Adversarial Examples for Electrocardiograms

Xintian Han, Yuxuan Hu, Luca Foschini, Lior Jankelson, Rajesh Ranganath

Introduction and Related Work. Among the physiological signals, ECG has seen some of the largest expansion in both medical and recreational applications. In parallel with the traditional 12 lead ECG, we are witnessing the rise of single-lead versions embedded in medical devices and wearable products. Devices such as the injectable Medtronic Linq monitor and the iRhythm Ziopatch wearable monitor are widely used in the diagnosis of cardiac arrhythmia, while smart watches marketed directly to consumers such as the Apple Watch Series 4 now feature a single lead ECG. Altogether, single lead ECG is expected to be used by tens of millions of Americans by the end of 2019 [7].

Meaningful use of the deluge of data being created requires automated methods. Many approaches to building predictive models from clinical data, including electrocardiogram (ECG), rely on deep learning. Examples include cheXnet for chest x-rays [11], deep survival analysis for coronary artery disease [12], and DeepPath for pathology [2]. Similar methods have also recently been cleared by the Food and Drug Administration. [9]

However, deep learning classifiers have been shown to be brittle to adversarial examples [4; 13], and very recently in medical related tasks [10; 3]. However, creating adversarial examples for physiological signals poses additional challenges. Naively attacking ECG deep learning classifiers with traditional methods creates examples presenting square waves artifacts that are not physiologically plausible. To remedy this, we develop a method to construct *smoothed* adversarial examples. The methods successfully creates *false negatives*: examples of symptomatic ECG indistinguishable to a human eye that get classified as normal by the model (Fig 1, Tab 1).

Methods. We construct adversarial examples for state of art deep learning methods in 2017 PhysioNet/CinC Challenge [1] that classify a single short ECG lead recordings to four types: normal sinus rhythm (Normal), atrial fibrillation (atrial fibrillation (AF)), an alternative rhythm (Other), or is too noisy to be classified (Noise). The challenge training set contains 8,528 single-lead ECG recordings lasting from 9s to about 60s. We split the training set randomly into a new training set (90%) and new test set (10%). We train the 13-layer convolutional network from [5] on the training set and get accuracy 0.88 and F1 score 0.87 of the three majority classes (Normal, AF and Other rhythm) on the test set which is comparable to the state of art ECG classification [5].

We create adversarial examples with test set. However, directly applying PGD to ECG classification will create very non-smooth signals which can be easily distinguished from real ECG by human eyes. We propose a method to train a smooth perturbation (TSP). We take the adversarial perturbation as the parameter θ and add it to the clean examples after convolving with a number of Gaussian kernels $G(s, \sigma)$. The resulting adversarial example could be written as a function of θ:

$$x_{\text{adv}}(\theta) = x + \frac{1}{m}\sum_{i}^{m} \theta \circledast G(s[i], \sigma[i]).$$

Then we use PGD to maximize the loss function L with respect to θ to get adversarial example for a given input-label pair (x, y):

$$\theta'_i = Clip_{0,\varepsilon}\{\theta'_{i-1} + \alpha \cdot \text{sign}(\nabla_\theta L(f(x_{\text{adv}}(\theta'_{i-1}), y)))\}.$$

Figure 1: Adversarial examples AF to Normal.

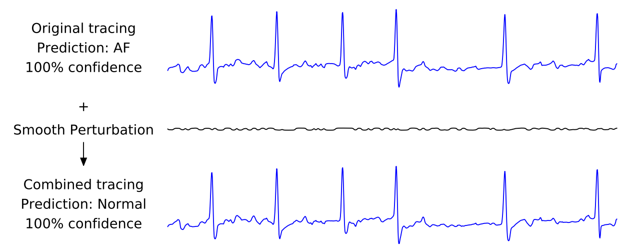

Table 1: Doctor's ability to find real ECG's and neural network's accuracy on adversarial examples as labeled by doctors

doctor's accuracy to distinguish		network accuracy
PGD 95/100	TSP 59/100	≤ 0.028

Discussion. We demonstrate here how adversarial examples may pose a real challenge for machine learning systems designed for ECG applications. Our findings are in line with recently published examples in other medical fields [3]. This misclassification susceptibility is important, since it may expose AI based systems to error induced by unexpected perturbations in signal, which could be environmental and unexpected. Moreover, it may enable malicious actors to change outcomes of clinical studies and insurance claims. This is especially relevant with the increased reliance on Real World Data (RWD) for health-care related decision making [6]. For example, in the near future raw ECG recordings in cardiovascular-related trials may come directly from study participants' smart watches as Patient Generated Health Data (PGHD). Finally, this type of interference may be particularly difficult to detect, given the indistinguishable change in ECG pattern. It is imperative to ensure a trusted chain of custody in both clinical use and RWD acquisition in order to prevent malicious actors to imperceptibly change the data to affect the outcome. Furthermore, and perhaps more importantly, it is important to develop and adopt analysis algorithms that are provably robust to adversarial attacks [8].

Acknowledgement We thank Dr. Wei-Nchih Lee for labeling adversarial examples.

References

[1] Gari D Clifford, Chengyu Liu, Benjamin Moody, Liwei H Lehman, 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. *Computing in cardiology*, 2017.

[2] Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L Moreira, Narges Razavian, and Aristotelis Tsirigos. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature medicine*, 24(10):1559, 2018.

[3] Samuel G Finlayson, Isaac S Kohane, and Andrew L Beam. Adversarial attacks against medical deep learning systems. *arXiv preprint arXiv:1804.05296*, 2018.

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

[5] 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. *Proceedings of Machine Learning Research*.

[6] Grace Hampson, Adrian Towse, William B Dreitlein, Chris Henshall, and Steven D Pearson. Real-world evidence for coverage decisions: opportunities and challenges. *Journal of comparative effectiveness research*, 7(12):1133–1143, 2018.

[7] International Data Corporation (IDC). Idc reports strong growth in the worldwide wearables market, led by holiday shipments of smartwatches, wrist bands, and ear-worn devices, March 2019. URL https://www.idc.com/getdoc.jsp?containerId=prUS44901819.

[8] 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.

[9] Dave Muoio. Roundup: 12 healthcare algorithms cleared by the FDA, November 2018. URL https://www.mobihealthnews.com/content/roundup-12-healthcare-algorithms-cleared-fda.

[10] Magdalini Paschali, Sailesh Conjeti, Fernando Navarro, and Nassir Navab. Generalizability vs. robustness: Adversarial examples for medical imaging. *arXiv preprint arXiv:1804.00504*, 2018.

[11] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017.

[12] Rajesh Ranganath, Adler Perotte, Noémie Elhadad, and David Blei. Deep survival analysis. In *Machine Learning for Healthcare Conference*, pages 101–114, 2016.

[13] 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.