v7.n10.2026

SCIENTIFIC JOURNAL - RECIMA21 ISSN 2675-6218

COMPARATIVE EVALUATION OF DEEP LEARNING MODELS AND EXPLAINABILITY METHODS FOR ECG HEARTBEAT CLASSIFICATION

COMPARAÇÃO DE MODELOS DE APRENDIZADO PROFUNDO E MÉTODOS DE EXPLICABILIDADE NA CLASSIFICAÇÃO DE BATIMENTOS DE ECG

COMPARACIÓN DE MODELOS DE APRENDIZAJE PROFUNDO Y MÉTODOS DE EXPLICABILIDAD PARA LA CLASIFICACIÓN DE LATIDOS DE ECG

Jessica Fernandes Lopes[1], Felipe Torquato Bressan[2], Rafael Staiger Bressan[3], Danilo Sipoli Sanches[4]

e7109152
https://doi.org/10.47820/recima21.v7i10.9152

PUBLISHED: 10/2026

ABSTRACT

This study compared a 1D-CNN architecture and a hybrid CNN-BiLSTM-Attention model for multiclass electrocardiographic heartbeat classification, incorporating Explainable Artificial Intelligence techniques to evaluate the faithfulness and temporal distribution of the evidence used by the models. Experiments used a dataset derived from the MIT-BIH Arrhythmia Database, organized into the Normal, Supraventricular, Ventricular, Fusion, and Unclassified classes. The protocol employed five-fold stratified cross-validation exclusively on the training set, epoch selection based on Macro F1 in an internal validation partition, and final evaluation on an untouched holdout test set. The 1D-CNN achieved a test Macro F1 of 0.8681 versus 0.8140 for the CNN-BiLSTM-Attention model, whereas the hybrid model achieved higher Balanced Accuracy, 0.9441 versus 0.9264, indicating higher average class-wise recall. Explainability analysis employed 1D Grad-CAM, Integrated Gradients, and temporal attention on a class-balanced subset of samples correctly classified by both architectures. Under the adopted perturbation protocol, Integrated Gradients showed positive mean faithfulness gains across all classes for both architectures, whereas Grad-CAM and temporal attention exhibited more class-dependent behavior. Low-to-moderate agreement among the methods indicated that different techniques may emphasize distinct temporal regions for the same correct prediction. Within this experimental setting, the additional complexity of the hybrid model did not translate into uniformly better predictive performance, supporting the complementary evaluation of predictive performance and explainability in ECG classification.

KEYWORDS: *Electrocardiography. Arrhythmias, Cardiac. Deep Learning. Explainable Artificial Intelligence. Multiclass classification.*

RESUMO

Este estudo comparou uma arquitetura 1D-CNN e um modelo híbrido CNN-BiLSTM-Attention para a classificação multiclasse de batimentos eletrocardiográficos, incorporando técnicas de Inteligência Artificial Explicável para avaliar a fidelidade e a distribuição temporal das evidências utilizadas pelos modelos. Os experimentos utilizaram um conjunto derivado da MIT-BIH Arrhythmia Database, organizado nas classes Normal, Supraventricular, Ventricular, Fusion e Unclassified. O protocolo empregou validação cruzada estratificada em cinco folds exclusivamente no conjunto de treinamento, seleção de época por Macro F1 em validação interna

[1] Ph.D. in Electrical Engineering from the Londrina State University (UEL), Brazil.
[2] Undergraduate Medical Student at Educational Foundation of the Municipality of Assis (FEMA), Brazil.
[3] M.Sc. in Computer Science from Federal University of Technology – Paraná (UTFPR), Brazil.
[4] Ph.D. in Electrical Engineering from the University of São Paulo (USP), Brazil.

This article is published in open access under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are credited.

e avaliação final em um conjunto de teste mantido como holdout. A 1D-CNN obteve Macro F1 de 0,8681 no teste, contra 0,8140 da CNN-BiLSTM-Attention, enquanto o modelo híbrido apresentou maior Balanced Accuracy, 0,9441 contra 0,9264, indicando maior recall médio entre as classes. A análise explicativa utilizou Grad-CAM 1D, Integrated Gradients e atenção temporal em um subconjunto balanceado de casos corretamente classificados por ambas as arquiteturas. Sob o protocolo de perturbação adotado, Integrated Gradients apresentou ganhos médios de fidelidade positivos em todas as classes para ambas as arquiteturas, enquanto Grad-CAM e atenção temporal exibiram comportamento mais dependente da classe. A concordância baixa a moderada entre os métodos indicou que diferentes técnicas podem destacar regiões temporais distintas para a mesma decisão correta. Neste cenário experimental, a maior complexidade do modelo híbrido não resultou em desempenho preditivo uniformemente superior, reforçando a avaliação complementar de desempenho e explicabilidade na classificação de ECG.

PALAVRAS-CHAVE: Eletrocardiografia. Arritmias Cardíacas. Aprendizagem Profunda. Inteligência Artificial Explicável. Classificação multiclasse.

RESUMEN

Este estudio comparó una arquitectura 1D-CNN y un modelo híbrido CNN-BiLSTM-Attention para la clasificación multiclase de latidos electrocardiográficos, incorporando técnicas de Inteligencia Artificial Explicable para evaluar la fidelidad y la distribución temporal de las evidencias utilizadas. Los experimentos utilizaron un conjunto derivado de la MIT-BIH Arrhythmia Database, organizado en las clases Normal, Supraventricular, Ventricular, Fusion y Unclassified. El protocolo empleó validación cruzada estratificada de cinco folds exclusivamente sobre el conjunto de entrenamiento, selección de época mediante Macro F1 en validación interna y evaluación final en un conjunto de prueba holdout. La 1D-CNN obtuvo un Macro F1 de 0,8681 en la prueba, frente a 0,8140 de la CNN-BiLSTM-Attention, mientras que el modelo híbrido presentó mayor exactitud balanceada, 0,9441 frente a 0,9264, indicando mayor sensibilidad media entre clases. El análisis explicativo utilizó Grad-CAM 1D, Integrated Gradients y atención temporal sobre un subconjunto balanceado por clase de muestras correctamente clasificadas por ambas arquitecturas. Bajo el protocolo de perturbación adoptado, Integrated Gradients presentó ganancias medias de fidelidad positivas en todas las clases para ambas arquitecturas, mientras que Grad-CAM y la atención temporal mostraron un comportamiento más dependiente de la clase. La concordancia baja a moderada entre los métodos indicó que diferentes técnicas pueden resaltar regiones temporales distintas para una misma predicción correcta. En este escenario experimental, la mayor complejidad del modelo híbrido no produjo un desempeño predictivo uniformemente superior, respaldando la evaluación complementaria del desempeño y la explicabilidad en la clasificación de ECG.

PALABRAS CLAVE*: Electrocardiografía. Arritmias Cardíacas. Aprendizaje Profundo. Inteligencia Artificial Explicable. Clasificación multiclase.*

1. INTRODUCTION

Electrocardiogram (ECG) analysis plays a central role in the identification of abnormalities in cardiac rhythm, conduction, and heartbeat morphology. However, the increasing volume of ECG data generated by monitoring systems makes manual analysis increasingly demanding, particularly in continuous or large-scale applications. This scenario has motivated the

This article is published in open access under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are credited.

development of automated approaches capable of supporting ECG classification while maintaining reliable performance across different heartbeat categories.

Deep learning enables discriminative representations to be learned directly from ECG signals. One-dimensional convolutional neural networks (1D-CNNs) capture local temporal and morphological patterns (Kiranyaz; Ince; Gabbouj, 2016), while deeper architectures have achieved strong results in arrhythmia detection (Hannun et al., 2019). Bidirectional Long Short-Term Memory (BiLSTM) networks and temporal attention can extend these approaches by modeling sequential dependencies and weighting internal representations.

A second challenge concerns the interpretability of model decisions. High predictive performance alone does not reveal which ECG regions contributed to a classification, motivating explainable artificial intelligence (XAI) approaches for examining model behavior. This concern is consistent with the need for transparency and human oversight in health applications (World Health Organization, 2021).

Despite the increasing use of recurrent and attention-based architectures for ECG classification, it remains unclear whether their additional complexity translates into consistent gains over simpler convolutional models when both are evaluated under the same data partitions and training conditions. A similar issue arises in explainability: visually plausible relevance patterns do not necessarily indicate that the highlighted regions influence the classifier's decision, and different explanation mechanisms may emphasize different portions of the same signal. These questions are particularly important in imbalanced multiclass settings, where aggregate performance can conceal substantial differences among classes.

This study aimed to compare the predictive performance and explanatory behavior of a compact 1D-CNN and a CNN-BiLSTM architecture with temporal attention for multiclass ECG heartbeat classification under equivalent data partitions and training conditions. Specifically, the study compared global and class-specific predictive metrics, evaluated explanation faithfulness through matched perturbations, assessed agreement among the applicable explanation mechanisms, and examined attribution over observed and zero-padded regions.

2. BACKGROUND AND RELATED WORK

The electrocardiogram (ECG) provides a temporal representation of cardiac electrical activity and is a fundamental source of information for detecting abnormalities in rhythm, conduction, and heartbeat morphology. The MIT-BIH Arrhythmia Database has become one of the main benchmarks for automated heartbeat classification and has supported decades of

This article is published in open access under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are credited.

research on signal processing and machine learning methods (Moody; Mark, 2001). Despite substantial advances, ECG classification remains challenging because heartbeat categories are typically highly imbalanced, allowing strong aggregate performance to coexist with substantially weaker recognition of less frequent classes. Recent reviews confirm the continued relevance of rigorous performance evaluation and methodological comparability in AI-based arrhythmia detection (Ayyub; Politis; Usman, 2025).

Deep learning has reduced dependence on handcrafted feature extraction by enabling discriminative representations to be learned directly from ECG signals. One-dimensional convolutional neural networks are particularly effective at capturing local temporal and morphological patterns and have established a strong foundation for heartbeat classification (Kiranyaz; Ince; Gabbouj, 2016). Larger deep networks have further demonstrated the potential of learned ECG representations for automated arrhythmia detection (Hannun et al., 2019). However, recent analyses emphasize that increasingly complex models may impose substantial computational costs, reinforcing the need to assess predictive gains together with model efficiency and practical applicability (Gupta et al., 2024).

Recurrent architectures provide a complementary mechanism for modeling sequential dependencies. Long Short-Term Memory networks were introduced to mitigate limitations in learning long-range dependencies in conventional recurrent networks (Hochreiter; Schmidhuber, 1997), while bidirectional recurrent models incorporate information from both directions of a sequence (Schuster; Paliwal, 1997). Combined CNN-BiLSTM architectures can therefore integrate local morphology with broader temporal context. Temporal aggregation can be further enhanced through additive attention, in which learned scores determine the relative contribution of sequential representations to the final context vector, following the general attention principle introduced by Bahdanau, Cho and Bengio (2015). These additions increase representational capacity, but also model complexity, and their advantage over simpler convolutional models must therefore be demonstrated under equivalent experimental conditions.

Kachuee, Fazeli and Sarrafzadeh (2018) used residual convolutional networks for five-class ECG heartbeat classification. More recently, Rani, Gupta and Mittal (2026) compared convolutional, recurrent and attention-based models using preprocessed 187-sample heartbeats organized into the same five classes. Their evaluation emphasized predictive performance and probabilistic calibration, whereas the present study examines perturbation-based faithfulness and agreement among explanation mechanisms. Evaluation design is equally important. Luz and Menotti (2011) showed that using heartbeats from the same patients for training and testing can favor reported classification performance. In a different setting involving 12-lead ECG datasets,

This article is published in open access under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are credited.

Leinonen et al. (2024) found that conventional K-fold cross-validation overestimated performance on unseen data sources, whereas leave-source-out evaluation provided less biased estimates. Together, these findings motivate distinguishing performance within a dataset from generalization to new patients and data sources.

Interpretability represents a parallel challenge. High predictive performance does not establish which components of an ECG contributed to a particular decision, an important limitation in biomedical applications where transparency, accountability, and human oversight are relevant requirements (World Health Organization, 2021). Explainable artificial intelligence (XAI) seeks to provide mechanisms for inspecting model behavior, but recent systematic reviews indicate that ECG explainability remains methodologically fragmented, with substantial heterogeneity in methods, evaluation criteria, and validation procedures (Manimaran et al., 2025; Taleban et al., 2026). This lack of standardized quantitative assessment limits direct comparison among explanation methods and weakens conclusions based exclusively on visual inspection.

The XAI techniques investigated in this study represent different explanatory mechanisms. Grad-CAM derives class-specific relevance from convolutional activations and their gradients, producing a temporal relevance profile when adapted to one-dimensional networks (Selvaraju et al., 2017). Integrated Gradients instead attributes importance directly to input positions by integrating gradients along a path between a reference baseline and the observed sample (Sundararajan; Taly; Yan, 2017). Because it operates at the input level, Integrated Gradients also provides a common attribution framework for comparing structurally different architectures. Temporal attention offers a third perspective by exposing the learned weighting of internal sequential representations. However, attention weights should not automatically be interpreted as explanations of model decisions, since high attention does not establish causal relevance of the corresponding input region (Jain; Wallace, 2019).

A central distinction in XAI concerns plausibility and faithfulness. An attribution may appear visually meaningful while having limited influence on the actual classifier output. Recent ECG-specific reviews identify fidelity, robustness, stability, and quantitative validation as important unresolved issues for reliable interpretation (Manimaran et al., 2025; Taleban et al., 2026). Perturbation-based evaluation can address part of this problem by measuring how model evidence changes when highly attributed positions are altered and comparing that response with matched controls. However, perturbation procedures must preserve the temporal nature of ECG data, since approaches designed for independent structured features may be inappropriate for sequential signals (Taleban et al., 2026). Accordingly, perturbation is treated here as an operational test of explanation faithfulness rather than as evidence of clinical or causal validity.

This article is published in open access under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are credited.

Interpretation is additionally affected by the representation supplied to the model. Fixed-length ECG segments may contain normalization, segmentation, and zero padding introduced during preprocessing. Explanation methods can therefore respond not only to the observed waveform but also to properties of the representation itself. This issue becomes particularly important when fiducial annotations are unavailable, because directly associating attribution peaks with the P wave, QRS complex, or other electrophysiological structures would introduce assumptions not supported by the processed data. In such settings, explanation analysis should remain grounded in observable properties of the representation and should clearly distinguish behavioral faithfulness from physiological or clinical validity.

The present study addresses a narrower gap. Rather than proposing a new architecture or explanation technique, it compares a compact convolutional model and a recurrent-attention alternative under identical data partitions and training controls. Explanations are generated for the same class-balanced set of heartbeats correctly classified by both models and examined through perturbation-based faithfulness, temporal distribution, and cross-method agreement. The analysis also considers attribution over zero-padding regions, an aspect that is particularly relevant when fixed-length preprocessed ECG representations are used.

3. MATERIALS AND METHODS

This study was designed as a controlled quantitative experiment for multiclass ECG heartbeat classification, combining predictive evaluation with post hoc explainability analysis. Experiments were conducted using the preprocessed representation of the MIT-BIH Arrhythmia Database provided by Kachuee, Fazeli and Sarrafzadeh (2018), derived from the original database described by Moody and Mark (2001). Each instance contains 187 temporal ECG samples and one class label, with signals sampled at 125 Hz and normalized to the [0,1] interval. Five heartbeat categories were considered: Normal (N), Supraventricular (S), Ventricular (V), Fusion (F), and Unclassified (Q).

The original training and test partitions were preserved throughout the study. The training set contains 87,554 heartbeats, comprising 72,471 N, 2,223 S, 5,788 V, 641 F, and 6,431 Q samples. The test set contains 21,892 heartbeats, including 18,118 N, 556 S, 1,448 V, 162 F, and 1,608 Q samples. Before training, both files were checked for dimensional consistency, missing or non-finite values, label validity, signal range, exact duplicates, and exact row overlap between the two partitions. The test set was never incorporated into cross-validation, epoch selection, or hyperparameter adjustment and remained untouched until final evaluation. Because the

This article is published in open access under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are credited.

preprocessed files do not retain patient identifiers, patient-level independence between partitions cannot be established and is therefore treated as a limitation of the experimental design.

Two neural architectures were evaluated. The first was a compact 1D-CNN composed of three convolutional blocks with 32, 64, and 128 filters, respectively, using kernel size 5, batch normalization, ReLU activation, and max pooling. The final convolutional representation was processed by global average pooling and a fully connected layer with five outputs. This model contains 52,677 trainable parameters. The second architecture combined two convolutional layers with 64 and 128 filters with a Bidirectional Long Short-Term Memory layer containing 64 hidden units in each direction. The resulting 128-dimensional temporal representations were processed by additive temporal attention, consisting of a 128-to-64 projection, hyperbolic tangent activation, scalar scoring, and softmax normalization. The weighted temporal representation was then forwarded to the final classifier. The CNN-BiLSTM-Attention model contains 150,149 trainable parameters.

Class imbalance was addressed using weighted cross-entropy, with class weights recomputed exclusively from the fitting partition available at each training stage. For class c, the balanced weight was defined as $w_c = \frac{N_{fit}}{Cn_c}$, where N_{fit} denotes the number of samples in the fitting partition, $C = 5$ is the number of classes, and n_c is the number of samples belonging to class c. Consistent with the weighted mean reduction used during training, the optimization objective was defined as

$$L = -\frac{\sum_{i=1}^{N} w_{y_i} \ln\left(p_{i,y_i}\right)}{\sum_{i=1}^{N} w_{y_i}},$$

where w_{y_i} denotes the weight associated with the true class of sample i, and p_{i,y_i} is the probability assigned by the model to that class. Both architectures were optimized using AdamW (Loshchilov; Hutter, 2019), with a learning rate of 10^{-3}, a weight decay of 10^{-4}, and a batch size of 256. The learning rate, weight decay, and batch size were fixed a priori and kept identical for both architectures so that the comparison would reflect architectural differences rather than architecture-specific optimization. These values were not optimized against the final holdout set, and the study does not assume that they represent globally optimal hyperparameters for either architecture. The holdout remained unavailable during model development and was used only after the training protocol and epoch-selection procedure had been fixed.

Model comparison was based on five-fold stratified cross-validation performed exclusively on the original training set. Within each outer fold, 10% of the corresponding fitting partition was reserved through stratified sampling for epoch selection. Both architectures used the same

This article is published in open access under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are credited.

internal split, deterministic minibatch ordering, and seed schedule. Each model was trained for the complete range of 100 epochs without early stopping, and Macro F1 was computed on the internal validation subset after every epoch. The epoch yielding the highest Macro F1 was selected, with the earliest epoch retained in the event of an exact tie. The model was then reinitialized and trained from scratch on the complete outer training partition for the selected number of epochs before evaluation on the corresponding outer validation fold. For final training, the median of the five selected epoch counts was computed independently for each architecture. The resulting values were 93 epochs for the 1D-CNN and 92 epochs for the CNN-BiLSTM-Attention. Each final model was subsequently reinitialized, trained on all 87,554 training samples, and evaluated once on the untouched 21,892-sample test set.

Macro F1 was adopted as the primary selection metric because each class contributes equally regardless of its prevalence. For $C = 5$ classes, it was defined as

$$F1_{Macro} = \frac{1}{C}\sum_{c=1}^{C} F1_c$$

with the class-specific F1 defined as:

$$F1_c = \frac{2P_cR_c}{P_c + R_c}$$

where P_c and R_c denote precision and recall for class c, respectively. Accuracy, Balanced Accuracy, Weighted F1, and Matthews Correlation Coefficient (MCC) were additionally reported, together with class-specific precision, recall, and F1. Uncertainty in the holdout Macro F1 difference was quantified using 5,000 paired class-stratified bootstrap resamples of test heartbeats, preserving both models' predictions for each sampled beat. Because patient identifiers are unavailable, the resulting interval quantifies beat-level holdout uncertainty and does not account for possible within-patient dependence.

Explainability analysis was conducted only after final model evaluation. To enable paired comparisons among explanation methods, eligibility was restricted to holdout samples correctly classified by both architectures. Up to 100 samples were deterministically selected from each class, resulting in a maximum class-balanced subset of 500 heartbeats. The per-class cap was adopted to prevent the predominance of Normal beats from dominating the XAI analysis while retaining enough observations for class-specific paired comparisons and bootstrap estimation. It also kept the computational cost of repeatedly computing Integrated Gradients and perturbation-based evaluations tractable. Because the Fusion class contained the smallest pool of heartbeats correctly classified by both models, the same upper limit was applied uniformly to all classes rather than increasing the sample size only for more prevalent categories. This design ensures

This article is published in open access under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are credited.

that the explanation methods are compared on exactly the same predictions, although it deliberately characterizes cases correctly classified by both models and therefore does not represent the entire holdout distribution. Three explanatory mechanisms were evaluated: 1D Grad-CAM (Selvaraju et al., 2017) applied to the third convolutional layer of the 1D-CNN, Integrated Gradients (Sundararajan; Taly; Yan, 2017) applied independently to both architectures, and the learned temporal attention weights of the hybrid model. Because temporal attention is defined over contextualized BiLSTM states at a lower temporal resolution, the resulting attention profile was linearly projected onto the 187-sample input grid for visualization and representation-level comparisons. The projected profile was treated as an input-aligned attention proxy rather than as a direct attribution to individual ECG samples.

Integrated Gradients used an all-zero vector as the reference baseline and retained signed attributions during computation. The zero reference was chosen because the input representation is normalized to the [0,1] interval and zero is already used by the preprocessing scheme outside the observed signal support. It therefore provides a consistent representation-level reference across all samples. This baseline is not intended to represent a physiologically neutral ECG, and conclusions derived from Integrated Gradients are restricted to the adopted input representation. Numerical integration followed the trapezoidal rule with an adaptive sequence of 64, 128, 256, 512, and 1,024 integration steps. For each explanation, completeness was assessed by comparing the summed attribution with the change in the target-class logit between the observed signal and the baseline. Integration stopped when either the absolute completeness error was at most 10^{-4} or the relative error was at most 2%; explanations that reached 1,024 steps without satisfying these criteria were retained and explicitly recorded rather than discarded. Because individual fiducial annotations are absent from the preprocessed representation, R peaks, QRS boundaries, or other physiological regions were not inferred. A trailing sequence of at least three consecutive zeros was considered padding, whereas isolated zeros were retained as part of the observed signal.

Explanation behavior was characterized using representation-grounded measures rather than assumed physiological landmarks. For each method, the analysis quantified the fraction of absolute attribution located on the observed signal support and on padding, normalized attribution entropy, normalized temporal center of mass, and mean attribution profiles over ten equally spaced relative-time bins. Cross-method agreement was assessed using Spearman rank correlation for three controlled comparisons: Grad-CAM versus the absolute Integrated Gradients attribution within the CNN, temporal attention versus Integrated Gradients within the hybrid model, and Integrated Gradients between the two architectures. These comparisons distinguish

This article is published in open access under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are credited.

disagreement caused by the explanation mechanism from disagreement associated with different model architectures.

Behavioral faithfulness was evaluated through paired perturbation. For each explanation, the 10% of observed positions with the largest attribution magnitude within the observed signal were replaced by zero, and the resulting decrease in the target-class logit was measured. A 10% perturbation fraction was adopted as a fixed perturbation budget for all methods and samples. The number of perturbed positions was computed over the observed signal support and rounded up to the nearest sample. This fraction was used as a pragmatic compromise between obtaining measurable changes in the target logit and preserving most of the original waveform; it was not tuned by model, class, or explanation method. The same fraction was used for attribution-guided and random masks so that faithfulness was assessed under matched perturbation budgets. The procedure was compared with R=20 random masks of identical size. The same random control masks were reused across explanation methods for the same heartbeat to maintain paired comparisons. The faithfulness gain was defined as

$$G_{fid} = \Delta z_{top} - \frac{1}{R}\sum_{r=1}^{R} \Delta z_{rand,r}$$

where Δz_{top} is the decrease in target-class logit after masking the highest-attribution positions and $\Delta z_{rand,r}$ is the decrease produced by the $r-th$ matched random perturbation. Positive values indicate that removing the positions identified as important by an explanation method affects the classifier more strongly than an equivalent random perturbation. This quantity was interpreted strictly as an operational measure of behavioral faithfulness and not as evidence of physiological causality or clinical validity. Descriptive 95% bootstrap intervals were also estimated for faithfulness statistics and paired differences among XAI methods over the selected heartbeats correctly classified by both models.

Reproducibility controls included a fixed global seed of 42 for Python, NumPy, and PyTorch. Deterministic execution was enabled for PyTorch and CUDA operations, nondeterministic acceleration settings were disabled, and data loading used explicit deterministic generators. These controls support reproducibility within equivalent environments, although bit-for-bit equivalence across different hardware and software configurations is not assumed.

This article is published in open access under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are credited.

4. RESULTS AND DISCUSSION

The two architectures exhibited distinct behavior across the development and final evaluation stages. Across the five outer cross-validation folds, the CNN-BiLSTM-Attention achieved higher mean values and lower fold-to-fold dispersion for all global predictive metrics recorded by the experimental pipeline. For the primary metric, Macro F1 was 0.8842 ± 0.0376 for the hybrid model and 0.8456 ± 0.0662 for the 1D-CNN. The hybrid architecture achieved the higher Macro F1 in four of the five folds. The selected epochs ranged from 72 to 96 for the 1D-CNN and from 80 to 98 for the hybrid model, yielding median final training durations of 93 and 92 epochs, respectively. Because no fold selected the predefined upper limit of 100 epochs, there is no indication that epoch selection was constrained by this boundary in the reported run.

The model ranking changed on the held-out test partition, which was excluded from training, cross-validation, and epoch selection in the final protocol. The 1D-CNN achieved a Macro F1 of 0.8681, compared with 0.8140 for the CNN-BiLSTM-Attention, whereas the hybrid architecture achieved higher Balanced Accuracy, 0.9441 versus 0.9264. Table 1 summarizes both the cross-validation variability and the final holdout performance.

Table 1. Predictive performance and model complexity.

Metric	1D-CNN	CNN-BiLSTM-Attention
Trainable parameters	52,677	150,149
CV Accuracy, mean ± SD	0.9537 ± 0.0339	**0.9756 ± 0.0095**
CV Balanced Accuracy, mean ± SD	0.9089 ± 0.0182	**0.9424 ± 0.0079**
CV Macro F1, mean ± SD	0.8456 ± 0.0662	**0.8842 ± 0.0376**
CV Weighted F1, mean ± SD	0.9599 ± 0.0246	**0.9772 ± 0.0076**
CV MCC, mean ± SD	0.8724 ± 0.0727	**0.9241 ± 0.0260**
Holdout Accuracy	**0.9680**	0.9568
Holdout Balanced Accuracy	0.9264	**0.9441**
Holdout Macro F1	**0.8681**	0.8140
Holdout Weighted F1	**0.9701**	0.9623
Holdout MCC	**0.9007**	0.8743
Final training epochs	93	92

The reversal in model ranking between cross-validation and the held-out test set deserves particular attention. During cross-validation, the hybrid architecture achieved higher Macro F1 in four of the five folds and showed lower fold-to-fold variability, suggesting more stable performance across the development partitions. This advantage, however, was not reproduced on the held-out

This article is published in open access under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are credited.

test set, where the 1D-CNN achieved a Macro F1 of 0.8681 compared with 0.8140 for the hybrid model. The paired class-stratified bootstrap yielded a Hybrid-minus-CNN difference of -0.0541, with a 95% interval from -0.0643 to -0.0440, which remained below zero under beat-level resampling.

This change in ranking should not be interpreted, by itself, as evidence that the hybrid architecture overfit the training data. Cross-validation and holdout evaluation reflect different partitions of the available dataset, and the absence of patient identifiers prevents analysis of whether differences in patient composition contributed to the observed shift. What the results do show is that the advantage of the hybrid model within the development data was not preserved on the fixed holdout partition. This reinforces the importance of retaining an untouched test set when comparing architectures, particularly when the ranking observed during model development is not consistent with the final evaluation.

The additional capacity of the hybrid model therefore did not correspond to uniformly better holdout performance. With 150,149 trainable parameters, the CNN-BiLSTM-Attention contains approximately 2.85 times as many parameters as the 52,677-parameter 1D-CNN. Its advantages in this experiment were associated with higher cross-validation means, lower fold-to-fold dispersion, and higher holdout Balanced Accuracy, whereas the compact CNN achieved higher holdout Accuracy, Macro F1, Weighted F1, and MCC. These results are compatible with the broader observation that increasing architectural complexity should be evaluated against measurable predictive gains and computational requirements rather than assumed to improve generalization (Gupta et al., 2024).

Class-specific performance provides additional context for the global results. As shown in Table 2, both architectures achieved similar high performance for Normal, Ventricular, and Unclassified beats, while more pronounced differences appeared for Supraventricular and Fusion beats.

Table 2. Class-specific precision, recall, and F1 on the held-out test set.

Class and metric	1D-CNN	CNN-BiLSTM-Attention
Normal (N)		
Precision	0.9921	**0.9957**
Recall	**0.9723**	0.9560
F1	**0.9821**	0.9755
Supraventricular (S)		
Precision	**0.5729**	0.5594
Recall	0.8201	**0.9065**

This article is published in open access under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are credited.

	F1	0.6746	**0.6918**
Ventricular (V)			
	Precision	**0.9245**	0.9141
	Recall	**0.9558**	0.9551
	F1	**0.9399**	0.9341
Fusion (F)			
	Precision	**0.6591**	0.3333
	Recall	0.8951	**0.9136**
	F1	**0.7592**	0.4884
Unclassified (Q)			
	Precision	**0.9803**	0.9707
	Recall	0.9888	**0.9894**
	F1	**0.9845**	0.9800

The Supraventricular class illustrates one aspect of the higher average class-wise recall observed for the hybrid architecture. Its recall increased from 0.8201 to 0.9065, with a modest increase in F1 from 0.6746 to 0.6918. Fusion presented a different trade-off. Recall increased slightly from 0.8951 to 0.9136, whereas precision decreased from 0.6591 to 0.3333, reducing F1 from 0.7592 to 0.4884. Inspection of the corresponding predictions showed that 256 Normal beats were classified as Fusion by the hybrid model, compared with 45 by the 1D-CNN. This pattern helps explain the coexistence of higher Balanced Accuracy and lower Macro F1 in the hybrid architecture: improved recall for some less frequent classes was accompanied by a higher false-positive burden, particularly for Fusion.

Fusion is therefore the class that most clearly exposes the different decision profiles of the two architectures. The hybrid model correctly recovered a slightly larger proportion of Fusion beats, but this gain in recall came with a substantial increase in false-positive assignments from the dominant Normal class. The resulting precision reduction, from 0.6591 to 0.3333, was large enough to offset the recall gain and explains much of the decrease in Macro F1. Because the training objective used class weighting, greater sensitivity to less frequent classes is expected to be encouraged; however, the present experiment does not isolate whether the specific Normal-to-Fusion shift was caused by class weighting, model capacity, or their interaction. An ablation study would be required to separate these effects.

This behavior also clarifies why Balanced Accuracy and Macro F1 lead to different impressions of the hybrid model. Balanced Accuracy rewards its higher mean class-wise recall, whereas Macro F1 penalizes the reduced precision associated with the additional Fusion

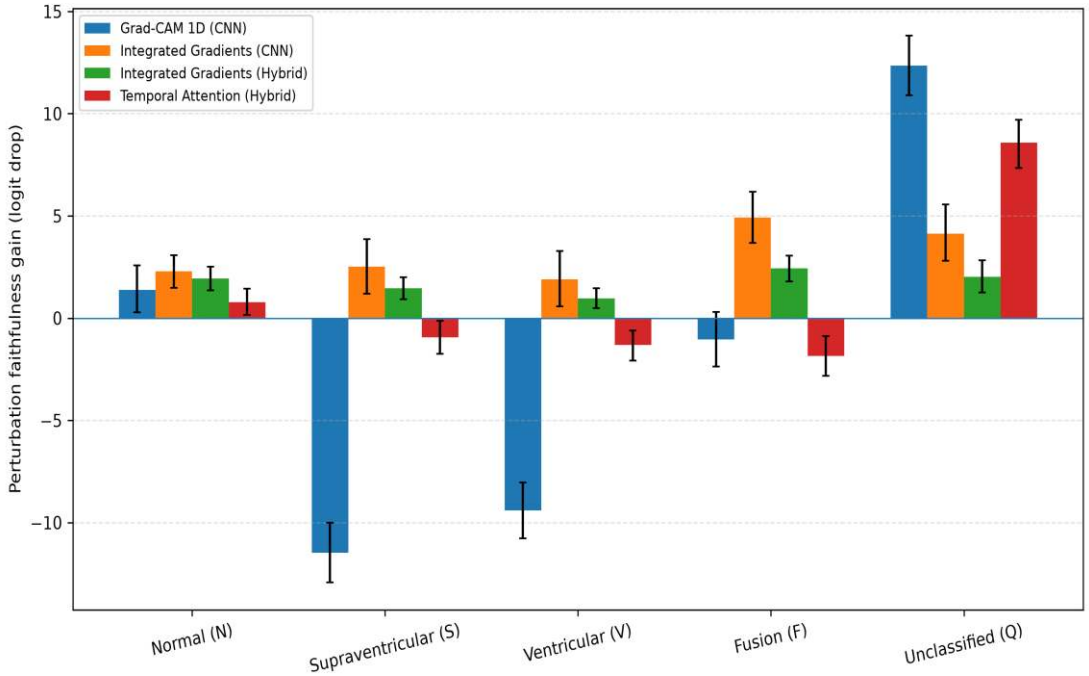
This article is published in open access under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are credited.

predictions. For this dataset, reporting both metrics is therefore necessary: the hybrid model is more sensitive across classes, but the compact CNN provides a more favorable precision-recall balance on the held-out partition.

The explainability analysis comprised 500 held-out beats correctly classified by both architectures, with 100 samples from each class. All classes reached the predefined sampling target; Fusion had the smallest eligible pool, with 144 heartbeats correctly classified by both models. Using identical samples across explanation methods supported paired comparisons, but it also introduces a deliberate selection toward cases successfully recognized by both models. Accordingly, the following XAI results characterize common-correct predictions and should not be generalized to the complete holdout distribution.

Perturbation-based faithfulness was used as the main quantitative assessment of the explanation methods. The target-class logit reduction produced by masking positions with the largest attribution magnitude was compared with the reduction produced by matched random masks. The class-specific mean gains and their 95% bootstrap intervals are presented in Figure 1.

Figure 1. Perturbation-based faithfulness gain across heartbeat classes. Error bars represent 95% descriptive bootstrap intervals for the selected test heartbeats correctly classified by both models.



This article is published in open access under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are credited.

Integrated Gradients produced positive mean faithfulness gains for every class in both architectures. Mean gains ranged from 1.89 to 4.91 for the 1D-CNN and from 0.96 to 2.43 for the hybrid model, and the corresponding 95% bootstrap intervals remained above zero for all class-model combinations in the analyzed subset. This indicates that, under the adopted perturbation procedure, positions emphasized by Integrated Gradients tended to affect the target-class logit more than matched randomly selected positions.

Grad-CAM and temporal attention exhibited a more class-dependent pattern. Grad-CAM produced positive mean gains for Normal and Unclassified beats and negative mean gains for Supraventricular, Ventricular, and Fusion beats. Temporal attention showed the same direction pattern, with positive values for Normal and Unclassified and negative values for the remaining three classes. These negative gains should not be interpreted as computational failures. Under the operational definition adopted here, they indicate that masking positions prioritized by the explanation method produced a smaller average target-logit reduction than matched random masking.

Paired comparisons further suggest that the relative behavior of the XAI methods depends on class and explanation mechanism. Within the CNN, Integrated Gradients yielded higher mean faithfulness than Grad-CAM for Supraventricular, Ventricular, and Fusion beats, whereas Grad-CAM yielded a higher value for Unclassified beats; for Normal beats, the paired bootstrap interval for the difference included zero. Within the hybrid model, Integrated Gradients yielded higher mean values than temporal attention for Normal, Supraventricular, Ventricular, and Fusion beats, whereas temporal attention yielded a higher value for Unclassified beats. These results do not establish universal superiority of a single XAI technique, but they support quantitative comparison alongside visual inspection, as also emphasized in recent ECG explainability reviews (Manimaran et al., 2025; Taleban et al., 2026).

The numerical approximation of Integrated Gradients was monitored independently through its completeness criterion. With adaptive integration up to 1,024 steps, 491 of 500 CNN explanations, corresponding to 98.2%, satisfied the predefined absolute or relative tolerance. For the CNN-BiLSTM-Attention, 473 of 500 explanations, or 94.6%, met the same criterion. Cases that did not satisfy the tolerance were retained and recorded rather than excluded. These results indicate adequate numerical convergence for most analyzed explanations, while completeness itself remains a mathematical consistency property and should not be interpreted as evidence of clinical validity.

Agreement among explanation mechanisms was generally low to moderate. Across classes, mean Spearman correlations ranged from approximately -0.04 to 0.17 between CNN

This article is published in open access under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are credited.

Grad-CAM and absolute Integrated Gradients, from 0.15 to 0.28 between hybrid temporal attention and Integrated Gradients, and from 0.16 to 0.43 when Integrated Gradients was compared across the two architectures. These results suggest that different explanation mechanisms, and even different architectures analyzed using the same attribution method, may emphasize distinct temporal regions despite producing the same correct class prediction.

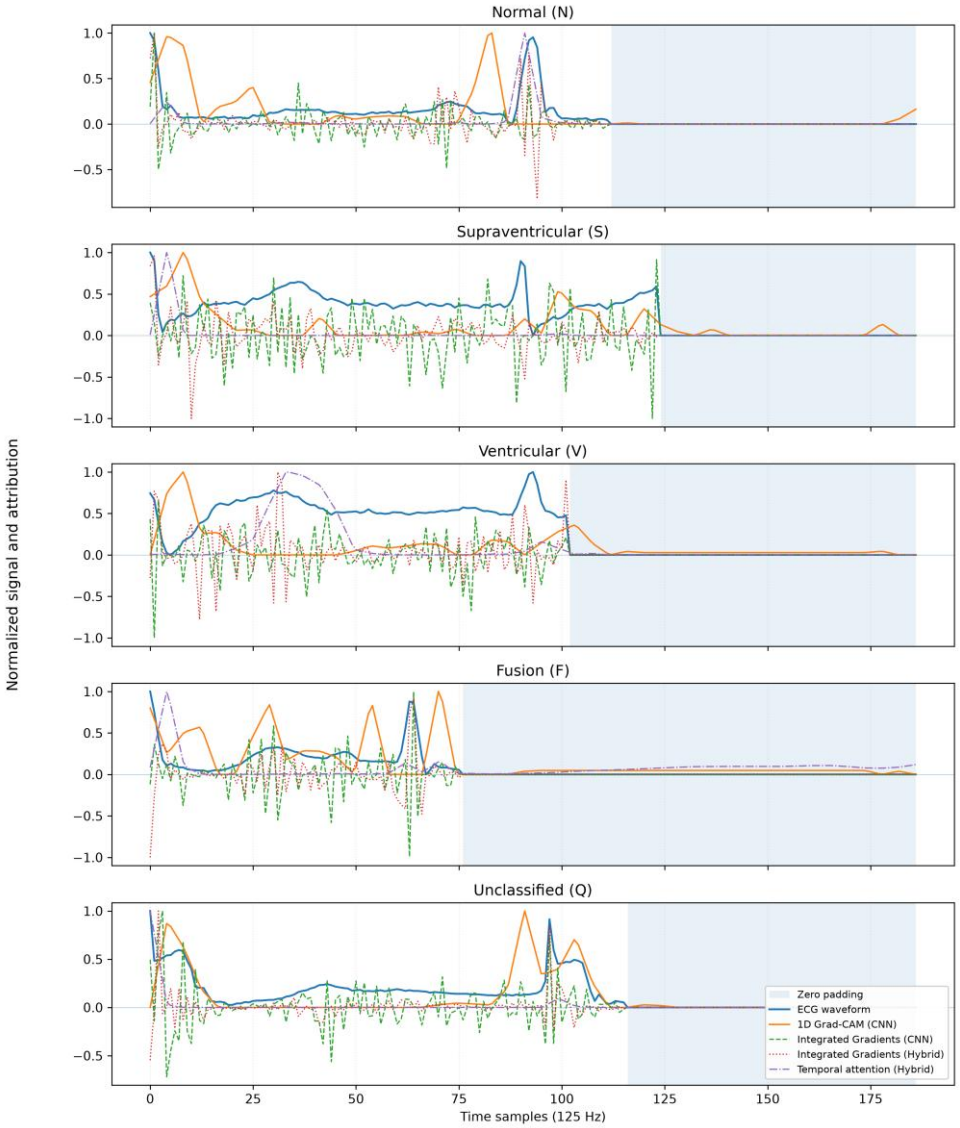
Padding also affected the interpretation of the explanation methods in different ways. Integrated Gradients assigned essentially no relevance to padded coordinates because these positions coincide with the all-zero reference used by the attribution method. This behavior follows from the formulation of Integrated Gradients and should not be interpreted as evidence that it intrinsically handles padding better than the other mechanisms. Temporal attention operates on latent BiLSTM representations and is therefore not subject to the same constraint.

For Fusion beats, approximately 55.0% of the mass of the projected attention profile overlapped positions identified as padding, compared with approximately 45.0% over the observed signal support. This does not imply that the classifier relied directly on zero-valued samples: the attention weights are computed at a lower temporal resolution over contextualized hidden states and are subsequently projected onto the 187-sample input grid. The result instead suggests that the learned temporal weighting can remain sensitive to features of the fixed-length representation, including the transition between observed signal and padded regions. Consequently, comparisons of padding attribution across Integrated Gradients and temporal attention should be interpreted as representation-level observations rather than as equivalent measures of physiological relevance. Mask-aware recurrent processing or variable-length representations would be useful extensions for determining whether this behavior persists when padding is explicitly excluded from temporal aggregation.

To complement the quantitative analyses, Figure 2 provides an illustrative comparison of temporal attribution patterns for common-correct heartbeats.

This article is published in open access under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are credited.

Figure 2. Illustrative temporal attributions produced by 1D Grad-CAM, Integrated Gradients, and temporal attention for common-correct ECG heartbeats. Shaded regions indicate zero padding and do not represent physiological regions of the ECG.



Overall, neither architecture showed a uniform advantage. The hybrid model was more stable across the cross-validation folds and achieved higher holdout Balanced Accuracy, whereas the substantially smaller 1D-CNN achieved higher holdout Macro F1, Accuracy, Weighted F1, and MCC. The XAI results likewise depended on both class and explanation mechanism. Integrated Gradients showed the most consistent faithfulness gains under the adopted

This article is published in open access under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are credited.

perturbation protocol, while Grad-CAM and temporal attention were more class-dependent. Low-to-moderate agreement among methods, together with the sensitivity of projected attention to padded regions, supports interpreting these outputs as complementary descriptions of model behavior rather than direct physiological explanations.

5. CONCLUSIONS

This study compared a compact 1D-CNN and a CNN-BiLSTM-Attention architecture for multiclass ECG heartbeat classification under a controlled experimental protocol that jointly considered predictive performance and explainability. The results indicate that the additional complexity of the hybrid model did not lead to uniformly better performance on the held-out test partition. Although the hybrid model achieved higher mean Macro F1 and lower variability during cross-validation, the 1D-CNN obtained higher Macro F1, accuracy, weighted F1, and MCC on the untouched holdout set. Conversely, the hybrid architecture achieved higher Balanced Accuracy, indicating higher mean class-wise recall but also a less favorable precision-recall trade-off for specific minority categories, particularly Fusion.

The evaluated XAI mechanisms were not interchangeable. Under the adopted zero baseline, perturbation budget, and sample-selection strategy, Integrated Gradients showed the most consistent positive faithfulness gains across classes and architectures, whereas Grad-CAM and temporal attention exhibited more class-dependent behavior. Agreement among methods remained low to moderate, indicating that different techniques may emphasize distinct temporal regions for the same prediction.

These findings should be interpreted within the limitations of the experimental design. The preprocessed dataset does not retain patient identifiers, preventing confirmation of patient-level independence between training and test partitions. Fiducial annotations are also unavailable, so attribution patterns cannot be directly associated with specific ECG structures such as the R peak or QRS complex. In addition, the XAI analysis was restricted to samples correctly classified by both architectures, which improves comparability among explanation methods but limits the conclusions to common-correct predictions. The reported explanations therefore characterize model behavior and should not be interpreted as evidence of physiological causality or clinical validity.

Overall, the study suggests that ECG classification should be assessed through multiple complementary dimensions rather than through predictive accuracy or model complexity alone. In the present dataset and experimental protocol, the compact 1D-CNN achieved higher holdout

This article is published in open access under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are credited.

Macro F1, Accuracy, Weighted F1, and MCC while using fewer parameters, whereas the CNN-BiLSTM-Attention achieved higher Balanced Accuracy and higher recall for Supraventricular and Fusion beats. Future work should determine whether these trade-offs persist across patient-level partitions, external datasets, and alternative model configurations. At the same time, explanation quality benefits from quantitative evaluation and comparison across multiple XAI mechanisms. Further research should also examine explanation behavior on misclassified cases and clinically annotated signals to determine whether the computational patterns identified here remain consistent under broader and more clinically representative conditions.

REFERENCES

AYYUB, Ahtisham; POLITIS, Christos; USMAN, Muhammad Arslan. A comprehensive review of AI-based detection of arrhythmia using electrocardiogram (ECG). **Computers in Biology and Medicine**, v. 196, pt. A, art. 110594, 2025.

BAHDANAU, Dzmitry; CHO, Kyunghyun; BENGIO, Yoshua. Neural machine translation by jointly learning to align and translate. In: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS (ICLR), 3., 2015, San Diego.

GUPTA, Utkarsh et al. A comprehensive review on efficient artificial intelligence models for classification of abnormal cardiac rhythms using electrocardiograms. **Heliyon**, v. 10, n. 5, 2024.

HANNUN, Awni Y. et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. **Nature Medicine**, v. 25, n. 1, p. 65-69, 2019.

HOCHREITER, Sepp; SCHMIDHUBER, Jürgen. Long short-term memory. **Neural Computation**, v. 9, n. 8, p. 1735-1780, 1997.

JAIN, Sarthak; WALLACE, Byron C. Attention is not explanation. In: CONFERENCE OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: HUMAN LANGUAGE TECHNOLOGIES, 2019, Minneapolis.

KACHUEE, Mohammad; FAZELI, Shayan; SARRAFZADEH, Majid. ECG heartbeat classification: a deep transferable representation. In: IEEE INTERNATIONAL CONFERENCE ON HEALTHCARE INFORMATICS (ICHI), 2018, New York.

KIRANYAZ, Serkan; INCE, Turker; GABBOUJ, Moncef. Real-time patient-specific ECG classification by 1-D convolutional neural networks. **IEEE Transactions on Biomedical Engineering**, v. 63, n. 3, p. 664-675, 2016.

LEINONEN, Tuija et al. Empirical investigation of multi-source cross-validation in clinical ECG classification. **Computers in Biology and Medicine**, v. 183, art. 109271, 2024.

This article is published in open access under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are credited.

LOSHCHILOV, Ilya; HUTTER, Frank. Decoupled weight decay regularization. In: INTERNATIONAL CONFERENCE ON LEARNING REPRESENTATIONS (ICLR), 7., 2019, New Orleans.

LUZ, Eduardo; MENOTTI, David. How the choice of samples for building arrhythmia classifiers impact their performances. In: ANNUAL INTERNATIONAL CONFERENCE OF THE IEEE ENGINEERING IN MEDICINE AND BIOLOGY SOCIETY, 33., 2011, Boston. **Proceedings [...]**. Boston: IEEE, 2011. p. 4988-4991.

MANIMARAN, Gouthamaan et al. Explainable deep learning based techniques for ECG-Based heart disease classification: A systematic literature review and future direction. **Computers in Biology and Medicine**, v. 199, art. 111324, 2025.

MOODY, George B.; MARK, Roger G. The impact of the MIT-BIH arrhythmia database. **IEEE Engineering in Medicine and Biology Magazine**, v. 20, n. 3, p. 45-50, 2001.

RANI, Jyoti; GUPTA, Shilpa; MITTAL, Vikas. Robust and Calibrated ECG Heartbeat Classification via Hybrid Convolutional, Temporal and Attention-Based Learning. **Applied Sciences**, v. 16, n. 9, art. 4393, 2026.

SCHUSTER, Mike; PALIWAL, Kuldip K. Bidirectional recurrent neural networks. **IEEE Transactions on Signal Processing**, v. 45, n. 11, p. 2673-2681, 1997.

SELVARAJU, Ramprasaath R. et al. Grad-CAM: Visual explanations from deep networks via gradient-based localization. In: IEEE INTERNATIONAL CONFERENCE ON COMPUTER VISION (ICCV), 2017, Venice.

SUNDARARAJAN, Mukund; TALY, Ankur; YAN, Qiqi. Axiomatic attribution for deep networks. In: INTERNATIONAL CONFERENCE ON MACHINE LEARNING (ICML), 34., 2017, Sydney. **Proceedings of Machine Learning Research**, v. 70, 2017. p. 3319-3328.

TALEBAN, Amirsajjad et al. Explainable artificial intelligence in electrocardiography: A systematic review. **Biomedical Signal Processing and Control**, v. 114, art. 109325, 2026.

WORLD HEALTH ORGANIZATION. **Ethics and governance of artificial intelligence for health**: WHO guidance. Geneva: World Health Organization, 2021.

This article is published in open access under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits unrestricted use, distribution, and reproduction in any medium, provided that the original author and source are credited.