ArNet-ECG: Deep Learning for the Detection of Atrial Fibrillation from the Raw Electrocardiogram

Noam Ben-Moshe[1,2], Shany Biton[2], Joachim A Behar[2]

[1]Faculty of Computer Science, Technion-IIT, Haifa, Israel
[2]Faculty of Biomedical Engineering, Technion-IIT, Haifa, Israel

Abstract

Introduction: *Atrial fibrillation (AF) is the most prevalent heart arrhythmia. AF manifests on the electrocardiogram (ECG) though irregular beat-to-beat time interval variation, the absence of P-wave and the presence of fibrillatory waves (f-wave). We hypothesize that a deep learning (DL) approach trained on the raw ECG will enable robust detection of AF events and the estimation of the AF burden (AFB). We further hypothesize that the performance reached leveraging the raw ECG will be superior to previously developed methods using the beat-to-beat interval variation time series. Consequently, we develop a new DL algorithm, denoted ArNet-ECG, to robustly detect AF events and estimate the AFB from the raw ECG and benchmark this algorithms against previous work.* ***Methods:*** *A dataset including 2,247 adult patients and totaling over 53,753 hours of continuous ECG from the University of Virginia (UVAF) was used.* ***Results:*** *ArNet-ECG obtained an F_1 of 0.96 and ArNet2 obtained an F_1 0.94.* ***Discussion and conclusion****: ArNet-ECG outperformed ArNet2 thus demonstrating that using the raw ECG provides added performance over the beat-to-beat interval time series. The main reason found for explaining the higher performance of ArNet-ECG was its high performance on atrial flutter examples versus poor performance on these recordings for ArNet2.*

1. Introduction

Atrial fibrillation (AF) is the most prevalent heart arrhythmia [1, 2]. Today, there is up to 13% of individuals with AF are misdiagnosed [3]. High gaps in AF management exist including the lack of effective decision support tools for risk prediction and diagnosis. On the ECG, AF is characterized by an irregular beat-to-beat time interval variation (RR-interval) resulting from the fibrillation of the atria which excites the atrioventricular node at a very high rate. This in turns, causes the atrioventricular node to fire in a chaotic and highly irregular manner and which consequently leads to an irregular polarization of the ventricles. Furthermore, it affects the morphology of the sinus waveform in such a way that the 'traditional' P wave is replaced with the appearance of low amplitude f-waves. Atrial flutter (AFL) is another supraventricular arrhythmia which has some similarities in its clinical management to AF and so we included AFL within the AF group as previously done by others [4]. AFL results from an abnormal circuit inside the right atrium, or upper chamber of the heart. AFL is characterized by regular RR beats and the presence of flutter waves. In this research we develop a new deep learning (DL) algorithm for AF events detection from long continuous raw ECG. We hypothesize that the raw ECG may yield superior performance compared to using the beat-to-beat time series derived from the ECG since the raw ECG holds both morphological information and rhythm information. This research aims to create a DL algorithm, denoted ArNet-ECG, for the detection of AF events from long continuous ECG recordings.

2. Methods

2.1. Dataset

The University of Virginia dataset (UVAF) was used. The UVAF [4, 5] consists of ECG recordings of patients for whom the University of Virginia health system physicians ordered Holter monitoring from December 2004 to October 2010. This dataset contains $n = 2,247$ annotated recordings of individual patients over the age of 18 years old. Recordings were sampled at 200 Hz. The median and interquartile age for the recordings were 57 (40-71), recording durations 24 (24-24) hours long. UVAF was used to train and evaluate the DL models. In order to automatically assess the quality of the raw ECG files and discard those that were too noisy, we adopted the R-peak quality criterion bSQI [6]. More information about the dataset is described in [7]. Out of the UVAF dataset 100 recordings from unique patients were selected and used as test set. All other recordings were used as training set and validation set. The test set was re-annotated

by an intern cardiologist. Specifically, supraventricular arrhythmias were annotated in three categories: (1) AF; (2) AFL; (3) Other supraventricular tachycardias such as Wolf-Parkinson-White and intranodal tachycardias and (4) other such as NSR that where not labelled as was done in [8]. Furthermore, for the specific purpose of the main experiments presented in this work the annotations for AF and AFL were grouped under the single label denoted as AF_l.

2.2. Preprocessing

Recordings of patients under 18 years old were also excluded. Corrupted recordings were excluded as described in Chocron et al. [7]. This resulted in a total number of 2,237 recordings used for our experiments. Each recording was divided into 30 seconds non-overlapping windows. A total of 6,424,793 windows were available from the UVAF after the exclusion criteria were applied. For each window, we defined its rhythm label as the reference label most represented over this window.

2.3. AF_l groups definition

Patient were divided into four groups based on their AFB as defined in Chocron et al [7], i.e Non-AF, AF_{mild}, AF_{mod} and AF_{sev}. The groups were defined as follows: Non-AF: total time spent in AF below 30 seconds. Mild AF (AF_{mild}): total time spent in AF above 30 seconds [9] and AFB below 4% [10], Moderate AF (AF_{mod}): AFB in the range 4-80%, Severe AF (AF_{sev}): AFB above 80%.

2.4. ArNet-ECG

We used a Residual Network (ResNet) architecture [11] as was previously used in the context of arrhythmia classification taking as input the RR time series [8]. The model takes a window of 30 seconds of raw ECG as input and outputs a binary value (AF_l or non-AF_l). ArNet-ECG architecture is a stack of residual 1D-Convolutional (CNN) blocks followed by dense layers. Each block has 1D-CNN layers with a Batch Normalization (BN) layers followed by a Rectified Linear Unit (ReLU) activation and a shortcut connection with Max Pooling. The output of the residual blocks is then passed to three successive dense layers. The last dense layer of size 1 outputs a binary prediction for the label of the window. The loss function used was a weighted binary cross-entropy loss which gave more weight to the AF windows since the classes are highly imbalanced. During training we optimized with Adam [12] algorithm and used dropout regularization. One of the advances of residual blocks is the skip connections that enhance the rate of information transferred throughout the network by connecting layers earlier in the network with layers later in the network as was suggested in [13].

2.5. Performance statistics

The performance of the models were assessed as in Chocron et al. [7]. The following statistics were computed to assess the models performance on the individual 30 seconds windows: sensitivity (Se), specificity (Sp), positive predictive value (PPV), the area under the receiver operating characteristic (AUROC) and the harmonic mean of the precision and recall (F_1). The threshold on the models output probabilities was defined as the point which maximizes the validation set F_1. The AFB and the absolute AFB estimation error (E_{AF} (%)) were defined as in Chocron et al. [7]. The AFB is defined as follows:

$$AFB = \frac{\sum_{i=1}^{N} l_i \times \mathbb{I}_i}{\sum_{i=1}^{N} l_i}$$

with N available windows, l_i length of the i^{th} window and $\mathbb{I}_i$ the unity operator equal to 1 for AF and zero otherwise. In the raw-ECG case, the length of each window is 30 seconds which corresponds to 6000 samples at a sampling frequency of 200Hz. The E_{AF} (%) is defined as:

$$E_{\text{AF}}(\%) = \frac{\sum_{i=1}^{N} l_i \times (\hat{y}_i - y_i)}{\sum_{i=1}^{N} l_i}$$

where y_i is a binary value representing the window label and $\hat{y}_i$ is the predicted binary label by the model when in for both 1 for AF and 0 otherwise. ArNet-ECG classifies 30 seconds windows whereas ArNet2 takes as input 60-beat windows. In order to benchmark both algorithms, their outputs were aligned to provide a classification (AF, non-AF) for each 5-sec window. The non-parametric Mann–Whitney rank test was used to determine whether there is a statistical significance if the F_1 between ArNet-ECG and ArNet2. For that purpose the F_1 was 1000 times computed on randomly sampled 80% of the test set. The procedure was used to obtain the distribution of the F_1 and the Mann–Whitney rank test was applied on these F_1 distributions.

3. Results

3.1. Model performances

Best performance on the validation set were achieved after training the model for two epochs. For single window classification ArNet-ECG reached an F_1 of 0.96 [95% confidence interval (CI): 0.961–0.962] and an AUROC of 0.99. The median and the interquartile range (Q1-Q3) of the absolute AFB estimation error was $|E_{AF}(\%)|$

0.41%(0.03%-2.45%) on the test set. The $|E_{AF}|(\%)$ distributions for the four AF severity levels is shown on Figure 1. Table 1 summarizes the results of ArNet2 and ArNet-ECG. ArNet-ECG F_1 performance was significantly higher than ArNet2 (P-Value < 0.001).

Model	F_1[95% CI]	AUROC	Se	Sp
ArNet2	0.94[0.944-0.945]	0.99	0.93	0.96
ArNet-ECG	0.96[0.961-0.962]	0.99	0.96	0.96

Table 1. Results for windows classification on the UVAF test set (n=100).

3.2. Error Analysis

We observe that ArNet-ECG was able to detect 98.77% (3060/3098) of the AFL windows. In the test set there were four patients with AFL. We compared results on all windows from these four patients as shown in Table 2. Thus ArNet-ECG was able to detect AFL windows significantly better than ArNet2 which often miss-classifies a large number of AFL windows. This is because the beat-to-beat is relatively regular in AFL versus AF. For that reason, it may be challenging to identify AFL events from the RR time series and rather the information on the presence of AFL will mostly be contained in the ECG waveform. In Figure 1 the median absolute AFB estimation error $|E_{AF}|(\%)$ for the test set per different AF severity labels of ArNet2 and ArNet-ECG is shown. For both ArNet2 and ArNet-ECG there were only errors higher than 45% in the AF_{sev} group. The patient denoted as 1 in Figure 1 got an $|E_{AF}|(\%)$ of 57.8% by ArNet-ECG but was accurately classified by ArNet2. This recording contains both AF and ventricular extrasystoles which might have caused the misclassification by ArNet-ECG. Figure 2 (A) shows an ECG segment misclassified by ArNet-ECG but correctly classified by ArNet2. In Figure 1 the patient denoted as 2 obtained an $|E_{AF}|(\%)$ of 0.1% by ArNet-ECG but $|E_{AF}|(\%)$ of 90.0% by ArNet2. This mistake is probably due to the fact that this patient has a low average heart rate at 63 bpm and high beat-to-beat variability which is different than the typical AF cases where both a high variability and high average heart rate are observed. In Figure 1 the patient noted as 3 obtained an $|E_{AF}|(\%)$ of 0.1% by ArNet-ECG but $|E_{AF}|(\%)$ of 47.3% by ArNet2. This patient had reference annotations of 100% AFL. Shown in Figure 2 (B) is an ECG segment classified correctly by ArNet-ECG but wrongly classified by ArNet2. The beat-to-beat variation in this example is rather regular because it is under an AFL event and consequently ArNet2 missed the abnormality.

4. Discussion and Conclusions

We introduce ArNet-ECG, an algorithm for the detection of AF events from the raw ECG. ArNet-ECG performance reached an F_1 of 0.96 which outperform STOTA performance obtained with ArNet2. Most AFL examples were correctly classified by ArNet-ECG. This can be explained by the fact that raw ECG was used as input to the network and thus even in the case of regular beat-to-beat interval variation in AFL the network learns to recognize flutter waves (f-waves). Future work will include the integration of a recurrent neural network unit to take into account the time dependency between consecutive windows as well as assessing the generalization of ArNet-ECG on external test sets including different ethnic group.

Model	F_1	AUROC	Se	Sp
ArNet2	0.70	0.97	0.54	0.99
ArNet-ECG	0.98	0.99	0.97	0.99

Table 2. Results for windows classification from patients with AFL from the UVAF test set (n=4).

Acknowledgments

This research was supported for NBM, SB and JAB by a grant (3-17550) from the Ministry of Science & Technology, Israel & Ministry of Europe and Foreign Affairs (MEAE) and the Ministry of Higher Education, Research and Innovation (MESRI) of France. NBM, SB and JAB also acknowledge the support of the Technion-Rambam Initiative in Artificial Intelligence in Medicine.

References

[1] Björck S, Palaszewski B, Friberg L, Bergfeldt L. Atrial fibrillation, stroke risk, and warfarin therapy revisited: a population-based study. Stroke 2013;44(11):3103–3108.

[2] Haim M, Hoshen M, Reges O, Rabi Y, Balicer R, Leibowitz M. Prospective national study of the prevalence, incidence, management and outcome of a large contemporary cohort of patients with incident non-valvular atrial fibrillation. Journal of the American Heart Association 2015; 4(1):e001486.

[3] Quer G, Freedman B, Steinhubl SR. Screening for atrial fibrillation: predicted sensitivity of short, intermittent electrocardiogram recordings in an asymptomatic at-risk population. EP Europace 2020;22(12):1781–1787.

[4] Carrara M, Carozzi L, Moss TJ, De Pasquale M, Cerutti S, Ferrario M, Lake DE, Moorman JR. Heart rate dynamics distinguish among atrial fibrillation, normal sinus rhythm and sinus rhythm with frequent ectopy. Physiological Measurement 2015;36(9):1873.

[5] Moss TJ, Lake DE, Moorman JR. Local dynamics of heart rate: detection and prognostic implications. Physiological Measurement 2014;35(10):1929.

[6] Behar J, Oster J, Li Q, Clifford GD. Ecg signal quality during arrhythmia and its application to false alarm reduc-

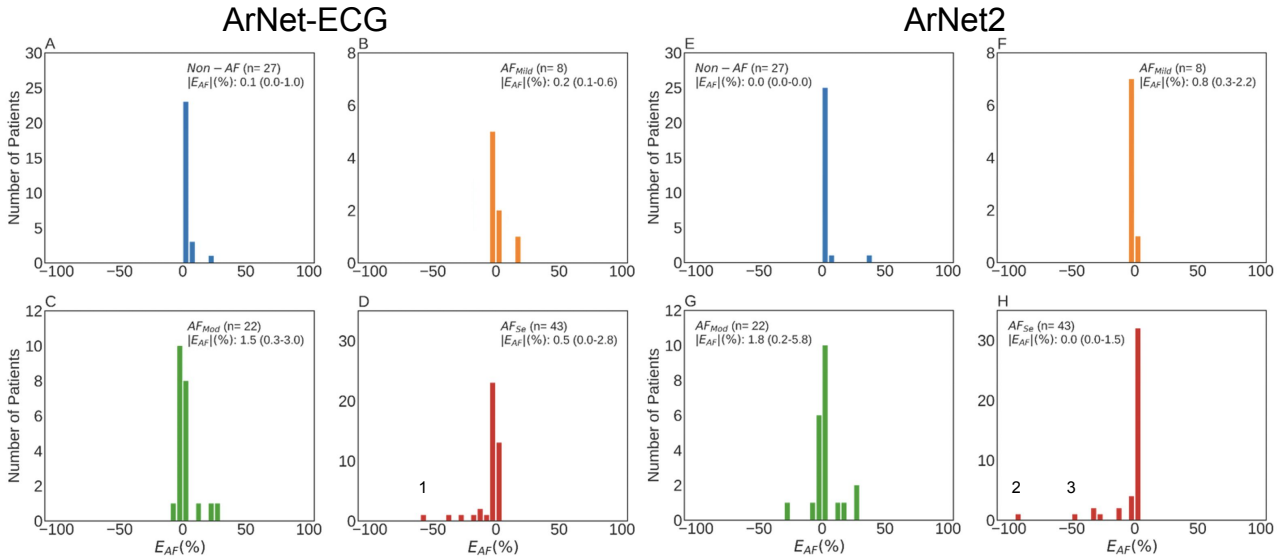

Figure 1. Histogram of the median absolute AFB estimation error $|\mathrm{E_{AF}}|(\%)$ for the test set per different $\mathrm{AF_I}$ severity labels: Non-AF (panel A and E), AF_{mil} (Panel B and F), AF_{mod} (panel C and G) and AF_{sev} (panel D and H). On the left results of ArNet-ECG this work) and on the right results of ArNet2 [8].

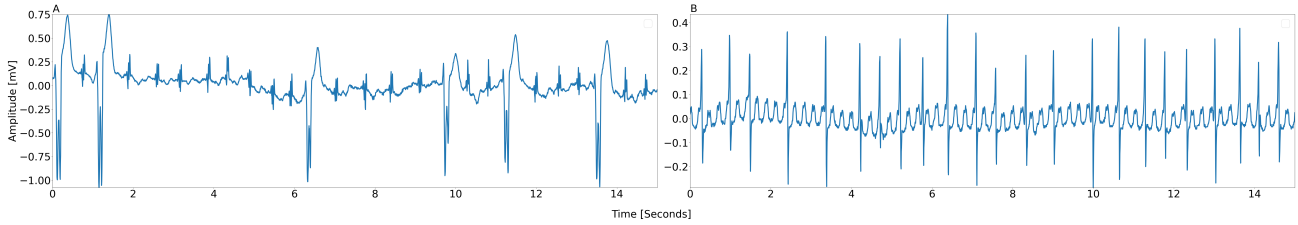

Figure 2. Error analysis examples. Panel A: example of an ECG window with a reference label of AF as well as containing premature ventricular contraction beats. This window was incorrectly classified by ArNet-ECG but correctly classified by ArNet2. Panel B: example of an ECG window with a reference label of AFL. This section was classified correctly by ArNet-ECG but mis-classified by ArNet2.

tion. IEEE Transactions on Biomedical Engineering 2013; 60(6):1660–1666.

[7] Chocron A, Oster J, Biton S, Mendel F, Elbaz M, Zeevi Y, Behar J. Remote atrial fibrillation burden estimation using deep recurrent neural network. IEEE Transactions on Biomedical Engineering 2020;.

[8] Biton S, Aldhafeeri M, Marcusohn E, Tsutsui K, Szwagier T, Elias A, Oster J, Sellal JM, Suleiman M, Behar JA. Generalizable and robust deep learning algorithm for atrial fibrillation diagnosis across ethnicities, ages and sexes. ArXiv Preprint arXiv220709667 2022;.

[9] Kirchhof P, Benussi S, Kotecha D, Ahlsson A, Atar D, Casadei B, Castella M, Diener HC, Heidbuchel H, Hendriks J, et al. 2016 esc guidelines for the management of atrial fibrillation developed in collaboration with eacts. Kardiologia Polska Polish Heart Journal 2016;74(12):1359–1469.

[10] Boriani G, Glotzer TV, Santini M, West TM, De Melis M, Sepsi M, Gasparini M, Lewalter T, Camm JA, Singer DE. Device-detected atrial fibrillation and risk for stroke: an analysis of¿ 10 000 patients from the sos af project (stroke prevention strategies based on atrial fibrillation information from implanted devices). European Heart Journal 2014; 35(8):508–516.

[11] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2016; 770–778.

[12] Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv14126980 2014;.

[13] Xiong Z, Stiles MK, Zhao J. Robust ecg signal classification for detection of atrial fibrillation using a novel neural network. In 2017 Computing in Cardiology. 2017; 1–4.

Address for correspondence:

Dr. Joachim Behar
jbehar@technion.ac.il
Faculty of Biomedical Engineering
Technion - Israel Institute of Technology
Haifa 3200003, Israel