Improvement Performance Deep Learning-based Multi-class ECG Classification Model with Limited Medical Dataset

Sanghoon Choi[1], Segyeong Joo[1]

[1]Department of Biomedical engineering, Asan Medical Institute of Convergence Science and Technology, Asan Medical Center, University of Ulsan College of Medicine, Seoul, Republic of Korea.

Abstract

Real-world medical data, such as electrocardiogram (ECG), often exhibits imbalanced data distributions, which presents a considerable difficulty for classification tasks. the use of augmented data remains controversial in the clinical field, as it is not real and may introduce biases into the training data learned by the generative model. In this study, we proposed a method for achieving optimal results without relying on augmentation techniques.

The experiments included using class weight (Experiment A-1), focal loss (Experiment A-2), balancing all classes to match the smallest class (Experiment B), and configuring subclasses (Experiment C). We employed the ResNet deep learning model for multi-class classification.

Experiment A used either focal loss or class weight and achieved high scores of 0.95 and 0.93 respectively. Focal loss had the best performance among the two methods.

We developed a method to improve the performance of an ECG classifier with limited data. Results show that properly weighting the loss function, specifically using focal loss, in a deep learning model is more effective than altering the amount of data to solve the issue of imbalanced data.

1. Introduction

Electrocardiogram (ECG) is a tool for diagnosis of cardiac diseases. Early and accurate detection of cardiac disease is important for intervention.

Recently, ECG classification using deep learning methods have been developed by collecting numerous ECG datasets [1-3]. However, real-world ECG dataset often exhibit data imbalance. Data imbalance in deep learning has substantial problem that can be biased majority classes as the model train for classification tasks. To address the challenge, several methods with data augmentation used to increase artificial signal of minority classes for balance each class have been proposed. These approaches have made significant contributions to enhancing the performance of deep learning models.

However, in medical field, the used of synthetic ECG signal has been controversial. Although the augmented data attributed to addressing data imbalance, these data set to use for training model are not real. In this study, we find the best results in multi class ECG classification not using an augmentation but the various methods that are changing loss function such as class weight and focal loss [4], setting balance dataset as minority class and configuration subclasses.

2. Methods

2.1. Datasets and Preprocessing

The datasets were collected a total of 5850 patients in Asan Medical Center from the Muse system (GE Healthcare, USA). This study was approved by the Institutional Review Board of the Seoul Asan Medical Center Hospital (IRB 2021-1259). The 12 lead ECG was recorded for 10s and sampled at 500Hz. The dataset for sampling rate of 250Hz or detaching one or more ECG leads were excluded. Also, the patients under the age of 18 were removed.

Table1. Amount of dataset in three Experiment

Class	Name of class	Imbalance dataset	Balance dataset	Subclass dataset
1	1st degree AV block or 2nd degree AV block (Mobits type 1)	279	279	885
2	PSVT	307	279	
3	High degree or complete AV block	299	279	
4	Irregular narrow QRS tachyarrhythmia	1451	279	1451
5	Sinus tachyarrhythmia	1132	279	1132
6	VPCs	2382	279	2382

In Table 1, there are three types of datasets, Imbalance dataset, Balance dataset, Subclass dataset, based on six

cardiovascular diseases. In imbalance dataset, the range of dataset is from 4.2% to 40%. The balance dataset is equally adjusted by the number of the lowest class, first or second-degree AV block (Mobitz type 1) as 279. In addition, the Subclass dataset combined minority classes, first or second-degree AV block (Mobitz type 1), PSVT and High degree or complete AV block, into a single class.

The preprocessing and normalization applied to dataset equally [5]. The Savitzky-Golay filter and low pass filter with a 4th order Butterworth and 50Hz of cutoff frequency used for removing baseline wondering caused by respiration or movements and high frequency noise such as power line interfere. Also, Min-max normalization was applied to scale the data to within [-1,1]. The normalization was calculated as (1):

$$Scale = \frac{data - \min(data)}{\max(data) - \min(data)} \times 2 - 1 \quad (1)$$

We conducted three experiments, Experiment A, B and C to verify for analyzing and comparing the performance of the developed method with those of the previous approaches (Figure 1).

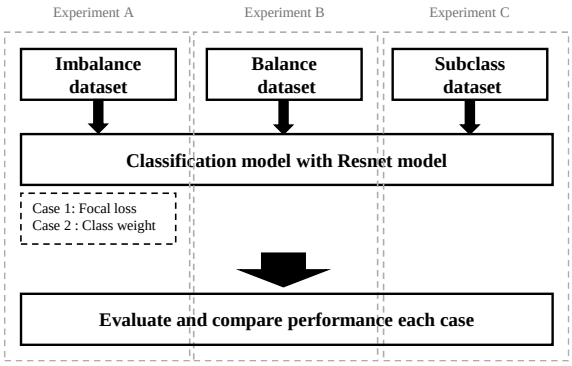

Figure 1. overview of proposed method.

In Experiment A with Imbalance dataset, we changed loss function of deep learning model as class weight and focal loss. Both techniques are weighting loss function based cross entropy function. The class weight is weighting loss function based the number of each class and calculated as (2):

$$Class\ weight = \frac{The\ number\ of\ data}{(class\ number \times each\ class\ data)} \quad (2)$$

Table 2 are the weights for each class. The majority classes are calculated a lower weight to the loss function, on the other hand, the minority classes assign a higher weight.

Another method is focal loss. In training phase, the easily classified examples means higher probability result than difficult-to-classify examples. The focal loss

Table 2. The values of class weight in each class

Class	Name of class	Class weight	The amount of data
1	1st degree AV block or 2nd degree AV block (Mobits type 1)	3.47	279
2	PSVT	3.07	307
3	High degree or complete AV block	3.39	299
4	Irregular narrow QRS tachyarrhythmia	0.67	1451
5	Sinus tachyarrhythmia	0.86	1132
6	VPCs	0.41	2382

function assigns lower weights to easily classified examples and more weight to difficult examples. The formulation of the focal loss function is as depicted as (3):

$$Focal\ Loss: -\alpha_{target}\left(1 - p_{target}\right)^{\gamma} \log p_{target} \quad (3)$$

The parameters α_{target} and $\left(1 - p_{target}\right)^{\gamma}$ govern the weighting of the loss function. The purpose of the $\left(1 - p_{target}\right)^{\gamma}$ term is to diminish the influence of the loss by a factor of γ. In this study, the value of α_{target} and γ were determined using the grid search technique (Table 3).

In Experiment B with Balance dataset, we experimented how Balance dataset which matched the number of data equally affects the performance of the deep learning model. The Experiment C was applied with a subclass dataset combined with minority classes. To identify the generality of the deep learning model from dataset, we used 5-fold cross validation.

Table 3. The parameters of focal loss

γ	α	F1 score
1	0.25	0.942
2	0.25	0.951
3	**0.25**	**0.963**
5	0.25	0.962

2.2. Deep learning model

We used deep learning model based Resent architecture for multi class classification [6]. The architecture of the ResNet model is illustrated in Figure 2. Each block consists of a convolutional layer, followed by a max pooling layer, and two residual blocks. A total of five such blocks were stacked. The initial layer and the first two blocks employed 16 convolution filters. The number of filters was doubled with each subsequent block. The kernel size decreased by a factor of two, starting from nine. The learning rate and dropout rate were set at

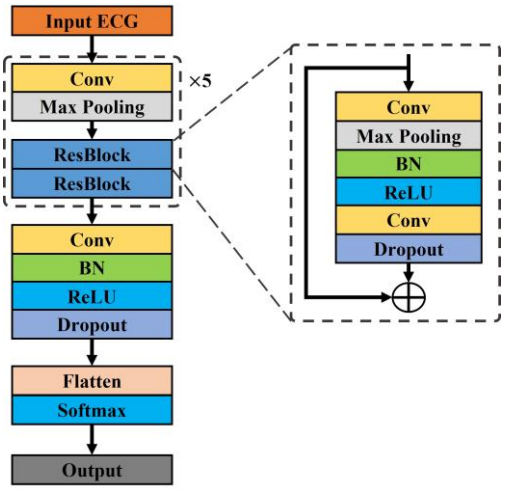

Figure 2. The ResNet model for classification

0.0005 and 0.1, respectively. The model underwent training for a total of 120 epochs.

2.3. Evaluation

The evaluation metrics were the accuracy, precision, recall and F1 score, which are calculated as follows:

$$Accuracy = \frac{TP+TN}{TP+FN+FP+TN} \quad (3)$$

$$Precision = \frac{TP}{TP+FP} \quad (4)$$

$$Recall = \frac{TP}{TP+FN} \quad (5)$$

$$F1\ score = \frac{2\times(Precision\times Recall)}{Precision+Recall} \quad (6)$$

TP denotes true positive, FP denotes false positives, FN denotes false negatives, and TN denotes true negatives.

3. Results

In Tabel 4, we evaluated the performance of each experiment as micro average score for evaluation metrics. The performance of balance dataset is in the range of 0.85 to 0.86, which represents that limiting the number of datasets as 279 recordings equally had not impacted the results. On the other hand, Changing loss function exhibit better result. However, the result of focal loss has the highest performance when comparing other methods with a score of F1 score at 0.97. Especially, in Tabel 5, the minority classes, class 1,2, and 3, were better performance in the focal loss than imbalance case. Figure 3 shows the normalized confusion matrix for each experiment.

Table 4. Results for Micro-Average score in each Experiment

Experiment		Precision	Recall	F1-score	Accuracy
Imbalance		0.91	0.91	0.91	0.91
Focal loss		**0.96**	**0.97**	**0.97**	**0.97**
Class weight		0.93	0.93	0.93	0.93
Balance		0.85	0.85	0.85	0.86
Subclass	1st	0.97	0.96	0.97	0.97
	2nd	0.86	0.86	0.86	0.86

Table 5. The results of each class for Imbalance case and Focal loss

Imbalance	Precision	Recall	F1-score	Support
1st degree AV block or 2nd degree AV block (Mobits type 1)	0.78	0.69	0.73	55
PSVT	0.82	0.92	0.92	61
High degree or complete AV block	0.6	0.93	0.7	58
Irregular narrow QRS tachyarrhythmia	0.97	0.88	0.92	286
Sinus tachyarrhythmia	0.98	0.98	0.98	222
VPCs	0.97	0.98	0.98	488
				0.91

focal loss	Precision	Recall	F1-score	Support
1st degree AV block or 2nd degree AV block (Mobits type 1)	0.88	0.84	0.86	55
PSVT	0.94	0.95	0.95	61
High degree or complete AV block	0.79	0.91	0.85	58
Irregular narrow QRS tachyarrhythmia	0.95	0.97	0.96	286
Sinus tachyarrhythmia	0.99	0.97	0.96	222
VPCs	1	0.97	0.98	488
				0.96

4. Discussion

In a limited medical data environment, we compared and verified the best performance of a heart disease classifier using deep learning with 12-lead resting ECGs without data augmentation. In significant data imbalance, a more effective approach is to consolidate the data into a compact dataset organized by classes and train it in a two-step manner. This involves prioritizing the utilization of a broader class representation over focusing solely on

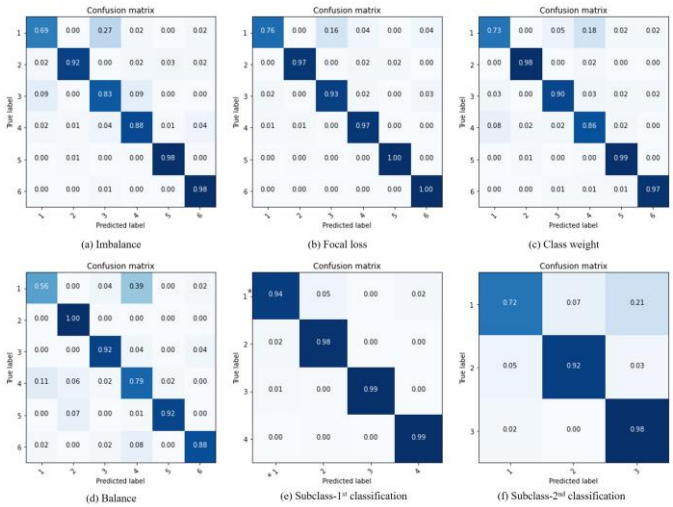

Figure 3. Confusion matrix for each experiment. In Fig3-(e), class 1* is subclass contained minor classes such as 1st degree AV block or 2nd degree AV block (Mobits type 1), PSVT, and High degree or complete AV block.

subclasses. However, it should be noted that the dataset containing six classes in this study may not fully capture the nuances of real-world medical data. It's imperative to contrast this dataset with the classification model's performance by integrating alternative arrhythmia configurations. Furthermore, our dataset segmentation was guided by a class-centric strategy rather than a subject-centric approach, which may impact the efficacy of classification tasks.

5. Conclusion

We have proposed a methodology to enhance the performance of an ECG classifier in a data-constrained context. Upon deploying the model across various scenarios, the top and bottom F1 scores were observed as 0.96 for the Inception net utilizing the focal loss, and 0.86 within a confined data environment while maintaining the same ratio. These findings underscore the significance of appropriately adjusting the loss function's weighting, particularly the focal loss function, within a deep learning model, which yields greater impact compared to manipulating the dataset size to address challenges posed by an imbalanced environment. In the domain of ECG class classification, minor changes in morphology correlate with reduced performance, while more pronounced morphology shifts align with improved performance. This study suggests a promising avenue for future research in developing an optimal classifier tailored to constrained medical contexts.

References

[1] Z. Liu and X. Zhang, "ECG-based heart arrhythmia diagnosis through attentional convolutional neural networks," in *2021 IEEE International Conference on Internet of Things and Intelligence Systems (IoTaIS)*, 2021: IEEE, pp. 156-162.

[2] G. Petmezas, K. Haris, L. Stefanopoulos, V. Kilintzis, A. Tzavelis, J. A. Rogers, A. K. Katsaggelos, and N. Maglaveras, "Automated atrial fibrillation detection using a hybrid CNN-LSTM network on imbalanced ECG datasets," *Biomedical Signal Processing and Control,* vol. 63, pp. 102194, 2021.

[3] P. Zhang, Y. Chen, F. Lin, S. Wu, X. Yang, and Q. Li, "Semi-Supervised Learning for Automatic Atrial Fibrillation Detection in 24-Hour Holter Monitoring," *IEEE Journal of Biomedical and Health Informatics,* vol. 26, no. 8, pp. 3791-3801, 2022.

[4] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980-2988.

[5] S. Chatterjee, R. S. Thakur, R. N. Yadav, L. Gupta, and D. K. Raghuvanshi, "Review of noise removal techniques in ECG signals," *IET Signal Processing*, vol. 14, no. 9, pp. 569-590, 2020.

[6] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770-778.

Address for correspondence:

Sanghoon Choi

sanghunc95@gmail.com

Segyeong Joo

26, Olympic-ro 43-gil, Songpa-gu, Seoul,

Republic of Korea(05506)

sgjoo@amc.seoul.kr