Forecast of paroxysmal atrial fibrillation using a deep neural network

Cédric GILON
IRIDIA
Université Libre de Bruxelles
Brussels, Belgium
cedric.gilon@ulb.be

Jean-Marie GRÉGOIRE
IRIDIA
Université Libre de Bruxelles
Brussels, Belgium
jean-marie.gregoire@ulb.ac.be

Hugues BERSINI
IRIDIA
Université Libre de Bruxelles
Brussels, Belgium
hugues.bersini@ulb.ac.be

Abstract—**Atrial fibrillation (AF) is the most common heart arrhythmia. It affects between 1% and 2% of the world population over 35 years old. This disease is linked to an increased risk of stroke and heart failure. AF is a progressive disease and, at first, paroxysmal AF episodes occur, last from seconds up to a week and then stop. The disease evolves to permanent state, where the heart is always in fibrillation and can't be corrected. Forecasting paroxysmal AF episode a few seconds or minutes before its onset remains a hard challenge, but could lead to new treatment methods. For this study, we constructed a new long-term electrocardiogram (ECG) database (24 to 96 hours), composed of 10484 ECG. As a result of a careful analysis by a cardiologist, 250 AF onsets of paroxysmal AF have been detected in 140 ECG. We developed a deep neural network (DNN) model, composed of convolutional neural network (CNN) layers and bidirectional gated recurrent units (GRU) as recurrent neural network (RNN) layers. The model was trained for a supervised binary classification distinguishing between heartbeats series (RR intervals) that precede an AF onset and series distant from any AF. The model achieved an average area under the receiver operating characteristic (ROC) curve of 0.74. We evaluated the impact of heartbeat window size given as input, and the time period between the heartbeats window and the AF onset. We found that an input window of 300 heartbeats gives the best results and, not surprisingly, the closer the window is from the AF onset, the better the results. We concluded that RR intervals series contains information about the incoming AF episode, and that it can be exploited to forecast such episode.**

Index Terms—**atrial fibrillation, heart rate variability, RR intervals, deep learning, deep neural network, convolutional neural network, recurrent neural network**

I. Introduction

Atrial fibrillation (AF) is the most common heart arrhythmia and the second most common heart disease after hypertension. This disease is characterised by irregular contractions of the atria, the two upper chambers of the heart. Between 30 and 50 million people are affected worldwide. The prevalence is estimated between 1% and 2% for the population over 35 years old, and AF is more frequent for people aged over 80 years old, for whom it is estimated between 10% and 17% [1] [2]. The number of affected patients continues to rise due to the global ageing of the world population. Indeed, the part of the population aged over 60 years old is expected to double by 2050 and is growing faster than all younger age groups [3].

AF is a progressive disease, and three types can be distinguished: paroxysmal AF, persistent AF and permanent AF. The disease will first be present in paroxysmal state, where AF episodes start randomly and last from seconds up to a week. The heart recovers to normal sinus rhythm (NSR) without the need for a medical intervention. The disease then evolves to persistent state, where episodes last more than 7 days. A medical intervention, either drugs or surgery, is required to help the heart to recover to a normal state. Finally, the last state of the disease is permanent AF. When the disease reaches this state, the heart is continuously in AF, and never recovers to AF.

One major danger of this disease is that it can be present but asymptomatic for years, before being revealed by one of its consequences. AF can lead to stroke, as blood clots can form in the heart and then be expelled into the body during AF episodes [4]. Patients with AF have an increased risk of stroke by a factor of 5. Other AF consequences are an increased risk of heart failure and death.

AF is diagnosed by a cardiologist using an electrocardiogram (ECG), in which signs of AF can be detected. In addition to ECG analysis, heart rate variability (HRV) was also studied for AF detection, as heartbeats become irregular during AF (Figure 1). HRV can be derived from the ECG and it corresponds to the series of time duration between heartbeats. This time duration is called RR interval and measured between two successive R waves, i.e. the most significant part of one ECG wave.

For AF detection purpose, cardiologists can record ECG on an opportunistic basis (e.g. during regular cardiac check), or a systematic ECG records can be made for a given population (e.g. for all patients over 70 years old). In clinical practice, if there exists a suspicion of AF for a patient, a long-term ECG is recorded using an Holter monitor. The patient will carry the portable electrocardiograph for several days, and the whole ECG will be analysed afterward by a medical software and a cardiologist to detect signs of AF. In addition, cardiologists can also rely on risk scores (e.g. CHA_2DS_2VASc [5] or CHARGE-AF [6]) to identify at-risk patients in the population. These scores are based on the combination of multiple clinical parameters of the patient (e.g. age, sex, diabetes, hypertension).

In recent years, machine learning (ML) algorithms and in particular deep learning (DL) algorithms have been used

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

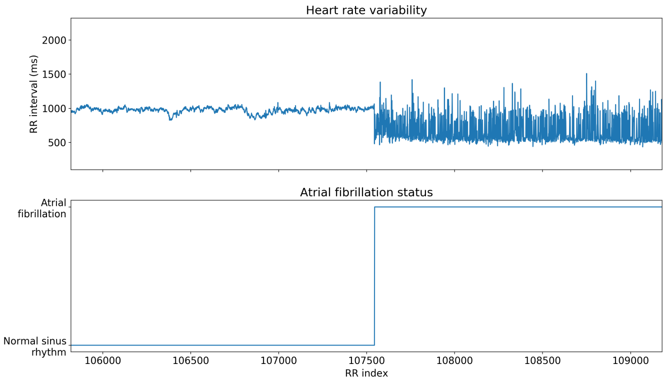

Figure 1. Heart rate transition from regular rhythm (normal sinus rhythm) to irregular rhythm (atrial fibrillation episode)

for automated detection of AF. These models have produced promising results, sometimes achieving a better accuracy than an average detection rate of cardiologists [7]. In addition, due to the increasing available computation power, the models are able to analyse several days of ECG in seconds, where it would take minutes or hours of diligent work for a cardiologist to perform similarly. In the coming years, artificial intelligence (AI) models could assist medical professionals in such repetitive tasks, to create a first analysis and detection of AF episodes that would then be confirmed by the cardiologist.

Studies for AF detection have been based on various ML and DNN models, composed of convolutional neural network (CNN) and/or recurrent neural network (RNN). Some studies are focusing on high-quality ECG (e.g. 200 Hz) [7], while other are focusing on HRV series [8] [9]. The datasets used are either public or private. The MIT-BIH Atrial Fibrillation Database [10] is one of the most used public datasets. It contains 25 long-term ECG with the corresponding annotations. It is available on Physionet [11] website.

Multiple companies are also offering AF detection features embedded in wearables, such as smartwatches [12] or portable ECG devices [13]. The device records ECG regularly or on user demands and analyses them to search for any sign of AF. The user is notified in case of abnormal heart rhythm and is invited to visit a cardiologist for further medical tests to refute or confirm the presence of AF. This type of use of new technology opens a new era for medical research where large-scale studies could be conducted using data collection from wearables. This could also lead to personalised medicine where models are trained for the needs of a specific patient. Special attention must be given to the model accuracy and false positive rate. Indeed, inducing extra workload for cardiologist instead of reducing it must be avoided.

Additionally to AF detection, other AF related tasks have been realised in the last years. Researchers have been looking for an AF signature in normal ECG signal as an alternative to AF risk score, to identify patients with a high likelihood to develop AF in the future. A deep CNN model was able to distinguish normal ECG from patients with AF from normal ECG from healthy patients up to a month before an AF crisis [14]. The area under the receiver operating characteristic (ROC) curve (AUC) for this classification equals 0.87. As a comparison, the CHARGE-AF risk score achieved an AUC of 0.75 [15].

Other researchers have been studying the possibility to forecast an oncoming AF using the seconds and minutes of ECG previous to the onset. The *PAF prediction challenge* [16] proposed by Physionet in 2001 made public a dataset for this task. It is composed of 100 pairs of 30 minutes ECG, half of the pairs are from healthy patients and the other half from patients with AF. The first goal was to make the distinction between pairs from healthy patients and patients with AF. The second goal is only for pairs from AF patients, and it was to determine which of the two ECG precedes the AF and which of the two is distant. For the second task, the winner of the challenge correctly identified the ECG preceding the AF in 22 of the 28 pairs of the test set, using an algorithm based on the premature atrial complex count (APC) [17]. More recently, a study proposed a feed-forward neural network model able to classify features extracted from the ECG. It was able to classify 55 of the 56 ECG from the 28 AF patients pairs [18]. A limitation of these results is the small training and testing dataset size. It makes it difficult to evaluate if the proposed models would generalise correctly when applied to a larger patients sample.

In summary, three AF related tasks can be distinguished in the literature:

- AF detection,
- patient identification,
- AF forecast.

The three tasks require the analysis of ECG or HRV, but the window of interest is different, as shown in Figure 2. For AF detection, this window in composed of all ECG records, with or without AF episodes. For patient identification, the windows of interest are the weeks before the first signs of AF episodes and the objective is to predict if an AF episode will occur or not in the following weeks or months. Finally, the windows of interest for AF forecast are the minutes and hours just before the AF onset.

In this research, we study the AF forecast but with a different goal than the one proposed for the *PAF prediction challenge*. Our model does not distinguish pairs of ECG but aims to predict if an individual ECG is before the onset or far from any AF.

In addition, we have chosen to limit the treated signal to the HRV time series just composed of RR intervals.

II. Materials and method

A. Dataset

A new database of long-term Holter ECG has been constituted. They were recorded from December 2009 to December 2018 in an outpatient clinic. All patients were included in the dataset, at the except of those with Cardiovascular Implantable Electronics Devices (CIED). It represents a total of 10484

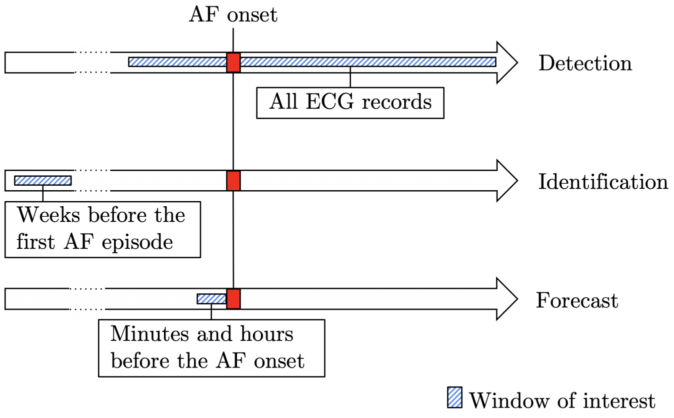

Figure 2. Windows of interest for AF tasks

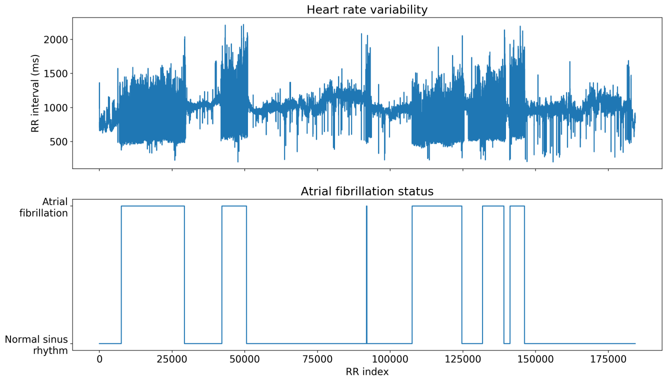

Figure 3. Complete Holter record with the RR intervals (top) and corresponding labels established by the cardiologist (bottom)

ECG. They were recorded using 2-channel Spiderview (200 Hz) Holter monitors. Each record lasts for at least 24 hours and up to 120 hours.

All the ECG were first analysed by a medical software and then by a cardiologist, to determine if the record presents signs of AF. If signs were found, the whole ECG was analysed to accurately establish, within 5 milliseconds, the starting and ending time of AF episodes. For each ECG with AF signs, a precise label file containing a value (AF or NSR) for each heartbeat was created (Figure 3). A medical software was used to extract the RR intervals and eliminate outliers from the records.

Only the RR intervals windows preceding the AF onset were used to construct the dataset, and not the whole Holter monitoring. Three parameters were used to extract relevant windows from the Holter database (Figure 4).

- The first parameter is the *window size*. It corresponds to the number of RR intervals in the window that will be given as input to the DNN model.
- The second parameter is the *distance*. It corresponds to the number of RR intervals between the window and the AF onset.

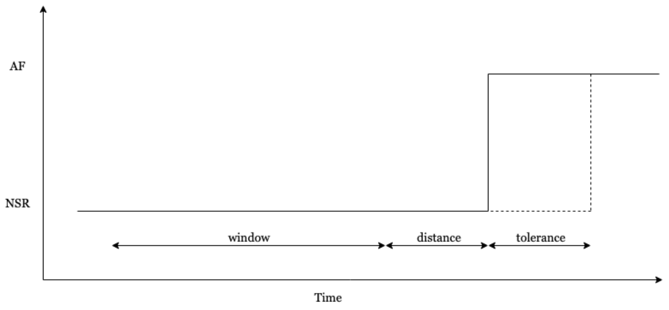

Figure 4. The three parameters used for RR intervals windows choice during dataset creation from Holter monitorings

TABLE I
MODEL LAYERS AND CORRESPONDING PARAMETERS. TOTAL NUMBER OF WEIGHTS IN THE NETWORK IS 151 901.

N°	Type	Parameters	Output shape
1	Input layer		300 x 1
2	1D convolution	filers: 100 kernel size: 3 stride: 1	298 x 100
3	1D convolution	filers: 100 kernel size: 3 stride: 1	296 x 100
4	Max pooling	pooling size: 2 stride: 2	148 x 100
5	Bidirectional GRU	units: 200	200
6	Dense	units: 1 activation: softmax	1

- Finally, the third parameter is the *tolerance*. Rather than attempting to predict if a given RR intervals window will or not lead directly to an AF onset, the tolerance is the number of RR intervals in which the AF crisis onset can take place.

The tolerance was fixed to 30 RR intervals. Therefore, instead of forecasting if a given window leads directly or not to an AF, the model forecasts if an AF will start or not in the next 30 beats after the window. For each AF onset, 30 windows can therefore be considered.

The impact of the variation of the distance and window size parameters on the model performance were studied. Three window sizes were chosen:

- a short window (60 RR intervals before AF),
- a medium window (300 RR intervals before AF),
- a large window (900 RR intervals before AF).

The distance was first set to 0 RR interval and then increased to 30, 60, 90 and up to 300 RR intervals before the AF onset.

For each case of the three chosen parameters, a dataset is created and balanced with negative examples, i.e. RR intervals windows distant from any AF signs and chosen randomly in the Holter monitoring. The corresponding task is therefore a supervised binary classification problem, where for a given RR intervals window, the model should be able to forecast whether or not an AF onset is likely to occur after the window.

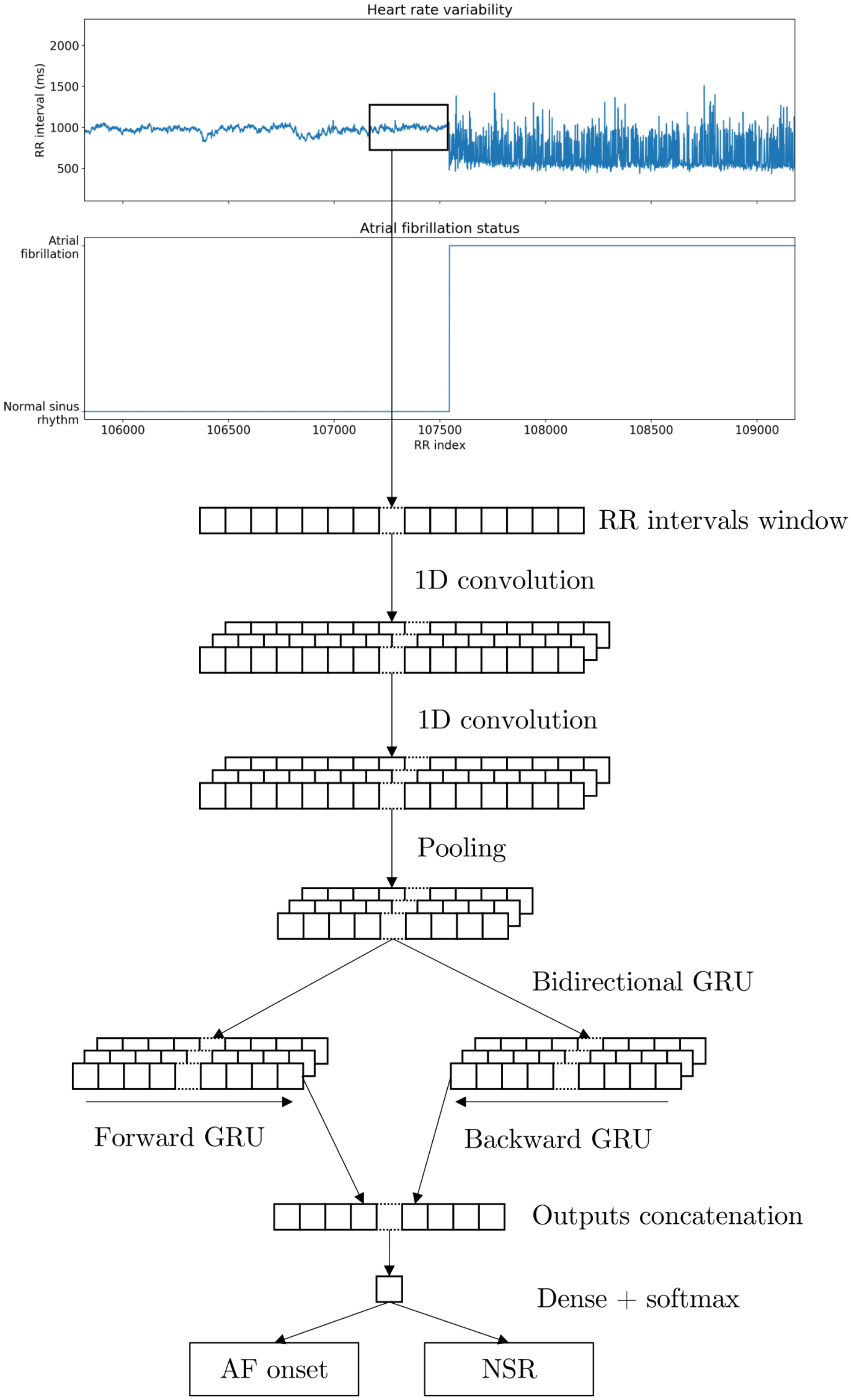

Figure 5. DNN model architecture. The RR intervals window is extracted from the ECG and is then analysed by a CNN composed of two 1 dimension convolution layers and a max pooling layer. A bidirectional GRU treats the output of the CNN and the final classification is done by a dense layer with a softmax activation.

B. Model architecture

The DNN model consists of several layers, as presented in Figure 5 and Table I. The input is a RR intervals window. The first two layers of the DNN are 1 dimension CNN layers, with a kernel of size 3 and a stride of 1. The number of filters is 100 and therefore 100 output features maps are produced. A max pooling layer with a pooling size of 2 is then used to reduce the dimension of the output before the next layer. A bidirectional gated recurrent unit (GRU) [19] layer is used to process the output. The bidirectional layer provides a rich context for the network. Indeed, at each time step in the series, the forward GRU contains all the past information while the backward one contains all future information from the series. GRU were preferred over long short-term memory (LSTM) [20] as they provided similar results and faster training due to the reduced number of parameters. The forward and backward results are concatenated and, finally, a dense layer is used for the final classification. The binary classification is either AF onset or NSR, depending on whether the model predicts that the RR intervals window precedes an AF onset or is very distant from any AF.

C. Model training

For each set of parameters (distance, tolerance and window size), a new dataset is constructed from the ECG database, and a new model is trained on this dataset. The process is repeated 200 times and the final metrics are averaged over the metrics from all runs.

The dataset is split into train, validation and test sub-datasets using a 7:1:2 ratio. It is important to note that the split is done at a patient level, i.e. if a patient Holter record contains multiple AF onsets, they can only be contained in the same split. Indeed, we wanted to test patients on completely new records to ensure the model generalisation capacity.

During the training phase, an early stopping mechanism was used to avoid overfitting. At the end of each epoch, the validation set is used to determine whether the model still improves and if the training should be continued or not. If the metrics did not improve for a certain number of epochs, the training is stopped and the best model weights are restored. This number of epochs is determined by the patience parameter, set to 25 epochs in our study.

The model was trained using the Adam optimizer [21], with a learning rate of 0.0001. The batch size was set to 128 samples. The model was implemented using Python 3.6 , Keras and Tensorflow. It was trained using a NVIDIA GeForce RTX 2080ti.

III. Results

A. Database composition

From the 10484 Holter monitorings, 550 show signs of AF. From those, 140 records show signs of paroxysmal AF, 391 records show signs of permanent or persistent AF and 19 records were rejected because of insufficient data quality or active pacemaker during the record (Figure 6). Most records presented only one (81 patients) or two (28 patients) AF episodes, but some contained up to 11 episodes (Figure 7).

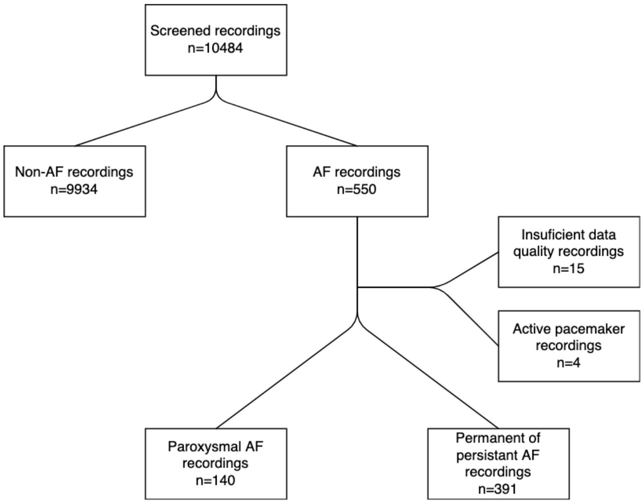

Figure 6. Composition of the Holter database (10484 recorded Holter monitorings). 140 show signs of paroxysmal AF and were used for the model training and testing.

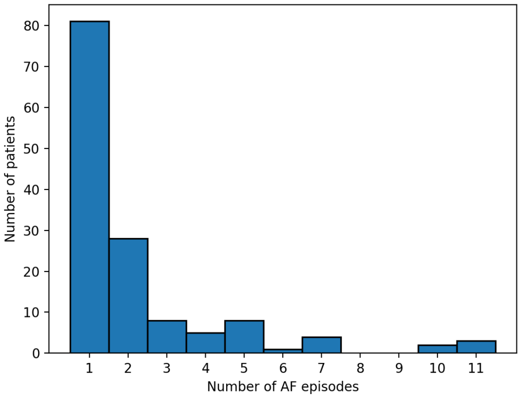

Figure 7. Distribution of the number of AF episodes in Holter recordings

In total, 308 AF episodes and 287 AF onset were recorded. The fact that the number of onsets is smaller than the number of episodes is due to the fact that some episodes had already begun before the monitoring was started. From theses 287 onsets, only 250 do not contain any sign of AF in the 300 RR intervals preceding the AF crisis and could be used for the study.

B. Window size parameter

The impact of the window size was studied. The three chosen window sizes (60 RR intervals, 300 RR intervals and 900 RR intervals) were tested on datasets with the distance parameter set to 0 RR intervals and the tolerance parameter set to 30 RR intervals. For each window size, the ROC curves for the test set was created and AUC we computed. Table II presents the results of the three choices and the 300 RR

TABLE II
METRICS FOR AF FORECAST WITH VARYING WINDOW SIZE (SMALL, MEDIUM AND LARGE WINDOWS)

Window size	Distance	AUC	95% CI
60	0	0.7152	0.7003 - 0.7300
300	0	0.7427	0.7279 - 0.7572
900	0	0.7123	0.6960 - 0.7284

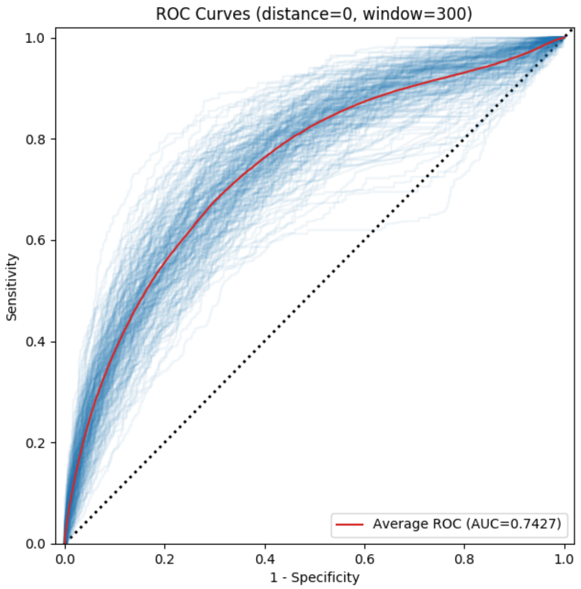

Figure 8. Average ROC on 200 runs with distance parameters set to 0 and window size parameter set to 300. Each blue line corresponds to one ROC for an independent run.

intervals window size gave the best performance. Therefore, this window size was selected for the tests of the distance parameter.

C. Distance parameter

The distance parameter was studied and we found, not so surprisingly, that the closer the RR intervals window is from the AF onset, the better the results become. The AUC equals 0.74 when the windows are next to the AF onset, i.e. distance of 0 (Figure 8), so just at the frontier. The AUC decreases as the window moves further away from the AF onset. As shown in Table III, the AUC is 0.62 at a distance of 300 RR intervals before the AF onset. The ROC evolution is presented in Figure 9. So even at such a time distance, there is still some weak signal of AF next appearance, and most of our future efforts will be to improve this prediction accuracy.

IV. DISCUSSION

In this study, we implemented, trained and evaluated a DNN composed of a 1 dimension CNN layers and a bidirectional GRU. We showed that the model was able to distinguish RR intervals windows preceding AF onset and RR intervals windows distant from any AF onset.

TABLE III
METRICS FOR AF FORECAST WITH VARYING DISTANCES

Window size	Distance	AUC	95% CI
300	0	0.7427	0.7279 - 0.7572
300	30	0.7115	0.6958 - 0.7269
300	60	0.7023	0.6866 - 0.7179
300	90	0.6818	0.6655 - 0.6979
300	300	0.6271	0.6099 - 0.6443

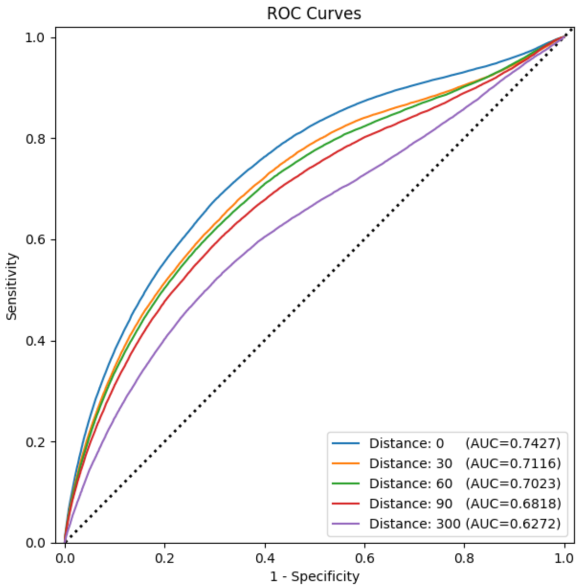

Figure 9. Evolution of the ROC with respect to the distance

We have built a new ECG database, which is larger than the one available on Physionet (250 AF onsets vs 28 AF onsets). During the ECG windows choice process, the tolerance parameter allowed us to considerate more of them for each AF onset and therefore to increase the dataset size. Indeed, for a tolerance set to 30 RR intervals, 30 windows can be considered for each AF onset. In total, each dataset with positive and negative examples is composed of about 15000 samples. However, it remains small for usual DNN model training, where a large number of parameters have to be optimised, based on a much bigger training set.

A second limitation is the reduced information, due to the use of RR intervals series in place of high quality ECG. Despite this, the model performance and its ability to derive features to distinguish between the two types of windows were already quite impressive. In the future, we aim to study the same dataset but with higher quality signal. The medical parameters of the patients, used to compute risk scores, could also be given as an input to the model to consider more insights about the patients.

Finally, it should be noted that the use of DNN for medical applications raises questions about results interpretation and explainability, a quite critical constraint in medicine. The size of the networks and the number of parameters makes it very difficult to interpret. Looking for simpler and more understandable ML classifiers, using well-known time and frequency features, could also be a way forward for this research.

V. Conclusion

Being able to forecast AF episodes a few seconds, minutes or hours before the onset is challenging, but could be a very important path to new treatments. Currently, patients with AF are put under anticoagulant treatment to reduce the risk of blood cloths formation, and therefore reduce the risk of potential stroke.

Implementing a ML model trained to recognise AF signatures in normal ECG before AF onsets could help to prevent the crisis before it happens, and reduce the risks and possible consequences induced by this AF episode. One key aspect to train this ML model is the number of records in the dataset. This dataset construction is still in progress, as we continue to increase the number of ECG. In addition, recordings from healthy patients and patients with other cardiac conditions should be included in the database. Randomly selected windows from those Holter monitorings should be included in the samples as negative examples to increase the model generalisation.

Nowadays, as we enter the new era of personalised medicine, the long-term view would be to enable smart wearables and pacemakers to be trained for a specific patient. The device could therefore be able to forecast the next AF onset and apply a preventive treatment to avoid the crisis and protect the patient in the best way.

References

[1] Chugh Sumeet S. *et al.*, “Worldwide Epidemiology of Atrial Fibrillation,” *Circulation*, vol. 129, no. 8, pp. 837–847, Feb. 2014.

[2] G. H. Mairesse *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),” *Europace: European Pacing, Arrhythmias, and Cardiac Electrophysiology: Journal of the Working Groups on Cardiac Pacing, Arrhythmias, and Cardiac Cellular Electrophysiology of the European Society of Cardiology*, vol. 19, no. 10, pp. 1589–1623, 2017.

[3] United Nations, “World Population Ageing,” *United Nation Reports - Economic and Social affairs*, 2015.

[4] B. Freedman *et al.*, “Screening for Atrial Fibrillation: A Report of the AF-SCREEN International Collaboration,” *Circulation*, vol. 135, no. 19, pp. 1851–1867, 2017.

[5] G. Y. H. Lip, R. Nieuwlaat, R. Pisters, D. A. Lane, and H. J. G. M. Crijns, “Refining Clinical Risk Stratification for Predicting Stroke and Thromboembolism in Atrial Fibrillation Using a Novel Risk Factor-Based Approach: The Euro Heart Survey on Atrial Fibrillation,” *Chest*, vol. 137, no. 2, pp. 263–272, Feb. 2010.

[6] Alonso Alvaro *et al.*, “Simple Risk Model Predicts Incidence of Atrial Fibrillation in a Racially and Geographically Diverse Population: the CHARGE-AF Consortium,” *Journal of the American Heart Association*, vol. 2, no. 2, 2013.

[7] A. Y. Hannun *et al.*, “Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network,” *Nature Medicine*, vol. 25, no. 1, pp. 65–69, Jan. 2019.

[8] R. S. Andersen, A. Peimankar, and S. Puthusserypady, “A deep learning approach for real-time detection of atrial fibrillation,” *Expert Systems with Applications*, vol. 115, pp. 465–473, Jan. 2019.

[9] O. Faust, A. Shenfield, M. Kareem, T. R. San, H. Fujita, and U. R. Acharya, “Automated detection of atrial fibrillation using long short-term memory network with RR interval signals,” *Computers in Biology and Medicine*, vol. 102, pp. 327–335, Nov. 2018.

[10] M. GB and M. RG., “A new method for detecting atrial fibrillation using R-R intervals,” *Computers in Cardiology*, vol. 10, pp. 227–230, 1983.

[11] Goldberger Ary L. *et al.*, “PhysioBank, PhysioToolkit, and PhysioNet,” *Circulation*, vol. 101, no. 23, pp. e215–e220, Jun. 2000.

[12] M. V. Perez *et al.*, “Large-Scale Assessment of a Smartwatch to Identify Atrial Fibrillation,” *New England Journal of Medicine*, vol. 381, no. 20, pp. 1909–1917, Nov. 2019.

[13] J. K. Lau *et al.*, “iPhone ECG application for community screening to detect silent atrial fibrillation: A novel technology to prevent stroke,” *International Journal of Cardiology*, vol. 165, no. 1, pp. 193–194, Apr. 2013.

[14] Z. I. Attia *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,” *The Lancet*, vol. 394, no. 10201, pp. 861–867, Sep. 2019.

[15] I. E. Christophersen *et al.*, “A comparison of the CHARGE–AF and the CHA2ds2-VASc risk scores for prediction of atrial fibrillation in the Framingham Heart Study,” *American heart journal*, vol. 178, pp. 45–54, Aug. 2016.

[16] G. Moody, A. Goldberger, S. McClennen, and S. Swiryn, “Predicting the onset of paroxysmal atrial fibrillation: the Computers in Cardiology Challenge 2001,” vol. 28, Feb. 2001, pp. 113–116.

[17] W. Zong, R. Mukkamala, and R. Mark, “A methodology for predicting paroxysmal atrial fibrillation based on ECG arrhythmia feature analysis,” in *Computers in Cardiology 2001. Vol.28 (Cat. No.01CH37287)*, Sep. 2001, pp. 125–128, iSSN: 0276-6547.

[18] E. Ebrahimzadeh, M. Kalantari, M. Joulani, R. S. Shahraki, F. Fayaz, and F. Ahmadi, “Prediction of paroxysmal Atrial Fibrillation: A machine learning based approach using combined feature vector and mixture of expert classification on HRV signal,” *Computer Methods and Programs in Biomedicine*, vol. 165, pp. 53–67, Oct. 2018.

[19] K. Cho *et al.*, “Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation,” *arXiv:1406.1078 [cs, stat]*, Sep. 2014, arXiv: 1406.1078.

[20] S. Hochreiter and J. Schmidhuber, “Long Short-term Memory,” *Neural computation*, vol. 9, pp. 1735–80, Dec. 1997.

[21] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” *arXiv:1412.6980 [cs]*, Jan. 2017, arXiv: 1412.6980 version: 8. [Online]. Available: http://arxiv.org/abs/1412.6980