An accurate and efficient method to train classifiers for atrial fibrillation detection in ECGs Learning by asking better questions

Wesselius, Fons J.; van Schie, Mathijs S.; de Groot, Natasja M.S.; Hendriks, Richard C.

DOI
10.1016/j.compbiomed.2022.105331

Publication date
2022

Document Version
Final published version

Published in
Computers in Biology and Medicine

Computers in Biology and Medicine 143 (2022) 105331

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

An accurate and efficient method to train classifiers for atrial fibrillation detection in ECGs: Learning by asking better questions

Fons J. Wesselius [a], Mathijs S. van Schie [a], Natasja M.S. de Groot [a,b,*], 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:
Machine learning
Algorithms
Classification
Atrial fibrillation
ECG signal Processing
Telemetry

ABSTRACT

Background: An increasing number of wearables are capable of measuring electrocardiograms (ECGs), which may help in early detection of atrial fibrillation (AF). Therefore, many studies focus on automated detection of AF in ECGs. A major obstacle is the required amount of manually labelled data. This study aimed to provide an efficient and reliable method to train a classifier for AF detection using large datasets of real-life ECGs.

Method: Human-controlled semi-supervised learning was applied, consisting of two phases: the *pre-training phase* and the *semi-automated training phase*. During pre-training, an initial classifier was trained, which was used to predict the classes of new ECG segments in the semi-automated training phase. Based on the degree of certainty, segments were added to the training dataset automatically or after human validation. Thereafter, the classifier was retrained and this procedure was repeated. To test the model performance, a real-life telemetry dataset containing 3,846,564 30-s ECG segments of hospitalized patients (n = 476) and the CinC Challenge 2017 database were used.

Results: After pre-training, the average F1-score on a hidden testing dataset was 89.0%. Furthermore, after the pre-training phase 68.0% of all segments in the hidden test set could be classified with an estimated probability of successful classification of 99%, providing an F1-score of 97.9% for these segments. During the semi-automated training phase, this F1-score showed little variation (97.3%–97.9% in the hidden test set), whilst the number of segments which could be automatically classified increased from 68.0% to 75.8% due to the enhanced training dataset. At the same time, the overall F1-score increased from 89.0% to 91.4%.

Conclusions: Human-validated semi-supervised learning makes training a classifier more time efficient without compromising on accuracy, hence this method might be valuable in the automated detection of AF in real-life ECGs.

1. Introduction

With the introduction of photoplethysmographic pulse waveform measurements in wearables, these consumer products have made their entrance into the early detection of heart rhythm disorders by monitoring the heart rate. In addition to photoplethysmography, an increasing number of smartwatches and other wearables are also capable of measuring electrocardiograms (ECGs), hence not only providing insight into the rate of cardiac contractions, but also in the electrical activation of the heart [1]. Given the fact that in the 4th quarter of 2019 alone already 118.9 million wearable devices were shipped worldwide [2] and the market is expected to grow further to a market value of $150 billion in 2026 [3], wearables could potentially play an increasingly important role in the early detection of heart rhythm disorders, in particular atrial fibrillation (AF), which, with an estimated prevalence of 2–4% in adults, is the most common sustained cardiac arrhythmia worldwide [4]. Although the ESC Guidelines [4] and EHRA consensus statements [5] recommend that a definite diagnosis of AF can only be established after an ECG recording has been reviewed by a physician, accurate wearable measurements could help in the early

Abbreviations: AF, Atrial Fibrillation; CinC, Computing in Cardiology; CS database, Cardiac Surgery database; ECG, Electrocardiogram; ECV database, Electrical Cardioversion database; EHRA, European Heart Rhythm Association; ESC, European Society of Cardiology; FN, False Negative; FP, False Positive; Hz, Hertz; mV, Millivolts; PoAF, Post-operative AF; s, Seconds; TN, True Negative; TP, True Positive.

* 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.2022.105331
Received 3 September 2021; Received in revised form 11 February 2022; Accepted 16 February 2022
Available online 19 February 2022

detection of AF.

1.1. Challenges in automated AF detection using ECGs

In the ECG, AF is characterized by irregular R-R intervals, absence of P waves and presence of fibrillatory waves. Already in the early 90s, multiple studies endeavored to automatically analyze ECGs in order to automatically detect AF [6–8]. During recent years, the number of studies focusing on automated detection of AF in ECGs steeply increased [9]. Automated AF detection comes with two important challenges. First, the need for good features that can accurately distinguish AF from other rhythms. Although the differences between normal sinus rhythm and AF are apparent, differentiating AF from other *irregular* rhythms (e.g. frequent ventricular extrasystoles or atrial extrasystoles, or sinus arrhythmia) is more complicated, since almost 40% of the proposed algorithms solely rely on the irregularity of ventricular activity as a feature of AF [9]. The incidence of cardiac arrhythmias in general is substantial, particularly in older patients or patients with cardiovascular comorbidities [10]. Furthermore, after cardiac surgery, the reported incidence of post-operative AF (PoAF) is even up to 60% [11]. In patients with PoAF, an increased long-term mortality and stroke incidence was observed, indicating an even higher relevance of early AF detection in these more complex ECGs.

Although new studies mainly innovate on the extracted features from ECGs and the used classification algorithms [9], the next challenge is related to the fact that accurate classification heavily depends on the presence of enough accurately labelled data. In 2017, the yearly Computing in Cardiology (CinC) Challenge focused on detecting AF in short term ECG recordings using a training set of 8,528 samples. Entries in the competition showed F1-scores up to 83.1%, indicating that it is possible to successfully train a classifier to automatically and accurately detect AF [12]. However, the expected advantage of more complex machine learning-based methods (e.g. convolutional neural networks and recurrent neural networks) was not observed. Clifford et al. state that this might be caused by the limited size of the training dataset [12]. However, manually labeling a large number of samples to generate a larger training set is time-consuming and hence might introduce inaccuracies.

1.2. Machine learning techniques

Traditionally, classifiers are trained using *supervised learning*, meaning that, in order to train an accurate classifier, all data has to be manually labelled [13]. As an alternative, *semi-supervised learning* can be used to automatically fill a training dataset based on classifier output without manual intervention, also called self-training. First an initial classifier is trained using manually classified data, after which the training set is augmented with automatically classified data and then re-trained [14]. Using this technique, once self-training has started, there is no guarantee that the augmented data is classified correctly. Furthermore, the classifier does not learn from human input anymore, hence erroneous classifications might introduce tunnel vision to a wrong class for certain combinations of features. Therefore, with semi-supervised learning for AF detection in complex ECGs, recordings containing a variety of different cardiac rhythms, noise levels, and artefacts should be represented in the initial training dataset. Another way to limit the required amount of labelled data is using *transfer learning*, in which an already trained classifier from a similar task is used as a starting point to train a classifier for a new task. De Cooman et al. showed that this method is fast and robust for the detection of seizures based on heart rate. First, they trained a classifier using offline patient-independent data and then used this classifier to analyze patient-specific data [15]. Lastly, *reinforcement learning* is a method which tries to learn from a dataset, but also tries to optimize its reaction to a situation in order to optimize the reward, which, in this case, is the classification accuracy, expressed as the F1-score [16].

1.3. Study aim

The output of a classifier does not necessarily solely consist of the *predicted class* of a newly analyzed sample, but can also contain information on the estimated *probability* of a sample being in a certain class [17]. This information could potentially be used to train the classifier behavior in a reinforcement learning-based approach. Furthermore, a transfer learning-based approach might reduce the required size of the training dataset. Therefore, the aim of this study was to develop and test an efficient and reliable method to train a classifier on a large dataset containing real-life telemetry data without manually classifying all samples, but by combining semi-supervised learning, reinforcement learning and transfer learning techniques, taking into account the degree of certainty of the classifier.

2. Methods

2.1. Dataset

Real-life post-operative telemetry data of 418 hospitalized patients who underwent various types of cardiac surgery (CS) was used (CS database). This dataset was augmented with real-life telemetry data of 58 patients who underwent electrical cardioversion (ECV) as AF treatment (ECV database). All data was acquired using a 12-lead ECG recorder with a sampling frequency of 200Hz, from which only lead II was used to train the classifier. Patients were distributed over a training/validation dataset and a hidden testing dataset with a ratio of 9:1. Long-term recordings were then split into a total of 3,846,564 non-overlapping segments of 30 s (CS database: 3,799,998 segments; ECV database: 46,566 segments; training/validation dataset: 3,474,361 segments; hidden testing dataset: 372,203 segments), which corresponds to the minimum clinical AF episode duration [4] and the average duration of segments in the CinC Challenge [12].

These segments of 30 s were annotated by an investigator experienced in ECG evaluations. First, as visualized in the flowchart (Supplementary Fig. 1), if the ECG could not be assessed due to noise or artefacts (caused by, for example, movement or pacemaker activity), it was classified as noise/artefact (Class ~). Next, if only sinus rhythm beats were present, it was classified as regular sinus rhythm (Class *Normal*). Otherwise, if AF occurred within the recording, it was classified as AF (Class *AF*). In all other cases (e.g. premature atrial or ventricular contractions, or atrial flutter), it was classified as other irregular rhythm (Class *Other*). In total, 3,829 signals were annotated as AF, 8,882 as regular sinus rhythm, 3,500 as other irregular rhythm, and 12,352 as noise/artefact. A hidden test dataset was created using 500 segments of each class. The other segments were used for training and validation purposes. Examples of the four different classes are visualized in Fig. 1.

Furthermore, the database from the CinC Challenge 2017 was used to study the effect of the proposed method on another database [12]. The updated labels from the CinC Challenge 2017 (v3) were used to determine the true class of each segment. The dataset was split into a training set containing 80% of the records and a testing set containing 20% of the records.

2.2. Classifier design

A classifier was designed to differentiate between the previously defined classes, similar to the CinC Challenge 2017 [12]. QRS-peak detection was performed based on the Pan Tompkins algorithm [18]. Next, P- and T-waves were detected using the method as described by Elgendi et al. [19]. As the aim of this study was not to find the optimal features for AF detection, the included features mainly describe the ECG in similar ways as previous studies [9]. For example, the ECG signal was described in terms of peak intervals and amplitudes per heartbeat. Furthermore, the R-R interval variability and the number of detected P-waves and T-waves was determined. Also, since P waves, QRS

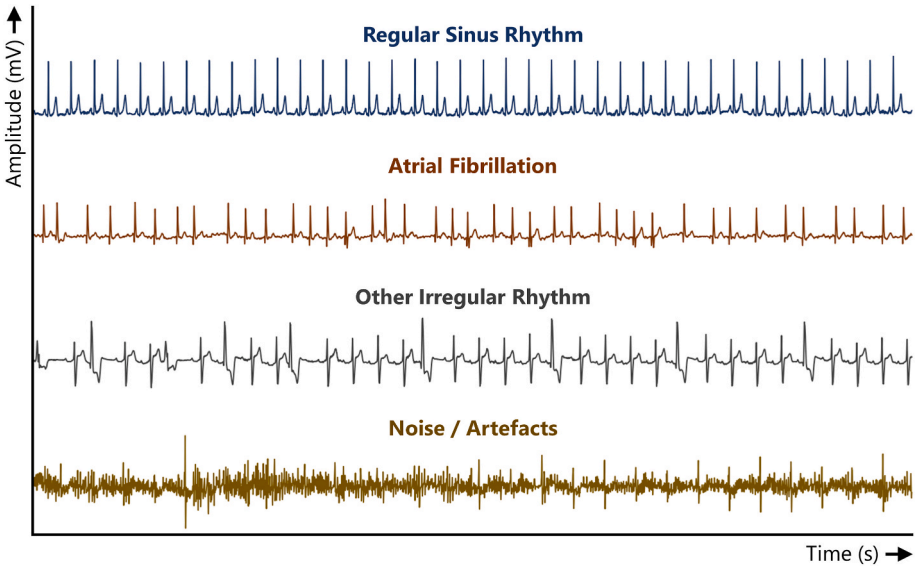


Fig. 1. Typical ECG segments of the four different classes: Regular Sinus Rhythm, Atrial Fibrillation, Other Irregular Rhythm, and Noise/Artefacts. Duration of the segments is 30 s. mV = Millivolts; s = Seconds.

complexes and T waves mainly contain frequencies of 0.5Hz–40Hz, the area under the frequency spectrum was analyzed to get information on the noise level [19]. More specifically, the area under the curve between 0.5Hz and 40Hz was compared to the area under the curve outside this range. As the focus of this paper is on how to efficiently train a classifier without manually labelling all data and not on how to find the optimal feature set, we describe the used features very briefly. A full overview of the used features is presented in Table 1. All calculations were performed using Python (version 3.8.3, 64-bits). Initial exploration showed that boosted decision trees resulted in the best performance. Moreover, multiple studies have reported good performance using boosted decision trees [20–22]. Therefore, the XGBoost Python module (version 1.5.2) was used to train a gradient boosted decision trees classifier based on the features [23].

2.3. General overview of the training model

A schematic overview of the proposed human-validated semi-supervised training process is visualized in Fig. 2. The proposed procedure consists of two phases: the *pre-training phase* (upper panel) and the *semi-automated training phase* (lower panel). During the *pre-training phase* a classifier was trained based on a training dataset containing manually labelled ECG segments. Next, during the *semi-automated training phase*, new ECG segments were semi-automatically classified. The classifier output did not only contain the predicted class, but also the estimated probability p of an ECG segment being of the predicted class – which is from here on called the *degree of certainty*. The degree of certainty was calculated as the mean predicted class probability of the trees in the random forest. The predicted class probability of a tree was calculated as the fraction of samples of the same class in a leaf. When the degree of certainty was above a certain threshold α, the new segment was automatically added to the training dataset. However, if the classifier was uncertain (i.e. $p < \alpha$), the user was asked to validate the predicted class manually, before it was added to the training dataset. Using this new training dataset, the classifier was retrained.

2.4. Phase I: Pre-training phase

The pre-training phase consists of several iterations. In the initialization ($i_I = 0^{\text{th}}$ iteration), ECG segments were manually labelled to create an initial training dataset containing at least $A = 10$ segments per class (a total of at least 40 segments). Then, in each subsequent iteration $i_I \geq 1$, if the training dataset expanded at least $A = 10$ ECG segments for each class compared to the previous training iteration, a temporary classifier was trained and validated. This temporary classifier was used to classify new segments parallel to manual classification in the next iteration $i_I + 1$ of the pre-training phase. Based on whether the classification was correct, the segment was added to either the *training dataset* (in case the classification was correct) or the *corrected training dataset* (in case the classification was incorrect). During the next training iteration, to learn from the previously misclassified ECG segments, at most 10% of the training set was filled with segments which were misclassified by the classifier from the previous training iteration, if available. This value corresponds to the commonly used value for the learning rate of 0.1 in deep learning approaches. If not enough segments were available in the corrected training dataset, all available segments were used. For each training iteration, 80% of the training dataset was used to train the classifier and 20% was used for validation purposes. Furthermore, the classifier was tested using the hidden testing dataset. The stopping criterion for this phase was defined as the smallest set of training data consisting of at least $B = 500$ segments.

2.5. Phase II: Semi-automated training phase

The semi-automated training phase consists of a theoretically unlimited number of iterations. During the initialization of this phase ($i_{II} = 0^{\text{th}}$ iteration), the trained XGBoost classifier from the pre-training phase was used. In order to increase the size of the training dataset, new segments were fed to the classifier. Only segments which the classifier could not classify with $p > \alpha$ were presented to the user for manual validation. All other segments were automatically added to the training set. Again, the classifier was retrained and revalidated when all classes contained at least $A = 10$ ECG segments more than during the last training iteration. This updated classifier was then used during the next iteration $i_{II} + 1$ of the semi-automated training phase.

2.6. Threshold determination

The threshold α was chosen based on the classifier performance for segments which could be classified with $p > \alpha$, balancing the F1-scores (represents classification accuracy) and the number of segments which could be automatically classified (represents time efficiency), as visu-

Table 1
Feature selection used to train the classifier.

Feature class	Feature	Statistical measure
R-R interval variability	Time between R-peaks	Mean, SD, CV, RMSSD, pNN5, pNN10, pNN50
	Ratio between R-R time intervals	Mean, SD, CV, RMSSD
	Poincaré plot of R-R time intervals	SD of points to regression line, SD of points to perpendicular line, shape of Poincaré points (SD/SD)
	Shannon entropy of R-R time intervals	Shannon entropy
	Amplitude difference between R-peaks	Mean, SD, CV, RMSSD, pNN5, pNN10, pNN50
	Ratio between R-R amplitude differences	Mean, SD, CV, RMSSD
	Poincaré plot of R-R amplitude differences	SD of points to regression line, SD of points to perpendicular line, shape of Poincaré points (SD/SD)
Intra-beat peak time intervals	Interval between P-peak and Q-peak	Mean, SD, CV
	Interval between Q-peak and S-peak	Mean, SD, CV
	Interval between Q-peak and T-peak	Mean, SD, CV
	Interval between S-peak and T-peak	Mean, SD, CV
Number of peaks per QRS-complex	Detected P-waves per QRS-complex	% of QRS-complexes with 0/1/>1 P-waves
	Detected T-waves per QRS-complex	% of QRS-complexes with 0/1/>1 T-waves
Peak amplitude differences	Amplitude difference between Q-peak and R-peak	Mean, SD, CV
	Amplitude difference between R-peak and S-peak	Mean, SD, CV
	Ratio between P-wave amplitude and QRS-complex amplitude	Mean, SD, CV
	Ratio between T-wave amplitude and QRS-complex amplitude	Mean, SD, CV
Autocorrelation	Autocorrelation of the ECG segment	Mean, SD, CV
	Peaks in the autocorrelation of the ECG segment	Mean, SD, CV, number of peaks relative to number of detected R-peaks
	Time between peaks in autocorrelation of the ECG segment	Mean, SD, CV
QRS-morphology	Number of different QRS-morphologies	Count
	Most common QRS-morphology	% of QRS-complexes
	Second most QRS-morphology	% of QRS-complexes
Noise level	Time in which no R-peak detection was reliably possible	% of total segment time
	Time in which no P- and T-wave detection was reliably possible	% of total segment time
Frequency analysis	Area under frequency plot between 0.5 and 40Hz	% of total area under frequency plot
	Area under frequency plot between 4 and 10Hz	% of total area under frequency plot
	Area under frequency plot below 0.5Hz	% of total area under frequency plot
	Area under frequency plot above 40Hz	% of total area under frequency plot

CV= Coefficient of Variance; ECG = Electrocardiogram; pNN5 = % of successive R-R time intervals greater than 5 ms; pNN10 = % of successive R-R time intervals greater than 10 ms; pNN50 = % of successive R-R time intervals greater than 50 ms; RMSSD = Root Mean Square of Successive Differences; SD= Standard Deviation.

alized in Fig. 3. Ideally, all ECG segments are automatically classified with an F1-score of 100%, which corresponds to the upper right corner of Fig. 3. In this study, threshold α was chosen based on visual inspection of the relation between the F1-score and the percentage of automatically classifiable segments to find an optimum based on the slope of the relation between the two parameters. The threshold α was determined directly after the pre-training phase only and remained constant for each subsequent training iteration.

2.7. Statistical outcome measures

After each training iteration, the hidden testing dataset – containing 500 segments per class from patients which were not used for training purposes – was used to test the performance of the classifier. Furthermore, after the stopping criterion was satisfied in the pre-training phase, classifier performance was evaluated as a function of the degree of certainty p of the classifier. Similarly, during the semi-automated training phase, classifier performance was evaluated when all classes contained at least 500 ECG segments more with respect to the last testing iteration (i.e. at 500 ECG segments, at 1000 ECG segments, at 1500 ECG segments, etcetera).

The classifier output for each class could either be true positive (TP), false positive (FP), false negative (FN), or true negative (TN). Classification accuracy for each class was described using the precision $P = TP/(TP + FP)$, recall $R = TP/(TP + FN)$, and the F1-score, which is the harmonic mean of the precision and recall: $F1 = 2 \times P \times R/(P + R)$.

2.8. Human-controlled semi-supervised learning on the CinC challenge 2017 dataset

The different classes in this CinC Challenge 2017 dataset are not equally represented and only 3.3% of all segments is classified as noise/artefacts. Therefore, unfortunately it was not possible to perform phase II of the proposed method. Instead, only phase I was applied by using 80% of the data to train the same classifier as used previously. Next, the testing set (20%) was used to test the classifier's performance using different thresholds for the degree of certainty α. Based on this relation, the threshold for α was determined based on visual inspection as described previously.

3. Results

3.1. Pre-training phase

After the pre-training phase, the classifier showed an average precision and recall in the hidden testing dataset of 89.2% and 89.0%, respectively. The corresponding average F1-score was 89.0%. Most inaccuracies were caused by the class containing non-AF irregular rhythms (F1-scores: 87.0% (Class *Normal*), 93.3% (Class *AF*), 78.6% (Class *Other*), 97.1% (Class ~)). The average degree of certainty of the classifier for the correct class was 88.0%. Again, most uncertainties were caused by the class containing non-AF irregular rhythms (degrees of certainty: 87.6% (Class *Normal*), 91.9% (Class *AF*), 77.1% (Class *Other*), 95.3% (Class ~)).

3.2. Threshold determination

As visualized in Fig. 4A, the number of ECG segments which could be classified with $p > \alpha$ decreased when increasing threshold α. As an effect, the ECG segments which could be classified with $p > \alpha$ were more accurately classified, as visualized in Fig. 4B, indicating a relation between the degree of certainty and the F1-score of the classifier. The threshold α was set by balancing the number of segments which could be automatically classified and the F1-score, as visualized in Fig. 4C. More specifically, α was chosen corresponding to the earliest point of significantly increasing curvature. In the first part of the plot up to 1,360

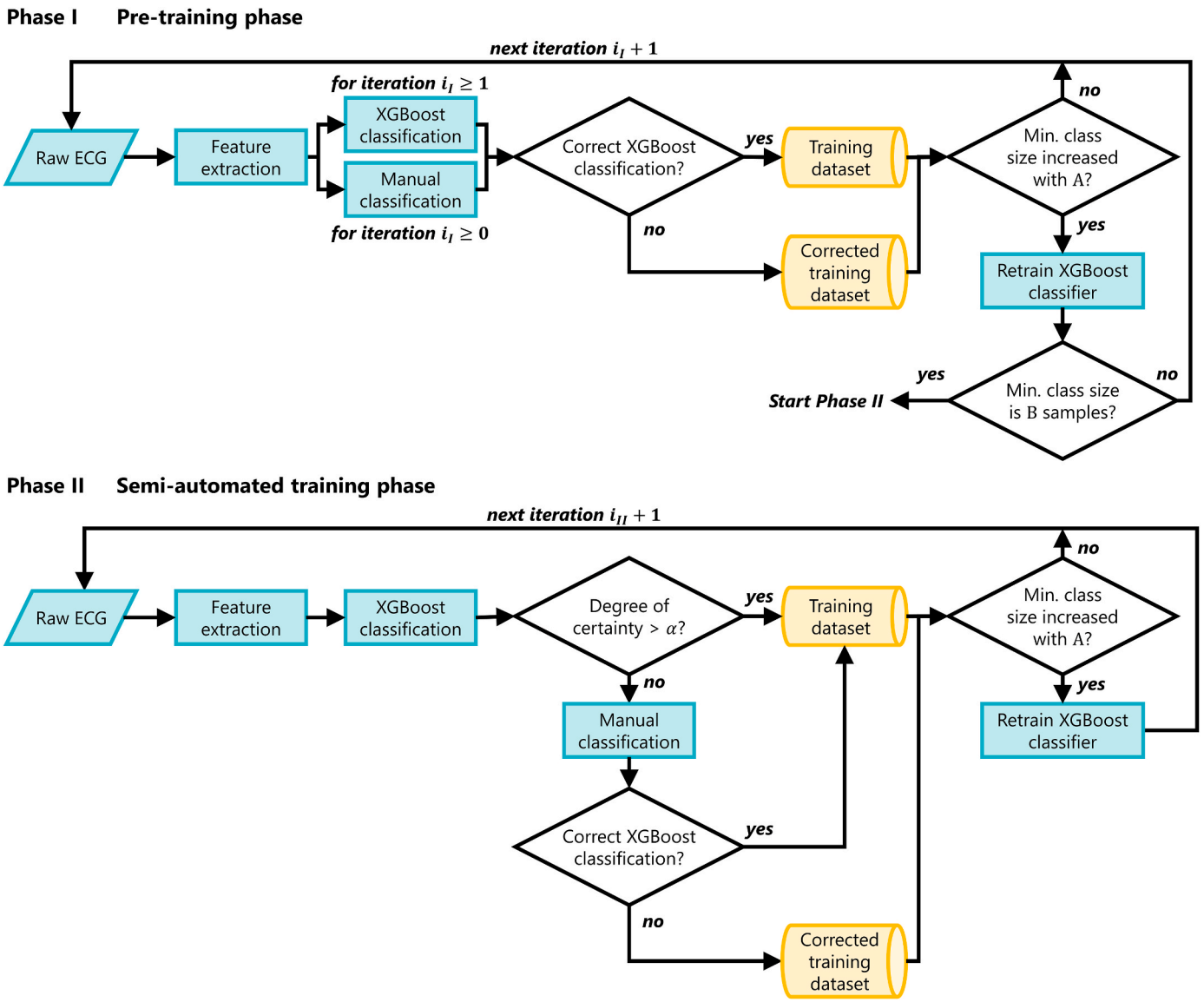


Fig. 2. Human-validated semi-supervised learning, consisting of two phases: the *pre-training phase* and the *semi-automated training phase*, as described in the text. In this study, $A = 10$ segments per class, $B = 500$ segments per class, and $\alpha = 99\%$. ECG = Electrocardiogram; α = Threshold for the estimated probability p of an ECG segment being of the predicted class.

segments, the F1-score decreases with 0.0017% per segment, while in the second part of the plot starting from 1,360 segments, the F1-score decreases more than 10 times as fast with 0.021% per segment. Therefore, based on visual inspection, the threshold α was set at a degree of certainty p of 99%, which corresponds to 68.0% of the testing dataset (= 1,360 ECG segments) being classified automatically with an average F1-score of 97.9%, as shown in Fig. 4A and B, respectively.

3.3. Semi-automated training phase

The average degree of certainty p of the classifier for each training iteration and F1-score of the classifier for the validation dataset is visualized in Fig. 5A and B, respectively. The degree of certainty p increases with increasing size of the training dataset. An increasing F1-score is observed up until a training dataset size of 1,500 ECG segments, after which the increase is less prominent. Similar to the results of the pre-training phase, the F1-score and degree of certainty p for the class containing non-AF irregular rhythms remains lower.

From training with 500 ECG segments to training with 3,000 ECG segments, the percentage of automatically classifiable ECG segments in the testing dataset with $p > 99\%$ increased from 68.0% to 75.8%. The average F1-score for *these segments* remains almost constant with values between 97.3% and 97.9%. The average F1-score for *all* segments in the hidden testing dataset increased from 89.0% to 91.4%. Table 2 summarizes the results for all testing iterations. Complete results for each training iteration using the validation and hidden testing dataset are presented in Supplementary Tables 1 and 2, respectively.

Although threshold α was fixed at 99%, the results for all other thresholds are visualized in Fig. 6. The upper panel again shows that more ECG segments can be automatically classified with more ECG segments in the training dataset. Furthermore, the average F1-score increases, as visualized in the middle panel. Lastly, the lower panel shows that the curve of the relation between the number of ECG segments which could be classified with $p > \alpha$ and the average F1-score shifts towards the upper right corner.

3.4. Human-controlled semi-supervised learning on the CinC challenge 2017 dataset

Using the CinC Challenge 2017 dataset, the average F1-score is

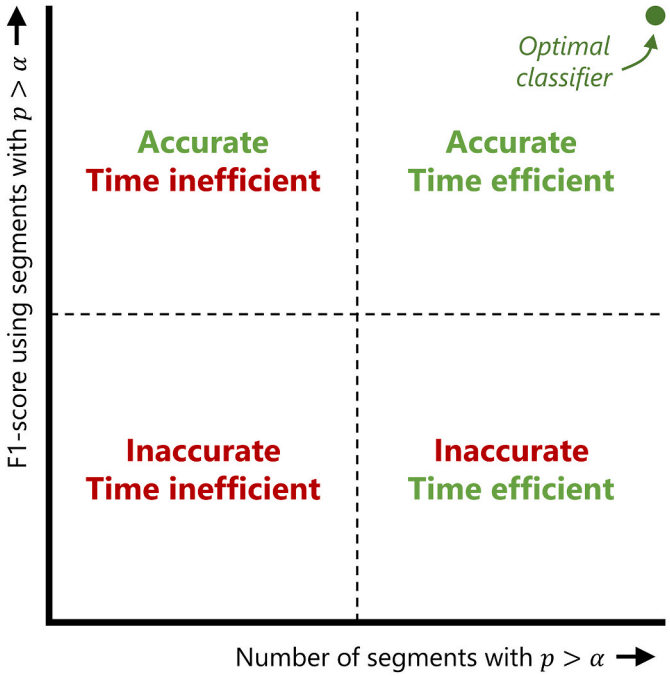


Fig. 3. Trade-off between time efficiency and classification accuracy. In the optimal situation (green dot), all segments are classified automatically and the classification accuracy is 100%. In any other situation, a trade-off has to be made between the classification accuracy of the automatically classified segments and the time efficiency. α = Threshold for the estimated probability p of an ECG segment being of the predicted class; p = Estimated probability of an ECG segment being of the predicted class.

67.0%. Whereas in the previous analysis only the class containing other arrhythmias showed a lower degree of certainty and lower F1-scores, using the CinC Challenge 2017 dataset, this is also observed for the class containing noise (degrees of certainty: 86.1% (Class *Normal*), 61.8% (Class *AF*), 54.9% (Class *Other*), 38.8% (Class ~); F1-scores: 84.9% (Class *Normal*), 69.7% (Class *AF*), 61.6% (Class *Other*), 52.0% (Class ~)).

Results for classifier performance using the CinC Challenge 2017 dataset are visualized in Fig. 7. Fig. 7A and B show similar trends as observed previously for the number of segments with $p > \alpha$ and the F1-score, respectively. Fig. 7C shows a different relation between the number of segments which can be classified with $p > \alpha$ and the F1-score. Instead of a slowly increasing slope, a sudden drop in F1-score is observed around 265 segments (= 15.5% of the testing set). The corresponding threshold α is 98.8% and segments with this degree of certainty show an F1-score of 95.5%.

4. Discussion

A new efficient and accurate method based on a combination of reinforcement learning-based and transfer learning-based methods was applied to train a classifier using a large set of real-life telemetry data of hospitalized patients. Transfer learning-based methods are applied by first training the classifier on a smaller dataset and then using this classifier to classify new ECG segments which are more difficult to classify. At the same time, using the degree of certainty of the classifier, the classifier adapts its behavior to optimize the final classification accuracy.

The major advantage of this new training method is the decreased workload for the user, making it less time-consuming compared to manually validating all ECG segments. In the current study, the workload decreased more than 3/4, since more than 75% of the segments could be classified automatically with p = 99%. Furthermore, hospitalized patients after cardiac surgery show many different combinations of rhythm abnormalities, noise levels, and artefacts, making it difficult to generalize a classifier for all cases. Using the degree of certainty p of the classifier as a gatekeeper before adding an ECG segment to the training set, the training method learns by asking the user when a new rhythm is encountered. Also, since during each training iteration previously misclassified data is used, the classifier learns from its mistakes which were corrected by the user. A previous study by Parvaneh et al. also shows an increase of the F1-score of 3.7% after manually checking all the disagreements between human input and classifier output [24]. However, instead of only retraining the classifier once, the current study iteratively updates the classifier and only asks input for ECGs of which the classifier is uncertain, hence increases time efficiency.

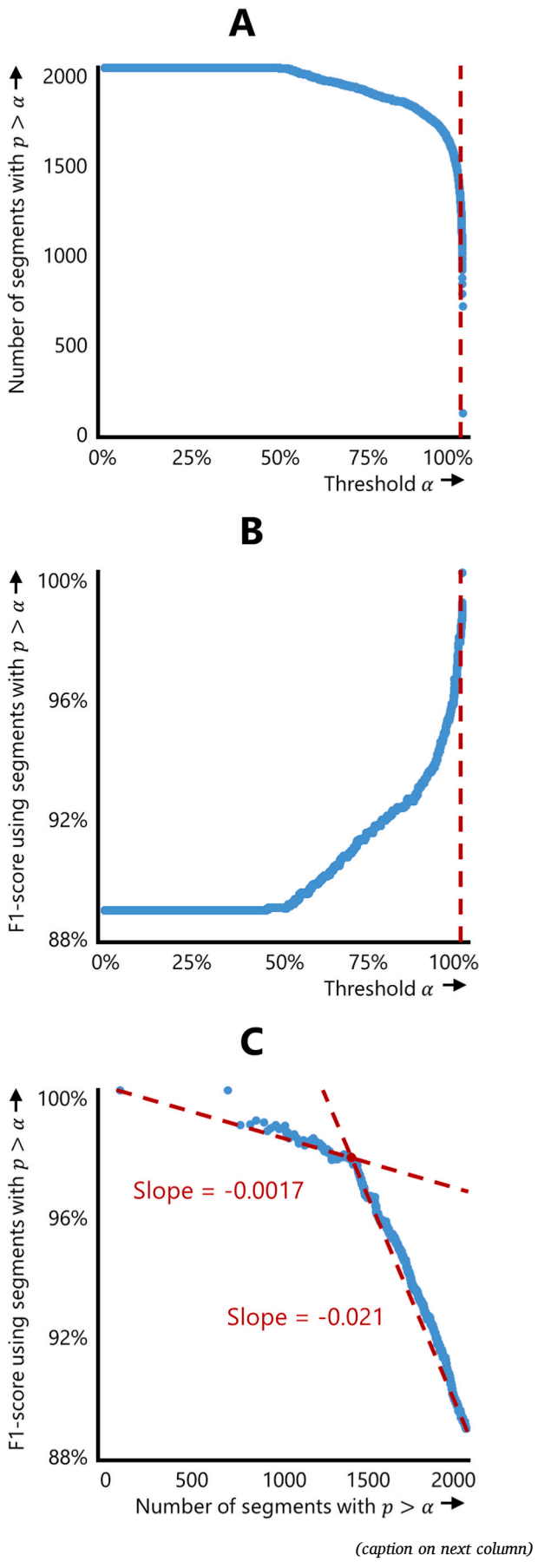
4.1. Database-dependent efficiency

The classifier was initially developed to be used on the real-life post-operative telemetry dataset and the real-life electrical cardioversion dataset. To study the effect of the proposed method on other datasets as well, phase I of the method was also applied on the CinC Challenge 2017 dataset. As demonstrated, the used classification algorithm is suboptimal for the CinC Challenge 2017 dataset, since the average F1-score is 67.0%, whereas previous studies report significantly higher F1-scores [12]. However, the relation between threshold α and the efficiency and F1-scores is still clear (Fig. 7A and B, respectively). A higher degree of certainty correlates with a higher classification accuracy. When comparing Figs. 4C and 7C a clear database-dependent efficiency can be observed. For the CinC Challenge 2017 database, the initial threshold α would be around 98.8%, resulting in automated classification of 15.5% of all segments, whereas using the CS and ECV databases 68.0% of the segments could be automatically classified. This shows that using the method on the CinC Challenge 2017 database improves efficiency less.

A highly probable cause for this large difference is the unequal distribution of the classes in the CinC Challenge 2017 dataset. In the first analysis, an initial classifier was trained using 2000 segments which were equally distributed over the four classes. Although more segments were available for initial training using the CinC Challenge dataset (n = 6,822), these were unequally distributed (59.2% normal sinus rhythm, 8.7% atrial fibrillation, 28.8% other arrhythmias, and only 3.3% noise/artefacts). Fig. 5A shows the increase in degree of certainty resulting from increasing the number of segments in the training set. Not only does the unequal distribution explain the lower efficiency, but it is also a plausible explanation for the low degree of certainty and the low F1-scores for the class containing noise/artefacts in the CinC Challenge 2017 dataset.

4.2. Supervised learning vs. semi-supervised learning vs. human-validated semi-supervised learning

Traditionally, classifiers are trained using a human-validated training dataset, hence the user should label all ECG segments of the training dataset manually [13]. Not only is this time-consuming, but also subject to more human errors. Although several databases containing annotated ECGs are available online (e.g. PhysioNet databases [25], the CinC Challenge 2017 database [12], and the database used by Attia et al. [26]), these databases are mainly focused on specific patient groups and mostly contain ECG segments without much variety in cardiac arrhythmias, noise levels, and artefacts [27]. As an alternative to using a human-validated training dataset, semi-supervised learning can be used to first train an initial classifier, after which the classifier trains itself [14]. The most important assumption using semi-supervised learning, is that the unlabeled data resembles the labelled data. For hospitalized patients after cardiac surgery, this would require the pre-training dataset to contain a large variety of segments containing different rhythm abnormalities and data quality, hence the pre-training dataset should still be relatively large. Instead, the proposed method uses the



(caption on next column)

Fig. 4. Threshold α determination after the pre-training phase. The final threshold ($\alpha = 99\%$) is indicated by the red dashed line in panel A and panel B and the intersection of the two dashed lines in panel C. Upper panel: The number of segments which can be classified with $p > \alpha$ decreases with increasing threshold α. Middle panel: The F1-score for the segments which are classified automatically increases with increasing threshold α. Lower panel: A trade-off between the number of segments which can be classified with $p > \alpha$ and the F1-score is made. After 1,360 segments, the average slope increases from 0.0017 to 0.021, hence in this study, the threshold was set where 1,360 segments could still be classified automatically ($\alpha = 99\%$). α = Threshold for the estimated probability p of an ECG segment being of the predicted class; p = Estimated probability p of an ECG segment being of the predicted class.

estimated probability p of a sample being in the predicted class as an indicator of the degree of certainty to avoid making this assumption. In this way, a new ECG segment, which is completely different from a previous segment, would not be classified with $p > \alpha$ and the user is asked to review the ECG segment manually. Therefore, as opposed to semi-supervised learning, the proposed method still learns from human input after the initial classifier was trained, hence avoiding tunnel vision.

4.3. How certain is certain?

The threshold for the degree of certainty was set at $p = 99\%$ based on visual inspection of the relation between the number of segments which could be automatically classified and the F1-score. Alternatively, the threshold could be set at $p = 0\%$, which would result in the special case of semi-supervised learning with self-training. Still, using this threshold, the average F1-score would be 89.0%. On the other end, the threshold could be chosen even higher than $p = 99\%$, resulting in a slightly higher F1-score, but less ECG segments which could be classified automatically, hence decreasing time efficiency.

In this study, the threshold was set by focusing on the slope of the relation between the number of segments and the F1-score of segments which could be automatically classified. Another option would be to determine the point with the smallest distance to the upper right corner of the graph, which represents the optimal classifier, as visualized in Fig. 3. In doing so, a decrease in the number of segments and a decrease in F1-score are equally penalized. However, using our method, the focus is on having a high F1-score, at the cost of a decrease in time efficiency. Therefore, the used method is preferable over the distance-based method when aiming for the most accurate classifier.

It should be noted that the value for α was set based on results of the testing dataset. Therefore, the presented results might be biased towards the used database. In order to find the optimal value for α – which might be different for each classification problem – it would be better to use an additional validation set.

4.4. Blinded manual classification vs. non-blinded manual validation

During the semi-automated training phase, segments of which the classifier could not determine the class with $p > \alpha$ were shown to the user for *validation*, hence a combination of 1) the ECG segment, 2) the predicted class and 3) the degree of certainty p was presented to the user. Another option would be to ask the user to manually *classify* the ECG segment blinded, since showing the predicted class to the user might introduce bias towards a certain class. This raises the question of whether this bias would negatively influence the training results. As shown in Fig. 4B, even in the special case of semi-supervised learning with self-training (threshold $\alpha = 0\%$), the average F1-score is 89.0%. The corresponding average recall and precision are 89.0% and 89.2%, respectively, indicating that even when the classifier is not sure of the predicted class, the predicted class is still correct in 89% of the cases. Therefore, in 89% of the cases, the introduced bias actually is a well-directed push towards the correct class, hence showing the predicted

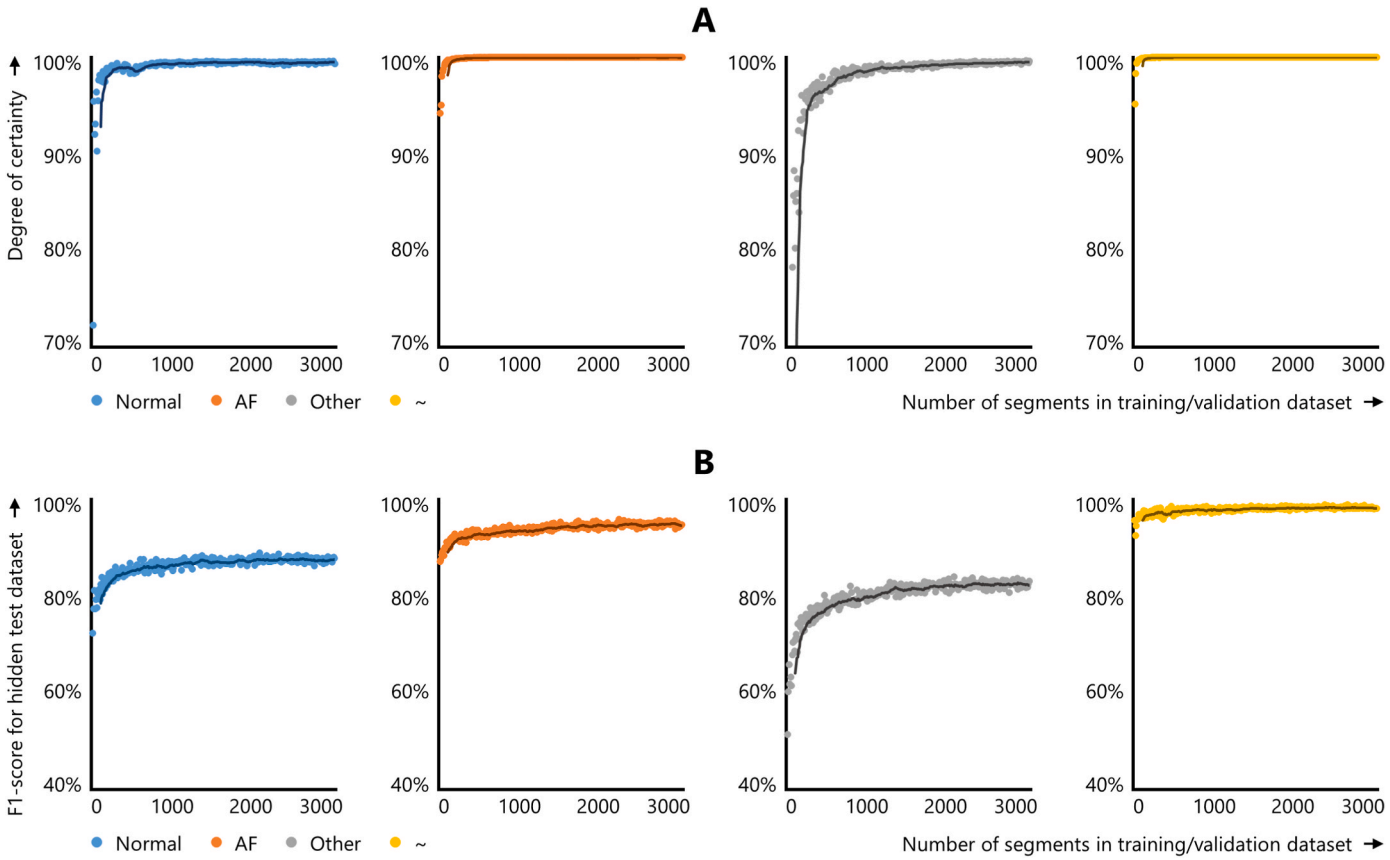


Fig. 5. Effect of increasing the size of the training dataset. Dots indicate averages per training iteration, solid lines indicate moving averages over 10 training iterations. Upper panel: Increasing the number of segments in the training dataset results in a higher degree of certainty for the classifier, represented by the estimated probability p of an ECG segment being of the predicted class. Lower panel: Increasing the number of segments in the training dataset results in a higher F1-score for the validation dataset. AF = Atrial fibrillation; Normal = Regular sinus rhythm; Other = Other irregular rhythms; ~ = Noise or artefacts.

Table 2
Results per testing iteration.

Testing iteration	Number of segments in training dataset	Number of segments with $p > 99\%$ (%)	Average recall (%)		Average precision (%)		Average F1-score (%)	
			All:	$>\alpha$:	All:	$>\alpha$:	All:	$>\alpha$:
1	**500**	1360 (68.0%)	89.0%	97.9%	89.2%	97.9%	89.0%	97.9%
2	**1000**	1475 (73.8%)	89.6%	97.4%	89.8%	97.3%	89.7%	97.3%
3	**1500**	1488 (74.4%)	91.6%	97.4%	91.5%	97.3%	91.6%	97.3%
4	**2000**	1511 (75.6%)	91.7%	97.9%	91.8%	97.8%	91.7%	97.9%
5	**2500**	1522 (76.1%)	91.5%	97.6%	91.6%	97.5%	91.6%	97.5%
6	**3000**	1515 (75.8%)	91.3%	97.6%	91.5%	97.4%	91.4%	97.5%

p = Estimated probability of an ECG segment being of the predicted class; All = Results based on the entire hidden testing dataset; $>\alpha$ = Results based on segments which could be classified with $p > \alpha$.

class actually might help in improving the classification accuracy. If, however, the classification accuracy would be poor for low degrees of estimated probability p, showing the predicted class likely has a negative influence, as might be the case for the analysis using the CinC Challenge 2017 dataset where the F1-score for the semi-supervised learning case was 67.0% (Fig. 7B).

4.5. Future perspectives

This study not only shows that human-validated semi-supervised learning results in an accurate classifier with a lower human workload, but also shows that the degree of certainty of the classifier increases with an increasing number of ECG segments in the training dataset and is related to classifier accuracy. Using the degree of certainty of the classifier in AF detection algorithms for the detection of AF episodes in long-term real-life telemetry data of hospitalized data might result in a more accurate detection, paving the way towards accurately determining the AF burden in these patients in terms of AF duration, number of episodes, and proportion of time an individual is in AF [28].

Furthermore, this method was now applied to real-life data of *hospitalized* patients. However, more and more *non-hospitalized* people also have their heart rhythm registered using wearables, showing similar problems of noisy measurements with artefacts and potentially other heart rhythms. Therefore, when training a classifier to automatically detect rhythm abnormalities in these recordings, human-validated semi-supervised learning might be feasible to decrease the human workload. In this case, human-supervised semi-supervised learning should not be applied by the end-user, since then the user should be a trained clinician,

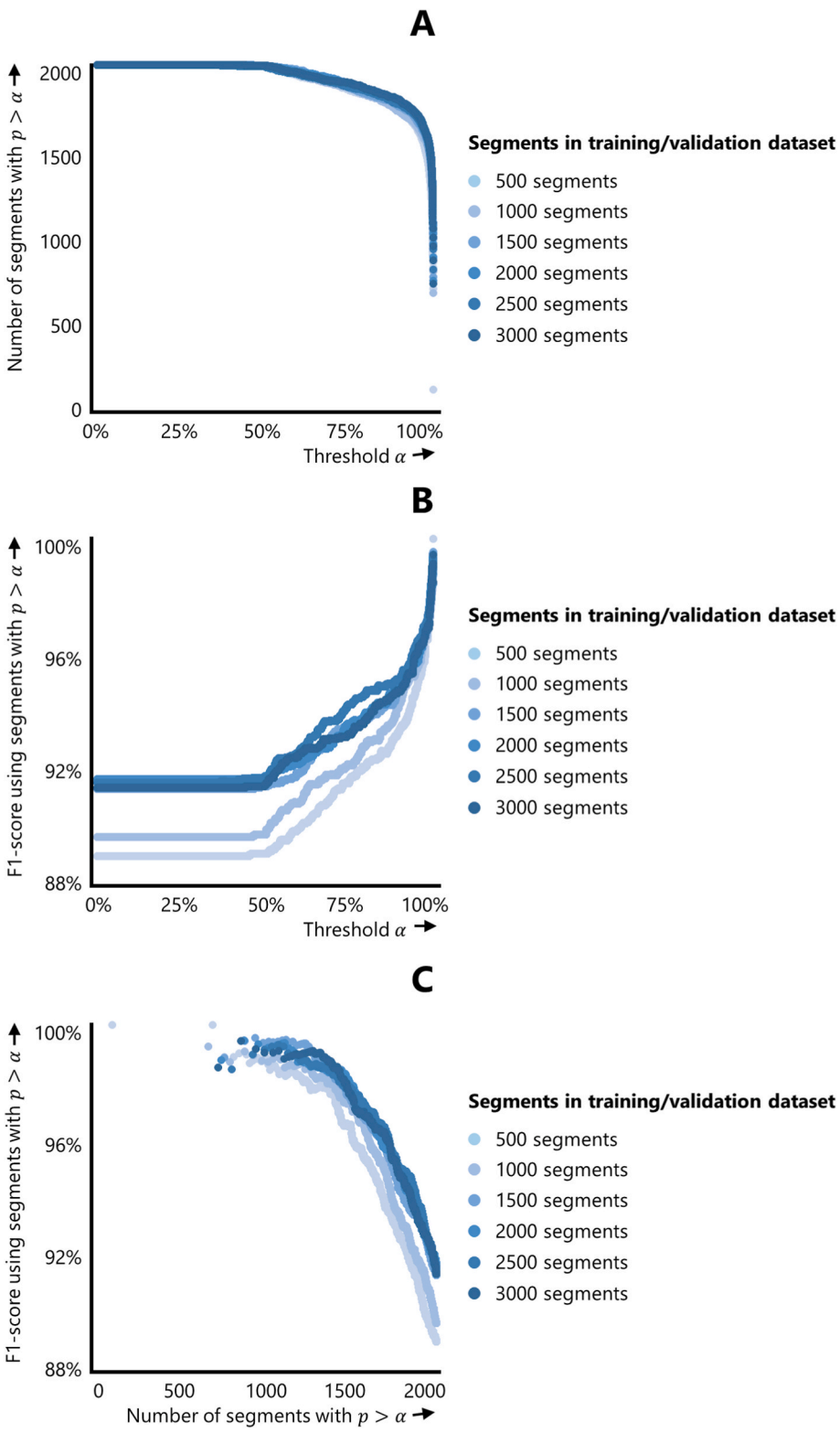
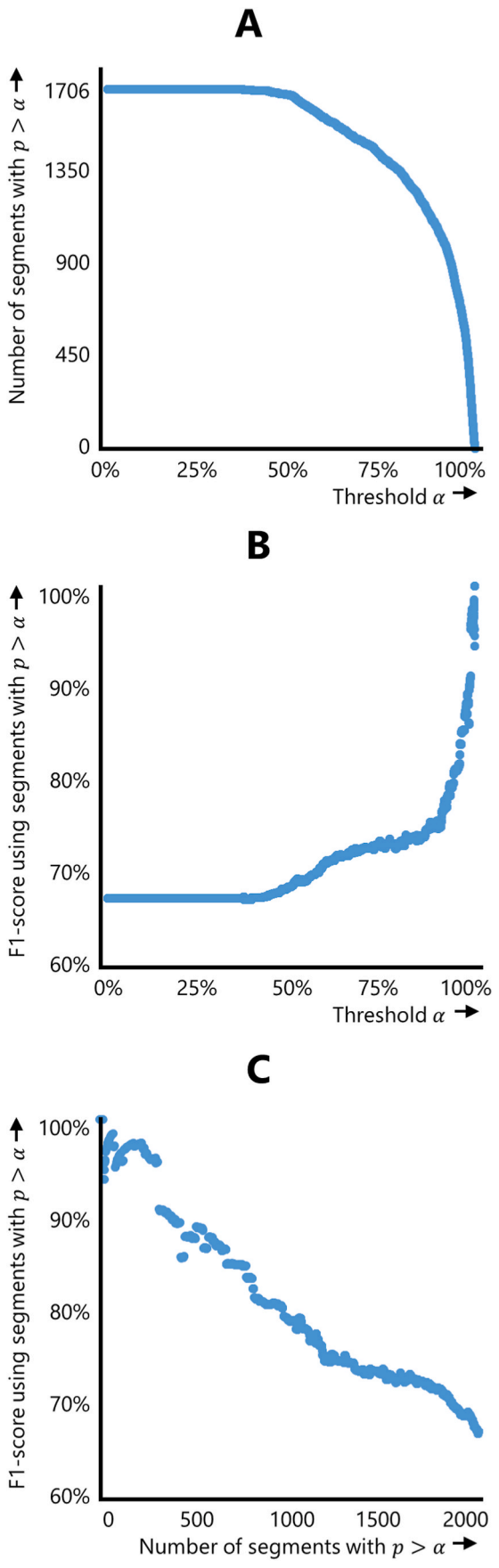


Fig. 6. Effect of increasing the size of the training dataset on results from the testing dataset for all thresholds α from 0% to 100% with steps of 0.01%. Upper panel: Relation between threshold α and number of segments with $p > \alpha$. Mainly for high values of α, the number of segments which can be classified with $p > \alpha$ increases with increasing size of the training dataset. Middle panel: Relation between threshold α and the F1-score. Mainly for low values of α, the F1-score of the classifier increases with increasing size of the training dataset. Lower panel: Relation between the number of segments with $p > \alpha$ and the F1-score. With increasing size of the training dataset, lines move towards the upper right corner of the graph, which is the optimal situation, as visualized in Fig. 3. This indicates that a better trade-off between classification accuracy and time efficiency is possible using a larger training dataset. α = Threshold for the estimated probability p of an ECG segment being of the predicted class; p = Estimated probability of an ECG segment being of the predicted class.

but could be used to reduce the workload for training a validated, accurate classifier using a large dataset.

5. Limitations

First, although technically not a limitation of this study in view of its methods, this study did not aim to reach the highest accuracy for AF detection, but to propose an efficient method of training an accurate classifier. For future studies, applying human-validated semi-supervised learning in accurate detection of cardiac arrhythmias, using other datasets with validated annotations (e.g. PhysioNet 2017 and 2021 datasets) is essential. Further, in this study only a single ECG lead was used. Using more leads – although not always possible due to data availability, especially in wearables – might improve classifier accuracy.



(caption on next column)

Fig. 7. Results for analysis using the CinC Challenge 2017 database. Upper panel: The number of segments which can be classified with $p > \alpha$ decreases with increasing threshold α. Middle panel: The F1-score for the segments which are classified automatically increases with increasing threshold α. Lower panel: Relation between the number of segments which can be classified automatically (efficiency) and the F1-score (accuracy). α = Threshold for the estimated probability p of an ECG segment being of the predicted class; p = Estimated probability p of an ECG segment being of the predicted class.

A dataset containing telemetry recordings of hospitalized patients was used, hence the classification results themselves should not be generalized to other wearable data of another population. First, telemetry data of hospitalized patients might have different quality compared to consumer wearable data. Also, the patient population using consumer wearables is different from the population of hospitalized patients. Nonetheless, since the incidence of various cardiac arrhythmias is lower in healthy and young individuals [10], applying the proposed method to train a classifier for AF detection in this group likely results in a similar or even better classifier performance. However, given the higher incidence of cardiac arrhythmias in elderly patients with cardiovascular comorbidities [10], the advantage of human-validated semi-supervised learning over classical semi-supervised learning most likely is less striking when applied to wearable ECGs from young and healthy individuals.

Although human-validated semi-supervised learning resulted in an accurate classifier, it still requires three parameters to be set manually, which could influence the performance. First, the threshold α is currently set based on visual inspection. More optimally, the threshold is automatically determined and adapts after each testing iteration. Next, the number of extra segments which is required to retrain the classifier was currently heuristically set at $A = 10$ per class and the number of segments required to start the semi-automated training phase at $B = 500$ per class. The effect of changing these thresholds was not investigated, but likely influences the time efficiency of the proposed method and the final classification accuracy. Lastly, during training, at most 10% of the data from the corrected training dataset was used, corresponding to the commonly used value for the learning rate of 0.1 in deep learning approaches. However, using too much data from the corrected training dataset might result in overtraining on the misclassified ECG segments. On the opposite, using too few segments from the corrected training dataset would reduce the effect of the classifier learning from its mistakes.

6. Conclusion

A new and efficient method, called human-validated semi-supervised learning, was proposed for training a classifier for large sets of ECG segments. This method makes training of an accurate classifier more time efficient without compromising on classification accuracy, hence increasing the size of the training dataset is less of an obstacle. Therefore, this method might be valuable when training a classifier based on large amounts of real-life ECG data of hospitalized patients showing varying cardiac rhythms, noise levels, and artefacts.

Sources of funding

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

Disclosures

None.

Declaration of interests

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.2022.105331.

References

[1] C.R. Lopez Perales, H.G.C. Van Spall, S. Maeda, A. Jimenez, D.G. Laţcu, A. Milman, et al., Mobile health applications for the detection of atrial fibrillation: a systematic review, EP Europace 23 (1) (2020) 11–28.

[2] IDC., Shipments of Wearable Devices Reach 118.9 Million Units in the Fourth Quarter and 336.5 Million for 2019, According to IDC, IDC Media Center, 2020 [Retrieved on 3 September 2021; Available from: https://www.idc.com/getdoc.jsp?containerId=prUS46122120.

[3] TCoTE. Communities, Smart wearables: reflection and orientation paper, Digit. Ind. Compet. Electron. Ind. 121 (17) (2016).

[4] 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, 2020.

[5] G.H. Mairesse, P. Moran, I.C. Van Gelder, C. Elsner, M. Rosenqvist, J. Mant, et al., Screening for atrial fibrillation: a European heart rhythm association (EHRA) consensus document endorsed by the heart rhythm society (HRS), Asia Pacific heart rhythm society (APHRS), and sociedad Latinoamericana de Estimulación Cardíaca y electrofisiología (SOLAECE), EP Europace 19 (10) (2017) 1589–1623.

[6] J. Slocum, A. Sahakian, S. Swiryn, Diagnosis of atrial fibrillation from surface electrocardiograms based on computer-detected atrial activity, J. Electrocardiol. 25 (1) (1992) 1–8.

[7] D. Cubanski, D. Cyganski, E.M. Antman, C.L. Feldman, A neural network system for detection of atrial fibrillation in ambulatory electrocardiograms, J. Cardiovasc. Electrophysiol. 5 (7) (1994) 602–608.

[8] T.F. Yang, B. Devine, P.W. Macfarlane, Artificial neural networks for the diagnosis of atrial fibrillation, Med. Biol. Eng. Comput. 32 (6) (1994) 615–619.

[9] F.J. Wesselius, M.S. Van Schie, N.M.S. De Groot, R.C. Hendriks, Digital biomarkers and algorithms for detection of atrial fibrillation using surface electrocardiograms: a systematic review, Comput. Biol. Med. 133 (2021) 104404.

[10] S. Khurshid, S.H. Choi, L.-C. Weng, E.Y. Wang, L. Trinquart, E.J. Benjamin, et al., Frequency of cardiac rhythm abnormalities in a half million adults, Circulation: Arrhythmia and Electrophysiology 11 (7) (2018), e006273.

[11] B. Maesen, J. Nijs, J. Maessen, M. Allessie, U. Schotten, Post-operative atrial fibrillation: a maze of mechanisms, Europace 14 (2) (2012) 159–174.

[12] 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. (2010) 44, 2017.

[13] K.W. Johnson, J. Torres Soto, B.S. Glicksberg, K. Shameer, R. Miotto, M. Ali, et al., Artificial intelligence in Cardiology, J. Am. Coll. Cardiol. 71 (23) (2018) 2668–2679.

[14] N.N. Pise, P. Kulkarni (Eds.), A Survey of Semi-supervised Learning Methods, International Conference on Computational Intelligence and Security; 2008 13-17 Dec. 2008, 2008.

[15] T. De Cooman, K. Vandecasteele, C. Varon, B. Hunyadi, E. Cleeren, W. Van Paesschen, et al., Personalizing heart rate-based seizure detection using supervised SVM transfer learning, Front. Neurol. 11 (145) (2020).

[16] A. Jonsson, Deep reinforcement learning in medicine, Kidney Dis. 5 (1) (2019) 18–22.

[17] A. Hosameldin, K.N. Asoke, Probabilistic Classification Methods. Condition Monitoring with Vibration Signals: Compressive Sampling and Learning Algorithms for Rotating Machines, IEEE, 2019, pp. 225–237.

[18] J. Pan, W.J. Tompkins, A real-time QRS detection algorithm, IEEE Trans. Biomed. Eng. 32 (3) (1985) 230–236.

[19] 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 (Basel) 3 (4) (2016) 26.

[20] 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 (10) (2018) 104006.

[21] P. Sodmann, M. Vollmer, N. Nath, L. Kaderali, A convolutional neural network for ECG annotation as the basis for classification of cardiac rhythms, Physiol. Meas. 39 (10) (2018) 104005.

[22] 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 (11) (2018) 114001.

[23] T. Chen, C. Guestrin, XGBoost: a scalable tree boosting system, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery, San Francisco, California, USA, 2016, pp. 785–794.

[24] 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 (8) (2018), 084003.

[25] 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 (23) (2000) E215–E220.

[26] 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 (10201) (2019) 861–867.

[27] 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.

[28] 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 (20) (2018) e623–e644.