Computers in Biology and Medicine 133 (2021) 104404

Contents lists available at ScienceDirect

Computers in Biology and Medicine

journal homepage: www.elsevier.com/locate/compbiomed

Digital biomarkers and algorithms for detection of atrial fibrillation using surface electrocardiograms: A systematic review

Fons J. Wesselius [a], Mathijs S. van Schie [a], Natasja M.S. De Groot [a,*], Richard C. Hendriks [b]

[a] Department of Cardiology, Erasmus Medical Center, Rotterdam, the Netherlands
[b] Faculty of Electrical Engineering, Mathematics and Computer Science, Delft University of Technology, Delft, the Netherlands

ARTICLE INFO

Keywords:
Atrial fibrillation
ECG signal Processing
Telemetry
Machine learning
Algorithms
Classification

ABSTRACT

Aims: Automated detection of atrial fibrillation (AF) in continuous rhythm registrations is essential in order to prevent complications and optimize treatment of AF. Many algorithms have been developed to detect AF in surface electrocardiograms (ECGs) during the past few years. The aim of this systematic review is to gain more insight into these available classification methods by discussing previously used digital biomarkers and algorithms and make recommendations for future research.

Methods: On the 14th of September 2020, the PubMed database was searched for articles focusing on algorithms for AF detection in ECGs using the MeSH terms *Atrial Fibrillation*, *Electrocardiography* and *Algorithms*. Articles which solely focused on differentiation of types of rhythm disorders or prediction of AF termination were excluded.

Results: The search resulted in 451 articles, of which 130 remained after full-text screening. Not only did the amount of research on methods for AF detection increase over the past years, but a trend towards more complex classification methods is observed. Furthermore, three different types of features can be distinguished: atrial features, ventricular features, and signal features. Although AF is an atrial disease, only 22% of the described methods use atrial features.

Conclusion: More and more studies focus on improving accuracy of classification methods for AF in ECGs. As a result, algorithms become increasingly complex and less well interpretable. Only a few studies focus on detecting atrial activity in the ECG. Developing innovative methods focusing on detection of atrial activity might provide accurate classifiers without compromising on transparency.

1. Introduction

Accurate detection of atrial fibrillation (AF) episodes in continuous rhythm registrations is essential in order to prevent complications and optimize treatment of AF. However, manual analysis of continuous rhythm registrations is time-consuming. For this reason, the amount of research focused on automated AF detection has increased over the past years.

Automated analysis of continuous rhythm registrations not only helps in detecting AF for treatment considerations, but also provides insightful data for research on the still not entirely unraveled mechanisms underlying AF. Available research mainly focusses on the presence or absence of AF in patients. Most often, differentiations in AF burden are made using the classification between paroxysmal, persistent, long-standing persistent or permanent AF. [1] As stated by Chen et al., more comprehensive information might be obtained by describing AF burden in terms of duration, number of episodes and/or proportion of time an individual is in AF during a monitoring period. [2] Accurate automated AF detection in continuous rhythm registrations is essential to acquire these measures from long-term electrocardiogram (ECG) readings.

Already numerous methods for optimal automated AF detection in ECGs have been proposed, making it more and more difficult to see the wood for the trees. Instead of proposing yet another new algorithm, with this review we aimed to provide an in-depth overview of the previously used algorithms and identify methodological gaps which potentially can be used to develop novel innovative AF detection algorithms.

* Corresponding author. Unit Translational Electrophysiology, Department of Cardiology Erasmus Medical Center Doctor, Molewaterplein 40 3015 GD, Rotterdam, the Netherlands.
E-mail address: n.m.s.degroot@erasmusmc.nl (N.M.S. De Groot).

https://doi.org/10.1016/j.compbiomed.2021.104404
Received 1 March 2021; Received in revised form 12 April 2021; Accepted 12 April 2021
Available online 15 April 2021

2. Search strategy

On the 14th of September 2020, the PubMed database was searched for articles focusing on algorithms for AF detection in ECG readings using the MeSH terms *Atrial Fibrillation*, *Electrocardiography* and *Algorithms*. Not all articles were indexed using MeSH terms, hence articles were also included if any of these terms or their synonyms were mentioned in the title or abstract. Additionally, title and abstract were screened for the terms *('detect*' or 'classifi*' or 'predict*')* and *('accuracy' or 'performance' or 'F1' or 'sensitivity' or 'specificity' or 'positive predictive value' or 'negative predictive value')*. Full search strategy is provided in Supplementary Appendix A. Only articles focusing on AF detection in humans were included. Articles which solely focused on differentiation of types of rhythm disorders or prediction of AF termination were excluded. Additional exclusion criteria are listed in Table 1.

The search resulted in 451 articles, of which 263 were excluded based on title and abstract. An additional 10 articles were not available in English and of 3 records no full-text was available. As summarized in Fig. 1, the remaining 175 full-text articles were screened for eligibility using the inclusion and exclusion criteria. A total of 130 articles remained after selection.

3. Study characteristics

Supplementary Appendix B lists all included studies with the used categories of classification methods and features and the classifier accuracy. As visualized in the upper panel of Fig. 2, research on methods for AF detection has increased over the past ten years from 20 studies up until 2009 to 108 studies up until 2019. In 2018, results from the PhysioNet/Computing in Cardiology (CinC) Challenge 2017 caused a strong increase in the number of studies (n = 30). [3] Most used databases for development and testing of AF detection algorithms are the Massachusetts Institute of Technology–Beth Israel Hospital (MIT-BIH) Arrhythmia Database [4,5], MIT-BIH Atrial Fibrillation Database [5,6], MIT-BIH Normal Sinus Rhythm Database [5], and the database used in the PhysioNet/CinC Challenge 2017 [5].

4. Classification methods for atrial fibrillation

Over the past years, not only *more research* has been conducted on the development of automated detection algorithms for AF in ECGs, but also *new strategies* have been applied. As visualized in the lower panel of Fig. 2, used classification methods can be grouped into six main categories: rule-based classification, decision tree(s), k-nearest neighbor (k-NN) classification, regression analysis, support vector machines (SVMs), and neural networks (NNs). Whereas older studies use more straightforward rule-based approaches, newer studies increasingly use NNs and SVMs, providing new ways of describing ECGs and improving the classification accuracy. In contrast to rule-based classifiers, which can be easily interpreted, these more complex methods appear more as a 'black box', hence decisions made by these classifiers are difficult to comprehend. Therefore, transparency and accuracy of the classifier should be carefully balanced.

Table 1
Inclusion and exclusion criteria.

Inclusion criteria	Exclusion criteria
Focus on automated AF detection in human using ECG	Focus on differentiation of types of atrial arrhythmias
(Clear) description of used algorithms	Focus on prediction of AF termination
Reporting performance measures	(Systematic) reviews
	Non-English articles

AF indicates atrial fibrillation; ECG indicates electrocardiogram.

4.1. Classification performance

A commonly used measure for accuracy of AF detection is the F1-score, which is calculated as the harmonic mean of the recall (i.e. sensitivity) and precision (i.e. positive predictive value). [3] The median F1-score of all studies was 94.0% [interquartile range (IQR): 93.1%–97.7%]. As can be seen from several studies testing their classifier on multiple databases, the chosen database has major influence on the achieved performance measures. Zhou. et al. showed a positive predictive value (PPV) of 92.3% for testing on the combination of the MIT-BIH Normal Sinus Rhythm Database and MIT-BIH Atrial Fibrillation Database, but a PPV of only 55.3% when testing on the MIT-BIH Arrhythmia Database. [7] Multiple other studies show similar variations in performance measures depending on the used testing database. [8–12]

5. Features of ECGs with AF

In general, classification methods require an input vector which contains features describing the ECG signal. Some features are derived from standard clinical protocols for ECG interpretation, for example P-wave presence, regularity of QRS-complexes, QRS-width and QT-time. Whilst these features are easy to interpret, more complex features can be used to further analyze and describe the ECG signal. NNs do not necessarily require preprocessing to extract features, since this method also allows raw ECG input. [13–26]

In total, 131 feature sets were described in 130 studies, where one study trained and validated a classifier with two different feature sets. On the ECG, AF is visually characterized by absence of P-waves or presence of f-waves, combined with irregular time intervals between QRS-complexes as a result of disorganized irregular atrial impulses activating the atrioventricular node. [1] Many studies use these characteristics in their classification method. As visualized in Fig. 3, features can be categorized into *atrial features*, *ventricular features*, and *signal features*.

For the AF classifiers, *atrial features* mainly focus on P-wave disappearance or f-wave appearance, *ventricular features* include mainly features describing irregularity of intervals between subsequent R-peaks (RR-intervals), and *signal features* describe further characteristics of the signal which cannot easily be related to cardiac electrophysiological characteristics and the clinical presentation and pathophysiology of AF (e.g. signal quality and frequency components).

Although AF is an atrial rhythm disorder, only 29 methods (22%) focus on atrial features while almost two third of the studies use ventricular features (86 methods, 66%), as shown in Fig. 3. Also, more than half of the studies use signal features (67 methods, 51%). More specifically, 19 methods (15%) used a combination of two categories of features; mostly ventricular features with either atrial or signal features (11 methods, 8%, and 7 methods, 5%, respectively). Only 1 method (1%) used a combination of atrial and signal features, without including ventricular features. A combined approach using all three categories of features was used in 16 methods (12%). As presented in Table 2, median F1-score was highest for classifiers focusing on ventricular features only (96.9% [IQR: 92.9%–98.1%]).

5.1. Atrial features

The underrepresentation of atrial features can be partially attributed to the challenging detection of atrial activity in the ECG due to relatively low amplitudes, diversity of waveforms and signal artefacts. [27] Therefore, especially in noisy signals, atrial activity detection is complex. Still, some methods incorporate atrial features to describe P-wave disappearance and f-wave appearance, or to describe the atrial wave morphology in general.

5.1.1. P-wave disappearance and f-wave appearance

Being among the most prominent features of an ECG during AF,

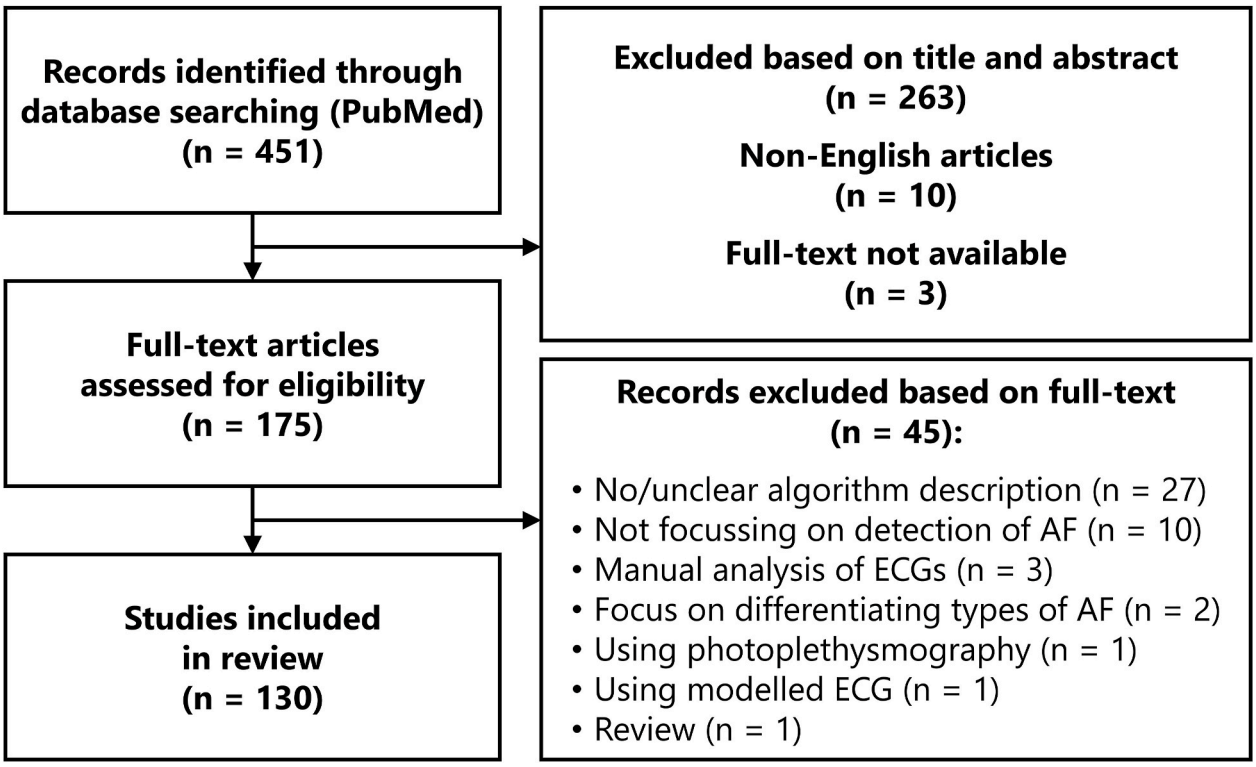

Fig. 1. Flowchart demonstrating selection of studies aiming to develop automated detection algorithms for atrial fibrillation (AF) in electrocardiograms (ECGs).

absence of P-waves and presence of f-waves are described by multiple methods. In the time-domain, the number of fluctuations in the TQ-interval is used to detect a pattern of quick changes in the atrial signal. [28,29] In the frequency domain, f-waves mainly result in peaks in the 4–10Hz frequency band, hence the area under the power spectral density curve within this frequency band is compared to the total area under the power spectral density curve as a measure of f-wave appearance. [30]

5.1.2. *Atrial wave morphology*

Amplitude of P-waves and P-wave duration directly reflect the electrophysiological characteristics of the atria. Furthermore, the time interval between atrial and ventricular activity, and the number of P-waves relative to the number of QRS-complexes provide information on the conduction from atria to ventricles. [31] Also, the time between P-waves is calculated as a measure of the atrial rate. [32]

Since P-waves are absent or have transformed into f-waves during AF, extracting basic morphological features is not always straightforward. Using computer algorithms, more complex morphological features are extracted, describing the statistics of the signal in terms of statistical measures like root mean square (RMS)-value, variance, skewness and kurtosis. [33,34]

5.2. *Ventricular features*

In contrast to the detection of atrial activity, automated detection of ventricular activity is more straightforward due to the more pronounced QRS-complexes in the ECG signal. Already in 1985, Pan and Tompkins proposed a QRS detection algorithm, which is still widely used in research. [35] In 2018, Liu et al. published a comparison between ten common automated QRS-detectors using more than 2 million beats. [36] From each detector, the accuracy was estimated in terms of an F1-score. Using a dataset containing high-quality ECG signals, all algorithms resulted in F1-scores larger than 99%. However, algorithms were highly dependent on the signal quality, as F1-scores decreased more than 25% when using ECG signals with the least optimal signal quality. Since ventricular features are dependent on the detection of ventricular activity, this relation between signal quality and detection rate has a direct effect on the accuracy of classifiers using these features.

Common features during manual analysis of ECGs are statistics of peak intervals, which reflect the propagation speed of cardiac activity. Furthermore, ventricular wave morphology is described in terms of durations and amplitudes of the QRS-complexes and T-waves. Additionally, computer algorithms are capable of processing more complex morphological features, focusing for example on the ratio between the amplitude of ventricular activity and atrial activity, statistical features of QRS-complexes, and correlation between beats. [31,34] Although these features are useful to describe the average ventricular activation in general, methods implementing ventricular characteristics for AF detection mainly focus on irregularity of RR-intervals, since this is an evident feature of AF and one of the main features used in clinical practice. Various methods to describe the variability in RR-intervals are used, ranging from calculation of common statistics (e.g. standard deviation (SD)) to more complex statistics (e.g. entropy).

5.2.1. *Standard deviation, coefficient of variation and RMSSD of RR-intervals*

A basic measure for variation is SD, but as pointed out by Sacha et al. in their review concerning the interaction between heart rate and heart rate variability, higher heart rates are associated with lower variance in RR-intervals. [37] With increasing heart rate, RR-intervals become smaller, hence variation in heart rate will have less effect on variation in RR-intervals. Therefore, the SD of RR-intervals should be divided by the average RR-interval to correct for differences in heart rate, resulting in the coefficient of variance. [38]

Another commonly used measure for variability in RR-intervals is the root mean square of successive differences (RMSSD). Instead of globally analyzing the RR-intervals, this method describes the average change in successive RR-intervals, hence this method is less prone to slowly changing RR-intervals, which would result in a higher overall SD of RR-

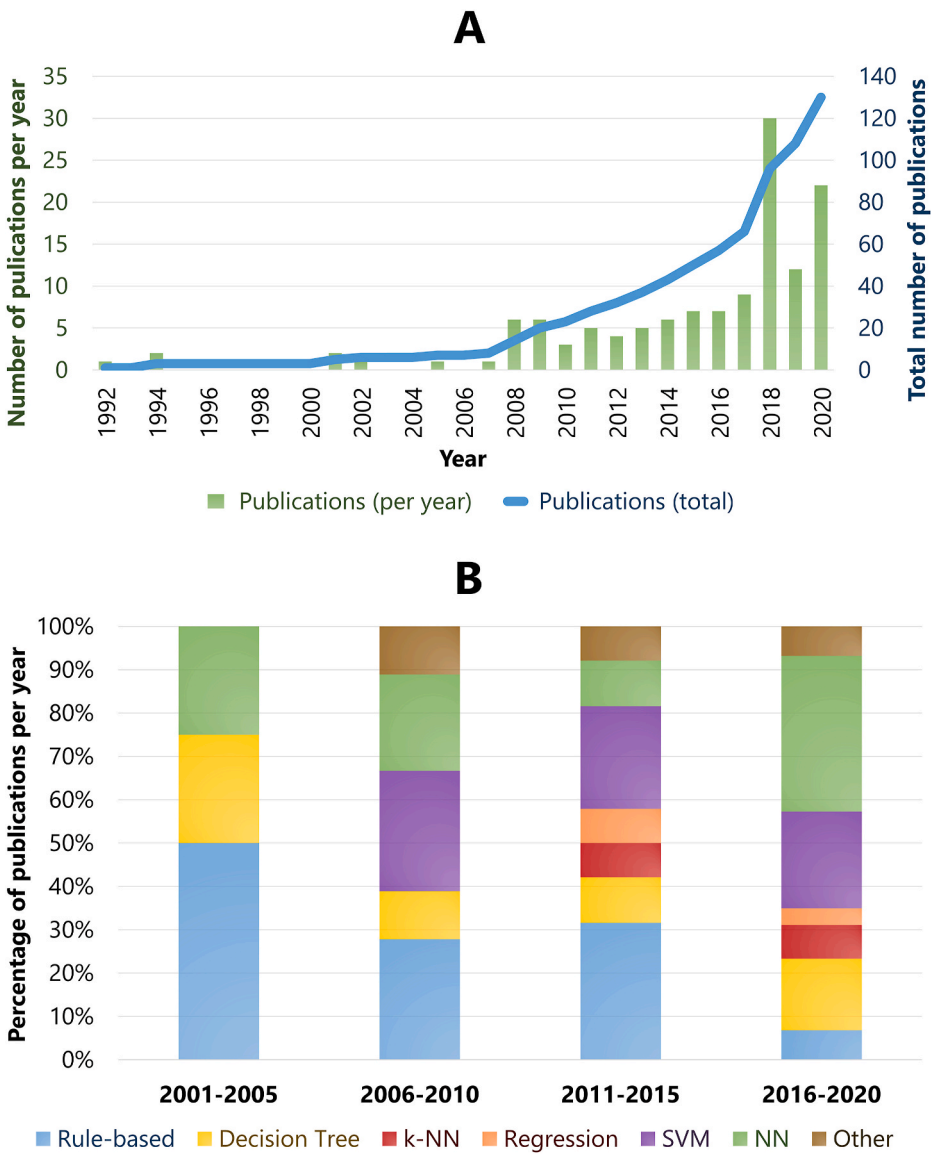

Fig. 2. A. The number of published articles on detection of atrial fibrillation per year. The bar chart indicates the number of published articles in a specific year, and the line chart shows the cumulative number of published articles. A steady increase in the number of published articles is observed. The sudden increase in 2018 is most probably caused by the PhysioNet/Computing in Cardiology Challenge (CinC) 2017.
B. Bar chart per period of five years demonstrating the percentage of published articles using neural networks (NN), support vector machines (SVM), rule-based classifiers, decision tree(s), regression analysis, k-nearest neighbors (k-NN) classifiers, and other methods. A shift in used classification methods towards neural networks and support vector machines is observed. Articles published before 2001 were excluded from the chart due to the low amount of studies (n = 3).

intervals. [39]

Additionally, the percentage of successive differences between RR-intervals which differ more than a certain amount of time is used to describe the RR-interval variability. [40] Commonly used thresholds are 5 ms, 10 ms and 50 ms, but any arbitrary threshold can be chosen.

5.2.2. Poincaré or Lorenz plots of RR-intervals

Similar to RMSSD, Poincaré or Lorenz plots are used to analyze *successive* RR-intervals. [41–48] However, instead of directly calculating a measure, the variability is visualized by plotting RR-intervals in a two-dimensional plane where the x-axis represents an RR-interval (RR_i) and the y-axis represents the subsequent RR-interval (RR_{i+1}), as visualized in Fig. 4. During AF, when variation of successive RR-intervals is higher compared to sinus rhythm (SR), data points are dispersed across a larger area around the average RR-interval. Furthermore, Park et al. show that patterns in Poincaré plots might be a useful feature to discriminate between AF and other arrhythmias resulting in irregular RR-intervals (e.g. premature ventricular beats). [47] Commonly used features from this graphical representation are the SD of distances from data points to the line perpendicular to the regression line where RR_i equals RR_{i+1} and the SD of distances from data points to the regression line itself.

5.2.3. Entropy measures of RR-intervals

The entropy of an RR-intervals series is another measure for the RR-interval irregularity. We describe two common definitions of entropy: the sample entropy and the Shannon entropy. [49] The sample entropy is used to describe the complexity of a time series, while the Shannon entropy expresses the amount of information or uncertainty.

The sample entropy of an RR-intervals series is defined as the probability that two matching RR-interval series will continue to match at the next RR-interval. [50] A match is defined as two RR-interval segments having corresponding data points within a certain small range, described by the tolerance factor r, as schematically visualized in Fig. 5. When two matching RR-interval series do not continue to match for the next RR-interval, the sample entropy increases, hence a higher sample entropy reflects that the next RR-interval is less predictable, i.e. indicates a higher variability of the signal.

The sample entropy is highly dependent on the tolerance factor, since the probability of two segments of RR-intervals matching increases with increasing tolerance factors. Therefore, Lake et al. proposed to correct the sample entropy by subtracting$\ln(2r)$. [50] Furthermore, they observed that heart rate and sample entropy add independent information to detect AF, hence proposed an optimization of the sample entropy for AF detection by subtracting the natural logarithm of the

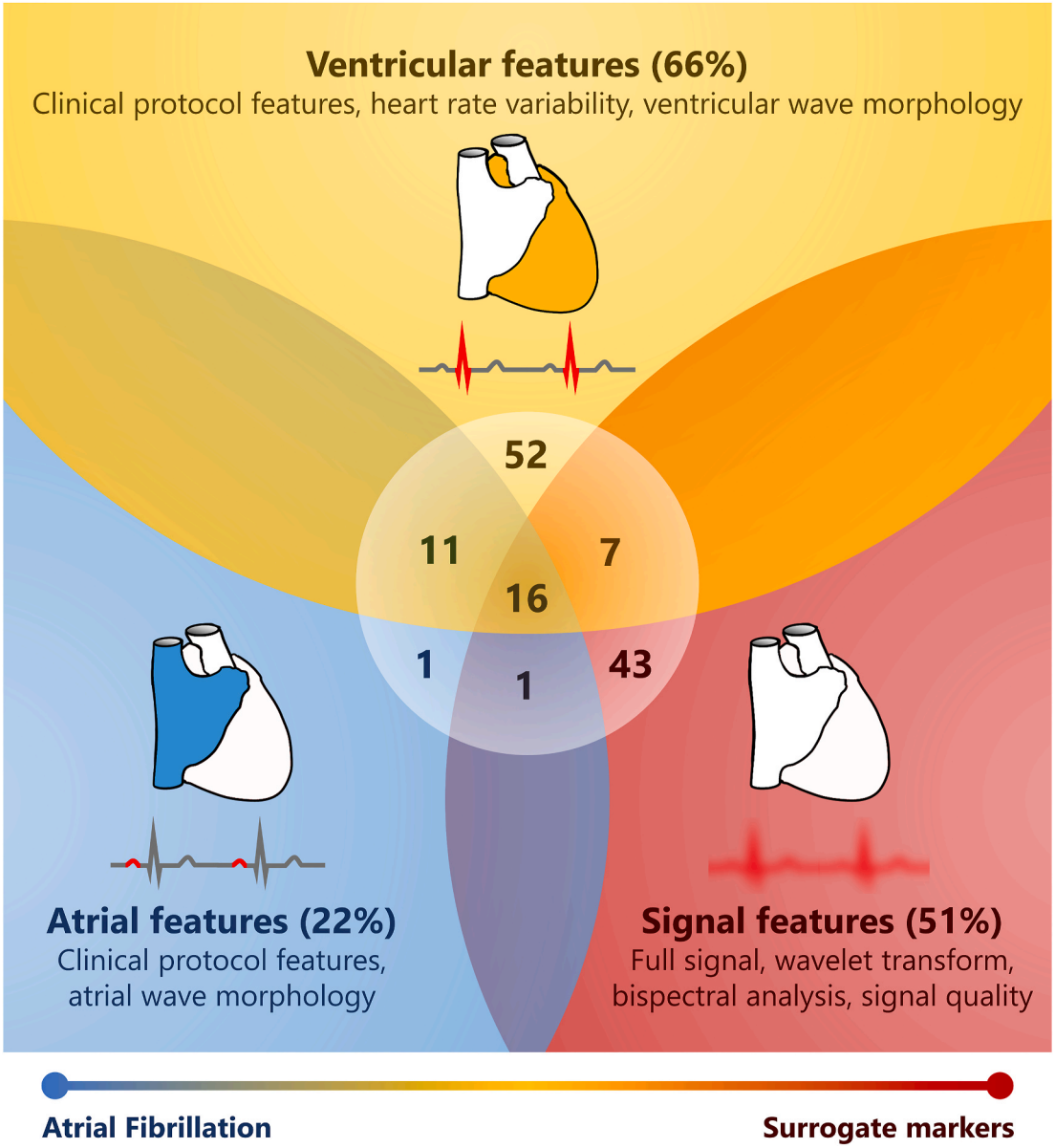

Fig. 3. Overview of the three feature categories which are used to detect atrial fibrillation in electrocardiograms. Yellow indicates ventricular features; blue indicates atrial features; red indicates signal features. Values in the white circle represent the number of studies using features from (a combination of) feature categories.

Table 2
F1-score for AF detection for each set of features.

Feature groups	Number of studies with F1-score (reported or calculated) [%]	Median F1-score [IQR]
Atrial features	1 [100%]	83.8%*
Ventricular features	38 [73%]	96.9% [92.9%–98.1%]
Signal features	34 [79%]	95.2% [83.6%–98.9%]
Atrial + ventricular features	10 [91%]	85.6% [79.8%–95.5%]
Atrial + signal features	1 [100%]	88.9%*
Ventricular + signal features	6 [86%]	91.1% [77.7%–97.7%]
Atrial + ventricular + signal features	13 [81%]	81.0% [78.3%–86.7%]
Overall	103 [79%]	94.0% [83.1%–97.7%]

F1-score is calculated as the harmonic mean of recall and precision. IQR indicates interquartile range. *Only one study, hence no IQR could be calculated.

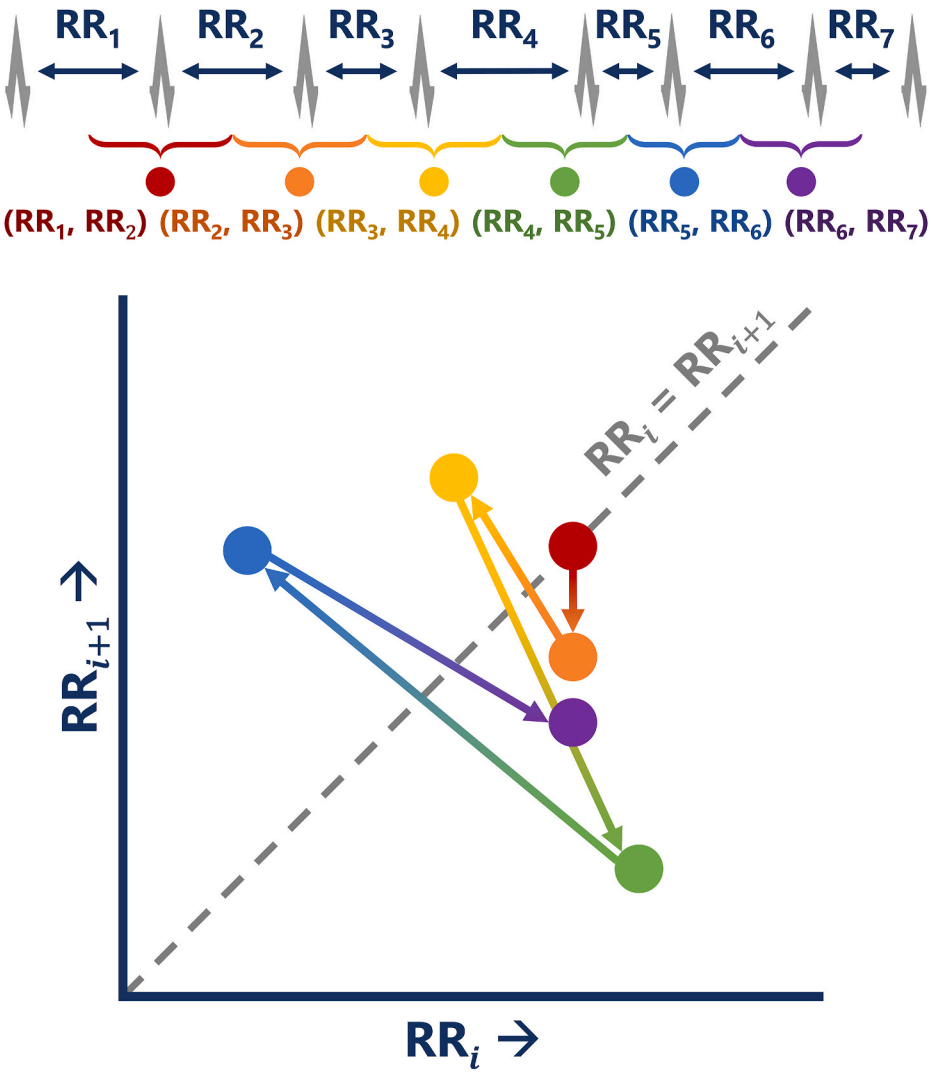

Fig. 4. Schematic visualization of Poincaré or Lorenz plot of RR-intervals. RR_i indicates the interval between the i-th R-peak and the subsequent R-peak.

mean RR-interval. The coefficient of the sample entropy (*CoSEn*) is then defined as:

$$CoSEn = SampEn - \ln\left(2r\right) - \ln\left(\overline{RR}\right) \quad (1)$$

where *SampEn* is the sample entropy, *r* is the tolerance factor, and $\overline{RR}$ is the mean RR-interval.

Alternatively, the Shannon entropy expresses the information or uncertainty of RR-intervals by describing the histogram of already observed RR-intervals in a single metric. [51] When all RR-intervals are similar (i.e. the histogram consists of one single bar), the Shannon entropy equals zero. In contrast, when the variation of RR-intervals is higher, the Shannon entropy increases, hence a higher Shannon entropy – like the sample entropy – indicates that the next RR-interval is less predictable, which is the case during AF.

5.2.4. Turning point ratio (TPR) of RR-intervals

A turning point (TP) of an RR-intervals series is defined as an RR-interval which is larger or smaller than both the preceding and succeeding RR-intervals (i.e. a local maximum or minimum). The turning point ratio (TPR) is given by dividing the total number of TP by the total length of the RR-interval series. [11] An ECG during SR, even in segments with an increasing or decreasing heart rate, will show relatively few turning points since RR-intervals are regular or steadily increasing or decreasing. During AF, however, RR-intervals are irregular, hence the TPR is expected to be higher.

5.2.5. Lyapunov exponent of RR-intervals

Lyapunov exponents describe the divergence of a system for two near equal inputs. Starting with multiple RR-intervals within a narrow range of values at different time points, the Lyapunov exponent is a measure for the variation between trajectories from those points, which is described as the mean distance between the trajectories. [40] For ECGs during SR, the trajectories are not expected to diverge significantly, whereas during AF, since RR-intervals are irregular, the trajectories will show more variation, hence the Lyapunov exponent is larger.

5.3. Signal features

In addition to these easily interpretable features, more abstract signal features describe the signal characteristics in terms of statistical measures, wavelet analysis, phase space analysis, Lyapunov exponents, bispectral analysis, and signal quality. These features are not specifically related to cardiac electrophysiology, but describe the signal more fundamentally. Furthermore, NNs allow the user to input an entire fixed-length ECG recording, after which the NN is trained to detect the most distinctive features, which are mostly abstract and difficult to interpret.

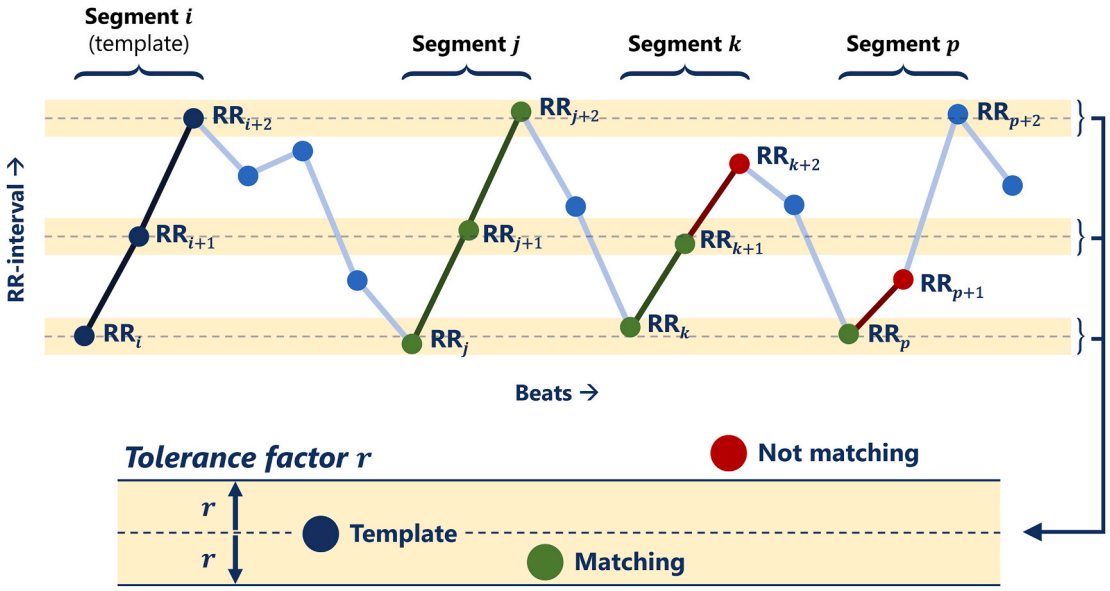

Fig. 5. Schematic visualization of matching used in sample entropy. Segment i is used as a template to compare with segments j, k and p. When setting the prior window length to one RR-interval, hence comparing only the first RR-interval of the segments, all three segments match the template. However, only segments j and k match for the next RR-interval, since RR_{p+1} is below the tolerance range. Increasing the prior window length from one to two RR-intervals results in segment p not matching the template, hence only segments j and k will be tested to match the next RR-interval. In this case, only segment j meets this requirement, as RR_{k+2} is below the tolerance range.

5.3.1. *Basic signal properties and statistical measures*

Signals can be described using basic signal properties and statistical measures, including the maximum and minimum amplitude, signal length, signal power, kurtosis, quartiles, and average value of the first derivative with respect to time. [42,52–54] It must be noted that classification based only on these measures is not expected to be feasible, since these features mainly describe the nature of the signal, and do not provide information on the source of the actual signal. [54]

5.3.2. *Power spectral analysis, bispectral analysis and wavelet transform*

Using the Fourier transform, the frequency spectrum of the signal can be computed by decomposing the signal into sinusoids with different frequencies, amplitudes, and phase shifts. Whilst ECGs containing regular SR are expected to produce narrow peaks in the Fourier spectrum at the fundamental frequency and the harmonics, due to irregularity an ECG containing AF might show more dispersion in the Fourier spectrum. [55]

Similar to the Fourier transform, bispectral analysis focusses on describing the signal in terms of the frequency components. However, with bispectral analysis the interaction between frequency components is analyzed. [56] Therefore, an additional layer of complexity is added. As a result, features extracted from bispectral analysis are not easily manually interpreted.

Although the Fourier transform is conceptually relatively simple, it has a bad trade-off between time-frequency localization; being very localized in the frequency domain, while very non-localized in time. The wavelet transform uses wavelets instead of sinusoids. Wavelets are designed to have limited duration, which can be varied. Therefore, they have an improved joint time-frequency localization. [57] Instead of transforming the signal to the frequency domain, the signal is transformed to *time*-frequency domain, which enables extraction of new signal features from this domain or using the transformed signal as input for NNs. Similar to the higher dispersion in the Fourier transform, a higher dispersion in the wavelet power spectrum is expected to be observed in ECGs during AF, which is mostly described in terms of increased entropy. [58,59]

5.3.3. *Phase space analysis*

Instead of analyzing signals in the time and frequency domain, they can be converted into a phase space, which describes the relation between the original signal and a delayed version of the signal. [60,61] Similar to the Poincaré representation of RR-intervals, using a time delay embedding a plot is generated of the signal amplitude at time t on the x-axis and signal amplitude at time $t + \tau$ on the y-axis, where τ is the chosen time delay. Moreover, using more dimensions with different time delays, even more complex relations can be described. [62] The density of data points in the phase space can be described by segmenting the phase space into smaller regions and counting the number of points in each region. [60,63] In ECGs during SR, the beat-to-beat trajectories of data points are expected to be almost similar. Therefore, data points will be concentrated in a limited number of regions. However, during AF, more variation might result in a larger dispersion of data points over all regions.

5.3.4. *Lyapunov exponents*

Additionally to the Lyapunov exponents from the RR-intervals, Lyapunov exponents can be calculated from the raw ECG signal. For analysis of Lyapunov exponents from raw ECGs, the signal is transformed into the phase space. [64] Next, a starting point is chosen and all points within a set radius are detected. The distance between the trajectories from all these points ("near equal inputs") is a measure for the divergence of the trajectories. Again, since the beat-to-beat trajectories of the ECG signal during SR are expected to be almost similar, low Lyapunov exponents are expected. Instead, during AF, as a result of more variation in trajectories, higher Lyapunov exponents are more likely to be observed.

5.3.5. *Signal quality*

Lastly, several methods use measures to describe the quality of the ECG signal. Multiple measures for signal quality have been proposed. Athif et al. describe signal quality in terms of correlation of each beat with a template beat. [65] Instead, Shao et al. focus on the amplitudes of isoelectric level, and Smisek et al. and Oster et al. compare outputs of multiple QRS-complex detectors. [43,44,66]

6. Discussion

A variety of features from ECG signals are available to describe both atrial and ventricular activity, and general signal characteristics. Using these features, an increasing number of studies focusses on extracting optimal digital biomarkers for AF detection from ECGs. A trend towards using more complex classifiers is observed, predominantly using ventricular features and signal features.

6.1. Comparing AF detection algorithms

The median F1-score for AF detection was 94.0%. Since all included studies used their own methods for detecting AF from ECGs, comparing the performances of the different classifiers might provide useful insight into the overall performances of classification methods. However, these results should be interpreted with great care, since, as pointed out by Ghodrati et al., used databases all focus on a specific group of patients. [12] This focus results in databases which are not representative of the entire patient population. Therefore, training a classifier using these databases might cause the classifier to overfit, hence in clinical practice the accuracy is lower due to, for example, differences in signal quality and prevalence of cardiac arrhythmias. Furthermore, the prevalence of AF in a certain database directly impacts the calculated accuracy measures, making it complex to compare classifiers which were developed and tested using different datasets. For example, the PPV is calculated as the percentage of AF classifications which are correct (i.e. percentage of times the classifier concludes an ECG segment contains AF and the reference label is AF). A higher AF prevalence inherently results in a higher PPV since the a priori probability for an AF classification is higher. Besides the mathematical difficulty of comparing these classifiers, bias could be introduced as a result of one database containing signals which are easier to classify (e.g. only high signal quality and only SR and AF) and the other database containing signals in which the classification is less straightforward (e.g. varying signal quality and multiple arrhythmias). As pointed out earlier, multiple studies in this review showed large variation in performance depending on the used testing database. This relation was only demonstrated for R-peak detection, but most likely there is an even stronger relation for detection of atrial features as a result of the lower amplitude of P-waves. As a result, classifier performance measures are not easily generalized and compared when different databases are used. [3] Moreover, variation in study aim most likely resulted in different trade-offs being made in the classifier design. For example, detecting whether a patient has AF or detecting the duration of AF episodes requires different approaches and studies with these aims will therefore report incomparable performance measures. Furthermore, not all studies use a hidden test set to determine the classifier performance measures, and instead report performance measures from k-fold cross-validation. In contrast to validation using a hidden testing database, no information about the generalizability of the classification methods is obtained using this validation method.

In 2017, the PhysioNet/CinC Challenge focused on AF detection from a single short ECG lead recording. [3] In this challenge, all participants used the same labeled dataset for training and validation (8528 recordings) and for testing (3658 recordings). The test set was hidden during the challenge. Therefore, all AF detection algorithms were tested on the same set of ECGs, eliminating the variation between datasets used in different studies. Still, no single optimal classification method was appointed during the challenge, potentially due to the dataset being too small to give complex approaches an advantage. [3]

Most optimally, all studies use the same testing dataset which is representative of a large patient population. Since this is currently not the case, the results in Table 2, showing the F1-scores per set of feature groups (atrial/ventricular/signal), should be interpreted with care. However, in the included studies, classifiers using only ventricular features seemed to result in the highest F1-scores. Most likely, this is caused by the fact that studies on optimal features for AF detection mostly focus on ventricular features, hence these are more evolved than atrial features.

6.2. Challenges in detecting atrial activity

Already in normal ECGs, a major challenge in detecting atrial activity is the relatively low signal amplitude compared to ventricular activity. During AF the disorganized atrial impulses do not result in a clear P-wave, but deterioration to f-waves with an even lower signal amplitude. Therefore, for intra-cardial mapping a more advanced method for the detection of atrial activity using QRS-T subtraction has been described by Salinet et al. [67] First, the QRS-complexes and T-waves are detected in the surface ECG. Next, the QT-pattern of the intra-cardiac signals is computed, which is then used subtract the ventricular activity from the intra-cardiac measurements. Similarly, Rieta et al. propose three methods to cancel out ventricular activity in intra-cardiac signals using the surface ECG. [68] Using the first method, template matching is applied to compute an average pattern, which is then subtracted from the original signal. Adaptive ventricular cancellation is more complex and aims to estimate the atrial ($a(t)$) and ventricular ($v(t)$) components of the signal $m(t)$, which is described by $m(t) = a(t) + v(t)$. The surface ECG is used as a reference for the ventricular activity. Using a filter based on this reference signal, $a(t)$ is estimated by minimizing the error signal $e(t) = a(t) + v(t) - \hat{v}(t)$, where $\hat{v}(t)$ is the filter output. Lastly, independent component analysis aims to separate the signal into an atrial source and ventricular source based on the assumption that the sources are mutually independent. Although these methods are applied to cancel out ventricular activity from intra-cardiac signals, in a similar way, using a combination of surface ECG leads, QRS-T subtraction could potentially be applied to improve the accuracy of atrial activity detection in surface ECGs.

Besides atrial activity being characterized by a relatively low signal amplitude compared to ventricular activity, noise is another obstacle in detecting atrial activity that also impacts the classification accuracy. [36] A recent review on ECG filtering techniques to eliminate power line interference shows that choosing *the* optimal filter technique is not straightforward, since each technique has its own advantages and disadvantages. [69] Therefore, they propose to use hybrid noise reduction methods which consist of combinations of multiple filtering techniques. More research on optimal filtering techniques for atrial activity detection in ECGs might further improve the detection of atrial activity.

6.3. Limitations

Since algorithm development aims to improve classifier accuracy, many studies solely report the performance measures of the final optimized classifier, without the various alternative classifiers which were trained and validated, but which turned out sub-optimal. This potential publication bias might have impacted the larger picture.

6.4. Future perspectives

AF burden is an emerging risk factor for ischemic stroke. [2] The 2020 ESC Guidelines for diagnosis and management of AF use the classification of AF with only five classes (first diagnosed, paroxysmal, persistent, long-standing persistent, and permanent). [1] However, the AF burden could be more accurately described using the duration, number of episodes and/or proportion of time an individual is in AF during a monitoring period. [2] Since manual analysis of continuous rhythm registrations is unfeasible, new methods should focus on transparent, yet accurate automated AF detection.

The current studies suggest that using only ventricular features gives the highest accuracy. However, research reporting on atrial activity during AF is relatively scarce. When more research is done on methods to optimally describe atrial activity, the transparency might improve without compromising on classifier accuracy. Although AF is an atrial

disease, more than 75% of the methods focus on ventricular features, signal features, or a combination of both. Whilst more complex classifiers are likely to be more accurate, interpretation of the algorithms is becoming more and more difficult due to complex feature sets and NNs processing the signal as a 'black box'. Less than 25% of the methods include features which are derived directly from atrial signal, since mainly due to suboptimal signal quality, atrial activity is difficult to detect. This obstacle might be overcome by using advanced filtering techniques and/or using multi-lead ECG signals, in contrast to most methods only using a limited number of ECG leads. More specifically, using multiple ECG leads in advanced QRS-T subtraction and noise removal methods might facilitate more accurate detection of atrial activity. By using atrial features, AF detection is performed closer to the source and decisions made by AF detection algorithms might remain more transparent.

More generally, due to differences between testing protocols of studies and potential publication bias, comparing performance measures of studies is not feasible. Therefore, we recommend to use a standardized testing protocol for all trained and validated classifiers with a standardized hidden testing database which contains different ECGs with various cardiac arrhythmias and varying signal quality and lead configurations. Currently available databases are the commonly used PhysioNet databases. [5] However, these are not the only databases containing large amounts of ECG data. For example, in 2019, Attia et al. used a dataset containing 1,000,000 12-lead ECGs of more than 200,000 patients from the Mayo Clinic ECG laboratory to train a classifier to detect patients with AF during SR. [70]

7. Conclusion

Over the past years, an increasing number of studies on AF detection have been performed. More and more studies focus on classification algorithms with complex features sets and non-transparent classifiers. Although AF is an atrial disease, less than 25% of the methods include features which are derived directly from atrial signal. Developing new innovative methods focusing on detection of atrial activity might provide accurate classifiers without compromising on transparency.

Sources of funding

N.M.S. de Groot, MD, PhD is supported by funding grants from CVON-AFFIP, The Netherlands [grant number 914728]; NWO-Vidi, The Netherlands [grant number 91717339]; Biosense Webster, USA [ICD 783454]; and Medical Delta, The Netherlands.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.compbiomed.2021.104404.

References

[1] G. Hindricks, T. Potpara, N. Dagres, E. Arbelo, J.J. Bax, C. Blomström-Lundqvist, et al., 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, European Heart Journal 2020 42 (5) (2020) 373–498.

[2] L.Y. Chen, M.K. Chung, L.A. Allen, M. Ezekowitz, K.L. Furie, P. McCabe, et al., Atrial fibrillation burden: moving beyond atrial fibrillation as a binary entity: a scientific statement from the American heart association, Circulation 137 (2018) e623–e644.

[3] G.D. Clifford, C. Liu, B. Moody, L.H. Lehman, I. Silva, Q. Li, et al., AF classification from a short single lead ECG recording: the PhysioNet/computing in Cardiology challenge 2017, Comput. Cardiol. 2017 (2010) 44.

[4] G.B. Moody, R.G. Mark, The impact of the MIT-BIH arrhythmia database, IEEE Eng. Med. Biol. Mag. 20 (2001) 45–50.

[5] A.L. Goldberger, L.A. Amaral, L. Glass, J.M. Hausdorff, P.C. Ivanov, R.G. Mark, et al., PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals, Circulation 101 (2000) E215–E220.

[6] G.B. Moody, R.G. Mark, A new method for detecting atrial fibrillation using R-R intervals, Comput. Cardiol. 10 (1983) 227–230.

[7] X. Zhou, H. Ding, B. Ung, E. Pickwell-MacPherson, Y. Zhang, Automatic online detection of atrial fibrillation based on symbolic dynamics and Shannon entropy, Biomed. Eng. Online 13 (2014) 18.

[8] Y. Li, X. Tang, A. Wang, H. Tang, Probability density distribution of delta RR intervals: a novel method for the detection of atrial fibrillation, Australas. Phys. Eng. Sci. Med. 40 (2017) 707–716.

[9] M.S. Islam, N. Ammour, N. Alajlan, H. Aboalsamh, Rhythm-based heartbeat duration normalization for atrial fibrillation detection, Comput. Biol. Med. 72 (2016) 160–169.

[10] X. Zhou, H. Ding, W. Wu, Y. Zhang, A real-time atrial fibrillation detection algorithm based on the instantaneous state of heart rate, PloS One 10 (2015), e0136544.

[11] S. Dash, K.H. Chon, S. Lu, E.A. Raeder, Automatic real time detection of atrial fibrillation, Ann. Biomed. Eng. 37 (2009) 1701–1709.

[12] A. Ghodrati, B. Murray, S. Marinello, RR interval analysis for detection of Atrial Fibrillation in ECG monitors, Conf Proc IEEE Eng Med Biol Soc 2008 (2008) 601–604.

[13] S.W.E. Baalman, F.E. Schroevers, A.J. Oakley, T.F. Brouwer, W. van der Stuijt, H. Bleijendaal, et al., A morphology based deep learning model for atrial fibrillation detection using single cycle electrocardiographic samples, Int. J. Cardiol. 316 (2020) 130–136.

[14] C.H. Hsieh, Y.S. Li, B.J. Hwang, C.H. Hsiao, Detection of atrial fibrillation using 1D convolutional neural network, Sensors 20 (2020).

[15] M.L. Huang, Y.S. Wu, Classification of atrial fibrillation and normal sinus rhythm based on convolutional neural network, Biomed Eng Lett 10 (2020) 183–193.

[16] P.R. Jeyaraj, E.R.S. Nadar, Atrial fibrillation classification using deep learning algorithm in Internet of Things-based smart healthcare system, Health Inf. J. 26 (2020) 1827–1840.

[17] Y. Ping, C. Chen, L. Wu, Y. Wang, M. Shu, Automatic detection of atrial fibrillation based on CNN-LSTM and shortcut connection, Healthcare 8 (2020).

[18] B.A. Teplitzky, M. McRoberts, H. Ghanbari, Deep learning for comprehensive ECG annotation, Heart Rhythm 17 (2020) 881–888.

[19] Q. Wu, Y. Sun, H. Yan, X. Wu, ECG signal classification with binarized convolutional neural network, Comput. Biol. Med. 121 (2020) 103800.

[20] X. Zhang, K. Gu, S. Miao, X. Zhang, Y. Yin, C. Wan, et al., Automated detection of cardiovascular disease by electrocardiogram signal analysis: a deep learning system, Cardiovasc. Diagn. Ther. 10 (2020) 227–235.

[21] K.S. Lee, S. Jung, Y. Gil, H.S. Son, Atrial fibrillation classification based on convolutional neural networks, BMC Med. Inf. Decis. Making 19 (2019) 206.

[22] A.L.P. Ribeiro, G.M.M. Paixão, P.R. Gomes, M.H. Ribeiro, A.H. Ribeiro, J. A. Canazart, et al., Tele-electrocardiography and bigdata: the CODE (clinical Outcomes in digital Electrocardiography) study, J. Electrocardiol. 57s (2019) S75–s78.

[23] X. Fan, Q. Yao, Y. Cai, F. Miao, F. Sun, Y. Li, Multiscaled Fusion of deep convolutional neural networks for screening atrial fibrillation from single lead short ECG recordings, IEEE J Biomed Health Inform 22 (2018) 1744–1753.

[24] R. Kamaleswaran, R. Mahajan, O. Akbilgic, A robust deep convolutional neural network for the classification of abnormal cardiac rhythm using single lead electrocardiograms of variable length, Physiol. Meas. 39 (2018), 035006.

[25] P.A. Warrick, M. Nabhan Homsi, Ensembling convolutional and long short-term memory networks for electrocardiogram arrhythmia detection, Physiol. Meas. 39 (2018) 114002.

[26] Z. Xiong, M.P. Nash, E. Cheng, V.V. Fedorov, M.K. Stiles, J. Zhao, ECG signal classification for the detection of cardiac arrhythmias using a convolutional recurrent neural network, Physiol. Meas. 39 (2018), 094006.

[27] M. Elgendi, M. Meo, D. Abbott, A Proof-of-concept study: simple and effective detection of P and T waves in arrhythmic ECG signals, Bioengineering 3 (2016) 26.

[28] Y. Chen, X. Wang, Y. Jung, V. Abedi, R. Zand, M. Bikak, et al., Classification of short single-lead electrocardiograms (ECGs) for atrial fibrillation detection using piecewise linear spline and XGBoost, Physiol. Meas. 39 (2018) 104006.

[29] X. Du, N. Rao, M. Qian, D. Liu, J. Li, W. Feng, et al., A novel method for real-time atrial fibrillation detection in electrocardiograms using multiple parameters, Ann. Noninvasive Electrocardiol. 19 (2014) 217–225.

[30] D. Marinucci, A. Sbrollini, I. Marcantoni, M. Morettini, C.A. Swenne, L. Burattini, Artificial neural network for atrial fibrillation identification in portable devices, Sensors 20 (2020).

[31] A. Mukherjee, A. Dutta Choudhury, S. Datta, C. Puri, R. Banerjee, R. Singh, et al., Detection of atrial fibrillation and other abnormal rhythms from ECG using a multi-layer classifier architecture, Physiol. Meas. 40 (2019), 054006.

[32] J. Zheng, H. Chu, D. Struppa, J. Zhang, S.M. Yacoub, H. El-Askary, et al., Optimal multi-stage arrhythmia classification approach, Sci. Rep. 10 (2020) 2898.

[33] N. Liu, M. Sun, L. Wang, W. Zhou, H. Dang, X. Zhou, A support vector machine approach for AF classification from a short single-lead ECG recording, Physiol. Meas. 39 (2018), 064004.

[34] M. Rizwan, B.M. Whitaker, D.V. Anderson, AF detection from ECG recordings using feature selection, sparse coding, and ensemble learning, Physiol. Meas. 39 (2018) 124007.
[35] J. Pan, W.J. Tompkins, A real-time QRS detection algorithm, IEEE Trans. Biomed. Eng. 32 (1985) 230–236.
[36] F. Liu, C. Liu, X. Jiang, Z. Zhang, Y. Zhang, J. Li, et al., Performance analysis of ten common QRS detectors on different ECG application cases, J Healthc Eng 2018 (2018) 9050812.
[37] J. Sacha, Interaction between heart rate and heart rate variability, Ann. Noninvasive Electrocardiol. 19 (2014) 207–216.
[38] A. Kennedy, D.D. Finlay, D. Guldenring, R.R. Bond, K. Moran, J. McLaughlin, Automated detection of atrial fibrillation using R-R intervals and multivariate-based classification, J. Electrocardiol. 49 (2016) 871–876.
[39] S. Bashar, M.B. Hossain, E. Ding, A. Walkey, D. McManus, K. Chon, Atrial fibrillation detection during sepsis: study on MIMIC III ICU data, IEEE J Biomed Health Inform 24 (11) (2020) 3124–3135.
[40] B.M. Asl, S.K. Setarehdan, M. Mohebbi, Support vector machine-based arrhythmia classification using reduced features of heart rate variability signal, Artif. Intell. Med. 44 (2008) 51–64.
[41] R. Czabanski, K. Horoba, J. Wrobel, A. Matonia, R. Martinek, T. Kupka, et al., Detection of atrial fibrillation episodes in long-term heart rhythm signals using a support vector machine, Sensors 20 (2020).
[42] M. Lown, M. Brown, C. Brown, A.M. Yue, B.N. Shah, S.J. Corbett, et al., Machine learning detection of Atrial Fibrillation using wearable technology, PloS One 15 (2020), e0227401.
[43] J. Oster, J.C. Hopewell, K. Ziberna, R. Wijesurendra, C.F. Camm, B. Casadei, et al., Identification of patients with atrial fibrillation: a big data exploratory analysis of the UK Biobank, Physiol. Meas. 41 (2020), 025001.
[44] M. Shao, Z. Zhou, G. Bin, Y. Bai, S. Wu, A wearable electrocardiogram telemonitoring system for atrial fibrillation detection, Sensors 20 (2020).
[45] A. Nguyen, S. Ansari, M. Hooshmand, K. Lin, H. Ghanbari, J. Gryak, et al., Comparative study on heart rate variability analysis for atrial fibrillation detection in short single-lead ECG recordings, Conf Proc IEEE Eng Med Biol Soc 2018 (2018) 526–529.
[46] T. Jeon, B. Kim, M. Jeon, B.G. Lee, Implementation of a portable device for real-time ECG signal analysis, Biomed. Eng. Online 13 (2014) 160.
[47] J. Park, S. Lee, M. Jeon, Atrial fibrillation detection by heart rate variability in Poincare plot, Biomed. Eng. Online 8 (2009) 38.
[48] H.D. Esperer, C. Esperer, R.J. Cohen, Cardiac arrhythmias imprint specific signatures on Lorenz plots, Ann. Noninvasive Electrocardiol. 13 (2008) 44–60.
[49] S.M. Pincus, Approximate entropy as a measure of system complexity, Proc. Natl. Acad. Sci. Unit. States Am. 88 (1991) 2297.
[50] D.E. Lake, 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. 300 (2010) H319–H325.
[51] D. Dharmaprani, L. Dykes, A.D. McGavigan, P. Kuklik, K. Pope, A.N. Ganesan, Information theory and atrial fibrillation (AF): a review, Front. Physiol. 9 (2018).
[52] M. Kropf, D. Hayn, D. Morris, A.K. Radhakrishnan, E. Belyavskiy, A. Frydas, et al., Cardiac anomaly detection based on time and frequency domain features using tree-based classifiers, Physiol. Meas. 39 (2018) 114001.
[53] J.A. Queiroz, A. Junior, F. Lucena, A.K. Barros, Diagnostic decision support systems for atrial fibrillation based on a novel electrocardiogram approach, J. Electrocardiol. 51 (2018) 252–259.
[54] A. Lemkaddem, M. Proenca, R. Delgado-Gonzalo, P. Renevey, I. Oei, G. Montano, et al., An autonomous medical monitoring system: validation on arrhythmia detection, Conf Proc IEEE Eng Med Biol Soc 2017 (2017) 4553–4556.
[55] G.H. Kruger, R. Latchamsetty, N.B. Langhals, M. Yokokawa, A. Chugh, F. Morady, et al., Bimodal classification algorithm for atrial fibrillation detection from m-health ECG recordings, Comput. Biol. Med. 104 (2019) 310–318.
[56] L. Khadra, A.S. Al-Fahoum, S. Binajjaj, A quantitative analysis approach for cardiac arrhythmia classification using higher order spectral techniques, IEEE Trans. Biomed. Eng. 52 (2005) 1840–1845.
[57] M. Priestley, Wavelets and time-dependent spectral analysis, J. Time Anal. 17 (2008) 85–103.
[58] I.H. Bruun, S.M.S. Hissabu, E.S. Poulsen, S. Puthusserypady, Automatic Atrial Fibrillation detection: a novel approach using discrete wavelet transform and heart rate variability, Conf Proc IEEE Eng Med Biol Soc 2017 (2017) 3981–3984.
[59] S. Asgari, A. Mehrnia, M. Moussavi, Automatic detection of atrial fibrillation using stationary wavelet transform and support vector machine, Comput. Biol. Med. 60 (2015) 132–142.
[60] A.S. Al-Fahoum, A.M. Qasaimeh, A practical reconstructed phase space approach for ECG arrhythmias classification, J. Med. Eng. Technol. 37 (2013) 401–408.
[61] C.S. Liu, W.K. Tseng, J.K. Lee, T.C. Hsiao, C.W. Lin, The differential method of phase space matrix for AF/VF discrimination application, Med. Eng. Phys. 32 (2010) 444–453.
[62] N. Marwan, M. Carmen Romano, M. Thiel, J. Kurths, Recurrence plots for the analysis of complex systems, Phys. Rep. 438 (2007) 237–329.
[63] S. Parvaneh, J. Rubin, A. Rahman, B. Conroy, S. Babaeizadeh, Analyzing single-lead short ECG recordings using dense convolutional neural networks and feature-based post-processing to detect atrial fibrillation, Physiol. Meas. 39 (2018), 084003.
[64] E.D. Übeyli, Adaptive neuro-fuzzy inference system for classification of ECG signals using Lyapunov exponents, Comput. Methods Progr. Biomed. 93 (2009) 313–321.
[65] M. Athif, P.C. Yasawardene, C. Daluwatte, Detecting atrial fibrillation from short single lead ECGs using statistical and morphological features, Physiol. Meas. 39 (2018), 064002.
[66] R. Smisek, J. Hejc, M. Ronzhina, A. Nemcova, L. Marsanova, J. Kolarova, et al., Multi-stage SVM approach for cardiac arrhythmias detection in short single-lead ECG recorded by a wearable device, Physiol. Meas. 39 (2018), 094003.
[67] J.L. Salinet Jr., J.P.V. Madeiro, P.C. Cortez, P.J. Stafford, G. André Ng, F. S. Schlindwein, Analysis of QRS-T subtraction in unipolar atrial fibrillation electrograms, Med. Biol. Eng. Comput. 51 (2013) 1381–1391.
[68] J.J. Rieta, F. Hornero, Comparative study of methods for ventricular activity cancellation in atrial electrograms of atrial fibrillation, Physiol. Meas. 28 (2007) 925–936.
[69] P.G. Malghan, M.K. Hota, A review on ECG filtering techniques for rhythm analysis, Research on Biomedical Engineering 36 (2020) 171–186.
[70] Z.I. Attia, P.A. Noseworthy, F. Lopez-Jimenez, S.J. Asirvatham, A.J. Deshmukh, B. J. Gersh, et al., An artificial intelligence-enabled ECG algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction, Lancet 394 (2019) 861–867.