Learning From Alarms: A Novel Robust Learning Approach to Learn an Accurate Photoplethysmography-Based Atrial Fibrillation Detector using Eight Million Samples Labeled with Imprecise Arrhythmia Alarms

Cheng Ding[1], ZhiCheng Guo[2], Cynthia Rudin[3], Ran Xiao[4], Amit Shah[5], Duc H. Do[6], Randall J Lee[7], Gari Clifford[1,8], Fadi B Nahab[9], Xiao Hu[1,4,8]

[1] Department of Biomedical Engineering, Georgia Institute of Technology, Atlanta, GA, USA
[2] Department of Electrical and Computer Engineering, Duke University, Durham, NC, USA
[3]Department of Computer Science, Duke University, Durham, NC, USA
[4]Nell Hodgson Woodruff School of Nursing, Emory University, Atlanta, GA, USA
[5]Department of Medicine, Emory University School of Medicine, Atlanta, GA, USA
[6]School of Medicine, University of California at Los Angeles, Los Angeles, CA, USA
[7]School of Medicine, University of California, San Francisco, CA, USA
[8]Department of Biomedical Informatics, Emory University School of Medicine, Atlanta, GA, USA
[9]Department of Neurology, Emory University School of Medicine, Atlanta, GA, USA

Abstract

Atrial fibrillation (AF) is a common cardiac arrhythmia with serious health implications. Passive monitoring using photoplethysmography (PPG) is desirable for long-term detection of AF. Deep neural networks (DNNs) show promise, but it requires massive training data with clean label, which is hard to obtain. To address the challenge, in this study, a large-scale dataset is created using PPG signals from hospital monitors, and each PPG signal is automatically annotated by concurrent alarms. However, the labels of collected PPG can be noisy because of the existence of false alarms. Then a novel loss function, cluster membership consistency (CMC) loss, is introduced to handle label noise caused by inaccurate PPG labels. The proposed approach shows superior performance in handling label noise and poor-quality signals. This novel approach is shown to be effective in further improving the model performance. It achieves superior or comparable results when evaluated against five different state of the art robust learning algorithms for noisy labels, while at the same time maintains computational efficiency advantage.

1. Introduction

Atrial fibrillation (AF) is a common irregular heart rhythm that can indicate serious health conditions [1]. Detecting AF in ambulatory settings is important for timely intervention. Photoplethysmography (PPG) is a suitable method for passive monitoring of AF due to its physiological basis and wide availability in wearable devices. However, existing AF detection algorithms using PPG often yield false positives, leading to unnecessary medical resource utilization. Deep neural networks (DNNs) show promise in accurate AF detection based on PPG [2], [3], but lack sufficient training data with well-annotated labels and struggle with the presence of other arrhythmias [4].

To address these challenges, this study introduces two innovations. First, a large-scale dataset consisting of approximately 8 million PPG strips from 24,100 patients is created by leveraging PPG signals collected from hospital monitors with built-in arrhythmia detection algorithms. However, the labels generated by these algorithms are imperfect. Second, a novel approach is proposed to achieve robust learning in the presence of label noise caused by inaccurate PPG labels. Specifically, we introduce the cluster membership consistency (CMC) loss, as shown in Fig. 1. This loss ensures a natural cluster structure in the latent space of the autoencoder used for signal processing. The deep neural network, trained with the CMC loss and cross-entropy loss, demonstrates superior performance in handling label noise and poor-quality signals compared to baseline models.

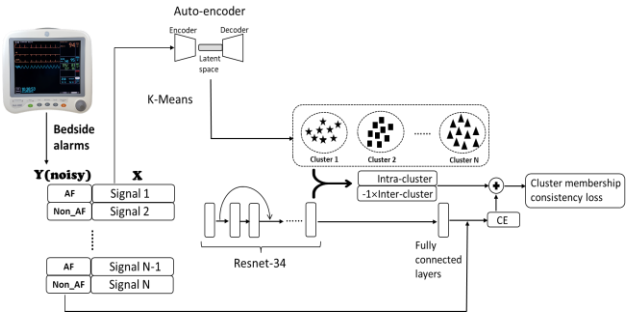


Figure 1. The workflow of CMC. First, cluster label is assigned to each sample by a pre-trained autoencoder. Then, the intra-cluster distance and inter-cluster distance of a single batch are computed based on these cluster labels. These distances are incorporated as additional loss terms, along with the conventional cross-entropy.

The contributions of this study include 1) a new approach to generate large labeled datasets using the concurrent alarm data; 2) the introduction of the CMC loss for robust learning, and 3) the demonstration of the proposed approach's effectiveness in handling label noise and poor-quality signals. This work provides valuable insights for PPG-based AF detection and has the potential for broader applications in deep learning.

2. Methods

2.1. Patient data

A comprehensive collection of physiological signals, including Electrocardiography (ECG) and PPG, at an academic medical centre in California, including cardiac arrhythmia alarms, and associated electronic health records (EHR) was conducted on a population of over 24,100 patients. Among these patients, a total of 2,834 individuals in the ICU were identified to have AF alarms from the bedside patient monitor. Moreover, out of these 2,834 patients, 1,650 had a well-documented ICD-10 code specifically indicating the presence of AF (coded as I48).

In accordance with our research methodology, we use 30 seconds for each photoplethysmography (PPG) signal, specifically centered at the start time of an AF alarm. This data acquisition approach is visually represented in Fig. 2. Similarly, for premature ventricular contraction (PVC) samples, we follow a similar procedure, with the exception that PVC alarms accompanied by AF alarms within a 30-second window from their onset are excluded from our dataset. Given that there is no explicit alarm for normal sinus rhythm (NSR), we define the existence of one 'NSR alarm' when the time interval between two consecutive alarms exceeds 30 seconds.

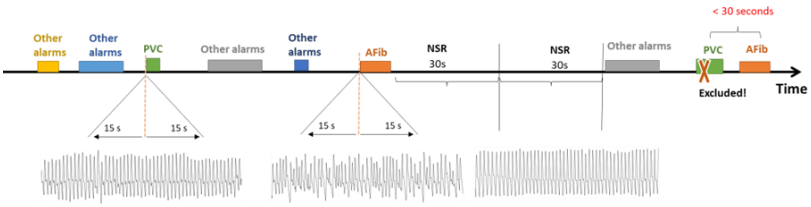


Figure 2. The illustration of the algorithmic labeling process for three classes. Each time an alarm is triggered, we store 15 seconds of data both before and after the trigger. Note the presence of normal PPG signals (labeled NSR) as well as the presence of many different alarms (VFIB, VT, PVC), and motion artifacts (labeled "artifacts"). On the right, we show that when an AF alarm is triggered within 30 seconds of a PVC alarm, we exclude the PVC alarm, because PVC and AF are difficult to tell apart.

Employing this specific criterion, we obtain a total of 4,070,350 30-second PPG segments originating from 2,985 patients who experienced AF. In contrast, for non-AF samples, we randomly select 2,436,460 NSR samples and 2,157,599 PVC samples from a pool of 24,100 patients, some of whom also exhibited AF alarms. Consequently, the sizes of the AF and non-AF sample sets are approximately balanced.

We selected the Stanford dataset as the test set, which is openly accessible, has been detailed in the publication referenced as [5]. In this study, the authors enlisted a total of 107 patients who had been clinically diagnosed with atrial fibrillation (AF), with a mean age of 68. Additionally, an additional 15 patients, with a mean age of 67, presenting paroxysmal AF, were also included in the dataset. Furthermore, data obtained from 41 healthy participants, undergoing exercise stress tests, with a mean age of 56, were incorporated as well.

2.3. Cluster membership consistency

Our proposed algorithm is founded on the principle that data sharing similar latent features should exhibit similar labels. Essentially, we assume that signals closely situated on the manifold of realistic signals possess smoothness properties. To achieve this, we utilize an autoencoder to map the manifold of real signals to a space where distances along the manifold can be measured as Euclidean distances.

Initially, we employ an unsupervised learning approach to cluster the training samples into a finite number of clusters. This unsupervised approach exclusively utilizes the PPG signals and disregards the labels, ensuring immunity to any label noise. The autoencoder compresses the signals, encoding the information into a space where Euclidean distances become meaningful measurements along the manifold of realistic signals.

Subsequently, during the supervised learning phase, we utilize the learned cluster membership to encourage small distances between points within the same cluster and large distances between points in different clusters. To achieve this, we augment the cross-entropy loss with two terms that constitute the cluster membership consistency loss.:

$$L_{intra} = \sum_{c=1}^{K}\left(\sum_{i=1}^{N_c}\sum_{j\neq i}^{N_c}\left\|\boldsymbol{F}(x_i^c;\theta) - \boldsymbol{F}(x_j^c;\theta)\right\|\right) \quad (1)$$

$$L_{inter} = -\sum_{c_1=1}^{K}\sum_{c_2\neq c_1}^{K}\left(\sum_{i=1}^{N_{c_1}}\sum_{j=1}^{N_{c_2}}\left\|\boldsymbol{F}(x_i^{c_1};\theta) - \boldsymbol{F}(x_j^{c_2};\theta)\right\|\right) \quad (2)$$

where the first term is the sum of distances among intra-cluster samples and the second term is the sum of distances among inter-cluster samples. Distances are computed both in a latent space being learned with defined by a function $\boldsymbol{F}$ that is parameterized by θ. This framework is flexible and yet has a small distance is easy to compute for any neural architecture and requires little extra computational cost for each training epoch.

The final loss is a weighted combination of the cross-entropy loss and two new CMC loss terms,

$$L = L_{CE} + \lambda_1 L_{intra} + \lambda_2 L_{inter} \quad (3)$$

are the hyperparameters selected by grid search through cross-validation.

3. Results

3.1. AF detection performance

As the key investigation in this study, we assess the efficacy of the proposed CMC loss against co-teaching [6], Early-Learning Regularization (ELR) [7], DivideMix [8], Iterative Noisy Cross-Validation (INCV) [9], and Sparse Over-Parameterization (SOP) [10], which are top runners in the 1st Learning and Mining with Noisy Labels Challenge at http://competition.noisylabels.com/. We tested each method on PPG data with both high and poor-quality subgroups, in addition to the entire dataset. We used AUROC (Area under the ROC Curve) as the metric for AF evaluation.

Table 1. Performance comparison between the proposed method with state-of-the-art algorithms on the test set.

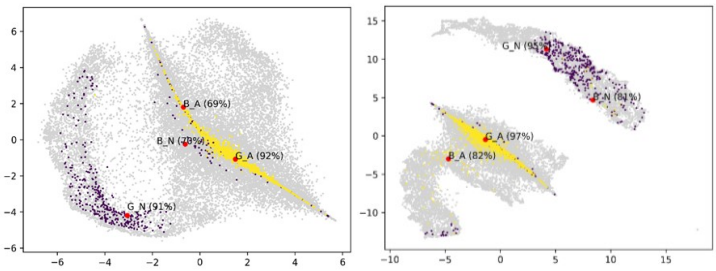
Methods	whole dataset	bad quality	good quality
CE	0.5853	0.5277	0.9051
SCE	0.5859	0.5359	0.7429
Co-teaching	0.5448	0.4626	0.8217
INCV	0.6082	0.5472	0.9245
DivideMix	**0.7437**	**0.6983**	0.9631
ELR	0.503	0.4807	0.8266
SOP	0.6633	0.5752	0.9614
CMC-2	0.7277	0.6779	0.9614
CMC-6	0.7416	0.6847	**0.9724**

The performance of the AF detection on the Stanford dataset is summarized in table 1, comparing seven baseline algorithms with the proposed CMC method using 2 and 6 clusters. The comparison across the two signal quality subgroups shows consistently better performance on the good quality subgroup compared to the bad quality subgroup. With all three groups, CMC produces either the best or the second-best performance among testing scenarios. Notably, CMC-6 demonstrates a strong performance on good quality signals and ranks second to DivideMix on bad quality signals.

3.2. Explore CMC latent space

We also explore the distribution of latent space learnt with CMC loss. We randomly selected 4 samples with correct labels, one sample each from each category - good quality AF, good quality Non-AF, bad quality AF and bad quality Non-AF, and generated their top-100 nearest neighbors from the training data based on the representations learnt from last convolutional layer. Then we manually annotate the selected neighbors to see the true label rate, as shown in Fig. 3. To obtain a big picture of the data distribution, we randomly selected 20,000 samples from the training data (grey dots), and 2,000 samples (yellow-AF, purple-Non-AF) from a clean-labelled dataset from our previous study.

As illustrated in Fig.3, we can observe that the true label rate for the selected 4 samples in CMC loss are all higher than in the CE loss, which shows that CMC will help gather the samples with same label together. Also, the true label rates of the two bad quality signals are lower than good quality signals. Moreover, for the selected bad quality Non-AF signal, CE is not able to separate it with the AF samples clearly, while CMC loss has push it far away from the AF samples cluster.



a. latent space from CE b. latent space from CMC

Figure 3. The distribution of latent space for CE and CMC loss. The latent space are calculated based the representation learnt from last convolutional layer. 100 neighbors (from training data) of each sample signal are selected and manually annotated for AF/Non-AF.

4. Discussion

Our study utilized cardiac arrhythmia alarms generated by bedside patient monitors to create the largest dataset for training PPG-based AF detection models. This extensive dataset consists of over 8 million 30-second PPG segments obtained from more than 24,000 hospitalized patients. To optimize learning from this dataset despite the presence of noisy labels, we propose a novel approach. Initially, we employ unsupervised learning to cluster the training samples, using the resulting cluster membership information to regulate the latent representation of PPG. This innovative technique significantly improves model performance and achieves results that are superior or comparable to five state-of-the-art robust learning algorithms designed for noisy labels. Moreover, our approach maintains a computational efficiency advantage.

Our experiments demonstrate the effectiveness of combining a large-scale auto-labeled dataset with our

robust learning approach for PPG AF detection. The key finding of our study is that enforcing cluster membership consistency can mitigate the impact of label errors in real-world datasets. Our method, alongside DivideMix, outperforms other algorithms, likely due to the incorporation of unsupervised information in both approaches. In our Cluster Membership Consistency (CMC) method, we derive cluster membership information through an unsupervised learning process before proceeding to supervised learning with noisy labels. DivideMix addresses the label noise issue as a semi-supervised problem, dividing the training data into clean and noisy subsets and re-labeling the noisy subsets by leveraging consensus from two pre-trained networks.

While ELR and SOP were top performers in a recent robust learning competition in computer vision, they exhibited inferior performance compared to our method and DivideMix in our study. This difference can be attributed to several factors. Firstly, ELR and SOP demonstrated superiority in experiments using simulated label errors, while our study incorporated real-world noise. Additionally, the scale of our dataset and model complexity may not satisfy the optimal conditions required for the SOP algorithm. Secondly, poor data quality may have influenced the results. Images in the datasets used in ELR and SOP studies likely possess a lower signal-to-noise ratio compared to PPG signals. ELR relies on loss calculation during training to assess the early learning stage, which can be unreliable due to noise in the data itself, thereby affecting signal quality. Our results challenge the assumption that early stopping effectively prevents overfitting caused by label noise, particularly in the presence of poor data quality. Lastly, our utilization of an autoencoder as a preliminary step for clustering contributes to our success. The autoencoder learns a compressed representation that can be reconstructed to the original input, effectively eliminating certain signal artifacts during the recovery process. The learned representation in our study demonstrates resilience to poor signal quality to some extent, potentially explaining our superior performance on subsets with subpar data quality.

5. Conclusion

In this study, we utilized cardiac arrhythmia alarms from patient monitors to create the largest dataset for training PPG-based AF detection models. The dataset consists of over 8 million 30-second PPG segments from 24,000+ hospitalized patients. We also introduce a novel approach that improves learning from this dataset by employing unsupervised clustering and regularization. Our approach achieves superior results compared to robust learning algorithms for noisy labels while maintaining computational efficiency. This demonstrates the effectiveness of combining a large auto-labeled dataset with our proposed approach for PPG AF detection.

Acknowledgments

This work was supported by NHLBI award under Grant R01HL166233.

References

[1] D. Ko, F. Rahman, R. B. Schnabel, X. Yin, E. J. Benjamin, and I. E. Christophersen, "Atrial fibrillation in women: epidemiology, pathophysiology, presentation, and prognosis," 2016, doi: 10.1038/nrcardio.2016.45.

[2] Y. Shen, M. Voisin, A. Aliamiri, A. Avati, A. Hannun, and A. Ng, "Ambulatory atrial fibrillation monitoring using wearable photoplethysmography with deep learning," in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery \& Data Mining*, 2019, pp. 1909–1916.

[3] C. Ding *et al.*, "Log-Spectral Matching GAN: PPG-based Atrial Fibrillation Detection can be Enhanced by GAN-based Data Augmentation with Integration of Spectral Loss; Log-Spectral Matching GAN: PPG-based Atrial Fibrillation Detection can be Enhanced by GAN-based Data Augmentation with Integration of Spectral Loss." [Online]. Available: https://github.com/chengding0713/Log-Spectral-matching-GAN.

[4] T. Pereira *et al.*, "Photoplethysmography based atrial fibrillation detection: a review," *npj Digital Medicine*, vol. 3, no. 1. Nature Research, Dec. 01, 2020, doi: 10.1038/s41746-019-0207-9.

[5] J. Torres-Soto and E. A. Ashley, "Multi-task deep learning for cardiac rhythm detection in wearable devices," *NPJ Digit. Med.*, vol. 3, no. 1, pp. 1–8, 2020.

[6] B. Han *et al.*, "Co-teaching: Robust Training of Deep Neural Networks with Extremely Noisy Labels." [Online]. Available: https://github.com/bhanML/Co-teaching.

[7] S. Liu, J. Niles-Weed, N. Razavian, and C. Fernandez-Granda, "Early-Learning Regularization Prevents Memorization of Noisy Labels."

[8] J. Li, R. Socher, and S. C. H. Hoi, "DIVIDEMIX: LEARNING WITH NOISY LABELS AS SEMI-SUPERVISED LEARNING." [Online]. Available: https://github.com/LiJunnan1992/DivideMix.

[9] P. Chen, B. Liao, G. Chen, and S. Zhang, "Understanding and utilizing deep neural networks trained with noisy labels," *36th Int. Conf. Mach. Learn. ICML 2019*, vol. 2019-June, pp. 1833–1841, 2019.

[10] S. Liu, Z. Zhu, Q. Qu, and C. You, "Robust Training under Label Noise by Over-parameterization." [Online]. Available: https://github.com/shengliu66/SOP.

Address for correspondence:

Xiao Hu, PhD
School of Nursing, Emory University
Room 207, 1520 Clifton Rd, Atlanta, GA 30322
Email address: xiao.hu@emory.edu