Defending Against Adversarial Attack in ECG Classification With Adversarial Distillation Training

Jiahao Shao
Tsinghua University
Shijia Geng
Zhaoji Fu
University of Science and Technology of China
Weilun Xu
HeartRhythm Medical
Tong Liu
Tianjin Medical University
Shenda Hong (hongshenda@pku.edu.cn)
National Institute of Health Data Science, Peking University https://orcid.org/0000-0001-7521-5127

Article

Keywords: deep learning, electrocardiograms, adversarial training, distillation, adversarial attack

Posted Date: April 7th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1522131/v1

DEFENDING AGAINST ADVERSARIAL ATTACK IN ECG CLASSIFICATION WITH ADVERSARIAL DISTILLATION TRAINING

Jiahao Shao
Department of Industrial Engineering
Tsinghua University
Beijing, China 100084
shaojh19@mails.tsinghua.edu.cn

Shijia Geng
HeartVoice Medical Technology
Hefei, China 230088
gengshijia@heartvoice.com.cn

Zhaoji Fu
School of Management
University of Science and Technology of China
Hefei, China 230026,
and HeartVoice Medical Technology
Hefei, China 230088
fuzj@mail.ustc.edu.cn

Weilun Xu
HeartRhythm Medical
Beijing, China 100020
xuweilun@heartrhythm.cn

Tong Liu
Department of Cardiology
Tianjin Institute of Cardiology
Second Hospital of Tianjin Medical University
Tianjin, China 300210
liutongdoc@126.com

Shenda Hong *
National Institute of Health Data Science
Peking University,
and Institute of Medical Technology
Health Science Center of Peking University
Beijing, China 100191
hongshenda@pku.edu.cn

ABSTRACT

In clinics, doctors rely on electrocardiograms (ECGs) to assess severe cardiac disorders. Owing to the development of technology and the increase in health awareness, ECG signals are currently obtained by using medical and commercial devices. Deep neural networks (DNNs) can be used to analyze these signals because of their high accuracy rate. However, researchers have found that adversarial attacks can significantly reduce the accuracy of DNNs. Studies have been conducted to defend ECG-based DNNs against traditional adversarial attacks, such as projected gradient descent (PGD), and smooth adversarial perturbation (SAP) which targets ECG classification; however, to the best of our knowledge, no study has completely explored the defense against adversarial attacks targeting ECG classification. Thus, we did different experiments to explore the effects of defense methods against white-box adversarial attack and black-box adversarial attack targeting ECG classification, and we found that some common defense methods performed well against these attacks. Besides, we proposed a new defense method based on adversarial distillation training (named CardioDefense) which comes from defensive distillation and can effectively improve the generalization performance of DNNs. The results show that our method performed more effectively against adversarial attacks targeting on ECG classification than the other baseline methods, namely, adversarial training, defensive distillation, Jacob regularization, and noise-to-signal ratio regularization. Furthermore, we found that our method performed better against PGD attacks with low noise levels, which means that our method has stronger robustness.

Keywords deep learning · electrocardiograms · adversarial training · distillation · adversarial attack

*Corresponding author.

1 Introduction

Electrocardiograms (ECGs) are widely used by clinicians to diagnose a range of cardiovascular diseases, which are the leading cause of death worldwide Mc Namara et al. [2019]. Owing to the development of technology and the increase in people's awareness to health, many companies have developed wearable devices that can measure single-lead ECG signals, such as the Huawei Watch GT2 Pro ECG and Apple Watch Series 4, which are worn by millions of people. Using these wearable devices, people can detect whether they have cardiovascular diseases before the disease becomes severe. However, it is impossible for clinicians to spend a considerable amount of time analyzing the large amount of ECG signals collected by these devices.

Deep neural networks (DNNs) are an economic alternative approach for classifying multi-lead ECG signals Ribeiro et al. [2020], Xu et al. [2018], Jain et al. [2020], single-lead ECG signals Lai et al. [2020], and even ECG images Sangha et al. [2022]. In addition, owing to the development of this technology, the accuracy of DNNs is comparable to that of professional cardiologists Hannun et al. [2019], Sinnecker [2020], Elul et al. [2021]. DNNs have been successfully used in many ECG analysis tasks Hong et al. [2020a], Somani et al. [2021], such as cardiovascular disease management Siontis et al. [2021], Fu et al. [2021], disease detection Attia et al. [2019], Erdenebayar et al. [2019], Ribeiro et al. [2020], mortality prediction Lima et al. [2021], Raghunath et al. [2020], Hong et al. [2020b], sleep staging Banluesombatkul et al. [2020], biometric human identification Labati et al. [2019], Hong et al. [2020c], and ECG-based non-invasive monitoring of blood glucose Li et al. [2021], indicating the effectiveness of DNNs in ECG analysis Hughes et al. [2021].

However, DNNs are vulnerable when facing adversarial noises involving perturbations that are imperceptible to the human eye. This phenomenon was first discovered by Szegedy et al. Szegedy et al. [2013] in the image classification field. Subsequently, researchers proposed certain convenient methods for generating adversarial perturbations, such as fast gradient sign method Goodfellow et al. [2014], basic iterative method Kurakin et al. [2016], projected gradient descent (PGD) Madry et al. [2017], and Carlini and Wagner (C&W) attacks Carlini and Wagner [2017]. These methods are mainly aimed at attacking DNNs for image classification, and they cannot be extended directly to DNNs for ECG signals, because the perturbations created by these methods are not physiologically plausible Han et al. [2020]. To attack DNNs for ECG signal classification, several white-box and black-box adversarial attack methods have been proposed recently. The white-box adversarial attack is generated by utilizing the inner structure knowledge of the target DNN, whereas the black-box adversarial attack does not have any knowledge regarding the network's inner structure. The white-box attack methods proposed by Han et al. Han et al. [2020] and Chen et al. Chen et al. [2020] are similar to PGD and C&W attacks, respectively. The only difference is that the perturbations created by smooth adversarial perturbation (SAP) proposed by Han et al. are smoothed through convolution, whereas those created by the attack method of Chen et al. are significantly limited by setting up an objective function to maximize the smoothness of the attack. Detecting the perturbations becomes difficult because of the restriction of the objective function or convolution processing. Lam et al. Lam et al. [2020] proposed a black-box attack called boundary attack, which improves the smoothness of perturbations by using a low-pass Hanning filter. In Figure 1, we plot a part of an original ECG signal sample and its counterparts that are attacked by PGD and SAP. We can see that the signal attacked by PGD is unnatural and not physiologically plausible, but it is difficult to distinguish the signal attacked by SAP from natural ECG signals.

To defend against adversarial attacks in ECG signal classification, Yang et al. Yang et al. [2020] applied the gradient-free trained sign activation neural network to classify ECG signals and found that the perturbations created by the HopSkipJump boundary-based black-box attack can fool the classification network and are visually distinguishable. Furthermore, because the network is gradient-free and white-box attacks mainly use gradient information to create adversarial attacks, the network is immune to traditional white-box attacks. Ma and Liang Ma and Liang [2020a] explored the effectiveness of three defense methods against PGD and SAP attacks, namely, adversarial training, Jacobian regularization (JR), and noise-to-signal ratio (NSR) regularization. The results showed that all three methods can improve the robustness of the DNNs for ECG classification against PGD and SAP attacks, and NSR has the best performance among these defense methods. However, both Yang et al. and Linhai et al. didn't completely explore the defense against the white-box and black-box adversarial attacks. In addition, the accuracy of the gradient-free trained sign activation network proposed by Yang et al. Yang et al. [2020] is lower than that of traditional DNNs on certain data, and it can be achieved or surpassed by traditional DNNs using certain defense methods.

In this study, we explored the defense against the white-box and black-box adversarial attacks which are aimed at ECG-based DNNs. Furthermore, SAP Han et al. [2020] is applied to represent the white-box attack and boundary attack Lam et al. [2020] is applied to represent the black-box attack. We defended ECG-based DNNs against SAP and boundary attack with common defense methods, such as adversarial trainingGoodfellow et al. [2014], defensive distillationPapernot et al. [2016], JRJakubovitz and Giryes [2018] and NSR regularizationMa and Liang [2020b], and found these methods performed well against SAP and boundary attack. Furthermore, defensive distillation can learn class-related knowledge, and transfer the knowledge from the first network to the second network to generalize the

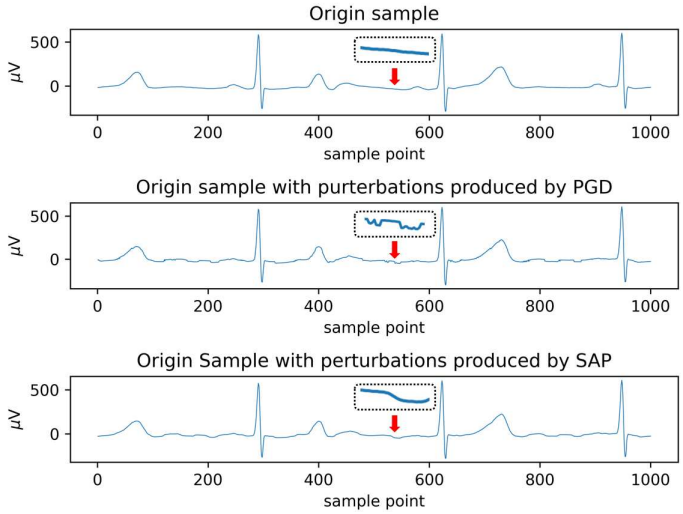

Figure 1: Comparison Between an Original ECG signal and That Attacked by PGD and SAP

classification ability. While SAP and boundary attack make adversarial ECG samples by adding small perturbations into original ECG samples, so that DNNs classify adversarial ECG samples into wrong categories, and those perturbations are so small that it is difficult for people to distinguish those adversarial samples. Therefore, we can regard those small perturbations as reasonable fluctuations of ECG signals, and add the ECG samples with those small perturbations into the training set of distillation network, which can make distillation network to further learn fluctuations of ECG signals and class-related knowledge to improve generalization ability. Based on this idea, we proposed a new method based on adversarial distillation training (named **CardioDefense**) in which we added adversarial samples into the training process of defensive distillation. The results show that CardioDefense outperforms JR, NSR regularization, defensive distillation and adversarial training under SAP and boundary attack.

Furthermore, although perturbations created by traditional adversarial attacks, such as PGD attacks, are not physiologically plausible, if we keep the level of noise, ϵ, low for PGD attacks, we will obtain fewer unnatural ECG segments. It is possible for hackers to attack ECG signals collected by wearable devices with low-noise PGD attacks because clinicians do not check these signals and people in general may not be able to recognize the attacked patterns. In addition, we can consider low-noise PGD attacks as a robust test for CardioDefense. Thus, we use the trained DNNs with different defense methods to classify ECG signals attacked by low-noise PGD attacks, and the results show that CardioDefense still performs well against this type of attack.

2 Results

2.1 Experimental model and data

We applied the 13-layer convolution network Goodfellow et al. [2018] as the base classification DNN model, which is one of the top-tier models in the 2017 PhysioNet/CinC Challenge. There are two kinds of defense methods, the first part consists of four traditional defense methods, including JRJakubovitz and Giryes [2018], NSR regularizationMa and Liang [2020b], adversarial training (**AT**)Goodfellow et al. [2014] and defensive distillation (**DD**)Papernot et al. [2016], and the second part is the method we proposed, CardioDefense, as well as its variants, including Init-CardioDefense (**Init**) and Dist-CardioDefense (**Dist**). The specific introduction and settings are described in Methods. Notably, we do not develop different classification DNNs for different defense methods. Our classification model is always a 13-layer convolution network Goodfellow et al. [2018], and these defense methods are only special settings to enhance the robustness of the classification DNN.

The experimental data are obtained from the publicly available training dataset of the 2017 PhysioNet/CinC Challenge Clifford et al. [2017]. All these ECG signals are single-lead, and their lengths are approximately 9–61 s. The available dataset contains 8528 ECG signals, including 5076 normal ECGs, 758 atrial fibrillation (AF) ECGs, 2415 other ECGs, and 279 noise samples. It is obvious that there is a severe category imbalance problem in the data, and to solve the problem, we duplicate the noise samples five times and double the size of the AF ECG samples.

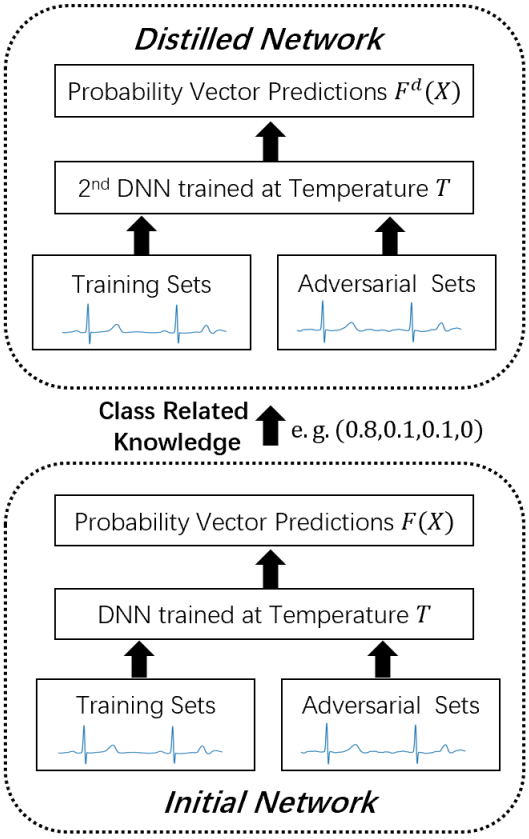

Figure 2: Adversarial Distillation Training Process of CardioDefense.

Because the length of the data is not fixed and it is not suitable for the training of the classification network, we first limit the length of the ECG signals to 9000 sampling points. For ECG signals with less than 9000 sampling points, we fill the same number of zeros on both sides of the data. For ECG signals with more than 9000 sampling points, we only take the first 9000 data points. During the process of training the DNN, the dataset after data expansion and length limitation is divided into two parts: 90% for the training set and the remaining 10% for the test set.

2.2 Defense Effects against SAP Attacks

Here, we first attacked original test samples by SAP($t = 20$, $t' = 40$) based on trained DNNs without defense methods, and then, we used each trained DNN with defense methods to classify these attacked test samples. Table 1 shows the average prediction accuracy, F1-score, decline ratio and corresponding standard deviation of five trained DNNs with each defense method in this situation. However, the attacker may know what defense methods we use, then train a similar model and attack it. Therefore, we are curious about the robustness of the trained DNNs with defense methods when the attacker uses themselves to make adversarial samples. In this paper, the former situation is called situation I, and the latter situation is called situation II. Table 2 shows the performance of the trained DNNs with defense methods in situation II.

Table 1 and Table 2 show that under SAP attacks, the accuracy of the classification DNN is significantly reduced (a decrease of about 50%) and the F1-score is reduced to 28%. From Table 1, we can see that all defense methods performed well in situation I, even JR which performed worst can control the decline ratio at 11%, and keep the accuracy ratio and F1-score at 0.7676, 0.6772 respectively. While in situation II, the performance of some defense methods was not good, including JR and DD, and the trained DNNs with JR even had a worse performance than the trained DNNs without any defense method. In these two situations, our proposed method, CardioDefense at $T = 1$, had the best defense performance, and it controlled the decline of model accuracy within the range of less than 1% in situation I and 5% in situation II. Besides, CardioDefense at $T = 1$ also had a good performance in F1-score, and it even increased the F1-score of the classification model a little in situation I.

Table 1: Comparison of Methods under SAP attack in Situation I

		Accuracy (Performance Drop)	f_1 score (Performance Drop)
	No Defense	0.4256±0.0727 (50.69%±8.42%)	0.2839±0.0656 (63.33%±8.39%)
Baselines	JR	0.7676±0.0270 (11.03%±3.16%)	0.6772±0.0281 (12.69%±3.06%)
	NSR	0.8359±0.0094 (3.39%±0.71%)	0.7200±0.0190 (5.75%±1.62%)
	DD	0.7684±0.0110 (11.36%±1.11%)	0.6475±0.0128 (15.85%±1.24%)
	AT	0.8551±0.0025 (1.03%±0.38%)	0.7498±0.0120 (2.11%±1.70%)
Proposed	Init-CardioDefense	0.8530±0.0044 (2.02%±0.72%)	0.7572±0.0114 (4.03%±1.74%)
	Dist-CardioDefense	0.8579±0.0082 (1.00%±0.53%)	0.7628±0.0135 (1.43%±0.80%)
	CardioDefense	**0.8631±0.0065 (0.67%±0.43%)**	**0.7737±0.0145 (-1.02%±1.84%)**

Table 2: Comparison of Methods under SAP Attacks in Situation II

		Accuracy (Performance Drop)	f_1 score (Performance Drop)
	No Defense	0.4256±0.0727 (50.69%±8.42%)	0.2839±0.0656 (63.35%±8.47%)
Baselines	JR	0.4223±0.0665 (51.08%±7.70%)	0.2958±0.0343 (61.82%±4.43%)
	NSR	0.7339±0.0161 (15.43%±1.89%)	0.5323±0.0342 (31.28%±4.42%)
	DD	0.5079±0.0246 (41.84%±2.89%)	0.3397±0.0249 (56.15%±3.22%)
	AT	0.8040±0.0065 (7.01%±0.67%)	0.6477±0.0163 (16.39%±2.10%)
Proposed	Init-CardioDefense	0.7656±0.0114 (11.30%±1.32%)	0.6035±0.0193 (22.09%±2.50%)
	Dist-CardioDefense	0.8148±0.0082 (5.60%±0.95%)	0.6817±0.0101 (12.00%±1.30%)
	CardioDefense	**0.8270±0.0046 (4.35%±0.36%)**	**0.6845±0.0127 (11.63%±1.64%)**

Furthermore, we can see that NSR regularization performs better than JR under SAP attacks and can maintain the accuracy of the classification model above 80% in situation I and above 70% in situation II, indicating that the defense effect of NSR is good. In our experiments, adversarial training exhibits outstanding performance, making the classification model maintain an accuracy ratio above 80% and an F1-score of more than 60% in both situations under SAP attacks. In addition, the performances of the methods that add adversarial samples into the training process of only one network of defensive distillation are excellent, such as Dist-CardioDefense and Init-CardioDefense. It is easy to understand that Dist-CardioDefense has a better defense effect than Init-CardioDefense: adversarial samples are added into the training process of the second network of DD for Dist-CardioDefense, and it is the second network that is used to classify ECG samples, whereas Init-CardioDefense puts adversarial samples in the first network of DD, which does not take the task of classify data. DD behaved well in situation I and had a similar performance to JR. While it did not perform well like JR in situation II, with only 50.79% accuracy ratios and 33.97% F1-scores, which denotes that DD and JR could defend against adversarial samples targeted at the trained DNNs without any defense method, but they couldn't defend against those adversarial samples targeted at the trained DNNs with themselves, that is to say, their robustness is not strong; on the contrary, other defense methods have good robustness.

To explore the defense effects of these defense methods under different-smoothing-degree SAP attacks, we changed the parameter t' of the SAP attack, which controls the convolution times, and set the parameter t' to $0, 10, 20, 30, 40$, respectively. The more convolution times, the smoother the attack noise becomes. As t' takes 0, SAP attack becomes PGD attack. The results in situation I and situation II are presented in Figure 3 and Figure 4, respectively. From these two figures, we can see that as the parameter t' changes from 0 to 10, the accuracy ratio and F1-score of the classification model with explored defense methods are improved, and the increase is smaller in situation I, but larger in situation II. However, as t' further increases, the accuracy ratio and F1-score do not change significantly, which means that the smoothness of the adversarial attack does affect the accuracy ratio and F1-score of the classification model with different defense methods. After 10 convolutions, the smoothness of the adversarial attack is high and does not change significantly with more convolution operations, which leads that the accuracy ratio and F1-score of the classification model with different methods do not change significantly. Besides, we can see that the trained DNNs with CardioDefense always had the best performance, and the order of defense effects for these defense methods remains the same under different convolution times.

2.3 Defense Effects against PGD Attacks

Similarly, we attacked each trained DNN with or without defense method using the PGD at $\epsilon = 10, t = 20, t' = 0$. The mean value with the standard deviation of the prediction accuracy and the F1-score for each defense method under

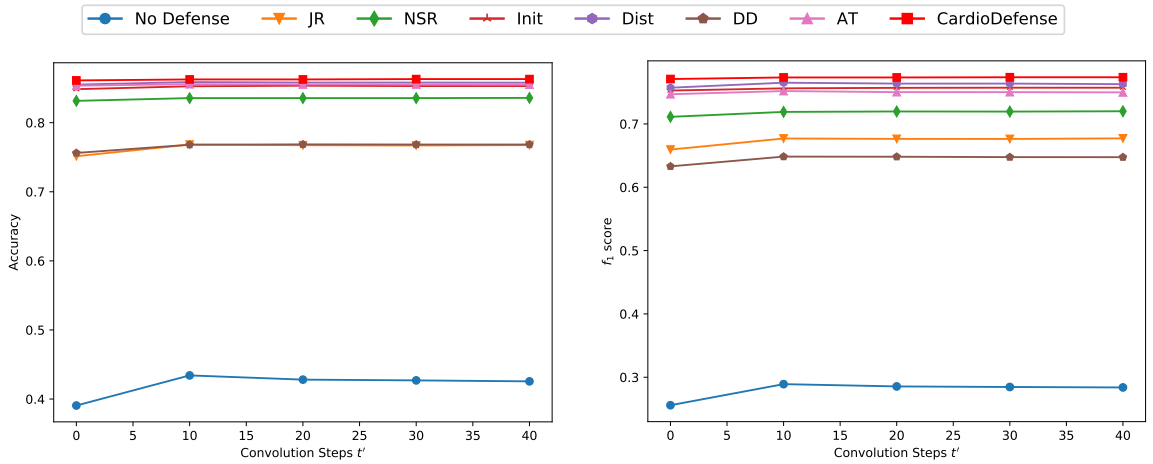

Figure 3: Performance of the Compared Methods Attacked by SAP under Different Convolution Steps in Situation I.

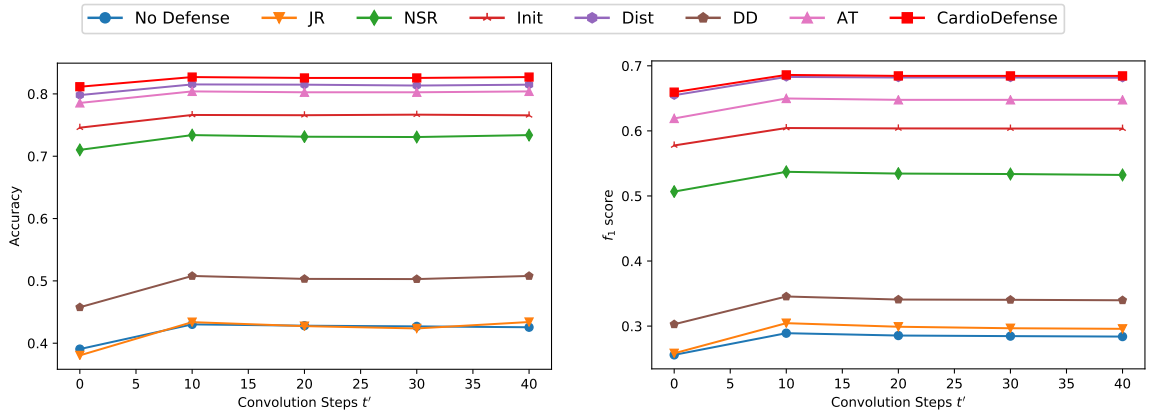

Figure 4: Performance of the Compared Methods Attacked by SAP under Different Convolution Steps in Situation II.

a PGD attack are shown in Table 3 and Table 4. From these tables, we can see that the accuracy and F1-score of the classification DNN with no defense are reduced to 39% and 25%, respectively, and the decline of these two metrics is more than 50%. The classification DNN with CardioDefense at $T = 1$ still had the best performance, and its accuracy ratio against PGD attacks is still over 80% in two situations, and its decline ratio of accuracy is lower than 1% in situation I. In addition, we can see that adversarial training had good performance against low noise level PGD attacks, and NSR behaved better than JR under PGD attacks, which is consistent with the results of Ma and Liang [2020a]. Furthermore, the performance of the trained DNNs with defense methods in situation I is better that that in situation II, and the performance of the trained DNNs with CardioDefense, Init-CardioDefense, Dist-CardioDefense, AT as well as NSR is more stable than that of the trained DNNs with JR and DD. In addition, the order of defense effects against low-noise PGD attacks is the same as that for SAP attacks in two different situations.

To further explore the performance of the explored defense methods under PGD attacks at different noise levels, we changed the parameter ϵ. Specifically, we set t' and t as 0 and 20, respectively, and parameter ϵ as $5, 10, 15, 20, 25$, and the corresponding results for two situations are shown in Figure 5 and Figure 6. We can see that the accuracy ratio and F1-score of the trained DNNs with different defense methods decreased with the increase in noise level for PGD attack in the beginning, and remained fixed until the end of our experiments in two situations. Moreover, the decrease of accuracy ratio and F1-score for the trained DNNs with defense methods in situation I is lower than that in situation II. Furthermore, these figures show that the trained DNNs with CardioDefense exhibited the best defense effects, indicating that although faced with PGD attacks within a certain low-level noise range, our method still has high defense effects. That is to say, our method has good robustness.

Table 3: Comparison of Methods under PGD Attacks in Situation I

		Accuracy (Performance Drop)	f_1 score (Performance Drop)
	No Defense	0.3906±0.0695 (54.80%±8.08%)	0.2558±0.0555 (66.94%±7.16%)
Baselines	JR	0.7515±0.0272 (12.91%±3.14%)	0.6595±0.0276 (14.96%±3.00%)
	NSR	0.8316±0.0118 (3.88%±0.98%)	0.7112±0.0232 (6.92%±1.92%)
	DD	0.7562±0.0109 (12.76%±1.09%)	0.6330±0.0117 (17.73%±1.06%)
	AT	0.8535±0.0030 (1.22%±0.45%)	0.7469±0.0153 (2.49%±2.19%)
Proposed	Init-CardioDefense	0.8485±0.0045 (2.53%±0.76%)	0.7527±0.0114 (4.59%±1.83%)
	Dist-CardioDefense	0.8551±0.0064 (1.33%±0.43%)	0.7572±0.0104 (2.15%±0.54%)
	CardioDefense	**0.8612±0.0050 (0.89%±0.26%)**	**0.7709±0.0114 (-0.67%±1.77%)**

Table 4: Comparison of Methods under PGD Attacks in Situation II

		Accuracy (Performance Drop)	f_1 score (Performance Drop)
	No Defense	0.3906±0.0704 (54.74%±8.16%)	0.2558±0.0555 (66.97%±7.17%)
Baselines	JR	0.3805±0.0756 (55.91%±8.76%)	0.2582±0.0462 (66.67%±5.97%)
	NSR	0.7102±0.0158 (17.72%±1.83%)	0.5067±0.0270 (34.58%±3.49%)
	DD	0.4577±0.0261 (46.97%±3.02%)	0.3028±0.0246 (60.90%±3.18%)
	AT	0.7855±0.0039 (9.00%±0.46%)	0.6191±0.0157 (20.07%±2.03%)
Proposed	Init-CardioDefense	0.7458±0.0157 (13.59%±1.82%)	0.5776±0.0237 (25.44%±3.06%)
	Dist-CardioDefense	0.7981±0.0057 (7.53%±0.66%)	0.6548±0.0137 (15.46%±1.77%)
	CardioDefense	**0.8115±0.0036 (5.98%±0.42%)**	**0.6595±0.0144 (14.86%±1.86%)**

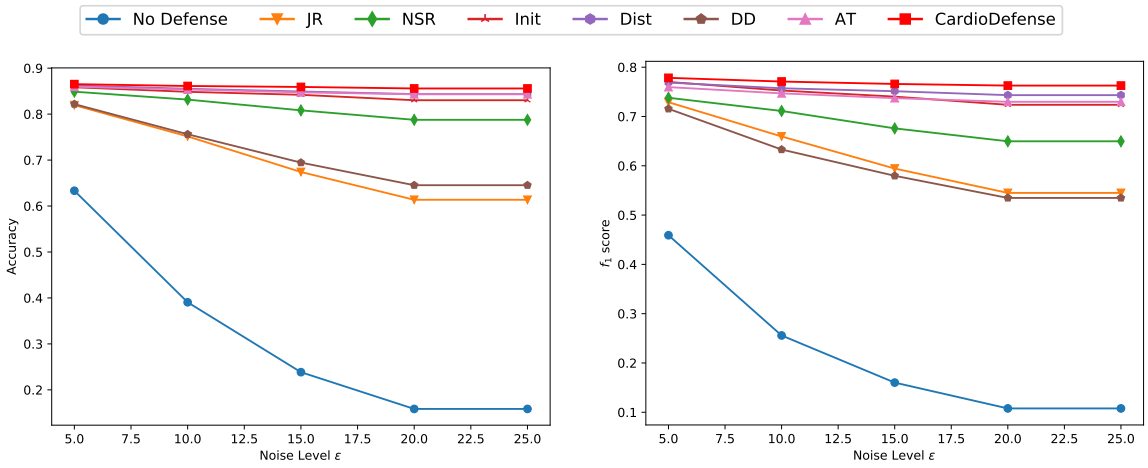

Figure 5: Performance of the Compared Methods Attacked by PGD under Different Noise Levels in Situation I.

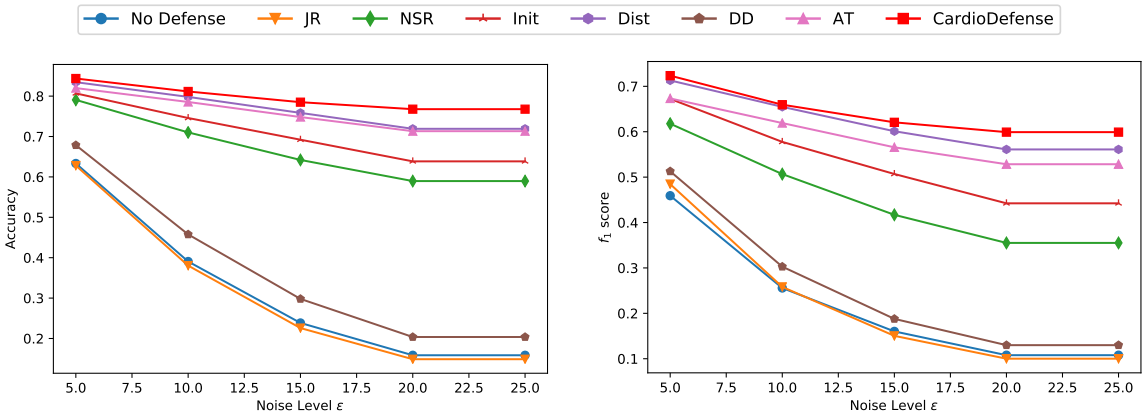

Figure 6: Performance of the Compared Methods Attacked by PGD under Different Noise Levels in Situation II.

Table 5: Comparison of Methods under Boundary Attacks

		Accuracy	f_1 score
Baselines	JR	0.8976±0.0298	0.8872±0.0375
	NSR	0.8824±0.0257	0.8614±0.0398
	DD	0.8877±0.0413	0.8702±0.0562
	AT	0.9014±0.0211	0.8694±0.0413
Proposed	Init-CardioDefense	**0.9163±0.0234**	**0.9023±0.0342**
	Dist-CardioDefense	0.9005±0.0315	0.8840±0.0513
	CardioDefense	0.9144±0.0238	0.8961±0.0304

2.4 Defense Effects against Boundary Attacks

Without knowing the structure of the classification model, boundary attack, one of target black-box attack, generates adversarial samples by adjusting the samples of the target category to keep a small difference from the samples to be attacked. With these adversarial samples, hackers can make the classifier produce errors that meet their expectations. Therefore, in order to comprehensively explore the effectiveness of these defense methods, we did experiments to explore the defense effects of these methods against boundary attack.

Specifically, because we have trained 5 classification DNNs without defense methods, we apply boundary attack to create adversarial samples based on these five DNNs and test data. Finally, we have created 291, 277, 300, 250 and 288 adversarial samples, respectively. Then, we use the trained DNNs with defense methods to classify these adversarial samples, and the results are showed as Table 5. It should be noted that we only explore the results in situation I here, because the goal of boundary attack is to create adversarial samples identified as target categories without considering the inner structure of the classification model, and if we attack the trained DNNs with each defense method by boundary attack, the generated adversarial samples will not be identified correctly by themselves. We can also see that all trained DNNs with explored defense methods had good performance, and CardioDefense also had excellent performance against boundary attack, which was close to the best one.

3 Discussions

From the results in situation I, we can see that although the adversarial samples created by SAP and boundary attack are hard to be distinguished from the natural ones directly by cliniciansHan et al. [2020], Lam et al. [2020], they are easily recognized by the trained DNNs with defense methods. Compared with the natural ECG signals, the corresponding adversarial signals created by SAP and boundary attack based on the trained DNNs without defense methods does not change dramatically, while the trained DNNs with defense methods are not sensitive to subtle changes, so the trained DNNs with defense methods performed well in situation I. However, if the adversarial signals are created based on the trained DNNs with defense methods like situation II, the performance of the trained DNNs with defense methods is not as good as that in situation I, which shows the robustness of different defense methods.

Specifically, the trained DNNs with CardioDefense have a high accuracy ratio and F1-score under SAP attack and boundary attack, and the defense effects of CardioDefense against SAP attack and boundary attack are better than many traditional defense methods, such as JR, NSR regularization, adversarial training, and defense distillation in different situations. At the same time, CardioDefense still performs well under low-noise PGD attacks, which have higher noise levels than SAP attacks. These phenomena show that our proposed model, CardioDefense, has better defense effects and stronger robustness.

In addition, the results show that adversarial training has good defense effects against SAP attack, boundary attack as well as PGD attack of low-level noise. In the training process of adversarial training, the classification model needs to classify the adversarial samples created by SAP, and it will be punished by a loss function if it classifies the adversarial samples mistakenly. At the same time, compared with the original ECG samples, the morphology of the adversarial samples created by SAP only has subtle changes, so it was not difficult for the classification model with adversarial training to learn the characteristics of the SAP. Due to the punishment mechanism and the characteristics of SAP that are easy to learn, the classification model with adversarial training is robust against SAP as well as boundary attack whose generated adversarial samples don't change dramatically compared with the original samples.

Furthermore, defensive distillation has much better defense effects in situation I than that in situation II, which can be concluded from the corresponding results and denotes that the robustness of defensive distillation is not good. Adversarial samples created by SAP are added into the training processes of both network of CardioDefense, the first network learns the morphological characteristics of the original ECG samples and adversarial samples, and then transmits this information to the second network of CardioDefense, which improves the generalization ability of the classification model. In addition, the second network still learns the characteristics of nature samples and adversarial samples, which further enhances the generalization and robustness of the model. These are the reasons why CardioDefense performs better than Init-CardioDefense and Dist-CardioDefense in most cases, and from the truth that Dist-CardioDefense performs better than Init-CardioDefense in situation II of SAP attack and PGD attack, we can see that the latter plays a more important role for the performance of CardioDefense.

4 Method

4.1 Defensive Distillation

To explain our proposed method, we first introduce its main body, defensive distillation. Initially, distillation learning was used exclusively to reduce the hierarchy of DNNs Hinton et al. [2015]. Generally, a large-scale DNN is first trained to learn the distribution of the sample data, and then, the labels of the training samples are changed to the probability of each category for the samples predicted by the large-scale DNN. Subsequently, a small-scale DNN network is trained with the changed training data without the loss of accuracy. The idea of this method is that the parameters learned by the DNN can represent the characteristics of data and the probability vector output from the softmax layer of the network contains some knowledge of the data. For example, if the probability values corresponding to two categories of samples predicted by a DNN are similar, it means that there are some similarities between these two types of data. Later, Nicolas et al. Papernot et al. [2016] proposed defensive distillation, in which the initial network is the same as the distilled network.

The core of defensive distillation is the temperature parameter, T, which is the only additional parameter compared with the ordinary DNN and is used in the normalization process of the softmax layer as follows.

$$F_i(X) = \frac{e^{z_i(X)/T}}{\sum_{l=0}^{N-1} e^{z_l(X)/T}}, i \in 0, 1, ..., N-1 \tag{1}$$

In the equation, $z_i(X)$ denotes the output logit of the last layer of the DNN for sample X corresponding to category i. For convenience, we use z_i to represent $z_i(X)$. Furthermore, $Z(X) = (z_0, z_1, ..., z_{N-1})$ denotes the output logit vector. The softmax layer normalizes the output logit vector, $Z(X)$, through Equation 1, and the result value, $F_i(X)$, denotes the probability that sample X belongs to category i. Here, we have $F(X) = (F_0(X), F_1(X), ..., F_{N-1}(X))$ as the output probability vector of the softmax layer for the DNN. We can see that the larger the T value is, the smaller the difference between the values of the probability vector, $F(X)$, becomes, and as $T \to \infty$, $F_i(X)$ converges to $\frac{1}{N}$. As $T = 1$, Equation 1 is the same as that of the traditional DNN.

Here, we use $Y(X)$, an indicator vector, to denote the labels of sample X, with the non-zero element in $Y(X)$ representing the correct category, and if X belongs to the first category, $Y(X)$ is like $(e.g., (1, 0, 0, ..., 0))$. Under defensive distillation, we need to train our classification DNN model twice, and both times, the model must be trained

from the beginning. For the first time, the loss function is as follows:

$$-\frac{1}{|\chi|}\sum_{X\in\chi}\sum_{i\in(0,\ldots,N-1)}Y_i(X)\log F_i(X), \tag{2}$$

where χ denotes the entire training set, and the goal is to minimize the loss function. Because $Y_i(X)$ has only one non-zero value, 1, Equation 2 can be changed as

$$-\frac{1}{|\chi|}\sum_{X\in\chi}\log F_l(X), \tag{3}$$

where l is the index of the correct label for sample X. For the second training, the labels for the training set are changed as the output probability vectors of the first trained DNN, which are called soft labels. The loss function is

$$-\frac{1}{|\chi|}\sum_{X\in\chi}\sum_{i\in(0,\ldots,N-1)}F_i(X)\log F_i^d(X). \tag{4}$$

The goal of the training is to minimize the loss function. Based on the study conducted by Nicolas et al. Papernot et al. [2016], we have the following.

$$\begin{aligned}\frac{\partial F_i(X)}{\partial X_j}\Big|_T &= \frac{\partial}{\partial X_j}\left(\frac{e^{z_i/T}}{\sum_{k=0}^{N-1}e^{z_k/T}}\right)\\ &= \frac{e^{z_i/T}}{Tg^2(X)}\left(\sum_{k=0}^{N-1}\left(\frac{\partial z_i}{\partial X_j}-\frac{\partial z_k}{\partial X_j}\right)e^{z_k/T}\right)\end{aligned}$$

This means that with an increase in T, the elements of the Jacobian matrix of F decrease. In other words, the gradients of our classification DNN decrease. Because gradients are used in SAP to create perturbations, it is more challenging for an attacker to create successful noises to fool the classification DNN as the gradients decrease. Therefore, defensive distillation makes our classification DNN model insensitive to small changes of the input ECG signals as the parameter T increases.

On the other side, for the training process of the first network of defensive distillation, the optimization mechanism is adjusting the parameter of the network to make $F(X)$ converge to $Y(X)$. This kind of training mechanism often makes DNNs with only one network over fit the training data. However, defensive distillation applies $F(X)$ output by the first network as the soft labels of training data for the second network, and the goal of optimization mechanism will make $F^d(X)$ converge to $F(X)$, but $F^d(X)$ can't be the same as $F(X)$ in practice, which improves the generalization ability of the second network.

4.2 CardioDefense

However, due to the limitation and contingency of nature ECG signals, it is difficult to learn class-related knowledge and data fluctuation well with only nature training data. Thus, we need to add some adversarial ECG samples created by SAP into the training process of defensive distillation, and we call the new adversarial defense method base on adversarial distillation training as CardioDefense.

Studies have shown that learning the characteristics of adversarial samples which is called adversarial training improves the robustness of the classification model. Szegedy et al. Szegedy et al. [2013] first found that DNNs are vulnerable to adversarial perturbations, and they used adversarial training to improve the robustness of the DNN. They found that it was better to use adversarial perturbations in the hidden layer. However, Goodfellow et al. Goodfellow et al. [2014] discovered that if the activation function of the hidden layer for the neural network is unbounded, such as the ReLU function, adding adversarial perturbations to the inputs is better. While it is time-consuming with adversarial samples created by PGD added into the training process of DNNs. To solve this problem, Shafahi et al. Shafahi et al. [2019] proposed a new training algorithm in which the gradient information is recycled to update the parameters of the network. In other words, every time the gradients of the adversarial samples are calculated, the network parameters are updated according to the gradients, and this process is repeated. In contrast, in traditional adversarial training, the network parameters are updated after the final calculation of the gradients of adversarial samples. In our study, the last perturbation generated by SAP is smoother than that generated in the intermediate process, and this can make the DNN find its real shortcomings. Thus, we use only the last perturbation generated by SAP instead of the large fluctuation one

generated in the intermediate process. Subsequently, we update the parameters of the DNN after the final calculation of the gradients of adversarial samples. Certain new adversarial training algorithms exist, such as max-margin adversarial training Ding et al. [2018] and increasing-margin adversarial training Ma and Liang [2020c], which are suitable for defending large perturbations. However, the noise level of SAP is not high, indicating that these new adversarial training methods are not suitable for defending SAP; therefore, we do not consider these new training algorithms.

The training process with adversarial samples as training data can be regarded as an optimization problem Madry et al. [2017], and the form is as follows.

$$\min_{\theta} E_{(x,y)\sim D}\left[\max_{\delta} L(\theta, x+\delta, y)\right] \tag{5}$$

The inner function describes the process of creating adversarial samples with δ representing the adversarial perturbations, and the target is to maximize the inner loss function. The outer function denotes the training process with adversarial samples, and the goal is to minimize the loss function by adjusting the value of the neural network parameters, θ. In this training process, the training set consists of only adversarial samples, which lack the training of the original samples. Therefore, we adopt the strategy with a mixture of adversarial samples and original samples as training data, and this is shown as follows.

$$\min_{\theta}\left(cE_{(x,y)}\left[L_{adv}(\theta, x_{adv}, y)\right] + (1-c)E_{(x,y)}\left[L(\theta, x, y)\right]\right) \tag{6}$$

Where x_{adv} denotes adversarial sample. In this way, we not only ensure the accuracy of the trained model on the natural samples, but also defend against attacks of adversarial samples by learning the characteristics of these two types of samples.

For adversarial attacks, if a hacker makes a classifier mistakenly identify a sample as a specified category, it is called a target attack. If the hacker makes the model classify a sample mistakenly without specifying the label of the error classification, it is called a non-target attack. In this study, we applied non-target SAP attack to create adversarial samples.

Furthermore, the first step to create adversarial samples for SAP is to create traditional adversarial samples using PGD. Generally, PGD creates adversarial samples by using multiple iterations and limits the difference between new adversarial samples and those created in the last iteration. We use $Clip_{x,\epsilon}(x')$ to represent limiting the maximum difference between x and x' to ϵ, where ϵ denotes the noise level, and the larger ϵ is, the greater the fluctuation of noise. We first set $x'_0 = x$; then, we have

$$x'_i = Clip_{x'_{i-1},\epsilon}(x'_{i-1} + \alpha sign(\bigtriangledown_{x'_{i-1}} L(f(x'_{i-1}, y)))) \tag{7}$$

After t iterations, traditional adversarial samples are created, and $x_{adv} = x'_t$. We define δ as the adversarial perturbation, which is the difference between the adversarial sample and the corresponding original sample. Then, SAP makes δ smooth through convolution, which is expressed as follows.

$$x_{adv}(\delta) = x + \frac{1}{m}\sum_{i}^{m} \delta \otimes K(s[i], \sigma[i]) \tag{8}$$

In Equation 8, $K(s[i], \sigma[i])$ denotes a Gaussian kernel of size $s[i]$ and standard deviation $\sigma[i]$. Next, replacing s[i] with 2M+1 and simplifying $K(s[i], \sigma[i])$ as K with $\sigma[i]$ as σ, we have

$$(\delta \otimes K)[n] = \sum_{m=1}^{2M+1} \delta[n-m+M+1] \times K[m] \tag{9}$$

and $K[m]$ is

$$K[m] = \frac{exp(-\frac{(m-M-1)^2}{2\sigma^2})}{\sum_{i=1}^{2M+1} exp(-\frac{(i-M-1)^2}{2\sigma^2})} \tag{10}$$

SAP uses a process similar to PGD to update perturbations, δ, by maximizing the loss function of the classification DNN. Similarly, we set $\delta'_0 = \delta$, and we have

$$\delta'_i = Clip_{\delta'_{i-1},\epsilon}(\delta'_{i-1} + \alpha sign(\bigtriangledown_{\delta'_{i-1}} L(f(x_{adv}(\delta'_{i-1}), y)))) \tag{11}$$

After t' steps, we obtain the final adversarial permutation, $\delta'_{t'}$, and the adversarial sample, $x_{adv} = x + \delta'_{t'}$.

In CardioDefense, we not only add adversarial samples created by SAP into the training process of the first network of CardioDefense to learn the class-related knowledge and data fluctuations, but also add them into the training process of the second network of CardioDefense to further improve the generalization ability of the classification model. The entire training process of the proposed method is shown in Figure 2, and the entire process is detailed in Algorithm 1.

Algorithm 1 The Adversarial Distillation Training Process of CardioDefense

Require: Training set (X_D, Y_D)
for epoch = 1 $\rightarrow E1$ **do**
  for each mini-batch (X, Y) of (X_D, Y_D) **do**
    Create adversarial samples (X_{adv}, Y) through Equation 7-Equation 11;
    Calculate initial DNN output logits of X and X_{adv};
    Calculate probability of each category of X and X_{adv} through Equation 1;
    Calculate loss through Equation 2 and mixed loss through Equation 6;
    According to the mixed loss, calculate gradients of initial DNN parameters;
    Update initial DNN parameters;
  end for
end for
Calculate the soft labels of X_D, Y'_D, generate new training set (X_D, Y'_D)
for epoch = 1 $\rightarrow E2$ **do**
  for each mini-batch (X, Y) of (X_D, Y'_D) **do**
    Create adversarial samples (X_{adv}, Y'_D) through Equation 7-Equation 11;
    Calculate distilled DNN output logits of X and X_{adv};
    Calculate probability of each category of X and X_{adv} through Equation 1;
    Calculate loss through Equation 3 and mixed loss through Equation 6;
    According to the mixed loss, calculate gradients of distilled DNN parameters;
    Update distilled DNN parameters;
  end for
end for

4.3 Variants of CardioDefense

In CardioDefense, we add adversarial samples into the training process of both networks of CardioDefense. It is interesting to investigate adding adversarial samples to only one of the two networks during the training process. Thus, we applied these two variants of our method as the other two defense methods. Furthermore, we abbreviate the two methods as **Init-CardioDefense** and **Dist-CardioDefense**, respectively, where **Init-CardioDefense** denotes the method with adversarial samples added in the training process of the first network of defensive distillation and **Dist-CardioDefense** denotes the method with adversarial samples added in the training process of the second network of defensive distillation.

4.4 Evaluations

In this study, we used the accuracy ratio and F1-score as performance metrics. Specifically, the accuracy ratio is calculated by dividing the number of truly classified samples by the total number of samples, and the F1-score is the harmonic mean of the F1-score from the classification type. Table 6 lists the counting rules for the numbers of different variables. The F1-score for each category of the ECG signal is defined as follows.

Normal: $F_{1N} = \frac{2 \times Nn}{\sum N + \sum n}$,

AF: $F_{1A} = \frac{2 \times An}{\sum A + \sum a}$,

Other: $F_{1O} = \frac{2 \times Oo}{\sum O + \sum o}$,

Noise: $F_{1P} = \frac{2 \times Pp}{\sum P + \sum p}$.

Here, the F1-score is calculated as

$$F_1 = \frac{F_{1N} + F_{1A} + F_{1O} + F_{1P}}{4} \tag{12}$$

4.5 Implementation Details

Here, we introduce the experimental implementation details. Based on the studies conducted by Ma and Liang Ma and Liang [2020a], the only parameter of JR, λ, is set as 44, due to its outstanding performance, and the two parameters of NSR regularization, ϵ_{max} and β, are set to 1. To make the classification model converge quickly, the regularization term of JR and NSR regularization and the NSR margin loss are not added to the training process until the 11th epoch.

Table 6: Evaluations of experiments.

		Prediction				
		Normal	AF	Other	Noise	Total
Ground-truth	Normal	Nn	Na	No	Np	$\sum N$
	AF	An	Aa	Ao	Ap	$\sum A$
	Other	On	Oa	Oo	Op	$\sum O$
	Noise	Pn	Pa	Po	Pp	$\sum P$
	Total	$\sum n$	$\sum a$	$\sum o$	$\sum p$	$\sum All$

The entire training process of CardioDefense is shown as Algorithm 1, and in the experiment, we set $E1 = E2 = 100$. The parameters s and σ of the Gaussian kernel are set as $\{5, 7, 11, 15, 19\}$ and $\{1, 3, 5, 7, 10\}$, respectively. In the process of creating adversarial samples, the iteration step to create PGD adversarial samples, t, is set to 5, and the iteration step to smooth the adversarial perturbation, t', is also set to 5. In addition, we set c in Equation 6 as 0.5 and α in both Equation 7 and Equation 8 as 1. For defensive distillation, the training epochs of the initial network and distilled network are also set as 100. For all other defense methods, the number of training epochs was 100. We set the training batch size of all models to 16 and applied the Adam optimizer with 0.001 as the initial learning rate.

In the process of creating adversarial samples to attack defense models, many parameters are kept the same as those in the training of our method, except that t and t' are set as variable parameters, so that we can know about the defense effect of those methods under different smoothing-degree sample attacks. In addition, we used a fixed test dataset to create adversarial samples. Furthermore, where the value of T is not described, it defaults to 1.

Moreover, to determine the accurate defense effects of different methods against SAP and PGD attacks, and avoid the results deviation caused by randomness, we train the classification DNN with each baseline defense method and CardioDefense five times.

5 Conclusion

In this study, we completely investigated the effects of defense methods against adversarial attacks targeting ECG classification deep neural networks. Furthermore, we proposed a novel defense method called CardioDefense, which involves adding adversarial samples into the training process of both networks of defensive distillation and is good at defending against adversarial attacks with small perturbations. The results of the experiments showed that it's easy to defend the adversarial attack targeting at ECG classification model without defense methods, and when the adversarial attacks attacked the classification model with defense methods, the performance of defense methods is not as good as before. Besides, CardioDefense has better defense effects against white-box attack including SAP attack and low-noise PGD attack which still have a higher level of noise than SAP, as well as black-box attack represented by boundary attack here.

In the future, we will explore the defense effects of gradient-free trained sign activation neural networks against SAP and evaluate more effective defense methods that require less training time but have better defense effects. In addition, we will also explore how to reduce the training time of the classification model with CardioDefense, or achieve a small loss of defense effect but significantly reduce training time. We also plan to extend our work to obtain more explainable results.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (No.62102008).

References

Kevin Mc Namara, Hamzah Alzubaidi, and John Keith Jackson. Cardiovascular disease as a leading cause of death: how are pharmacists getting involved? *Integrated pharmacy research & practice*, 8:1, 2019.

Antônio H Ribeiro, Manoel Horta Ribeiro, Gabriela MM Paixão, Derick M Oliveira, Paulo R Gomes, Jéssica A Canazart, Milton PS Ferreira, Carl R Andersson, Peter W Macfarlane, Wagner Meira Jr, et al. Automatic diagnosis of the 12-lead ecg using a deep neural network. *Nature communications*, 11(1):1–9, 2020.

Sean Shensheng Xu, Man-Wai Mak, and Chi-Chung Cheung. Towards end-to-end ecg classification with raw signal extraction and deep neural networks. *IEEE journal of biomedical and health informatics*, 23(4):1574–1584, 2018.

Piyush Jain, Pranjali Gajbhiye, RK Tripathy, and U Rajendra Acharya. A two-stage deep cnn architecture for the classification of low-risk and high-risk hypertension classes using multi-lead ecg signals. *Informatics in Medicine Unlocked*, 21:100479, 2020.

Dakun Lai, Yuxiang Bu, Ye Su, Xinshu Zhang, and Chang-Sheng Ma. Non-standardized patch-based ecg lead together with deep learning based algorithm for automatic screening of atrial fibrillation. *IEEE Journal of Biomedical and Health Informatics*, 24(6):1569–1578, 2020.

Veer Sangha, Bobak J Mortazavi, Adrian D Haimovich, Antônio H Ribeiro, Cynthia A Brandt, Daniel L Jacoby, Wade L Schulz, Harlan M Krumholz, Antonio Luiz P Ribeiro, and Rohan Khera. Automated multilabel diagnosis on electrocardiographic images and signals. *Nature Communications*, 13(1):1–12, 2022.

Awni Y Hannun, Pranav Rajpurkar, Masoumeh Haghpanahi, Geoffrey H Tison, Codie Bourn, Mintu P Turakhia, and Andrew Y Ng. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nature medicine*, 25(1):65–69, 2019.

Daniel Sinnecker. A deep neural network trained to interpret results from electrocardiograms: better than physicians? *The Lancet Digital Health*, 2(7):e332–e333, 2020.

Yonatan Elul, Aviv A Rosenberg, Assaf Schuster, Alex M Bronstein, and Yael Yaniv. Meeting the unmet needs of clinicians from ai systems showcased for cardiology with deep-learning–based ecg analysis. *Proceedings of the National Academy of Sciences*, 118(24), 2021.

Shenda Hong, Yuxi Zhou, Junyuan Shang, Cao Xiao, and Jimeng Sun. Opportunities and challenges of deep learning methods for electrocardiogram data: A systematic review. *Computers in Biology and Medicine*, 122:103801, 2020a.

Sulaiman Somani, Adam J Russak, Felix Richter, Shan Zhao, Akhil Vaid, Fayzan Chaudhry, Jessica K De Freitas, Nidhi Naik, Riccardo Miotto, Girish N Nadkarni, et al. Deep learning and the electrocardiogram: review of the current state-of-the-art. *EP Europace*, 23(8):1179–1191, 2021.

Konstantinos C Siontis, Peter A Noseworthy, Zachi I Attia, and Paul A Friedman. Artificial intelligence-enhanced electrocardiography in cardiovascular disease management. *Nature Reviews Cardiology*, 18(7):465–478, 2021.

Zhaoji Fu, Shenda Hong, Rui Zhang, and Shaofu Du. Artificial-intelligence-enhanced mobile system for cardiovascular health management. *Sensors*, 21(3):773, 2021.

Zachi I Attia, Peter A Noseworthy, Francisco Lopez-Jimenez, Samuel J Asirvatham, Abhishek J Deshmukh, Bernard J Gersh, Rickey E Carter, Xiaoxi Yao, Alejandro A Rabinstein, Brad J Erickson, et al. An artificial intelligence-enabled ecg algorithm for the identification of patients with atrial fibrillation during sinus rhythm: a retrospective analysis of outcome prediction. *The Lancet*, 394(10201):861–867, 2019.

Urtnasan Erdenebayar, Yoon Ji Kim, Jong-Uk Park, Eun Yeon Joo, and Kyoung-Joung Lee. Deep learning approaches for automatic detection of sleep apnea events from an electrocardiogram. *Computer methods and programs in biomedicine*, 180:105001, 2019.

Emilly M Lima, Antônio H Ribeiro, Gabriela MM Paixão, Manoel Horta Ribeiro, Marcelo M Pinto-Filho, Paulo R Gomes, Derick M Oliveira, Ester C Sabino, Bruce B Duncan, Luana Giatti, et al. Deep neural network-estimated electrocardiographic age as a mortality predictor. *Nature communications*, 12(1):1–10, 2021.

Sushravya Raghunath, Alvaro E Ulloa Cerna, Linyuan Jing, David P VanMaanen, Joshua Stough, Dustin N Hartzel, Joseph B Leader, H Lester Kirchner, Martin C Stumpe, Ashraf Hafez, et al. Prediction of mortality from 12-lead electrocardiogram voltage data using a deep neural network. *Nature medicine*, 26(6):886–891, 2020.

Shenda Hong, Yanbo Xu, Alind Khare, Satria Priambada, Kevin Maher, Alaa Aljiffry, Jimeng Sun, and Alexey Tumanov. Holmes: health online model ensemble serving for deep learning models in intensive care units. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1614–1624, 2020b.

Nannapas Banluesombatkul, Pichayoot Ouppaphan, Pitshaporn Leelaarporn, Payongkit Lakhan, Busarakum Chaitusaney, Nattapong Jaimchariyatam, Ekapol Chuangsuwanich, Wei Chen, Huy Phan, Nat Dilokthanakul, et al. Metasleeplearner: A pilot study on fast adaptation of bio-signals-based sleep stage classifier to new individual subject using meta-learning. *IEEE Journal of Biomedical and Health Informatics*, 25(6):1949–1963, 2020.

Ruggero Donida Labati, Enrique Muñoz, Vincenzo Piuri, Roberto Sassi, and Fabio Scotti. Deep-ecg: Convolutional neural networks for ecg biometric recognition. *Pattern Recognition Letters*, 126:78–85, 2019.

Shenda Hong, Can Wang, and Zhaoji Fu. Cardioid: Learning to identification from electrocardiogram data. *Neurocomputing*, 412:11–18, 2020c.

Jingzhen Li, Igbe Tobore, Yuhang Liu, Abhishek Kandwal, Lei Wang, and Zedong Nie. Non-invasive monitoring of three glucose ranges based on ecg by using dbscan-cnn. *IEEE Journal of Biomedical and Health Informatics*, 25(9): 3340–3350, 2021.

J Weston Hughes, Jeffrey E Olgin, Robert Avram, Sean A Abreau, Taylor Sittler, Kaahan Radia, Henry Hsia, Tomos Walters, Byron Lee, Joseph E Gonzalez, et al. Performance of a convolutional neural network and explainability technique for 12-lead electrocardiogram interpretation. *JAMA cardiology*, 6(11):1285–1295, 2021.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Alexey Kurakin, Ian Goodfellow, Samy Bengio, et al. Adversarial examples in the physical world. 2016.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. IEEE, 2017.

Xintian Han, Yuxuan Hu, Luca Foschini, Larry Chinitz, Lior Jankelson, and Rajesh Ranganath. Deep learning models for electrocardiograms are susceptible to adversarial attack. *Nature medicine*, 26(3):360–363, 2020.

Huangxun Chen, Chenyu Huang, Qianyi Huang, Qian Zhang, and Wei Wang. Ecgadv: Generating adversarial electrocardiogram to misguide arrhythmia classification system. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3446–3453, 2020.

Jonathan Lam, Pengrui Quan, Jiamin Xu, Jeya Vikranth Jeyakumar, and Mani Srivastava. Hard-label black-box adversarial attack on deep electrocardiogram classifier. In *Proceedings of the 1st ACM International Workshop on Security and Safety for Intelligent Cyber-Physical Systems*, pages 6–12, 2020.

Zhibo Yang, Yanan Yang, Yunzhe Xue, Frank Y Shih, Justin Ady, and Usman Roshan. Accurate and adversarially robust classification of medical images and ecg time-series with gradient-free trained sign activation neural networks. In *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 2456–2460. IEEE, 2020.

Linhai Ma and Liang Liang. Enhance cnn robustness against noises for classification of 12-lead ecg with variable length. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 839–846. IEEE, 2020a.

Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE symposium on security and privacy (SP)*, pages 582–597. IEEE, 2016.

Daniel Jakubovitz and Raja Giryes. Improving dnn robustness to adversarial attacks using jacobian regularization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 514–529, 2018.

Linhai Ma and Liang Liang. Improve robustness of dnn for ecg signal classification: a noise-to-signal ratio perspective. *arXiv preprint arXiv:2005.09134*, 2020b.

Sebastian D Goodfellow, Andrew Goodwin, Robert Greer, Peter C Laussen, Mjaye Mazwi, and Danny Eytan. Towards understanding ecg rhythm classification using convolutional neural networks and attention mappings. In *Machine learning for healthcare conference*, pages 83–101. PMLR, 2018.

Gari D Clifford, Chengyu Liu, Benjamin Moody, H Lehman Li-wei, Ikaro Silva, Qiao Li, AE Johnson, and Roger G Mark. Af classification from a short single lead ecg recording: The physionet/computing in cardiology challenge 2017. In *2017 Computing in Cardiology (CinC)*, pages 1–4. IEEE, 2017.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7), 2015.

Ali Shafahi, Mahyar Najibi, Mohammad Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. Adversarial training for free! *Advances in Neural Information Processing Systems*, 32, 2019.

Gavin Weiguang Ding, Yash Sharma, Kry Yik Chau Lui, and Ruitong Huang. Mma training: Direct input space margin maximization through adversarial training. *arXiv preprint arXiv:1812.02637*, 2018.

Linhai Ma and Liang Liang. Increasing-margin adversarial (ima) training to improve adversarial robustness of neural networks. *arXiv preprint arXiv:2005.09147*, 2020c.