Detection of Cardiac Arrhythmias from Varied Length Multichannel Electrocardiogram Recordings Using Deep Convolutional Neural Networks

Marwen Sallem¹, Adnen Saadaoui², Amina Ghrissi³ and Vicente Zarzoso³

Abstract

Automatic identification of different arrhythmias helps cardiologists better diagnose patients with cardiovascular diseases. Deep learning algorithms are used for the classification of multichannel ECG signals into different heart rhythms. The study dataset includes a cohort of 43101 12-lead ECG recordings with various lengths. Two options were tested to standardize the recordings length: zero padding and signal repetition. Downsampling the recordings to 100 Hz allowed to handle the problem of different sampling frequencies of data coming from different data sources. We design a deep one-dimensional convolutional neural network (CNN) called VGG-ECG. VGG-ECG is a 13-layer fully CNN for multilabel classification. The test score values (S) on the local and the hidden test sets are respectively 0.47 and 0.37 compared to S;0.01 for baseline models with logistic regression.

1. Introduction

Cardiovascular diseases represent an underlying cause of death worldwide as they are responsible for about one million deaths annually in the United States alone. Hence, contributing to the automatic identification of different arrhythmias with deep learning (DL) tools would help cardiologists better diagnose patients with cardiovascular diseases. For this reason, the current work aims to classify multichannel ECG signals into 27 heart rhythms using deep learning (DL) as part of Physionet challenge 2020 [2] [3] [4]. The study dataset provided by the challenge includes a cohort of 43101 12-lead ECG recordings. The main issues encountered in the classification task consist in the varied lengths of multichannel ECG recordings added to the acquisition different characteristics of samples coming from different data sources.

We design a deep one dimensional (1D) convolutional neural network (CNN) inspired by VGG16 architecture [5] [6] and regularized with dropout [7]. Data preprocessing and preparation is also performed. Long recordings are truncated and only the first 3000 time samples are kept to alleviate the training load. Two options are tested to standardize the recordings’ lengths: zero padding and signal repetition.

The designed solution is expected to help cardiologists distinguish 27 scored cardiac rhythms out of 111.

2. Methods

2.1. Classification Metrics

A variety of metrics is used to evaluate the classification performance. In the unofficial phase of the challenge, metrics used include: F2-score, G2-score and geometric mean (GM). The metrics formulas combine precision and recall [4]. In the official phase, one single score (S) is computed. S generalizes the traditional accuracy metric by giving full credit to correct diagnoses and penalizing wrong diagnosis, with respect to similarities between arrhythmias. $s = \sum_i w_{ij}a_{ij}$, where $a_{ij}$ is an element of the confusion matrix equal to the number of samples classified as belonging to category $c_i$ but that truly belong to $c_j$. Moreover, different weights $W = [w_{ij}]$ are assigned to a couple of classes $(c_i, c_j)$ based on their similarity levels. The higher the values of the scores are the better our model is [8].

2.2. VGG-ECG

VGG-ECG is a 13-layer fully CNN composed of blocks of 2 and 3 successive convolutional layers (Conv) alternated by average pooling layers. 1D Conv layers have a receptive filed of dimensions $1 \times 3$ to scan information along the temporal dimension of the leads. It is denoted Conv3 as shown in the Fig. [1] This architecture proved efficient in a similar multichannel ECG classification task [5]. Dropout with a drop probability equal to 0.3 is used to regularize the model and avoid overfitting [7]. ADAM optimizer with learning rate equal to $10^{-4}$ [9] is used for training. Binary cross-entropy loss is also used to allow multilabel classification of recordings in more than one class $(c_i)$. In order to avoid overfitting, training is stopped when the value of validation loss does not improve significantly. Then the...
trained model reaching the best performance is selected. Random seeds are fixed in order to make experiments reproducible when training on GPUs.

3. Study Dataset

The public dataset afforded by the challenge committee for training and validation is huge and composed of 43101 12-lead ECG recordings coming from four different data sources. Recordings have different lengths varying from 30 s to 30 min. The sampling frequencies ($F_s$) are also different as described below:

1. **Source 1** is composed of two sets of 12-lead ECGs of lengths between 6 and 60 s with $F_s=500$ Hz. This data represents the preliminary dataset afforded for the unofficial submission phase of Physionet challenge. The first set contains 6877 recordings coming from 3699 male (M) and 3178 female (F). The second set contains 3453 recordings coming from 3453 M and 1610 F.

2. **Source 2** contains 75 recordings of 30 min length each and sampled at 257 Hz.

3. **Source 3** is composed of two sets of 12-lead ECGs of length 10 s each. The sets include 549 recordings of 377 M and 139 F ($F_s=1000$ Hz) and 21837 recordings of 11379 M and 10458 F ($F_s=500$ Hz) respectively.

4. **Source 4** contains 10344 12-lead ECG recordings coming from 5551 M and 4793 F of length 10 s sampled at 500 Hz.

4. Experiments and Results

First experiments are conducted on the preliminary dataset afforded for the unofficial submission phase of Physionet challenge. Then final results are conducted on the full dataset described in Sec. 3.

4.1. Preliminary Results

In order to handle class imbalance, class weights of the neural networks nodes are configured to ensure balance. Data preprocessing includes zero padding, considering only the first 18000 samples of the padded recordings. The training scores F2-score, G2-score and geometric mean (GM) after 5-fold cross validation are respectively 0.77, 0.58, 0.67 with zero padding and 0.74, 0.54, 0.63 with signal repetition. The scores’ standard deviations are below 0.02. By analyzing the confusion matrix, we notice that ST-segment elevation (STE) class has poor sensitivity (0.43). It can be explained by the low number of samples (220) compared to atrial fibrillation, for instance, that has 1221 samples and high sensitivity (0.93). Results on the test set are F2-score=0.77, G2-score=0.55 and GM=0.65.

4.2. Final Results

Regarding the difference of the recordings’ lengths that vary from 3 s and 30 min, we propose to consider only the first 3000 temporal samples and ignore the rest. We expect the first part of each recording to contain sufficient information about the heart rhythm type. Signals that are shorter than 3000 samples will be zero padded. Truncating the first 3000 samples followed by downsampling allows to alleviate considerably the computational load of training and validation process.

In order to handle the problem of different sampling frequencies we aim to standardize $F_s$ by downsampling all recordings to a standard frequency ($F_d$). We benchmark the classification performance on the test set with respect to three values of $F_d$ as shown in Tab. 1.

<table>
<thead>
<tr>
<th>$F_d$ (Hz)</th>
<th>50</th>
<th>100</th>
<th>150</th>
</tr>
</thead>
<tbody>
<tr>
<td>$S$</td>
<td>0.25</td>
<td>0.47</td>
<td>0.44</td>
</tr>
</tbody>
</table>

Table 1. Effect of $F_d$ on test classification scores.

After that, the model is submitted to the challenge committee to be assessed on a hidden test set. The results are shown in Fig. 1. VGG-ECG reaches $S=0.37$ compared to baseline models (Baseline I and II), proposed by the challenge, which scores’ values do not exceed 0.1. This result can be explained by the ability of CNNs to capture complex hidden features from data compared to baseline models that are based on linear regression [10].

**Implementation:** DL learning experiments are conducted with the use of Python 3.7 programming language.

![VGG-ECG architecture](image-url)
and Tensorflow 2.3 deep learning library within Google Cloud Platform (GCP), running on n1-standard-4 (4 vCPUs, 15 GB memory) machine and NVIDIA Tesla T4 Virtual Workstation GPU. Wandb [11] ML tool is used to track the training performance and visualize statistics about the model.

5. Conclusions and perspectives

The classification of multichannel ECG recordings using DL helps cardiologists to automatically detect different arrhythmia types added to sinus rhythm. In order to handle the varied length recordings and the imbalanced dataset issues, we benchmark several solutions. Two options are tested to standardize the recordings length: zero padding and signal repetition. Moreover, all signals are downsampled to 100 Hz in order to standardize the different sampling frequencies. Several settings are also assessed like adding residual layers and customizing classification thresholds but did not enhance the performance. The multilabel classification performed considers only the 27 scored classes, as stated by the challenge guidelines, and ignores the remaining classes. By benchmarking the classification performance of several architectures, we opt for a deep 1D CNN model inspired by VGG16 for multilabel classification with binary cross entropy loss.

The test score obtained on our test set is equal to 0.47. However, the score on the hidden test set is equal to 0.37 which outperforms significantly baseline models.

On the first hand, further work will investigate data augmentation [12] [13] technique to handle the issue of imbalanced dataset ratio between different arrhythmias. On the second hand, we will add LSTM and attention layers to the CNN model as they are expected to learn automatically salient dependencies from ECG time series [5] [14].

Acknowledgments


References

[4] [GitHub of Physionet 2020 Challenge]

Address for correspondence:
Marwen Sallem
sallem.marwen@yahoo.fr
Paris, France.