Morphology Features Self-Learned by Explainable Deep Learning for Atrial Fibrillation Detection Correspond to Fibrillatory Waves

Alexander Hammer, Hagen Malberg, Martin Schmidt

Institute of Biomedical Engineering, TU Dresden, Dresden, Germany

Abstract

The main challenge in utilizing deep learning (DL) for clinical diagnostic support is its lack of explainability and interpretability. Recent approaches aim to explain DL decisions from electrocardiogram (ECG) analysis by tracing model explanations back to beat segments. Fibrillatory (F) waves, as a main characteristic of atrial fibrillation (AF), are irregularly distributed over the signal and have not yet been considered. Using 477 publicly available AF ECGs, we systematically investigated the relationship between F waves and reliable model explanations. F waves were detected using peak detection after removing beat-aligned QT templates. We employed a convolutional neural network, derived from an explainable ECG architecture (xECGArch), which uses self-learned morphology features for AF detection. Analysis of variance revealed an increased mean relative relevance (rR) of the F waves compared to the rR of the full waveform (+13.5 %, p<.001). When limiting F wave detection to areas without overlap with other morphologic features, the rR increased by 13.1 % and exceeded the rR of the full waveform by 28.3 % (p<.001). The rR peaking in the flank facing the QRS complex indicates the importance of the distance to QRS complexes for differentiation from P waves. For the first time, we attributed DL explanations to F waves, improving DL interpretability, which is essential for clinical use.

1. Introduction

Atrial fibrillation (AF) is the most prevalent cardiac arrhythmia globally [1]. Early detection and treatment of AF can reduce the risk of morbidity and mortality, which are significantly increased without intervention [1]. However, AF is often paroxysmal in its early stages [2], which makes diagnosis challenging. The use of mobile electrocardiography devices, along with automated AF detection algorithms, can enhance the probability of early detection of AF episodes. Deep learning (DL) approaches have achieved high accuracy in detecting AF on single-lead electrocardiograms (ECGs), indicating significant potential in mobile applications [3]. However, DL approaches lack explainability in their decision-making process due to their black-box nature, which reduces trustworthiness and their value for diagnostic support [4]. An increasing number of papers are utilizing explainability approaches for artificial intelligence (xAI) to elucidate the black box. However, few papers have aimed to validate the causality of model explanations through systematic investigations. Previous systematic investigations have taken either a pseudo-quantitative template beat-based approach (e.g., [5]) or have analyzed the mean relevance of beat segments (e.g., [6]). Fibrillatory (F) waves, one of the main characteristics of AF in the ECG, superimpose irregularly on the ECG [7]. Thus, their relevance for classification cannot be investigated systematically using existing methods. We hypothesize that DL approaches can learn features corresponding to F waves and use them for AF detection. To investigate this, we present a method for systematically analyzing the relevance of F waves to DL methods. Furthermore, we investigate the relationship between model explanations and the occurrence of F waves in an extensive public data set.

2. Methods

2.1. Data material and model explanations

We used the short-term model of the explainable ECG architecture (xECGArch), which was recently introduced [5]. xECGArch combines two 1-dimensional convolutional neural networks (CNNs) with different time horizons. The short-term model has an receptive field of 0.6 s and primarily learns morphological features, while the long-term model with an receptive field of 10 s primarily learns rhythmic features [5].

xECGArch was trained and validated on 9 854 10-second single-lead ECGs (Einthoven II) from 4 public databases (DBs) to discriminate AF from non-AF ECGs and tested on 986 unseen ECGs from the same DBs. For our investigations, we used the 507 ECGs of the test set containing AF and excluded 30 misclassified ECGs. We

extracted model explanations in terms of sample-wise relevance for classification using deep Taylor decomposition (DTD), as DTD provided the most trustworthy explanations for this scenario in a systematic comparison of 13 xAI methods using perturbation [5]. We then determined relative relevance (rR) values by recording-wise normalization to the respective maximum relevance.

2.2. F wave detection

To examine the relation between F waves and model explanations, we extracted the signal component of estimated atrial activity $\hat{s}_{aa}$, then detected F waves using peak detection and examined the model explanation within the F waves. For extracting $\hat{s}_{aa}$, we subtracted beat-adjusted QT templates that were generated by template delineation using iterative two-dimensional signal warping (i2DSW) [8, 9], applied in line with [10] from the preprocessed signal. If two consecutive beats were more than the mean beat distance plus 2.5 times the standard deviation apart, the area in between was interpolated linearly to compensate for missed beats. Furthermore, we padded signal edges preceding or following the first or last detected beat with the first or last valid value. Subsequently, the residual signal was band-pass filtered between $5 - 10$ Hz, which is the approximate frequency of F waves, resulting in the remaining signal $\hat{s}_{aa}$.

For irregularly aperiodic occurring waves, we assumed that $\hat{s}_{aa}$ corresponds to F waves and executed a peak detection. We utilized auto-correlation with a maximum lag of 3 s to check for periodicity in $\hat{s}_{aa}$. Subsequently, we counted the number of lag indices n_p in which the auto-correlation peaked, ensuring a minimum height r_h^{thr} of

$$r_h^{thr} = \max(1.5 \cdot \bar{r}_h, 0.2), \tag{1}$$

a minimum prominence r_p^{thr} of

$$r_p^{thr} = \max(2.8 \cdot \bar{r}_h, 0.3), \tag{2}$$

and an interbeat interval of at least 0.33 s. $\bar{r}_h$ was defined as the mean correlation value of all lag indices where the auto-correlation peaked. Threshold values were determined empirically. From a threshold value of $n_p > 2$, $\hat{s}_{aa}$ was assumed to contain periodic P waves, and the ECGs were excluded for F wave detection. The threshold value was determined using the original test dataset [5] to most effectively differentiate ECGs with AF from non-AF ECGs.

Subsequently, F waves were detected in the remaining ECGs using the MATLAB function `findpeaks`. We set the maximum peak width to $w_{max}^F = {}^1\!/_{5\,\text{Hz}} \cdot 1.25 = 0.25$ s, which corresponds to 1.25 times the width of F waves when oscillating at 5 Hz. In a grid search, we examined the optimal peak prominence in the interval $20 : 3 : 50\,\mu$V and the polarity of the F waves (positive: peak detection, negative: valley detection) to maximize the median of the mean rR values of all F waves. We defined the on- and offsets of the F waves as the preceding and following troughs, for peak detection, or peaks, for trough detection. Since F waves are usually particularly visible in signal sections without cardiac excitation-related waves, such as the TQ segment (T wave end of beat n to Q wave of beat $n + 1$), we examined the rR of the F waves within the entire signal (F_{Full}) as well as only within the TQ segments (F_{TQ}).

2.3. F wave relevance analyses

As we established previously, F waves superimpose on the entire signal waveform to varying degrees [7], just as the rR values of our model do. To investigate whether the rR values within the F waves differ from chance, we compared the mean rR within all F waves of a recording with the mean rR over the entire signal or with the mean rR within all TQ segments. This was done for both F_{Full} and F_{TQ}. For statistical evaluation, we used a one-way analysis of variance (ANOVA) and Student's t-tests for post-hoc analyses, using Tukey-Kramer alpha-error correction. To evaluate the effect size, we calculated Cohen's *d*.

We also examined the distribution of rR values within F waves to gain further insight into the model's decision-making process. To do this, we created a template F wave by averaging all peak-aligned F waves and followed the same procedure for the rR values. Because CNNs are known to focus on edges, and because F waves are often asymmetric, i.e., have flanks of different steepness, we also created the template F wave only for right- or left-skewed F waves. Furthermore, for each template, we calculate the coefficients of variability of the rR across F waves.

3. Results

The rR, averaged across all F waves in a recording, was highest for a prominence of 41 and negative polarity (trough detection), reaching an rR of 0.278 a.u. in the median, when considering the full waveform. Under these assumptions, F waves were found in 277 ($58.1\,\%$) recordings, taking into account the full waveform, and in 225 ($47.2\,\%$) recordings, considering only the TQ segments.

The rR averaged over recordings is shown in Figure 1. The ANOVA revealed significant differences ($F = 49.3, p < .001$) between the rR averaged across different signal parts. Overall, the mean rR within the TQ segments is significantly increased compared to the mean rR within the full waveform by $+0.043$ a.u. ($p < .001$, $d = .939$). In both cases, F_{Full} and F_{TQ}, the mean rR within F waves is significantly increased compared to the mean rR within the full waveform ($+0.034$ a.u., $p < .001$, $d = .468$) or the full TQ segments ($+0.027$ a.u., $p < .001$, $d = .333$).

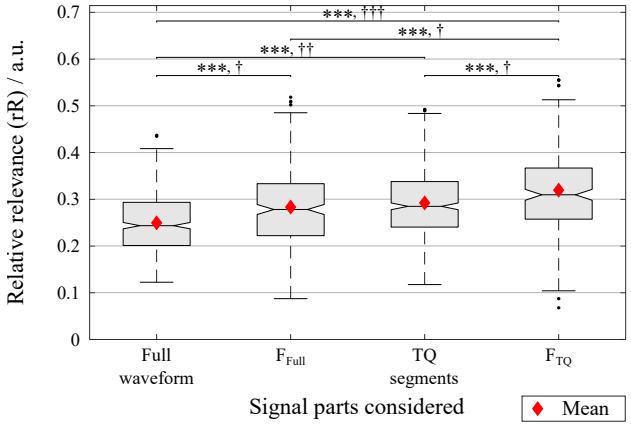


Figure 1: Mean relative relevance (rR) in full waveform and F waves extracted from the full waveform (F_{Full}) versus TQ segments only and F waves extracted from TQ segments only (F_{TQ}). Significance levels for group differences according to Student's *t*-tests (***: $p < .001$) and effect sizes according to Cohen's d (†: $0.2 \leq |d| < 0.5$, ††: $0.5 \leq |d| < 0.8$, †††: $|d| \geq 0.8$) were highlighted.

Figure 2 shows the model explanations as saliency maps. The color of each sample represents its relevance to the AF classification. Upon examining the signal example in Figure 2a, it is evident that QRS complexes are of irregular rR, while strongly pronounced and easily recognizable F waves are consistently considered particularly relevant (e.g., at the beginning of the signal and between QRS complexes 2 and 3, 4 and 5, and 6 and 7). Figures 2b–2d display the distributions of the rR over the averaged F waves, including all F waves, only left-skewed F waves, and only right-skewed F waves. The rR is at its highest on the descending flank of the F waves and decreases rapidly on the ascending flank. In the case of left-skewed F waves, the rR is slightly shifted to the right, causing the maximum to extend to the trough. Due to the shift of the maximum rR to the right, the left-skewed F waves provide increased relevance values on the ascending flank compared to the right-skewed F waves. The variability of the rR is lowest on the flanks and increases in the valley and towards the outside. The minima of the rR variability for left-skewed F waves are on the ascending flank and for right-skewed F waves on the descending flank.

4. Discussion and Conclusion

For the first time, we systematically investigated the correspondence between explanations for a DL model for AF detection and beat-independent morphological features, specifically F waves. To achieve this, we selected the short-term model from xECGArch. This model is designed to learn morphological features and has demonstrated a high accuracy of $94.0\,\%$ in detecting AF on unknown test data. In addition, we utilized DTD, an xAI method that has been validated for reliability in this scenario, to extract highly trustworthy model explanations.

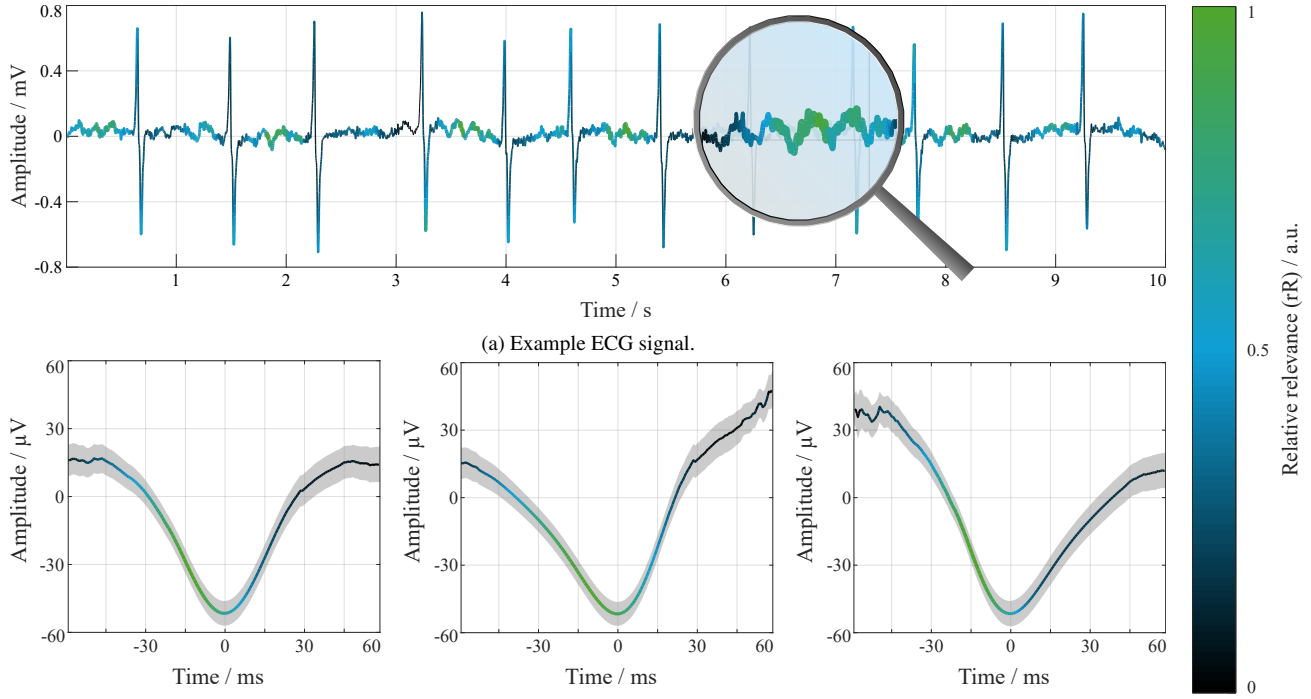
Previous systematic relevance analyses [5,6,11] suggest that the area between the T wave end of a beat and the beginning of the QRS complex of the following beat is of increased relevance. However, previous approaches cannot explain these effects, as they average relevance values over the beats, smoothing out effects that irregularly occur over the entire waveform, similar to F waves. Our findings confirm that the TQ segments have an increased relevance for AF detection compared to the mean relevance across the entire waveform. This relevance further increases when only the F waves within the TQ segments are considered. This suggests that the increased relevance in the TQ segments results from F waves, which are important for classification and stick out in this area. Noise, which can be observed best in the TQ segments, may also be a contributing factor if it is overrepresented in a group by chance and is, therefore, utilized by the model to distinguish between AF and non-AF. However, the diverse data set and filtering $\hat{s}_{aa}$ between 5–10 Hz, and setting the maximum F wave width to $w^F_{max} = 0.25$ s can largely rule out this possibility.

The F waves have a relatively low rR of $\sim$0.3 a.u. on average. This results from the uneven distribution of rR over the waves, with the left flank being particularly relevant (see Figure 2). When averaging the rR over the entire F wave, the significantly increased rR of one flank is moderated. It is well-known that CNNs primarily learn flanks for classification. Figures 2c and 2d demonstrate that the used CNN identifies the flank pointing to the subsequent QRS complex as the relevant flank, rather than the steeper flank that is easier to detect. This may be due to the importance of the distance to the following QRS complex in distinguishing F waves from regular P waves. This assumption is supported by previous findings [5,6], indicating that the area immediately before the QRS complex, where the P wave or PQ segment is located, has relatively low relevance values for classification as AF, while the left flank of the QRS complex is particularly relevant.

For the first time, we have been able to attribute the explanations for DL-based AF detection to F waves as irregularly distributed morphological features across the signal through systematic investigations. This contributes to an improved interpretability of DL methods' explanations. The elucidation of the black box is an important step towards the clinical applicability of DL-based diagnostic support systems.

Acknowledgments

This study was supported by grants from the European Union's Horizon 2020 research and innovation program (TIMELY, No. 101017424).



(a) Example ECG signal.

(b) F wave template (all).

(c) F wave template (left-skewed).

(d) F wave template (right-skewed).

Figure 2: Saliency maps, representing the sample-wise relative relevance (rR) to the model for detecting atrial fibrillation, color-coded from black to green (not relevant to highly relevant). Figure (a) displays an example ECG signal while figures (b)–(d) show the mean rR across all F waves versus left or right-skewed F waves only. The mean rR was scaled to a range of 0 to 1 o enhance contrast. The gray areas in (b)–(c) represent the coefficients of variability of the rR across the F waves.

References

[1] Chugh SS, et al. Worldwide Epidemiology of Atrial Fibrillation. Circulation 2014;129(8):837–847.

[2] Padfield GJ, et al. Progression of paroxysmal to persistent atrial fibrillation: 10-year follow-up in the Canadian Registry of Atrial Fibrillation. Heart Rhythm 2017;14(6):801–807.

[3] Stracina T, Ronzhina M, Redina R, Novakova M. Golden Standard or Obsolete Method? Review of ECG Applications in Clinical and Experimental Context. Front Physiol 2022;13:867033.

[4] Holzinger A, Langs G, Denk H, Zatloukal K, Müller H. Causability and Explainability of Artificial Intelligence in Medicine. WIRES DATA MIN KNOWL 2019;9(4):e1312.

[5] Goettling M, Hammer A, Malberg H, Schmidt M. xECGArch: a trustworthy deep learning architecture for interpretable ECG analysis considering short-term and long-term features. Sci Rep 2024;14:13122.

[6] Ivaturi P, Gadaleta M, Pandey AC, Pazzani M, Steinhubl SR, Quer G. A Comprehensive Explanation Framework for Biomedical Time Series Classification. IEEE J Biomed Health Inform 2021;25(7):2398–2408.

[7] Hammer A, Malberg H, Schmidt M. Towards the Prediction of Atrial Fibrillation Using Interpretable ECG Features. In Comput. Cardiol., volume 49. Tampere, FIN, 2022; 1–4.

[8] Schmidt M, Baumert M, Porta A, Malberg H, Zaunseder S. Two-Dimensional Warping for One-Dimensional Signals—Conceptual Framework and Application to ECG Processing. IEEE Trans Signal Process 2014;62(21):5577–5588.

[9] Schmidt M, Baumert M, Malberg H, Zaunseder S. Iterative two-dimensional signal warping—Towards a generalized approach for adaption of one-dimensional signals. Biomed Signal Process Control 2018;43:311–319.

[10] Hammer A, Scherpf M, Ernst H, Weiß J, Schwensow D, Schmidt M. Automatic Classification of Full- And Reduced-Lead Electrocardiograms Using Morphological Feature Extraction. In Comput. Cardiol., volume 48. Brno, CZ, 2021; 1–4.

[11] Hammer A, Goettling M, Malberg H, Linke A, Richter S, Mangner N, Schmidt M. Fusion of automatically learned rhythm and morphology features matches diagnostic criteria and enhances AI explainability. npj Artif Intell 2024; under review.

Address for correspondence:

Alexander Hammer
Institute of Biomedical Engineering, TU Dresden
Fetscherstr. 29, 01307 Dresden, Germany
alexander.hammer@tu-dresden.de