Detecting AF Burden Using 1-D CNNs

Hasan Khan | CS3033 083

Abstract

1-D CNNs have recently been used to classify various classes of Arrhythmias using ECG sequences. In this paper, a 1-D CNN is used to predict the Atrial Fibrillation (AF) burden of ECG sequences, which is a useful continuous metric that helps gauge AF severity and can be predictive for risk of near term stroke.

1 Introduction

Anywhere from 2.7-6.0 million Americans suffer from Atrial Fibrillation, the leading type of heart arrhythmia (1). Atrial Fibrillation is characterized by an irregular heartbeat; if left untreated, it can lead to blood clots, stroke and other serious heart complications. Historically, cardiologists and technicians manually scan Electrocardiogram (ECG) recordings of patients to determine the presence of AF in patients, often with the assistance of peak detection and interval analysis algorithms. However, these algorithms have a high error rate (2), and in recent years, a plethora of statistical models and machine learning techniques such as Convolutional Neural Networks (CNNs) (3) and Markov Models (4), among others, have been successfully applied to arrhythmia and AF detection problems, often out preforming traditional methods. These models are attractive because they can help save cardiologist time, increase accessibility to AF diagnosis, and are especially effective for continuous monitoring when paired with wearable technologies that can measure ECG signals.

However, these techniques are often focused on the classification and sub typing of ECG sequences. In this project, the objective is modified to calculate the AF Burden of each sequence instead, a continuous variable defined as the proportion of an ECG sequence that displays AF characteristics. AF Burden has been shown to be associated with near term stroke risk (5), and depending on the duration of the ECG, can be used to help classify between variants of AF, such Paroxysmal AF (episodic AF) and Persistent AF (chronic AF).

This project builds on previous AF classification techniques, modifying them slightly for the purposes of predicting AF Burden. ECG signals from the The 4th China Physiological Signal Challenge (6) are used as the input data for the model, from which the project is inspired.

2 Method

2.1 Data

Labeled ECG signal data from The 4th China Physiological Signal Challenge is used for this project. All data is publicly available and hosted on PhysioNet (7).

The dataset is comprised of 1436 ECG signals recorded using 12-lead Holter or 3-lead wearable ECG monitoring devices. The ECG signals are variable in length, and extracted from lead I and lead II of the long-term dynamic ECGs sampled at 200 Hz. The ECG signals stem from 105 patients in total; 49 of the readings come from AF affected patients (22 PAF patients) and 56 readings come from non AF patients.

Each record originally includes the raw ECG signal, an annotation of signal timestamps for detected peaks, and two labels: one overall sequence label consisting of 3 classes (non atrial fibrillation, paroxysmal atrial fibrillation, persistent atrial fibrillation) and one granular label consisting of AF

start and stop indexed at the heartbeat level. A True AF burden label is generated for each record using the ranges provided in the granular label and annotated peak timestamps.

2.2 Process

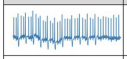
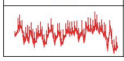
The objective of this project is to predict AF Burden. Rather than target for AF Burden directly as a regression problem, the problem is simplified to a binary classification task.

The following is a high level description of the methodology used. In the first step, all ECG signals are normalized to lie between 0 and 1, and then broken down into labeled, equal sized n-second "chunks", where n is evaluated over 30, 10, 5 (i.e increasing granularity). Sequence labels consist of 3 classes: non atrial fibrillation, paroxysmal atrial fibrillation, and persistent atrial fibrillation, and are accordingly transformed into chunk labels, which consist of 2 classes: AF and Non AF. See figure 1 (top tables) for a visual depiction of this transformation. The data is then split into training and testing sets, and a 1-D CNN is trained on the chunked signals of the training data. The model then makes label predictions for signal chunks in the test set.

Once the chunk level predictions are generated on the test data, the predictions are aggregated back to the sequence level[1], and for each sequence a Predicted AF Burden is derived from the proportion of that sequences' chunks that are labeled or not labeled with AF. The exact calculation for the Predicted AF Burden of each sequence is as follows, where 200 is the sampling frequency in hertz and chunksize is 30, 10, or 5 seconds depending on the data variant:

$$\text{Predicted AF Burden} = \frac{\text{\# AF Labeled Chunks} \times \text{chunk size} \times 200}{\text{Signal Length}}$$

See figure 1 (bottom table) for an example of some key columns from the annotated test data. The main motivation for this approach (rather than simply modeling for AF Burden directly) is that generating granular, chunk-level AF predictions gives more information on the ranges of AF for a given sequence. Note that this is a related task and the primary challenge in the The 4th China Physiological Signal Challenge, but is not explicitly explored in the scope of this project.

Signal	Signal Label	Beat Level Annotations	True AF Burden*
	Normal	[Non AF, Non AF ... Non AF, Non AF]	0.0
	Paroxysmal AF	[Non AF, Non AF, ... AF, AF]	0.5
	Persistent AF	[AF, AF, ... AF, AF]	1.0

Chunked Signal	Chunked Label
	Normal
	Normal
	Normal
	Normal
	AF
	AF
	AF
	AF

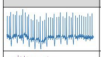
Signal	True Signal Label	True AF Burden	Predicted Chunk Labels	Predicted AF Burden
	Normal	0.0	[Non AF, Non AF, Non AF]	0.0
	Paroxysmal AF	0.5	[AF, AF]	1.0
	Persistent AF	1.0	[Non AF, AF, AF]	0.66

Figure 1: Top: A visualization of the data chunking transformation. All signal images are snippets of signals from the training set. Chunked labels are derived from the provided beat level annotations. Bottom: A mock example of the annotated test data output (note: signals are recycled from above; not the same datapoints).

[1]Chunk labels are concatenated into a list; see the "Predicted Chunk Labels" column in figure 1

2.3 Data Variants & Train-Test Split

During the chunking process, three variants of the original dataset are created: one variant is split into 30-second chunks, another into 10-second chunks, and a third into 5-second chunks. Different n-second chunk sizes were evaluated to determine whether the model could better approximate AF Burden values with finer grained splits (thus resulting in more precise AF Burden ratios) without compromising on the accuracy of the chunk level binary classification (which should theoretically be impacted by chunk size).

Both datasets are split into training and testing sets using a 70/30 split on the full sequences; n-second chunks from 996 sequences are used for the training set, and 444 sequences are used for the testing set. Note that care is taken to maintain all chunks from a single record remain in either the train set or the test set (i.e the chunks of one signal are not split between the train and the test).

2.4 Modeling and Evaluation

A modification of a preexisting 1-D Convolutional Neural Network architecture (8) was the selected model for this project. The model is comprised of 9-layers: 4 convolution layers, 1 max pooling layer and 2 dense layers. A dropout of 0.3 is used as a regularization method to improve on generalization error and reduce over fitting. ReLu activation functions are applied at each layer to overcome the vanishing gradient problem. A batch size of 16 is used, and each model is run on 3 epochs. A binary cross entropy loss function is chosen since the task is binary classification. See figure 2 for a visualization of this architecture.

For the chunk classification task, accuracy and F1 scores are computed on the test set. For AF Burden prediction, a mean absolute error (MAE) is generated on the test set.

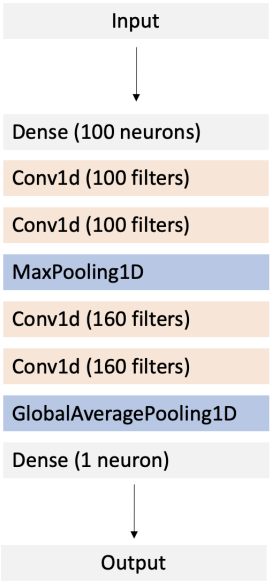


Figure 2: Model architecture

3 Results

Table 1: Chunk Classification Results

Chunk size	Accuracy	F1
30-second chunk size	0.9283	0.9281
10-second chunk size	0.9583	0.9582
5-second chunk size	0.9707	0.9706

Table 2: AF Burden MAE

Chunk size	Mean Absolute Error
30-second chunk size	0.118
10-second chunk size	0.083
5-second chunk size	0.046

The results for the classification task and the AF Burden prediction task are shown in the tables above. The results show that the model run on data chunked at the smallest granularity (5 seconds) obtained the best classification results and lowest MAE scores for AF Burden. This may result from the fact that splitting the original dataset into 5 second chunks naturally results in more train/test data than splitting the original data on larger granularities; it may also suggest that longterm signal dependencies (i.e more than 5 seconds) along an ECG sequence do not make much of an impact on the classification task.

Furthermore, scatter plots comparing the True AF Burden with the predicted AF Burden for each datapoint in the test set show the directionality of AF Burden error and serve to visually confirm the MAE scores. The diagonal represents a perfect classification (i.e true burden = predicted burden). The dataset with the smallest chunk sizes (n=5) achieves the best AF Burden scores (observe the tighter fit to the diagonal than n=10 and n=30), in line with intuition discussed in Section 2.3.

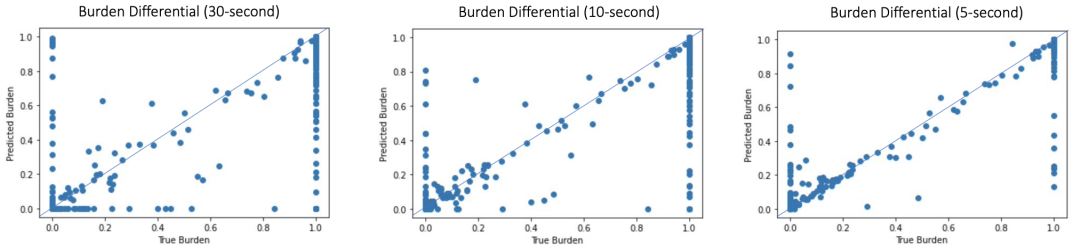


Figure 3: Burden differentials

The combination of high accuracy and F1 scores suggest that the model performs well on the chunk classification task. The low MAE scores on AF burden prediction suggest that AF Burden prediction is a tractable problem. Still, checking the test data against more evaluation metrics will be needed to draw any further conclusions on the AF Burden prediction task.

Admittedly, both the chunk classification task and the AF burden prediction task do not have baselines results upon which to gauge the above results against. Future work on this task would ideally include a baseline model result or label annotations from a panel of cardiologist to get a better understanding of model performance.

4 Discussion

This project represents a first step toward approximating AF burden values from ECG records. Immediate caveats to be addressed include evaluating different model architectures and conducting hyper-parameter tuning to find faster, more efficient models without decreasing the chunk classification accuracy and low AF burden MAE scores achieved. A comparison between the binary classification approach used in this project and directly modeling for AF Burden given the same data would be a relevant area of exploration. Additionally, predicted labeled chunks and AF burden values could be utilized to predict the exact start and stop times for AF events in individual signals. Furthermore, a simple multi-class classification of the given signals could be performed to categorize the signals as Paroxysmal AF, Persistent AF or Non AF.

Future steps could involve conducting similar AF Burden prediction tasks on data collected using Photoplethysmography(PPG), which has been shown to generate positive results for AF detection and classification (9), and is commonly found in wearable technologies.

References

[1] M. K. Chung, L. L. Eckhardt, L. Y. Chen, H. M. Ahmed, R. Gopinathannair, J. A. Joglar, P. A. Noseworthy, Q. R. Pack, P. Sanders, K. M. Trulock, and null null, "Lifestyle and risk factor modification for reduction of atrial fibrillation: A scientific statement from the american heart association," *Circulation*, vol. 141, no. 16, pp. e750–e772, 2020.

[2] A. Shah and S. Rubin, "Errors in the computerized electrocardiogram interpretation of cardiac rhythm," *Journal of electrocardiology*, vol. 40, pp. 385–90, 09 2007.

[3] P. Rajpurkar, A. Y. Hannun, M. Haghpanahi, C. Bourn, and A. Y. Ng, "Cardiologist-level arrhythmia detection with convolutional neural networks," *CoRR*, vol. abs/1707.01836, 2017.

[4] D. Coast, R. Stern, G. Cano, and S. Briller, "An approach to cardiac arrhythmia analysis using hidden markov models," *IEEE transactions on bio-medical engineering*, vol. 37, pp. 826–36, 10 1990.

[5] L. Y. Chen, M. K. Chung, L. A. Allen, M. Ezekowitz, K. L. Furie, P. McCabe, P. A. Noseworthy, M. V. Perez, and M. P. Turakhia, "Atrial fibrillation burden: Moving beyond atrial fibrillation as a binary entity: A scientific statement from the american heart association," *Circulation*, vol. 137, no. 20, pp. e623–e644, 2018.

[6] W. Xingyao, M. Caiyun, Z. Xiangyu, G. Hongxiang, C. Gari, and C. Liu, "Paroxysmal atrial fibrillation events detection from dynamic ecg recordings: The 4th china physiological signal challenge 2021," *PhysioNet*, 2021.

[7] A. L. Goldberger, L. A. N. Amaral, L. Glass, J. M. Hausdorff, P. C. Ivanov, R. G. Mark, J. E. Mietus, G. B. Moody, C.-K. Peng, and H. E. Stanley, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000 (June 13). Circulation Electronic Pages: http://circ.ahajournals.org/content/101/23/e215.full PMID:1085218; doi: 10.1161/01.CIR.101.23.e215.

[8] M. Chan, "1d-cnn-for-ecg-classification." `https://github.com/manduchan/1D-CNN-for-ECG-Classification`, 2020.

[9] M. Voisin, Y. Shen, A. Aliamiri, A. Avati, A. Hannun, and A. Ng, "Ambulatory atrial fibrillation monitoring using wearable photoplethysmography with deep learning," 2018.