Automated Comprehensive Interpretation of 12-lead Electrocardiograms Using Pre-trained Exponentially Dilated Causal Convolutional Neural Networks

Max N Bos[1,2]*, Rutger R van de Leur[1,3]*, Jeroen F Vranken[1,2], Deepak K Gupta[2],
Pim van der Harst[1], Pieter A Doevendans[1,3,4], René van Es[1]

[1] University Medical Center Utrecht, Utrecht, The Netherlands
[2] Informatic Institute, University of Amsterdam, Amsterdam, The Netherlands
[3] Netherlands Heart Institute, Utrecht, The Netherlands
[4] Central Military Hospital, Utrecht, The Netherlands
$*$ These authors contributed equally.

Abstract

Correct interpretation of the electrocardiogram (ECG) is critical for the diagnosis of many cardiac diseases, and current computerized algorithms are not accurate enough to provide automated comprehensive interpretation of the ECG. This study aimed to develop and validate the use of a pre-trained exponentially dilated causal convolutional neural network for interpretation of the ECG as part of the 2020 Physionet/Computing in Cardiology Challenge. The network was pre-trained on a physician-annotated dataset of 254,044 12-lead ECGs. The weights of the pre-trained network were partially frozen, and the others were finetuned on the challenge dataset of 42,511 ECGs. 10-fold cross-validation was applied and the best performing model in each fold was selected and used to construct an ensemble. The proposed method yielded a cross-validated area under the receiver operating curve (AUROC) of 0.939 ± 0.004 *and a challenge score of* 0.565 ± 0.005. *Evaluation on the hidden test set resulted in a score of* 0.417*, placing us 7th out of 41 in the official ranking (team name UMCUVA). We demonstrated that an ensemble of exponentially dilated causal convolutional networks and pre-training on a large dataset of ECGs from a different country and device manufacturer performs excellent for interpretation of ECGs.*

1. Introduction

The 12-lead electrocardiogram (ECG) is a fundamental diagnostic tool in clinical practice and used to diagnose a wide range of possibly life-threatening cardiac abnormalities. Automated diagnosis of the ECG could be of great support in clinical practice, especially when expert knowledge is not readily available.

Conventional computerized algorithms that use handcrafted features have not been able to reach sufficient accuracy for automated comprehensive ECG interpretation and overreading by a physician remains necessary [1]. Recent studies have shown promising results using deep neural networks (DNNs) for comprehensive automated ECG interpretation [2–4]. However, these approaches use proprietary datasets or limited publicly available datasets, provide no comparison with existing architectures and only classify a small selection of diagnoses. The PhysioNet/Computing in Cardiology Challenge (CinC) 2020 aimed to address this problem by providing a large dataset of publicly available ECGs, in which participants submitted open-source algorithms which were compared on a hidden test set [5].

Different DNN architectures have been proposed for the automated interpretation of ECGs, but exponentially dilated causal convolutions have never been used [4]. These convolutions have the advantage that they take the temporal nature of the ECG into account, while efficiently learning long-range dependencies in time series [6]. Furthermore, other works have discussed the prospects of using large task-related ECG datasets to pre-train ECG classifiers for transfer learning [7]. In this study, we propose a combination of transfer learning and an exponentially dilated causal convolutional network for automated comprehensive interpretation of the ECG.

2. Methods

2.1. Challenge Data

The training and hidden testing datasets combined raw ECG data from five different sources in China, Russia, Germany, and the United States and are described in detail in the the PhysioNet/CinC 2020 challenge paper [5]. The training dataset consisted of 42,511 12-lead ECGs, of which 32,181 ECGs were 10-second length, while the remaining 10,330 ECGs were varying between 6 and 60 seconds. For the latter, only the first 10 seconds were extracted for ECGs with a duration greater than 10 seconds. ECGs with a length shorter than 10 seconds were zero-padded per training batch. All ECGs were resampled to 500 Hz using linear interpolation. All ECGs were interpreted and annotated using one or more of the 111 SNOMED-CT codes for ECGs. For the challenge only the 27 most prevalent codes were used, of which three pairs regarding right bundle branch block, premature atrial contraction, and premature ventricular contractions, were considered equivalent. Therefore, 24 classes were used for training.

2.2. Pre-training Data

We investigated the application of transfer learning using a dataset acquired by the University Medical Center Utrecht (UMCU). The dataset consisted of 254,044 500 Hz-sampled 10-second 12-lead ECGs from patients between 18 and 85 years old, recorded using the General Electric MAC 5500 Resting ECG acquisition and analysis system (GE Healthcare, Chicago, IL, USA). All ECGs were systematically annotated by a physician as part of the regular clinical workflow. These free text annotations were structured using a text mining algorithm as described before and mapped to the 24 classes in the challenge dataset [3]. We constructed a randomly sampled UMCU test dataset of 16,698 ECGs to evaluate the resulting DNN during the pre-training process. Table 1 lists the 24 classes and the corresponding number of ECGs present in the CinC and UMCU training dataset.

2.3. Model Architecture

We constructed a deep convolutional neural network with 1-dimensional exponentially dilated causal convolutions (Causal CNN, Figure 1). Based on the method described by Franceschi et al. [8], we built an architecture composed of several causal convolutional blocks, transforming the $12 \times L$-sized ECG data to 216 L-dimensional feature maps, where each point in a feature map is based on a history of 383 sample points, including itself. Subsequently, we employ a 1-dimensional adaptive max pooling layer to squeeze the temporal dimension resulting in a 216-dimensional representation, followed by a 216-to-24 linear layer with sigmoid activation to allow for multi-label classification. Each causal convolution block consists of a combination of causal convolutions, weight normalizations, leaky ReLUs and residual connections. The causal convolution is a result of first applying a convolution and thereafter truncating the output, to remove future timepoints. The residual connection is only used when up-

Diagnosis	UMCU dataset	CinC dataset
1st degree AV block	10560 (2.10%)	2394 (3.98%)
atrial fibrillation	19229 (3.82%)	3485 (5.74%)
atrial flutter	2714 (0.54%)	313 (0.52%)
bradycardia	19573 (3.89%)	277 (0.46%)
complete RBBB	15854 (3.15%)	3069 (5.10%)
incomplete RBBB	15854 (3.15%)	1611 (2.68%)
LAFB	7237 (1.44%)	1806 (3.00%)
left axis deviation	20374 (4.05%)	6086 (10.11%)
LBBB	7322 (1.46%)	1041 (1.73%)
low QRS voltages	7950 (1.58%)	556 (0.92%)
NICD	7189 (1.43%)	996 (1.65%)
pacing rhythm	4211 (0.84%)	299 (0.50%)
PAC	11857 (2.36%)	1935 (3.21%)
PVC	10441 (2.08%)	551 (0.91%)
prolonged PR interval	10560 (2.10%)	340 (0.56%)
prolonged QT interval	8323 (1.65%)	1513 (2.51%)
Q-wave abnormal	25969 (5.16%)	1013 (1.68%)
right axis deviation	2739 (0.54%)	427 (0.71%)
sinus arrhythmia	7660 (1.52%)	1238 (2.06%)
sinus bradycardia	18328 (3.64%)	2359 (3.92%)
sinus rhythm	179786 (35.73%)	20766 (34.48%)
sinus tachycardia	24440 (4.86%)	2390 (3.97%)
T-wave abnormal	60321 (11.99%)	4673 (7.76%)
T-wave inversion	4661 (0.93%)	1111 (1.84%)

Table 1. Number of ECGs present for each class in the constructed UMCU and CinC training datasets. A single ECG can belong to multiple classes. AV: atrioventricular, RBBB: right bundle branch block, LAFB: left anterior fascicular block, LBBB: left bundle branch block, NICD: nonspecific intraventricular conduction delay, PAC: premature atrial complex, PVC: premature ventricular complex.

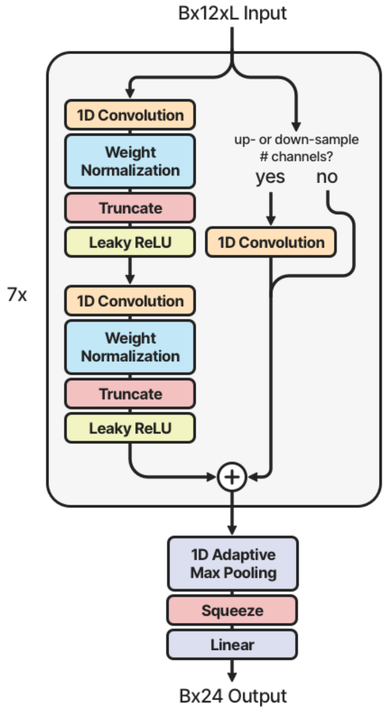

Figure 1. Illustration of the model architecture. B: batch size, L: waveform input length.

sampling the number of input channels. The dilation parameter used in the causal convolutional layer is doubled in each subsequent causal convolution block from 1 to 64. The first convolutional layer transforms 12 input channels to 108 output channels and thereafter the number of channels is kept constant at 108 for the first six consecutive causal convolution blocks. The seventh causal convolution block used 216 output channels in the causal convolution layers. All causal convolutions used a kernel size of 3. The kernel size and the number of causal convolution blocks were selected such that the resulting receptive field would be sufficient to capture the normal duration of a cardiac cycle from P-wave onset to T-wave offset, namely 383 sample points or an equivalent duration of $383 \cdot 2ms = 766ms$ at 500 Hz. As a result of employing causal convolutions and subsequently max pooling, the length of the input can be variable.

2.4. Model Training & Inference

We first trained the architecture for multi-label classification of the 24 classes using the UMCU training dataset that is described in Section 2.2. We applied 10-fold iterative stratification for multi-label data [9] to obtain 10 differing training and test splits of the challenge dataset described in Section 2.1. The parameters in some of the first causal convolution blocks of the final pre-trained model were frozen and only the remaining parameters of the model were trained further. This resulted in a set of 10 models, which were used during inference to obtain final probability scores by computing the mean over the individual model probability outputs. All ECGs for inference were resampled to 500 Hz using linear interpolation and the full length (up to 10 seconds) was used.

We optimized the network parameters of the architecture using a weighted focal loss function to handle class imbalance, and Adam as the optimization algorithm [10, 11]. We assigned a weight to the positive samples of each class equal to the class imbalance ratio to force equal attribution to the loss of both class samples. These weights were multiplied by a factor between 0 and 1, to be able to optimize the influence of weighting between no weighting and full weighting. The used batch size was 128. Early stopping was performed when the CinC challenge metric score on the corresponding evaluation dataset had not increased for five consecutive epochs.

Hyperparameters of the model were selected using manual tuning. We assessed pre-training and different values for the learning rate, number of frozen blocks, gamma value of the focal loss and the weighting factor. The model with the highest mean challenge metric over the cross-validation results was chosen. The final model used pre-training with 5 frozen blocks, a learning rate of 0.001, gamma of 2 and weighting factor of 0.5. All model training processes, and the architecture as described in Section 2.3 were implemented using the PyTorch package (version 1.3).

Hyperparameter			**Cross-validation**
pre-train	**frozen**	**lr**	**CinC metric**
no	0	0.001	0.542 ± 0.008
yes	5	0.001	0.565 ± 0.005
yes	6	0.0001	0.546 ± 0.007
yes	7	0.0001	0.499 ± 0.006

Table 2. 10-fold cross-validation performances on the CinC dataset of Causal CNN variants using different hyperparameter value combinations. The hyperparameters *pre-train*, *frozen*, and *lr* indicate whether the model was pre-trained on the UMCU dataset, the number of consecutive causal convolutional blocks for which the parameters were frozen, and the learning rate, respectively.

2.5. Statistical analysis

Overall algorithm discriminatory performance was assessed using the macro-averaged area under the receiver operating curve (AUROC), class-specific AUROCs, and the CinC challenge metric [5]. Cross-validation performance results are presented with the standard deviation. All statistical analyses were performed using Python (version 3.7).

3. Results

The obtained cross-validation performance results of the Causal CNN with pre-training and without pre-training and various hyperparameter values are presented in Table 2. The best approach with five frozen blocks had an AUROC cross-validation score of 0.939 ± 0.004. The final AUROC and CinC challenge metric of this approach on the hidden CinC test set were 0.915 and 0.417, respectively, archieving place 7 out of 41 in the official ranking (team name UMCUVA). The challenge metric ranged from 0.298 to 0.643 across the four different hidden test sets. The AUROC ranged from 0.751 to 0.992 for the different diagnoses, with the worst performance for T-wave inversions, T-wave abnormalities and premature ventricular complexes and the best performance for sinus tachycardia, pacing and sinus bradycardia.

4. Discussion

In this study we demonstrated that an ensemble of exponentially dilated causal convolutional networks performs excellent for comprehensive multi-label interpretation of 12-lead ECGs. In addition, we showed that pre-training on a large dataset of ECGs is feasible and improves accuracy when trained on a relatively small dataset.

This is the first study to use exponentially dilated causal convolutional networks for an ECG classification task. Other works on ECGs have examined the usage of CNN-based architectures that are developed for classification of images [2–4]. However, the application of exponentially dilated causal convolutions is more sensible for time series, such as ECGs, as they take the temporal nature of the data into account and use increasing receptive fields from which ECG features could be extracted [8]. Earlier studies showed that this technique outperforms recurrent neural networks, another technique that allows for increasing receptive fields, in terms of both efficiency and prediction performance [6]. Furthermore, Strodthoff et al. [7] have shown that transfer learning can be applied to improve ECG classifiers. Likewise, we have demonstrated that transfer learning, using the data available at the UMCU, to a dataset recorded using a different ECG device and acquired from an ethnically and geographically different population, can be effective for improving ECG classification.

This study has several limitations to address. Although our proposed neural network allows for ECG input of variable length, this study only used ECGs of maximum 10 seconds. It may be interesting to further investigate the application of training on ECGs of variable duration. Moreover, instead of zero-padding samples shorter than 10 seconds, one may consider constructing mini-batches of ECGs with similar lengths. Furthermore, we only assessed the focal loss method with different gamma values and weighting factors to account for the severe class imbalance in this dataset. However, discriminatory performance is still worse in the smaller classes. Overall performance might be improved by implementing other methods to combat class imbalance, and by including ECGs of differing sampling rates to the training data.

The PhysioNet/CinC 2020 dataset is a step towards development of deep-learning based automated comprehensive ECG interpretation algorithms. Unfortunately, the dataset included too few ECGs for the majority of the 111 available classes, and attention was primarly towards the 24 classes evaluated in this study. Future studies should focus on gathering sufficient data of rare ECG diagnoses to provide completely comprehensive ECG interpretations.

5. Conclusion

A combination of transfer learning and exponentially dilated causal convolutions shows excellent performance for comprehensive interpretation of 12-lead ECGs. Our algorithm had a CinC challenge metric score of 0.417 and achieved place 7 out of 41 in the official ranking (team name UMCUVA).

Acknowledgments

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research. We have no conflicts of interest to disclose.

References

[1] Schläpfer J, Wellens HJ. Computer-Interpreted Electrocardiograms. Journal of the American College of Cardiology 2017;70(9):1183–1192.

[2] Ribeiro AH, Ribeiro MH, Paixão GM, Oliveira DM, Gomes PR, Canazart JA, Ferreira MP, Andersson CR, Macfarlane PW, Wagner M, Schön TB, Ribeiro ALP. Automatic Diagnosis of the 12-lead ECG using a Deep Neural Network. Nature Communications 2020;11(1):1760.

[3] Van de Leur RR, Blom LJ, Gavves E, Hof IE, Van der Heijden JF, Clappers NC, Doevendans PA, Hassink RJ, Van Es R. Automatic Triage of 12-Lead ECGs Using Deep Convolutional Neural Networks. Journal of the American Heart Association 2020;9(10):e015138.

[4] Hong S, Zhou Y, Shang J, Xiao C, Sun J. Opportunities and Challenges of Deep Learning Methods for Electrocardiogram Data: A Systematic Review. Computers in Biology and Medicine 2020;122(6):103801.

[5] Perez Alday EA, Gu A, Shah A, Robichaux C, Wong AKI, Liu C, Liu F, Rad AB, Elola A, Seyedi S, Li Q, Sharma A, Clifford GD, Reyna MA. Classification of 12-lead ECGs: the PhysioNet/Computing in Cardiology Challenge 2020. Physiological measurement 2020;In Press.

[6] Bai S, Zico Kolter J, Koltun V. An Empirical Evaluation of Generic Convolutional and Recurrent Networks for Sequence Modeling. arXiv e prints 2018;arXiv:1803.01271.

[7] Strodthoff N, Wagner P, Schaeffter T, Samek W. Deep Learning for ECG Analysis: Benchmarks and Insights from PTB-XL. IEEE Journal of Biomedical and Health Informatics 2020;1–1.

[8] Franceschi JY, Dieuleveut A, Jaggi M. Unsupervised Scalable Representation Learning for Multivariate Time Series. In Advances in Neural Information Processing Systems. 2019; 4652–4663.

[9] Sechidis K, Tsoumakas G, Vlahavas I. On the Stratification of Multi-Label Data. Machine Learning and Knowledge Discovery in Databases 2011;145–158.

[10] Lin T, Goyal P, Girshick R, He K, Dollár P. Focal loss for dense object detection. In 2017 IEEE International Conference on Computer Vision (ICCV). 2017; 2999–3007.

[11] Kingma DP, Ba J. Adam: A Method for Stochastic Optimization. arXiv e prints 2014;arXiv:1412.6980.

Address for correspondence:

René van Es, PhD
Heidelberglaan 100
PO Box 85500, 3508 GA, Utrecht, The Netherlands
r.vanes-2@umcutrecht.nl