Considerations on Performance Evaluation of Atrial Fibrillation Detectors

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

***Abstract*—*Objective:* A large number of atrial fibrillation (AF) detectors have been published in recent years, signifying that the comparison of detector performance plays a central role, though not always consistent. The aim of this study is to shed needed light on aspects crucial to the evaluation of detection performance. *Methods:* Three types of AF detector, using either information on rhythm, rhythm and morphology, or segments of ECG samples, are implemented and studied on both real and simulated ECG signals. The properties of different performance measures are investigated, for example, in relation to dataset imbalance. *Results:* The results show that performance can differ considerably depending on the way detector output is compared to database annotations, i.e., beat-to-beat, segment-to-segment, or episode-to-episode comparison. Moreover, depending on the type of detector, the results substantiate that physiological and technical factors, e.g., changes in ECG morphology, rate of atrial premature beats, and noise level, can have a considerable influence on performance. *Conclusion:* The present study demonstrates overall strengths and weaknesses of different types of detector, highlights challenges in AF detection, and proposes five recommendations on how to handle data and characterize performance.**

***Index Terms*—Atrial fibrillation, deep learning, detection, expert-crafted detection, performance evaluation, performance measures.**

I. Introduction

THE recent interest in deep learning (DL) has led to an avalanche of atrial fibrillation (AF) detectors, e.g., [1]–[17]. As a consequence, the problem of how to evaluate and compare performance between different detectors, whether based on DL or expert-crafted features, is brought into focus. To outline a framework for evaluation that not only ensures a fair comparison but also goes beyond reporting overall performance measures is therefore essential.

While public databases facilitate the comparison of detector performance, conclusions should be made with caution for a number of reasons. Rather than using the entire database, certain detectors have been evaluated on a subset, e.g., by excluding poor-quality signal segments or omitting segments for the purpose of balancing the datasets.

Depending on the approach taken to comparing detector output to database annotations, i.e., beat-to-beat [18]–[24], segment-to-segment [16], [22], [25]– [29], or episode-to-episode comparison [24], [28], [30], [31], the performance can differ considerably. Although only results computed using the same approach must be compared, this is not always the case.

To express performance in terms of statistical measures, e.g., sensitivity and specificity, is common practice. However, the use of performance measures should be accompanied by results uncovering detector properties. For example, by investigating what signal scenarios cause frequent false alarms, weaknesses in detector design can be more efficiently addressed. Such understanding can be gained by means of simulated ECG signals which, in contrast to real signals, offer control of principal quantities such as type and level of noise, rate of atrial premature beats (APBs), and AF burden, i.e., the percentage of time a patient spends in AF during the monitored period [32], [33]. The interest in brief AF episodes (< 30 s) and their association with future risk of stroke [34], [35], motivates the simulation of signals with varying episode length to enrich the understanding of performance.

Since existing studies on AF detection offer very little insight on how well episode patterns are captured, further studies are needed that investigate the influence of missed and falsely detected episodes on pattern characterizing parameters, e.g., minimal AF episode duration [36], clustering of AF episodes [37], and temporal distribution of AF episodes [38]. The need for episode pattern analysis, complementing the analysis of AF burden, is emphasized in recent clinical guidelines, e.g., by determining the density of episodes per unit of time [39], [40].

The present paper addresses aspects crucial to the evaluation of AF detector performance, leading up to a set of investigation-based recommendations on how to handle data and characterize performance. For the purpose of illustrating differences in performance, three types of AF detector are implemented (Sec. III) and studied on both real and simulated ECG signals (Sec. II).

Manuscript received October 21, 2020; revised January 28, 2021; accepted March 16, 2021. Date of publication March 22, 2021; date of current version October 20, 2021. This work was supported by the European Regional Development Fund under Grant agreement 01.2.2-LMT-K-718-03-0027 with the Research Council of Lithuania (LMTLT) and Swedish Research Council under Grant 2016-03382. *(Corresponding author: Monika Butkuvienė.)*

Monika Butkuvienė is with the Biomedical Engineering Institute, Kaunas University of Technology, 44249 Kaunas, Lithuania (e-mail: monika.butkuviene@ktu.lt).

Andrius Petrėnas and Andrius Sološenko are with the Biomedical Engineering Institute, Kaunas University of Technology, Lithuania.

Alba Martín-Yebra and Leif Sörnmo are with the Department of Biomedical Engineering, Lund University, Sweden.

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

Digital Object Identifier 10.1109/TBME.2021.3067698

The results shed light on the suitability of different performance measures for AF detection, the basis for comparing detector output to annotations, and the importance of investigating the influence of physiological and technical factors on performance (Sec. V). Besides discussing the results, Sec. VI provides an overview of how important aspects are dealt with in the literature.

II. ECG Signals With Paroxysmal Atrial Fibrillation

A. Clinical Signals

The Saint Petersburg Atrial Fibrillation Database (SPAFDB) consists of 36 three-lead (V_1, V_6, Y) ambulatory recordings, lasting from 1 to 7 days and amounting to a total of 158 days of monitoring [37]. In total, SPAFDB consists of 2370 manually annotated AF episodes which account for 19% of the total monitoring time.

The publicly available MIT–BIH Atrial Fibrillation Database (AFDB) consists of 23 10-h, two-lead ambulatory ECG recordings from patients with AF, mostly paroxysmal [41]. The Long-Term AF Database (LTAFDB) consists of 84 24-h two-lead ambulatory ECG recordings acquired in patients with paroxysmal or persistent AF [41]. In total, AFDB consists of 297 manually annotated AF episodes which account for 43% of the total monitoring time, while LTAFDB consists of 7329 manually annotated AF episodes which account for 59% of the total monitoring time. It should be noted that the leads are not specified for any of the two databases.

B. Simulated Signals

To investigate the influence of various physiological and technical factors on performance, simulated ECGs in paroxysmal AF are used [32]. The model produces 12-lead ECGs composed of real signal components randomly selected from three datasets, each consisting of ventricular rhythm, atrial activity (f-waves or P-waves), and QRST complexes. Accounting for the switching between non-AF and AF, these components, together with noise, are summed to produce simulated signals. The noise, being the sum of baseline wander, muscle noise, and electrode movement artifacts, is scaled to the desired root mean square (RMS) value. Two of the model parameters are set based on the overall characteristics of AFDB, namely AF burden to 0.37 and median episode length to 167 beats. The APB rate is set to 0.05 and noise RMS level to 0.02 mV, whereas the remaining model parameters have their default values [32].

III. Atrial Fibrillation Detectors Under Comparison

In the literature, three types of AF detector can be discerned, those using only rhythm, both rhythm and morphology, and segments of ECG samples as input. The first two types require prior QRS detection, here accomplished by the wavelet-based detector described in [42], whereas the third type does not. In the following, one representative of each detector type is considered with the aim to reveal overall strengths and weaknesses.

A. Rhythm-Based Detector

Rhythm-based detection makes use of that AF episodes are manifested by irregular RR intervals which often are associated with increased heart rate. The implemented detector, designed to detect brief AF episodes, includes blocks for ectopic beat filtering, bigeminy suppression, characterization of RR interval irregularity, and signal fusion [21]. The detector is used to process all three ECG databases and simulated signals, employing the parameter values in [21].

B. Rhythm- and Morphology-Based Detector

Four parameters serve as input to the rhythm- and morphology-based detector, capable of detecting AF episodes as short as 8 beats [31]: 1. Rhythm irregularity, quantified by the rhythm-based detector described in Sec. III.A; 2. P-wave absence, quantified by computing the normalized ratio of the rectified signal in the PQ interval to that of the TQ interval; 3. f-wave presence, quantified by the squared and summed error between different PR intervals; and 4. noise level, quantified by the spectral entropy ratio-weighted RMS value of the extracted f-wave signal. The latter three parameters are determined from an f-wave signal, extracted using an echo state network [43]. The parameters are fed to a fuzzy logic classifier producing a fuzzy output, i.e., a value between 0 and 1, reflecting the likelihood of AF being present in the sliding detection window. The detector requires two ECG leads, i.e., one with negligible atrial activity (e.g., V_6) and another with atrial activity (e.g., V_1). The detector is used to process SPAFDB and simulated signals, employing the parameter values in [31].

C. DL-Based Detector

A DL-based detector is implemented using a 1D convolutional neural network (CNN), whose structure is inspired by those described in [16], [44]. The ECG signal is preprocessed using a band-pass filter (0.5–40 Hz) to remove baseline wander and high-frequency noise. The CNN is composed of two convolutional layers and one fully connected layer. Both convolutional layers rely on 128 kernels with a stride of one, followed by a 1×32 average-pooling layer with a stride of 32. The fully connected layer consists of 256 neurons with a rectified linear unit activation function and two output neurons with a softmax activation function. To mitigate the risk of overfitting, all layers are followed by dropout layers with probabilities of 0.5. The outputs of the convolutional layers are batch-normalized. The DL-based detector is trained using the gradient-based Adam optimizer [45], with a learning rate of 0.001 and a batch size of 128.

The detector was trained on two-thirds of AFDB and validated on the remaining one-third, using the lead with the most negative S-wave which reasonably well mimics V_1 of the test databases. To increase the number of segments for training, each signal was divided into 30-s segments with 50% overlap. Poor-quality segments were eliminated based on sample skewness and kurtosis as proposed in [46]. In total, 358 segments out of 59185 were eliminated due to poor quality. The resulting training dataset consists of 25169 segments assigned to AF and 33658 segments to non-AF. To equalize the signal amplitude across a recording, the modulus of each segment was taken and normalized to the interval $[0, 1]$. The detector is tested on SPAFDB and simulated

signals, again with each signal divided into 30-s segments but without any overlap.

IV. Performance Evaluation

A. Annotation Comparison

The predominant approach to processing annotations is to compare the detector output to the annotations on a beat-to-beat basis. Another approach is to compare L-beat segments, where a segment is assigned to AF when at least 50% of the L detected beats are in agreement with the beat annotations. Yet another approach is to count the number of correctly detected AF episodes: an episode is considered correctly detected when the overlap between the detector output and the episode annotation exceeds a predefined threshold, e.g., 50%.

In the following, these three approaches to processing annotations are referred to as beat-to-beat, segment-to-segment, and episode-to-episode comparison, respectively.

B. Performance Measures

Performance is invariably evaluated by determining the number of correctly detected AF cases (true positives, *TP*), the number of correctly detected non-AF cases (true negatives, *TN*), the number of falsely detected AF cases (false positives, *FP*), and the number of missed AF cases (false negatives, *FN*). Depending on the type of annotation comparison, "case" refers to either beat, segment, or episode. From these four counts, the well-known performance measures, such as sensitivity (*Se*), specificity (*Sp*), positive predictive value (*PPV*), negative predictive value (*NPV*), and accuracy (*Acc*) are computed.

Other measures include balanced accuracy (Acc_B), F_1 score, and Matthews correlation coefficient (*Mcc*), defined by

$$Acc_B = \frac{1}{2}(Se + Sp), \tag{1}$$

$$F_1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN}, \tag{2}$$

$$Mcc = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}, \tag{3}$$

respectively. The measures Acc_B and F_1 both take values in the interval $[0, 1]$, where 1 means perfect detection and 0.5 random detection. In its original definition, *Mcc* takes values in the interval $[-1, 1]$, however, to facilitate a comparison between performance measures, *Mcc* is normalized to the interval $[0, 1]$ [47].

V. Results

A. Analysis of Performance Measures

Fig. 1(a) shows an annotated AF pattern composed of just a few episodes, together with the output of the rhythm-based detector composed of numerous false detections making up for 2% of the total number of beats. Using beat-to-beat comparison, the receiver operating characteristic (ROC) shown in Fig. 1(b) suggests near-perfect performance. However, due to the huge

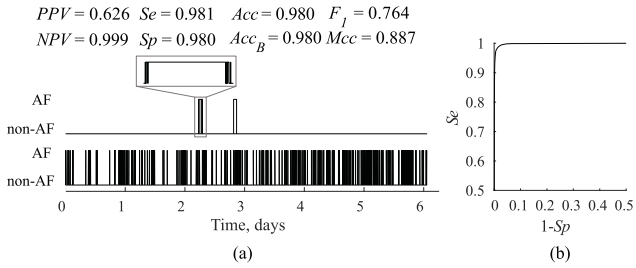

Fig. 1. (a) Annotated AF pattern from SPAFDB (upper panel), output of the rhythm-based detector (lower panel), and (b) corresponding ROC. The performance measures are computed using beat-to-beat comparison. The annotated pattern consists of 8 episodes with a median episode length of 113 beats, while the detector-produced pattern consists of 518 episodes with a median episode length of 15 beats.

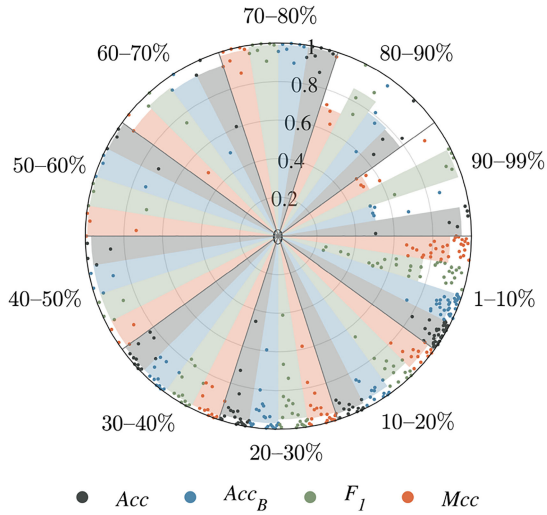

Fig. 2. The effect of data imbalance, expressed as AF burden, on different performance measures. The arc of the circle indicates AF burden. The dots in each colored sector show the values of a performance measure obtained for different AF patterns, using the rhythm-based detector and beat-to-beat comparison. The radius of the colored sector represents the median of the values of a performance measure.

imbalance between non-AF and AF beats (96.7% are non-AF), such a conclusion is misleading. Since 98.1% of the AF beats and 98.0% of the non-AF beats are correctly detected, the false AF detections have negligible influence on the ROC. In terms of performance measures, *Acc* and Acc_B are insensitive to data imbalance and therefore indicate higher performance, whereas F_1 and *Mcc* are sensitive and therefore indicate lower performance, see Fig. 1(a).

To shed further light on data imbalance, the performance of the rhythm-based detector is studied on 103 recordings from SPAFDB, AFDB, and LTAFDB; 40 recordings with AF burden $<1\%$ and $>99\%$ were excluded. Fig. 2 shows that imbalance has only a minor effect on *Acc*, Acc_B, F_1, and *Mcc* when AF burden is between 10% (negative imbalance) to 80% (positive imbalance). Interestingly, only F_1 and *Mcc* are influenced by a negative imbalance of 1–10%, while *Acc* and Acc_B remain essentially unchanged. Since the sectors 80–90% and 90–99% contain very few values, no meaningful observations can be made.

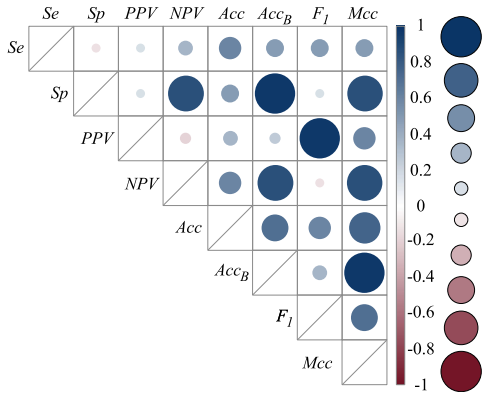

Fig. 3. Pearson correlation coefficient for different performance measures, obtained using the rhythm-based detector and beat-to-beat comparison.

The information carried by the different performance measures is investigated by correlation analysis, again using the rhythm-based detector and beat-to-beat comparison on 103 recordings. Fig. 3 shows that *Sp*, *NPV*, Acc_B, and *Mcc* are strongly correlated ($r > 0.8$) with each other, while *PPV* and F_1 do not correlate with *Sp*, *NPV*, Acc_B, or *Mcc*. The measure *PPV* correlates strongly with F_1 since both are determined by the number of false positives. On the other hand, *Sp* and *NPV* are strongly correlated due to the fact that missed AF beats are uncommon in rhythm-based AF detection, thus reducing *NPV*.

B. Influence of Annotation Comparison

Fig. 4 shows how the type of annotation comparison influences performance. For the rhythm-based and the rhythm- and morphology-based detectors, episode-to-episode comparison indicates much lower performance for all measures than do the other two types of comparison. However, for the rhythm- and morphology-based detector, the difference in performance is less pronounced. While the segment-to-segment comparison indicates the best performance, the difference relative to beat-to-beat comparison is negligible.

Fig. 5 shows how segment length influences performance using segment-to-segment comparison. As expected, performance deteriorates as the length shortens due to that shorter manifestations of noise and sporadic ectopic beats cause more false detections.

Fig. 6 shows how the overlap percentage between detected and annotated episodes influences performance using episode-to-episode comparison. As expected, *Se* and *NPV* decrease and *Sp* and *PPV* increase as the overlap percentage increases since fewer episodes are detected. However, the intersection point between the *Se*/*NPV* and *Sp*/*PPV* curves differs considerably for the two types of detector, being 15% for the rhythm-based and 48% for the rhythm- and morphology-based.

C. Factors Influencing AF Detection Performance

1) Lead Selection: Detection accuracy as a function of processed lead is presented in Fig. 7. The performance of the DL-based detector depends heavily on lead, with the best performance obtained for V_1, i.e., the one used for training, then dropping dramatically for the other leads with lower f-wave amplitude. The performance of the expert-crafted detectors is largely independent of selected lead. It should be noted that the lead dependence of the rhythm-based detector is due to that the performance of the QRS detector is lead-dependent.

2) APB Rate: Detection accuracy as a function of APB rate is presented in Fig. 8. The performance of the rhythm-based and the DL-based detectors drops rapidly as the APB rate increases, whereas the rhythm- and morphology-based detector performs well even at high APB rates thanks to the inclusion of morphologic information. For the rhythm-based detector, the drop in performance is expected as this type of detector is known to poorly discriminate AF from irregular rhythms with APBs.

3) Noise Level: Detection accuracy as a function of noise level is presented in Fig. 9. The performance of the rhythm-based and the rhythm- and morphology-based detectors drops rapidly when the noise level exceeds 0.15 mV, largely attributed to the drop in performance of the QRS detector. Although less dependent on noise level, the DL-based detector performs considerably worse at lower noise levels than the other two detectors. It should be recalled that the performance of the DL-based detector does not depend on QRS detection.

VI. Discussion

The recent, rapid progress in AF detector development is driven by clinical relevance and advancements in medical technologies. However, this development comes with the challenge of adequately evaluating and comparing performance relative to published detectors.

A. Performance Evaluation

In the present study, three types of annotation comparison are considered. The results show that performance depends on the selected type, notably that episode-to-episode comparison indicates much poorer performance than do the other two types of comparison (Fig. 4). Even when the same type is used, segment length (Fig. 5) and episode overlap (Fig. 6) influence performance, e.g., increasingly poorer when shorter segments are used. Therefore, a meaningful comparison can only be made when these aspects are taken into consideration. If not, conclusions on detector superiority, which tend to be common in the literature, cannot and should not be drawn.

Another aspect which deserves consideration is that neither beat-to-beat nor segment-to-segment comparison indicates the number of detected AF episodes. Obviously, episode-to-episode comparison is more appropriate to use, especially when dense episode patterns are the subject of analysis. However, this type of comparison is rarely used, likely because it results in lower performance figures (Fig. 4). For DL-based detectors, segment-to-segment comparison is the preferred choice since deep neural networks typically do not rely on heartbeat timing; the segment length is usually related to what is deemed the shortest detectable episode. DL-based detectors requiring heartbeat timing include [11], [17], [48].

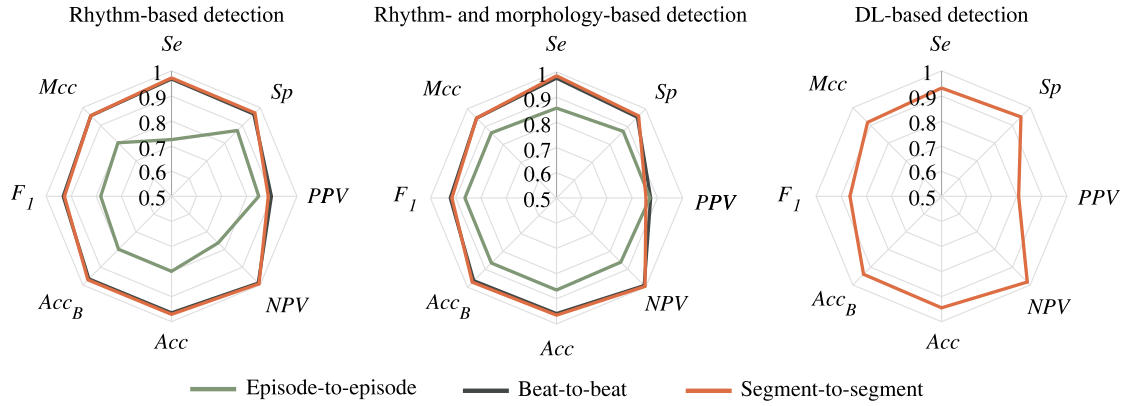

Fig. 4. Detector performance using episode-to-episode (50% overlap), beat-to-beat, and segment-to-segment (30 s) annotation comparison. Note that only segment-to-segment comparison can be used to describe the performance of the DL-based detector, since the detector structure does not lend itself to the other two types of comparison. The results are obtained using SPAFDB.

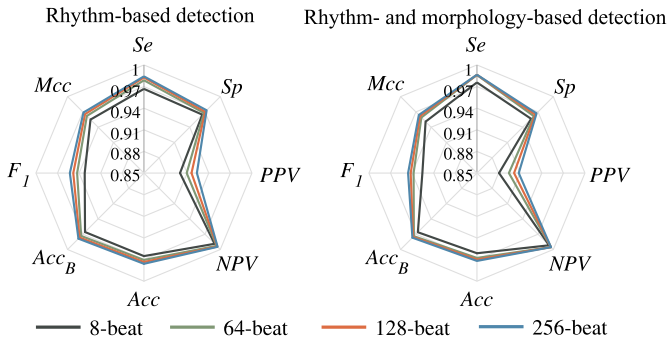

Fig. 5. Influence of segment length on detector performance using segment-to-segment comparison. The DL-based detector is not included since it was trained to process 30-s segments. The results are obtained using SPAFDB.

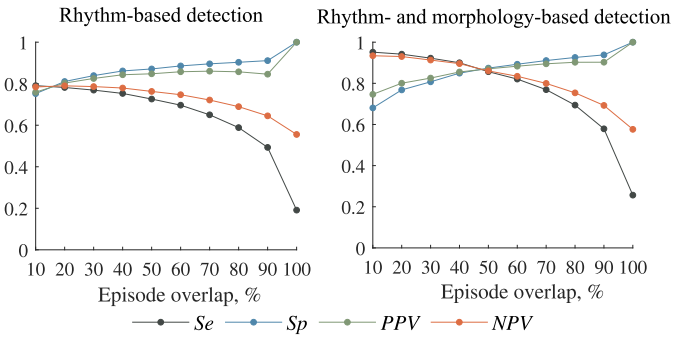

Fig. 6. Influence of episode overlap on detector performance using episode-to-episode comparison. The results are obtained using SPAFDB.

Since no consensus has been established on what measures should be used to report on performance, an important aim of the present paper is to facilitate such a consensus by highlighting various properties of measures commonly used in the literature. Since AF detection represents a binary problem, it is intimately associated with the 2×2 confusion matrix defined by the counts *TP*, *FN*, *TN*, and *FP*, cf. Sec. IV.B. Combinations of these four counts have been used to define performance measures, with *Se*, *Sp*, *PPV*, and *NPV* as the most popular [21], [28], [49], [50]. None of these measures can, however, be considered fully informative as their respective definitions involve only two counts of the confusion matrix [47]. Joint use of all four measures provides richer information on performance, but also renders a comparison of performance more complicated. Therefore, it is understandable that the use of a single overall performance measure, e.g., *Acc*, F_1, *Mcc*, has become popular [51], [52]. However, a single overall performance measure hides important properties. It is well-known that *Acc*, being a popular measure in AF detection, tends to inflate performance for imbalanced datasets [47], [53], [54], cf. Fig. 2. By comparing F_1 with *Mcc*, it should be highlighted that *Mcc* depends on the number of samples correctly classified as true negatives, while F_1 does not. Since *Mcc* indicates good performance only when most AF episodes and most non-AF "episodes" are correctly detected, we recommend the use of *Mcc* instead of F_1 or *Acc* when evaluating overall performance.

The area under the ROC, known as the area-under-the-curve (AUC), is another single overall performance measure popular in many studies, e.g., [12], [21], [22], [25], [55], [56]. Unfortunately, the AUC results from integrating *Se* and *Sp* not only in regions of operational interest, but also in regions of no clinical interest [47], [57], [58]. Hence, we recommend that AUC is disregarded when reporting on performance, while it may be used to provide better understanding of how different parameter settings influence performance [21], [25], [55].

B. Factors Influencing Performance

The influence of various physiological and technical factors on performance is rarely investigated in the literature, despite the fact that essential information on detector properties can be uncovered, cf. Sec. V.C. Situations in which performance degrades deserve particular attention.

Since the performance of the DL-based detector depends heavily on the lead selected for processing (Fig. 7(b)), the

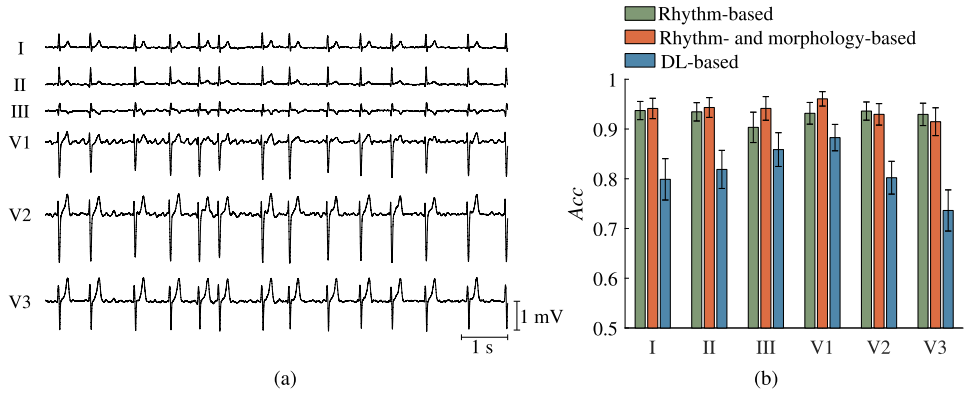

Fig. 7. (a) Simulated multi-lead ECG during AF and (b) detection accuracy (*Acc*) as a function of lead selection. The results are based on 100 simulated 1-h ECGs and presented as mean ± confidence interval (CI) (95%).

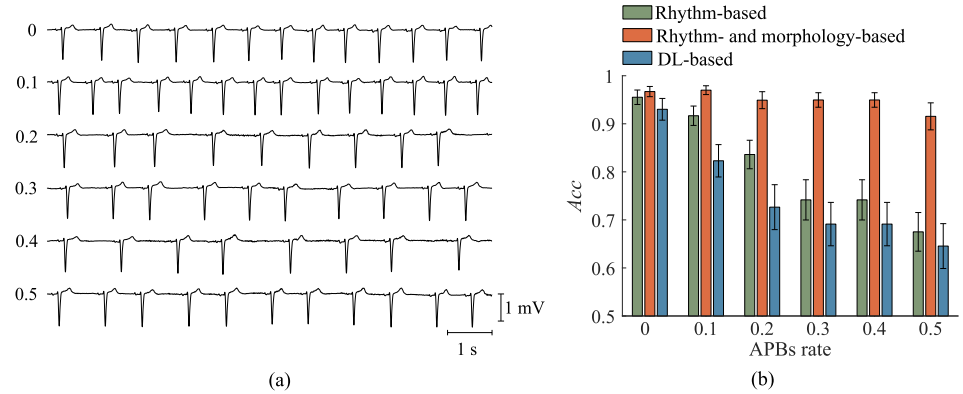

Fig. 8. (a) Simulated ECGs with different APB rates and (b) detection accuracy (*Acc*) as a function of APB rate. The results are based on 100 simulated 1-h ECGs and presented as mean ± CI (95%).

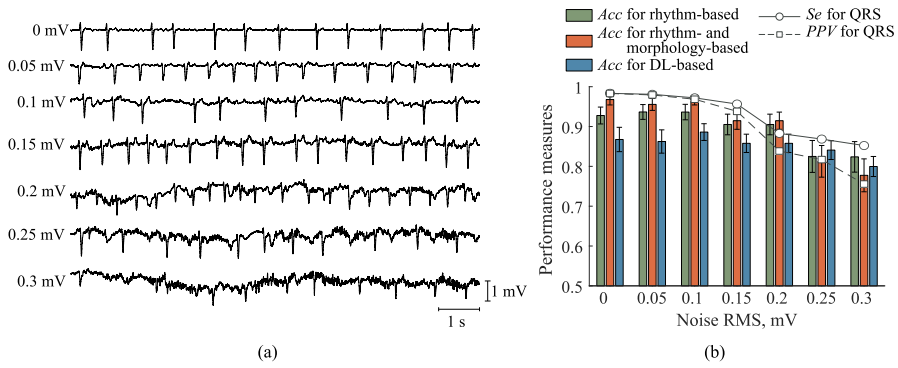

Fig. 9. (a) Simulated ECGs (V_1) during AF with different noise levels (RMS) and (b) detection accuracy (*Acc*) and QRS detection performance (*Se* and *PPV*) as a function of noise level. The results are based on 100 simulated 1-h ECGs and presented as mean ± CI (95%).

datasets used for training and testing should consist of recordings from the same lead to achieve the best performance. In the present study, lead V_1 was used since its f-waves are more prominent than in any of the other leads of the standard 12-lead ECG. When using different databases for training and testing, it is not only important to use a similar lead, but also to avoid differences in measurement equipment and large variation in signal quality. These observations are probable reasons why nearly all DL-based detectors in the literature have been tested using cross-validation on the training database, see Table I. As for the two expert-crafted detectors, their performance is not nearly as sensitive to lead selection (Fig. 7(b)).

Since P- and f-wave information is part of the decision process, the performance of the rhythm- and morphology-based

TABLE I
COMPARISON OF DL-BASED AF DETECTORS

Authors	All records of AFDB used	Different patients in training & test sets	Testing on other databases	Plots of problem signals
Xia et al., 2018 [15]	No	No	No	No
Faust et al., 2018 [14]	Yes	Yes	No	No
He et al., 2018 [13]	No	Yes	No	No
Andersen et al., 2019 [11]	Yes	No	Yes	Yes
Lai et al., 2019 [10]	Yes	No	No	No
Dang et al., 2019 [9]	Yes	No	No	No
Fujita et al., 2019 [8]	No	No	No	No
Wang, 2020 [7]	Yes	No	No	No
Jin et al., 2020 [6]	No	Yes	No	No
Huang et al., 2020 [5]	No	No	No	No
Zhang et al., 2020 [4]	Yes	No	No	No
Ghosh et al., 2020 [3]	No	No	No	No
Shi et al., 2020 [2]	No	No	No	No
Mousavi et al., 2020 [17]	Yes	Yes	No	No
Percentage "Yes"	**50%**	**29%**	**7%**	**7%**

TABLE II
COMPARISON OF EXPERT-CRAFTED AF DETECTORS

Authors	All records of AFDB used	Different patients in training & test sets	Testing on other databases	Plots of problem signals
Dash et al., 2009 [29]	No	Yes	Yes	Yes
Babaeizadeh et al., 2009 [28]	Yes	Yes	Yes	Yes
Lake et al., 2011 [59]	Yes	Yes	Yes	Yes
Lian et al., 2011 [27]	Yes	Yes	Yes	Yes
Huang et al., 2011 [50]	Yes	Yes	Yes	No
Shouldice et al., 2012 [60]	Yes	Yes	Yes	No
Carvalho et al., 2012 [61]	Yes	No	No	No
Lee et al., 2013 [26]	No	Yes	Yes	Yes
Zhou et al., 2014 [55]	Yes	Yes	Yes	No
Ródenas et al., 2015 [20]	No	No	No	No
Asgari et al., 2015 [25]	Yes	No	No	Yes
Petrėnas et al., 2015 [31]	Yes	Yes	Yes	Yes
Zhou et al., 2015 [62]	Yes	Yes	Yes	No
Percentage "Yes"	**77%**	**77%**	**77%**	**54%**

detector remains largely unchanged for an increasing APB rate [31], cf. Fig. 8(b). On the other hand, as expected, the performance of the rhythm-based detector deteriorates considerably since decisions are based on RR interval information. More surprising is that the performance of the DL-based detector also deteriorates considerably, a behavior likely explained by a training process that identifies rhythm irregularity as a prominent AF feature; however, a representative training database with a greater variety of cardiac rhythms than that of AFDB should help improving performance [63]. In addition to investigating performance as a function of APB rate, other AF-masquerading arrhythmias, e.g., bi-/trigeminy, atrial tachycardia, and atrial flutter, deserve to be investigated as well.

The influence of missed and falsely detected QRS complexes on AF detector performance is rarely reported in the literature. In many studies, QRS detection is assumed to be perfect simply because the database annotations on QRS occurrence times serve as the starting point for AF detection [27], [29], [50], [55], [59]. However, in practice, ECGs are often noisy, e.g., when recorded under ambulatory conditions, and, therefore, QRS detection is far from perfect. As evidenced by Fig. 9, the performance of the expert-crafted AF detectors deteriorates at higher noise levels because of deteriorating QRS detector performance. In addition, for rhythm- and morphology-based detectors, noise enters through P- and f-wave features, thus calling for their careful use at higher noise levels.

AF detector performance as a function of different AF episode length also deserves attention since performance will deteriorate as the length becomes increasingly shorter. For example, when the median episode length of simulated signals decreased from 120 to 30 beats, the accuracy of the rhythm-based and the rhythm- and morphology-based detectors decreased from 0.84 to 0.65 and from 0.92 to 0.80, respectively [32].

C. Comparing Detector Performance

A meaningful comparison of performance requires that the datasets for training and testing are handled in the same way across studies. Firstly, all records of the database should be used, meaning that no records should be excluded due to poor signal quality or as a means to obtain balanced datasets [64]. Secondly, testing should be done on a database different from the one used for training so that performance can be established on unseen data. Thirdly, the same patient should not appear in

TABLE III
STRENGTHS AND WEAKNESSES OF DIFFERENT TYPES OF DETECTOR

Detector type	Strengths	Weaknesses
Rhythm-based	Low computational demands Well-suited for implementation in wearable devices	Difficult to distinguish AF from other irregular rhythms Depends on QRS detection performance
Rhythm- and morphology-based	Well-suited for distinguishing AF from other irregular rhythms	P- and f-wave features are sensitive to noise Depends on QRS detection performance
DL-based	No need for expert-crafted features No need for QRS detection Performance can be improved using training dataset with non-AF arrhythmias	Sensitive to changes in ECG morphology Unclear how the detector generalizes to unseen data Detection is data-driven and thus lacks interpretability Complex networks are computationally demanding

both the training and the test datasets. Though not critical to a comparison, it is highly desirable to provide insight on what particular problem situations cause performance to deteriorate, e.g., by presenting examples of motion artefacts and non-AF arrhythmias.

Tables I and II show to what extent DL-based and expert-crafted detectors, respectively, comply with the above-mentioned requirements; the listed detectors were all evaluated on AFDB. It is obvious that a comparison of performance can be highly misleading as data handling differs among the studies. Only 7 out of 14 (50%) of the DL-based detectors were tested on all records of AFDB, whereas 10 out 13 (77%) of the expert-crafted detectors; it should be noted that the records excluded in [13], [29], and [26] were motivated by incorrect annotations. Similarly, as few as 4 (29%) of the DL-based detectors used different patients in the training and the test sets, whereas 10 (77%) of the expert-crafted detectors. The effect of using different patients in the training and the test sets is illustrated by a recent study which reported excellent performance of the proposed DL model for AF detection ($Se = 0.991$, $Sp = 0.985$) when the same patient appeared in both sets [17]; however, when the sets contained different patients, the performance was mediocre ($Se = 0.905$, $Sp = 0.797$). Concerning testing on a database other than that used for training, only 1 DL-based detector (7%) complied with this requirement, whereas 10 (77%) of the expert-crafted detectors. Interestingly, the performance of that particular DL-based detector was found to drop dramatically when tested on another database (Se remained unchanged at 0.990 while Sp dropped from 0.970 to 0.860 [11]), thus offering a possible reason why different training and test databases have been shunned in the literature. Concerning plots of problem signals, again only 1 study (7%) on DL-based detection provided such information, whereas 7 (54%) of the studies on expert-crafted detectors.

A summary of strengths and weaknesses of rhythm-based, rhythm- and morphology-based, and DL-based detectors are presented in Table III.

D. Future Challenges in AF Detection

Today, the ECG can be acquired over an extended time period so that detailed characterization of AF episode patterns can be accomplished. The resulting information may be used to understand the significance of AF triggers and the development of complications such as stroke.

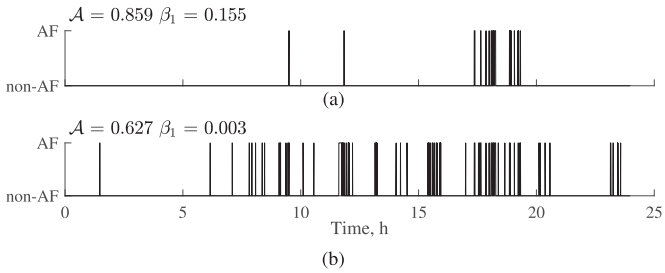

Fig. 10. Influence of false alarms on parameters characterizing AF patterns: (a) annotated AF pattern from LTAFDB and (b) detector-produced pattern.

AF patterns can be characterized by the heuristic parameter $\mathcal{A}$, defined so that patterns with a single short AF episode cluster are associated with values close to 1, while episode patterns spread out evenly over the entire monitoring period with values close to 0 [38]. Another approach to characterizing AF episode patterns is through history-dependent point process modeling, using an alternating, bivariate Hawkes self-exciting model recently introduced in [37]. The model parameter β_1 defines the exponential decay of the point process intensity function and provides information on episode clustering. Clustered AF episode patterns are associated with smaller β_1 values.

Sophisticated analysis of episode patterns implies higher demands on detection performance. As illustrated by Fig. 10, the annotated episodes differ considerably from those produced by the detector, and as a result, $\mathcal{A}$ and β_1 will differ considerably as well.

Unfortunately, episode analysis is made difficult in recordings containing noisy segments as AF detection becomes unreliable. Rather than simply discarding such segments from further analysis, as is often done, future research should focus on improving electrode technology and algorithms for signal processing and machine learning to ensure more reliable characterization of episode patterns.

E. Limitations

One detector representative for each of the three main types found in the scientific literature, i.e., rhythm-based, rhythm- and morphology-based, and deep learning-based, have been studied. Another type of detector is the one relying solely on atrial information; however, this type was not considered as it

is known to perform poorly in noisy signals [72]. While other representatives could have been chosen, the aim of the present study is to identify structure-dependent aspects on performance evaluation, not to grade the performance of different detectors, therefore making the choice of representatives less critical.

In certain applications, e.g., wearable devices, computational complexity needs to be considered when evaluating performance. Since complexity is detector-specific, such considerations have been left out of the present study. Nonetheless, it may be noted that the structure of a rhythm-based detector is typically less complex than that of a DL-based. For example, the rhythm-based detector in [21] requires 8 multiplications per RR interval, whereas the DL-based detector in [11], with its 159841 trainable parameters, evidently requires many more multiplications as well as dramatically more memory.

The DL-based detector was trained on ECG segments whose quality was determined from sample skewness and kurtosis [46]. More recently, other approaches to quality assessment have been proposed designed specifically for use in AF detection [73]–[76]. These approaches may lead to better training results and, ultimately, better detection performance.

Modern sensor technology have helped form a new paradigm of long-term AF monitoring relying on the analysis of photoplethysmographic (PPG) signals. As a result, a large number of PPG-based AF detectors have been published, e.g., [19], [30], [63], [65]–[71], [23]. While PPG-based detection was not addressed in the present paper, the considerations made on performance evaluation of ECG-based detectors apply to PPG-based detectors as well.

VII. Recommendations

From the implications of the results as well as from reviewing a large number of recent, peer-reviewed papers, the present study leads up to the following five recommendations on evaluating detector performance:

1) To use different datasets for training and testing, and to ensure that the two datasets do not contain signals from the same patient.
2) To substantiate the approach taken to annotation comparison and, if applicable, report segment length and episode overlap.
3) To evaluate performance in terms of *Mcc*, rather than *Acc* or F_1, and to include *Se*, *Sp*, and *PPV* so as to facilitate a comparison to the many published detectors; AUC should be left out when reporting performance.
4) To evaluate the influence of physiological and technical factors on performance, including lead selection, APB rate, noise level, and AF-masquerading arrhythmias.
5) To pay special attention to detection performance when the aim is to characterize AF episode patterns.

Acknowledgment

The authors would like to thank Pablo Laguna, University of Zaragoza, Spain, for helpful comments.

References

[1] W. Cai *et al.*, "Accurate detection of atrial fibrillation from 12-lead ECG using deep neural network," *Comput. Biol. Med.*, vol. 116, 2020, Art. no. 103378. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2019.103378

[2] H. Shi *et al.*, "An incremental learning system for atrial fibrillation detection based on transfer learning and active learning," *Comput. Methods Programs Biomed.*, vol. 187, 2020, Art. no. 105219. [Online]. Available: https://doi.org/10.1016/j.cmpb.2019.105219

[3] S. K. Ghosh *et al.*, "Detection of atrial fibrillation from single lead ECG signal using multirate cosine filter bank and deep neural network," *J. Med. Syst.*, vol. 44, no. 114, pp. 1–15, 2020. [Online]. Available: https://doi.org/10.1007/s10916-020-01565-y

[4] H. Zhang *et al.*, "SS-SWT and SI-CNN: An atrial fibrillation detection framework for time-frequency ECG signal," *J. Healthc. Eng.*, vol. 2020, 2020, Art. no. 7526825. [Online]. Available: https://doi.org/10.1155/2020/7526825

[5] M.-L. Huang and Y.-S. Wu, "Classification of atrial fibrillation and normal sinus rhythm based on convolutional neural network," *Biomed. Eng. Lett.*, vol. 10, pp. 183–193, 2020. [Online]. Available: https://doi.org/10.1007/s13534-020-00146-9

[6] Y. Jin *et al.*, "Multi-domain modeling of atrial fibrillation detection with twin attentional convolutional long short-term memory neural networks," *Knowl.-Based Syst.*, vol. 193, 2020, Art. no. 105460. [Online]. Available: https://doi.org/10.1016/j.knosys.2019.105460

[7] J. Wang, "A deep learning approach for atrial fibrillation signals classification based on convolutional and modified elman neural network," *Future Gener. Comput. Syst.*, vol. 102, pp. 670–679, 2020. [Online]. Available: https://doi.org/10.1016/j.future.2019.09.012

[8] H. Fujita and D. Cimr, "Computer aided detection for fibrillations and flutters using deep convolutional neural network," *Inf. Sci.*, vol. 486, pp. 231–239, 2019. [Online]. Available: https://doi.org/10.1016/j.ins.2019.02.065

[9] H. Dang *et al.*, "A novel deep arrhythmia-diagnosis network for atrial fibrillation classification using electrocardiogram signals," *IEEE Access*, vol. 7, pp. 75 577–75 590, 2019. [Online]. Available: https://doi.org/10.1109/ACCESS.2019.2918792

[10] D. Lai *et al.*, "An automatic system for real-time identifying atrial fibrillation by using a lightweight convolutional neural network," *IEEE Access*, vol. 7, pp. 130 074–130 084, 2019. [Online]. Available: https://doi.org/10.1109/ACCESS.2019.2939822

[11] R. S. Andersen, A. Peimankar, and S. Puthusserypady, "A deep learning approach for real-time detection of atrial fibrillation," *Expert Syst. Appl.*, vol. 115, pp. 465–473, 2019. [Online]. Available: https://doi.org/10.1016/j.eswa.2018.08.011

[12] A. Y. Hannun *et al.*, "Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network," *Nat. Med.*, vol. 25, no. 1, pp. 65–69, 2019. [Online]. Available: https://doi.org/10.1038/s41591-018-0268-3

[13] R. He *et al.*, "Automatic detection of atrial fibrillation based on continuous wavelet transform and 2D convolutional neural networks," *Front. Physiol.*, vol. 9, 2018, Art. no. 1206. [Online]. Available: https://doi.org/10.3389/fphys.2018.01206

[14] O. Faust *et al.*, "Automated detection of atrial fibrillation using long short-term memory network with RR interval signals," *Comput. Biol. Med.*, vol. 102, pp. 327–335, 2018. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2018.07.001

[15] Y. Xia *et al.*, "Detecting atrial fibrillation by deep convolutional neural networks," *Comput. Biol. Med.*, vol. 93, pp. 84–92, 2018. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2017.12.007

[16] B. Pourbabaee, M. J. Roshtkhari, and K. Khorasani, "Deep convolutional neural networks and learning ECG features for screening paroxysmal atrial fibrillation patients," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 48, no. 12, pp. 2095–2104, Dec. 2018. [Online]. Available: https://doi.org/10.1109/TSMC.2017.2705582

[17] S. Mousavi, F. Afghah, and U. R. Acharya, "HAN-ECG: An interpretable atrial fibrillation detection model using hierarchical attention networks," *Comput. Biol. Med.*, vol. 127, 2020, Art. no. 104057. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2020.104057

[18] I. A. Marsili *et al.*, "Implementation and validation of real-time algorithms for atrial fibrillation detection on a wearable ECG device," *Comput. Biol. Med.*, vol. 116, 2020, Art. no. 103540. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2019.103540

[19] A. Sološenko *et al.*, "Detection of atrial fibrillation using a wrist-worn device," *Physiol. Meas.*, vol. 40, no. 2, 2019, Art. no. 025003. [Online]. Available: https://doi.org/10.1088/1361-6579/ab029c

[20] J. Ródenas *et al.*, "Wavelet entropy automatically detects episodes of atrial fibrillation from single-lead electrocardiograms," *Math 1*, vol. 17, no. 9, pp. 6179–6199, 2015. [Online]. Available: https://doi.org/10.3390/e17096179

[21] 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. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2015.01.019

[22] S. Ladavich and B. Ghoraani, "Rate-independent detection of atrial fibrillation by statistical modeling of atrial activity," *Biomed. Signal Process. Control*, vol. 18, pp. 274–281, 2015. [Online]. Available: https://doi.org/10.1016/j.bspc.2015.01.007

[23] J. Lee *et al.*, "Atrial fibrillation detection using an iPhone 4S," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 1, pp. 203–206, 2013. [Online]. Available: https://doi.org/10.1109/TBME.2012.2208112

[24] K. Jiang *et al.*, "High accuracy in automatic detection of atrial fibrillation for holter monitoring," *J. Zhejiang Univ. Sci. B*, vol. 13, no. 9, pp. 751–756, 2012. [Online]. Available: https://doi.org/10.1631/jzus.B1200107

[25] S. Asgari, A. Mehrnia, and M. Moussavi, "Automatic detection of atrial fibrillation using stationary wavelet transform and support vector machine," *Comput. Biol. Med.*, vol. 60, pp. 132–142, 2015. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2015.03.005

[26] J. Lee *et al.*, "Time-varying coherence function for atrial fibrillation detection," *IEEE Trans. Biomed. Eng.*, vol. 60, no. 10, pp. 2783–2793, 2013. [Online]. Available: https://doi.org/10.1109/TBME.2013.2264721

[27] J. Lian, L. Wang, and D. Muessig, "A simple method to detect atrial fibrillation using RR intervals," *Amer. J. Cardiol.*, vol. 107, no. 10, pp. 1494–1497, 2011. [Online]. Available: https://doi.org/10.1016/j.amjcard.2011.01.028

[28] S. Babaeizadeh *et al.*, "Improvements in atrial fibrillation detection for real-time monitoring," *J. Electrocardiol.*, vol. 42, no. 6, pp. 522–526, 2009. [Online]. Available: https://doi.org/10.1016/j.jelectrocard.2009.06.006

[29] S. Dash *et al.*, "Automatic real time detection of atrial fibrillation," *Ann. Biomed. Eng.*, vol. 37, no. 9, pp. 1701–1709, 2009. [Online]. Available: https://doi.org/10.1007/s10439-009-9740-z

[30] J. Wasserlauf *et al.*, "Smartwatch performance for the detection and quantification of atrial fibrillation," *Circ. Arrhythm. Electrophysiol.*, vol. 12, no. 6, 2019, Art. no. e00 6834. [Online]. Available: https://doi.org/10.1161/CIRCEP.118.006834

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

[32] A. Petrėnas *et al.*, "Electrocardiogram modeling during paroxysmal atrial fibrillation: Application to the detection of brief episodes," *Physiol. Meas.*, vol. 38, no. 11, pp. 2058–2080, 2017. [Online]. Available: https://doi.org/10.1088/1361-6579/aa9153

[33] A. Sološenko *et al.*, "Modeling of the photoplethysmogram during atrial fibrillation," *Comput. Biol. Med.*, vol. 81, pp. 130–138, 2017. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2016.12.016

[34] R. C. Seet, P. A. Friedman, and A. A. Rabinstein, "Prolonged rhythm monitoring for the detection of occult paroxysmal atrial fibrillation in ischemic stroke of unknown cause," *Circulation*, vol. 124, no. 4, pp. 477–486, 2011. [Online]. Available: https://doi.org/10.1161/CIRCULATIONAHA.111.029801

[35] A. Kishore *et al.*, "Detection of atrial fibrillation after ischemic stroke or transient ischemic attack: A systematic review and meta-analysis," *Stroke*, vol. 45, no. 2, pp. 520–526, 2014. [Online]. Available: https://doi.org/10.1161/STROKEAHA.113.003433

[36] R. Mahajan *et al.*, "Subclinical device-detected atrial fibrillation and stroke risk: A systematic review and meta-analysis," *Eur. Heart J.*, vol. 39, no. 16, pp. 1407–1415, 2018. [Online]. Available: https://doi.org/10.1093/eurheartj/ehx731

[37] 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, 2021 [Online]. Available: https://doi.org/10.1109/TBME.2020.2995563

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

[39] 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, 2021. [Online]. Available: https://doi.org/10.1055/s-0040-1716408

[40] 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, 2021.

[41] 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. e 215–e220, 2000. [Online]. Available: https://doi.org/10.1161/01.cir.101.23.e215

[42] 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, Apr. 2004. [Online]. Available: https://doi.org/10.1109/TBME.2003.821031

[43] A. Petrėnas *et al.*, "An echo state neural network for QRST cancellation during atrial fibrillation," *IEEE Trans. Biomed. Eng.*, vol. 59, no. 10, pp. 2950–2957, 2012. [Online]. Available: https://doi.org/10.1109/TBME.2012.2212895

[44] U. Erdenebayar *et al.*, "Automatic prediction of atrial fibrillation based on convolutional neural network using a short-term normal electrocardiogram signal," *J. Korean Med. Sci.*, vol. 34, no. 7, 2019, Art. no. e 64. [Online]. Available: https://doi.org/10.3346/jkms.2019.34.e64

[45] D. Kingma and J. Ba., "Adam: A method for stochastic optimization," 2014. [Online]. Available: https://arxiv.org/abs/1412.6980

[46] A. Ghaffari *et al.*, "Segmentation of holter ECG waves via analysis of a discrete wavelet-derived multiple skewness-kurtosis based metric," *Ann. Biomed. Eng.*, vol. 38, no. 4, pp. 1497–510, 2010. [Online]. Available: https://doi.org/10.1007/s10439-010-9919-3

[47] D. Chicco and G. Jurman, "The advantages of the matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation," *BMC Genomics*, vol. 21, no. 6, pp. 1–13, 2020. [Online]. Available: https://doi.org/10.1186/s12864-019-6413-7

[48] S. W. Baalman *et al.*, "A morphology based deep learning model for atrial fibrillation detection using single cycle electrocardiographic samples," *Int. J. Cardiol.*, vol. 316, pp. 130–136, 2020. [Online]. Available: https://doi.org/10.1016/j.ijcard.2020.04.046

[49] O. Andersson *et al.*, "A 290 mV Sub-V_T ASIC for real-time atrial fibrillation detection," *IEEE Trans. Biomed. Circuits Syst.*, vol. 9, no. 3, pp. 377–386, 2015. [Online]. Available: https://doi.org/10.1109/TBCAS.2014.2354054

[50] C. Huang *et al.*, "A novel method for detection of the transition between atrial fibrillation and sinus rhythm," *IEEE Trans. Biomed. Eng.*, vol. 58, no. 4, pp. 1113–1119, 2011. [Online]. Available: https://doi.org/10.1109/TBME.2010.2096506

[51] M. Rizwan, B. M. Whitaker, and D. V. Anderson, "AF detection from ECG recordings using feature selection, sparse coding, and ensemble learning," *Physiol. Meas.*, vol. 39, no. 12, 2018, Art. no. 124007. [Online]. Available: https://doi.org/10.1088/1361-6579/aaf35b

[52] M. Shao *et al.*, "Detection of atrial fibrillation from ECG recordings using decision tree ensemble with multi-level features," *Physiol. Meas.*, vol. 39, no. 9, 2018, Art. no. 094008. [Online]. Available: https://doi.org/10.1088/1361-6579/aadf48

[53] Q. Gu, L. Zhu, and Z. Cai, "Evaluation measures of the classification performance of imbalanced data sets," in *Proc. Intern. Symp. Intell. Comput. Applic.*, vol. 51, pp. 461–471, 2009. [Online]. Available: https://doi.org/10.1007/978-3-642-04962-0

[54] M. Bekkar, H. K. Djemaa, and T. A. Alitouche, "Evaluation measures for models assessment over imbalanced data sets," *J. Inf. Eng. Appl.*, vol. 3, no. 10, pp. 27–38, 2013.

[55] X. Zhou *et al.*, "Automatic online detection of atrial fibrillation based on symbolic dynamics and shannon entropy," *Biomed. Eng. Online*, vol. 13, no. 18, pp. 1–18, 2014. [Online]. Available: https://doi.org/10.1186/1475-925X-13-18

[56] P. Langley *et al.*, "Accuracy of algorithms for detection of atrial fibrillation from short duration beat interval recordings," *Med. Eng. Phys.*, vol. 34, no. 10, pp. 1441–1447, 2012. [Online]. Available: https://doi.org/10.1016/j.medengphy.2012.02.002

[57] J. M. Lobo, A. Jiménez-Valverde, and R. Real, "AUC: A misleading measure of the performance of predictive distribution models," *Glob. Ecol. Biogeogr.*, vol. 17, no. 2, pp. 145–151, 2008. [Online]. Available: https://doi.org/10.1111/j.1466-8238.2007.00358.x

[58] B. Hanczar *et al.*, "Small-sample precision of ROC-related estimates," *Bioinformatics*, vol. 26, no. 6, pp. 822–830, 2010. [Online]. Available: https://doi.org/10.1093/bioinformatics/btq037

[59] D. E. Lake and J. R. Moorman, "Accurate estimation of entropy in very short physiological time series: The problem of atrial fibrillation detection in implanted ventricular devices," *Am J. Physiol. Heart Circ. Physiol.*, vol. 300, no. 1, pp. H 319–H25, 2011. [Online]. Available: https://doi.org/10.1152/ajpheart.00561.2010

[60] R. B. Shouldice, C. Heneghan, and P. de Chazal, "Automatic detection of paroxysmal atrial fibrillation," in *Atrial Fibrillation - Basic Research and Clinical Applications*, J.-I. Choi, Ed. London, U.K.: InTechOpen Limited, 2012, ch. 7, pp. 125–146. [Online]. Available: https://doi.org/10.5772/26860

[61] P. Carvalho *et al.*, "Model-based atrial fibrillation detection," in *ECG Signal Processing, Classification and Interpretation*, A. Gacek and W. Pedrycz, Eds. London, U.K.: Springer, 2012, pp. 99–133. [Online]. Available: https://doi.org/10.1007/978-0-85729-868-3

[62] X. Zhou *et al.*, "A real-time atrial fibrillation detection algorithm based on the instantaneous state of heart rate," *PLoS ONE*, vol. 10, no. 9, 2015, Art. no. e0 136544. [Online]. Available: https://doi.org/10.1371/journal.pone.0136544

[63] T. Pereira *et al.*, "Photoplethysmography based atrial fibrillation detection: A review," *NPJ Digit. Med.*, vol. 3, no. 1, pp. 1–12, 2020. [Online]. Available: https://doi.org/10.1038/s41746-019-0207-9

[64] L. Sörnmo *et al.*, "Letter regarding the article 'Detecting Atrial Fibrillation by Deep Convolutional Neural Networks by X. et al.'," *Comput. Biol. Med.*, vol. 100, pp. 41–42, 2018. [Online]. Available: https://doi.org/10.1016/j.compbiomed.2018.06.027

[65] S. K. Bashar *et al.*, "Atrial fibrillation detection from wrist photoplethysmography signals using smartwatches," *Sci. Rep.*, vol. 9, no. 1, pp. 1–10, 2019. [Online]. Available: https://doi.org/10.1038/s41598-019-49092-2

[66] S. Fallet *et al.*, "Can one detect atrial fibrillation using a wrist-type photoplethysmographic device?" *Med. Biol. Eng. Comput.*, vol. 57, no. 2, pp. 477–487, 2019. [Online]. Available: https://doi.org/10.1007/s11517-018-1886-0

[67] S. Kwon *et al.*, "Deep learning approaches to detect atrial fibrillation using photoplethysmographic signals: Algorithms development study," *JMIR Mhealth Uhealth*, vol. 7, no. 6, 2019, Art. no. e12770. [Online]. Available: https://doi.org/10.2196/12770

[68] G. H. Tison *et al.*, "Passive detection of atrial fibrillation using a commercially available smartwatch," *JAMA Cardiol.*, vol. 3, no. 5, pp. 409–416, 2018. [Online]. Available: https://doi.org/10.1001/jamacardio.2018.0136

[69] A. G. Bonomi *et al.*, "Atrial fibrillation detection using a novel cardiac ambulatory monitor based on photo-plethysmography at the wrist," *J. Amer. Heart Assoc.*, vol. 7, no. 15, 2018, Art. no. e00 9351. [Online]. Available: https://doi.org/10.1161/JAHA.118.009351

[70] L. M. Eerikäinen *et al.*, "Comparison between electrocardiogram- and photoplethysmogram-derived features for atrial fibrillation detection in free-living conditions," *Physiol. Meas.*, vol. 39, no. 8, 2018, Art. no. 084001. [Online]. Available: https://doi.org/10.1088/1361-6579/aad2c0

[71] V. D. A. Corino *et al.*, "Detection of atrial fibrillation episodes using a wristband device," *Physiol. Meas.*, vol. 38, no. 5, pp. 787–799, 2017. [Online]. Available: https://doi.org/10.1088/1361-6579/aa5dd7

[72] L. Sörnmo Ed., *Atrial Fibrillation From an Engineering Perspective*, Berlin, Germany: Springer, 2018. [Online]. Available: https://doi.org/10.1007/978-3-319-68515-1

[73] J. Oster and G. D. Clifford, "Impact of the presence of noise on RR interval-based atrial fibrillation detection," *J. Electrocardiol.*, vol. 48, no. 6, pp. 947–951, 2015. [Online]. Available: https://doi.org/10.1016/j.jelectrocard.2015.08.01

[74] B. Taji, A. D. C. Chan, and S. Shirmohammadi, "False alarm reduction in atrial fibrillation detection using deep belief networks," *IEEE Trans. Instrum. Meas.*, vol. 67, no. 5, pp. 1124–1131, 2018. [Online]. Available: https://doi.org/10.1109/TIM.2017.2769198

[75] M. Henriksson *et al.*, "Model-based assessment of f-wave signal quality in patients with atrial fibrillation," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 11, pp. 2600–2611, 2018. [Online]. Available: https://doi.org/10.1109/TBME.2018.2810508

[76] S. K. Bashar *et al.*, "Noise detection in electrocardiogram signals for intensive care unit patients," *IEEE Access*, vol. 7, pp. 88357–88368, 2018. [Online]. Available: https://doi.org/10.1109/ACCESS.2019.2926199