2021 43rd Annual International Conference of the
IEEE Engineering in Medicine & Biology Society (EMBC)
Oct 31 - Nov 4, 2021. Virtual Conference

An Approach for Deep Learning in ECG Classification Tasks in the Presence of Noisy Labels

Xinwen Liu, Huan Wang, Zongjin Li*

Abstract— ***Cardiovascular disease (CVD) is a serial of diseases with global leading causes of death. Electrocardiogram (ECG) is the most commonly used basis for CVD diagnosis due to its low cost and no injury. Due to the great performance shown in classification tasks with large-scale data sets, deep learning has been widely applied in ECG diagnosis. Manual labeling is a time-consuming and labor-intensive job, which makes it error-prone and easy to labeled wrongly. These noisy labels cause deterioration in performance since deep neural network is easy to over-fitting with noisy labels. However, currently, only limited studies have been concerned with this problem. To alleviate the performance degradation caused by noisy labels, we come up with an optimization method combining data clean and anti-noise loss function. Our method filters the noisy data by data-clean method, followed by training the network with boot-hard loss function. The experiment is carried on MIT-BIH arrhythmia database and we take a 1-D CNN model for test. The result indicates that our optimization method can produce an effective improvement for noisy label problems when the proportion of incorrect labels ranging from 10% to 50%.***

Clinical Relevance— ***The proposed algorithm can be potentially applied to deal with the noisy label problem in ECG diagnosis task.***

I. INTRODUCTION

Cardiovascular disease (CVD) is a general term of a serial of diseases serving as a global leading cause of death in recent a few years [1]. ECG is the most commonly used tool to diagnose CVDs due to its advantage of no injury and low cost. It is applied to record the electrical activity of the heart during each cardiac cycle, so ECG can contain much information of the heart and the abnormality or irregularity heart activity [2]. In this way, corresponding cardiovascular diseases or potential health problems can be diagnosed or predicted based on ECG.

Deep learning has achieved significant performance in the field of physiological signal, especially the application of convolution neural network (CNN). Under this circumstance, electrocardiogram (ECG), as one of the most common physiological signals in clinical medicine, has been the focus of much research over last decades as well. CNN is widely applied in ECG classification tasks and proved to be outstanding compared with traditional methods. Xu et al. [4] proposed a CNN network with coupled-convolution structure, achieving a heartbeat classification accuracy higher than 99%. Lang et al.[6] proposes a new neural network architecture for ECG diagnosis based on Deform-CNN. The overall diagnostic accuracy rate of this architecture in the 12-lead ECG data of CPSC-2018 can reach 86.3%. Yibo et al. [7] [8, 9] came up with a temporal attention mechanism-based CNN network which can reach great result in atrial fibrillation. All these studies were carried out on standard datasets, where the data is labeled carefully and verified to be correct.

Good performance of CNN is based on large-scale and the correct information to learn. However, ECG datasets are constructed by manual labeling which is an overwhelming task even for experts, making it error prone and possible to label the ECG segments wrongly, which lead to label noise. CNN can easily over-fit with the wrongly labeled training data, leading to obvious performance degradation. In fact, it is noisy label that serves as the most salient factor of reducing ECG classification accuracy [10].

To solve the problem of noisy label, basic regularization approach such as early stop and dropout can be used to eliminate over-fitting with incorrect label. However, their effectiveness cannot be guaranteed because they may prevent the reduction of training loss [11]. Prior knowledge can also serve as one approach to this issue, but it is unable to work in reality. In fact, limited attention is drawn to label noise problem in ECG diagnosis field. Pasolli et al. [12] proposed a genetic algorithm based optimal subset research where the data outside the optimal subset are deleted. Though efficient is the method, the highest label noise level of this research is only 20%. Li et al. [13] used 5 different machine learning classifiers in cross validation to remove the mislabeled data.

Though these methods are effective, one common drawback of the methods mentioned above is their expensive computational load, which makes them hard to be realized in practice diagnosis. In this paper, we managed to mitigate the performance degradation of CNN caused by noisy labels by combining data clean and anti-label noise loss function. It is proved that our method is able to improve the ECG classification accuracy at the label noise level ranging from 10% to 50%. We conducted the experiment on the MIT-BIH arrhythmia database with a 1-D CNN network.

This paper is organized as follows. Section 2 details the method of our approach and Section 3 shows the experiment results. A brief conclusion is drawn in Section 5.

II. METHODOLOGY

In this section of the paper, the principle of applied method will be introduced. The proposed method has two steps: data clean and boot-hard loss function, and they will be explained respectively.

This work was supported by Sichuan Science and Technology Program (2020108) and Glasgow College, University of Electronic Science and Technology of China. The experimental procedures involving human subjects described in this paper were approved by the Institutional Review Board. (Corresponding author: Huan Wang).

Xinwen Liu and Zongjin Li, Huan Wang are with the Glasgow College, University of Electronic Science and Technology of China, Chengdu, 611731, China. Huan Wang is also with the Department of Industrial Engineering, Tsinghua University, China (email: wh.huanwang@gmail.com).

978-1-7281-1178-0/21/$31.00 ©2021 IEEE

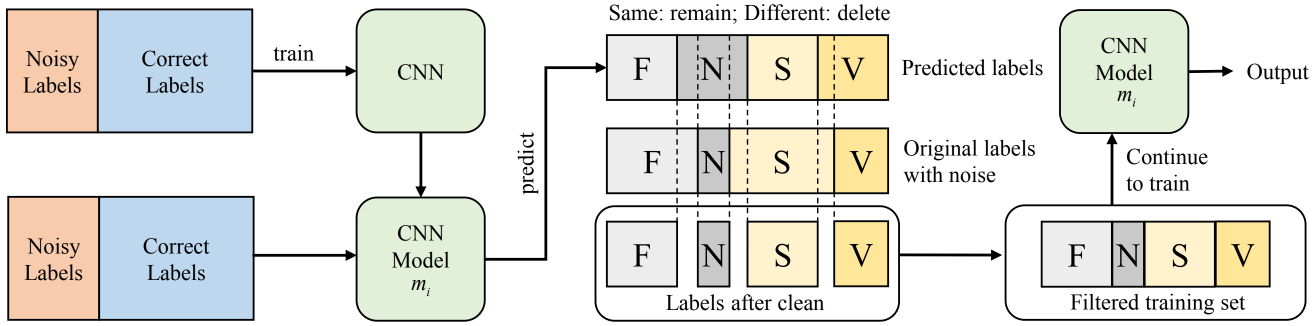

Figure 1 Procedures of the implementation of data clean.

A. *Data Clean Method*

In the process of training, the CNN network often performs a learning of correct data at first, and then prone to over-fit to incorrect-labeled data easily, leading to an obvious decline in accuracy. This can be ascribed to the reason that deep neural networks prefer to learn simple representations by subspace dimensional compression, followed by the over-fitting of noisy labels by dimensional expansion [14]. We make use of this characteristic of CNN to reduce the proportion of data with uncorrected labels in the training set, thus alleviate its over-fitting with the noisy labels.

We first train the CNN and utilize the model of the i^{th} epoch, m_i, to predict the labels of data in the training set. Suppose that predicted labels are correct, data with inconsistent labels are deleted after comparison. The remaining data form a filtered training set, to continue the training of m_i. To find the most proper m_i, a range of k models are saved and experimented. For the one with the best result, we choose the corresponding model as the proper one to implement data clean method. It needs to be mention that only cross-entropy loss function is used in the network in process of data clean.

The proposed data clean method filters out the data that are predicted to be with noisy labels and use them to update CNN in the follow-up training. F**ig. 1** illustrates the procedures of how data clean works in details.

B. *Boot-Hard Loss Function*

Boot-hard loss function is a loss function that developed based on 'bootstrapping', meaning pulling up oneself only by its own bootstraps [15]. By utilizing boot-hard, the network managed to own the capability of judging the consistency of noisy labels by reducing its attention paid on uncorrected labels [16].

The main idea of boot-hard is to update the prediction target dynamically, based on the model's current state. The final targets are produced by the 'hard' version convex combination of current predicted labels and noisy training ones. To be precise, cross-entropy objective is still used while new regression targets are generated in each mini-batch on the basis of current state of the model.

Bootstrapping can be seen as an instance of a method in which given predicted class probability p given x, the regression target is modulated by a softmax temperature parameter T in the model:

$$P(p_j=1\,|\,x)=\frac{\exp(T(\sum w_{ij}^{(1)}x_i+b_j^{(1)}))}{\sum_{j'}\exp(T(\sum w_{ij'}^{(1)}x_i+b_{j'}^{(1)}))} \qquad (1)$$

where w and b denote weight and bias in network neuron. When T=∞, hard bootstrapping is recovered.

Boot-hard uses Maximum A Prosteriori (MAP) estimate of class probability p to adjust regression targets, which is denoted as z_k. Then loss function can be obtained：

$$L_{hard}(p,t)=\sum_{k=1}^{l}[\beta t_k+(1\text{-}\beta)z_k]\log p_k \qquad (2)$$

where t_k refers to training target of data and β is a parameter can be adjusted.

This may result in an Expectation Maximization Algorithm (EM) like algorithm when used with mini-batch gradient descent: estimate the 'true' label based on predicted labels and given ones and in E-step, and update parameters that can better predict targets in M-step.

C. *Implementation Details*

To demonstrate the effectiveness of our method, we performed experiments on MIT-BIH Arrhythmia Database with a 11-layer 1-D CNN network whose structure is shown in **TABLE I**. Adam optimizer and learning rate of 0.00006 is adopted. Each model is trained for 100 epochs and the batch size is set to 32. Since data imbalance of the database may cause a poor accuracy, we applied oversampling to alleviate the problem. Our experiment is carried out with Nvidia GPU 2070s, using Keras framework and Tensorflow as backend. Since the ECG recordings are too long to be used directly, we cropped all recordings into pieces with a length of 250s for one single heartbeat.

Assume that all origin data obtained in MIT-BIH Arrhythmia Database is correct, we added label noise by replacing the correct labels with any one of other categories randomly. The noise level is the proportion that the number of replacements takes up in all samples.

TABLE III RESULTS OF TRAINING WITH NOISY LABELS

Noise Level (%)		No noisy label	10	20	30	40	50
Accuracy	*Best*	98.36%	96.88%	91.52%	80.31%	73.30%	53.11%
	Last	98.04%	95.66%	84.95%	71.63%	57.60%	45.48%

TABLE IV RESULTS OF USING DATA CLEAN AT FIRST TEN EPOCHS WHEN NOISE LEVEL IS 40%

Clean epoch		1	2	3	4	5	6	7	8	9	10
Accuracy	*Best*	76.36%	74.52%	75.84%	76.80%	73.13%	74.94%	75.02%	74.59%	76.39%	76.09%
	Last	74.65%	65.67%	73.68%	75.25%	70.12%	73.32%	72.86%	67.57%	68.38%	70.12%

TABLE I DETAILS OF 1-D CONVOLUTIONAL NETWORK

No.	Layers	Layer Details	Padding
1	Conv1	3×1, 32	valid
2	Conv2	3×1, 32	same
3	MaxPool1	4×1	valid
4	Conv3	3×1, 64	same
5	Conv4	3×1, 64	valid
6	MaxPool2	2×1	same
7	Conv5	3×1, 128	valid
8	MaxPool3	2×1	same
9	Dense	512	
10	Dense	512	
11	Dense	4	

For every experiment result, we recorded the average score of five trails and adopted accuracy as the evaluation standard, which is denoted as:

$$Accuracy = \frac{TP + TN}{TP + FN + TN + FP} \tag{3}$$

where *FP*, *FN*, *TP* and *TN* represent the number of data that is false positive, false negative, true negative and true positive respectively.

III. Experiment And Results

In this section, we present the experiment results of proposed method and comparative experiment to prove the effectiveness of proposed approach.

A. Data Description

All experiments are implemented on the MIT-BIH Arrhythmia Database, which contains 48 fully annotated 30-minute two-lead ECGs with 360Hz sampling rate. These records are obtained from 47 subjects who were 25 men between the ages of 32 and 89 and 22 women between the ages of 23 and 89. Five classes of ECG heartbeats are included N (normal beat), S (supra-ventricular arrhythmia), V (complex ventricular contraction), F (fusion of ventricular and normal beat), and Q (unclassified beat). Since Q beats only take up quite small part in all records, we only consider N, S, V and F beats classification in our experiment. The number of each type is shown in **TABLE II**.

TABLE II NUMBER OF ALL TYPES OF HEARTBEATS

Type	S	N	F	V
Quantity	2927	89841	802	7008

B. Training under Noisy Label

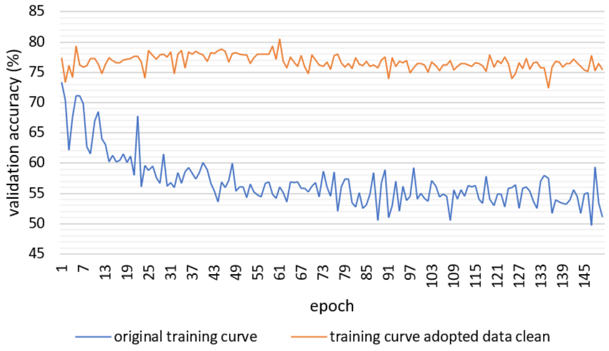

Figure 2 Training curves of models that utilized data clean method and the one not respectively

In this part, to show the influence of noisy labels on CNN, the 1-D CNN model's performance with noisy label is shown. We trained the basic CNN for all noise levels and noise free, serving as a comparison for the performance of our proposed method and demonstrating the over-fitting in the training process. Here we added label noise at different levels (l = 10%, 20%, 30%, 40% and 50%) to the training set.

The classification accuracy is presented **TABLE III**, where *best* represents the scores of the epoch with optimal validation accuracy and *last* represents the scores of the last epoch. By comparing the results of *best* and *last*, the over-fitting situation of the network in the training process can be demonstrated by the obvious difference in accuracy between them. To be precise, there is a difference of 45% by adding label noise level of 50%, which means label noise has a significant impact on network. Therefore, models' robustness to noisy label is important in practical applications.

C. The Effectiveness of Data Clean

In this experiment, we applied data clean to the model with cross-entropy loss function. By applying data clean at the first ten models separately, we further discuss the epoch where data clean can achieve the best results. Since it will be too complex if we list the exploration process for all 5 noise levels, here we only take l = 40% as an example. **TABLE IV** represents the accuracy obtained by best model and last model respectively. By comparing the results in **TABLE III** and **TABLE IV**, the improvement brought by data clean is shown. When l= 40%, an increase of 5% takes place. The gap between the results of *best* and *last* can show that data clean managed to reduce model's over-fitting to data with noisy labels. The training curve of models can also prove this (**Fig. 2**). For all noise levels, we adopted the 4th epoch as data clean epoch. The improvement in accuracy and over-fitting problem is caused by reducing the proportion of incorrect labeled data in the data set in the process of data clean.

D. The Effectiveness of Boot-hard Loss Function

In this experiment, boot-hard loss function is applied to the network on the basis of the implementation of data clean. The parameter β of boot-hard function is adjusted to find the most effective one. Since the range of β is 0 to 1, we carried out experiment on 0.1, 0.2, 0.4 and 0.8. Similar with the last part, here we also take l = 0.4 as an example and the relevant data can be found in **TABLE V**. It manages to achieve an accuracy of 80% through our method, which is the same with the accuracy when l = 0.3. Moreover, we found that when β is set to 0.1, model performs well for all noise levels in the experiments. Boot-hard loss function managed to further improve the model's performance by reducing the attention that network pays on wrongly labeled data, allowing the model to have a coherent learning of training set.

Table V RESULTS OF BOOT-HARD LOSS FUNCTION WHEN NOISE LEVEL IS 40%

β	0.1	0.2	0.4	0.8
Accuracy *Best*	80.52%	78.04%	74.76%	76.18%
Last	75.75%	74.22%	71.76%	73.55%

E. Comparison with method existed

TABLE VI RESULTS OF PROPOSED METHOD AND MAE

Noise level (%)		**10**	**20**	**30**	**40**	**50**
Proposed	***Best***	98.11%	97.84%	86.66%	80.52%	60.99%
	Last	97.53%	97.25%	83.52%	75.75%	55.29%
MAE	***Best***	92.16%	83.28%	61.71%	52.60%	36.50%
	Last	82.97%	56.67	24.51%	24.76%	25.01%

In this part, we show the superiority of our proposed method by comparing its result with MAE loss function. It needs to be clarified that β is set to 0.1 and the 4th epoch is adopted for data clean for all noise levels in our method. The results are reported in **TABLE VI.** For the noise level of 10% and 20%, the accuracy of the best model can be recovered to around 98%, which is nearly the accuracy when there is no label noise in the training set. For noise levels higher than 20%, a general increase of more than 6% can be reached for best models and more than 10% for last ones. Even for the noise level of 50%, the *best* accuracy can increase by 7.88%. Generally, our method is effective for all noise levels.

Besides, we compare the performance of our method with Mean Absolute Error (MAE) loss function which is widely used due to its good performance when there are outliers in the training data [17]. All experimental settings are the same for MAE and proposed method. We simply change the loss function of original CNN from cross-entropy to MAE. Corresponding results of MAE are also recorded in **TABLE VI**. We can see that the model with MAE shows even lower robustness than that with cross-entropy when dealing with ECG signals. Its accuracy of the last model is nearly random classification when the noise level is no less than 30%. This maybe for the reason that MAE makes the network unable to learn effective information when there are too many error labels, leading to a random guess on test set.

IV. CONCLUSION

In this study, we managed to explore an approach that can improve the accuracy when facing label problem in ECG classification tasks. We use data clean method and boot-hard loss function to improve the classification accuracy by reducing the proportion of wrongly labeled data in the training set and reducing the network's attention paid to the data with uncorrected labels. In this process, the proper epoch for data clean method and corresponding parameters of boot-hard loss function are explored and verified by experiments. By experimenting it on MIT-BIH Arrhythmia Database with a 1-D CNN model, our proposed method is proved to be efficient in reducing noisy labels' influence on the performance of network, even when the noise level reach 50%.

References:

[1]. Heart Disease and Stroke Statistics-2019 Update: A Report From the American Heart Association. Circulation, 2019.

[2]. Liu, X., et al., Deep learning in ECG diagnosis: A review. Knowledge-Based Systems, 2021. 227: p. 107187.

[5]. Acharya, U.R., et al., A deep convolutional neural network model to classify heartbeats. Computers in Biology and Medicine, 2017. 89.

[6]. Qin, L., et al., An End-to-End 12-Leading Electrocardiogram Diagnosis System Based on Deformable Convolutional Neural Network With Good Antinoise Ability. IEEE Transactions on Instrumentation and Measurement, 2021. 70: p. 1-13.

[7]. Gao, Y., H. Wang and Z. Liu, An end-to-end atrial fibrillation detection by a novel residual-based temporal attention convolutional neural network with exponential nonlinearity loss. Knowledge-Based Systems, 2021. 212: p. 106589.

[8]. Gao, Y., H. Wang and Z. Liu, An end-to-end atrial fibrillation detection by a novel residual-based temporal attention convolutional neural network with exponential nonlinearity loss. Knowledge-Based Systems, 2021. 212: p. 106589.

[9]. Gao, Y., H. Wang and Z. Liu, An end-to-end atrial fibrillation detection by a novel residual-based temporal attention convolutional neural network with exponential nonlinearity loss. Knowledge-Based Systems, 2021. 212: p. 106589.

[10]. Dubois, K.N., DEEP MEDICINE: How Artificial Intelligence Can Make Healthcare Human Again. Perspectives on Science and Christian Faith, 2019. 71.

[11]. Tanaka, D., et al. Joint Optimization Framework for Learning with Noisy Labels. in 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2018.

[12]. Pasolli, E. and F. Melgani, Genetic algorithm-based method for mitigating label noise issue in ECG signal classification. Biomedical Signal Processing & Control, 2015. 19: p. 130-136.

[13]. Li, Y. and W. Cui, Identifying the mislabeled training samples of ECG signals using machine learning. Biomedical signal processing and control, 2019. 47(JAN.): p. 168-176.

[14]. Wang, Y., et al., Symmetric Cross Entropy for Robust Learning with Noisy Labels. arXiv, 2019.

[15]. Davison, A.C. and D.V. Hinkley, Bootstrap Methods and their Application. 1997: Cambridge University Press.

[16]. Reed, S., et al., Training Deep Neural Networks on Noisy Labels with Bootstrapping. Computer Science, 2014.

[17]. Prashanth, T., S. B and S. Saha, Deriving the Lipschitz constant for MAE loss function. 2019.