Characterization of Atrial Fibrillation Episode Patterns: A Comparative Study

Monika Butkuvienė, Andrius Petrėnas, Alba Martín-Yebra, Vaidotas Marozas, and Leif Sörnmo, *Fellow, IEEE*

***Abstract*— *Objective:* The episode patterns of paroxysmal atrial fibrillation (AF) may carry important information on disease progression and complication risk. However, existing studies offer very little insight into to what extent a quantitative characterization of AF patterns can be trusted given the errors in AF detection and various types of shutdown, i.e., poor signal quality and non-wear. This study explores the performance of AF pattern characterizing parameters in the presence of such errors.**

***Methods:* To evaluate the performance of the parameters AF aggregation and AF density, both previously proposed to characterize AF patterns, the two measures mean normalized difference and the intraclass correlation coefficient are used to describe agreement and reliability, respectively. The parameters are studied on two PhysioNet databases with annotated AF episodes, also accounting for shutdowns due to poor signal quality.**

***Results:* The agreement is similar for both parameters when computed for detector-based and annotated patterns, which is 0.80 for AF aggregation and 0.85 for AF density. On the other hand, the reliability differs substantially, with 0.96 for AF aggregation but only 0.29 for AF density. This finding suggests that AF aggregation is considerably less sensitive to detection errors. The results from comparing three strategies to handle shutdowns vary considerably, with the strategy that disregards the shutdown from the annotated pattern showing the best agreement and reliability.**

***Conclusions:* Due to its better robustness to detection errors, AF aggregation should be preferred. To further improve performance, future research should put more emphasis on AF pattern characterization.**

***Index Terms*— Paroxysmal atrial fibrillation, detection, agreement, reliability, performance evaluation.**

I. Introduction

The episode patterns of paroxysmal atrial fibrillation (AF) are largely unexplored despite that they vary considerably with respect to occurrence, duration, and clustering. At the same time, certain types of pattern indeed may carry valuable information on disease progression and risk of complications. For instance, reduction of flow velocity in the left atrial appendage is associated with increased risk of thrombus formation [1]–[3]. Furthermore, the flow velocity decreases as AF progresses from shorter to longer episodes [4]. Understanding how the characteristics of episode patterns relate to thrombus formation may improve AF management beyond the current, rather simplistic criterion of an episode lasting at least 30 s to be diagnostic of clinical AF [5].

The problem of how to temporally characterize AF episodes received certain attention two decades ago. Most of the studies analyzed inter-episode or inter-detection intervals in terms of statistical distributions under the assumption that episodes are statistically independent [6]–[9]. While inter-episode dependence was statistically established for some patients in [6], no episode pattern characterization was performed in those patients.

Several years later, two related parameters were proposed to characterize AF episode patterns, referred to as AF density [10] and AF aggregation [11]. Both parameters are based on the same idea, namely to quantify the deviation between the observed episode pattern and a template pattern consisting of evenly spread episodes once the two patterns have been subject to a parameter-specific transformation. For AF density, the pattern is transformed into times needed to develop different proportions of the total AF burden, whereas, for AF aggregation, the pattern is transformed into a cumulative distribution of the unnormalized AF burden. By means of examples it was shown that the parameters can distinguish patterns with episodes evenly spread across the observation interval, i.e., the total monitoring period, from patterns with episodes aggregated to a small part of the observation interval. The clinical significance of AF density and AF aggregation remains largely to be demonstrated.

Recently, history-dependent point process modeling was proposed to characterize the alternating transition times from non-AF to AF, and vice versa, using a novel bivariate Hawkes self-exciting model [12]. A transition increases the likelihood of observing additional transitions in the near future, thus allowing clustered episode patterns to be modeled. The maximum likelihood estimator was derived and employed to find the model parameters from observed data. Using long-term ECG databases, the goodness-of-fit analysis showed that the model fit the AF patterns in most recordings. In a subsequent study, a subset of the model parameters, describing episode intensity and degree of episode clustering, were shown to

Manuscript received xx xx, 2023. This work was supported by the European Regional Development Fund (01.2.2-LMT-K-718-03-0027) under grant agreement with the Research Council of Lithuania (LMTLT).

M. Butkuvienė and A. Petrėnas are with the Biomedical Engineering Institute, Kaunas University of Technology, Kaunas, Lithuania. (e-mail: monika.butkuviene@ktu.lt)

A. Martín-Yebra is with BSICoS Group, Aragon Institute of Engineering Research (I3A), University of Zaragoza, Zaragoza, Spain.

V. Marozas is with the Biomedical Engineering Institute and Department of Electronics Engineering, Kaunas University of Technology, Kaunas, Lithuania.

L. Sörnmo is with the Department of Biomedical Engineering, Lund University, Lund, Sweden.

play a role in pre-ablation risk assessment, demonstrating the clinical significance of AF episode pattern characterization [13]; in that study, AF density, in combination with AF burden, was also studied but did not play a similar role. To obtain parameter estimates with acceptable accuracy, at least 10 episodes were deemed necessary [12]. With regard to AF density and AF aggregation, both developed within a non-statistical framework, no such constraint needs to be imposed.

So far, there has been limited insight on how much a quantitative characterization can be trusted given the presence of shutdowns for a period of time due to poor signal quality and non-wear which leads to reduced AF detection performance. Accordingly, for the first time in the literature, this study investigates the feasibility to characterize patterns in the presence of errors manifested by falsely detected, missed, merged, and split episodes.

This paper is organized as follows. Section II describes the two above-mentioned parameters whose properties are investigated using the databases described in Sec. III. Section IV describes the performance measures agreement and reliability employed to quantify the influence of errors in the detector-based episode pattern relative to the annotated episode pattern. The results are presented in Sec. V, followed by a discussion on issues related to pattern characterization, and final conclusions.

II. Methods

AF detection is required before AF aggregation and AF density can be computed. Here, detection is accomplished using a rhythm-based detector [14], relying on the assumption that AF episodes are manifested by irregular RR intervals which often are accompanied by an increase in heart rate. The detector, designed with special reference to detect brief AF episodes, includes blocks for ectopic beat filtering, bigeminy suppression, characterization of RR interval irregularity, and signal fusion. RR intervals are determined using a wavelet-based QRS detector [15].

The assessment of ECG signal quality can be performed using either a separate algorithm (adopted here and described in Sec. II-B) or an algorithm built-in into the AF detector. Either way, a strategy for handling the presence of shutdowns, including poor-quality segments, has to be invoked before the computation of the AF pattern characterizing parameters (Fig. 1).

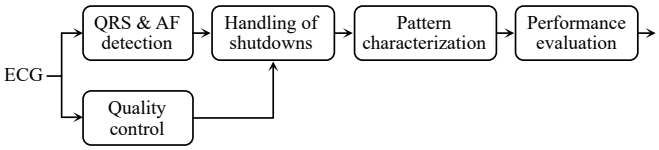

Fig. 1: Block diagram of the processing steps performed to evaluate AF pattern characterizing parameters.

A. AF pattern characterizing parameters

The unnormalized AF burden at time n in a window of length l is given by,

$$\tilde{b}(n,l) = \sum_{k=n}^{\min(n+l-1,N_{\mathrm{RR}})} o_{\mathrm{AF}}(k), \quad n,l = 1,\ldots,N_{\mathrm{RR}}, \tag{1}$$

where the binary function $o_{\mathrm{AF}}(n)$ indicates whether the n:th RR interval $r(n)$ is in AF or not,

$$o_{\mathrm{AF}}(n) = \begin{cases} 1, & r(n) \in \mathrm{AF}, \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

and N_{RR} is the number of RR intervals in the observation interval. Thus, the number of beats in AF is given by $N_{\mathrm{AF}} = \sum_{n=1}^{N_{\mathrm{RR}}} o_{\mathrm{AF}}(n)$, where $N_{\mathrm{AF}} < N_{\mathrm{RR}}$ since at least one beat must be in non-AF. The window length becomes increasingly shorter at the end of the observation interval $[1, N_{\mathrm{RR}}]$ to ensure that the window does not extend beyond N_{RR}.

Basic information on the cumulative distribution of the unnormalized burden can be obtained by determining the maximum of $\tilde{b}(n,l)$ with respect to n,

$$b(l) = \max_{n=1,\ldots,N_{\mathrm{RR}}} \tilde{b}(n,l). \tag{3}$$

The parameter *AF aggregation* $\mathcal{A}$ [11] is defined by the sum of the absolute deviations between $b(l)$ and a template, linear cumulative distribution associated with AF episodes evenly spread out over the observation interval,

$$\mathcal{A} = \frac{2}{N_{\mathrm{RR}} N_{\mathrm{AF}}} \sum_{l=1}^{N_{\mathrm{RR}}} \left| b(l) - l \frac{N_{\mathrm{AF}}}{N_{\mathrm{RR}}} \right|. \tag{4}$$

The sum is normalized by $2/(N_{\mathrm{RR}} N_{\mathrm{AF}})$ to ensure that $\mathcal{A}$ is in the range from 0 to 1. A value close to 1 indicates an accumulation of episodes, characteristic of a pattern with one or a few short highly aggregated episodes, while a value close to 0 indicates a pattern composed of evenly spread episodes.

The minimum contiguous time l_p required to develop a certain proportion p of the total unnormalized burden N_{AF} is determined by finding the times when $b(l)$ changes. The time n_1 corresponds to $p = 1$, n_2 to $p = 2$, and so on until $p = N_{\mathrm{AF}}$, where $n_1 < \cdots < n_{N_{\mathrm{AF}}}$.

The parameter *AF density* $\mathcal{D}$ [10] is defined similarly to $\mathcal{A}$ but with the difference that $b(l)$ is replaced by l_p, and, accordingly, the sum of absolute deviations ranges from 1 to N_{AF},

$$\mathcal{D} = \frac{2N_{\mathrm{RR}}}{(N_{\mathrm{AF}}+1)(N_{\mathrm{RR}}-N_{\mathrm{AF}})} \sum_{p=1}^{N_{\mathrm{AF}}} \left| \frac{l_p}{N_{\mathrm{RR}}} - \frac{p}{N_{\mathrm{AF}}} \right|. \tag{5}$$

To ensure that $\mathcal{D}$ is in the range from 0 to 1, a normalization factor other than the one in (4) is needed. A value of $\mathcal{D}$ close to 1 indicates an aggregation of burden, characteristic of patterns with a single episode irrespective of whether it is short or long, while a value close to 0 indicates a pattern for which the burden is evenly spread out over the observation interval.

Thus, while $\mathcal{A}$ reflects how the observed, cumulative distribution of the unnormalized burden deviates from a template,

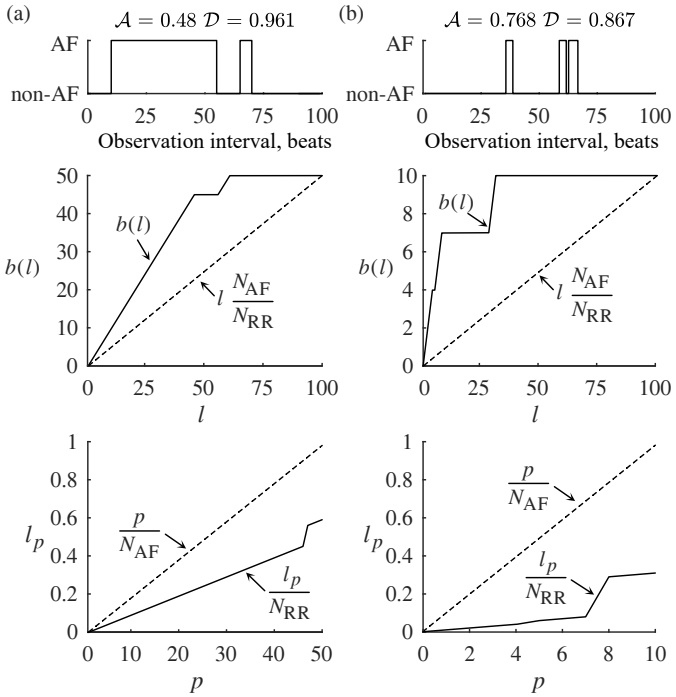

Fig. 2: Illustration of AF patterns (top) and the two functions which together define $\mathcal{A}$ (middle) and $\mathcal{D}$ (bottom). The two patterns are related to (a) a large difference between $\mathcal{A}$ and $\mathcal{D}$ and (b) a small difference.

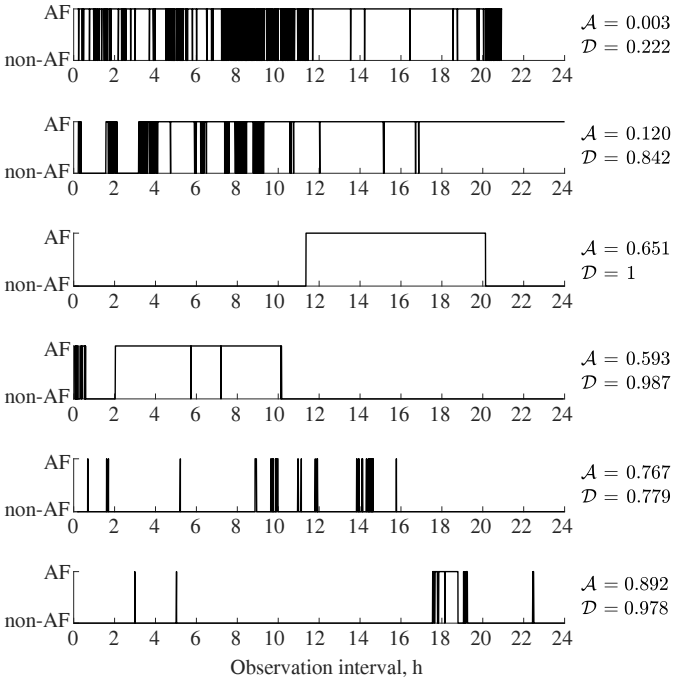

Fig. 3: Examples of AF episode patterns taken from recordings 204, 115, 56, 111, 120, and 03 (top to bottom), which all are part of the Long-Term Atrial Fibrillation Database (LTAFDB) [16].

linear cumulative distribution of the unnormalized burden, $\mathcal{D}$ reflects how the observed, minimum contiguous time required to develop a proportion p of the total burden deviates from a template, linear progression of the minimum contiguous time.

The two functions inside the absolute value of the respective parameter definitions in (4) and (5) are illustrated in Fig. 2.

Figure 3 illustrates essential differences between $\mathcal{D}$ and $\mathcal{A}$. In particular, for an observation interval with one episode, $\mathcal{D}$ is always equal to 1 irrespective of whether the episode lasts the entire observation interval or just a fraction of it, whereas $\mathcal{A}$ depends on the relation between episode duration and observation interval duration, and increases as episode duration decreases.

B. Handling of poor-quality segments

The occurrence of poor-quality segments implies that the AF pattern will contain gaps due to analysis shutdown, which can be either ignored or filled by conjecturing the rhythm(s). Before computing $\mathcal{A}$ and $\mathcal{D}$, the following three strategies to handle poor-quality segments are investigated:

1) Poor-quality segments are ignored in both the annotated and the detector-based patterns.
2) Poor-quality segments are set to non-AF in the detector-based pattern. The annotated pattern remains unchanged.
3) Poor-quality segments are set to a context-dependent rhythm in the detector-based pattern. If non-AF occurs both before and after a poor-quality segment, the whole segment is set to non-AF, and vice versa. On the other hand, if non-AF occurs before and AF occurs after, the first half of the segment is set to non-AF and the second half to AF, and vice versa. The annotated pattern remains unchanged.

Figure 4 illustrates the three different strategies to handle AF patterns with shutdown.

ECG signal quality is assessed by analyzing non-overlapping 10-s segments using the following criteria [17]: the heart rate is within the range 40–180 bpm, none of the RR intervals exceed 3 s, and the ratio between the longest and the shortest RR interval is less than 2.2. If any of the criteria is not met, the signal quality is considered poor. If all criteria are met, QRS waveform template matching is performed to further assess the signal quality; a cross-correlation coefficient below 0.66 signifies poor quality.

III. Materials

The pattern characterizing parameters are investigated using the publicly available MIT–BIH Atrial Fibrillation Database (AFDB) and Long-Term Atrial Fibrillation Database (LTAFDB) [16]. The former database consists of 25 two-lead ambulatory ECG recordings, each lasting 10-h, from patients with paroxysmal or persistent AF. In total, AFDB consists of 297 manually annotated AF episodes, accounting for 38% of the database. The latter database consists of 84 two-lead ambulatory ECG recordings, each lasting 24–25-h, from patients with paroxysmal or persistent AF. In total, LTAFDB consists of 7,317 manually annotated AF episodes, accounting for 50% of the database. Recordings without AF episodes (1 in LTAFDB) or with AF only (2 in AFDB and 12 in

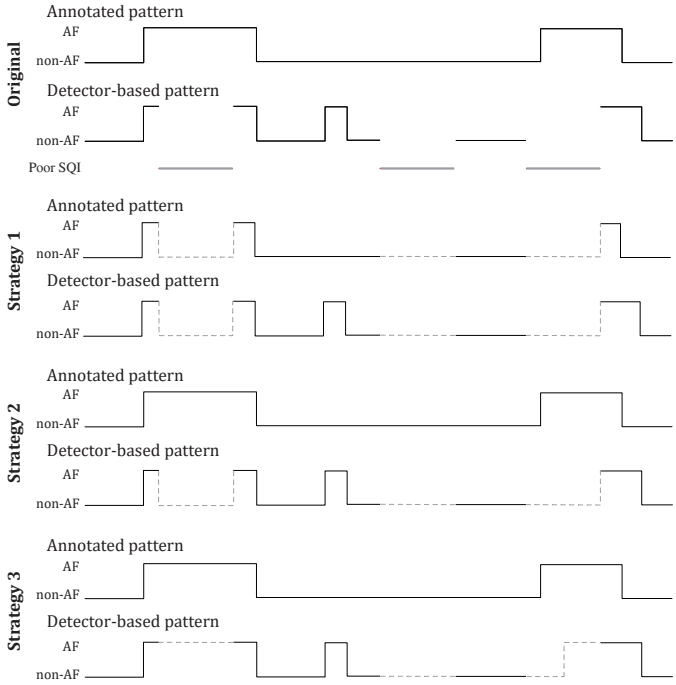

Fig. 4: Illustration of the three different strategies to handle shutdowns when comparing a detector-based AF pattern to the corresponding annotated pattern. The gray, thicker lines indicate segments with poor quality. The dashed line results from application of the indicated strategy.

LTAFDB) were excluded from the analysis, thus resulting in 94 recordings with AF.

Figure 5 presents histograms of episode duration for AFDB and LTAFDB. For AFDB, the median episode duration is 168 beats, and 35% of all episodes are shorter than 100 beats. In contrast, LTAFDB has a much shorter median episode duration of 18 beats, and 80% of all episodes are shorter than 100 beats. Thus, brief episodes are far more common in LTAFDB than in AFDB, and, indeed, the majority of episodes in LTAFDB are shorter than 30 beats.

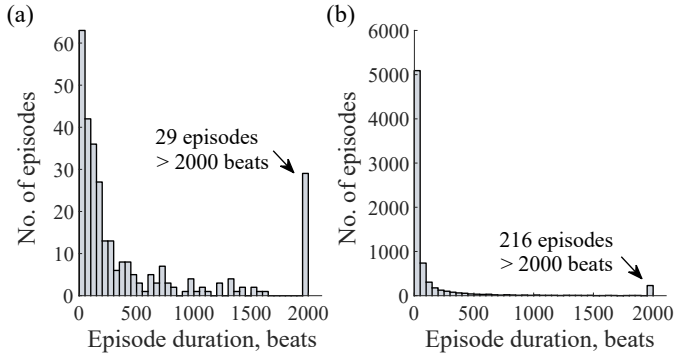

Fig. 5: Histograms of AF episode duration for (a) MIT–BIH Atrial Fibrillation Database (AFDB) and (b) LTAFDB.

IV. Performance Evaluation

A. Agreement and reliability

When dealing with a detection problem, e.g., detection of AF episodes, performance is typically evaluated by comparing on a beat-to-beat basis the detector output to the annotations, resulting in counts of true/false positives and true/false negatives which in turn make it possible to compute measures like sensitivity, specificity, and accuracy [18]. Since episode pattern characterization does not represent a detection problem, other performance measures need to be employed. In the present study, the parameter values characterizing the detector-based and the annotated patterns are compared to each other with regard to *agreement* and *reliability*, quantified by the mean normalized difference and the intraclass correlation coefficient, respectively, as performance measures.

For each of the M recordings in AFDB and LTAFDB, the pattern characterizing parameter x, $x \in \{\mathcal{A}, \mathcal{D}\}$, are computed for the detector-based and the annotated patterns, yielding $x(1), ..., x(M)$ and $x_r(1), ..., x_r(M)$, respectively. The mean normalized difference K_x is defined by,

$$K_x = 1 - \frac{1}{M} \sum_{i=1}^{M} \frac{|x(i) - x_r(i)|}{x(i) + x_r(i)}, \tag{6}$$

where $0 \leq K_x \leq 1$ with 1 indicating perfect agreement.

The intraclass correlation coefficient I_x is defined by [19], [20],

$$I_x = \frac{\sigma_r^2}{\sigma_r^2 + \sigma_e^2}, \tag{7}$$

where σ_r^2 is the variance of $x_r(i)$, σ_e^2 is the variance of the difference between $x(i)$ and $x_r(i)$, and $0 \leq I_x \leq 1$ with 1 indicating perfect reliability.

The complementarity of K_x and I_x is illustrated in Fig. 6 for $M = 10$. In Fig. 6(a), $x(i)$ is identical to $x_r(i)$ for eight AF patterns, whereas the error in $x(i)$ is 50% for the remaining two patterns, resulting in high agreement (K_x = 0.93) but low reliability (I_x = 0.43). In Fig. 6(b), $x(i)$ comes with a 50% error for all 10 patterns, resulting in low agreement (K_x = 0.67) but high reliability (I_x = 0.94).

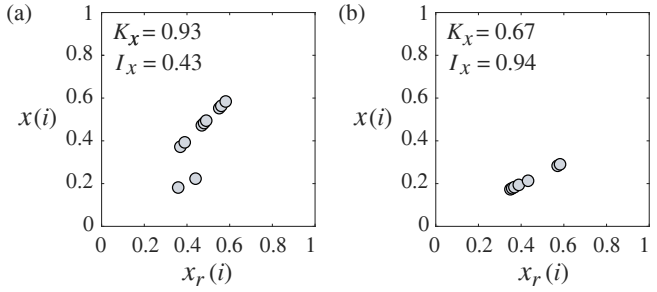

Fig. 6: Two scenarios illustrating the complementary information conveyed by the measures agreement K_x and reliability I_x: $x(i)$ comes with a 50% error for (a) two patterns and (b) for all 10 patterns.

B. Shutdown time

Based on the American National Standard for testing and reporting performance results of cardiac rhythm algorithms [21], a shutdown is defined as the period of time when an AF detector is disabled. Based on this definition, segments excluded due to poor signal quality or non-wear of the device are referred to as shutdowns.

Figure 7 presents histograms of shutdown time due to poor signal quality in AFDB and LTAFDB. In total, the shutdown time takes 8% of the total time of the databases. For AFDB and LTAFDB, the median shutdown time is 16 and 17 beats, respectively.

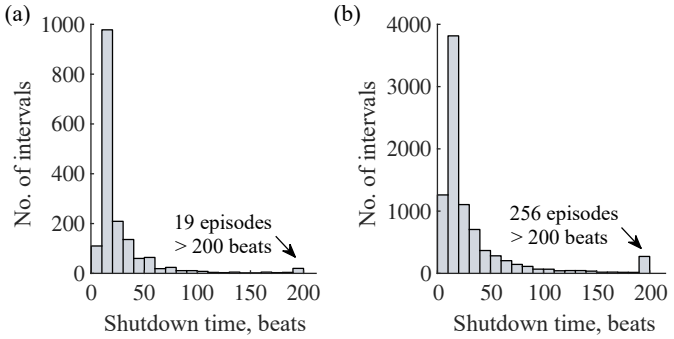

Fig. 7: Histograms of shutdown times for (a) AFDB and (b) LTAFDB.

V. Results

A. Association between parameters computed from detector-based and annotated patterns

Using AFDB and LTAFDB together, Fig. 8 presents for both $\mathcal{A}$ and $\mathcal{D}$ the association between values obtained from detector-based and annotated patterns. The parameter $\mathcal{D}$ is considerably more sensitive to errors in episode patterns than $\mathcal{A}$, reflected by the results that $\mathcal{A}$ and $\mathcal{A}_r$ show a strong correlation ($r = 0.90$), while the correlation between $\mathcal{D}$ and $\mathcal{D}_r$ is low ($r = 0.26$).

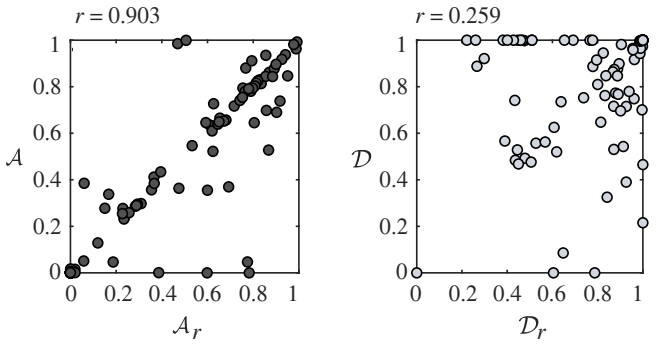

Fig. 8: Scatter plots of the pattern characterizing parameters $\mathcal{A}$ and $\mathcal{D}$ when obtained from annotated patterns, then denoted $\mathcal{A}_r$ and $\mathcal{D}_r$, and detector-based patterns, then denoted $\mathcal{A}$ and $\mathcal{D}$. The Pearson correlation coefficient r is given in each plot.

As illustrated by the examples in Figs. 9(a) and (b), both $\mathcal{A}$ and $\mathcal{D}$ are similarly sensitive to missed and falsely detected episodes provided that the episode duration is of the same order as that of the annotated episodes. However, as illustrated in Figs. 9(c) and (d), $\mathcal{D}$ is much more influenced by splitting and merging of AF episodes than $\mathcal{A}$.

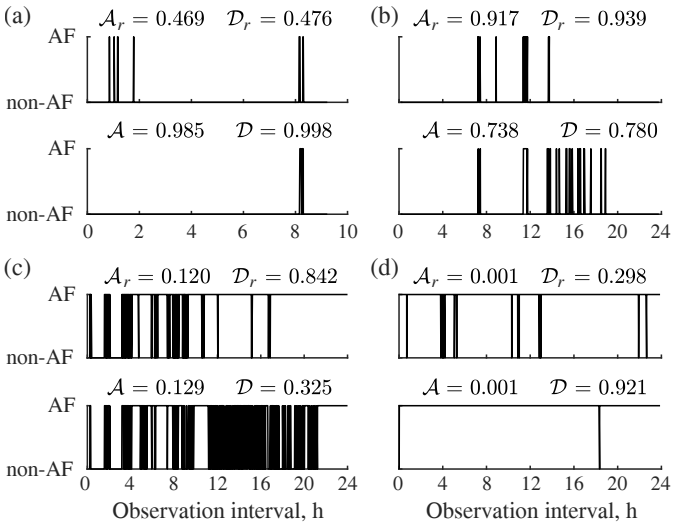

Fig. 9: Examples of how differences between the annotated AF pattern (top panel) and the detector-based pattern (bottom panel) influence $\mathcal{A}$ and $\mathcal{D}$ in situations when AF episodes are (a) missed, (b) falsely detected, (c) splitted, and (d) merged.

B. Association between parameters using patterns defined on a beat or a time basis

In the present study, following the definitions used in [10] and [11], AF patterns are defined on a beat basis, i.e., the index n in (2) refers to the n:th RR interval. Alternatively, patterns can be defined on a time basis (in seconds) with the advantage of accounting for changes in heart rate. While patterns may differ slightly as illustrated in Fig. 10(a), both parameters assume similar values regardless of whether time or beat is used as basis, see Fig. 10(b) ($r = 0.996$ and $r = 0.997$ for $\mathcal{A}_r$ and $\mathcal{D}_r$, respectively).

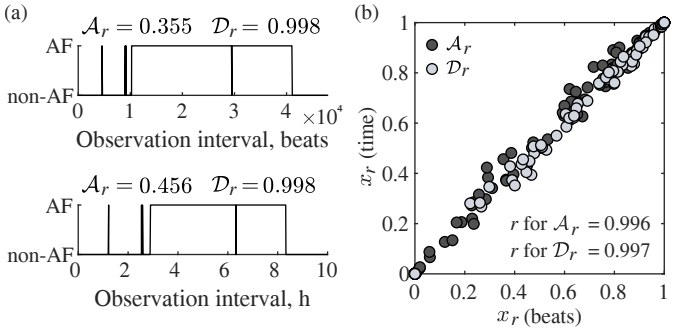

Fig. 10: (a) Examples of AF pattern defined on a beat basis (top panel) and time basis (bottom panel), and (b) scatter plot of the pattern characterizing parameters $\mathcal{A}_r$ and $\mathcal{D}_r$.

C. Agreement and reliability

Although a high positive correlation is necessary to establish agreement, it is not a sufficient condition. To complement the

correlation results in Fig. 8, Fig. 11 presents the performance measures agreement, i.e., $K_{\mathcal{A}}$ and $K_{\mathcal{D}}$, and reliability, i.e., $I_{\mathcal{A}}$ and $I_{\mathcal{D}}$, for the three comparison strategies to handle shutdowns. The results vary considerably between strategies, with strategy #1, i.e., ignoring the shutdown from the annotated pattern, as the best with respect to all four performance measures; however, strategy #1 tends to favor performance since shutdowns are excluded from both annotated and detector-based patterns, cf. the discussion. Concerning strategies #2 and #3, $K_{\mathcal{D}}$ is about 1.3 times larger than $K_{\mathcal{A}}$, however, $I_{\mathcal{D}}$ is 2–4 times lower than $I_{\mathcal{A}}$.

In addition, Fig. 11 illustrates the impact of using detected RR intervals instead of annotated ones as input to the AF detector. The results show that $K_{\mathcal{A}}$ decreases from 0.80 to 0.74 and $K_{\mathcal{D}}$ from 0.85 to 0.77, whereas $I_{\mathcal{A}}$ decreases from 0.96 to 0.92 and $I_{\mathcal{D}}$ from 0.29 to 0.15. Thus, the agreement is similar for $\mathcal{A}$ and $\mathcal{D}$, whereas the reliability differs substantially in favor of $\mathcal{A}$.

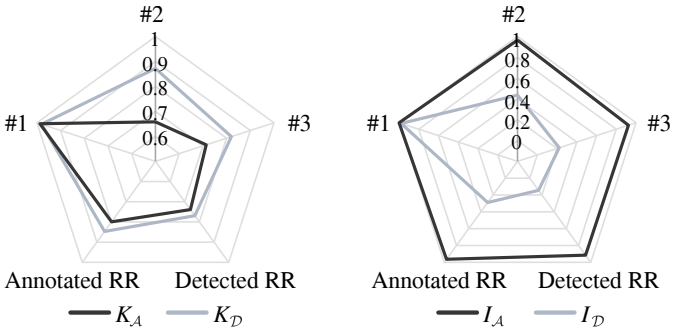

Fig. 11: The performance measures agreement K (left) and reliability I (right) for $\mathcal{A}$ and $\mathcal{D}$ using the three different comparison strategies when RR intervals obtained from QRS detection are used as input to the AF detector. The two values at the bottom of the diagrams are computed when RR intervals from either annotation or QRS detection are used as input.

D. AF density and AF aggregation in relation to AF burden

Figure 12 shows the association between $\mathcal{D}_r$ and $\mathcal{A}_r$ for different ranges of AF burden $\mathcal{B}$. For the databases taken together, $\mathcal{D}_r$ and $\mathcal{A}_r$ show a moderately strong correlation ($r = 0.63$), however, the correlation becomes very strong ($r = 0.97$) when $\mathcal{B}_r < 0.2$. Since $\mathcal{A}$ characterizes the aggregation of AF episodes relative to the observation interval, it is nearly 0 for AF patterns with a very large AF burden ($\mathcal{B}_r > 0.8$). In contrast, $\mathcal{D}_r$ varies widely, explained by the fact that $\mathcal{D}$ does not account for the observation interval. No association is observed between $\mathcal{D}_r$ and the number of AF episodes, varying from 1 to 1044.

Figure 13(a) shows that AF patterns with either low or high AF burden dominate AFDB and LTAFDB. It is obvious from Fig. 13(b) that $\mathcal{A}_r$ is strongly, negatively correlated with AF burden $\mathcal{B}_r$ for $\mathcal{B}_r > 0.5$, whereas no such association exists between $\mathcal{D}_r$ and $\mathcal{B}_r$. This result is expected since the time periods before the first and after the last AF episodes are ignored when computing $\mathcal{D}$.

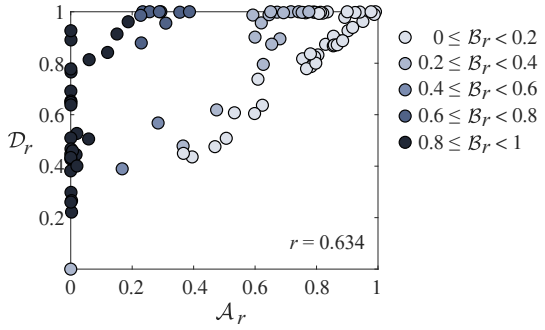

Fig. 12: Association of $\mathcal{D}_r$ and $\mathcal{A}_r$. Shading of circles represent different $\mathcal{B}_r$.

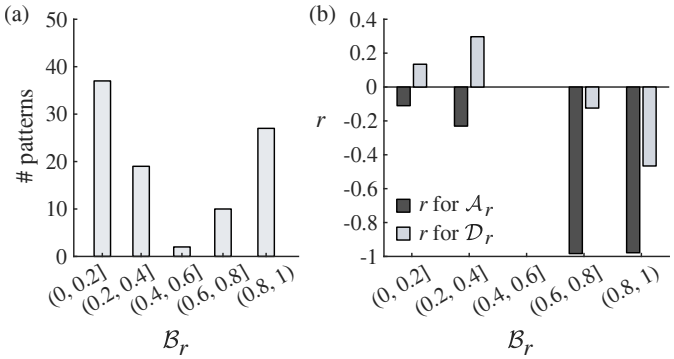

Fig. 13: (a) The number of AF patterns for a particular range of $\mathcal{B}_r$, and (b) correlation r of either $\mathcal{A}_r$ or $\mathcal{D}_r$ with $\mathcal{B}_r$. Note that r is not computed for the interval (0.4 0.6] as it contains only two AF patterns.

VI. Discussion

Performance evaluation of AF pattern characterizing parameters is considerably more challenging than performance evaluation of an AF detector which represents the problem invariably addressed in the literature. Indeed, even a few missed or falsely detected episodes can influence the parameters. Rather than just focusing on achieving increasingly better AF detection performance, as is currently the case, future research should broaden the horizon and put more emphasis on developing methods for AF pattern characterization.

Recently, a debate has emerged over AF progression and risk of complication, suggesting that risk is better predicted by parameters accounting for AF characteristics. For instance, the 4S-AF structured pathophysiology-based characterization scheme, already included in clinical guidelines on AF diagnosis and management [5], considers severity of AF burden next to stroke risk, symptom severity, and atrial substrate severity [22], [23]. Growing evidence points towards a link between AF burden and increased risk of thrombus formation [24], [25]; however, the threshold of AF burden is unclear and varies considerably among studies [23], [26], [27].

Out of the available approaches to pattern characterization, episode clustering is only accounted for by the Hawkes model and then provided by the model parameter defining the exponential decay of the point process intensity function [12]. As noted in Introduction, a limitation with this approach is that at least 10 episodes need to be observed to achieve

acceptable statistical accuracy. Since the observation interval in AFDB and LTAFDB is relatively short, i.e., 10 h and 24–25 h, respectively, the number of episodes was less than 10 in as many as 58% of the recordings. Therefore, the Hawkes parameters were not considered in the present study.

Using the complementary measures agreement and reliability, AF aggregation was found to be considerably more reliable than AF density. The interpretation of agreement is straightforward, i.e., a high agreement means that the parameter value characterizing the detector-based pattern is close to the value obtained for the annotated one. On the other hand, the reliability depends on data heterogeneity, and, therefore, it will differ among databases even for the same error variance. For this reason, a comparison of reliability obtained for different databases should be interpreted with caution.

The majority of AF detectors have been developed with the aim of achieving the best sensitivity or specificity, sometimes at the expense of ignoring methodological errors [28], [29]. However, addressing specific challenges, such as detecting brief AF episodes [14], [30], necessitate an emphasis on factors which can cost performance. Similarly, judging pattern characterizing parameters solely on the basis of agreement and reliability may unjustly diminish the parameters designed for a specific task. Although AF burden exhibits the highest level of agreement and reliability, it merely represents the relative time spent in AF. On the other hand, parameters characterizing the temporal distribution of AF episodes or the degree of clustering may be influenced by even a single false positive, resulting in reduced agreement and reliability. Thus, the performance of pattern characterizing parameters should be evaluated in the context of their intended task.

AF detectors are usually evaluated using publicly available databases such as AFDB, LTAFDB, and the MIT–BIH Arrhythmia database, where the times of the QRS complexes have been annotated. However, errors in QRS detection should not be overlooked since they also reduce AF detection performance [29]. Our study showed that the use of detected RR intervals instead of annotated ones, reduced the agreement and the reliability with up to 10%. Given that observation interval may last for weeks, QRS annotation becomes an overwhelming task to accomplish. Therefore, it is preferable to report detector-based results to avoid reporting over-optimistic performance.

Our previous study showed that AF patterns are influenced by the type of AF detector [31]: rhythm-based [14] and rhythm- and morphology-based [30] detectors, both types developed to detect brief episodes, tended to split longer episodes, whereas a segment-based deep learning detector did the opposite, namely to merge short episodes [31]. As shown in Fig. 9, AF aggregation is robust to episode splitting and merging, while AF density is influenced to a large extent. The latter parameter decreases for a split pattern and increases for a pattern with merged episodes. In connection with this observation, it should be noted that AF density probably assumed lower values due to the use of a rhythm-based detector. Therefore, the detector's propensity to influence the pattern should be considered before evaluating pattern characterizing parameters.

Today, implantable devices and modern wearables, such as ECG patches or smartwatches, are the only alternatives ensuring convenient long-term monitoring [32]–[34]. Obviously, poor signal quality or termination of monitoring will result in an intermittent AF pattern with repercussions on the pattern characterizing parameters. A shutdown is an important issue to consider since inclusion or elimination of lost segments influence the results dramatically, as demonstrated in the present study. Commonly, studies evaluating AF detector performance tend to exclude poor quality segments [35]–[37], thus exaggerating performance. To shed light on this issue, we included comparison strategy #1 which excludes shutdowns from both annotated and detector-based patterns. Unsurprisingly, such a strategy leads to markedly better agreement and reliability at the expense of a distorted reference pattern. While the exclusion of shutdowns may be tolerable when investigating AF detector performance, this is not the case in studies investigating methods for pattern characterization.

The main limitation of the present study is that the pattern characterizing parameters were explored using databases with recordings of relatively short duration. As a result, shutdowns were only attributed to poor signal quality, while various technical (e.g., hardware, software, connectivity) and user-related (e.g., non-wear, user error) issues are also commonly encountered during activities of daily living [38]. Availability of longer-duration AF patterns would allow exploring a larger variety of parameters such as those of the Hawkes model.

VII. Conclusions

The results show that AF aggregation is considerably less sensitive to detection errors than AF density, and, therefore, AF aggregation should be preferred. To improve performance, future research should put more emphasis on AF pattern characterization rather than just focusing on achieving increasingly better AF detection performance.

References

[1] M. Hautmann *et al.*, "Left atrial appendage thrombus formation, potential of resolution and association with prognosis in a large real-world cohort," *Sci. Rep.*, vol. 13, no. 1, p. 889, 2023, https://doi.org/10.1038/s41598-023-27622-3.

[2] M. Handke *et al.*, "Left atrial appendage flow velocity as a quantitative surrogate parameter for thromboembolic risk: determinants and relationship to spontaneous echocontrast and thrombus formation–a transesophageal echocardiographic study in 500 patients with cerebral ischemia," *J. Am. Soc. Echocardiogr.*, vol. 18, no. 12, pp. 1366–1372, 2005, https://doi.org/10.1016/j.echo.2005.05.006.

[3] N. Al-Saady, O. Obel, and A. Camm, "Left atrial appendage: structure, function, and role in thromboembolism," *Heart*, vol. 82, no. 5, pp. 547–554, 1999, https://doi.org/10.1136/hrt.82.5.547.

[4] M. Petersen *et al.*, "Left atrial appendage morphology is closely associated with specific echocardiographic flow pattern in patients with atrial fibrillation," *Europace*, vol. 17, no. 4, pp. 539–545, 2014, https://doi.org/10.1093/europace/euu347.

[5] G. Hindricks *et al.*, "2020 ESC Guidelines for the diagnosis and management of atrial fibrillation developed in collaboration with the European Association of Cardio-Thoracic Surgery (EACTS) The Task Force for the diagnosis and management of atrial fibrillation of the European Society of Cardiology (ESC) developed with the special contribution of the European Heart Rhythm Association (EHRA) of the ESC," *Eur. Heart J.*, vol. 42, no. 5, pp. 373–498, 2020, https://doi.org/10.1093/eurheartj/ehaa612.

[6] G. S. Greer *et al.*, "Random and nonrandom behavior of symptomatic paroxysmal atrial fibrillation," *Am. J. Card.*, vol. 64, no. 5, pp. 339–342, 1989, https://doi.org/10.1016/0002-9149(89)90531-6.

[7] A. M. Gillis and M. S. Rose, "Temporal patterns of paroxysmal atrial fibrillation following DDDR pacemaker implantation," *Am. J. Cardiol.*, vol. 85, no. 12, pp. 1445–1450, 2000, https://doi.org/10.1016/s0002-9149(00)00792-x.

[8] W. F. Kaemmerer, M. S. Rose, and R. Mehra, "Distribution of patients? Paroxysmal atrial tachyarrhythmia episodes: implications for detection of treatment efficacy," *J. Cardiovasc. Electrophysiol.*, vol. 12, no. 2, pp. 121–130, 2001, https://doi.org/10.1046/j.1540-8167.2001.00121.x.

[9] L. A. Shehadeh, L. S. Liebovitch, and M. A. Wood, "Temporal patterns of atrial arrhythmia recurrences in patients with implantable defibrillators: Implications for assessing antiarrhythmic therapies," *J. Cardiovasc. Electrophysiol.*, vol. 13, no. 4, pp. 303–309, 2002, https://doi.org/10.1046/j.1540-8167.2002.00303.x.

[10] E. I. Charitos *et al.*, "Atrial fibrillation density: A novel measure of atrial fibrillation temporal aggregation for the characterization of atrial fibrillation recurrence pattern," *Appl. Cardiopulm. Pathophysiol.*, vol. 17, no. 1, pp. 3–10, 2013.

[11] M. Šimaitytė *et al.*, "Quantitative evaluation of temporal episode patterns in paroxysmal atrial fibrillation," in *Proc. Comput. Cardiol.*, vol. 45, 2018, pp. 1–4, https://doi.org/10.22489/CinC.2018.059.

[12] M. Henriksson *et al.*, "Modeling and estimation of temporal episode patterns in paroxysmal atrial fibrillation," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 1, pp. 319–329, 2020, https://doi.org/10.1109/TBME.2020.2995563.

[13] J. Saiz-Vivó *et al.*, "Atrial fibrillation episode patterns as predictor of clinical outcome of catheter ablation," *Med. Biol. Eng. Comput.*, vol. 61, pp. 317–327, 2022, https://doi.org/10.1007/s11517-022-02713-x.

[14] A. Petrėnas, V. Marozas, and L. Sörnmo, "Low-complexity detection of atrial fibrillation in continuous long-term monitoring," *Comput. Biol. Med.*, vol. 65, pp. 184–191, 2015, https://doi.org/10.1016/j.compbiomed.2015.01.019.

[15] J. P. Martínez *et al.*, "A wavelet-based ECG delineator: evaluation on standard databases," *IEEE Trans. Biomed. Eng.*, vol. 51, no. 4, pp. 570–581, 2004, https://doi.org/10.1109/TBME.2003.821031.

[16] A. L. Goldberger *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000, https://doi.org/10.1161/01.cir.101.23.e215.

[17] C. Orphanidou *et al.*, "Signal-quality indices for the electrocardiogram and photoplethysmogram: derivation and applications to wireless monitoring," *IEEE J. Biomed. Health Inform.*, vol. 19, no. 3, pp. 832–838, 2014, https://doi.org/10.1109/JBHI.2014.2338351.

[18] L. Sörnmo (ed.), "Atrial fibrillation from an engineering perspective," *Springer*, 2018, https://doi.org/10.1007/978-3-319-68515-1.

[19] T. K. Koo and M. Y. Li, "A guideline of selecting and reporting intraclass correlation coefficients for reliability research," *J. Chiropr. Med.*, vol. 15, no. 2, pp. 155–163, 2016, https://doi.org/10.1016/j.jcm.2016.02.012.

[20] D. Liljequist, B. Elfving, and K. Skavberg Roaldsen, "Intraclass correlation–a discussion and demonstration of basic features," *PLoS One*, vol. 14, no. 7, p. e0219854, 2019, https://doi.org/10.1371/journal.pone.0219854.

[21] "ANSI/AAMI EC57:2012/(R)2020: Testing and reporting performance results of cardiac rhythm and ST segment measurement algorithms," 2020, https://array.aami.org/doi/pdf/10.2345/9781570204784.ch1.

[22] Y. Guo *et al.*, "4S-AF scheme and ABC pathway guided management improves outcomes in atrial fibrillation patients," *Eur. J. Clin. Invest.*, vol. 52, no. 6, p. e13751, 2022, https://doi.org/10.1111/eci.13751.

[23] T. S. Potpara *et al.*, "The 4S-AF scheme (stroke risk; symptoms; severity of burden; substrate): A novel approach to in-depth characterization (rather than classification) of atrial fibrillation," *Thromb. Haemost.*, vol. 121, no. 3, pp. 270–278, 2020, https://doi.org/10.1055/s-0040-1716408.

[24] S.-Y. Yang *et al.*, "Atrial fibrillation burden and the risk of stroke: A systematic review and dose-response meta-analysis," *World J. Clin. Cases*, vol. 10, no. 3, pp. 939–953, 2022, https://doi.org/10.12998/wjcc.v10.i3.939.

[25] D. S. Chew *et al.*, "Arrhythmic burden and the risk of cardiovascular outcomes in patients with paroxysmal atrial fibrillation and cardiac implanted electronic devices," *Circ. Arrhythm. Electrophysiol.*, vol. 15, no. 2, p. e010304, 2022, https://doi.org/10.1161/CIRCEP.121.010304.

[26] P. Zimetbaum *et al.*, "Role of atrial fibrillation burden in assessing thromboembolic risk," *Circ. Arrhythm. Electrophysiol.*, vol. 7, no. 6, pp. 1223–1229, 2014, https://doi.org/10.1161/CIRCEP.114.001356.

[27] G. L. Botto *et al.*, "Impact of the pattern of atrial fibrillation on stroke risk and mortality," *Arrhythmia Electrophysiol. Rev.*, vol. 10, no. 2, pp. 68–76, 2021, https://doi.org/10.15420/aer.2021.01.

[28] L. Sörnmo *et al.*, "Letter regarding the article 'Detecting atrial fibrillation by deep convolutional neural networks by Xia et al.'," *Comput. Biol. Med.*, vol. 100, pp. 41–42, 2018, https://doi.org/10.1016/j.compbiomed.2018.06.027.

[29] M. Butkuvienė *et al.*, "Considerations on performance evaluation of atrial fibrillation detectors," *IEEE Trans. Biomed. Eng.*, vol. 68, no. 11, pp. 3250–3260, 2021, https://doi.org/10.1109/TBME.2021.3067698.

[30] A. Petrėnas *et al.*, "Detection of occult paroxysmal atrial fibrillation," *Med. Biol. Eng. Comput.*, vol. 53, no. 4, pp. 287–297, 2015, https://doi.org/10.1007/s11517-014-1234-y.

[31] M. Butkuvienė *et al.*, "Atrial fibrillation episode patterns and their influence on detection performance," in *Proc. Comput. Cardiol.*, vol. 48, 2021, pp. 1–4, https://doi.org/10.23919/CinC53138.2021.9662847.

[32] Z. Kalarus *et al.*, "Searching for atrial fibrillation: looking harder, looking longer, and in increasingly sophisticated ways. An EHRA position paper," *Europace*, vol. 25, no. 1, pp. 185–198, 2023, https://doi.org/10.1093/europace/euac144.

[33] A. B. e Gala *et al.*, "NICE atrial fibrillation guideline snubs wearable technology: a missed opportunity?" *Clin. Med.*, vol. 22, no. 1, pp. 77–82, 2022, https://doi.org/10.7861/clinmed.2021-0436.

[34] L. Zhu *et al.*, "Atrial fibrillation detection and atrial fibrillation burden estimation via wearables," *IEEE J. Biomed. Health Inform.*, vol. 26, no. 5, pp. 2063–2074, 2022, https://doi.org/10.1109/JBHI.2021.3131984.

[35] P.-C. Chang *et al.*, "Atrial fibrillation detection using ambulatory smartwatch photoplethysmography and validation with simultaneous holter recording," *Am. Heart J.*, vol. 247, pp. 55–62, 2022, https://doi.org/10.1016/j.ahj.2022.02.002.

[36] M. Dörr *et al.*, "The WATCH AF trial: SmartWATCHes for detection of atrial fibrillation," *JACC Clin. Electrophysiol.*, vol. 5, no. 2, pp. 199–208, 2019, https://doi.org/10.1016/j.jacep.2018.10.006.

[37] S. K. Bashar *et al.*, "Atrial fibrillation detection from wrist photoplethysmography signals using smartwatches," *Sci. Rep.*, vol. 9, no. 1, p. 15054, 2019, https://doi.org/10.1038/s41598-019-49092-2.

[38] S. Cho *et al.*, "Factors affecting the quality of person-generated wearable device data and associated challenges: rapid systematic review," *JMIR mHealth uHealth*, vol. 9, no. 3, p. e20738, 2021, https://doi.org/10.2196/20738.