Two Will Do: CNN With Asymmetric Loss, Self-Learning Label Correction, and Hand-Crafted Features for Imbalanced Multi-Label ECG Data Classification

Cristina Gallego Vázquez[1], Alexander Breuss[1], Oriella Gnarra[1,3], Julian Portmann[2], Giulia Da Poian[1]

[1]Sensory-Motor Systems (SMS) Lab, Department of Health Sciences and Technology , ETH Zurich
[2]Department of Computer Science, ETH Zurich
[3]Sleep-Wake-Epilepsy-Center, Department of Neurology, Bern University Hospital (Inselspital)

Abstract

In this work, we present a machine learning approach that is able to classify 30 cardiac abnormalities from an arbitrary number of electrocardiogram (ECG) leads. Features extracted by a deep convolutional neural network are combined with hand-crafted features (demographic, morphological, and heart rate variability metrics) and fed into a multilayer perceptron. We employ an Asymmetric Loss (ASL) function, which enables the model to focus on hard but under-represented samples. To mitigate the issue of ground-truth mislabeling and to provide robustness, we investigate the use of a self-learning label correction method that iteratively estimates correct labels during training. Our team SMS+1 placed 7th on the unseen test set, with an overall challenge score of 0.51, and 0.52, 0.45, 0.50, 0.50, and 0.49 for twelve-, six-, four-, three-, and two-lead, respectively. Our model maintains similar diagnostic potential on both standard twelve-lead ECGs and reduced-lead ECGs.

1. Introduction

Detecting cardiac abnormalities as early as possible is a crucial task in which automated electrocardiogram signal interpretation plays an important role. In recent years, Deep Learning (DL) has been widely applied in many areas, including healthcare, and has shown high accuracy in ECG arrhythmia classification [1]. The most successful types of DL models are restricted Boltzmann machines, stacked autoencoder, Convolutional Neural Networks (CNNs), and Deep Belief Networks [2]. Compared to traditional approaches, DL-based approaches can automatically learn informative feature representations [1]. However, it can also be beneficial to incorporate expert knowledge, represented by hand-crafted features [3, 4].

All authors contributed equally to this work.

The 2020 PhysioNet/CinC Challenge focused on classifying 27 cardiac abnormalities from twelve-lead ECG, the most standard diagnosis screening system for a variety of cardiac arrhythmias [5]. This year's challenge focuses on the ability to achieve similar multi-class classification performance with a reduced set of leads, motivated by the limited accessibility of twelve-lead ECG devices. Most severe diseases occur rarely but are very important to be detected by the model, and high-data quality acquisition including expert annotations are difficult to acquire. Therefore, two of the biggest challenges when applying DL to ECG data are the imbalance and noisy nature of the labels arising from incorrectly labeled recordings [4].

2. Methods

In this work, we present a deep learning architecture for multi-label classification of 30 ECG findings. Our network combines hand-crafted features ('wide') with ECG features extracted via a neural network ('deep'). For encoding deep features, we employ a deep neural architecture built by interleaving nonlinear convolutional blocks, which allow modeling patterns at different time scales. We employed an Asymmetric Loss (ASL) function, which enables the model to focus on hard, but under-represented, samples. A final set of fully connected layers combine both the 'wide' and 'deep' features to produce multilabel classifications of ECG findings. To mitigate the issue of ground-truth mislabeling and to provide robustness, we investigate the use of a self-learning label correction method that does not require a fully correctly labeled dataset, but iteratively estimates corrected labels during training. The model architecture is illustrated in Figure 1.

2.1. Dataset

The 2021 PhysioNet/CinC Challenge datasets include annotated twelve-lead ECG recordings from six different
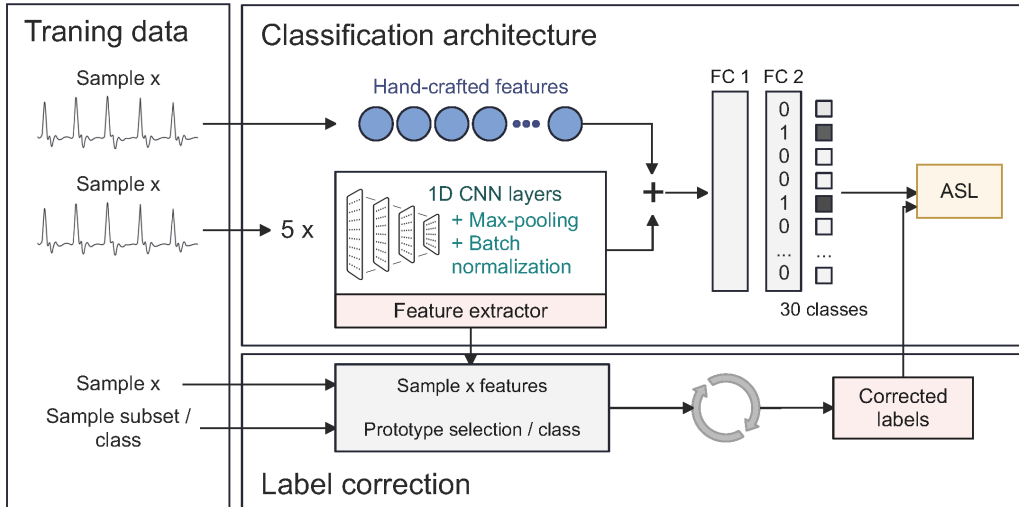
Figure 1. Architecture of the model: Training data is first resampled before it is passed to a 1D CNN and combined with hand-crafted features. The feature set obtained from the 1D CNN is also used as an input to a label correction phase that iteratively estimates corrected labels during training by identifying prototypes of every class that are likely labeled correctly. The output of the label correction is then combined with the output after the convolution step.

sources [6]. These datasets include over 100,000 twelve-lead ECG recordings with over 88,000 ECGs shared publicly as training data. An analysis of the demographic data shows a low percentage of missing data (0.27% for age followed by 0.03% for sex). Age has negative skewness of -0.74 with a mean of 59.23 (std. 18.39). Sex is balanced between males (55%) and females (45%). The durations of the recordings range from a minimum of 5 seconds to a maximum of 30 minutes, but 92% of recordings are 10 seconds long. The datasets have unbalanced classes. For example, sinus rhythm represents 22% of the labels while complete left bundle branch block appears only in 0.16% of the recordings.

2.2. Classification architecture

2.2.1. Extraction of hand-crafted ECG features

As recordings from separate hospitals and devices can have different sampling rates, we first resample each recording to 250 Hz. Then we apply an finite impulse response (FIR) bandpass filter between 3 - 45 Hz. We extract a set of features per lead, including morphological features (P and T wave amplitude, PR interval, etc.) [7], heart rate variability features in time, frequency, and non-linear domains (RMSSD, pNN60, spectral power density pertaining to low and high frequency band, T and P wave permutation and approximation entropy, etc.) [8], and a subset of features used by the 2020 PhysioNet Challenge winners [3]. The high prevalence (92%) of ten seconds long recordings

influenced the choice of the selected hand-crafted features. We also include demographic features (age and gender) as shown in Table 1. Collectively, these features are concatenated with the learned outputs from the "Deep" portion of the model (c.f. chapter 2.2.2).

Table 1. Hand-crafted features complementing the 1D-CNN used in the challenge.

	Hand-crafted features
Demographic	Age, Sex
Morphological[7]	QRS/P and QRS/T duration PR interval, BPM P and T wave amplitude
2020 PhysioNet challenge winners[3]	Heart rate, RR interval P and T wave approximate entropy (median)
Heart rate variability (HRV)[8]	HRV time domain (14 features) HRV frequency domain (9 features) HRV non linear domain (29 features)

2.2.2. 1D-CNN

The deep component of our model consists of a series of convolution operations and two fully connected feed-forward networks. The one-dimensional convolution operations are applied to the original ECG waveform segments

to extract a latent space representation of the signals. The dimension settings of the layers are listed in Table 2 . The stride and kernel of the max-pooling layers after each convolution are set to four. Batch normalization is also applied after each convolution. A one-dimensional adaptive max-pooling is performed to the output before it is combined with the hand-crafted features for the two fully convolution operations.

Table 2. Deep Learning model settings.

Layer	In	Kernel	Stride/Padding	Out
CNN 1	leads	5	1/2	16
CNN 2	16	5	1/2	32
CNN 3	32	5	1/2	64
CNN 4	64	5	1/2	128
CNN 5	128	5	1/2	256
FC 1	256 + feat			512
FC 2	512			30

2.2.3. Asymmetric loss function

A trained model with imbalanced data may make predictions with high precision and low recall, being severely biased towards the more represented classes. In medical applications, where it is important to avoid false negatives, this is an issue. We employed an Asymmetric Loss (ASL) for multi-label classification [9], which enables the model to focus on hard, but under-represented samples, and also deals with potential mislabeled samples. This loss function contains two complementary asymmetric mechanisms that work differently on well-represented and under-represented samples and dynamically adjust the asymmetry levels throughout the training. The ASL uses two focusing hyperparameters to modify the contribution of easy samples to the loss function (γ_+, γ_-). In our work we set $\gamma_+ = 1$ and $\gamma_- = 3$.

2.3. Implementation Details

During model training, we monitor the average Challenge score and use early stopping when the validation Challenge score stops improving for 4 epochs. This approach is used for both, the first epochs without label correction, and when the label correction phase is performed after the model has been pre-trained. We use an Adam optimizer ($\beta_1 = 0.9, \beta_2 = 0.999$ and $\epsilon = 10^{-9}$) with a learning rate of 0.001. The batch size is set to 10. The complete model consists of 382.412 trainable parameters and it is trained on the 2021 PhysioNet/CinC challenge datasets and no other external data sources.

2.4. Self-learning label correction

In a real-world dataset, most of the "ground truth annotations" come from human experts, which are subjective, mistake-prone, and introduce bias in the data. Learning from noisy labels reduces model performance and it is still a challenge in DL. Following the work from [10], we include a self-learning label correction module to our model. This approach doesn't require a fully correctly labeled dataset as it iteratively estimates corrected labels during training. The main idea is to identify prototypes from the set of samples of each class that have a high chance of being correctly labeled. This is done once the model has already been trained with the original noisy dataset, so network features can be extracted from each sample to identify prototypes based on similarity and density measures. Each sample is then compared to the prototypes of each class and a corrected label is assigned if needed.

2.5. Evaluation

We evaluate the performance of the proposed method based on the "challenge score", as described in [6]. An ablation study is performed to investigate the effect of the ASL, hand-crafted features and, the LC phase added to the 1D-CNN. For this purpose, the dataset is split into train, validation, and test sets (60%, 20%, 20%).

3. Results

The results of the ablation study aimed at investigating the effect of each of the additional components added to the 1D-CNN are reported in Table. 3. The largest drop in performance occurs when removing the ASL component from the model reducing the challenge score by 14%. Removing the features has a marginal effect on the performance with a 1% drop in the 12, 4, and 2 lead local testings. A small increase in performance can be appreciated for the 6 and 3 lead combinations.

The label correction phase decreases the performance from 7%, for the 12 lead, up to 14% for the 2 lead.

When tested on the hidden validation set, we observe that including the features leads to a small, but consistent increase in performance. The model without label correction achieves a challenge score of 0.57, 0.58, 0.57, 0.56, and 0.57 for twelve, six, four, three, and two-leads, respectively. By adding the label correction the score dropped by 13-11%, a bigger decrease than in the local testing.

Finally, the best performing model on the validation test placed 7th in the challenge when evaluated on the test data, with an overall challenge score of 0.51, and 0.52, 0.45, 0.50, 0.50, and 0.49 for twelve-, six-, four-,three-, and two-lead, respectively. On the undisclosed data-set, an overall score of 0.42 would place our method 4th.

Table 3. Local challenge score obtained for the ablation study on the train set, from the full model, model without hand-crafted features and model without label correction.

Model	Challenge Score				
	12 lead	6 lead	4 lead	3 lead	2 lead
Full	**0.66**	0.63	**0.65**	0.60	**0.65**
w/o Feat.	0.65	**0.65**	0.64	**0.64**	0.64
w/o ASL	0.52	0.52	0.52	0.51	0.51
Including LC Phase					
Full	0.59	0.53	0.53	0.54	0.51
w/o Feat.	0.59	0.53	0.53	0.54	0.51

Table 4. Challenge scores for our final selected entry (team SMS+1) using validation on the public training set, 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.

Leads	Training	Validation	Test	Ranking
12	0.66	0.57	0.52	6
6	0.63	0.55	0.45	16
4	0.65	0.57	0.50	7
3	0.60	0.56	0.50	8
2	0.65	0.57	0.49	8

4. Discussion and Conclusion

Results clearly show that the asymmetric loss was most crucial to the performance of our model. We believe this is because the ASL introduces different weights for false positives and false negatives, which remedies the imbalance of positive and negative labels.

The added hand-crafted features slightly improve the model performance, we believe that this is due to the fact that some features, such as the entropy features, are likely hard to compute for a convolutional network.

In both, the ablation study and the official submission, there was no improvement in the results when performing the label correction phase. We assume that the reason is that in the original work the labels were much noisier [10], leading to more aggressive label-correction parameters. Additionally, the authors of [10] were dealing with a single-label classification problem, where relabeling was deciding on which class a label belongs to. For the multi-label classification problem considered in this work, the number of possible labels is exponential in the number of classes, making the task of finding a new label much harder, due to the search space being larger. Further investigation is therefore necessary.

To conclude, the best performance across all leads is achieved by the 1D-CNN in combination with the ASL and the manually handcrafted features. Our model shows a consistent ability in detecting a variety of cardiac abnormalities, on standard 12 lead ECGs as well as on various reduced-lead ECGs. Two will do!

Acknowledgments

References

[1] Ebrahimi Z, Loni M, Daneshtalab M, Gharehbaghi A. A review on deep learning methods for ECG arrhythmia classification. Expert Systems with Applications Sep. 2020; X(7):100033.

[2] S.M. M, Kambhamettu C, Barner KE. A novel application of deep learning for single-lead ECG classification. Computers in biology and medicine Aug. 2018;99:53–62.

[3] Natarajan A, Chang Y, Mariani S, Rahman A, Boverman G, Vij S, Rubin J. A wide and deep transformer neural network for 12-lead ECG classification. Computing in Cardiology Sep. 2020;1–4.

[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 Jul. 2020;122:103801.

[5] Alday EAP, Gu A, Shah AJ, Robichaux C, Wong AKI, Liu C, Liu F, Rad AB, Elola A, Seyedi S. Classification of 12-lead ECGs: the Physionet/Computing in Cardiology Challenge 2020. Physiological measurement Dec. 2020; 41(12):124003.

[6] Reyna MA, Sadr N, Alday EAP, Gu A, Shah AJ, Robichaux C, Rad AB, Elola A, Seyedi S, Ansari S, et al. Will two do? varying dimensions in electrocardiography: The Physionet/Computing in Cardiology Challenge 2021. Computing in Cardiology 2021;48:1–4.

[7] Ravichandran K, Chiasson D, Oyedele K. Classification of Electrocardiogram Anomalies 2013;Available at https://cs229.stanford.edu/proj2013/.

[8] Shaffer F, Ginsberg JP. An overview of heart rate variability metrics and norms. Frontiers in public health 2017;5:258.

[9] Ben-Baruch E, Ridnik T, Zamir N, Noy A, Friedman I, Protter M, Zelnik-Manor L. Asymmetric loss for multi-label classification. arXiv preprint arXiv200914119 Sep 2020;.

[10] Han J, Luo P, Wang X. Deep Self-Learning From Noisy Labels. Proceedings of the IEEECVF International Conference on Computer Vision 2019;5138–5147.

Address for correspondence:

Gallego Vázquez, Cristina
TAN E 5.2, Tannenstrasse 1, 8006 Zürich
cristina.gallegovazquez@hest.ethz.ch