Ensemble Learning of Modified Residual Networks for Classifying ECG with Different Set of Leads

Federico M Muscato[1], Valentina D A Corino[1], Luca T Mainardi[1]

[1]Department of Electronics, Information and Bioengineering, Politecnico di Milano, Milan, Italy

Abstract

The automatic detection and classification of cardiac abnormalities can assist physicians in making diagnoses, saving costs in modern healthcare systems. In this study we present an automatic algorithm for classification of cardiac abnormalities included in the CinC's challenge 2021 dataset consisting of twelve-lead, six-lead, three-lead, and two-lead ECGs (team: Polimi_1). For each set of leads an ensemble of three deep learning models, trained on three different subsets, was developed. These subsets, obtained by splitting the recordings with the most frequent classes, had more balanced distributions for training and were used to train the 3 classifiers. The trained models were modified Residual Networks with a Squeeze-and-Excitation module. This module is based on the intuition of channel attention: the basic idea of this approach is to apply a weight to the Convolutional channels based on their relevance in learning before propagating to the next layer. For evaluation, we submitted our model to the official phase of the PhysioNet/Computing in Cardiology Challenge 2021. The model received scores of 0.47, 0.46, 0.45, 0.48 and 0.45 (ranked 14th, 13th, 15th, 10th, and 13th out of 39 teams) on 12-lead, 6-lead, 4-lead, 3-lead, 2-lead hidden test set, respectively; placing us in the 11th position for the mean of the 12-lead, 3-lead, and 2-lead scores.

1. Introduction

Cardiovascular disease is the leading cause of death worldwide and the electrocardiogram (ECG) is a major tool in their diagnoses [1]. Early treatment can prevent serious cardiac events and the ECG can play an important role in screening [2]. In particular the 12-lead ECG is used as the primary clinical tool to diagnose changes in heart conditions. Deep neural networks (DNNs) have recently achieved great success especially in tasks such as image classification [3] and speech recognition [4], and there are great expectations in their application in health care and clinical practice. Recent studies [5] showed their applicability with ECG recordings and in particular of one class of DNNs: Convolutional Neural Networks (CNNs). These are networks that aim at learning a compressed representation (encoding) of an input dataset with an approach similar to the biological sensorial processing of the visual cortex whose cells are sensitive to small sub regions of the visual field called receptive fields. However, it is still an open question if DNNs would be useful in a more complex setting with an inhomogeneous dataset, including many different rhythms to be classified [6]. Furthermore, there is no evidence that a reduced subset of leads could obtain similar performances to the twelve leads configuration. Indeed, it is important to define which reduced subset can capture the wider range of diagnosis.

The aim of this work is the development of five different end-to-end DNNs respectively from twelve-lead, six-lead, four-lead, three-lead, and two lead ECG recordings. These models identify 30 different diagnoses directly from raw ECG signals using the annotated datasets available for the 2021 Computing in Cardiology Challenge [7] (team name: *Polimi_1).*

2. Materials and Methods

2.1. Data

The available data consisted of 88,253 recordings from six different databases. Each ECG recording was acquired in a hospital clinical setting, but the sample frequency varies between the different dataset with values from 257 Hz to 1 kHz. Every ECG recording is accompanied with demographic information like age and sex and also the diagnoses. Each recording could have more than one diagnosis, in fact it was a multi label challenge. The organizers of the Challenge claimed that the quality of the label depended on the clinical or research practices and that they were generated by machine, over read by a single cardiologist, and finally determined by multiple cardiologists.

Regarding the labels, the training data contained 133 diagnoses, but the challenge consisted in evaluating 30 of them with the following different lead combinations:

- 12: I, II, III, aVR, aVL, aVF, V1, V2, V3, V4, V5, V6;
- 6: I, II, III, aVR, aVL, aVF;
- 4: I, II, III, V2;
- 3: I, II, V2;

Page 1 ISSN: 2325-887X DOI: 10.22489/CinC.2021.143

- 2: I, II.

2.2. Challenge Score

In this challenge the evaluation metric is a generalized accuracy called “challenge score”. This scoring metric is new and was developed by the organizers of the challenge [6]. This new metric awards partial credit to misdiagnoses of predictions that result in similar outcomes or treatment as the true actual diagnosis. This approach is important because it reflects the clinical reality in which some errors are more harmful than others. Also, this metric does not penalize too much misdiagnosis between classes that have similar responses.

2.3. Pre-Processing

All the recordings were resampled to minimum frequency of 257 Hz. To allow a fixed input size in the model each ECG was set to be 4096 points (approximately 16 seconds). This was done by truncating the part exceeding 4096 samples for longer signals and zero padding the shorter signals.

Since the dataset was unbalanced (Figure 1) in respect of Normal Sinus Rhythm (NSR) and Sinus Bradycardia (SB) labels (in particular the NSR occurred three times more with respect to the other diagnosis), the recordings with these labels were split into three different subsets. Each of these subsets was stacked with all the recordings that were not labelled as NSR and SB. In this way a more balanced distribution for training was obtained, as presented in Figure 2. The frequencies of NSR and SB are highlighted in red.

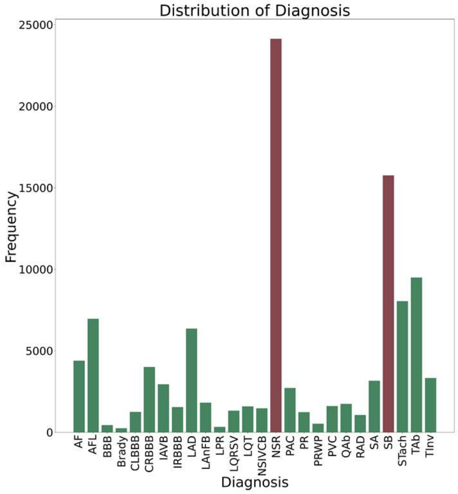

Figure 1 Training set diagnosis distribution

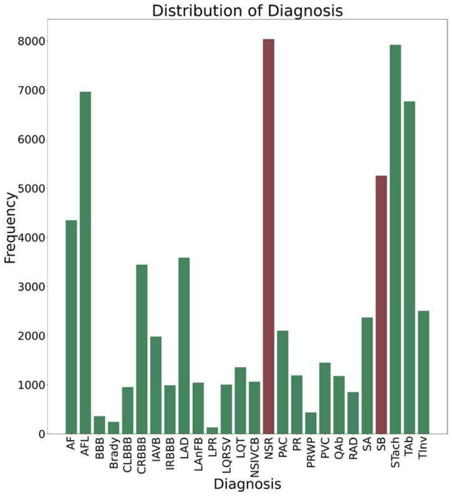

Figure 2 New Training Set diagnosis distribution with 3 splits of NSR and SB labelled recording

Since four pairs of classes were scored as if they were the same class by the weights of the challenge score these pairs were considered to be identical and simply combined. In this way the number of classes for the model were decreased to 26. Furthermore, every signal that did not have any positive scored labels after the removal of unscored labels was discarded.

Finally, in some recordings, some leads were inverted or missing and for this reason, in order to have a better generalization capability, a data augmentation procedure was developed. In details, it consisted in random applications of this technique: summation of a gaussian noise in the signal of leads randomly selected; changing the position between two or more leads; flipping the signal of one, two or more leads.

2.4. Model architecture

The model used in this work is a modified ResNet [8] that receives an input of fixed length of 4096 samples along the channels according to the number of leads required. This model consists in one convolutional layer followed by N = 8 residual block each of which contain two convolutional layers and a squeeze and excitation block [9] (see Figure 3). This module is based on the intuition of channel attention: the idea of this approach is to apply a weight to the channels based on their importance before propagating to the next layer.

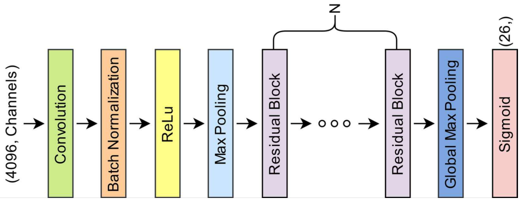

Figure 3 The modified ResNet used in this work

In the first convolution layer and inside every two residual blocks a dilated convolution was performed. This operation was performed with a dilation rate that was doubled every time starting from the value of 4. Dilated convolutions, as shown in Figure 4, are used to increase the receptive field of the network, i.e., the region in the input space that a CNN feature is affected by. Indeed, a larger receptive field improves the capability of comprehensive consideration [10]. Due to technical issues this last implementation was not present in the submission.

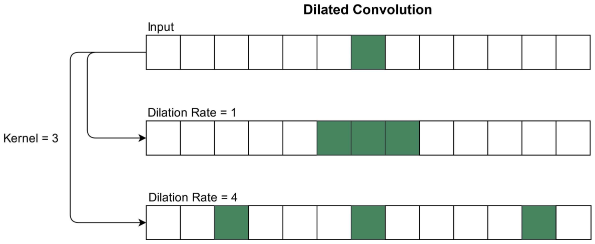

Figure 4 1D dilated convolution with a kernel size of 3 and dilation rates 1 and 4. The green block in the input signal (first row) indicates the unit of interest. In the output signals (second and third row) the green blocks show the receptive field for each different dilation rate.

The first convolutional layer and the two initial Residual blocks have 64 convolutional filters. The number of filters increases by a factor of two for every second Residual blocks. Also, in these blocks the feature dimension is halved after the max pooling. In every Residual block a dropout with factor of 0.5 is performed.

2.5. Threshold optimization

The prediction thresholds used by the network for assigning binary values for classification of every class were 0.5 as given by the sigmoid function. This choice of thresholds did not perform very well so they were optimized. The solution that gave the best results consisted in an optimization, which maximized the challenge metric, starting from different thresholds values found with a grid-search in two separate steps. First a general value within 0 and 1 with steps 0.1 score was fixed for all the classes which gave the best score. Then, by setting all the other values of other classes, the threshold for each class was progressively updated by searching separately for the value between 0 and 1 with step of 0.01 which maximized the score.

Due to technical issues, a less complex optimization was implemented in the submission. This solution consisted only in optimizing the thresholds starting from the value found in a single step. The general starting point, indeed, was set as the value within 0 and 1 with steps of 0.01 which, applied for every class, maximized the challenge metric.

2.6. Ensemble

Ensemble learning builds a set of diversified models and combines them. Theoretically and empirically, numerous studies [11] have demonstrated that ensemble learning usually yields higher accuracy than individual models; a collection of weak models (inducers) can be combined to produce a single strong ensemble model. In this work an ensemble was computed by the majority voting of three different models each of them trained on one of the three different subsets with a more balanced distribution.

3. Results

Due to technical issues the presented model could not have been submitted successfully, therefore a version without data augmentation, dilated convolution and with a more simpler threshold optimization was scored in the official phase. This solution obtained the challenge score in the hidden test set of 0.47, 0.46, 0.45, 0.48 and 0.45 on 12-lead, 6-lead, 4-lead, 3-lead, 2-lead respectively as shown in Table 1. These results placed us in the 11th position for the “all-lead” score, which is computed as the mean of the 12-lead, 3-lead, and 2-lead scores.

Leads	Validation	Test	Ranking
12	0.593	0.47	14
6	0.577	0.46	13
4	0.591	0.45	15
3	0.595	0.48	10
2	0.582	0.45	13

Table 1 Challenge scores for our final selected entry (team Polimi_1) on the hidden validation set, on the hidden test set and the ranking obtained in this last set.

In Table 2 the challenge scores on an offline validation set are presented; this set is made up of records belonging to the same datasets as those used for validation by the challenge organizers. It could be noted that the results obtained in this set by the submitted model are similar to the ones obtained in the official phase in the “validation set”. Therefore, the better results obtained by the final model could be considered consistent for comparison. This solution, indeed, improved all the scores for all the different lead combinations.

Leads	Submitted Model	Final Model
12	0.589	0.611
6	0.580	0.600
4	0.594	0.613
3	0.587	0.617
2	0.582	0.611

Table 2 Challenge Score on the offline validation set.

4. Discussion and Conclusions

The aim of this work was to develop an end-to-end model which uses raw ECG signals without filtering. In this way the capability of the model to handle noisy signals is addressed to demonstrate its applicability in realistic scenarios. For each of the five lead combinations an ensemble of three models trained on each subset were developed. This solution was implemented in order to address the great imbalance of classes presented in the dataset. Thanks to the ensemble, indeed, a more powerful model was obtained: this model had a better capability of generalization because it included the learning from all the three models. The imbalance of classes was also addressed with the optimization of the threshold for each class. This implementation was particularly important because the classes were not only unbalanced between each other, but it was far more common to have negative labels than positive ones.

A limitation of this work was using only 4096 samples of the signal. This choice was made because it sufficiently covered the major part of the signals from the available dataset and in order to take the most advantage of parallel processing power of GPU and reduce a lot of training time. Such a procedure in fact was determined by the computational limitations due to the width of the dataset.

The submitted model achieved the scores in the hidden test which are shown in Table 1. It is worth mentioning that this set was obtained combining four different hidden sets. The first two hidden test sets were from sources used also in the training set, and thus their scores were higher. On the contrary, the last two were from totally hidden undisclosed sets, and their scores were much lower suggesting that the model failed to migrate to a completely different data set effectively.

However, the final model developed in this work could not have been submitted. This model showed great performances demonstrating room for improvement: indeed, it could be asserted that with this approach the results would have been overall improved also in the validation set of the challenge and hopefully also in the hidden test set. The offline validation set in which these results are obtained has recordings from the same datasets used as validation by the organizers and the results obtained by the same models are very similar. This could be observed confronting the score of the submission with the score illustrated in the “Submitted Model” column of Table 2. Since the results of the “Final Model” are higher in all the combinations, it could be inferred that techniques of data augmentation, grid-search threshold optimization and the use of a higher dilation rate showed promising performances in classifying a higher amount of diagnosis with different characteristics.

References

[1] Benjamin E, Muntner P, Alonso A, Bittencourt M, Callaway C, Carson A, et al. Heart disease and stroke statistics 2019 update: a report from the American Heart Association. Circulation 2019; 139(10): e56.

[2] Kligfield P, Gettes LS, Bailey JJ, Childers R, Deal BJ, Hancock EW et al. Recommendations for the standardization and interpretation of the electro-cardiogram: part I. Journal of the American college of Cardiology 2007; 49(10): 1109–1127.

[3] Krizhevsky A, Sutskever I, and Hinton GE. ImageNet classification with deep convolutional neural networks. Communications of the ACM 2017; 60(6): 84–90.

[4] Hinton G, Deng L, Yu D, Dahl G, Mohamed A, Jaitly N, et al. Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. IEEE Signal Processing Magazine 2012; 29(6): 82–97.

[5] Ribeiro AH, Ribeiro MH, Paixao GM, Oliveira DM, Gomes PR, Canazart JA, et al. Automatic diagnosis of the 12-lead ECG using a deep neural network. Nature Communications 2020; 11(1): 1–9.

[6] Perez Alday EA, Gu A, Shah A, Robichaux C, Wong AKI, Liu C, et al. Classification of 12-lead ECGs: the Physionet/Computing in Cardiology Challenge 2020. Physiological Measurement 2020.

[7] Reyna MA, Sadr N, Perez Alday EA, Gu A, Shah A, Robichaux C, et al. Will two do? Varying dimensions in electrocardiography: the PhysioNet/Computing in Cardiology Challenge 2021. Computing in Cardiology 2021; 48:14.

[8] He K, Zhang X, Ren S, Sun J. Delving deep into rectifiers: surpassing human-level performance on imagenet classification. In Proceedings of the IEEE International Conference on Computer Vision (ICCV). 2015; 1026–1034.

[9] Hu J, Shen L, Sun G. Squeeze-and-Excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR). 2018; 7132–7141.

[10] Cai W, Hu D. QRS complex detection using novel deep learning neural networks. IEEE Access. 2020; 8: 97082–97089.

[11] Xiao Y, Wu J, Lin Z, Zhao X. A deep learning-based multi-model ensemble method for cancer prediction. Computer Methods and Programs in Biomedicine. 2018; 153: 1–9.

Address for correspondence:

Federico Maria Muscato
Department of Electronics, Information and Bioengineering
Via Camillo Golgi, 39, 20133 Milano, Italy
federicomaria.muscato@polimi.it