Interpretability of Machine Learning Models for Cardiac Arrhythmia Classification

Bruno Torres Marques
Federal University of Ceara
Quixadá, Ceará, Brazil
brunotores@alu.ufc.br

Regis Pires Magalhães
Federal University of Ceara
Quixadá, Ceará, Brazil
regismagalhaes@ufc.br

Lívia Almada
Federal University of Ceara
Quixadá, Ceará, Brazil
livia.almada@ufc.br

João Paulo do V. Madeiro
Federal University of Ceara
Fortaleza, Ceará, Brazil
jpaulo.vale@dc.ufc.br

César Lincoln C. Mattos
Federal University of Ceará
Fortaleza, Ceará, Brazil
cesarlincoln@dc.ufc.br

José Macedo
Federal University of Ceara
Fortaleza, Ceará, Brazil
jose.macedo@ufc.br

Abstract

Context: Cardiovascular diseases, particularly cardiac arrhythmias, cause global mortality. Electrocardiograms (ECG) are essential for diagnoses. However, the automated analysis of ECG signals using machine learning faces interpretability challenges, which are crucial for accepting these models in medical practice. **Problem:** The classification of cardiac arrhythmias using machine learning faces resistance in the medical field due to the black-box nature of these models, which hinders the understanding of decisions and reduces professional trust. There is a need for interpretable models that reveal the critical factors in diagnosis, fostering the reliable use of artificial intelligence in healthcare. **Solution:** This work proposes an interpretable model for classifying cardiac arrhythmias, using machine learning to identify and visually explain the ECG features contributing to the diagnosis, with explanations at the model, class, and specific signal levels. **IS Theory:** This work is based on the Technology Acceptance Model (TAM), which suggests that "perceived usefulness" and "perceived ease of use" influence the intention to use a system. In healthcare, these perceptions relate to the model's ability to provide clear and helpful explanations, facilitating medical professionals' adoption of artificial intelligence (AI). **Method:** Instead of using raw biosignals, extracted ECG features are employed to enhance interpretability. This approach provides model-agnostic explanations at both local and global levels. Interpretability techniques are applied to clarify the contribution of each feature to the diagnosis. **Summary of Results:** Features such as the variability and median of RR and PR intervals and the signal-to-noise ratio of ECG signals are crucial for accurate arrhythmia classification. **Contributions and Impact on IS:** The approach allows for understanding the influence of specific ECG features on diagnosis, helping to identify patterns that support classification decisions and promoting the adoption of AI with trust and responsibility in medical practice.

CCS Concepts

• **Do Not Use This Code → Generate the Correct Terms for Your Paper**; *Generate the Correct Terms for Your Paper*; Generate the Correct Terms for Your Paper; Generate the Correct Terms for Your Paper.

Keywords

Heartbeat classification, Electrocardiogram, Explainable artificial intelligence (XAI).

1 Introdução

Cardiovascular diseases represent a major global health problem, leading the causes of death in several countries [10]. According to [11], cardiovascular diseases have become the leading cause of death worldwide, accounting for an estimated 17.9 million lives in 2019. These include cardiac arrhythmias, commonly detected by health professionals using an electrocardiogram (ECG). The ECG assesses the heart's electrical activity using electrodes, usually placed in the patient's chest. Any morphological variation in the ECG signals can indicate specific diagnoses about the patient's heart condition and the possible presence of diseases, which makes the processing of the ECG signals crucial to detect cardiac arrhythmias.

Clinicians and researchers widely use the process of analyzing the ECG signal to diagnose many heart muscle disorders. It provides critical insights into patient conditions. The propagation of the signal during the heart muscle's repolarization and depolarization process produces these recordings. The conduction of the potential in the cardiac tissues to the body's surface is measured using electrodes [18].

The ECG recording can be manually interpreted by a physician, who may identify certain anomalies indicative of potential pathological conditions. With the advancement of computer science in the biomedical field, researchers began analyzing ECG data through computational algorithms and software programs capable of uncovering trends and cardiac information that might otherwise go unnoticed by physicians [6].

The external structure of the ECG signal reveals crucial information about its underlying data. The amplitude, which represents the strength of the signal, and the duration of each part of the signal are fundamental to manually analyzing ECG. Nonetheless, manual analysis is an arduous and time-consuming task due to the abundance of data, which increases the risk of losing important information. Therefore, the use of information systems with advanced computational methods, such as classification techniques, can simplify this process and improve the accuracy of clinical diagnoses [18].

Research into Artificial Intelligence (AI) in health information systems is growing worldwide. Countries such as the United States

of America (USA), the United Kingdom, and Israel are at the forefront of promoting AI research. The challenges of applying such methods involve achieving high reliability and accuracy. These are necessary to provide services and maintain data confidentiality [22].

The use of machine learning (ML) models in healthcare is gaining attention. However, the actual adoption of these systems is still in its early stages [16]. A survey conducted on 85 hospitals in 2017 revealed that only 5% planned to implement AI solutions for healthcare purposes, and there was significant uncertainty about when to start using AI. The main barriers to AI use have included the lack of support from executives and doctors and trust issues [3, 21]. [16] point out the general lack of interpretability of ML models results in resistance to adoption and hinders their acceptance in medical diagnosis, where accountability (and responsibility) are crucial.

Making ML classifiers more interpretable is a challenging task [16]. For example, to "explain" the predictions produced by decision trees, one could consider following the path of the tree. However, a decision tree with many nodes could make this process complicated and the related explanations almost incomprehensible to humans [12]. In addition, integrating ML models into clinical support systems presents a challenging task. These models are usually based on uncertain, unbalanced, heterogeneous, and noisy datasets, which often contain a high number of features but are not large enough to model complex patient systems accurately [4, 5, 16].

Given this context, this paper proposes an approach for building an interpretable model to help detect and classify heart diseases in ECG data. Our work presents an explainable approach to interpreting how a given ML model performs the detection and classification of heart disease from ECG data. Importantly, our proposal has the following differences: (i) it is agnostic regarding the learning algorithm; (ii) it involves both global and local explanations; (iii) it considers specific features extracted from ECG data, which are themselves better interpretable than the raw signals.

Therefore, the research questions guiding this study are:

- Q1. Which explainer provide the best interpretability in the classification of arrhythmias for a given machine learning model?
- Q2. What are the most important characteristics for classifying arrhythmias at a global level and how do the values of these characteristics influence the classification?
- Q3. Which characteristics of ECG signals are most relevant in classifying different heart rhythms?

The remainder of this article is organized as follows: Section 2 defines the problem being addressed. Section 3 reviews related work in the field. Section 4 describes the data and methods used, including the data set, evaluation of models, evaluation of explainers, and interpretation of the ML model. Section 5 presents the results and discussion, covering model selection, explainers evaluation, global explanations, class-level explanations, and local explanations. Finally, Section 6 concludes the study and outlines future work.

2 Problem Definition

Given an ECG signal represented by X, the initial process involves extracting a set of features, denoted by $F(X) = \{c_1, c_2, \ldots, c_n\}$, which are numerical values that capture different aspects of the ECG signal. Each c_i represents a specific characteristic extracted from the signal, such as amplitude, duration, and heart rate, among others, which are relevant for identifying patterns related to cardiac arrhythmias. These characteristics are essential for understanding the complexity of the ECG signal and help distinguish normal patterns from abnormal patterns that indicate the presence of arrhythmias.

The problem of interpretability arises when the classifier model uses the set of features $\{c_1, c_2, \ldots, c_n\}$ as the basis for its predictions about the presence or absence of arrhythmias. Interpretability refers to understanding and explaining how each feature c_i contributed to the model's decision.

Understanding the accuracy of a model's predictions and feature contributions in medical applications is crucial. Additionally, we need to know how the model combined these features to generate the final output.

3 Related Work

The interpretability of models for classifying cardiac arrhythmias in ECG is challenging. Although several researchers are working on solving this problem, explainability solutions that detail the behavior of a model have not yet been provided. [16] presents an approach to making machine learning models for ECG classification more interpretable. They propose model-agnostic explanation methods—capable of explaining predictions from any machine learning model—and local explanation methods, which detail predictions for individual instances. For evaluation, the study used the MIT-BIH dataset [15].

The paper [19] presents a framework for the explainability of deep learning-based ECG classification models, addressing the gap between the complexity of the models and the need for accessible interpretations for clinicians. According to the authors, the proposal integrates a symbolic system with a deep learning model, enabling the generation of explanations in clinical language that reflect medical concepts, such as the presence or absence of P waves and irregular rhythms. The framework is designed to be scalable and adaptable to different classes of arrhythmias, promoting the implementation of artificial intelligence in clinical practice.

The work proposed by [1] presents two new interpretability frameworks for deep neural network models trained for ECG classification. These frameworks are based on *post-hoc* methods, which explain the decisions of machine learning models. They use the ECG signal's P, QRS, and T waves to detect cardiac arrhythmias and atrial fibrillation. The method uses a cloaking technique, in which the samples for each wave are set to zero, and the percentage change in the model output is calculated after this cloaking. After the percentage variation is calculated, the relative importance of the waves for the classification of the ECGs is determined. The relevance measures of this study showed which waves are most important in classifying cardiac arrhythmias and atrial fibrillation, validating the effectiveness of the proposed new interpretability frameworks.

The study [17] investigates the application of explainable artificial intelligence (XAI) methods in classifying electrocardiogram

(ECG) data, with a focus on the ST-CNN-5 model. The results indicate that the ST-CNN-5 model showed a slight improvement in accuracy compared to established models but exhibited lower values of specificity and area under the curve. SHAP (SHapley Additive ex-Planations) [14], GradCAM (Gradient-weighted Class Activation Mapping) [20], and LIME (Local Interpretable Model-agnostic Explanations) methods were employed for interpretability analysis. The study underlines the effectiveness of SHAP in highlighting crucial ECG features, while GradCAM showed improvements in explanation compared to the former model. The research emphasizes the importance of thoroughly evaluating the capabilities and limitations of the extended model compared to other models, aiming to provide more reliable and interpretable predictions for ECG data classification.

The approaches proposed in the previous works differ from those proposed in this paper, mainly in terms of the characteristics used for classification, the explainability technique, and the scope (local or global) and generality (agnostic or specific) of the explanations.

Table 1 summarizes this information, comparing the present work with previously presented studies. Driven by this motivation, this work proposes using features extracted from ECG signals instead of electrical biosignals. Therefore, this approach makes it possible to provide explanations in a more comprehensive scope compared to related work, as model-agnostic explanations are provided in local and global scopes.

Another key factor distinguishing this work from the others is the application of SHAP algorithms for model explainability. While some studies rely on techniques such as Grad-CAM, input exclusion masks, relevance measures, and wavelets, this paper adopts SHAP as the main approach to provide explanations at both local and global scopes. SHAP is a game theory-based technique that assigns fair contributions of each feature to the model's prediction, allowing for a more comprehensive and reliable understanding of the model's decisions in classifying cardiac arrhythmias in ECGs [14].

4 Data and methods

This section describes the dataset and the methodology utilized in this study. The methodology consists of the following steps: 1) the feature engineering process to extract interpretable features for model training; 2) the model training and selection; 3) the evaluation and selection of explainers; and, finally, 4) the interpretation of the arrhythmias classifier.

4.1 Dataset

This study uses the PhysioNet dataset [9], which contains 12,186 single-lead ECG recordings ranging in duration from 9 seconds to just over 60 seconds [7].

The PhysioNet dataset [9] comprises four classes representing different heart rate variations. The first class is Normal, which reflects the ideal heart rhythm with regular electrical activity. The second class is Atrial Fibrillation, characterized by irregular and disorganized beats. The Other Rhythm class covers a variety of abnormal heart rhythms, such as atrioventricular blocks, extrasystoles, and ventricular rhythms. Finally, the Very Noisy class includes ECG recordings with excessive interference, making it difficult to classify the heart rhythm accurately.

4.2 Feature engineering

A total of 222 characteristics were extracted from the ECG signals, which were categorized to reflect different aspects of the signal. Table 2 summarizes these characteristics, which include measurements of duration, amplitude, and signal variation, as well as statistical parameters calculated from the Wavelet transform.

These characteristics play a fundamental role in the analysis and interpretation of ECG signals, allowing for a better classification and understanding of the different cardiac rhythms.

4.3 Evaluation of models

This step comprises the generation of the arrhythmia classifier using as descriptor variables the set of interpretable features.

As model generation is not the main focus of this work, and due to the great potential of AutoML tools, we use an automated approach to simplify the model development process. AutoML automates both the selection of the most suitable algorithms and the optimized adjustment of their hyperparameters, facilitating the modeling process. The AutoML tool selected was AutoGluon, an open-source library developed by AWS (Amazon Web Services) in partnership with Apache MXNet.

According to [8], AutoGluon stands out for being faster, more robust, and more accurate in classification and regression tasks when compared to other AutoML tools such as TPOT, H2O, AutoWEKA, auto-sklearn, and Google AutoML Tables. Using this tool allowed us to efficiently explore different model architectures, optimizing performance and maximizing the accuracy of the results obtained.

For the model generation, the data set resulting from the extraction of attributes was divided into 80% for training and 20% for testing. This division was stratified to ensure that the distribution of each class was preserved in both subsets, maintaining the balance of the classes throughout the evaluation process.

The models' performance is measured using the accuracy and the F1-score. The accuracy offers a global measure of the predictive ability of models, which is given by the rate of correct predictions. The F1-score balances the accuracy and precision metrics, which is particularly useful for unbalanced datasets. Based on these metrics, the best classifier is then selected for interpretation.

4.4 Evaluation of explainers

This phase selects the model explainer among six SHAP (Shapley Additive exPlanations)[14] algorithms used for explanation. SHAP is a game theory-based technique that assigns contributions of each feature, named SHAP values, to the model's prediction, which allows an understanding of the model's decisions. The SHAP algorithms were used to evaluate the explainers: (i) Permutation, which calculates the importance of features by randomly permuting the observations of a single feature and measuring the impact on the model's predictions; (ii) Permutation Partition, similar to Permutation, but calculates the impact of permutations only on the positive parts of the model's predictions; (iii) Partition, which divides the feature space into partitions and calculates the contribution of each partition to the model's prediction; (iv) Random, a method that uses random values to explain the model's predictions; (v) Tree, which efficiently calculates Shapley values for tree-based models; and (vi)

Table 1: Comparison between related works and the proposed work.

Works	Features used	Explainability Technique	Scope of Explanations	Generality of Explanations
[16]	ECG	PSI, LIME, SHAP, Random	Local	Model Agnostic
[19]	ECG	Domain concepts	Local	Model Specific
[1]	ECG	Relevance Measures, Wavelet	Local	Model Specific
[17]	ECG	SHAP, GradCAM, LIME	Local	Model Agnostic
This Work	Features extracted from ECGs	SHAP	Local and Global	Model Agnostic

Table 2: Characteristics extracted from ECG signals.

Category	Features
Signal Duration	1. Duration of the processed signal (s).
Heartbeats	2. Number of detected heartbeats. 3. Heartbeats detected per time interval (beats/s).
Signal-to-Noise Ratio (SNR)	4-8. Percentiles of the Signal-to-Noise (SNR) Ratio (5th, 25th, 50th, 75th, 95th)
Amplitude and Intervals	9-14. Mean, standard deviation, and median of the *QRS* amplitude and *RR* interval 15-20. Mean, standard deviation, and median of the *QRS* duration and *P* waves 21-26. Mean, standard deviation, and median of the *PR* interval duration and *T* wave amplitude
Modeling Errors	39-71. Relative modeling error for the *QRS* complexes, *T* wave, and *P* wave
Spectral Power	72-78. Spectral power of the signal in the specified frequency bands
Wavelet Transform	79-222. Statistical parameters of the Wavelet transform (mean, maximum, minimum, variance, skewness, percentiles)

Tree Which Approximate, which uses approximations to speed up the calculation of Shapley values for tree models.

We evaluate the explainers using some metrics [2, 13], including (i) Explanation Error, which measures the discrepancy between the model's predictions and the predictions explained by the explainability algorithms; (ii) Compute Time, which evaluates the time needed to calculate the explanations for the model's predictions; and (iii) metrics such as Keep (positive), Remove (positive), Keep (negative), Remove (negative), Keep (absolute) and Remove (absolute), which assess the importance of keeping or removing positive, negative or sign-independent features for the model's predictions.

4.5 Interpretation of machine learning model

In this step, we use the chosen SHAP algorithm to interpret the machine learning model to understand how the features have impacted the model decision.

The machine learning model interpretation is divided into global, class, and local levels. The most significant attributes that impact the model's predictions are identified globally, revealing the characteristics that influence the overall predictions. At the class level, the influence of specific features on the prediction of each class of Arrythymia is examined, highlighting the different importance of attributes according to the class analyzed. Finally, at the local level, the analysis focuses on individual instances, exploring the reasons and mechanisms that lead the model to make specific Arrythymia for an ECG instance.

We analyze the top 10 features with better SHAP values for each interpretation level to provide a comprehensive understanding of how the features extracted from ECG signals influence the model's predictions in different contexts.

5 Results and Discussion

This section presents the results of our experiments, including model selection for arrhythmia classification, evaluation of explainers based on the chosen model, and interpretation of the classifier considering the most significant features using the selected explainer.

5.1 Results of Model Selection

In our experiments, the AutoML tool AutoGluon was used to train and optimize various classification models based on the extracted features. AutoGluon automates the processes of model selection and hyperparameter tuning, enabling efficient exploration of different models. In this work, we used the AutoGluon "best_quality" preset, which prioritizes achieving the best possible performance through comprehensive optimization strategies. The tool trained a total of 61 models, including LightGBM, CatBoost, Random Forest, Extra Trees, and XGBoost.

In addition to automated model selection, AutoGluon employs Bayesian Optimization to efficiently explore the hyperparameter space, balancing the trade-off between exploration and exploitation. This approach helps in identifying the most promising hyperparameter configurations, enhancing model performance without exhaustive search [23].

The models were evaluated based on the weighted F1-score, which effectively handles class imbalance. Table 3 summarizes the best models of each type trained by AutoGluon.

Table 3: Best models of each type trained by AutoGluon.

Model	F1-score	Accuracy
XGBoost	0.8180	0.8243
LightGBM	0.8131	0.8214
CatBoost	0.8049	0.8126
Random Forest	0.7795	0.7877
Extra Trees	0.7728	0.7877

The best-performing model was XGBoost, with an F1-score of 0.8180 and an accuracy of 0.8243, as shown in Table 3. Other models, such as LightGBM and Extra Trees, exhibited F1-scores of 0.8131 and 0.7728, respectively, as demonstrated in the table, which indicates the performance diversity among the models trained.

5.2 Results of the Explainability Evaluation

This section addresses research question RQ1, which explores which explainer provides the best interpretability when classifying arrhythmias using the selected model. The main metric for ranking explainers was the Explanation Error. Other metrics were also considered, as no single metric fully captures the performance of an attribution explanation method. The results are shown in Table 4.

The Permutation Partition explainer achieved the best result for the Explanation Error, with a value of 0.2783. Although the Permutation explainer outperformed the Permutation Partition in several metrics, including "Keep Positive" (0.8335), "Remove Positive" (-0.5660), "Keep Negative" (-0.6084), "Remove Negative" (1.0146), and "Remove Absolute" (0.9586), the results were very close. Since the Explanation Error was the decisive criterion, the Permutation Partition was selected as the preferred explainer.

One important exception was the "Compute Time" metric, where the Tree Approximation explainer was the fastest, with a time of 0.0064. This indicates that while the Permutation and Permutation Partition explainers provide the highest quality explanations, they incur a higher computational cost compared to the Tree Approximation method.

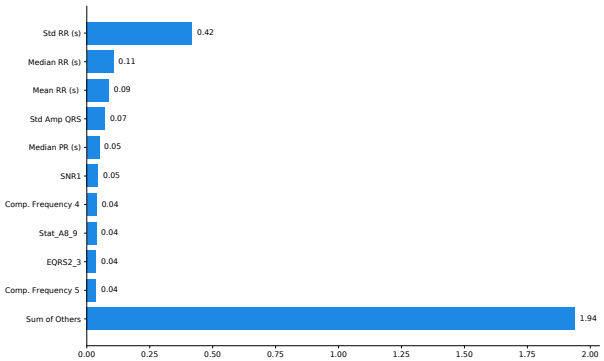
These findings highlight the effectiveness of both the Permutation and Permutation Partition explainers, depending on the specific metric being considered. In subsequent experiments, we used the Permutation algorithm to obtain SHAP values, as it provided the best overall explainability results.

5.3 Explanations at a Global Level

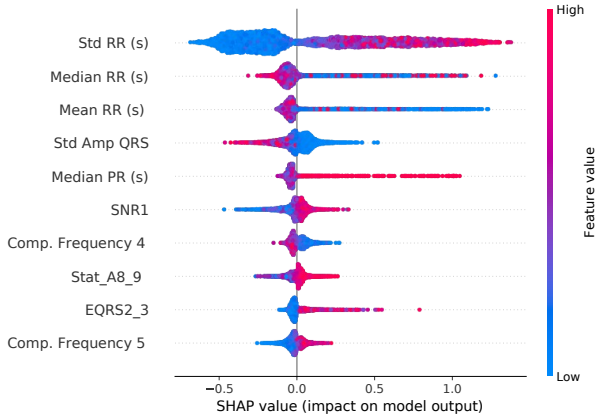
Research question RQ2 is addressed in this section, which investigates the most important characteristics for arrhythmia classification at a global level and their influence on classification.

Two visual approaches were employed to understand the influence of these characteristics on the model. Figure 1 shows the importance of the top ten features, measured by SHAP values, and their impact on model predictions.

As shown in Figure 1a, the most important feature is the Standard Deviation of RR Interval (STD RR (s)), with an average SHAP value of 1.94, followed by Median RR (s) and Mean RR (s). The median duration of the PR interval (Median PR (s)), 5th percentile of the Signal-to-Noise Ratio estimation for each beat (SNR1), spectral power of the signal in the 8–20 Hz component (Comp. Frequency 4), one of the statistics of wavelet decomposition coefficients used (Stat_A8_9), and STD Amplitude QRS (STD Amp QRS) are also highly important.



(a) Average SHAP Value



(b) Density

Figure 1: Ten most important features for the model.

Additionally, the "sum of other features" represents the cumulative contribution of all other features not ranked among the top ten. While these features individually contribute less to the model, their total influence remains significant in the overall predictions.

Figure 1b provides further insight into how the values of these features impact the model's output. High values of STD RR (s) and Median PR (s) are associated with increased SHAP values, indicating a stronger impact on predictions, whereas lower values of Mean RR (s) and STD Amp QRS suggest a smaller effect on model output.

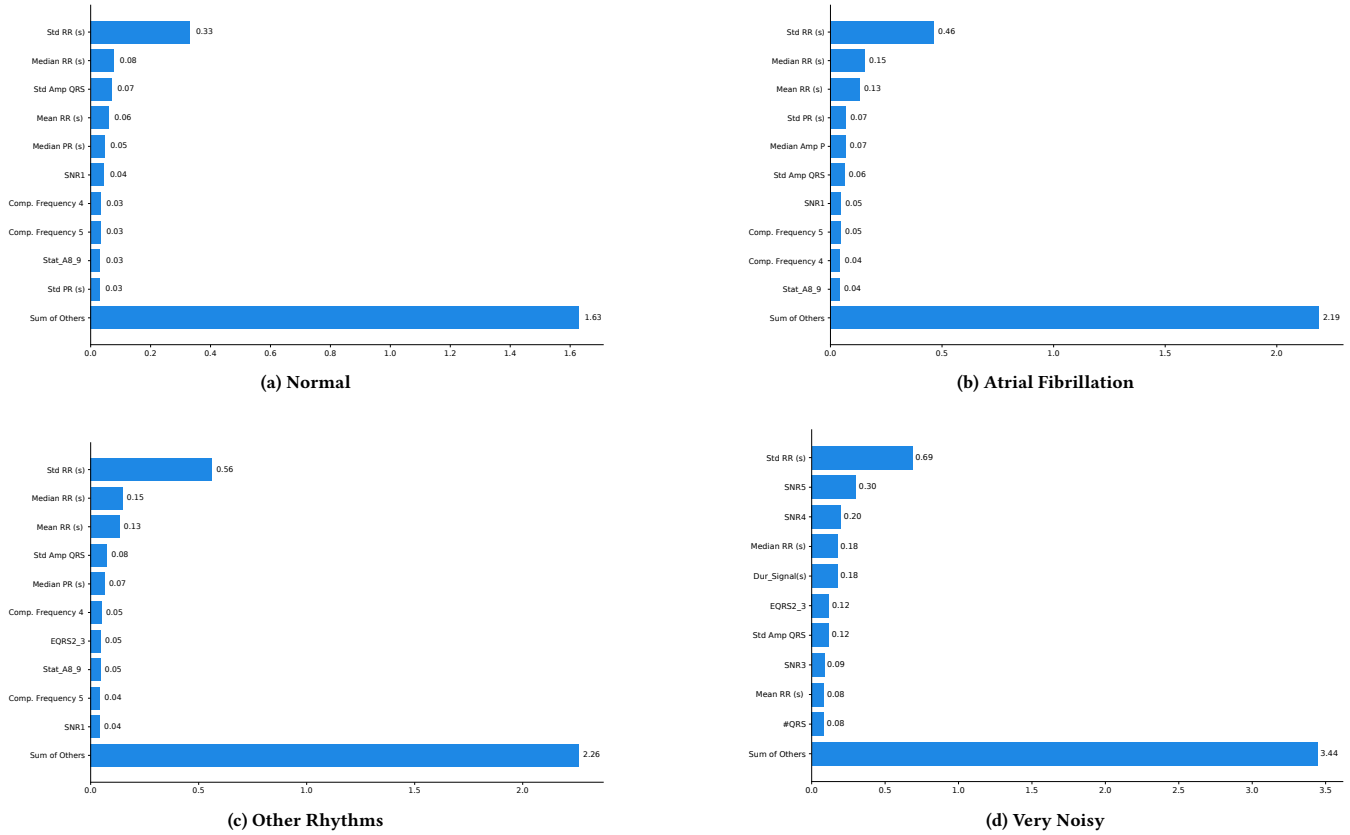
The overall analysis shows that, while certain features stand out, the cumulative contribution of other minor features also plays a relevant role in the model's overall behavior, especially in more complex classification scenarios.

5.4 Explanations at the Class Level

Research question RQ3 examines the most relevant features for classifying different heart rhythms. The analysis for each class is presented in Figure 2.

Table 4: Evaluation of the explainers using various metrics.

Explainer	Explanation Error	Compute Time	Keep Positive	Remove Positive	Keep Negative	Remove Negative	Keep Absolute	Remove Absolute
Permutation part.	**0.2783**	6.3782	0.8298	**-0.5660**	-0.6077	1.0072	**-0.8712**	0.9512
Permutation	0.3180	6.8237	**0.8335**	-0.5654	**-0.6084**	**1.0146**	-0.8681	**0.9586**
Partition	0.3841	7.6365	0.1748	-0.2394	-0.1923	0.3140	-0.4531	0.5573
Random	0.4577	0.1448	-0.0087	0.0707	-0.0008	0.0586	-0.4398	0.5671
Tree approx.	2.2692	**0.0064**	-0.0998	0.6019	0.4692	-0.0051	-0.8612	0.9469
Tree	2.2983	0.0210	-0.0969	0.6106	0.4821	-0.0048	-0.8585	0.9467



(a) Normal

(b) Atrial Fibrillation

(c) Other Rhythms

(d) Very Noisy

Figure 2: Mean of the ten most contributing features for each class.

Class-level analysis reveals that RR Interval Variability (*Std RR (s)*), RR Interval Median (*Median RR (s)*), and Signal-to-Noise Ratio for the first channel (*SNR1*) are significant for distinguishing between normal and abnormal heart rhythms. In the normal rhythm class, as shown in Figure 2a, *Std RR (s)* has the highest mean SHAP value (0.33), followed by *Median RR (s)* (0.08) and *Std Amp QRS* (0.07).

For the Atrial Fibrillation class, as shown in Figure 2b, *Std RR (s)* stands out with a mean SHAP value of 0.46, followed by *Median RR (s)* (0.15) and *Mean RR (s)* (0.13). Other important features include the *Median P-Wave Amplitude* (*Median Amp P*) and the Standard Deviation of the PR Interval (*Std PR (s)*), both with SHAP values of 0.07. The *Signal-to-Noise Ratio for the first channel* (*SNR1*) also contributes with a value of 0.05.

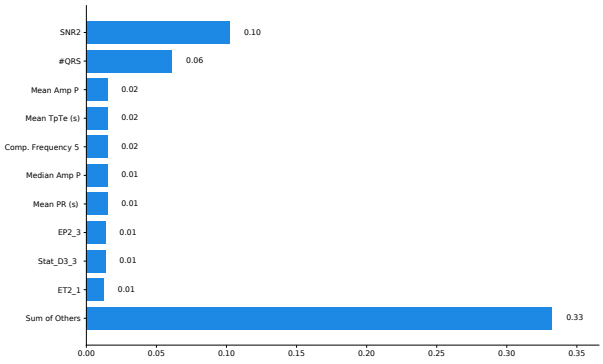
In the Other Rhythms class, shown in Figure 2c, *Std RR (s)* is the most important feature (0.56), followed by *Median RR (s)* (0.15) and *Mean RR (s)* (0.13).

Finally, for very noisy data, as represented in Figure 2d, *Std RR (s)* has the highest mean SHAP value (0.69), followed by *SNR5* (0.30) and *SNR4* (0.20). The presence of these noise-related features indicates their importance in classifying records with high interference.

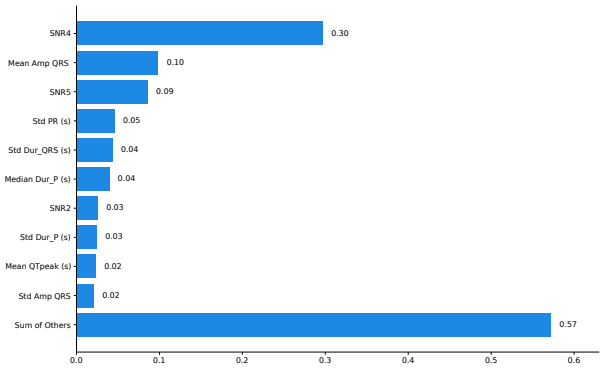
5.5 Explanations at the Local Level

Based on the proposed methodology, it was possible to obtain local explanations that provide valuable insights into specific predictions. Local explanations are fundamental for detailed analysis of individual cases, allowing for comparisons between different types of samples, such as patients with and without Atrial Fibrillation. This approach enables more in-depth case studies, offering a clearer understanding of the characteristics that directly influence the model's decisions in each specific instance.

In Figure 3, we present the ten characteristics that most contributed to the classification of a normal case and a case of Atrial Fibrillation. These explanations facilitate comparative analysis between different types of samples, helping to identify the patterns that are key to the classification.



(a) Normal Case



(b) Atrial Fibrillation

Figure 3: Ten most important features for the classification of a normal case and a case of Atrial Fibrillation.

These local explanations allow doctors and researchers to understand better how the model makes its decisions and which characteristics are most associated with certain conditions. In the example of a typical case, characteristics such as SNR2, Mean PR (s), and Median Amp P played relevant roles in the classification. In contrast, in the case of Atrial Fibrillation, characteristics like Std Amp QRS, Mean QTpeak (s), and Std Dur_P (s) were decisive for diagnosing the condition. Therefore, local explanations can be used in case studies, facilitating the interpretation and reliability of the model's predictions.

6 Conclusions and future work

The interpretability of machine learning models has become fundamental in applying highly complex algorithms, such as Histogram-based Gradient Boosting, in the classification of cardiac rhythms in ECG. This study identified critical ECG features such as RR interval variability, median values of RR and PR intervals, and the signal-to-noise ratio, which play a significant role in the model's predictive accuracy. These insights support confidence and acceptance among healthcare providers and patients and highlight how information systems can enhance clinical decision-making by providing interpretable and actionable data insights.

The challenges in interpreting these models include the intrinsic complexity of advanced algorithms, making it difficult to explain their decisions directly. This work overcomes these challenges by exploring the use of SHAP as an effective explainability technique, allowing a detailed analysis of the importance of each feature in the classification of heart rhythms. Such interpretability methods bridge the gap between complex ML models and their practical application in information systems, enabling professionals to leverage these systems with greater trust.

Furthermore, this work stands out by employing an approach that uses the extracted features instead of the ECG signals, bringing explanations at global, class, and local levels, which related works still need to achieve.

Utilizing metrics for explainability evaluation is also crucial, as it ensures that the interpretative results are understandable and contextually useful in health systems, thus increasing their usefulness for professionals and stakeholders.

However, one of the main limitations of this study was the excessive processing time required when applying advanced explainability techniques to weighted ensemble models. Although these models demonstrated high performance in arrhythmia classification, interpreting their decisions using techniques such as SHAP incurred a high computational cost. In some cases, generating the explanations took weeks to complete, making their application infeasible in scenarios requiring rapid responses, such as real-time clinical decision support systems. This limitation highlights the need to explore more efficient interpretability approaches that balance accuracy and computational feasibility without compromising the model's practical utility.

For future work, this study recommends deploying and validating these models within clinical information systems to assess real-world effectiveness and user adoption among healthcare professionals. Additionally, exploring other explainability techniques, such as generating rule-based explanations or more complex models, could improve the interpretability of machine learning models applied to cardiovascular health. These advancements in interpretability contribute significantly to the reliability and practical integration of AI-based information systems in health contexts, promoting broader adoption and trust in AI-driven clinical solutions.

Acknowledgments

We would like to thank Samsung for supporting and funding this research, which enabled us to develop and carry out this work. The company's support was fundamental to the progress of the research and analysis activities, contributing significantly to the results achieved.

References

[1] Matteo Bodini, Massimo W Rivolta, and Roberto Sassi. 2021. Opening the black box: interpretability of machine learning algorithms in electrocardiography. Philosophical Transactions of the Royal Society A 379, 2212 (2021), 20200253.

[2] Philine Lou Bommer, Marlene Kretschmer, Anna Hedström, Dilyara Bareeva, and Marina M.-C. Höhne. 2024. Finding the Right XAI Method—A Guide for the Evaluation and Ranking of Explainable AI Methods in Climate Science. Artificial Intelligence for the Earth Systems 3, 3 (2024), e230074. https://doi.org/10.1175/AIES-D-23-0074.1

[3] Federico Cabitza, Andrea Campagner, and Clara Balsano. 2020. Bridging the "last mile" gap between AI implementation and operation: "data awareness" that matters. Annals of translational medicine 8, 7 (April 2020), 501.

[4] Federico Cabitza, Andrea Campagner, and Luca Maria Sconfienza. 2020. As if sand were stone. New concepts and metrics to probe the ground on which to build trustable AI. BMC Medical Informatics and Decision Making 20, 1 (11 Sep 2020), 219. https://doi.org/10.1186/s12911-020-01224-9

[5] Federico Cabitza, Davide Ciucci, and Raffaele Rasoini. 2019. A Giant with Feet of Clay: On the Validity of the Data that Feed Machine Learning in Medicine. In Organizing for the Digital World, Federico Cabitza, Carlo Batini, and Massimo Magni (Eds.). Springer International Publishing, Cham, 121–136.

[6] Davide Chicco, Anna Karaiskou, and Maarten De Vos. 2024. Ten quick tips for electrocardiogram (ECG) signal processing. PeerJ Computer Science 10 (2024), e2295. https://doi.org/10.7717/peerj-cs.2295

[7] Gari D Clifford, Chengyu Liu, Benjamin Moody, Li-Wei H Lehman, Ikaro Silva, Qiao Li, A E Johnson, and Roger G Mark. 2018. AF Classification from a Short Single Lead ECG Recording: the PhysioNet/Computing in Cardiology Challenge 2017.

[8] Nick Erickson, Jonas Mueller, Alexander Shirkov, Hang Zhang, Pedro Larroy, Mu Li, and Alexander Smola. 2020. AutoGluon-Tabular: Robust and Accurate AutoML for Structured Data. arXiv:2003.06505 [stat.ML] https://arxiv.org/abs/2003.06505

[9] A. Goldberger, L. Amaral, L. Glass, J. Hausdorff, P. C. Ivanov, R. Mark, ..., and H. E. Stanley. 2000. PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. Circulation 101, 23 (2000), e215–e220. https://www.ahajournals.org/doi/10.1161/01.cir.101.23.e215

[10] Yola Jones, Fani Deligianni, and Jeff Dalton. 2020. Improving ECG Classification Interpretability using Saliency Maps. , 675-682 pages. https://doi.org/10.1109/BIBE50027.2020.00114

[11] Eunchan Kim, Jaehyuk Kim, Juyoung Park, Haneul Ko, and Yeunwoong Kyung. 2023. TinyML-Based Classification in an ECG Monitoring Embedded System. Computers, Materials & Continua 75, 1 (2023), 1751–1764. https://doi.org/10.32604/cmc.2023.031663

[12] Zachary C. Lipton. 2018. The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. Queue 16, 3 (jun 2018), 31–57. https://doi.org/10.1145/3236386.3241340

[13] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su-In Lee. 2020. From local explanations to global understanding with explainable AI for trees. Nature Machine Intelligence 2, 1 (01 Jan 2020), 56–67. https://doi.org/10.1038/s42256-019-0138-9

[14] Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems (Long Beach, California, USA) (NIPS'17). Curran Associates Inc., Red Hook, NY, USA, 4768–4777.

[15] George B. Moody and Roger G. Mark. 2001. The impact of the MIT-BIH Arrhythmia Database. IEEE Engineering in Medicine and Biology Magazine 20 (2001), 45–50. https://api.semanticscholar.org/CorpusID:18619383

[16] Inês Neves, Duarte Folgado, Sara Santos, Marília Barandas, Andrea Campagner, Luca Ronzio, Federico Cabitza, and Hugo Gamboa. 2021. Interpretable heartbeat classification using local model-agnostic explanations on ECGs. Computers in Biology and Medicine 133 (2021), 104393. https://doi.org/10.1016/j.compbiomed.2021.104393

[17] Jaya Ojha, Hårek Haugerud, Anis Yazidi, and Pedro G. Lind. 2024. Exploring Interpretable AI Methods for ECG Data Classification. In Proceedings of the 5th ACM Workshop on Intelligent Cross-Data Analysis and Retrieval (ICDAR '24). Association for Computing Machinery, New York, NY, USA, 11–18. https://doi.org/10.1145/3643488.3660294

[18] M. Ramkumar, C. Ganesh Babu, K Vinoth Kumar, D Hepsiba, A. Manjunathan, and R. Sarath Kumar. 2021. ECG Cardiac arrhythmias Classification using DWT, ICA and MLP Neural Networks. Journal of Physics: Conference Series 1831, 1 (mar 2021), 012015. https://doi.org/10.1088/1742-6596/1831/1/012015

[19] Amit Sangroya, Suparshva Jain, Lovekesh Vig, C. Anantaram, Arijit Ukil, and Sundeep Khandelwal. 2022. Generating Conceptual Explanations for DL based ECG Classification Model. https://doi.org/10.32473/flairs.v35i.130681

[20] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. , 618-626 pages. https://doi.org/10.1109/ICCV.2017.74

[21] T Sullivan. 2024. Half of Hospitals to Adopt Artificial Intelligence Within 5 Years. Healthcare IT News. https://www.healthcareitnews.com/news/half-hospitals-adopt-artificial-intelligence-within-5-years Accessed: 2024-11-14.

[22] Mudrakola Swapna, Uma Maheswari Viswanadhula, Rajanikanth Aluvalu, Vijayakumar Vardharajan, and Ketan Kotecha. 2022. Bio-Signals in Medical Applications and Challenges Using Artificial Intelligence. https://doi.org/10.3390/jsan11010017

[23] Jia Wu, Xiu-Yun Chen, Hao Zhang, Li-Dong Xiong, Hang Lei, and Si-Hao Deng. 2019. Hyperparameter Optimization for Machine Learning Models Based on Bayesian Optimizationb. Journal of Electronic Science and Technology 17, 1 (2019), 26–40. https://doi.org/10.11989/JEST.1674-862X.80904120

Received 20 February 2007; revised 12 March 2009; accepted 5 June 2009