Explainable AI Analysis of a Prediction Model for Detecting Premature Atrial and Ventricular Complexes

Pedro A Moreno-Sánchez[1,*], Guadalupe García-Isla[2], Valentina Corino[2], Mark van Gils[1], Luca Mainardi[2]

[1] Faculty of Medicine and Health Technology, Tampere University, Tampere, Finland
[2] Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milano, Italy.

Abstract

The relationship between premature atrial complexes (PACs) and cardiovascular diseases remains elusive, with existing PAC detectors based on beat classification demonstrating low sensitivity. PAC detectors based on Machine Learning (ML) often lack interpretability, impeding their adoption by cardiologists. Using Explainable AI (XAI) techniques, such as SHapley Additive exPlanations (SHAP), this study enhances the interpretability of a highly accurate (avg. accuracy: 0.985) PAC detector. This detector utilizes a random forest (RF) classifier for normal (N), supraventricular (S), and ventricular (V) beats using ECG-derived features, including heart rate variability (HRV) and QRS complex morphology. Our findings reveal that RR interval features predominantly influence the detection of N and S classes, while QRS morphology critically impacts V class predictions. By refining the model to use only 14 key features from an original set of 185, we developed simpler, surrogate models by using RF and decision trees. Although there is a slight decline in performance—most notably in the sensitivity and positive predictive value (PPV) for S class detection—these models maintain substantial predictive power, hence, underscoring the potential of XAI in building interpretable PAC detectors.

1 Introduction

Premature atrial complexes (PACs) were traditionally viewed as benign, however, recent research suggests a correlation between frequent PACs and the initial onset of Atrial fibrillation (AF). Other studies propose PACs as a direct cause of stroke, cardiac tissue deterioration and implicated in left ventricular remodeling [1].

Manually annotating PACs in long-term electrocardiogram (ECG) recordings is labor-intensive and requires specialized expertise. There is a pressing need for a highly sensitive PAC detector capable of automating this process, because existing solutions are primarily beat classifiers with low PAC detection sensitivity [2]. Such a detector would facilitate research into the effects of PACs on the onset of AF and cardiac tissue remodelling, and assess stroke risk. Moreover, improving the accuracy of this detector could reduce the rate of false positives in arrhythmia detection, where several PAC beats often mimic AF. [3].

Computer-assisted ECG interpretation has progressed from traditional feature detection and rule-based classification to more advanced, data-driven approaches like machine learning (ML) and deep learning (DL). These methods have improved accuracy and integrated multiple modalities, yielding better predictions of cardiovascular outcomes than traditional techniques [4]. However, their clinical adoption is constrained by challenges such as data collection, workflow integration, external validation, and their opaque "black box" nature, which obscures decision-making processes and hinders trust among clinicians. To mitigate this, Explainable Artificial Intelligence (XAI) aims to clarify how AI systems make decisions, enhancing their understandability and facilitating broader clinical acceptance and trust.

The scientific literature on often highlights that ECG-based ML models, like deep learning architectures and ensemble trees, suffer from opaque decision-making processes. Despite the importance of Explainable Artificial Intelligence (XAI) for the adoption of these models, there are still few studies focused on this critical aspect [4]. It's noteworthy that the XAI techniques in the reviewed studies are mainly post-hoc, applied after training to interpret model predictions. These methods highlight the importance of ECG-derived features in the AI model, aiming to explain how these features influence the model's decision-making processes.

This paper tackles the noted research gaps: the lack of an automated PAC detector and the absence of explainability in ML solutions for early ECG-based diagnosis. We present an explainability analysis of a PAC detector, developed by the authors in [5], which uses feature processing and Random Forest classifier. From the analysis, we identify key features for classifying normal, supraventricular, and ventricular beats. These features are then used to assess the performance of various surrogate models.

2 Material and methods

2.1 Data

This study utilized two public PhysioNet databases [6]: the long-term ST database (LTSTDB) and the supraventricular database (SVDB), which include 2-lead ECG signals recorded at 250 Hz and 128 Hz over durations of 21–24 hours and 30 minutes, respectively. Chosen for their high quantity of PACs and manual annotations, the datasets were merged to utilize their complementary PAC representations for model training and testing.

Table 1 lists the number of beats per class in each database, using five categories: Normal (N), Supraventricular (S), Ventricular (V), Junctional (J), and Unclassifiable (Q) beats. In our study, S beats encompassed atrial premature beats (A), aberrated atrial premature beats (a), and PACs (S) annotations. V class included premature ventricular contractions (V), ventricular-normal beat fusions (F), and ventricular escape beats (E). N class comprised normal beats (N), bundle branch block beats (B), and atrial escape beats (e). J and Q classes were omitted from further analysis due to their low representation in the databases. Throughout this paper, N, S, and V refer to these classifier categories, with S specifically denoting PACs and V including ventricular-related beats (V, E, and F).

2.2 Feature extraction and model

This study examines the explainability of a previously developed PAC detector model, depicted in Figure 1. The model's pipeline includes four phases: signal preprocessing, ECG feature extraction, feature preprocessing, and classification using a Random Forest classifier [5].

The signal preprocessing homogenizes the sampling frequencies of the datasets and removes powerline and high-frequency noise. Subsequently, 185 features are extracted to characterize key ECG properties, including heart rate variability (HRV) and wave morphology. HRV features such as RR intervals, differences between consecutive RR intervals, and their mean and standard deviations are computed for each beat and its neighbors. Additional metrics include the percentage of successive interval differences exceeding thresholds from 10 to 50 ms and the root mean square of successive differences (RMSSD), contributing to a total of 41 HRV features.

Morphological features of the P wave, QRS complex, PR segments, and entire beat were extracted using a fixed window centered on the R peak. Before extracting these segments, three intra-patient templates were generated using 80 (long-term), 40 (mid-term), and 4 (short-term) neighboring beats for comparison. This process yielded 72 morphological features, including the maximum cross-correlation value for each segment against the templates, the logarithm of the cross-correlation value, and the median standard deviation of the beats. Similarly, an additional 72 morphological features were derived from a filtered version of the ECG using the discrete wavelet transform (DWT).

The model used a patient-wise 10-fold cross-validation method, ensuring no patient's beats were divided between training and testing sets. The dataset was divided into ten subsets, each maintaining a consistent proportion of S class beats to account for their imbalanced distribution. During each cross-validation cycle, nine subsets were used for training and one for testing. To prevent patient-specific bias, a total of 30,000 beats was set for training (10,000 for each class). The classification performance of the test set is detailed in Table 2.

2.3 Ensemble trees classifiers

Ensemble trees, renowned for their stability, robustness across different dataset sizes, and solid predictive performance, are highly favored in modern ML classifiers. These models improve prediction by aggregating various

Table 1. Simplified beat annotations per PhysioNet database.

PAC classes	N			S			V			J		Q
Beat annotations	B	N	E	A	A	S	V	F	E	J	j	Q
LTSTDB	88,720	6,727,000	22	5,482	29	30,820	39,840	476	71	1	6	2
SVDB	1	162,100	0	0	1	12,090	9,930	23	0	9	0	80

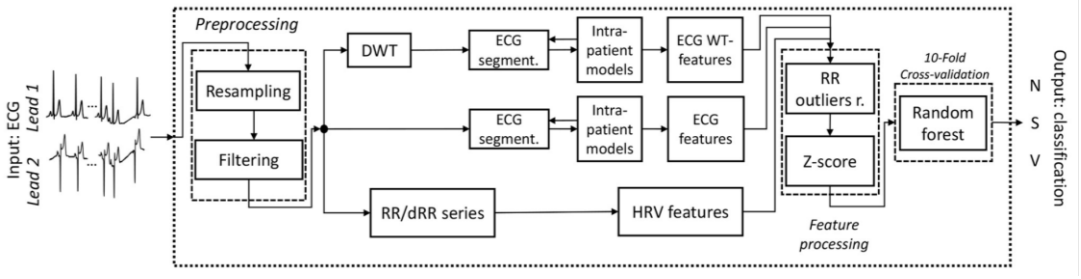

Figure 1. PAC detector classifier structure. Figure is freely adapted from [4]

decision trees, each contributing to the final model to enhance performance over individual estimators.

Random Forest (RF), a prominent ensemble tree method, uses the bagging method to train its decision trees, randomly selecting features at each node split, which helps prevent overfitting. In this study, alongside RF, raw decision trees (DT) are also used, employing their intuitive, tree-like structure that facilitates easy interpretation and visualization of the decision paths, making them an effective tool for developing a surrogate model of the PAC detector. [7].

2.4 XAI-SHAP

To analyze the explainability of the original PAC detector, we used the post-hoc XAI technique SHAP (Shapley Additive Explanations), which calculates an additive importance score for each prediction. Utilizing principles from coalitional game theory, SHAP assesses the contribution of each feature to the prediction, providing a signed importance score that shows both the weight and direction of a feature's impact on the predicted outcome[8].

3 Results

3.1 XAI results

The explainability of the original PAC detector was assessed through calculating SHAP values for each dataset instance. For each fold, a SHAP bar plot was created to show the absolute relevance of the most influential features in classifying the three beat classes. An example from fold #1 is illustrated in Figure 2.
The bar plots did not reveal clear repetitive patterns across the ten folds to identify the most relevant features for predicting the three classes. Consequently, we combined all SHAP values from the ten folds to create a comprehensive global bar plot, displayed in Figure 3. This plot highlights that the features with the highest mean absolute contributions—dRRi0, dRRin1, and RRi0—have significant impact on classes N and S but are less influential for class V. For class V, other features such as QRS_xcorr_Raw_10_10_L1 or QRS_xcorr_Raw _40_40_L2 play more critical roles.

Table 2. Classification performance of original model expressed as the percentage mean and standard deviation (in parentheses), averaged across the three classes

PAC classes	Acc.	Sens.	Spec.	PPV.	NPV.
N	97.63 (1.55)	97.62 (1.57)	97.8 (1.33)	99.97 (0.02)	39.75 (15.36)
S	98.46 (1.08)	89.94 (8.68)	98.49 (1.10)	31.97 (15.90)	99.95 (0.03)
V	98.99 (0.93)	90.02 (7.75)	99.06 (0.97)	40.74 (27.19)	99.93 (0.08)

Acc: Accuracy, Sens: Sensitivity, Spec: Specificity, PPV. Positive Predictive Values, NPV: Negative Predictive Values.

Additionally, we grouped the features by typologies for the bar plot analysis, defining five categories: RR interval-wise, Beats-feature, QRS-complex, P and PR points. Figure 4 illustrates their influence on the three classes. The impact on model output for each class, in descending order, is as follows: for class N, RR, Beat, QRS, P, and PR; for class S, RR, QRS, Beat, P, and PR; for class V, QRS, Beat, RR, P, and PR, with P and PR having equal impact.

3.2 Features selection

The importance of the features across the folds was used to select the most relevant ones for developing the surrogate models. By visually inspecting the 10 bar plots, we set a SHAP value threshold of 0.075 to identify the most crucial features.
Out of 44 features deemed relevant by SHAP, 14 exceeded this threshold, significantly reducing the original count of 185 features used in the PAC detector.

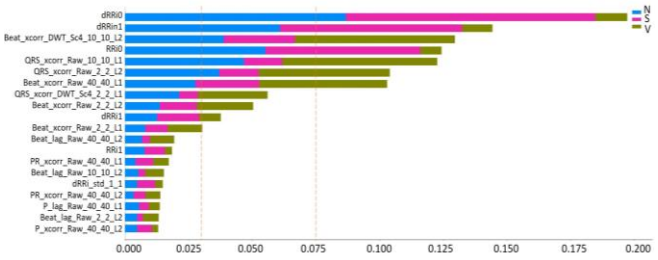

Figure 2. Example of bar plot for fold 1.

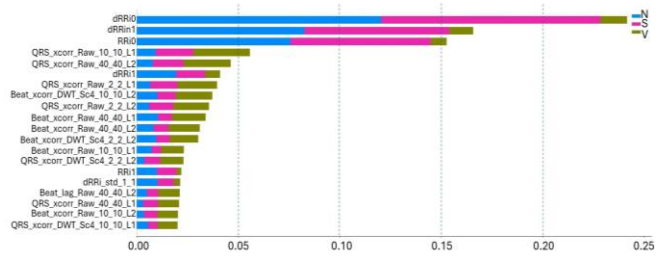

Figure 3. 10-fold global bar plot for individual features

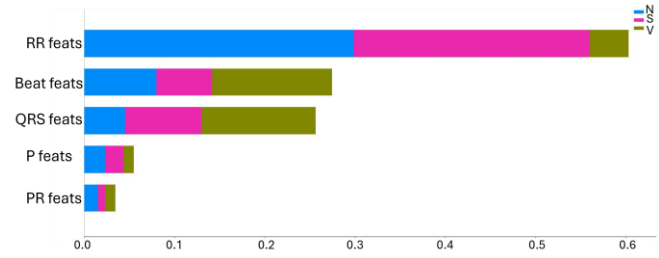

Figure 4: 10-fold global bar plot for ECG points categories

3.3 Performance of surrogate models

Using the 14 features identified by SHAP, we retrained the original PAC detector pipeline, with performance detailed in Tables 3 and 4. Aiming for an explainable PAC

detector, we replaced the RF classifier with a DT, known for its transparency. The performance of this fully-explainable surrogate model is presented in Table 5. Differences in performance compared to the original model were statistically tested using the Wilcoxon Signed-Rank Test ($p<0.05$).

Table 3. Classification performance of RF surrogate model with 44 features. *=Significant difference with original model. Underlined= a decrease in 5% respect to original model

PAC classes	Acc.	Sens.	Spec.	PPV.	NPV.
N	97.19* (1.76)	97.18* (1.79)	97.30 (3.20)	99.97 (0.03)	35.27* (14.61)
S	98.35 (1.21)	88.33 (10.72)	98.39 (1.24)	30.93 (16.43)	99.94 (0.05)
V	98.66* (1.14)	89.96 (8.62)	98.72 (1.17)	34.40* (26.88)	99.93 (0.08)

Table 4. Classification performance of RF surrogate model with 14 features.

PAC classes	Acc.	Sens.	Spec.	PPV.	NPV.
N	96.99* (1.78)	96.97* (1.80)	97.89 (1.19)	99.97 (0.02)	33.89 (16.88)
S	98.06* (1.16)	89.11 (8.65)	98.10* (1.18)	25.64* (14.71)	99.95 (0.04)
V	98.75 (1.07)	89.68 (8.36)	98.80 (1.10)	36.36 (26.76)	99.93 (0.08)

Table 5: Classification performance of DT surrogate model with 14 features

PAC classes	Acc.	Sens.	Spec.	PPV.	NPV.
N	93.34* (1.82)	93.32* (1.87)	93.81* (6.71)	99.92* (0.07)	16.52* (7.86)
S	96.11* (1.54)	79.53* (16.03)	96.19* (1.56)	13.08* (8.37)	99.88* (0.10)
V	96.94* (1.41)	85.94* (8.23)	97.03* (1.46)	18.84* (20.40)	99.89* (0.15)

4 Discussion and conclusions

This study demonstrated the potential of addressing the explainability of PAC detectors in ECG analysis. By integrating SHAP with RF and DT, we identified key features that influence model predictions. This significantly reduced the number of original ECG features (from 185 to 14) enhancing the interpretability capabilities of the detector. The transition to a DT for the surrogate model benefits with transparency and simplicity the understanding of the decision-making process. Despite a slight decline in performance—most notably in the sensitivity and PPV for S and V class detection—these models maintain substantial predictive power. Thus, this study represents a significant step towards building explainable AI tools in cardiology facilitating their clinical adoption. Future work should focus on inspecting the performance differences between the two inner databases (LTSTDB and SVDB), and extracting the decision rules from the DT to optimize both accuracy and explainability, ensuring that they can effectively support clinical decision-making in real-world settings.

Acknowledgments

This work is funded by the project PerCard (Personalised Prognostics and Diagnostics for Improved Decision Support in Cardiovascular Diseases) in ERA PerMed supported by the Research Council of Finland (decision number 351846), under the frame of ERA PerMed; and by the Fondazione Regionale della Ricerca Biomedica (FRRB) under the frame of ERA PerMed. L.M. is an additional member of the PNRR-PE-AI FAIR project funded by the Next Generation EU program.

References

[1] B. Huang *et al.*, "Relation of premature atrial complexes with stroke and death: Systematic review and meta-analysis," *Clinical Cardiology*, vol. 40, no. 11, pp. 962–969, 2017, doi: 10.1002/clc.22780.

[2] M. Llamedo and J. P. Martínez, "Heartbeat Classification Using Feature Selection Driven by Database Generalization Criteria," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 3, pp. 616–625, Mar. 2011, doi: 10.1109/TBME.2010.2068048.

[3] L. Sörnmo, A. Petrėnas, and V. Marozas, "Detection of Atrial Fibrillation," in *Atrial Fibrillation from an Engineering Perspective*, L. Sörnmo, Ed., Cham: Springer International Publishing, 2018, pp. 73–135. doi: 10.1007/978-3-319-68515-1_4.

[4] P. A. Moreno-Sánchez *et al.*, "ECG-based data-driven solutions for diagnosis and prognosis of cardiovascular diseases: A systematic review," *Computers in Biology and Medicine*, vol. 172, p. 108235, Apr. 2024, doi: 10.1016/j.compbiomed.2024.108235.

[5] G. García-Isla, L. Mainardi, and V. D. A. Corino, "A Detector for Premature Atrial and Ventricular Complexes," *Front. Physiol.*, vol. 12, p. 678558, Jun. 2021, doi: 10.3389/fphys.2021.678558.

[6] A. L. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet," *Circulation*, vol. 101, no. 23, pp. e215–e220, Jun. 2000, doi: 10.1161/01.CIR.101.23.e215.

[7] P. A. Moreno-Sanchez, "Development of an Explainable Prediction Model of Heart Failure Survival by Using Ensemble Trees," in *2020 IEEE International Conference on Big Data (Big Data)*, Dec. 2020, pp. 4902–4910. doi: 10.1109/BigData50022.2020.9378460.

[8] S. M. Lundberg *et al.*, "Explainable machine-learning predictions for the prevention of hypoxaemia during surgery," *Nature Biomedical Engineering*, vol. 2, no. 10, Art. no. 10, Oct. 2018, doi: 10.1038/s41551-018-0304-0.

Address for correspondence:

Pedro A. Moreno-Sánchez. Sähkötalo Building Korkeakoulunkatu 3, 33720 Tampere, Finland. *pedro.morenosanchez@tuni.fi*