Transferability and Adversarial Training in Automatic Classification of the Electrocardiogram with Deep Learning

Arvid Eriksson[1], Thomas B Schön[2], Antônio H Ribeiro[2]

[1] KTH Royal Institute of Technology, Stockholm, Sweden
[2] Uppsala University, Uppsala, Sweden

Abstract

Automatic electrocardiogram (ECG) analysis using deep neural networks has seen promising results in recent years. However, these models are susceptible to domain shifts when they are applied to new data distributions not previously trained on. We investigate how training on worst-case artificial samples through adversarial training can help promote models that can easily be molded through fine-tuning to new datasets. We compare the area under the precision-recall curve (AUPRC) for the classification of atrial fibrillation using two cohorts: we use PTB-XL for training and CODE-15% for fine-tuning and evaluating the models. Our results show that adversarially trained models on ECG data yield higher transferability when fine-tuned on new datasets compared to normally trained models (0.732 vs. *0.685 AUPRC). We also note that they even have the ability to supersede models solely trained on the new dataset using more total time and data (0.732* vs. *0.685 AUPRC). Our work thus paves the way for the training of general models that can be applied to different types of new settings with high performance.*

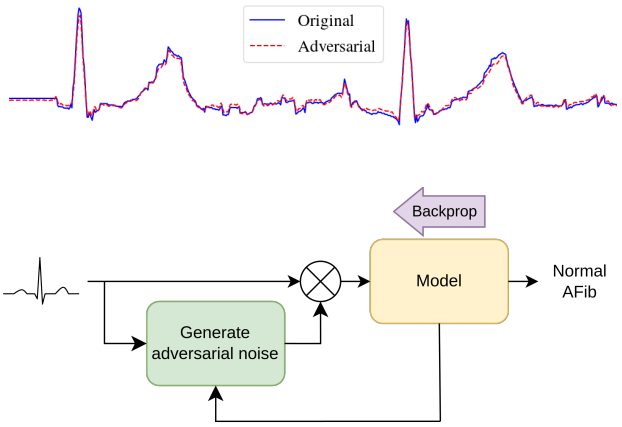

Figure 1: *Top:* Comparison of normal ECG and corresponding adversarial example with a perturbation size of $\varepsilon = 0.01$. A normally trained model predicts with Prob > 0.99 that the ECG is normal while for the adversarially modified ECG, the model predicts atrial fibrillation with Prob > 0.99. *Bottom:* Illustration of adversarial training. The method aims to make the model resistant to such disturbances.

1. Introduction

Cardiovascular diseases account for one-third of the deaths worldwide and the electrocardiogram (ECG) is a major tool in their diagnoses [1]. Deep neural networks have recently achieved striking success in the automatic analysis of the ECG [2, 3] and while these models bring great promise they also usually require large amounts of data to be developed ($n > 10,000$). The ECG is useful in a wide variety of settings from basic care to the emergency department and data might not be available in the exact setting that we are interested in developing a deep neural network-based solution. This suggests the need for models that are general and robust and can be applied to scenarios slightly different than they have been developed for, including different populations, different hospitals, and different equipment.

In this paper, we study the possibility of using adversarial training to improve model transferability. Adversarial training is a technique that involves training the model using worst-case disturbed artificial samples instead of the original ones [4]. One example of such a disturbed sample is shown in Figure 1 (top), it is contaminated with a worst-case disturbance, called an adversarial attack since it is deliberately chosen to maximize the model error. These types of disturbances are known to severely affect state-of-the-art deep learning models used for ECG classification [5].

Adversarial training is one of the most effective approaches for deep learning models to defend against adversarial attacks [4, 6]. Adversarial training is summarized in Figure 1 (bottom) and is formulated as a min-max problem, searching for the best solution to the worst-case attacks. Different types of adversarial training have seen pre-

vious success when applied to ECG models [7–9]. These methods have shown adversarial training to improve the robustness of adversarial attacks targeted toward ECGs.

Our work investigates whether adversarial training enhances the transferability of ECG deep learning models. This has been shown in other domains, such as in image classification [10]. We study both the effect of adversarially training a model as well as how training a model on one dataset using adversarial training can affect the fine-tuned performance on another dataset.

2. Methods

2.1. Datasets

We use two datasets to develop and evaluate our model atrial fibrillation classification models.

CODE-15% [12]. The Clinical Outcomes in Digital Electrocardiography (CODE) dataset was developed with the database of digital ECG exams of the telehealth network of Minas Gerais and a detailed description of the cohort can be obtained at [12].

The data set was collected between 2010 and 2017 from 811 counties in the state of Minas Gerais, Brazil, and consists of 2,322,513 12-lead ECG records from 1,676,384 different patients. A subset of 15% of this data set is available online [12] and is the variant used for this paper, containing 6,504 positive atrial fibrillation samples and 313,337 negative samples.

PTB-XL [13] is a dataset consisting of 21,799 12-lead ECG records collected between 1989 and 1996 from 18,869 different patients using devices from Schiller AG. It covers a larger range of diagnostic statements than the CODE dataset, with additional types of arrhythmias and diagnostics annotated, and contains 1,362 positive atrial fibrillation samples and 18,239 negative.

2.2. Model

Data preprocessing. The ECG signals have been resampled such that all ECGs have the same sampling frequency of 400 Hz. Each input ECG has 4,096 time samples for each of the 12 standard ECG leads. Original signals of a shorter time span have been extended through zero-padding. The output data comprises binary scalar variables corresponding to positive or negative diagnoses.

Architecture. The deep learning model consists of a residual neural network (ResNet) adapted to uni-dimensional signals and includes convolutional layers both before and within the residual blocks. We make use of the same network architecture as [2], where the CODE-15% data set was utilized to classify multiple ECG abnormalities; we refer to that work for further details and note that we have modified the final output layer in adaptation to our binary classification.

2.3. Adversarial attacks and training

Adversarial attacks. In this framework, an adversary changes the input to the model to maximize the prediction error. It is formulated as a maximization problem, where the perturbation δ that will be added is computed through the problem

$$\arg\max_{\delta \in S} \mathcal{L}(x + \delta, y)$$

where $S = \{\delta : \|\delta\|_\infty \leq \varepsilon\}$ and ε is the adversarial radius and $\mathcal{L}(x, y)$ is the model loss for an ECG sample and its corresponding diagnosis.

One traditional approach to solve this problem is to use projected gradient descent (PGD). PGD adversarial attacks are state-of-the-art for evaluating robustness and adversarial training [6]. It generates adversarial examples by applying projected gradient descent, with step size α (see Equation 1), to solve the optimization problem. It has the following iteration using Π as the projection operator

$$x^{t+1} = \Pi_{S+x}(x^t + \alpha \operatorname{sgn}(\nabla_x \mathcal{L}(x^t, y))) \tag{1}$$

In this paper, we use APGD as an improved version of PGD as described in [14] which makes the step size adaptively scheduled during iterations in an explore-exploit-like fashion.

Adversarial training. Adversarial training can be defined as a min-max problem where the best solution for the worst-case attacks is sought after:

$$\min_\theta \frac{1}{n} \sum_{i=1}^{n} \max_{\delta \in S} \mathcal{L}(x_i + \delta, y_i) \tag{2}$$

where $(x_i, y_i) \in \mathcal{D}$ where $\mathcal{D}$ is the training set consisting of ECG samples and corresponding diagnoses.

In practice, adversarial training consists of training on adversarial examples constructed through adversarial attacks which estimate the maximizing perturbation δ. PGD, or variants such as APGD, is the most often used estimator and is also used for evaluating the adversarial robustness of a model.

Implementation. Similar to [9] we let the loss during adversarial training be the average between the standard and adversarial loss where the adversarial loss is defined as $\mathcal{L}(x_{adv}, y)$ where x_{adv} is generated through 10-iteration APGD with random initialization.

Adversarial training is only initialized after an initial warm-up period where the model is trained normally. A scaling of the adversarial radius ε using an exponential schedule is then employed with a ε_{min} and ε_{max} which represent the minimum and maximum epsilon respectively during scheduling. This lets the model converge before introducing the adversarial examples. All models are trained using the AdamW optimizer [15] using a learning rate of 1×10^{-3} and using binary cross-entropy loss predicting

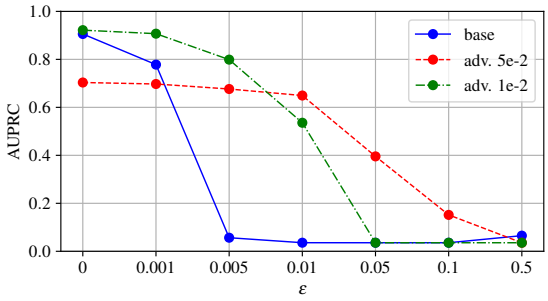

Figure 2: Robust performance for baseline and adversarially trained model with $\varepsilon_{max} = 0.05$ and $\varepsilon_{max} = 0.01$ respectively.

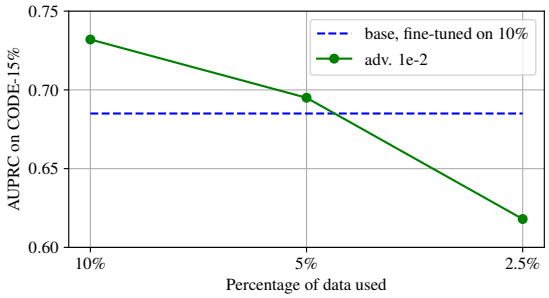

Figure 3: Performance of fine-tuned adversarially trained model with $\varepsilon_{max} = 0.05$ for different amounts of training data fine-tuned on. Baseline is a normally trained model fine-tuned on 10% of the CODE-15% dataset.

atrial fibrillation. The fine-tuning experiments used an identical setup besides using a learning rate of 1×10^{-4}.

2.4. Experiment description.

The experiments consisted of initial robustness comparisons of normally and adversarially trained models on PTB-XL and fine-tuning a normally trained and adversarially trained model from the PTB-XL dataset on CODE-15% and comparing them with a baseline on the task of classifying atrial fibrillation.

Robustness. Three PTB-XL models were trained using an 80/10/10 training, validation, and test split for 20 epochs. Adversarial training was done for two models with a warm-up of 10 epochs with a common $\varepsilon_{min} = 1 \times 10^{-3}$ and $\varepsilon_{max} = 1 \times 10^{-2}$ and $\varepsilon_{max} = 5 \times 10^{-2}$ respectively. The last model was trained normally as a baseline.

The robust performance of these three models was then evaluated by measuring the performance using the area under the precision-recall curve (AUPRC) for 10-APGD attacks for different ε (as ROC curves can be optimistic for unbalanced datasets [16]). The results can be seen in Figure 2.

Transferability. The three models were then fine-tuned on the CODE-15% dataset where their performance was compared to a baseline model. The fine-tuned models were trained using 10% of CODE-15%, while the baseline was trained on 90% of the dataset. The remaining 10% was used for evaluation and comparison of the models. The results can be seen in Table 1. We also attempt fine-tuning an adversarially trained model using only 5% and 2.5% of CODE-15% for a set number of iterations and compare this to the fine-tuned baseline, see Figure 3. Experiments were performed using an RTX 3080 using PyTorch.

3. Results

Figure 2 shows the robust performance of the three models trained on PTB-XL. Note how the heavily adversarially trained model (red) is robust even for higher ε. A higher ε during adversarial training corresponds to higher robustness for larger perturbations but lower clean performance.

Table 1 shows the specifications of each fine-tuned model and their performance on the CODE-15% test set. Note the higher AUPRC for the fine-tuned models from the adversarially trained PTB-XL model. The transferability of the adversarially trained models seems to be generally higher than the corresponding non-adversarially trained model which indicates more general representations. The results even show that an adversarially trained and fine-tuned model can outperform a normally trained model on the CODE-15% dataset that utilizes significantly more total data and time.

Figure 3 shows that the adversarially trained model only needs roughly half of the data when fine-tuning to match the baseline fine-tuned model, further suggesting higher transferability of adversarially trained models.

4. Discussion

This study shows the impact of adversarial training through the scope of transferability. We show that adversarial training improves robustness, often at the cost of performance on clean data, while also generating models that are more suitable for fine-tuning. These robustness results are in line with previous studies in computer vision [9].

The high transferability of the adversarially trained models shows a potential opportunity for creating general ECG models through adversarial training which can then be fine-tuned on smaller datasets while retaining high performance. This is key in clinical domains as it can allow the production of a general ECG model which can then easily be adapted to specific hospitals or equipment.

Adversarially training models on large numbers of data, however, can be computationally expensive. This indicates the need for future work to investigate whether the above observations hold when using faster forms of adversarial

Table 1: Performance of various models on the CODE-15% dataset measured by area under the precision-recall curve (AUPRC). Under different configurations of training in PTB-XL training and fine-tuning on CODE-15%. The interquartile range of the AUPRC reported under parenthesis is computed using bootstrap with $n = 500$.

Training					Fine-tuning				AUPRC
Dataset	Epochs	Size	Adv. train	t (min)	Dataset	Epochs	Size	t (min)	
CODE-15%	10	311196	no	99	-	-	-	-	0.685 (0.675-0.700)
PTB-XL	20	17418	no	15	CODE-15%	10	34583	10	0.685 (0.673-0.699)
PTB-XL	20	17418	yes, $\varepsilon_{max} = 0.01$	40	CODE-15%	10	34583	10	0.732 (0.721-0.746)
PTB-XL	20	17418	yes, $\varepsilon_{max} = 0.05$	40	CODE-15%	10	34583	10	0.684 (0.674-0.698)

training such as using FastAT [17] or GradAlign [18].

One should also note that the definition of robustness improvements used in adversarial training is not necessarily the same as the robustness to domain shifts. Future work should also investigate ECG-specific problems such as how smoothing as used in [5] to introduce indistinguishable adversarial examples can be used for adversarial training and its impact on transferability. Finally, more extensive studies using more tasks besides the classification of atrial fibrillation and additional cohorts are needed to confirm our findings.

Acknowledgments

TBS is financially supported by the Swedish Research Council, by the *Wallenberg AI, Autonomous Systems and Software Program (WASP)* funded by Knut and Alice Wallenberg Foundation, and by the *Kjell och Märta Beijer Foundation*.

References

[1] Mensah GA, Fuster V, Murray CJL, Roth GA, et al. Global burden of cardiovascular diseases and risks, 1990-2022. Journal of the American College of Cardiology 2023; 82(25):2350–2473.

[2] Ribeiro AH, Ribeiro MH, Paixão GMM, Oliveira DM, Gomes PR, Canazart JA, Ferreira MPS, Andersson CR, Macfarlane PW, Meira Jr. W, Schön TB, Ribeiro ALP. Automatic diagnosis of the 12-lead ECG using a deep neural network. Nature Communications 2020;11(1):1760.

[3] Lu L, Zhu T, Ribeiro AH, Clifton L, Zhao E, Zhou J, Ribeiro ALP, Zhang YT, Clifton DA. Decoding 2.3 million ECGs: Interpretable deep learning for advancing cardiovascular diagnosis and mortality risk stratification. European Heart Journal Digital Health 2024.

[4] Madry A, Makelov A, Schmidt L, Tsipras D, Vladu A. Towards deep learning models resistant to adversarial attacks. ICLR 2018;

[5] Han X, Hu Y, Foschini L, Chinitz L, Jankelson L, Ranganath R. Deep learning models for electrocardiograms are susceptible to adversarial attack. Nature Medicine 2020; 26(3):360–363.

[6] Bai T, Luo J, Zhao J, Wen B, Wang Q. Recent advances in adversarial training for adversarial robustness. In Zhou ZH (ed.), IJCAI 2021.

[7] Hossain KF, Kamran SA, Tavakkoli A, Ma X. ECG-ATK-GAN: Robustness against adversarial attacks on ECGs using conditional generative adversarial networks. Applications of Medical Artificial Intelligence, 2022; 68–78.

[8] Shao J, Geng S, Fu Z, Xu W, Liu T, Hong S. CardioDefense: Defending against adversarial attack in ECG classification with adversarial distillation training. Biomedical Signal Processing and Control 2024;91:105922.

[9] Ma L, Liang L. 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) 2020.

[10] Deng Z, Zhang L, Vodrahalli K, Kawaguchi K, Zou JY. Adversarial training helps transfer learning via better representations. Advances in Neural Information Processing Systems, volume 34, 2021; 25179–25191.

[11] Ribeiro AH, Ribeiro MH, Paixão GM, Oliveira DM, Gomes PR, Canazart JA, Ferreira MP, Andersson CR, Macfarlane PW, Jr. WM, Schön TB, Ribeiro ALP. CODE dataset, 2021.

[12] Ribeiro AH, Paixao GM, Lima EM, Horta Ribeiro M, Pinto Filho MM, Gomes PR, Oliveira DM, Meira Jr W, Schon TB, Ribeiro ALP. CODE-15%: a large scale annotated dataset of 12-lead ECGs, June 2021. URL `https://doi.org/10.5281/zenodo.4916206`.

[13] Wagner P, Strodthoff N, Bousseljot RD, Samek W, Schaeffter T. PTB-XL, a large publicly available electrocardiography dataset.

[14] Croce F, Hein M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. ICML 2020

[15] Loshchilov I, Hutter F. Decoupled weight decay regularization. ICLR, 2019

[16] Davis J, Goadrich M. The relationship between Precision-Recall and ROC curves. ICML 2006.

[17] Wong E, Rice L, Kolter JZ. Fast is better than free: Revisiting adversarial training, January 2020.

[18] Andriushchenko M, Flammarion N. Understanding and improving fast adversarial training, October 2020.

Address for correspondence:

Arvid Eriksson
KTH, SE-100 44 Stockholm Sweden
arveri@kth.se