Automated Diagnosis of Reduced-lead Electrocardiograms using a Shared Classifier

H.T. Jessen[1,2], R.R. van de Leur[1,3], P.A. Doevendans[1,3], R. van Es[1]

[1]Department of Cardiology, University Medical Center Utrecht, Utrecht, the Netherlands
[2]Informatics Institute, University of Amsterdam, Amsterdam, the Netherlands
[3]Netherlands Heart Institute, Utrecht, the Netherlands

Abstract

Portable ECG devices with a reduced number of leads are increasingly being used in clinical practice. As part of the PhysioNet/Computing in Cardiology Challenge 2021, this study aims to develop an algorithm for automated diagnostis of reduced-lead ECGs. We compared separate baseline classifiers for the different lead-subsets with our newly proposed shared classifier. The different models were pre-trained on a physician-annotated dataset of 269,726 12-lead ECGs. Fine-tuning was done on the challenge dataset, consisting of 87,746 ECGs. Even though different models showed promising results on the internal pre-training dataset, optimal scores were achieved by the baseline model on the hidden validation set. Our classifiers received scores of 0.559, 0.540, 0.534, 0.543, and 0.530 (ranked 91th, 88th, 99th, 90th, and 92th out of 252 teams) for the 12-lead, 6-lead, 4-lead, 3-lead, and 2-lead versions of the hidden validation set.

1. Introduction

The 12-lead electrocardiogram (ECG) is an essential diagnostic tool in clinical practice. Considering the increasing popularity of portable ECG devices, current ECG classifiers need to be adjusted for reduced lead sets. The Physionet/Computing in Cardiology Challenge (CinC) 2021 [1] aimed to address this problem by providing a large dataset of publicly available 12-, 6-, 4-, 3- and 2-lead ECGs, in which open-source algorithms were compared on a hidden test set. Recently, exponentially dilated causal convolutions were proposed for analysis of 12-lead ECGs, as they take the temporal nature of the ECG into account and are able to learn long-range dependencies [2]. In this study, we propose that this architecture could be enriched with a shared latent space for the different reduced lead sets, to minimize the difference in performance.

2. Methods

2.1. Data

The training data consisted of 88.243 12-lead ECGs from 5 different hospitals in different countries [3–8]. Signal lengths ranged from 5 seconds to 30 minutes and sampling frequencies from 257 Hz to 1000 Hz. Each ECG was resampled to 500 Hz and only the first 10 seconds were used. ECGs shorter than 10 seconds were zero-padded on the right. Each ECG was annotated with one or more SNOMED-CT codes, and the 30 most prevalent were used to calculate the challenge metric. Four pairs of diagnoses were considered equivalent, if either of them was present the class was said to be active. This resulted in a total of 26 distinct classes. The goal of the challenge is to classify these diagnoses on different lead subsets. The following lead subsets were used:

- Twelve leads: I, II, III, aVR, aVL, aVF, V1-V6
- Six leads: I, II, III, aVR, aVL, aVF
- Four leads: I, II, III, V2
- Three leads: I, II, V2
- Two leads: I, II

As the training dataset only consists of 12-lead ECGs, when training for a specific lead subset, the corresponding leads were extracted from the 12-lead ECGs.

All models were pre-trained on an internal dataset of the University Medical Center Utrecht (UMCU). The dataset contained 287.442 10-second ECGs sampled at 500 Hz, recorded using the General Electric MAC 5500 Resting ECG acquisition and analysis system (GE Healthcare, Chicago, IL, USA). The ECGs were physician annotated as part of the standard clinical workflow. A random subset of 17716 ECGs was used for validation, the rest for pre-training. The classes with the corresponding occurrences in the pre-training- and training datasets are shown in Table1.

Diagnosis	UMCU	CinC
Atrial fibrillation	20696 (7.67%)	4222 (6.01%)
Atrial flutter	2963 (1.1%)	6734 (9.59%)
Bundle Branch Block	0 (0.00%)	405 (0.58%)
Bradycardia	21181 (7.85%)	235 (0.33%)
LBBB	7757 (2.88%)	1191 (1.7%)
RBBB	17614 (6.53%)	3858 (5.50%)
1st degree av block	11503 (4.26%)	2838 (4.04%)
IRBBB	17614 (6.53%)	1505 (2.14%)
Left axis deviation	21599 (8.01%)	6125 (8.73%)
LAFB	7672 (2.84%)	1756 (2.5%)
Low qrs voltages	8408 (3.12%)	1276 (1.82%)
NICD	7613 (2.82%)	1420 (2.02%)
Sinus rhythm	190375 (70.58%)	23023 (32.8%)
PAC	12824 (4.75%)	2617 (3.73%)
Pacing rhythm	4341 (1.61%)	1158 (1.65%)
PRWP	6212 (2.3%)	519 (0.74%)
PVC	11237 (4.17%)	1543 (2.2%)
Prolonged pr interval	11503 (4.26%)	323 (0.46%)
Prolonged qt interval	8670 (3.21%)	1496 (2.13%)
Qwave abnormal	27265 (10.11%)	1662 (2.37%)
Right axis deviation	2863 (1.06%)	1031 (1.47%)
Sinus arrhythmia	8372 (3.10%)	3068 (4.37%)
Sinus bradycardia	19808 (7.34%)	15195 (21.65%)
Sinus tachycardia	25835 (9.58%)	7692 (10.96%)
T wave abnormal	64224 (23.81%)	9349 (13.32%)
T wave inversion	4817 (1.79%)	3194 (4.55%)

Table 1. Number of ECGs with the corresponding percentage per class in both the UMCU and the CinC training datasets. LBBB: left bundle branch block, RBBB: right bundle branch block, IRBBB, incomplete right bundle branch block, NICD: nonspecific intraventricular conduction delay, PAC: premature atrial complex, PRWP: poor r wave progression, PVC: premature ventricular complex.

2.2. Approaches

The chosen approach is an extension on the work of [2], proposed in the CinC challenge 2020. The core model is the exponentially dilated causal convolutional neural network, consisting of a convolution backbone followed by an Adaptive Pooling layer to combine the features over the temporal dimension, after which a linear layer created the final output. The sigmoid function is used to convert each of the 26 outputs between 0 and 1. All methods use the focal loss [9] during training, which is an extension of the cross-entropy loss to focus on the wrongly predicted samples. To correct for the class imbalance, each class is weighted by dividing the maximum number of positive samples from any class by the number of positive samples from the weighted class by the. To make the actual predictions, a threshold of 0.5 was used, all values above were set as positive, all others were set as negative.

2.2.1. Baseline

The exact architecture as proposed by [2] was used as a baseline, with only the number of input channels altered to allow classification on the different reduced-lead subsets. This approach has 5 completely separate classifiers, one for each subset. No additional tuning of model hyperparameters was performed. Next, the baseline was extended by adding the age and sex of the patient to the latent representation before the final linear layer. The age is added directly (a default of 60 was used in case of missing values), for sex the values 0 and 1 were used to encode female and male respectively (with 1 being the default for missing values).

2.2.2. Shared classifier

In contrast to the baseline where 5 separate models are used, we propose the use of a shared classifier. In this setting, each of the reduced-lead subsets had a separate convolutional backbone which produces an latent representation of the ECG. This latent embedding is subsequently entered into a multilayer percepton model to predict the diagnostic classes. This part of the model shares its weights with the all the corresponding reduced lead sets of the same ECG during training. When the algorithm is deployed, it will also work when only a single lead set is entered. The number of linear layers was manually tuned using short runs on a subset of the data, which showed that a 3-layer classifier outperformed the single layer classifier. It is expected that the shared classifier results in a reduced performance on the 12-lead, and possibly 6-lead, ECG due to the sharing of the classifier. However, this reduced performance on the higher order subsets should come with a corresponding increase in performance on the lower order subsets.

To overcome the possible reduced performance of the higher-lead subsets, we also experiment with the use of a pair-wise shared classifier. Instead of having a single shared classifier, each of the reduced-lead subsets had a shared classifier with the 12-lead backbone. For the 12-lead ECG, the baseline 12-lead model was used, as the shared classifiers all decreased the 12-lead CinC challenge metric.

The shared classifier was extended to also use a shared latent space between the different backbones. Therefore, the Euclidean distance between the latent representation after the backbone for each of reduced-lead subsets and the 12-lead representation were tracked during training, which were all around 10 and did not decrease over time. The average of these distances was added to the loss function with a small weight (0.0001, 0.001 and 0.01 were tried). This additional loss term did decrease the average euclidean distance, but also decreased performance for all of the lead

subsets so much that it was decided not to use this additional loss term in the final version.

2.3. Training and Inference

For pre-training and training, the same procedures were followed, the difference being the dataset used. The Adam optimizer [10] was used with a batch size of 64. For both pre-training and training, early stopping was used after five consecutive epochs without the CinC challenge metric increasing on the validation dataset. The model with the highest validation score was stored and used as the final model. Early stopping was used due to time restrictions, but this made the training procedure unstable. There were cases were the early stopping was reached after 6 epochs, whereas in a next training run with the exact same settings it took 19 epochs to converge, resulting in a higher challenge score.

The shared classifier was trained gradually where at each training step one of the reduced-lead classifiers was trained. Predictions were made and evaluated for a lead subset, the loss was backpropagated and used to update the parameters of the backbone and the shared classifier. At the next step, the next lead subset would be used.

The learning rate 0f 0.001 and the freezing of the first 5 convolutional blocks of the Causal CNN after pre-training were selected based on the results from [2].

3. Results

The CinC challenge scores for the different lead subsets after pre-training and fine-tuning are shown in Tables 2 and 4 respectively. Overall, the CinC challenge scores are the highest on the UMCU validation set. The baseline outperformed the other models, after which the baseline was pre-trained longer until 10 consecutive epochs without improvement on the validation set, these results are shown as Baseline and Baseline 10. The optimal model was the baseline with the longer pre-training and the final results of that model are shown in table 5.

Model	12	6	4	3	2
Baseline	0.547	0.550	0.543	0.582	0.528
Baseline 10	0.670	0.587	0.658	0.641	0.599
Baseline + age and sex	0.632	0.588	0.641	0.630	0.573
Shared classifier	0.451	0.452	0.616	0.416	0.481
Pair-wise shared classifier	x	0.477	0.534	0.512	0.477

Table 2. Validation CinC scores on UMCU dataset after pre-training.

Method	12	6	4	3	2
Baseline	0.965	0.938	0.963	0.957	0.947
Baseline 10	0.967	0.951	0.964	0.966	0.950
Baseline + age and sex	0.959	0.951	0.958	0.965	0.950
Shared classifier	0.950	0.911	0.941	0.936	0.907
Pair-wise shared classifier	x	x	x	x	x

Table 3. Validation AUC scores on UMCU dataset after pre-training.

Model	12	6	4	3	2
Baseline	0.554	0.526	0.534	0.538	0.527
Baseline 10	0.559	0.540	0.534	0.543	0.530
Baseline + age and sex	0.539	0.537	0.532	0.540	0.527
Shared classifier	0.409	0.432	0.478	0.450	0.434
Pair-wise shared classifier	x	x	x	x	x

Table 4. Validation CinC scores on CinC dataset after fine-tuning.

Leads	Training	Validation	Test	Ranking
12	0.670	0.559	x	91
6	0.587	0.540	x	88
4	0.658	0.534	x	99
3	0.641	0.543	x	90
2	0.599	0.530	x	92

Table 5. Challenge scores for our final selected entry (team UMCU) on the validation set of the UMCU dataset after pre-training, repeated scoring on the hidden validation set, and one-time scoring on the hidden test set as well as the ranking on the hidden test set.

4. Discussion and Conclusions

In the present study we investigated whether "Two will do it" to identify 26 clinical diagnoses from reduced lead ECGs. The results from our study indicate that reducing the number of ECG leads from 12 to 2 only slightly affects CinC challenge classification performance. The baseline model with longer pre-training therefore outperformed all other methods (4). Surprisingly there is a very limited degradation in performance with a reduction of used leads. This could be due to the inherent nature of the ECG, a single lead captures electrical activity from the whole heart. Especially for rhythm related disorders it makes sense that classifying does not get more difficult with fewer leads, as the rhythm of the ECG is captured in each of the leads. Another possibility is that this lack of reduced performance is due to the way the CinC challenge score is calculated. It might be that an increasing number of leads makes the

model more certain of its prediction, the challenge score however only looks at the binary predictions without taking the certainty into consideration.

This study was the first to use a shared classifier to improve performance on reduced-lead ECGs by exploiting the additional information from the corresponding 12-lead ECG. Even though the absolute scores resulting from the shared classifier are lower compared to the baseline, we see that all the reduced-lead subsets score higher than the 12-lead ECG, with the 4- and 6-lead ECGs scoring the highest (4). A limitation of the current approach of training 5 backbones together with the shared classifier, is that all final 6 models were selected based on when the average validation loss over the different subsets was optimal. When looking at the validation curves, we saw that the different subsets achieved the best scores after different epochs. In retrospect, the choice could have been made to keep the training process the same, but store the backbone and the state of the shared classifier separately when each of the subsets achieve the highest validation set. This would only work for fine-tuning the model, as otherwise there would be different versions of the shared classifier after pre-training.

In all cases, the CinC challenge metric is higher on the UMCU validation set after pre-training compared to the validation scores after finetuning on the CinC dataset. This could be due to the larger training dataset for the UMCU dataset or the distribution of the scored classes. When looking at the baseline extended with the age and sex, we see that this improves the scores quite a bit on the UMCU dataset, whereas this is not the case on the CinC dataset. This could be due to the fact that most age and sex information is already encoded in the ECG itself.[11]

Not only the varying scores, with reduced-lead subsets scoring higher than the original 12-lead in some settings, but also the varying number of epochs during training indicate that the time limitation has a strong effect on the final performance. When the baseline was pre-trained longer, the challenge score increased for all the subsets.

As a possible improvement on the shared classifier, we experimented with using the pair-wise shared classifier, however, performance is not improved a lot compared with the shared classifier. Also, the pair-wise shared classifiers require multiple training runs, greatly increasing training time.

Overall, we can state that the classification performance of ECG disorders measured by the CinC challenge metric only reduces slightly when decreasing the number of leads. The current baseline architecture with exponentially dilated causal convolutions outperforms our shared classifier, although the later does show the expected behaviour of better performance for the reduced-leads compared to the original 12-lead ECG.

References

[1] 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;48:1–4.

[2] Bos MN, van de Leur RR, Vranken JF, Gupta DK, van der Harst P, Doevendans PA, et al. Automated comprehensive interpretation of 12-lead electrocardiograms using pretrained exponentially dilated causal convolutional neural networks. Computing in Cardiology 2020 2020;1–4.

[3] Liu F, Liu C, Zhao L, Zhang X, Wu X, Xu X, et al. An Open Access Database for Evaluating the Algorithms of Electrocardiogram Rhythm and Morphology Abnormality Detection. Journal of Medical Imaging and Health Informatics 2018;8(7):1368—1373.

[4] Tihonenko V, Khaustov A, Ivanov S, Rivin A, Yakushenko E. St Petersburg INCART 12-lead Arrhythmia Database. PhysioBank PhysioToolkit and PhysioNet 2008;Doi: 10.13026/C2V88N.

[5] Bousseljot R, Kreiseler D, Schnabel A. Nutzung der EKG-Signaldatenbank CARDIODAT der PTB über das Internet. Biomedizinische Technik 1995;40(S1):317–318.

[6] Wagner P, Strodthoff N, Bousseljot RD, Kreiseler D, Lunze FI, Samek W, et al. PTB-XL, a Large Publicly Available Electrocardiography Dataset. Scientific Data 2020;7(1):1–15.

[7] Zheng J, Cui H, Struppa D, Zhang J, Yacoub SM, El-Askary H, et al. Optimal Multi-Stage Arrhythmia Classification Approach. Scientific Data 2020;10(2898):1–17.

[8] Zheng J, Zhang J, Danioko S, Yao H, Guo H, Rakovski C. A 12-lead Electrocardiogram Database for Arrhythmia Research Covering More Than 10,000 Patients. Scientific Data 2020;7(48):1–8.

[9] Lin TY, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection, 2017. URL http://arxiv.org/abs/1708.02002. Cite arxiv:1708.02002.

[10] Kingma DP, Ba J. Adam: A method for stochastic optimization, 2015. URL http://arxiv.org/abs/1412.6980.

[11] Attia ZI, Friedman PA, Noseworthy PA, Lopez-Jimenez F, Ladewig DJ, Satam G, et al. Age and Sex Estimation Using Artificial Intelligence From Standard 12-Lead ECGs. Circulation Arrhythmia and Electrophysiology 2019;12(9):e007284. ISSN 1941-3149.

Address for correspondence:

Rene van Es
Heidelberglaan 100, 3584 CX Utrecht
r.vanes@umcutrecht.nl