Classification of 12-Lead Electrocardiograms Using Residual Neural Networks and Transfer Learning

Sardar Ansari[1,†], Christopher E Gillies[1,†], Brandon Cummings[1,†], Jonathan Motyka[1,†], Guan Wang[1,†], Kevin R Ward[1,†], Hamid Ghanbari[1,†]

[1]University of Michigan, Ann Arbor, MI, USA
[†]Michigan Center for Integrative Research in Critical Care, Ann Arbor, MI, USA

Abstract

This article concerns the PhysioNet/Computing in Cardiology Challenge 2020 which focused on building computational methods to identify cardiac abnormalities from 12-lead ECGs. Our team, MCIRCC, utilized a large secondary dataset of 12-lead ECGs obtained from the Section of Electrophysiology at the University of Michigan, called the MUSE dataset, to pre-train multiple residual neural networks that were later re-trained on the challenge dataset. To do so, the diagnosis statements that existed in our dataset were utilized to assign the same labels to our ECGs as the challenge data. After parameter optimization, we selected a subset of top performing models and created an ensemble model that achieved a challenge validation score of 0.616*, and full test score of* 0.141*, placing us* 27^{th} *out of 41 teams in the official ranking.*

1. Introduction

The electrocardiogram (ECG) is widely used for the diagnosis and monitoring of various cardiovascular diseases and cardiac abnormalities [1]. However, manual interpretation of ECG recordings is laborious and requires inspection by trained clinical personnel [2]. Machine learning models may enable automatic classification of cardiac abnormalities and reduce interpretation time and healthcare costs. In 2020, the PhysioNet and Computing in Cardiology held a challenge to tackle this problem by improving the performance of automatic interpretation algorithms for 12-Lead ECGs [3][1]. Our challenge entry utilized a large cohort of 12-lead ECGs from the University of Michigan to pre-train a residual neural network and re-train it on the challenge data to fine-tune the model, as described below.

[1]A preprint of this article can be found here: https://www.medrxiv.org/content/10.1101/2020.08.11.20172601v1

2. Methods

2.1. MUSE Dataset

The approach that was utilized in this study involved pre-training of residual networks using a large dataset of 12-lead ECGs obtained from the Section of Electrophysiology at the University of Michigan. The dataset contained 1,277,298 records collected from 374,321 patients from 1990 to 2012 (with a few exceptions). Each recording was 10 seconds long and was sampled at 250Hz or 500Hz. All ECGs were resampled at 250Hz before processing. Each ECG was first analyzed and labeled by the MUSE software at the time of recording. The automatically generated diagnoses were then reviewed and corrected (if necessary) by a cardiologist as part of routine clinical care.

These labels existed in the dataset in a semi-structured format, i.e., majority of the labels were broken down into separate segments, each segment often representing a single or a combination of arrhythmias and diagnoses. Our team leveraged these labels and used prepositions, conjunctions and adjectives such as *and*, *with*, *likely* and *frequent* to further break down each segment. The resulting segments were then filtered to only include phrases with at least 50 instances. Each selected phrase was searched in the Unified Medical Language System (UMLS) metathesaurus [2] to find the correposnding Concept Unique Identifiers (CUI). The CUIs with SNOMED CT (SCT) code mappings were selected from the search results. If any of the SCT codes existed in the list of the challenge labels, they were selected; otherwise, the SCT graph was traversed up (towards the root) to find matches between the ancestors and the challenge labels. This allowed us to match diagnoses in the MUSE and challenge datasets that had different levels of granularity.

The resulting mapping between the diagnosis statements from the MUSE dataset and the challenge labels was then manually inspected and corrected. The final mapping was

[2]https://uts.nlm.nih.gov/home.html

SCT Code	Abbreviation	Physionet Count	MUSE Count
270492004	IAVB	2394	76559
164889003	AF	3475	74352
164890007	AFL	314	10987
426627000	Brady	288	0
713427006	CRBBB[1]	683	64668
713426002	IRBBB	1611	33149
445118002	LAnFB	1806	32308
39732003	LAD	6086	110471
164909002	LBBB	1041	37631
251146004	LQRSV	556	57480
698252002	NSIVCB	997	6415
10370003	PR	299	28684
284470004	PAC[2]	1729	32789
427172004	PVC[3]	188	54083
164947007	LPR	340	0
111975006	LQT	1513	64728
164917005	QAb	1013	0
47665007	RAD	427	8554
59118001	RBBB[1]	2402	64668
427393009	SA	1240	62985
426177001	SB	2359	171613
426783006	NSR	20846	789961
427084000	STach	2402	140853
63593006	SVPB[2]	215	32789
164934002	TAb	4673	97719
59931005	TInv	1112	42
17338001	VPB[3]	365	54083

Table 1. The list of scored classes in the challenge and their frequency in the challenge and MUSE datasets. Classes with equal superscripts are considered identical in the challenge. The full description of classes can be found in [3].

used to label each ECG recording with the SCT codes that exist in the challenge's list of scored classes. The list of classes and the number of instances in each dataset are shown in Table 1. Some of the challenge classes were absent in the MUSE dataset, including Brady, LPR and QAb.

2.2. Classifying ECGs in the MUSE Dataset

2.2.1. Architecture

The MUSE ECGs and their labels were used to train a residual neural networks (ResNet) with various architectures. Figure 1 depicts the general structure of the network, composed of an input layer, followed by a convolutional layer and residual blocks. Each residual block was composed of a max pooling layer followed by n convolutional layers and a residual connection. The residual block was

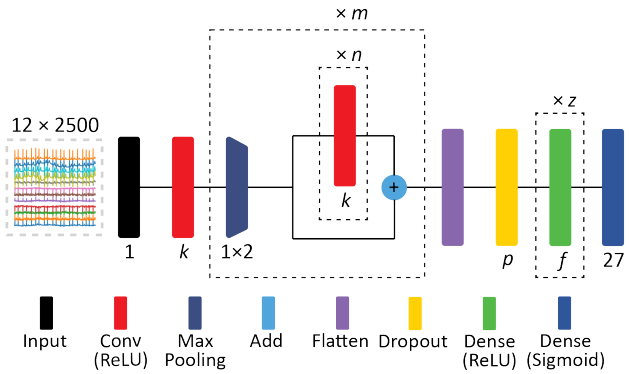

Figure 1. The architecture of the ResNet models used in this study. The models were composed of residual blocks (the outer dashed box) which included a max pooling layer (stride of 2 in the time dimension) follow by n convolutional layers (inner dashed box). The residual block was repeated m times. The other parameters that defined the architecture were the number of filters in each convolutional layer (k), the size of each filter ($12 \times s$; not shown), dropout probability (p), and number and size of dense layers (z and f, respectively).

repeated m times, followed by a flattening layer, a dropout layer with probably p and z dense layers of size f with ReLU activation function. Finally, a dense layer with sigmoid activation function was applied to the output binary classification scores. All convolutional layers had k filters of size $12 \times s$, with the first dimension spanning the 12 leads of ECG. Manual parameter selection was performed by varying n from 2 to 6, m from 4 to 10, k from 16 to 64, s from 3 to 11 with step size of 2, p from 0 to 0.5 with step size of 0.25, z from 0 to 2 and f from 32 to 128.

2.2.2. Training and Testing

The MUSE data was randomly divided into three subsets for training (60%), validation (20%) and testing (20%). The training was conducted using Tensorflow 2.3.0 and its implementation of Keras. A batch size of 128, binary cross entropy loss function and Adam optimizer were used. The training was performed for 100 epochs. The challenge metric was calculated after each epoch and training was terminated early if the metric did not improve by at least 0.01 for 3 consecutive epochs. The learning rate was reduced by a factor of 0.1 (unless it dropped below 0.0001) if challenge metric did not improve for 2 consecutive epochs. A total of 166 networks with different architectures were trained and the eight top performing models were selected and applied to the challenge data.

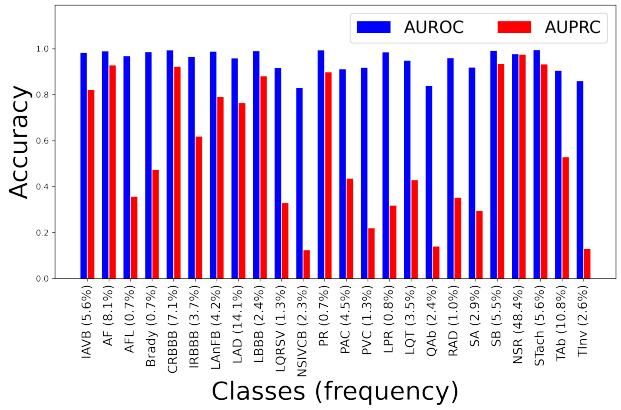

Figure 2. The area under the receiver operating characteristics curve (AUROC) and area under the precision-recall curve (AUPRC) for each class when the ensemble model was applied to the challenge public dataset. The numbers in the parentheses are the fraction of recordings with that label.

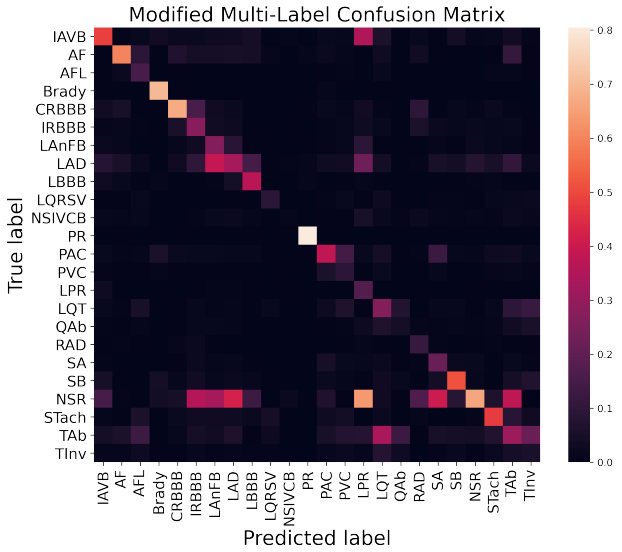

Figure 3. The modified multi-label confusion matrix as calculated by the challenge evaluation function. The matrix was formed by calculating the outer product of true and predicated label vectors for each ECG, normalizing by the total number of positive predictions and false negatives for that ECG, and adding up the resulting matrices across ECGs. Each row of the matrix was then divided by its sum.

2.3. Modeling the Challenge Data

2.3.1. Preprocessing

The ECGs from the six datasets provided by the challenge had different baseline levels and frequency compositions. To equalize the histograms of these datasets, multiple preprocessing steps were applied to each ECG recording. First, all input ECGs were resampled at 250Hz. Then, a double median filter was applied to the ECG leads to remove the baseline wonder, i.e., each ECG lead was first filtered with a median filter of length 200ms, followed by another median filter of length 600ms. The ECG leads were also low-pass filtered using a 5^{th} order Butterworth filter and a cutoff frequency of 40Hz. To address the variable length of input ECGs, we selected the first 10s of an ECG record if it was longer than 10s; otherwise, the ECG was zero padded to 10s.

2.3.2. Training and Testing

Each of the eight selected models that were trained on the MUSE data was transferred and retrained on the challenge data. For each model, the last layer (dense layer with 27 nodes and sigmoid activation function) was removed and replaced by a dropout layer (probably=0.5), followed by dense layers with 128 and 32 neurons and ReLU activation function, and a dense layer with 27 neurons and sigmoid activation function. The six datasets in the challenge data were combined and then divided into separate datasets for training (60%), validation (20%) and testing (20%). The training parameters were similar to the ones used for training on the MUSE dataset (see Section 2.2.2). None of the model weights were frozen; hence, full retraining of the weights was allowed.

After training, the validation dataset was used to find the best threshold for the class scores, by finding the value that led to the highest challenge score. The threshold was used to obtain binary classifications from the output of each model and calculate the performance on the test. The five top scoring models were then used to build an ensemble model for classification of the challenge ECGs. The class scores for the ensemble model were obtained by calculating the median of the scores generated by individual models, while the ensemble binary classifications were obtained by calculating the mode of the individual binary labels.

3. Results

The parameters for the best performing models on the MUSE dataset are shown in Table 2. The models were selected according to the challenge metric and achieved scores ranging from 0.642 to 0.731. The selected models were then modified and retrained on the challenge data, as described in Section 2.3. The resulting ensemble model was applied to the challenge public (training) data and area under the receiver operating characteristics curve (AUROC) and area under the precision-recall curve (AUPRC) were calculated for each class. The results are illustrated in

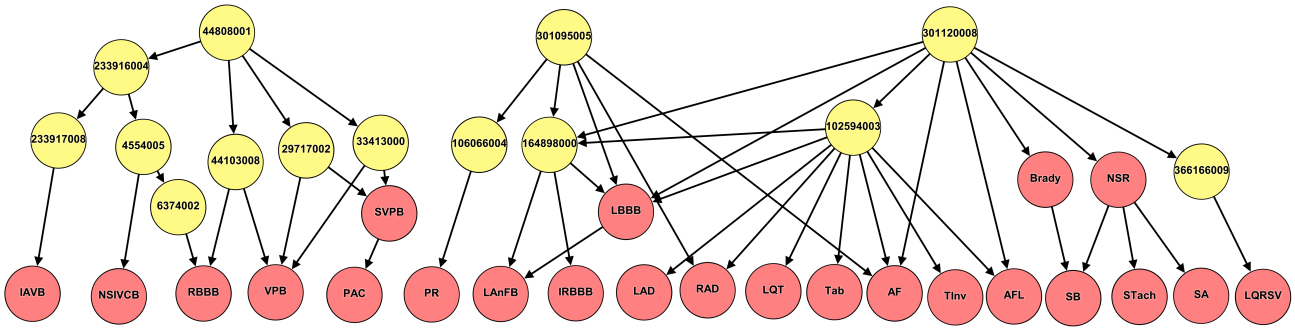

Figure 4. The SCT codes for the classes that were scored in the challenge and their parent. An arrow indicates a parent-child relationship. The nodes in red were scored in the challenge, while the ones in yellow were not and are shown to depict the graph structure. The graph does not include all parent nodes for any given child.

Figures 2. The AUROCs ranged from 0.83 to 0.99, while the AUPRCs ranged from 0.12 to 0.97. Figure 3 depicts the normalized multi-class confusion matrix as defined in [3]. Some classes such as NSIVCB, QAb and TInv were frequently misclassified, which is consistent with the results in Figure 2. The model was then scored on the challenge hidden dataset and achieved a score of 0.616 on the validation set and 0.141 on the full test set (Team MCIRCC), ranking 27th among 41 teams in the official ranking.

4. Discussion and Conclusions

The results indicate high AUROC and AUPRC for some classes (e.g., AF, CRBBB, PR, SB, NSR and STach), while other classes achieved low levels of AUPRC (e.g., NSIVCB, QAb and TInv). Notably, the classes that had no or very low representation in the MUSE dataset (Brady, LPR, QAb, TInv), were among the ones with lowest AUPRC.

We also attempted to include the patient demographics (age and gender) as inputs to the dense layers of the challenge model. This did not improve the performance of the classification models. In addition, we used the ResNet models trained on the MUSE data (excluding the final sigmoid layer) to extract features from the challenge ECGs. These features were used as inputs to train XGBoost models to classify the recordings. The results were similar to the ones reported above using dense layers. We also experimented with freezing different layers of the MUSE models before retraining on the challenge data. The best results were obtained when the entire model was retrained.

Some of the classes that were scored in this challenge had parent-child relationship with each other, as shown in Figure 4. However, the ECGs belonging to the child nodes were not consistently labeled in relation to the parents. Hence, this may have resulted in conflicting information being presented to the models, potentially diminishing the performance of the models.

Model	Parameters							Challenge Metric
	k	s	n	m	p	z	f	
1	16	7	2	8	0	0	0	0.689
2	16	7	3	8	0	0	0	0.7
3	32	7	2	6	0	0	0	0.642
4*	32	7	2	6	0	0	0	0.671
5*	32	7	2	7	0	0	0	0.687
6*	32	7	2	9	0	0	0	0.703
7*	32	7	3	9	0	0	0	0.731
8*	64	7	2	6	0	0	0	0.673

Table 2. The parameters for the eight top-scoring models **trained and tested on the MUSE dataset**. *These models achieved the highest scores after retraining on the public challenge data and were included in the ensemble model.

References

[1] Kligfield P. The centennial of the Einthoven electrocardiogram. Journal of Electrocardiology 2002;35(4):123–129.

[2] Padayachee C, Sear C, Challa P, Jenkins C, Whitman M. Can the computer tell me what's wrong with my heart? Early day lessons from digital hospital and ECG interpretation. Heart Lung and Circulation 2018;27:S303–S304.

[3] Perez Alday EA, Gu A, Shah A, Robichaux C, Wong AKI, Liu C, Liu F, Rad BA, 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; under review.

Address for correspondence:

Sardar Ansari
2800 Plymouth Road, Bldg. 10-A112, Ann Arbor, MI 48109
sardara@umich.edu