Reproducibility of machine learning models for paroxysmal atrial fibrillation onset prediction

Cédric Gilon[1], Jean-Marie Grégoire[1,2], Jérome Hellinckx[1], Stéphane Carlier[2], Hugues Bersini[1]

[1] IRIDIA, Université Libre de Bruxelles, Belgium
[2] Département de Cardiologie, Université de Mons, Belgium

Abstract

Atrial fibrillation (AF) is the most common heart arrhythmia. Paroxysmal AF onset prediction is a more complex task than screening AF. Published methods using the AFPDB database show excellent results, suggesting that paroxysmal AF onset prediction is possible with machine learning (ML) models using heart rate variability (HRV) parameters.
Aims *To understand if AF onset prediction is possible using previously published methods. Reproduce results of published studies using the Physionet database.*
Methods *We searched the literature for all articles on paroxysmal AF onset prediction. We analysed in depth 3 methodology using ML methods to replicate their results.*
Results *With the information available in the publication, we were unable to reproduce the results presented by the authors with differences up to 20%. For each publication, we explored different scenarios with multiple splits and parameters choice for the model.*
Conclusion *Reproducibility of the models and results is becoming a key aspect of ML research and authors must describe and make available the whole methods required to achieve their results.*

1. Introduction

Atrial fibrillation (AF) is the most common heart arrhythmia. This disease is linked to an increased risk of stroke, heart failure and death. During paroxysmal AF, crisis starts and stops with no known warning sign. In 2001, Physionet launches the PAF Prediction Challenge [1] to understand if incoming signs of AF onset can be detected in the 30-minute window preceding the start of the crisis. This question has continued to be explored by multiple research teams. The results presented in the publications using database from the challenge are optimistic with accuracy values around 90% [2–4]. The reproducibility of published works is not always straightforward without access to the original code. Recently, machine learning (ML) based methods have shown great results in multiple fields including AF detection and AF onset prediction. Reviews have shown that ML publications contain errors or missing information in the methodology, data leakage from the train split to the test split making the results difficult or sometimes impossible to reproduce [5]. It is recommended to have the most independent data split to ensure a good generalisation of the prediction [6].

In this paper, we have analysed and reproduced the method of 3 publications to understand if the results are reproducible. The materials and methods are presented in Section 2, the results in Section 3, the discussion in Section 4 and the conclusion in Section 5.

2. Materials and method

2.1. PAF Prediction Challenge Database

The PAF Prediction Challenge Database (AFPDB) has become the standard dataset for AF onset prediction. This dataset is composed of 200 ECGs records from 100 patients, with 2 records per patient. The duration of each record is 30 minutes and the sampling frequency is 128 Hz. For AF patients, one record is preceding an AF and the other is distant from any AF sign (with at least 45 minutes before and 30 minutes after). For healthy patients, the two records are normal sinus rhythm (NSR). For the challenge, the dataset was composed of a train set (50 patients with 25 AF patients) and a test set (50 patients with 28 AF patients). In total, the dataset is composed of 53 records preceding AF and 147 records distant from any AF sign.

2.2. Models from previous work

We searched the literature for all AF onset prediction publications. We found 33 papers published between 2001 and 2022 and selected 3. Table 1 summarised the selected approaches.

Ref.	Date	Model	Window	Features
[2]	2012	SVM	30	bispectral frequential non-linear
[3]	2018	SVM	5	time-domain bispectral frequential non-linear
[4]	2018	KNN	5	time-domain frequential

Table 1: Selected models for AF onset prediction

Model SVM-30

A SVM classifier is proposed in [2]. They used the 30-minute window of the signal. After preprocessing and HRV extraction, they used frequential features, bispectral features and non-linear features (sample entropy and Poincaré plot-extracted features) from the HRV signal. They used the initial train-test split from the challenge, but they restricted their dataset to use only AF patients. In total, it represents 50 ECGs for training and 56 ECGs for testing. The best results were achieved using $C = 1000$ and $\gamma = 3.6$ for the SVM. The selected features are two frequential features (LF, HF), six bispectral features (e1, e2, h1, h2, h3, h4) and four non-linear features (sampen, sd1, sd, ratio sd1/sd2).

Model SVM-5

[3], they also proposed an SVM classifier using features from 5-minute windows. The HRV is extracted from ECGs. After correction features from time-domain, frequential features and bispectral features are used. Genetic algorithm (GA) is used to select features and the final set is composed of temporal features (NN50, pNN50), non-linear features (SampEn, SD2), frequential features (AR-LF) and bispectral features (LL-H1, ROI-WCOB). They used a 10-fold cross validation (CV) to validate their results using the 106 ECGs from the 53 AF patients.

Model KNN

In [4], they used a K-nearest neighbours (KNN) model. The 30-minute records are split into 5-minute windows with 50% overlapping windows. They used the train dataset, with the expect of the record *n27*. In total, it represents 74 NSR records and 25 AF records. To compute their classifier performance, they used 10-fold CV. Using the ECGs signal, they extracted HRV and they used temporal, frequential and non-linear features. To select features, they used a GA where every feature usage is encoded as one bit. They present multiple models with results for different dataset splits, feature selection and k value for KNN models. The best model is using $k = 3$ and 5 features (RMSSD, FFT_LF, FFT_VLF, FFT_Total).

Ref.	Accuracy	Sensitivity	Specificity
[2]	-	96.30%	93.10%
[3]	87.7%	86.8%	88.7%
[4]	90.0%	92.0%	88.0%

Table 2: Reported results for AF onset prediction

2.3. Reproduction

We reproduce the method presented in each paper. We create multiple scenarios by variation the dataset choice, the dataset split and the model parameters. Each scenario was run 1000 times and reported the average accuracy, sensitivity and specificity with 95% confidence interval (CI).

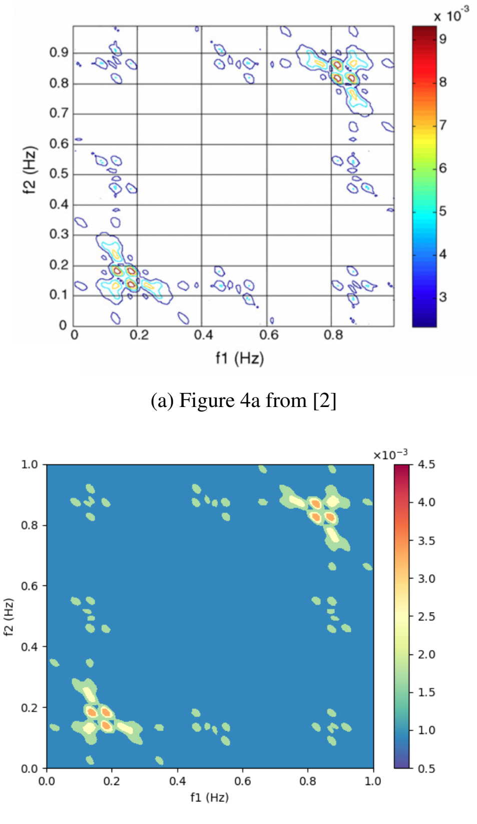
From [2], we used two SVM with 30-minute window. The first with C=1000 and γ=3.6 as presented in the paper and a second SVM with C chosen in $[0.1, 1, 10, 100, 1000, 10000]$ and γ chosen in $[10, 3.6, 1, 0.1, 0.01, 0.001, 0.0001]$. We used the initial train-test split. We also tested features if standardisation increase the results.

From [3], we used the SVM model with features extracted from 5-minute window. We used the features selected by the GA. We used variable C and γ as for model SVM-30. We tested two types of window choice: only the last five minutes of the records or all 5-minute windows available from the 30-minute window (with 50% overlap). We tested two datasets: the first with all AF patients and the second with the whole dataset. We used 10-fold CV at patient level.

From [4], we used a KNN model with features selected by the GA from 5-minute window. We used two types of cross validation, either at record level or at patient level. We tested three dataset splits: the train with only AF patients, the train with all patients and finally the whole dataset. We used two types of cross validation, either at record level or at patient level, i.e. the two records of one patient should be contained in the same split. We run 10-fold CV.

3. Results

The performances of the reproduced models were lower than those presented in the selected publications. They are presented in Table 2.



(a) Figure 4a from [2]

(b) Reproduction

Figure 1: Reproduction of the contour plot of biamplitude for 30-minute window (record p03 of AFPDB)

Model SVM-30

For the model of [2], we achieved an accuracy of 78.57%. The results are presented in Table 3 for the fixed parameters and in Table 4 for the variable parameters. We were able to reproduce the bispectral plot presented in the publication as shown in Figure 1 to validate our method. The model seems to overfit on one class when no standardisation is used. The results were better using variable C and γ.

Model SVM-5

For the model of [3], we achieved 74.45% of accuracy. The results with the various scenarios are presented in Table 5. Results on the whole dataset were better in accuracy but lower in sensitivity.

Model KNN

For the model of [4], we achieved 75.62% of accuracy for KNN with $k = 3$. The results are presented in Table 6. The distribution of the sensitivity and specificity of each CV is presented in Figure 2. The results using splits are record level are better but could be linked to the data leakage between the train and test set.

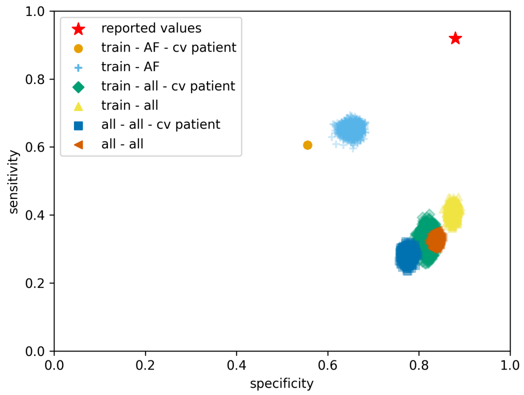


Figure 2: Sensitivity and specificity for KNN with k =3 in the different scenarios

4. Discussion

Papers giving too optimistic and non-reproducible results can be the source of a rejection of ML methods by clinicians [7]. We reviewed three publications about AF onset prediction. We reproduce the methods, but our results did not match those reported by the authors, with differences up to 20% in accuracy. In addition, some part of the methodology remains undefined, and we did not receive answers when we contacted the authors. Researchers are now proposing frameworks to help authors to include all the required materials and method information to better reproduce their work [6]. We think that a good step forward could be to open-source the code created by the author to re-train the exact model.

5. Conclusion

ML models need to be more detailed if the reported results must be reproducible. The use of larger databases is mandatory for this type of prediction, as splits on a small dataset can be the cause of results variation. Progress must be made before the clinical adoption and use of algorithms to predict paroxysmal AF onset.

6. Code and data

The code and data are available in this repository: https://github.com/cedricgilon/paf-challenge-reproducibility. The data and labels are also available on the Physionet website.

dataset train	dataset test	patient	# ECGs	norm.	accuracy(%)	sensitivity (%)	specificity (%)
train	test	AF	50	no	50.0 (50.0-50.0)	100.0 (100.0-100.0)	0.0 (0.0-0.0)
train	test	AF	50	yes	44.64 (44.64-44.64)	78.57 (78.57-78.57)	10.71 (10.71-10.71)
train	test	AF+NSR	108	no	72.0 (72.0-72.0)	0.0 (0.0-0.0)	100.0 (100.0-100.0)
train	test	AF+NSR	108	yes	70.0 (70.0-70.0)	0.0 (0.0-0.0)	97.22 (97.22-97.22)

Table 3: Results for SVM-30 model with $C = 1000$ and $\gamma = 3, 6$

dataset train	dataset test	patient	# ECGs	norm.	accuracy (%)	sensitivity (%)	specificity (%)
train	test	AF	50	no	50.0 (50.0-50.0)	100.0 (100.0-100.0)	0.0 (0.0-0.0)
train	test	AF	50	yes	47.92 (47.58-48.26)	59.28 (57.72-60.83)	36.56 (34.8-38.32)
train	test	AF+NSR	108	no	72.0 (72.0-72.0)	0.0 (0.0-0.0)	100.0 (100.0-100.0)
train	test	AF+NSR	108	yes	69.67 (69.46-69.87)	5.81 (5.35-6.27)	94.5 (94.06-94.95)

Table 4: Results for SVM-30 model with variable C and γ

dataset	windows	patient	# ECGs	CV	accuracy	sensitivity	specificity
train+test	last	AF	106	patients	53.17 (52.87-53.47)	56.89 (56.37-57.4)	50.91 (50.17-51.64)
train+test	last	AF+NSR	200	patients	72.33 (72.16-72.5)	10.06 (9.45-10.67)	94.79 (94.39-95.19)
train+test	all	AF	106	patients	62.84 (62.46-63.22)	61.33 (60.45-62.21)	64.39 (63.83-64.96)
train+test	all	AF+NSR	200	patients	74.45 (74.35-74.54)	17.95 (16.83-19.07)	94.82 (94.43-95.2)

Table 5: Results for SVM-5 model

dataset	patient	# ECGs	CV	accuracy (%)	sensitivity (%)	specificity (%)
train	AF	50	patients	58.11 (58.11-58.11)	60.61 (60.61-60.61)	55.61 (55.61-55.61)
train	AF	50	record	65.07 (65.01-65.13)	65.12 (65.03-65.2)	65.21 (65.13-65.28)
train	AF+NSR	100	patients	69.58 (69.53-69.64)	33.34 (33.2-33.48)	82.06 (82.0-82.11)
train	AF+NSR	100	records	75.62 (75.58-75.65)	40.8 (40.71-40.89)	87.44 (87.4-87.47)
train+test	AF+NSR	200	patients	64.45 (64.41-64.48)	28.31 (28.22-28.39)	77.65 (77.6-77.69)
train+test	AF+NSR	200	records	70.03 (70.01-70.06)	32.65 (32.59-32.72)	83.63 (83.6-83.66)

Table 6: Results for KNN model with *k*=3

References

[1] Moody G, Goldberger A, et al. Predicting the onset of paroxysmal atrial fibrillation: the Computers in Cardiology Challenge 2001. In Computers in Cardiology 2001. Vol.28 (Cat. No.01CH37287). Rotterdam, Netherlands: IEEE. ISBN 978-0-7803-7266-5, 2001; 113–116.

[2] Mohebbi M, Ghassemian H. Prediction of paroxysmal atrial fibrillation based on non-linear analysis and spectrum and bispectrum features of the heart rate variability signal. Computer Methods and Programs in Biomedicine January 2012; 105(1):40–49. ISSN 01692607.

[3] Boon K, Khalil-Hani M, Malarvili M. Paroxysmal atrial fibrillation prediction based on HRV analysis and non-dominated sorting genetic algorithm III. Computer Methods and Programs in Biomedicine January 2018;153:171–184. ISSN 01692607.

[4] Narin A, Isler Y, et al. Early prediction of paroxysmal atrial fibrillation based on short-term heart rate variability. Physica A Statistical Mechanics and its Applications November 2018;509:56–65. ISSN 03784371.

[5] Vandewiele G, Dehaene I, et al. Overly Optimistic Prediction Results on Imbalanced Data: a Case Study of Flaws and Benefits when Applying Over-sampling. Artificial Intelligence in Medicine January 2021;111:101987. ISSN 09333657. ArXiv:2001.06296 [cs, eess, stat].

[6] Walsh I, Fishman D, et al. DOME: recommendations for supervised machine learning validation in biology. Nature Methods October 2021;18(10):1122–1127. ISSN 1548-7091, 1548-7105.

[7] Shah RU, Bress AP, Vickers AJ. Do Prediction Models Do More Harm Than Good? Circulation Cardiovascular Quality and Outcomes April 2022;15(4). ISSN 1941-7713, 1941-7705.

Corresponding author:

Cédric Gilon
cedric.gilon@ulb.be