Reduced-Lead ECG Classifier Model Trained with DivideMix and Model Ensemble

Hiroshi Seki[1], Takashi Nakano[1], Koshiro Ikeda[1], Shinji Hirooka[1], Takaaki Kawasaki[1], Mitsutomo Yamada[1], Shumpei Saito[1], Toshitaka Yamakawa[2], and Shimpei Ogawa[1]

[1]AMI inc., Kagoshima, Japan [2]Kumamoto University, Kumamoto, Japan

Abstract

Automatic diagnosis of multiple cardiac abnormalities from reduced-lead electrocardiogram (ECG) data is challenging. One of the reasons for this is the difficulty of defining labels from standard 12-lead data. Reduced-lead ECG data usually do not have identical characteristics of cardiac abnormalities because of the noisy label problem. Thus, there is an inconsistency in the annotated labels between the reduced-lead and 12-lead ECG data. To solve this, we propose deep neural network (DNN)-based ECG classifier models that incorporate DivideMix and stochastic weight averaging (SWA). DivideMix was used to refine the noisy label by using two separate models. Besides DivideMix, we used a model ensemble technique, SWA, which also focuses on the noisy label problem, to enhance the effect of the models generated by DivideMix. Our classifiers received scores of 0.623, 0.593, 0.606, 0.612, and 0.601 (ranked 99th, 99th, 99th, 99th, and 99th, respectively, out of 100 teams) for the 12-lead, 6-lead, 4-lead, 3-lead, and 2-lead versions, respectively, of the hidden validation set with the challenge evaluation metric.

1. Introduction

Cardiovascular disease is a leading cause of global mortality [1]. As the electrocardiogram (ECG) can record the electrical activity of the heart non-invasively, there are a lot of studies on the automatic diagnosis of cardiac abnormalities from ECG. PhysioNet/Computing in Cardiology Challenge 2021 focuses on the classification of cardiac abnormalities from reduced-lead ECGs [2, 3].

Real-world data are annotated by multiple human labelers with different skill levels. The annotation quality harms the performance of machine learning. The annotation quality is also affected by different annotation rules of each hospital. Therefore, there are many works on the noisy-label problem to train a robust model from noisy real-world datasets [4–6].

Likewise, the reduced-lead ECG classification can be regarded as the noisy label problem because the reduction of certain ECG leads hinders the detection of important characteristics of cardiac abnormalities.

In this paper, we propose DNN-based ECG classifier models that are robust to annotation inconsistency. These models incorporate DivideMix [5] and stochastic weight averaging (SWA) [6]. We used *EfficientNet* [7] which consists of 1D-CNN (convolutional neural network) for multi-label classification.

2. Multi-class Classification with DivideMix

2.1. Base Classifier

A multi-class classifier based on the DNN takes $|\mathcal{C}|$-dimensional ECG time-series data as input $X_c = (x_{c,t} \in \mathbb{R}^{|\mathcal{C}|} | t = 1, \ldots, T)$, and predicts the probabilities $Y = (y_1, \cdots, y_N)$ where T is the sequence length, N is the number of diagnoses, and y_i is the i-th label that takes 0 (negative) or 1 (positive). The dimension $\mathcal{C}$ represents the lead combination. In this study, we trained the classifier models to identify diagnoses from reduced-lead ECG sets: $c \in (\mathcal{C}_2, \mathcal{C}_3, \mathcal{C}_4, \mathcal{C}_6, \mathcal{C}_{12})$, where $\mathcal{C}_2$ = (I, II), $\mathcal{C}_3$ = (I, II, V2), $\mathcal{C}_4$ = (I, II, III, V2), $\mathcal{C}_6$ = (I, II, III, aVR, aVL, aVF), and $\mathcal{C}_{12}$ is the standard 12-leads [3].

EfficientNet first generates a sequence of hidden representations $\boldsymbol{h}_{c,1:T'} = (h_{c,t} \in \mathbb{R}^{|H|} | t = 1 \cdots T', T' \leq T)$ by taking the ECG signal X_c, where $h_{c,t}$ is the $|H|$-dimensional hidden vector at frame t. These hidden representations are then passed to a global max-pooling layer to obtain a fixed-length representation h_c. We represent these neural network modules as follows:

$$h_c = \text{pool}(f_{\text{effnet}}(X_c)), \tag{1}$$

where $\text{pool}(\cdot)$ and $f_{\text{effnet}}(\cdot)$ are the global max-pooling function and the *EfficientNet* module.

In multi-class classification, the posterior probability of diagnoses is calculated using a softmax layer with addi-

tional fully connected layers (MLP):

$$\hat{Y} = P(Y|X_c) = \text{softmax}(\text{MLP}(h_c)). \quad (2)$$

We define $f_{\text{cls}} = \text{MLP}(\text{pool}(f_{\text{effnet}}(\text{X}_c)))$ for simplicity. As our task is multi-label classification, we replace the softmax function with the sigmoid function in the next section.

2.2. Division of Training Data

Empirically, DNNs first learn to predict clean samples (expected to be high annotation quality), and later memorize noisy ones (expected to be poor annotation quality) [8]. By exploiting this observation, *DivideMix* [5] splits the training data into a set of clean samples and the one of noisy samples using two-component Gaussian Mixture Models (GMM). In the training stage, two models, $f_{\text{cls};\theta_m}(m = 1, 2)$, are trained in parallel. In other words, the training data that are split by $f_{\text{cls};\theta_1}$ are used for the training of $f_{\text{cls};\theta_2}$ in the next epoch, and vice versa.

In *MixMatch* [9], the noisy samples are used as unlabeled data. In the case of multi-class classification, two networks make predictions using the softmax function, and the posterior probabilities are averaged to create a new label:

$$\begin{aligned} u^{\text{nl}}_{c,\theta_m} &= \text{softmax}(f_{\text{cls};\theta_m}(\text{X}^{\text{nl}}_c)), m = 1, 2 \\ u^{\text{nl}}_c &= \text{Sharpen}((u^{\text{nl}}_{c,\theta_1} + u^{\text{nl}}_{c,\theta_2})/2.0), \end{aligned} \quad (3)$$

where X^{nl}_c is the ECG data estimated as a noisy label (nl) and Sharpen is a function introduced in [5].

3. Proposed Method

3.1. Multi-Label Label Refinement

In this section, we describe the modification of *DivideMix* for the multi-label classification. First, the softmax function in Eq. (3) is replaced with the sigmoid, and it is interpolated with the ground-truth label without the *sharpening* operation:

$$\begin{aligned} u^{\text{nl}}_{c,\theta_m} &= \text{sigmoid}(f_{\text{cls};\theta_m}(X^{\text{nl}}_c)), \\ u^{\text{nl}}_c &= \lambda_n(u^{\text{nl}}_{c,1} + u^{\text{nl}}_{c,2})/2.0 + (1 - \lambda_n)Y, \end{aligned} \quad (4)$$

where λ_n is the interpolation coefficient, and Y is the ground-truth label. In the experiment, λ_u was set to 0.5. The label of clean sample u^{cl}_c is updated as:

$$u^{\text{cl}}_{c,\theta_m} = \lambda_{\text{gmm}}Y + (1 - \lambda_{\text{gmm}}) \cdot \text{sigmoid}(f_{\text{cls};\theta_m}(X^{\text{cl}}_c)), \quad (5)$$

where λ_{gmm} is the probability which is estimated as clean by GMM. In contrast to the case of clean samples, the pseudo-label for the m-th network is estimated using the $m-$th network to reduce the training time.

Second, the sample-wise loss l is updated to a binary cross-entropy loss l_n and averaged over all labels:

$$\begin{aligned} l &= \text{mean}(\{l_1, \cdots, l_n \cdots, l_N\}), & (6) \\ l_n &= -\Big\{Y_n \log(\hat{Y_n}) + (1 - Y_n) \log(1 - \hat{Y_n}), \Big\} & (7) \end{aligned}$$

where Y and $\hat{Y}$ are the reference and estimated labels, respectively, and n is the label index. Because the reduction of certain leads hinders the detection of one part of diagnostic characteristics but not all diagnoses, it is our future work to model label-dependent losses.

3.2. Non-Sequential Manifold MixUp

Related works on image classification tasks apply *MixMatch* to the input data domain. However, it is not clear whether the interpolation of time-series data of different lengths affects the model training. Therefore, we generate the fixed-length hidden vector by using the max-pooling function and apply *manifold-MixUp* [10].

Let X^{cl}_c and Y^{cl} denote a pair of clean ECG sample and reference label, X^{nl}_c and Y^{nl} denote a pair of noisy ECG sample and reference label. The interpolation of the hidden vectors is then represented as follows:

$$\begin{aligned} h^{\text{cl}}_{c,\theta_m} &= \text{pool}(f_{\text{effnet};\theta_m}(\text{X}^{\text{cl}}_c)), \\ h^{\text{nl}}_{c,\theta_m} &= \text{pool}(f_{\text{effnet};\theta_m}(\text{X}^{\text{nl}}_c)), \\ h^{\text{mix}}_{c,\theta_m} &= \lambda_{\text{mix}}h^{\text{cl}}_{c,\theta_m} + (1 - \lambda_{\text{mix}})h^{\text{nl}}_{c,\theta_m}, \end{aligned} \quad (8)$$

where λ_{mix} is the coefficient used in *MixUp* sampled from a beta distribution. The interpolation of the label is:

$$u^{\text{mix}}_{c,\theta_m} = \lambda_{\text{mix}}u^{\text{cl}}_{c,\theta_m} + (1 - \lambda_{\text{mix}})u^{\text{nl}}_c, \quad (9)$$

and the objective function is defined as:

$$\begin{aligned} \mathcal{L} = L_x + L_u &= \text{BCE}(\text{MLP}(\mathcal{X}(h^{\text{mix}}_{c,\theta_m})), \mathcal{X}(u^{\text{mix}}_{c,\theta_m})) \\ &+ \text{L}_2(\text{MLP}(\mathcal{U}(h^{\text{mix}}_{c,\theta_m})), \mathcal{U}(u^{\text{mix}}_{c,\theta_m})) \end{aligned} \quad (10)$$

where BCE and L_2 are the binary cross-entropy and mean squared loss functions with the sigmoid function, and $\mathcal{X}$ and $\mathcal{U}$ are dummy functions which divide the samples into clean/noisy samples.

3.3. Model Ensemble

The model ensemble is a technique that combines predictions calculated by multiple classifiers for variance reduction (discussed in a context of a bias-variance tradeoff). Under the proposed framework, the two models were trained in parallel. Therefore, these two models can be used for model ensemble.

Table 1. *EfficientNet* model architecture. Each line describes a sequence of 1D convolution layer or Fused-MBConv (mobile inverted residual bottleneck convolution) modules consists of k-size kernels. The first convolutional layer of each stage has stride shown in the 3rd column and the followings use 1. #Channels is the number of output channels of each stage.

Stage	Operator	Stride	#Channels	#Layers
0	Conv, k: 7	2	32	1
1	Fused-MBConv2, k: 5	2	32	2
2	Fused-MBConv1, k: 5	2	64	1
3	Fused-MBConv2, k: 7	2	128	2
4	Fused-MBConv1, k: 7	2	128	1
5	Fused-MBConv2, k: 7	2	256	2
6	Fused-MBConv2, k: 7	2	256	2
7	Conv, k: 1	1	512	1

Stochastic weight averaging (SWA) [6] creates a new model by averaging the model weights sampled at different stages of training. In the experiment, we applied SWA to the two models to generate two averaged models. The result of final prediction is an arithmetic mean of the posterior probabilities calculated by the four models.

3.4. Model Architecture

Table 1 shows the model architecture based on EfficientNet [7]. Fused-MBConv is a sequence of 1) 1D convolution layer, 2) squeeze-and-excitation module, and 3) point-wise convolution layer. 1) The input tensor (W, C) is expanded to (W', 2C) at the first convolution layer followed by batch normalization (BN) and the Mish function [11] where W, C are the width and channel sizes. 2) In the squeeze-and-excitation module, channel statistics are summarized by a pooling function, and its dimension is reduced to C/4. This embedded feature is expanded to C followed by a sigmoid function for channel-wise attention. 3) Lastly, point-wise convolution and BN are used to update the output channel size.

Natarajan, et al., [12] proposed *wide*-and-*deep* Transformer neural networks. This approach uses a Transformer network to compute a fixed-length representation. It is fused with hand-crafted ECG features on top of the Transformer network to incorporate expert knowledge. Likewise, we used age, gender, and RR-interval-related features extracted from lead II as the *wide* features. These *wide* features are concatenated before the point-wise convolution to condition the all Fused-MBConv blocks.

4. Experimental Setup and Results

4.1. Feature Extraction

We used the CPSC database [13], INCART database [14], PTB database [15], PTB-XL database [16], Chapman-Shaoxing Database [17], Ningbo Database [18], and other databases [2, 3].

Table 2. Challenge scores for our final selected entry (team ami_kagoshima) using 10-fold cross validation on the public training set, 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.

Leads	Training	Validation	Test	Ranking
12	0.701 ± 0.006	0.623	???	???
6	0.686 ± 0.003	0.593	???	???
4	0.693 ± 0.006	0.606	???	???
3	0.693 ± 0.005	0.612	???	???
2	0.685 ± 0.006	0.601	???	???

All ECG signals were resampled to 500 Hz and normalized to a range of $[-1, 1]$ by min-max normalization for each lead. We extracted 15 seconds of ECG data from a random starting point and applied zero-padding when the duration was shorter than 15 seconds. When the duration (before zero-padding) τ was longer than 10 seconds, we decreased its duration randomly by sampling from the uniform distribution U(10, τ) to make the network learn duration-independent prediction.

We used stratified 10-fold cross-validation and averaged over 10 challenge metric scores for each reduced (2, 3, 4, 6, and 12) leads setup [3] to test the effectiveness of the proposed method. No additional processing was added to the different lead combinations. The Welch t-test was used for the statistical test.

4.2. Optimization

• Baseline Model: We used the model described in Section 3.4. The number of output units was set to 24 which corresponds to diagnoses scored by the Physionet 2021 Challenge. We used the Adam algorithm [19] and minimized the binary cross-entropy loss. The model was trained for 40 epochs with a batch size of 240. As the *wide* feature, we extracted age, gender, and RR-interval-related features computed by biosppy [20] and hrv [21]. It is passed to 4-layer fully connected layers with BN and a Mish function [11] followed by the Fused-MBConv module. 2-layer fully connected layers were used as the MLP introduced in Eq. (10). The predicted posterior probabilities were converted to positive or negative based on a fixed threshold of 0.3.

• Proposed Model: The model was trained for 40 epochs with a batch size of 160. The first two epochs were trained as the baseline model, and the other epochs were trained under the proposed framework. SWA was applied for the last 13 epochs. The number of expectation-maximization algorithm iterations used for GMM training was set to 10. All the models were trained from scratch.

4.3. Results

The averaged challenge-scores of the baseline method were 0.682, 0.667, 0.676, 0.673, and 0.664 on the 12-, 6-, 4-, 3- and 2-leads ECG data, respectively. Table 2 shows the challenge scores of the proposed method. Our results on 10-fold cross-validation were $0.701(2.8\%)^{(***)}$, $0.686(2.8\%)^{(***)}$, $0.693(2.5\%)^{(**)}$, $0.693(3.0\%)^{(***)}$, and $0.685(3.2\%)^{(***)}$ on the 12-, 6-, 4-, 3- and 2-leads ECG data, respectively[1]. The values given in the parentheses represent relative improvements.

5. Discussion and Conclusion

In this paper, we have proposed reduced-lead ECG classifiers based on *DivideMix* and SWA. As the reduction of certain ECG leads hinders the cardiac electrical signal, it is expected to degrade the classification performance when the available leads are limited. We can see that the challenge scores of the baseline and proposed models decreased linearly except for the 6-leads setup. The proposed method have obtained relatively large improvements on 2- and 3-leads setups. It is considered that the proposed method alleviated performance degradation owing to the poor annotation quality. Future work is detailed diagnoses-level investigations of the performance changes caused by the reduction of available lead combinations.

References

[1] Virani SS, Alonso A, Aparicio HJ, Benjamin EJ, Bittencourt MS, Callaway CW, et al. Heart Disease and Stroke Statistics – 2021 Update: a Report from the American Heart Association. Circulation 2021;143(8):e254–e743.

[2] 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;41.

[3] 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.

[4] Karimi D, Dou H, Warfield SK, Gholipour A. Deep learning with noisy labels: exploring techniques and remedies in medical image analysis. arXiv preprint 2019; arXiv:1912.02911.

[5] Li J, Socher R, Hoi SC. DivideMix: Learning with Noisy Labels as Semi-supervised Learning. In International Conference on Learning Representations. 2020; .

[6] Izmailov P, Podoprikhin D, Garipov T, Vetrov D, Wilson AG. Averaging Weights Leads to Wider Optima and Better Generalization. Uncertainty in Artificial Intelligence 2018; .

[7] Tan M, Le QV. EfficientNetV2: Smaller Models and Faster Training. In International Conference on Machine Learning. 2021; .

[8] Arpit D, Jastrzębski S, Ballas N, Krueger D, Bengio E, Kanwal MS, et al. A Closer Look at Memorization in Deep Networks. In International Conference on Machine Learning. PMLR, 2017; 233–242.

[9] Berthelot D, Carlini N, Goodfellow I, Papernot N, Oliver A, Raffel C. MixMatch: A Holistic Approach to Semi-Supervised Learning. In Advances in Neural Information Processing Systems. 2019; .

[10] Verma V, Lamb A, Beckham C, Najafi A, Mitliagkas I, Lopez-Paz D, et al. Manifold Mixup: Better Representations by Interpolating Hidden States. In International Conference on Machine Learning. 2019; 6438–6447.

[11] Misra D. Mish: A Self Regularized Non-Monotonic Activation Function. arXiv preprint 2019;arXiv:1908.08681.

[12] Natarajan A, Chang Y, Mariani S, Rahman A, Boverman G, Vij S, et al. A Wide and Deep Transformer Neural Network for 12-Lead ECG Classification. In Computing in Cardiology. IEEE, 2020; 1–4.

[13] 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.

[14] 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.

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

[16] 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.

[17] 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.

[18] 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.

[19] Kingma DP, Ba JL. Adam: A method for stochastic optimization. arXiv preprint 2014;arXiv:1412.6980.

[20] Carreiras C, Alves AP, Lourenço A, Canento F, Silva H, Fred A, et al. BioSPPy: Biosignal processing in Python, 2015–.

[21] Bartels R, Peçanha T. HRV: a Pythonic package for Heart Rate Variability Analysis. Journal of Open Source Software 2020;5(51):1867.

Address for correspondence:

Hiroshi Seki (hseki@ami.inc)
302, 2-13 Higashi-Sengoku, Kagoshima, Japan

[1] ***: $p < 0.001$, **: $p < 0.01$