Identification of patients with Atrial Fibrillation, a Big Data exploratory analysis of the UK Biobank

Julien Oster[1], Jemma C. Hopewell[2], Klemen Ziberna[3], Rohan Wijesurendra[3], Christian F. Camm[2], Barbara Casadei[3], Lionel Tarassenko[4]

[1]IADI, U1254, INSERM, Université de Lorraine, Nancy, France
[2]CTSU, Nuffield Department of Population Health, University of Oxford
[3]Cardiovascular Medicine, University of Oxford
[4]Institute of Biomedical Engineering, Dept of Engineering Science, University of Oxford, Oxford OX3 7DQ, United Kingdom

E-mail: `lionel.tarassenko@eng.ox.ac.uk`

Abstract.

Atrial Fibrillation (AF) is the most common cardiac arrhythmia, with an estimated prevalence of around 1.6% in the adult population. The analysis of the Electrocardiogram (ECG) data acquired in the UK Biobank represents an opportunity to screen for AF in a large sub-population in the UK.

The main objective of this paper is to assess ten machine-learning methods for automated detection of subjects with AF in the UK Biobank dataset. Six classical machine-learning methods based on Support Vector Machines are proposed and compared with state-of-the-art techniques (including a deep-learning algorithm), and finally a combination of a classical machine-learning and deep learning approaches. Evaluation is carried out on a subset of the UK Biobank dataset, manually annotated by human experts.

The combined classical machine-learning and deep learning method achieved an F1 score of 84.8% on the test subset, and a Cohen's Kappa coefficient of 0.83, which is similar to the inter-observer agreement of two human experts.

The level of performance indicates that the automated detection of AF in patients whose data have been stored in a large database, such as the UK Biobank, is possible. Such automated identification of AF patients would enable further investigations aimed at identifying the different phenotypes associated with AF.

Keywords: Electrocardiogram, Atrial Fibrillation, Biobank, Big Data, machine learning, signal processing

Submitted to: *Physiol. Meas.*

1. Introduction

Cardiovascular diseases (CVD) are associated with four millions deaths in Europe each year, which represents 46% of annual deaths (Nichols et al. 2013). Most deaths caused by CVD are premature and could be prevented by changes in lifestyle. The Framingham Heart Study was a precursor in the collection and analysis of prospective longitudinal data for this endeavour (Kannel & McGee 1979). A very large prospective study, with over 500,000 participants aged between 40 and 69 years, UK Biobank, has been collecting extensive phenotype and genotype data for the last 12 years (Sudlow et al. 2015, Collins 2012, Palmer 2007). A study using this large dataset has recently demonstrated its value by comparing 5-year mortality predictors (Ganna & Ingelsson 2015). Among the collected data, the electrocardiogram (ECG) has been recorded at rest and during exercise on a sub-population of 100,000 participants.

Atrial Fibrillation (AF) is the most common cardiac arrhythmia, with a prevalence increasing with age. However it remains quite difficult to evaluate the true prevalence of AF, since arrhythmia can be asymptomatic. Different studies have estimated its prevalence in the adult population (aged 45+) to be around 2.0% (95% Confidence Interval 1.6% - 2.4%) (1.6% in women and 2.4% in men) (Davis et al. 2012). The prevalence of AF increases exponentially with age. For instance, among people aged 65 and over the prevalence of AF in the UK rises to approximately 7% (Jones et al. 2014, Fuster et al. 2011). A recent study has highlighted the power of home-based long-term cardiac monitoring for the early detection of AF (Steinhubl et al. 2018). The detection of AF was higher (6.7% vs 2.6%) with active home-based monitoring (compared to a control group without home-based monitoring). Earlier detection of AF was associated with increased anticoagulant treatment, but no statistical difference in AF-related hospitalisation. AF is a major risk factor for stroke, its presence leading to a five-fold higher risk (Wolf et al. 1991). The analysis of the ECG data acquired in the UK Biobank study represents an opportunity to screen for AF in a large middle-aged sub-population in the UK.

The automated detection of AF has an extensive literature, recently reviewed from an engineering perspective (Sörnmo 2018). AF is characterised by a highly irregular ventricular response (or heart rate), but also by ECG morphological changes such as the presence of small f-waves (small oscillations with a dominant frequency around 5Hz with amplitudes of few μV), which replace the P wave. Automated techniques were first designed based on extracting hand-crafted features reflecting either morphological (presence of P wave or f waves) or rhythm-based characteristics. The latter techniques have mainly consisted in the extraction of predictability and regularity features of the RR intervals for a given window (Lake & Moorman 2011, Sarkar et al. 2008). Machine learning approaches have recently been proposed to combine multiple features (Colloca et al. 2013). The PhysioNet/Computing in Cardiology (CinC) challenge 2017 provided the scientific community with the largest publicly available dataset of annotated (single-lead, mobile-phone derived) ECG signals yet ($\approx 10,000$ subjects). The challenge was

aimed at the identification of subjects with AF (Clifford et al. 2017). Standard machine learning approaches not involving deep learning were proposed with innovative hand-crafted features (including morphological features (such as presence of P or f waves, QRS width,...), temporal features (regularity of RR intervals,...), or abductive signal interpretation) (Teijeiro et al. 2017, Datta et al. 2017, Behar et al. 2017), but the challenge also offered the first glimpse of deep learning techniques applied to this problem (Andreotti et al. 2017), inspired by recent success in applying deep learning approaches to rhythm classification (Hannun et al. 2019).

The main objective of this paper was to assess ten machine learning methods for the automated detection of subjects with AF in a sub-population of 100,000 UK Biobank participants. Six classical machine learning approaches, with hand-crafted features as the inputs to a Support-Vector-Machine classifier, were proposed, and compared with three methods proposed for the 2017 CinC challenge. A strategy combining two methods was also evaluated.

2. Material and method

2.1. Data

In a sub-study of the UK Biobank study, 97,952 ECGs were obtained in 95,182 subjects (2,770 subjects had the test twice) using a 4-lead electrocardiograph device (CAMUSB 6.5, Cardiosoft v6.51, with two electrocardiograph electrodes on each upper limb). The ECG leads were recorded with a 500Hz sampling frequency at 5μV per Least Significant Bits (LSB) resolution. No manual annotations of the recordings were provided. The data were stored and exported in xml (Extended Markup Language) files; however this process unfortunately failed on a number of occasions. Ultimately, the raw ECG signals could only be retrieved for a subset of this population (78.8%); 77,202 ECG signals were extracted from 75,778 individual subjects.

ECG acquisition was performed before and during a submaximal exercise test on a stationary bicycle (eBike Comfort Ergometer, General Electric, firmware version 1.7) the intensity of which was adapted to the individual (based on age, gender, weight and medical history). The study protocol consisted of three consecutive phases, including a 15-second period of rest, a 6-minute period of increasing activity (cycling) and finally a 1-minute recovery period. Figure 1 shows the typical evolution of the heart rate during the ECG exercise test.

Subjects were stratified into different cardiovascular risk categories according to multiple factors, subjects with higher risk (9,611 subjects) only undergoing a two-minute resting ECG, as shown by the histogram of the recording length in figure 1. All the available ECG data were included in this study, meaning that both resting and exercise ECG were processed and analysed.

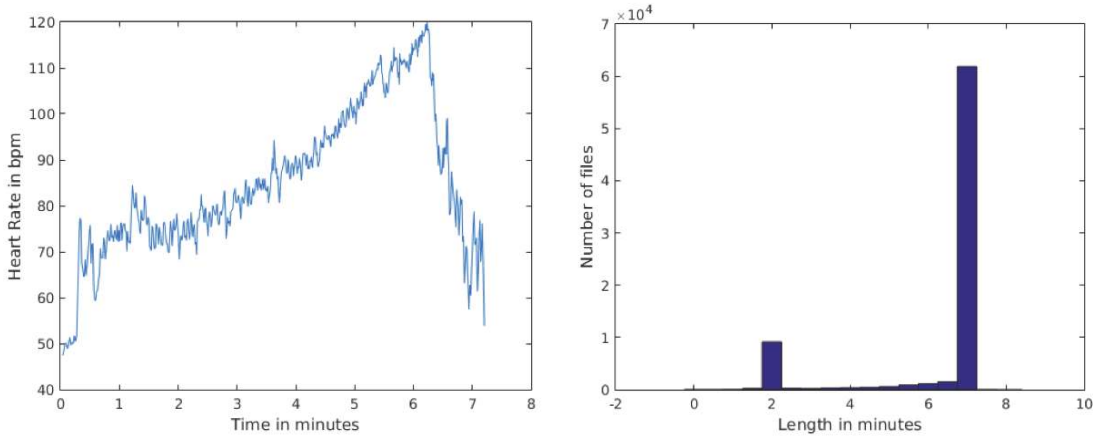

Figure 1. Heart rate during the exercise protocol (left). Histogram of the length of the 77,202 ECG recordings available (right).

2.2. Method

The proposed methodology will be described in the following sections, and is illustrated in figure 2. A subset of the UK Biobank data was first selected for manual annotations by three experts. Automated AF detection algorithms were then trained on the 2017 PhysioNet/CinC challenge database and then evaluated on the subset of manually annotated UK Biobank data.

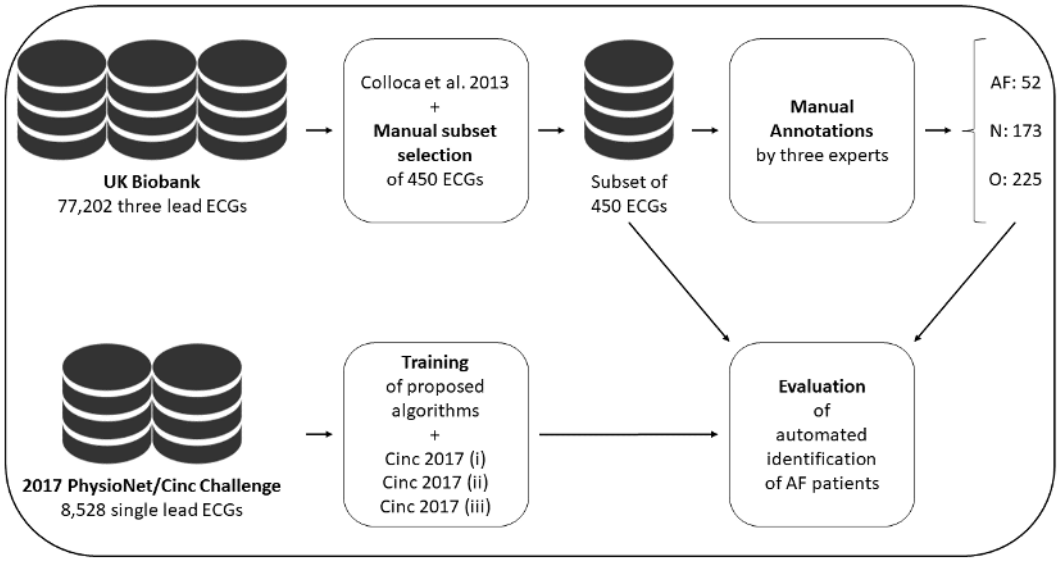

Figure 2. Flowchart of the proposed methodology. The top row illustrates the expert annotation process including the manual subset selection, while the bottom row represents the automated identification of AF patients with the evalutation performed on a subset of the UK Biobank database, and the training of the evaluated techniques performed on the 2017 PhysioNet/Cinc challenge database.

2.2.1. Expert annotations Heartbeat detection and rhythm classification can be validated by comparing the results of automated analysis with manual annotations generated by experts. The size of this database makes it prohibitively difficult to recruit multiple expert cardiologists to visually inspect and manually annotate all the available ECGs. It was therefore decided to select a small subset of the population, whose ECG were then manually annotated. The selection of these files is an important issue as the prevalence of AF is relatively low. An entirely random selection of files to annotate would not necessarily ensure sufficient representation of the AF sub-population.

As a first selection step, an automated method was applied in order to identify potential subjects with AF. This automated pre-selection technique combines predictability and regularity features of the RR intervals within a given window (Lake & Moorman 2011, Sarkar et al. 2008), using a machine learning approach (Colloca et al. 2013). A measure of the predictability of the signal can be derived from its Sample Entropy as suggested previously (Lake & Moorman 2011). Irregular rhythms are characterised by (rhythm) specific patterns in the Poincaré plot representation of the first-order derivative of the RR interval time-series (Sarkar et al. 2008). A Support Vector Machine (SVM) classifier was trained on the freely available MIT Atrial Fibrillation database (Goldberger et al. 2000). A grid-search was conducted to find the optimal parameters (window size, C, γ,...) for classifying a window as being positive. The classifier was trained to obtain a 99% sensitivity score on the training set. Applying this approach on overlapping windows will lead to a large number of false positives (given the high sensitivity of the approach and the multiple analyses on a single record, the possibility of false detection on a single record is increased), but will aim to identify the majority of subjects with AF correctly. This first selection step identified more than 5,000 ECG records. These were first reviewed by a non-expert ECG specialist, who divided them as showing potential atrial arrhythmia, ventricular arrhythmia, or presumed normal rhythm.

450 files were then selected for visual inspection and manual annotations by human experts to estimate the level of accuracy for the identification of patients with AF episodes: 150 files with apparent atrial arrhythmia, 150 files with apparent ventricular arrhythmia, 50 files with apparent normal rhythm, and finally 100 randomly selected files that had not been pre-reviewed.

The following set of rules was adopted for the classification of arrhythmic episodes following visual inspection and annotation; an episode was classified as:

- Supraventricular tachycardia, when there were at least three consecutive narrow QRS complexes ($< 120ms$) with a Heat Rate (HR) over 100 beats per minute (bpm).
- Atrial fibrillation or flutter if the characteristic arrhythmia lasted at least three consecutive heartbeats.
- Ventricular bigeminy if, for at least four consecutive beats, a normal QRS complex was followed by a premature ventricular complex (PVC).

- Ventricular trigeminy if, for at least six consecutive beats, two normal QRS complexes were followed by a PVC.
- Ventricular and atrial ectopic beats were ignored if isolated, but if more than one ectopic beat was present over a ten-second window, the episode was annotated as "ventricular ectopics" or "atrial ectopics".
- In all other cases, the rhythm was classified as normal sinus rhythm, which included sinus tachycardia (present during exercise).

Two experts visually analysed and manually annotated this subset of 450 ECG records. Finally, a third expert was invited to adjudicate whenever there was a discrepancy between the AF annotations of the first two experts.

The inter-observer agreement between the first two experts, who acted independently, was evaluated with Cohen's Kappa coefficient, with special emphasis on the AF annotations.

The 450 subjects were then classified in three groups:

(i) Atrial Fibrillation, when there is at least one AF episode.

(ii) Other rhythm, when there is at least one episode of ventricular or atrial ectopics.

(iii) Normal rhythm in all other cases.

2.2.2. Automated Data Analysis This subsection describes the automated analysis of the ECG data. The processing starts with the detection of R peaks, which will allow the extraction of both a Signal Quality Index and the RR interval time-series. The latter will then be used to extract features that will be used as inputs to the AF classifier. A flowchart of the processing is given in figure 3.

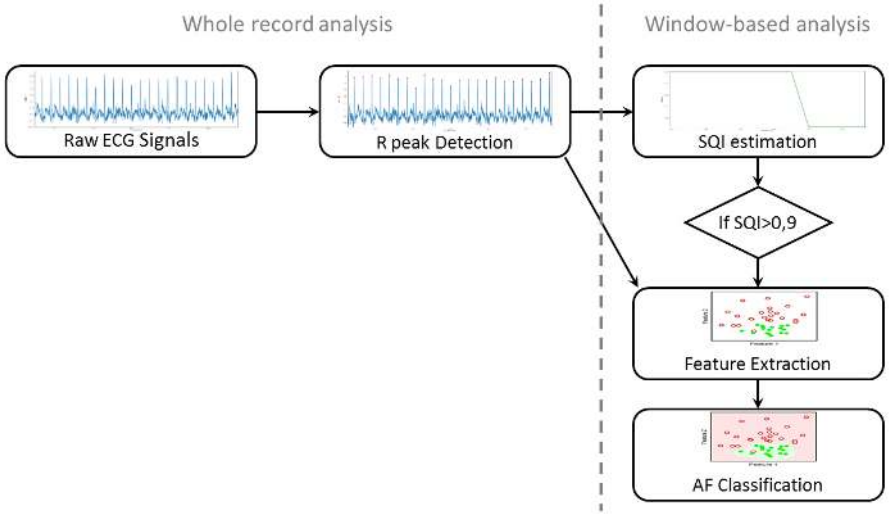

Figure 3. Flowchart of the data processing.

R-peak detections Analysis of ECG data usually starts with the estimation of the R-peak position. This task is normally relatively easy, as the R-peak is the most distinguishable feature of the ECG signal. The data can however be extremely noisy during exercise recordings, with large artefacts due to electrode motion and muscle artefacts superimposed on the ECG signal. The PhysioNet/Computing in Cardiology challenge in 2014 was dedicated to the problem of R-peak detection in multimodal data (Silva et al. 2015). Although the UK Biobank fitness dataset only contains ECG data, the solutions developed during the 2014 competition are also well suited for multi-lead recordings, and hence we used the Challenge-winning approach (Johnson et al. 2014, Johnson et al. 2015). R-peaks were estimated on each of the three leads, using a peak energy detector based on a modified Pan-Tompkins algorithm. A second QRS detector, freely available on PhysioNet and called "gqrs" was applied to each lead. This detector consists of a matched filter, with a set of custom-built heuristic rules (Goldberger et al. 2000). The level of agreement between the two detectors was used as a proxy for the signal quality index (SQI). This simple SQI assessed every second on a 10-second sliding window for each lead is robust to pathological rhythm (Behar et al. 2013). Detection was then performed on the lead having the highest SQI. A further test was carried out in order to avoid double detections of the same peak.

Rhythm classification, Atrial Fibrillation The proposed technique takes inspiration from Colloca's technique (Colloca et al. 2013). A rhythm-based approach was adopted as analysis of the ECG morphology is not possible during exercise due to the presence for instance of muscle artefacts, and accurate extraction of f-waves is impossible. It was decided to restrict the analysis to only five features: the first consists in a coefficient derived from the sample entropy of the RR-interval time series, denoted COSEn and proposed in (Lake & Moorman 2011). The four other features are presented in (Sarkar et al. 2008), and are derived from the Poincaré representation of the RR-interval time series (See figure 4). This Poincaré representation space is then divided in several bins following a Union-Jack pattern, and the number of heartbeats in each bin is counted. We included the evidence of AF (AFE), which is a combined coefficient of three subfeatures and was proposed by the authors for the classification of AF (Sarkar et al. 2008). We also included the following three subfeaures: (i) the numbers of beats in the Origin bin (OrC), representing the number of normal heartbeats; (ii) Irregularity Evidence (IrE), which counts the number of irregular heartbeats; and (iii) Premature Atrial Contraction Evidence (PACE), which counts the number of heartbeats which seem indicative of atrial arrhythmia. The features based on the heart rate were excluded given the inherent presence of tachycardia during exercise. A SVM with a radial basis function (RBF) kernel was used as the classifier, which was trained on the 2017 PhysioNet/Computing in Cardiology challenge training dataset (Clifford et al. 2017). The classifiers were trained using cross-validation. Random search was performed to determine the SVM hyperparameters: soft margin constant C and RBF kernel hyperparameter γ.

The AF detector was applied to 60-s sliding windows with a 30s overlap. The use

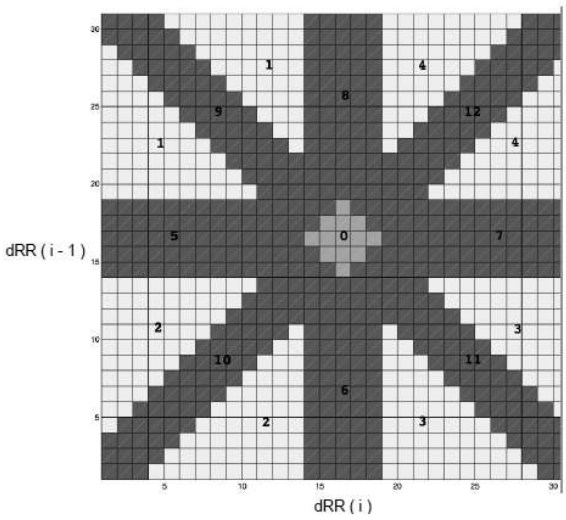

Figure 4. Poincaré representation of the first derivative of the RR intervals. The origin bin is number 0. Irregularity evidence is derived by counting the number of heartbeats in bins numbered 1 to 12. PACE is derived by comparing the number of heartbeats in bins 5, 6, 7, 8, 10 and 12 with the number of heartbeats in bins 1, 2, 3 and 4.

of sliding windows should provide increased sensitivity, but will also be accompanied with an increased false positive rate. The length of the window was set to 60s to be comparable to the 65 beat window proposed by (Colloca et al. 2013) and which was shown to give good classification performance.

Six different versions of the proposed approach were implemented. All of these versions are based on the same SVM classifer but contain extra post-processing stages built to reduce the false positive rate:

(i) Applying the SVM classifier without any post-processing step, later denoted as "proposed (raw)".

(ii) The impact of noise on RR-based AF detection was recently demonstrated in (Oster & Clifford 2015), the level of performance decreasing with increasing level of noise. The authors suggested the use of an SQI to exclude noisy segments and ensure a certain level of performance. The analysis of noisy windows was therefore excluded; the 10^{th} percentile of the SQI over the window was compared with a threshold of 0.9 as suggested by (Oster & Clifford 2015). This version will be denoted as "proposed (sqi)"

(iii) The use of a sliding window could introduce an increase in the false positive rate. The SVM classifier was trained to allow for the detection of short episodes of AF (or occult paroxysmal atrial fibrillation). Nevertheless, the shorter the AF episodes targeted, the more likely the classifier is to detect noisy episodes as well as other types of arrhythmia (Petrėnas et al. 2015). In order to ensure longer segments are detected, or that occasional false detections are suppressed, AF episodes were only considered if at least two consecutive windows were classified as AF by the

above version of the algorithm (proposed (sqi)). This iteration will be denoted as "proposed (persistence)". It has to be noted that the minimum duration of an AF episode for automated detection is not known. The specification of a minimum duration criterion could be investigated by using synthetic data generated by a realistic model for example.

(iv) It was shown that AF detection approaches based on RR-based features such as the COSEn or the Poincaré plot are over-sensitive to normal rhythms with frequent ectopy, either atrial or ventricular (Colloca et al. 2013, Petrėnas et al. 2015). There is therefore a need to develop an intermediate method to reject individuals identified as having AF, but whose arrhythmia alarms are triggered by the presence of premature ventricular contractions. Although RR-based approaches have been suggested to discriminate between AF and ectopics (Carrara et al. 2015), a morphological analysis was performed to detect ventricular heartbeats. This analysis was based on the two novel methods presented by (Oster & Tarassenko 2016). The first one is based on state-of-the-art features (including RR-based temporal features, morphological features such as QRS width or maximal vector of the QRS loop and its angle) (Llamedo & Martínez 2011, Llamedo & Martínez 2012, Llamedo 2015) and a SVM classifier; while the second one applied Bayesian filtering (or Switching Kalman Filter -SKF-) to classify the heartbeats according to their morphology, with automated annotation of a small number of clusters (Oster & Tarassenko 2016).
If at least 10 ventricular beats were detected within a given AF window (as detected by the version "proposed (sqi)), the window was rejected. This led to two versions of the proposed method denoted "proposed (morphology SVM)" or "proposed (morphology SKF)" according to the ventricular heartbeat detection algorithm being used.

(v) Finally, the proposed method was also applied to the whole ECG recording instead of being applied to sliding windows. This will be denoted as "proposed (whole signal)". This approach will enable the assessment of how the application to short time windows allows for increased sensitivity but also an increased number of false positives.

State-of-the-art algorithms - Cinc 2017 The proposed processing process was then compared with state-of-the-art approaches, and especially some of the 2017 PhysioNet/Cinc challenge entries:

(i) one of the top entries with a challenge score (average of three F1 scores) of 0.83 (Datta et al. 2017) used hand-crafted features with a boosting algorithm. This technique consisted in a two-stage classifier, the first one for separating Noisy and AF from Normal and Other rhythms ECG signals. The second stage consisted of two classifiers trained to identify the cardiac rhythm. The processing starts with the automated identification and suppression of noisy segments from ECG signals,

based on spectrogram analysis. The authors then extract more than 150 features from the ECG signal: morphological features, AF features (including CosEn (Lake & Moorman 2011) and AFE (Sarkar et al. 2008)), heart rate variability (HRV) features, spectral features and statistical features. These features are then selected using minimum redudancy maximum relevance, and finally the classifier is trained using the adaBoost technique.

(ii) another entry used hand-crafted features as inputs to a cascaded SVM (with RBF kernel) classifier approach (score 0.80) (Behar et al. 2017). This technique consisted in the extraction of almost 50 features for (i) signal quality (ii) ECG morphology (iii) predictability of the RR intervals and (iv) HRV. A cascaded approach was then taken, by training three consecutive SVM classifiers, the first identifying normal rhythm, the second AF rhythms, and the last distinguishing noisy signals from other rhythms. Each SVM used a radial basis function kernel, and was trained using repeated cross-validation, and random search for the hyperparameters (C and γ).

(iii) a deep-learning approach, combining convolutional and recurrent networks, was trained on the 2017 PhysioNet/Cinc challenge dataset and obtained a score of 0.85 during cross-fold validation using the same training set (Vogt 2018). The network architecture consisted in a convolutional neural network (CNN), with 15 layers of a block with Batch Normalisation, a convolution layer, a ReLu activation layer and average pooling. The CNN was followed by a Long Short-Term Memory (LSTM) network with 64 hidden units. This network contained about 3.7 million parameters to be optimised. A two-stage procedure was used for training the overall network. First the CNN was trained by replacing the LSTM with Global Max Pooling. This enabled the CNN to learn a good representation of the ECG signal. A second training phase consisted in training the LSTM layer, while also fine tuning the CNN (by adjusting the learning rate).

All three methods were applied to 60s-windows with an overlap of 30s using only the first lead.

Finally the aggregation of multiple entries was shown to be able to increase the classification performance for the 2017 PhysioNet/Computing in Cardiology challenge (Clifford et al. 2017). We decided to combine the best performing state-of-the-art technique (Vogt 2018) with the best performing proposed technique (proposed (whole signal)) to assess the potential of such a voting approach. Unanimous voting is needed to identify a subject with AF, and the method will later be denoted as "combined". This approach is equivalent to the combination of a high sensitivity screening tool with a second technique with a high Positive Predictive Value (PPV) for discarding most of the false detections.

The different methods for AF classification were then compared using the sensitivity, the positive predictive value, and the F1 score, which is the harmonic mean of Se and PPV. Maximising F1 leads to a good compromise between Se and PPV.

3. Results

3.1. Manual annotations

Table 1 summarises the manually-annotated classes for the subset of 450 subjects of the UK Biobank population.

Class	Normal rhythm	Atrial Fibrillation	Other Rhythm	Total
Number of subjects	173	52	225	450

Table 1. Manually annotated labels for the subset of 450 subjects from the UK Biobank

Table 2 highlights the inter observer agreement for different rhythms. Cohen's kappa coefficient for inter-observer agreement on AF was 0.78. On this small subset, experts disagreed for 23 subjects with suspected AF episodes (out of 450). Further analysis showed that the disagreement came from potential AF episodes in short and noisy segments. One expert tended to annotate these short suspected AF episodes with a low level of confidence (as self-rated), whereas the second tended to discard these episodes as noise. The adjudication process identified 3 additional subjects with suspected AF.

The final expert annotations in this study classified 52 subjects with AF out of 450 subjects. The 12% prevalence of AF in this small dataset is clearly an over-estimate, but is due to non-random selection of files in the subset.

	AF		Ventricular ectopics		Atrial ectopics	
	P_{obs2}	N_{obs2}	P_{obs2}	N_{obs2}	P_{obs2}	N_{obs2}
P_{obs1}	49	8	181	6	76	27
N_{obs1}	15	378	38	225	54	293

Table 2. Inter-observer agreement on different rhythm annotations, where P_{obsi} (resp. N_{obsi}) stands for positive (resp. negative) subjects as identified by the i^{th} expert

An example of a file for which the experts disagreed is depicted on figure 5. In the zoomed-in example, a short episode of supraventricular arrhythmia is visible, for which the presence or not of a P wave is difficult to determine.

3.2. Automated Identification of AF patients

The AF detection statistics on this sub-population are given in table 3. It can be seen that all techniques have a relatively high sensitivity, but that comes at the cost of a relatively poor positive predictive value. The proposed SVM (raw) technique offers the best sensitivity; however as shown by the addition of a check on the signal quality this is counterbalanced by a large number of false positives generated by noise. Adding morphological hand-crafted features to discard ventricular ectopics (SVM (iv) a and b),

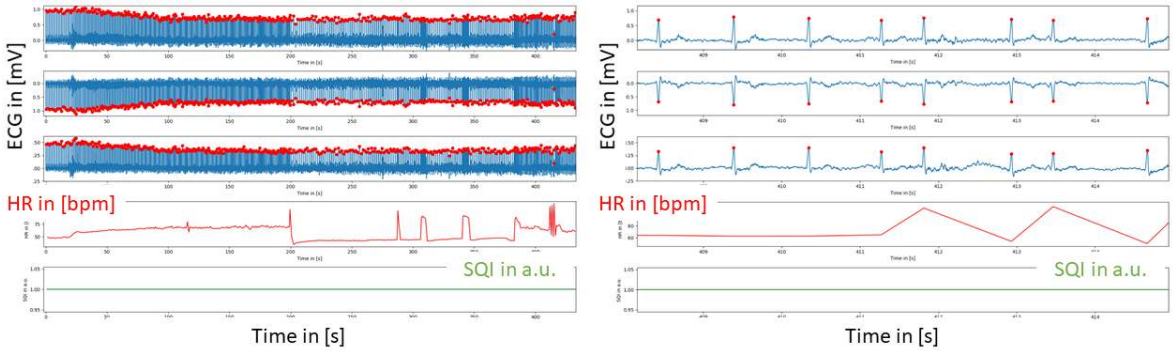

Figure 5. Example of signals with disagreement between the two expert annotators. (From top to bottom: ECG leads 1 to 3, instantaneous Heart Rate and signal quality index.)

Cinc 2017 (i) (Behar et al. 2017) and CinC 2017 (ii) (Datta et al. 2017) increases the PPV and the F1 score reaches values near 60%. The SVM-based and CinC 2017 (i) and (ii) methods have comparable performance.

The deep-learning technique (Cinc 2017 (iii)) outperforms the classical SVM method with hand-crafted features, obtaining a PPV of 60%, and an F1 score of 74%. Applying the SVM-based classifier on the whole record (SVM (v)) does significantly decrease the number of false positives, with a PPV of 83.3%, and leads to an increased F1 score (80%) despite a reduction in the sensitivity (77%). As shown by (Clifford et al. 2017), combining several techniques increases the overall performance, with a combined method (SVM (v) combined with Cinc 2017 (iii)) reaching an F1 score of 85%.

	Method	Se %	PPV %	F1 %
SVM (i)	proposed (raw)	**100**	28.6	44.4
SVM (ii)	proposed (sqi)	90.4	38.5	54.0
SVM (iii)	proposed (persistence)	88.5	50.0	63.9
SVM (iv) a	proposed (morphology SKF)	86.5	47.4	61.2
SVM (iv) b	proposed (morphology SVM)	78.9	51.3	62.1
SVM (v)	proposed (whole record)	76.9	83.3	80.0
CinC 2017 (i)	(Behar et al. 2017)	94.2	35.8	51.9
CinC 2017 (ii)	(Datta et al. 2017)	92.3	43.2	58.9
CinC 2017 (iii)	(Vogt 2018)	96.2	60.2	74.1
SVM (v) + CinC 2017 (iii)	combined	75.0	**97.5**	**84.8**

Table 3. AF detection performance on the manually annotated sub-population of 450 subjects.

Cohen's Kappa coefficient for the combined method is 0.83, which is similar to (even slightly higher than) the inter-observer agreement reached by two experts. The confusion matrix for the combined method is given in table 4.

	SVM (v) proposed (whole record)		CinC 2017 (iii) (Vogt 2018)		SVM (v) + CinC 2017 (iii) combined	
	P_{pred}	N_{pred}	P_{pred}	N_{pred}	P_{pred}	N_{pred}
P_{true}	40	12	50	2	39	13
N_{true}	8	390	33	365	1	397

Table 4. Confusion tables for three different approaches, where P_{true} (resp. N_{true}) stands for positive (respectively negative) AF subjects as observed by human experts and P_{pred} (respectively N_{pred}) are positive (respectively negative) AF subjects as predicted by automated techniques.

Examples of files where the different methods disagreed are shown in figures 6 to 8. Some methods, including the deep-learning method (CinC 2017 (iii)) are able to recognise ventricular ectopics in the first example (figure 6). In figure 7 supraventricular arrhythmia associated with a high level of noise makes it hard to decide whether P waves are present or not, but the deep-learning method (CinC 2017 (iii)) again correctly identifies the record as an example of non-AF. Similarly, in figure 8, the only method that correctly identifies the record as an example of non-AF is the deep-learning method (Cinc 2017 (iii))

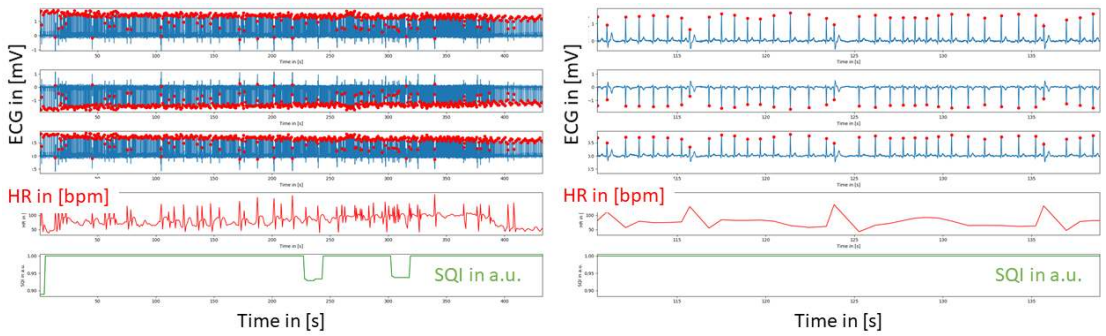

Figure 6. Example of signals with disagreement between the different methods. (From top to bottom: ECG leads 1 to 3, instantaneous Heart Rate and signal quality index. Left-hand side is the depiction of the whole signal, while the righ-hand side is a zoomed-in section of the signal.) CinC 2017 (i) (Behar et al. 2017) and CinC 2017 (iii) (Vogt 2018) correctly identify this record as non-AF.

4. Discussion

All the methods investigated in this paper (classical SVM-based approaches and the state-of-the-art methods at the 2017 Computing in Cardiology conference, including a deep-learning method) were trained on the 2017 PhysioNet/Computing in Cardiology challenge. The UK Biobank data, which we used as evaluation or test data, was recorded during a stress-test exercise, which has two main implications, (i) the presence of high-level noise makes the use of morphological features quite difficult as f waves can hardly

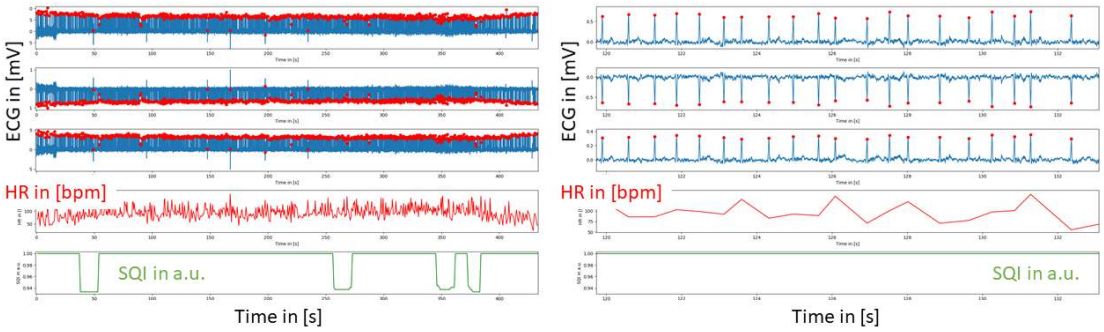

Figure 7. Example of signals with disagreement between the different methods. (From top to bottom: ECG leads 1 to 3, instantaneous Heart Rate and signal quality index. Left-hand side is the depiction of the whole signal, while the righ-hand side is a zoomed-in section of the signal.) CinC 2017 (ii) (Datta et al. 2017) and CinC 2017 (iii) (Vogt 2018) only correctly identify this record as non-AF.

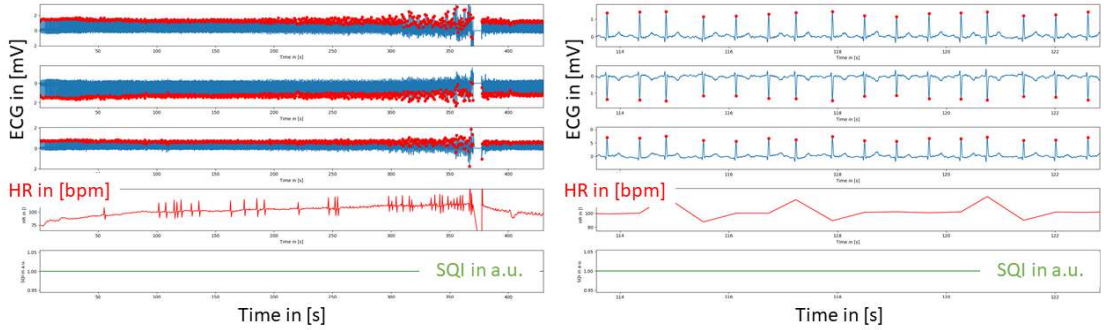

Figure 8. Example of signals with disagreement between the different methods. (From top to bottom: ECG leads 1 to 3, instantaneous Heart Rate and signal quality index. Left-hand side is the depiction of the whole signal, while the righ-hand side is a zoomed-in section of the signal.) CinC 2017 (iii) (Vogt 2018) correctly identifies this record as non-AF.

be distinguished from muscle artefacts; (ii) as most subjects were resting, exercising and recovering during the ECG recording, the signal is non-stationary. The latter will have an impact on features measuring the regularity of the interval (Liu et al. 2018). Ideally, the machine-learning algorithms should have been trained on a subset of the UK Biobank dataset or at least some kind of transfer learning (Pan et al. 2010) ought to be used when training takes place on a dataset with a signal-to-noise ratio known to be different to that of the test set.

One of the main issues for the analysis of the UK Biobank data lies in the size of the dataset. Expert label annotations or interpretation of such data is very difficult to obtain as these tasks are time-consuming. Furthermore, regardless of the skill level, there will always be disagreements between experts, especially when the data in question is noisy (due to exercise-induced noise) or incomplete (missing leads or no knowledge of the patient history).

The results show that window-based hand-crafted rhythm-based features as the

inputs to a classical machine learning algorithm (SVM) provide limited performance with a maximum F1 value of 64%. The addition of morphological analysis to hand-crafted morphological features did not lead to improvements in F1 score, 62% (SVM (iv) b) or 58.9% (CinC 2017 (ii) (Datta et al. 2017)).

Analysis of the whole record with hand-crafted rhythm-based features as the inputs to the SVM classifier was shown to lead to an acceptable F1 score of around 80%. The use of overlapping windows leads to a higher sensitivity but at the cost of higher numbers of false positives. Analysing the whole record can however be problematic in cases when an AF episode is present in a very noisy record, for which very few RR intervals can be estimated. An alternative approach would consist in excluding noisy segments from the analysis, and only performing a single classification per record.

A deep-learning approach (Vogt 2018), trained on almost 10,000 subjects from the CinC 2017 Challenge, automatically extracts features, representing both morphological and rhythm-based information. This approach yields a PPV of more than 60% while only missing 2 AF subjects in the validation subset. Applying this approach on the whole dataset would yield too many false positives, but the deep-learning algorithm can be combined with a classical machine-learning approach to increase performance.

Classical machine-learning techniques rely on the development of carefully hand-crafted features. This crafting of features depends on the expertise of the data scientists, used to incorporate prior knowledge of the pathology; in the case of AF the chaotic heart rhythm associated with the pathology. The use of hand-crafted features offers the advantage of an easier interpretability of decisions, as individual feature values can give clinicians clues to the decision process.

Deep-learning algorithms are purely data-driven solutions, meaning that almost no prior knowledge is required to develop a working solution (apart from the network architecture). They require a large amount of data to be trained effectively, however, when they are (as can be seen when trained with the large dataset of the 2017 CinC challenge) they can achieve excellent performance. The drawback of these techniques lies in the fact that no insights on the decision process can be easily extracted, although there has been recent progress in the interpretability of deep-learning networks (Vogt 2018).

The combined method (SVM (v) + CinC 2017 (iii)) yields a good level of performance, with sensitivity, PPV and F1 score at around 85%. Cohen's Kappa coefficient relating the prediction from the combined method to the expert labels reaches almost 83%, which is slightly higher than the inter-observer agreement between two human experts. The PPV value for the combined method is also high, showing that patients who have been identified as having AF do indeed have the arrhythmia. Further analysis of the automatically identified AF patients will help to identify different phenotypes in order to refine the risk factors for AF.

In a further test, the approach which combines the SVM-based method applied to the entire record with the deep-learning method was applied to the whole UK-Biobank dataset, and 629 recordings were identified as containing AF episodes, which represents a prevalence of 0.8%. One explanation for this low prevalence comes from the fact that

the recordings are relatively short, and that paroxysmal AF can easily be missed in such short recordings. A recent study on home-based monitoring for AF detection has shown that many AF episodes can only be detected after more than one week of monitoring (Steinhubl et al. 2018), making the case for long-term home cardiac monitoring.

Acknowledgments

This research has been conducted using the UK Biobank Resource under Application Number 9119.

This study has been supported by the NIHR Oxford Biomedical Research Centre, *Technology and Digital Health theme*. JO was also funded by the Royal Academy of Engineering through a Newton International Fellowship Alumni programme (AL/180051). BC and RW are funded by the British Heart Foundation and the NIHR Oxford Biomedical Research Centre. KZ and CFC are funded by the UK Medical Research Council. JCH is supported by a fellowship from the British Heart Foundation (FS/14/55/30806).

References

Andreotti, F., Carr, O., Pimentel, M. A., Mahdi, A. & De Vos, M. (2017). Comparing Feature-Based Classifiers and Convolutional Neural Networks to Detect Arrhythmia from Short Segments of ECG, *Computing* **44**: 1.

Behar, J. A., Rosenberg, A. A., Yaniv, Y. & Oster, J. (2017). Rhythm and Quality Classification from Short ECGs Recorded Using a Mobile Device, *Computing* **44**: 1.

Behar, J., Oster, J., Li, Q. & Clifford, G. D. (2013). ECG signal quality during arrhythmia and its application to false alarm reduction, *IEEE Transactions on Biomedical Engineering* **60**(6): 1660–1666.

Carrara, M., Carozzi, L., Moss, T. J., de Pasquale, M., Cerutti, S., Ferrario, M., Lake, D. E. & Moorman, J. R. (2015). Heart Rate Dynamics distinguish among Atrial Fibrillation, Normal Sinus Rhythm and Sinus Rhythm with frequent Ectopy, *Physiological Measurement* **36**(9): 1873.

Clifford, G. D., Liu, C., Moody, B., Lehman, L.-w. H., Silva, I., Li, Q., Johnson, A. & Mark, R. G. (2017). AF classification from a short single lead ECG recording: The Physionet Computing in Cardiology Challenge 2017, *Proceedings of Computing in Cardiology* **44**: 1.

Collins, R. (2012). What makes UK Biobank special?, *The Lancet* **379**(9822): 1173–1174.

Colloca, R., Johnson, A. E., Mainardi, L. & Clifford, G. D. (2013). A Support Vector Machine approach for reliable detection of atrial fibrillation events, *Computing in Cardiology Conference (CinC), 2013*, IEEE, pp. 1047–1050.

Datta, S., Puri, C., Mukherjee, A., Banerjee, R., Choudhury, A. D., Singh, R., Ukil, A., Bandyopadhyay, S., Pal, A. & Khandelwal, S. (2017). Identifying Normal, AF and other Abnormal ECG Rhythms using a cascaded binary classifier, *Computing* **44**: 1.

Davis, R. C., Hobbs, F. R., Kenkre, J. E., Roalfe, A. K., Iles, R., Lip, G. Y. & Davies, M. K. (2012). Prevalence of Atrial Fibrillation in the general population and in high-risk groups: the ECHOES study, *Europace* **14**(11): 1553–1559.

Fuster, V., Rydén, L. E., Cannom, D. S., Crijns, H. J., Curtis, A. B., Ellenbogen, K. A., Halperin, J. L., Kay, G. N., Le Huezey, J.-Y., Lowe, J. E. et al. (2011). 2011 ACCF/AHA/HRS focused updates incorporated into the ACC/AHA/ESC 2006 Guidelines for the management of patients with atrial fibrillation: a report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines developed in partnership with the European

Society of Cardiology and in collaboration with the European Heart Rhythm Association and the Heart Rhythm Society, *Journal of the American College of Cardiology* **57**(11): e101–e198.

Ganna, A. & Ingelsson, E. (2015). 5 year mortality predictors in 498 103 UK Biobank participants: a prospective population-based study, *The Lancet* **386**(9993): 533–540.

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., Mietus, J. E., Moody, G. B., Peng, C.-K. & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals, *Circulation* **101**(23): e215–e220.

Hannun, A. Y., Rajpurkar, P., Haghpanahi, M., Tison, G. H., Bourn, C., Turakhia, M. P. & Ng, A. Y. (2019). Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network, *Nature medicine* **25**(1): 65.

Johnson, A. E., Behar, J., Andreotti, F., Clifford, G. D. & Oster, J. (2014). R-peak estimation using multimodal lead switching, *Computing in Cardiology Conference (CinC), 2014*, IEEE, pp. 281–284.

Johnson, A. E., Behar, J., Andreotti, F., Clifford, G. D. & Oster, J. (2015). Multimodal heart beat detection using signal quality indices, *Physiological measurement* **36**(8): 1665.

Jones, C., Pollit, V., Fitzmaurice, D., Cowan, C. et al. (2014). The management of atrial fibrillation: summary of updated NICE guidance, *BMJ* **348**: g3655.

Kannel, W. B. & McGee, D. L. (1979). Diabetes and cardiovascular disease: the Framingham study, *Jama* **241**(19): 2035–2038.

Lake, D. E. & Moorman, J. R. (2011). Accurate estimation of entropy in very short physiological time series: the problem of atrial fibrillation detection in implanted ventricular devices, *American Journal of Physiology-Heart and Circulatory Physiology* **300**(1): H319–H325.

Liu, C., Oster, J., Reinertsen, E., Li, Q., Zhao, L., Nemati, S. & Clifford, G. D. (2018). A comparison of entropy approaches for AF discrimination, *Physiological measurement* .

Llamedo, M. (2015). ECG-Kit: A Matlab toolbox for cardiovascular signal processing.
URL: *http://marianux.github.io/ecg-kit/*

Llamedo, M. & Martínez, J. P. (2011). Heartbeat classification using feature selection driven by database generalization criteria, *Biomedical Engineering, IEEE Transactions on* **58**(3): 616–625.

Llamedo, M. & Martínez, J. P. (2012). An automatic patient-adapted ECG heartbeat classifier allowing expert assistance, *Biomedical Engineering, IEEE Transactions on* **59**(8): 2312–2320.

Nichols, M., Townsend, N., Scarborough, P. & Rayner, M. (2013). Cardiovascular disease in Europe: epidemiological update, *European heart journal* **34**(39): 3028–3034.

Oster, J. & Clifford, G. D. (2015). Impact of the presence of noise on RR intervals-based Atrial Fibrillation detection, *Journal of Electrocardiology* .

Oster, J. & Tarassenko, L. (2016). Automated ECG ventricular beat detection with switching Kalman filters, *Computing in Cardiology Conference (CinC), 2016*, IEEE, pp. 37–40.

Palmer, L. J. (2007). UK Biobank: bank on it, *The Lancet* **369**(9578): 1980–1982.

Pan, S. J., Yang, Q. et al. (2010). A survey on transfer learning, *IEEE Transactions on knowledge and data engineering* **22**(10): 1345–1359.

Petrėnas, A., Sörnmo, L., Lukoševičius, A. & Marozas, V. (2015). Detection of occult paroxysmal Atrial Fibrillation, *Medical & biological engineering & computing* **53**(4): 287–297.

Sarkar, S., Ritscher, D. & Mehra, R. (2008). A detector for a chronic implantable atrial tachyarrhythmia monitor, *Biomedical Engineering, IEEE Transactions on* **55**(3): 1219–1224.

Silva, I., Moody, B., Behar, J., Johnson, A., Oster, J., Clifford, G. D. & Moody, G. B. (2015). Robust detection of heart beats in multimodal data, *Physiological measurement* **36**(8): 1629.

Sörnmo, L. (2018). *Atrial Fibrillation from an Engineering Perspective*, Springer.

Steinhubl, S. R., Waalen, J., Edwards, A. M., Ariniello, L. M., Mehta, R. R., Ebner, G. S., Carter, C., Baca-Motes, K., Felicione, E., Sarich, T. et al. (2018). Effect of a home-based wearable continuous ECG monitoring patch on detection of undiagnosed Atrial Fibrillation: the mSToPS

randomized clinical trial, *JAMA* **320**(2): 146–155.

Sudlow, C., Gallacher, J., Allen, N., Beral, V., Burton, P., Danesh, J., Downey, P., Elliott, P., Green, J., Landray, M. et al. (2015). UK Biobank: an Open Access resource for identifying the causes of a wide range of complex diseases of middle and old age, *PLoS medicine* **12**(3): 1–10.

Teijeiro, T., García, C. A., Castro, D. & Félix, P. (2017). Arrhythmia classification from the abductive interpretation of short single-lead ECG records, *Comput. Cardiol* **44**: 1–4.

Vogt, N. (2018). CNNs, LSTMs, and Attention Networks for Pathology Detection in Medical Data, *arXiv preprint arXiv:1912.00852* .

Wolf, P. A., Abbott, R. D. & Kannel, W. B. (1991). Atrial fibrillation as an independent risk factor for stroke: the Framingham Study., *Stroke* **22**(8): 983–988.