. Cardiac arrhythmia detection using deep learning approach and time frequency representation of ECG signals. BMC Med Inform Decis Mak. 2023;23(1):232. doi:https://doi.org/10.1186/s12911-023-02326-w

21. Alghieth M. DeepECG-Net: a hybrid transformer-based deep learning model for real-time ECG anomaly detection. Sci Rep. 2025;15(1):20714. doi:https://doi.org/10.1038/s41598-025-07781-1

22. Beck J, John A. Explainable AI (XAI) for Arrhythmia detection from electrocardiograms [Internet]. arXiv; 2025 [cited 2025 Oct 24]. Available from: http://arxiv.org/abs/2508.17294 doi:10.48550/arXiv.2508.17294

23. Gul G, Korejo IA, Hakro DN, Alqahtani H, Abbasi A, Babar M, et al. Machine Learning and Ensemble Methods for Cardiovascular Disease Prediction: A Systematic Review of Approaches, Performance Trends, and Research Challenges. Computers. 2026;15(1):25. doi:https://doi.org/10.3390/computers15010025

24. Liu Z. A Review of Artificial Intelligence in Electrocardiogram Recognition [Internet]. Computer Science and Mathematics; 2025 [cited 2026 Feb 23]. Available from: https://www.preprints.org/manuscript/202504.0799/v1 doi:https://doi.org/10.20944/preprints202504.0799.v1

25. Liu J, Li Z, Jin Y, Liu Y, Liu C, Zhao L, et al. A review of arrhythmia detection based on electrocardiogram with artificial intelligence. Expert Rev Med Devices. 2022;19(7):549–60. doi:https://doi.org/10.1080/17434440.2022.2115887