Knowledge, Machine Learning and Atrial Fibrillation: More Ingredients for a Tastier Cocktail

Tomas Teijeiro[1]

[1] Embedded Systems Laboratory (ESL), EPFL, Lausanne, Switzerland

Abstract

Fifty years after the publication of the first algorithms for the automatic detection of Atrial Fibrillation (AF), this cardiac condition is still the most studied from the computer science and engineering perspectives. Machine learning techniques are widely applied to a variety of problems, including detection, characterization, prediction and simulation, in general with promising results. In the last years, the Big Data + Deep Learning binomial is getting most of the attention in academia and industry, but in many occasions this approach fails on capitalizing all the knowledge adquired in previous decades of research.

This article, written as a companion to the keynote with the same title presented in the CinC 2020 conference, tries to illustrate the importance of exploiting expert knowledge and classical approaches in sinergy with the most advanced deep learning methods, which by themselves have fundamental limitations. The discussion is built around the AF detection problem and the conclusions extracted from the Physionet/CinC Challenge 2017, but the main points are generally applicable to problems for which humans have a better answer than computers, and this answer can be described.

1. Introduction

The application of computer science to study atrial fibrillation has been explored for more than 50 years. Probably one of the first references is the work by Gersch et. al. [1], aimed at AF classification. Interestingly, this is still the most studied problem nowadays, even if that first paper reported a classification accuracy of 100%.

AF is one of the cardiac conditions with a simpler diagnosis procedure based on the ECG signal, that can be summarized in three conditions [2]:

- Irregularly irregular heart rate.
- Absence of P waves.
- Presence of f waves.

This apparently simple description makes it particularly suitable for formalization, and together with the high performance obtained by the first proposed methods [3] has led sometimes to a consideration of "easy problem".

However, this is just the tip of the iceberg, and a simple bibliographic query[1] shows that just in the last 20 years, almost 275 000 new publications have arisen on the topic of AF detection. It is important to note that under this generic naming there is a wide range of problems conditioned on many variables, and with incomparable difficulties. We may target different AF types (paroxysmal, persistent, permanent, . . .), faced just with normal sinus rhythms or with other possible concurrent arrhythmias, using different signals for detection (ECG, PPG, respiration, . . .), in a short-term or a continuous monitoring scenario, targeting an embedded implementation in a wearable device or a server-oriented deployment, etc [2].

Moreover, there are plenty of other problems related to AF management beyond detection, and that are attracting a lot of interest from the computer science community. Here we can mention the extraction and characterization of f waves, that have found to be extremely valuable for the prediction of treatment outcomes [4] or the prognosis of patients undergoing catheter ablation [5] in the context of personalized interventions. Other interesting topics are the development of AF simulation models [2], patient risk stratification based on the electronic health record [6], or medication management [6].

Regarding the computational techniques used to tackle these problems, the evolution seems to be clearly guided by the availability of data. Initially, efforts focus on finding parameters that show a statistically significant variation during the target condition [2]. The search is limited to a few parameters that have a supporting clinical hypothesis, and the validation is done on a small population from the same hospital or health centre (usually less than 20 subjects). Then, when more data becomes available, ideally in a public fashion, there is an explosion of explorative approaches trying to exploit all the available data with sophisticated methods, focusing on optimizing a performance metric. This is the stage in which we can find almost any combination of features, preprocessing methods,

[1] Search in Google Scholar with query ("atrial fibrillation" detection)

and machine learning algorithms, curiously most of them claiming their superiority over the others. Finally, since we are in the Deep Learning era, this technique is nowadays the culmination of this exploration, receiving most of the academic interest even if in practice it rarely proved an advantage over more classical approaches [2, 7].

In the following sections we will motivate the need to consider formalized expert knowledge as a key part of machine learning-based methods for AF management, taking as a reference the AF detection problem on ECG signals. We will illustrate how pure data-driven approaches based on deep learning face fundamental limitations, particularly with small and medium-size datasets, and how hybrid approaches integrating domain-specific knowledge can provide advantages not only in terms of model interpretability, but also if we just focus on model performance.

2. Intrinsic limitations of data-drive approaches: A toy example

To illustrate the fundamental downside of approaches that attempt to build models of cardiac events solely from data, we performed a simple experiment in which we tried to solve the following problem:

Classification Problem: *Given a single-lead ECG segment of 30 seconds, and its derived RR sequence* $\{RR_1, \ldots, RR_n\}$ *measured in milliseconds, the segment is classified as* ***positive*** *if and only if:*

$$\exists i \in \{1, \ldots, n-2\} \mid max(RR_i, RR_{i+1}, RR_{i+2}) < 500 \vee min(RR_i, RR_{i+1}, RR_{i+2}) > 1000. \quad (1)$$

According to this definition, any ECG segment can be unambigously classified as ***positive*** or not, with perfect class separability. It is an undoubtedly easy classification problem, but it requires to perform four basic operations on the input signal:

1. Detection of QRS complexes.
2. Differentiation to extract the RR series.
3. Counting up to 3.
4. Performing a logical OR operation.

We tried to solve this toy problem with a neural network architecture that has demonstrated good performance in arrhythmia detection from the raw ECG [8], and using the MIT-BIH Arrhythmia database as source data [9]. The 46 recordings containing the MLII lead were selected, and splitted into two groups of 37 and 9 for training/testing. Then, each recording was cut in segments of 30 seconds, with 29 seconds of overlap between consecutive segments, and assigned a binary label according to equation 1. The resulting total size of the training set was 65712.

The prevalence of the ***positive*** label both in the training and test sets was around 25%, which is a not so high unbalance for this type of problems, and therefore no specific actions were taken to correct this. For training, we used the same optimizer and hyperparameter values described in [8].

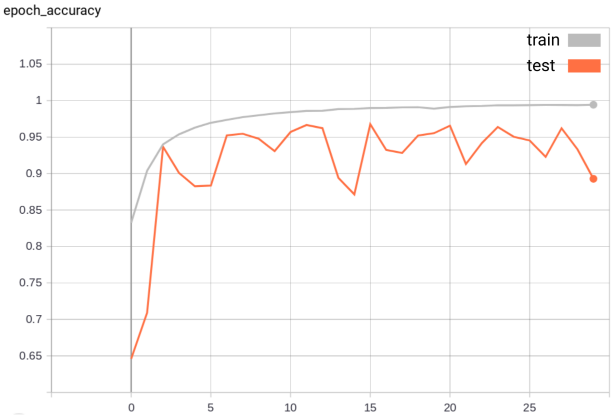


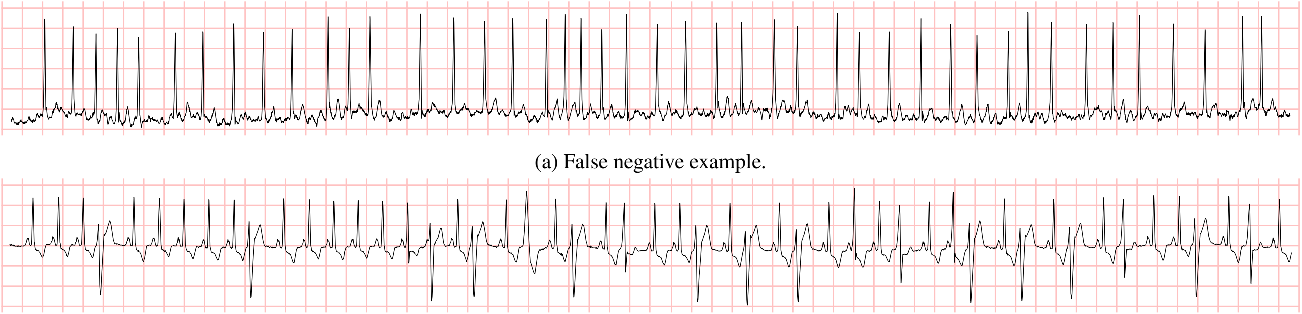
Figure 1: Learning curve of the neural network on a toy example problem.

Figure 1 shows the learning curve during 30 epochs. We can see that the validation accuracy rapidly converges to a range between 90% and 95%, and if we use the F1 score to take into account the class unbalance the value is around 0.82. We believe this is still a quite remarkable result considering that the network has been trained exclusively with the raw ECG, but it illustrates the difficulties to fit some basic quantitative conditions that usually guide humans when they perform such sort of classification tasks.

If we take a look at the classification errors of the final model, we can see that in this particular case the main difficulty relies on learning the rule of *"three consecutive RR intervals"*. Figure 2a shows an example of a segment wrongly classified as negative, probably because it has exactly 3 consecutive RR intervals under the 500 ms threshold. On the other side, Figure 2b shows an example of a false positive. It is a segment with many ectopic beats, leading to many RR intervals under the 500 ms threshold, but these are isolated. Interestingly, even if the labels distribution in the training set is biased towards the negative class (75%), most of the classification errors are false positives (1536 vs 177 false negatives).

3. The importance of hybrid strategies: Lessons from the Physionet Challenge 2017

The Physionet/CinC challenge 2017 is the latest public testbed for the evaluation of algorithms for AF detection in short single-lead ECG recordings. Algorithms should classify segments of 30 seconds duration in one of four classes (normal rhythm, AF, other rhythm or noise) using a traning set of 8528 records, while the evaluation was per-


(a) False negative example.

(b) False positive example.

Figure 2: Illustrative misclassified examples by the neural network.

formed by the challenge organizers on a hidden test set of 3658 records.

Figure 3 shows the performance obtained by the top-25 algorithms in the follow-up stage of the Challenge [10], grouped by three categories:

- **Classical ML**: Solutions relying on handcrafted features linked to expert knowledge, and a classical machine learning model (typically ensembles of decision trees via gradient boosting or random forest, or support vector machines).
- **Pure DL**: Neural network models (typically convolutional or recurrent neural network architectures) trained directly on the raw ECG signal.
- **Hybrid with DL**: Hybrid solutions that combine Classical ML and DL models using handcrafted features linked to expert knowledge.

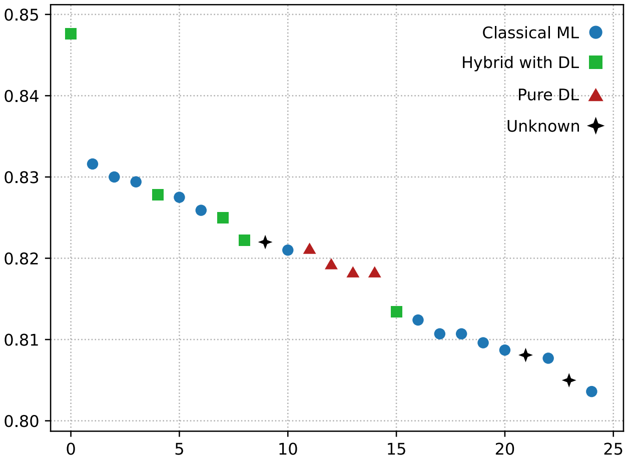


Figure 3: Performance of the top-25 algorithms in the Physionet/CinC Challenge 2017.

The first conclusion we can get is that machine learning definitely has an advantage over knowledge-based systems and single or multi-variate statistical methods (no such methods were proposed, or achieved the top-25 rank), particularly when there are no consistent definitions of the target labels. The ability to automatically fit a decision function can help not only to optimize the score on a particular target dataset, but also to highlight the inconsistencies among different human labelers [11], potentially leading to an improved definition of the target conditions.

On the other hand, it is interesting to see that Pure DL approaches are all clustered around the same positions and with minimum score differences among them. Apparently, the network type and architecture, the employed optimizer or the selected hyperparameters are of very little importance, and the main limitation seems to come from not exploiting the domain knowledge embedded in the features employed by other approaches. These features are mostly encoding the morphological and rhythm information that is known to be related to AF, and for this information to arise just from the raw ECG it would probably be necessary a dataset several orders of magnitude larger [12].

Finally, regarding the best-performing solution [13], which I was lucky to participate in its development, we can highlight two main aspects that were not explored in any of the other approaches: 1) the interpretation of the ECG in multiple abstraction levels prior to the feature extraction, and 2) the partial relabeling of the training set to improve its internal consistency. The interpretation is done in a pure knowledge-based fashion, and it ends up describing the ECG as a sequence of waves, and also as a sequence of rhythms. This makes it possible to create features that are more abstract, such as for example *"proportion of the record length in which the patient showed a non-regular rhythm"*.

As for the relabeling of the training set, it demonstrates the importance of having internal consistency between features and labels, which can be summarized as follows: *"If your features don't grasp the information required to distinguish two examples, it is better that they have the same label"*. The strategy followed to relabel was based on the errors made by the classifier during cross-validation, and following the assumption that if an error is due to a mislabeled example, changing the label would improve the classification performance for two classes. On the contrary, if the error is due to an outlier, changing the label would only improve the performance on that class, but penalize the new one.

4. Conclusion

We do not know if in another 50 years atrial fibrillation will continue to pose unsolved challenges in computer science and engineering. What seems likely is that machine learning will be behind the solution to many of the problems that will be crossed off the list. However, since the human cardiovascular system will continue to behave in exactly the same way as it does today, and as it has for the past 150 000 years, all the effort we devote to formalizing that knowledge and apply it to problem solving will result in more robust, reliable, interpretable, and trustworthy systems.

Acknowledgments

This work has been supported by the H2020 DeepHealth Project (GA No. 825111).

References

[1] Gersch W, Eddy DM, Dong E. Cardiac arrhythmia classification: A heart-beat interval-Markov chain approach. Computers and Biomedical Research 1970;3(4):385–392. ISSN 00104809.

[2] Sörnmo L (ed.). Atrial Fibrillation from an Engineering Perspective. Series in BioEngineering. Cham: Springer International Publishing, 2018. ISBN 978-3-319-68513-7.

[3] Moody G, Mark R. A new method for detecting atrial fibrillation using R-R intervals. Computers in Cardiology 1983; 227–230.

[4] Bollmann A, Kanuru NK, McTeague KK, Walter PF, DeLurgio DB, Langberg JJ. Frequency analysis of human atrial fibrillation using the surface electrocardiogram and its response to ibutilide. American Journal of Cardiology 1998;81(12):1439–1445. ISSN 00029149.

[5] Alcaraz R, Hornero F, Rieta JJ. Electrocardiographic Spectral Features for Long-Term Outcome Prognosis of Atrial Fibrillation Catheter Ablation. Annals of Biomedical Engineering 2016;44(11):3307–3318. ISSN 15739686.

[6] Siontis KC, Yao X, Pirruccello JP, Philippakis AA, Noseworthy PA. How Will Machine Learning Inform the Clinical Care of Atrial Fibrillation? Circulation Research 2020; 127(1):155–169. ISSN 0009-7330.

[7] Clifford GD. The Future AI in Healthcare: A Tsunami of False Alarms or a Product of Experts? arXiv Preprint 2020; URL `http://arxiv.org/abs/2007.10502`.

[8] Warrick P, Homsi MN. Cardiac arrhythmia detection from ecg combining convolutional and long short-term memory networks. In Computing in Cardiology, volume 44. IEEE Computer Society, 2017; 1–4.

[9] Moody GB, Mark RG. The impact of the MIT-BIH arrhythmia database, 2001.

[10] Physionet challenge 2017: Full test set scores for follow-up entries released, 2018. URL `https://groups.google.com/g/physionet-challenges/c/qA2iUfQmRtc/m/OHQPImxEAwAJ`.

[11] Clifford GD, Liu C, Moody B, Lehman LH, Silva I, Li Q, Johnson AE, Mark RG. AF classification from a short single lead ECG recording: The PhysioNet/computing in cardiology challenge 2017. In Computing in Cardiology, volume 44. IEEE Computer Society, 2017; 1–4.

[12] Attia ZI, Noseworthy PA, Lopez-Jimenez F, Asirvatham SJ, Deshmukh AJ, Gersh BJ, Carter RE, Yao X, Rabinstein AA, Erickson BJ, Kapa S, Friedman PA. 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 2019; 394(10201):861–867. ISSN 01406736.

[13] Teijeiro T, García CA, Castro D, Félix P. Abductive reasoning as the basis to reproduce expert criteria in ECG Atrial Fibrillation identification. Physiological Measurement 2018;39(8).

Address for correspondence:

Tomas Teijeiro
tomas.teijeiro@epfl.ch