LETTERS
https://doi.org/10.1038/s41591-020-0791-x

nature medicine

Check for updates

Deep learning models for electrocardiograms are susceptible to adversarial attack

Xintian Han[1]✉, Yuxuan Hu[2], Luca Foschini[3], Larry Chinitz[2], Lior Jankelson[2] and Rajesh Ranganath[1,4,5]✉

Electrocardiogram (ECG) acquisition is increasingly widespread in medical and commercial devices, necessitating the development of automated interpretation strategies. Recently, deep neural networks have been used to automatically analyze ECG tracings and outperform physicians in detecting certain rhythm irregularities[1]. However, deep learning classifiers are susceptible to adversarial examples, which are created from raw data to fool the classifier such that it assigns the example to the wrong class, but which are undetectable to the human eye[2,3]. Adversarial examples have also been created for medical-related tasks[4,5]. However, traditional attack methods to create adversarial examples do not extend directly to ECG signals, as such methods introduce square-wave artefacts that are not physiologically plausible. Here we develop a method to construct smoothed adversarial examples for ECG tracings that are invisible to human expert evaluation and show that a deep learning model for arrhythmia detection from single-lead ECG[6] is vulnerable to this type of attack. Moreover, we provide a general technique for collating and perturbing known adversarial examples to create multiple new ones. The susceptibility of deep learning ECG algorithms to adversarial misclassification implies that care should be taken when evaluating these models on ECGs that may have been altered, particularly when incentives for causing misclassification exist.

Cardiovascular diseases represent a major health burden, accounting for 30% of deaths worldwide[7]. The electrocardiogram (ECG) is a simple and non-invasive test used for the screening and diagnosis of cardiovascular disease. It is widely available in multiple medical device applications, including standard 12-lead ECGs, Holter recorders and monitoring devices[8]. In recent years, there has been further growth in ECG utilization in the form of single-lead ECGs, which are used in miniature implantable medical devices and wearable medical consumer products such as smart watches. These single-lead ECGs, such as the one incorporated in the Apple Watch Series 4, were predicted to be worn by tens of millions of Americans by the end of 2019 (https://www.idc.com/getdoc.jsp?containerId=prUS44901819). Moreover, consumer-wearable devices are utilized to collect data in clinical studies, such as the Health eHeart study (https://www.ucsf.edu/news/2018/02/409806/wearables-could-catch-heart-problems-elude-your-doctor) and the Apple Heart Study (https://www.acc.org/latest-in-cardiology/articles/2019/03/08/15/32/sat-9am-apple-heart-study-acc-201). Large studies that make use of patient-generated health data (PGHD) are expected to become more frequent after the recent release by the Food and Drug Administration (FDA) of a set of guidelines and tools to collect real-world data (RWD) from research participants via apps and other mobile health sources (https://www.fda.gov/media/120060/download). Having clinicians analyze such a large number of ECGs is impractical.

Recently, driven by the introduction of deep learning methodologies, automated systems have been developed that allow rapid and accurate ECG classification[1]. In the 2017 PhysioNet Challenge for atrial fibrillation classification using single-lead ECGs, multiple efficient solutions utilized deep neural networks[9]. Deep learning has been shown to be susceptible to adversarial examples both in general[2,3] and very recently in medical applications[4]. Adversarial examples in ECGs have been independently discovered by others[10]. In contrast to the results of ref. [10], this Letter develops a model-based smoothed attack and explores the existence of adversarial examples by constructing a sampling process for them. Part of this Letter's contribution is to establish a mathematical construction of adversarial examples for ECGs that align with human expert evaluation.

We obtained ECGs from the publicly available 2017 PhysioNet/CinC Challenge[6]. The goal of the challenge was to classify single-lead ECG recordings to four types: normal sinus rhythm (Normal), atrial fibrillation (AF), an alternative rhythm (Other) or noise (Noise). The challenge dataset contained 8,528 single-lead ECG recordings lasting from 9 s to ~60 s, including 5,076 Normal, 758 AF, 2,415 Other and 279 Noise examples. We used 90% of the dataset for training and 10% for testing.

We used a 13-layer convolutional network[11] that won the 2017 PhysioNet/CinC Challenge. We evaluated both accuracy and F1 score. F1 score ranges from 0 to 1, and a high F1 score indicates good network performance, with high true positive and true negative rates. The model achieved an average accuracy rate of 0.88 and F1 score of 0.87 for the ECG classes (Normal, AF and Other) on the test set, which is comparable to state-of-the-art ECG classification systems[11].

Adversarial examples are designed to cause a machine learning algorithm to make a mistake. An adversarial example is made by adding a small perturbation to the input of the machine learning algorithm that changes the prediction on the input, while also ensuring it still looks like a real input[3]. These kinds of adversarial example have been successfully created in the field of medical imaging classification[4].

Traditional adversarial attack algorithms add a small imperceptible perturbation to lower the prediction accuracy of a machine learning model. However, attacking ECG deep learning classifiers with traditional methods creates examples that display square-wave artefacts that are not physiologically plausible (Extended Data Fig. 1).

[1]Center for Data Science, New York University, New York, NY, USA. [2]Leon H. Charney Division of Cardiology, NYU Langone Health, New York, NY, USA. [3]Evidation Health, Inc., San Mateo, CA, USA. [4]Department of Population Health, NYU Langone Health, New York, NY, USA. [5]Courant Institute of Mathematics, New York University, New York, NY, USA. ✉e-mail: xintian.han@nyu.edu; rajeshr@cims.nyu.edu

Table 1 | Success rate of the targeted smooth attack method

		Target class			
		Normal (%)	AF (%)	Other (%)	Noise (%)
Original class	Normal	–	57	55	13
	AF	74	–	87	22
	Other	72	76	–	20
	Noise	79	64	57	–

The original class is the class into which the network classifies the signal before the adversarial attack. The target class is the class into which the adversarial attack aimed to make the network classify the signal after adding. The success rate is calculated as the percentage of examples from the original class that were misclassified by the network to the target class after the adversarial attack.

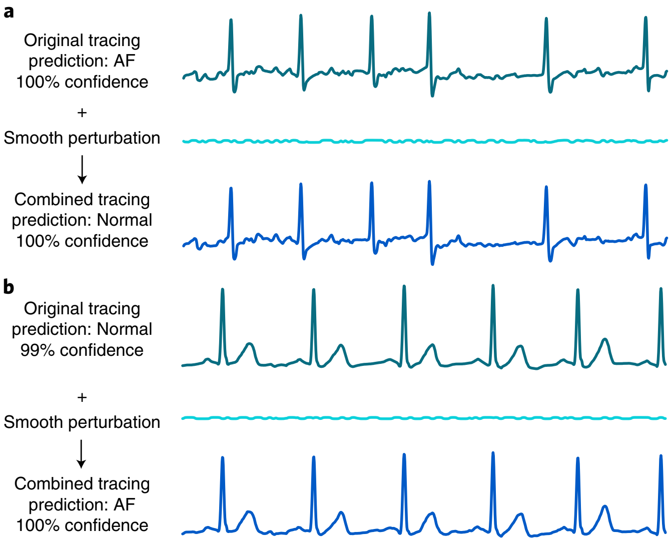


Fig. 1 | Demonstration of disruptive adversarial examples. a, Example of an original ECG tracing that was correctly diagnosed by the network as atrial fibrillation (AF) with 100% confidence, but, after the addition of smooth perturbations, was diagnosed wrongly as normal sinus rhythm (Normal) with 100% confidence. **b**, Example of an original ECG tracing that was correctly diagnosed by the network as Normal with 100% confidence, but after the addition of smooth perturbations was diagnosed wrongly as AF. Perturbation and tracing voltages are plotted on the same scale.

By taking a weighted average of nearby time steps, we crafted smooth adversarial examples that cannot be distinguished from original ECG signals but will still fool the deep network to make a wrong prediction (see Methods).

We generated adversarial examples on the test set. We transformed the test examples to make the network change the label of Normal, Other and Noise to any other label. For AF, we altered the AF test examples so that the deep neural network classifies them as Normal. We can also alter Normal, AF, Other and Noise to any given label. The results are presented in Table 1. Misdiagnosis of AF as Normal may increase the risk of AF-related complications such as stroke and heart failure. We showcase the generation of adversarial examples in Fig. 1.

After adversarial attacks, 74% of the test ECGs originally classified correctly by the network are now assigned a different diagnosis, ultimately showing that deep ECG classifiers are vulnerable to adversarial examples. To assess how the generated signals would be classified by human experts, we invited one board-certified medicine specialist and one cardiac electrophysiology specialist to diagnose whether signals generated by our methods and original ECGs come from the same class. Figure 2a shows that almost all of the modified signals were judged as belonging to the same class as the original signal. This shows that the deep network failed to correctly classify most of the newly generated examples, when a human would have assigned only 1.4% of them to a different class.

We also invited the clinical specialists to distinguish ECG signals from the adversarial examples generated by our smooth method and the traditional attack method based on 'projected gradient descent' (PGD)[12,13]. The question we asked was 'Which one in the pair is the real ECG?' The calculated probability of correctly identifying ECGs is the number of correct answers they obtained for each pair over the number of pairs we showed them. The doctors were not shown adversarial examples beforehand.

Figure 2b shows that the adversarial examples generated by our method are significantly harder for clinicians to distinguish from the original ECG than the traditional attack method. On average, the clinicians were able to correctly identify the smoothed adversarial examples from their original counterpart 62% of the time. (The electrophysiology specialist was slightly more accurate at 65% versus 59%.) PGD examples are easier for clinicians to detect because of square-wave discontinuity artefacts that are not physiologically plausible. These discontinuities also appear in PGD examples of images, but in images they are hidden by the resolution and color channels.

Here, we provide a construction that shows that adversarial examples are not rare. In particular, we show that it is possible to create more examples that remain adversarial by adding a small amount of Gaussian noise to an original adversarial example and then smoothing the result. We repeated this process 1,000 times and found that the deep neural network still incorrectly classified all 1,000 new, adversarial examples. Adding Gaussian noise could still produce adversarial examples on 87.6% of the test examples from which adversarial examples were generated. We plotted all of the newly crafted adversarial examples, which form a band around the original ECG signal, as shown in Fig. 3. The signals in the band may intersect. We chose pairs of intersecting signals and concatenated the left half of one signal with the right half of the other to create a new example. We found that signals created by concatenation are also adversarial examples. We also sampled random values in the band for each time step and then smoothed them to create new adversarial examples. These different perturbations on adversarial examples all led to new examples that remained mislabeled. This means that the adversarial examples should not be considered as rare isolated cases, in that from a single adversarial example, many more can be created.

The use of machine learning algorithms as a healthcare tool for clinical interpretation and prediction is seeing an unprecedented surge. A search in PubMed for the phrases 'electrocardiogram' AND ('machine learning' OR 'artificial intelligence') yields over 1,200 publications. Specifically, deep learning has been utilized recently to create algorithms that predict the ejection fraction[14], predict the susceptibility to QT prolongation in patients with normal QT intervals (https://www.alivecor.com/research/investigational-qt/artificial-intelligence-and-deep-neural-networks/) and identify patients with hyperkalemia[15]—all based on the ECG and demographics, without any additional clinical information. This promising ability of deep learning algorithms to reduce the cost or improve the performance of complex and laborious daily clinical challenges is creating significant incentive for rapid implementation and approval as practical clinical tools. Correspondingly, 23 machine learning algorithms, many that use deep learning, have been approved by the FDA for medical use in 2018 alone, a 283% increase from 2017 (https://medicalfuturist.com/fda-approvals-for-algorithms-in-medicine/). Products for arrhythmia classification with single-lead ECGs such as the Apple Watch, which sold over 20 million units

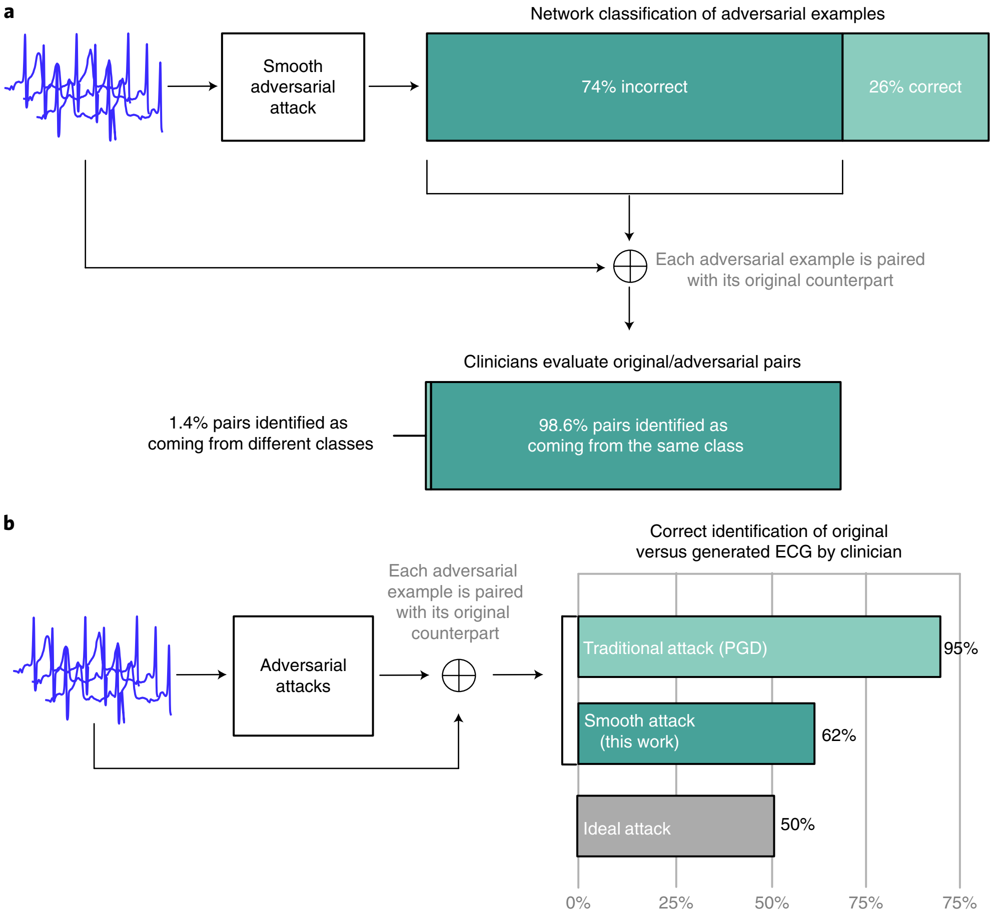


Fig. 2 | Accuracy of the network in classifying adversarial examples and clinician success rate in distinguishing authentic ECGs from adversarial examples. a, Top: schematic showing that a smooth adversarial attack generated adversarial examples of ECG tracings that were misclassified in 74% of cases. Bottom: schematic showing that the surveyed clinicians concluded that 246.5/250 pairs of adversarial examples and the original ECGs belonged to the same class. **b**, Schematic showing the success rate of ECG interpretation experts in distinguishing between 100 pairs of original ECGs and adversarial examples generated by the traditional attack method (PGD), the smooth attack method and the ideal attack method. The ideal attack method creates signals that clinicians cannot distinguish completely from the original signals.

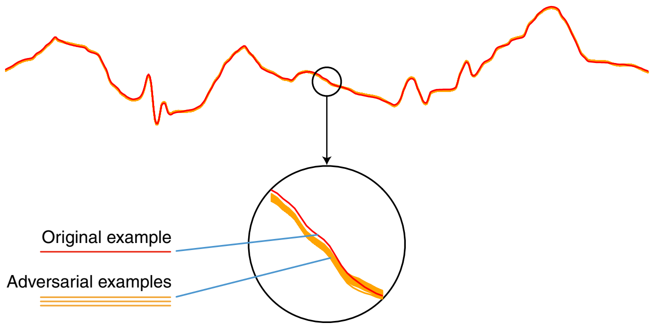


Fig. 3 | Perturbing a known adversarial example to generate multiple new ones. Schematic showing that 1,000 different adversarial examples can be generated from the original ECG signal by adding small Gaussian noise and smoothing. The newly generated adversarial examples, as well as the original ECG signal, are plotted at the top. A portion of the original ECG signal and adversarial examples is enlarged in the circle below; the newly generated examples form a wide band around the original example.

in 2018 alone (https://ww.9to5mac.com/2019/02/27/apple-watch-sales-2018/), and the Alivecor Kardia, which uses deep learning and has recorded over 25 million ECGs (https://www.alivecor.com/press/press_release/alivecor-data-yield-reaches-25mm-ecgs/), are increasingly being adopted. Hence, it is imperative to understand the limitations and vulnerabilities of the deep learning algorithms used to detect arrhythmia from ECGs.

In this work, we demonstrate the ability to add imperceptible perturbations to ECG tracings to create adversarial examples that fool a deep neural network classifier into assigning the examples to an incorrect rhythm class. Moreover, we show that such examples are not rare.

These findings question the safety of using deep learning in analyzing ECGs at a scale where millions of tests may be run every week by widespread consumer devices. To increase robustness to adversarial examples, it is crucial that classification methods for ECGs, especially those intended to operate without human supervision, generalize well to new examples. However, generalization may be a significant challenge, because different environments and different devices can introduce unknown perturbations to the signal. Thus, ensuring safe generalization would require obtaining labeled data from each new environment and new device.

One way to protect against adversarial examples is adversarial training. Adversarial training works by generating adversarial examples repeatedly during model training based on the current model and adding them to the training batch used to improve the model[12]. However, such approaches can only protect against known adversarial examples, created with a given specific attack method, and are not guaranteed to protect against future attack methods (see Methods for detailed discussions). A more direct approach would be to certify deep neural networks for robustness with mathematical proofs[16,17], as suggested for other safety-critical domains, such as the aviation industry[18].

The possibility to construct even a single adversarial example may still enable malicious actors to inject small perturbations into RWD that are indistinguishable to the human eye. Indistinguishability matters to malicious actors to ensure they cannot be discovered by human auditors. The ability to create adversarial examples is an important issue, with future implications including robustness to the environmental noise of medical devices that rely on ECG interpretation (for example, pacemakers and defibrillators)[19], the skewing of data to alter insurance claims[4] and the introduction of intentional bias into clinical trials. For example, imagine a large clinical trial intended to assess the effect of a treatment on reducing arrhythmias. Such a trial could use a pretrained neural network to identify how many arrhythmias occurred. Attacking this pretrained network could inject bias into the clinical trial by changing ECGs to reduce the number of documented arrhythmias. To prevent such possibilities, it is paramount that platforms for the collection and analysis of RWD implement principles from 'trusted computing' to provide trusted data provenance guarantees that can certify that data has not been tampered with from device acquisition to any downstream analysis[20]. In this vein, closed systems without access to the raw ECG reduce malicious actors' practical ability to attack systems. For devices where only the test signal can be modified, it is still possible to train a different network and generate adversarial examples using the blackbox attack from ref. [12]. For commercial devices such as the Apple Watch, malicious actors can potentially get access to many forward passes and construct examples using only the forward passes to solve the optimization problem used to construct smoothed adversarial examples[21]. If access to the full model is available, gradient-based attacks can be directly implemented.

One thing to note is that the lack of robustness observed is not inherent to the use of statistical methods to classify ECGs. Humans tend to be more robust to small perturbations because they use coarser visual features to classify ECGs, such as the R–R interval and the P-wave morphology. These features change less under small perturbations and generalize better to new domains. To automate the classification of more complex ECG tracings, it may be useful to incorporate hierarchical coarse pattern-dependent classification models along with deep learning to not only increase robustness to adversarial attacks but also to improve network accuracy. Additionally, regularizing deep networks to prefer coarser features can improve robustness. Tree-based learning algorithms using coarser features will also be more robust to attacks.

In conclusion, with this work, we do not intend to cast a shadow on the utility of deep learning for ECG analysis, which undoubtedly will be useful to handle the volumes of physiological signals requiring processing in the near future. This work should, instead, serve as an additional reminder that machine learning systems deployed in the wild should be designed with safety and reliability in mind[22], with a particular focus on training data curation and provable guarantees on performance.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41591-020-0791-x.

Received: 1 June 2019; Accepted: 5 February 2020;
Published online: 9 March 2020

References

1. Hannun, A. Y. et al. Cardiologist-level arrhythmia detection and classification in ambulatory electrocardiograms using a deep neural network. *Nat. Med.* **25**, 65–69 (2019).
2. Szegedy, C. et al. Intriguing properties of neural networks. In *International Conference on Learning Representations* (ICLR, 2014).
3. Biggio, B. et al. Evasion attacks against machine learning at test time. In *Machine Learning and Knowledge Discovery in Databases* Vol. 8190 (eds Blockeel, H., Kersting, K., Nijssen, S. & Železný, F.) 387–402 (Springer, 2013).
4. Finlayson, S. G. et al. Adversarial attacks on medical machine learning. *Science* **363**, 1287–1289 (2019).
5. Paschali, M., Conjeti, S., Navarro, F. & Navab, N. Generalizability vs. robustness: investigating medical imaging networks using adversarial examples. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 493–501 (Springer, 2018).
6. Clifford, G. D. et al. AF Classification from a short single lead ECG recording: the PhysioNet/Computing in Cardiology Challenge 2017. In *2017 Computing in Cardiology* (*CinC*) 1–4 (IEEE, 2017).
7. Kelly, B. B. & Fuster, V. *Promoting Cardiovascular Health in the Developing World: a Critical Challenge to Achieve Global Health* (National Academies Press, 2010).
8. Kennedy, H. L. The evolution of ambulatory ECG monitoring. *Prog. Cardiovasc. Dis.* **56**, 127–132 (2013).
9. Hong, S. et al. ENCASE: an ENsemble ClASsifiEr for ECG classification using expert features and deep neural networks. In *2017 Computing in Cardiology* (*CinC*) 1–4 (IEEE, 2017).
10. Chen, H., Huang, C., Huang, Q. & Zhang, Q. ECGadv: generating adversarial electrocardiogram to misguide arrhythmia classification system. Preprint at *arXiv* https://arxiv.org/pdf/1901.03808.pdf (2019).
11. Goodfellow, S. D. et al. Towards understanding ECG rhythm classification using convolutional neural networks and attention mappings. In *Proceedings of the 3rd Machine Learning for Healthcare Conference* 83–101 (PMLR, 2018).
12. Madry, A., Makelov, A., Schmidt, L., Tsipras, D. & Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations* (ICLR, 2018).
13. Kurakin, A., Goodfellow, I. & Bengio, S. Adversarial machine learning at scale. In *International Conference on Learning Representations* (ICLR, 2017).
14. Kwon, J.-m et al. Development and validation of deep-learning algorithm for electrocardiography-based heart failure identification. *Korean Circ. J.* **49**, 629–639 (2019).
15. Galloway, C. D. et al. Development and validation of a deep-learning model to screen for hyperkalemia from the electrocardiogram. *JAMA Cardiol.* **4**, 428–436 (2019).
16. Singh, G., Gehr, T., Mirman, M., Püschel, M. & Vechev, M. Fast and effective robustness certification. In *Advances in Neural Information Processing Systems* 10802–10813 (NIPS, 2018).
17. Cohen, J., Rosenfeld, E. and Kolter, Z. Certified Adversarial Robustness via Randomized Smoothing. In *International Conference on Machine Learning* 1310–1320 (ICML, 2019).
18. Julian, K. D., Kochenderfer, M. J. & Owen, M. P. Deep neural network compression for aircraft collision avoidance systems. *J. Guid. Control Dyn.* **42**, 598–608 (2018).
19. Nguyen, M. T., Van Nguyen, B. & Kim, K. Deep feature learning for sudden cardiac arrest detection in automated external defibrillators. *Sci. Rep.* **8**, 17196 (2018).
20. Lyle, J. & Martin, A. *Trusted Computing and Provenance: Better Together* (Usenix, 2010).
21. Uesato, J., O'Donoghue, B., Kohli, P. and Oord, A. Adversarial Risk and the Dangers of Evaluating Against Weak Attacks. In *International Conference on Machine Learning* 5025-5034 (ICML, 2018).
22. Saria, S. & Subbaswamy, A. Tutorial: safe and reliable machine learning. Preprint at https://arxiv.org/pdf/1904.07204.pdf (2019).

Methods

Description of the traditional attack methods. Two traditional attack methods are the 'fast gradient sign method' (FGSM)[23] and PGD[12,13]. These are white-box attack methods based on the gradients of the loss used to train the model with respect to the input. Both FGSM and PGD can be used for targeted and untargeted attacks. Targeted attacks force the network to output a specific incorrect label, while untargeted attacks force the network to make any wrong classification. Untargeted attacks usually minimize the probability of the true class, while targeted attacks maximize the probability of the target class.

Denote our input entry **x**, true label *y*, classifier (network) *f* and loss function *L*(*f*(**x**),*y*). We describe FGSM and PGD in the following.

- Untargeted attack

- FGSM: FGSM is a fast algorithm. For an attack level ϵ, FGSM sets

$$\mathbf{x}_{\text{adv}} = \mathbf{x} + \epsilon \text{sign}(\nabla_{\mathbf{x}} L(f(\mathbf{x}), y))$$

The attack level is chosen to be sufficiently small so as to be undetectable.

- PGD: PGD is an improved version that uses multiple iterations of FGSM. Define $\text{Clip}_{\mathbf{x},\epsilon}(\mathbf{x}')$ to project each $\mathbf{x}'$ back to the infinity norm ball by clamping the maximum absolute difference value between $\mathbf{x}$ and $\mathbf{x}'$ to ϵ. Beginning by setting $\mathbf{x}'_0 = \mathbf{x}$, we have

$$\mathbf{x}'_i = \text{Clip}_{\mathbf{x},\epsilon}\left(\mathbf{x}'_{i-1} + \alpha \text{sign}\left(\nabla_{\mathbf{x}} L\left(f\left(\mathbf{x}'_{i-1}\right), y\right)\right)\right) \tag{1}$$

After T steps, we get our adversarial example $\mathbf{x}_{\text{adv}} = \mathbf{x}'_T$.

- Targeted attack (target class *t*)

- FGSM: For an attack level ϵ, FGSM sets

$$\mathbf{x}_{\text{adv}} = \mathbf{x} - \epsilon \text{sign}(\nabla_{\mathbf{x}} L(f(\mathbf{x}), t))$$

- PGD: Beginning by setting $\mathbf{x}'_0 = \mathbf{x}$, we have

$$\mathbf{x}'_i = \text{Clip}_{\mathbf{x},\epsilon}\left(\mathbf{x}'_{i-1} - \alpha \text{sign}\left(\nabla_{\mathbf{x}} L\left(f\left(\mathbf{x}'_{i-1}\right), t\right)\right)\right)$$

Unlike untargeted attacks, the gradient is subtracted. After T steps, we get our adversarial example $\mathbf{x}_{\text{adv}} = \mathbf{x}'_T$.

In this Letter, we use targeted attacks to change AF to Normal and untargeted attacks on classes besides AF.

Our smooth attack method. To smooth the signal, we use convolution. Convolution takes the weighted average of one position of the signal and its neighbors:

$$(\mathbf{a} \circledast \mathbf{v})[n] = \sum_{m=1}^{2K+1} \mathbf{a}[n-m+K+1] \times \mathbf{v}[m]$$

where **a** is the objective function and **v** is the weight or kernel function. In our experiment, the weights are determined by a Gaussian kernel. Mathematically, if we have a Gaussian kernel of size 2*K*+1 and standard deviation σ, we have

$$\mathbf{v}[m] = \frac{\exp\left(-\frac{(m-K-1)^2}{2{*}\sigma^2}\right)}{\sum_{i=1}^{2K+1} \exp\left(-\frac{(\mathrm{i}-\mathrm{K}-1)^2}{2{*}\sigma^2}\right)}$$

We can easily see that when σ goes to infinity, the convolution with the Gaussian kernel becomes a simple average; when σ goes to zero, the convolution becomes the identity function. Instead of taking an adversarial perturbation and then convolving it with the Gaussian kernels, we could create adversarial examples by optimizing a smooth perturbation that fools the neural network. We introduce our method of training 'smooth adversarial perturbations' (SAP). In our SAP method, we take the adversarial perturbation as the parameter $\boldsymbol{\theta}$ and add it to the clean examples after convolving with a number of Gaussian kernels. We denote **K**(*s*,σ) to be a Gaussian kernel with size *s* and standard deviation σ. The resulting adversarial example can be written as a function of $\boldsymbol{\theta}$:

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

In our experiment, we let *s* be {5, 7, 11, 15, 19} and σ be {1.0, 3.0, 5.0, 7.0, 10.0}.

Then we try to maximize the loss function with respect to $\boldsymbol{\theta}$ to get the adversarial example in an untargeted attack. We still use PGD, but this time on $\boldsymbol{\theta}$:

$$\theta'_i = \text{Clip}_{\mathbf{0},\epsilon}\left(\theta'_{i-1} + \alpha \text{sign}\left(\nabla_{\theta} L\left(f\left(\mathbf{x}_{\text{adv}}\left(\theta'_{i-1}\right), y\right)\right)\right)\right) \tag{2}$$

There are two major differences between update equations (2) and (1). In equation (2), we update $\boldsymbol{\theta}$, not $\mathbf{x}_{\text{adv}}$, and clip around zero, not the input **x**. In practice, we initialize the adversarial perturbation $\boldsymbol{\theta}$ to be the one obtained from PGD $(\epsilon = 10, \alpha = 1, T = 20)$ on **x** and run another PGD $(\epsilon = 10, \alpha = 1, T = 40)$ on $\boldsymbol{\theta}$.

For targeted attacks (target class *t*), the update is

$$\theta'_i = \text{Clip}_{\mathbf{0},\epsilon}\left(\theta'_{i-1} - \alpha \text{sign}\left(\nabla_{\theta} L\left(f\left(\mathbf{x}_{\text{adv}}\left(\theta'_{i-1}\right), t\right)\right)\right)\right)$$

If we take the same combination of convolution on the adversarial examples generated in PGD to create smooth adversarial examples, 71% of the originally correctly classified test ECGs are assigned different labels, which is worse than our smooth attack method (74%). The idea of optimizing the parameters of a smooth model could be expanded to other models, such as differential equation models of ECGs[24], to find adversarial examples that more closely match human physiology.

Existence of adversarial examples. Our experiments are designed to show that adversarial examples are not rare. We only discuss untargeted attacks, but it is easy to extend our analysis to targeted attacks. We denote the original signal *x* and the generated adversarial example $\mathbf{x}_{\text{adv}}$.

First, we generate Gaussian noise $\boldsymbol{\delta}$ such that $\delta[i] \sim \mathcal{N}(0, 25)$ and then add it to the adversarial examples. To make sure the new examples are still smooth, we smooth the perturbation by convolving with the same Gaussian kernels as in our smooth attack method. We then clip the perturbation to make sure that it is still in the infinity norm ball. The newly generated example is

$$\mathbf{x}'_{\text{adv}} = \mathbf{x} + \text{Clip}_{\mathbf{0},\epsilon}\left(\frac{1}{m}\sum_{i=1}^{m}(\mathbf{x}_{\text{adv}} + \delta - \mathbf{x}) \circledast \mathbf{K}(s[i], \sigma[i])\right)$$

We repeat the process of generating new examples 1,000 times. These newly generated examples are still adversarial examples. Some of them may intersect. For each intersected pair, we concatenate the left part of one example and the right part of the other to create new adversarial examples. Denote $\mathbf{x}_1$ and $\mathbf{x}_2$ to be a pair of adversarial examples that intersect. Suppose they intersect at time step *t* and the total length of the example is *T*. The new hybrid example $\mathbf{x}'$ satisfies

$$\mathbf{x}'[1:t] = \mathbf{x}_1[1:t]; \quad \mathbf{x}'[t+1:T] = \mathbf{x}_2[t+1:T]$$

where [1:*t*] means from time step 1 to time step *t*. All the newly concatenated examples are still misclassified by the network.

The 1,000 adversarial examples form a band. To emphasize that all the smooth signals in the band are still adversarial examples, we sample uniformly from the band to create new examples. Denote max[*t*] and min[*t*] to be the maximum value and minimum value of 1,000 samples at time step *t*. To sample a smooth signal from the band, we first sample a uniform random variable $\mathbf{a}[t] \approx \mathcal{U}(\min[t], \max[t])$ for each time step *t* and then we smooth the perturbation. The example generated by uniform sampling and smoothing is

$$\mathbf{x}'_{\text{adv}} = \mathbf{x} + \text{Clip}_{\mathbf{0},\epsilon}\left(\frac{1}{m}\sum_{i=1}^{m}(\mathbf{a} - \mathbf{x}) \circledast \mathbf{K}(s[i], \sigma[i])\right)$$

We repeat this procedure 1,000 times. All of the newly generated examples still cause the network to make the wrong diagnosis. We visualize the three procedures to show the existence of adversarial examples in Extended Data Fig. 2.

Limitations of adversarial training. Adversarial training[12] is a more effective method to build robust models than including adversarial examples in the training data. However, adversarial training does well only on small image datasets like MNIST[25], not larger ones like CIFAR10[26]. For CIFAR10, even dynamically including adversarial examples while training the model will not lead to a robust model[27]. In addition, there is no formal guarantee that adversarial training implemented with PGD can converge to the saddle point of the infinity norm minimax formulation of adversarial training. For example, switching to a higher-order optimizer may produce different adversarial examples not captured by PGD-based adversarial training.

Statistics and reproducibility. Figure 1a,b was generated for 50 AF signals and 124 normal sinus rhythms. Figure 3 was generated twice. Extended Data Fig. 1 was generated for 40 examples. We obtained similar results for the examples we generated.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The dataset can be accessed from https://physionet.org/challenge/2017/.

Code availability

The code is available from https://github.com/XintianHan/ADV_ECG.

References

23. Goodfellow, I. J., Shlens, J. & Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations* (ICLR, 2015).
24. McSharry, P. E., Clifford, G. D., Tarassenko, L. & Smith, L. A. A dynamical model for generating synthetic electrocardiogram signals. *IEEE Trans. Biomed. Eng.* **50**, 289–294 (2003).
25. LeCun, Y., Bottou, L., Bengio, Y. & Haffner, P. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).
26. Krizhevsky, A. & Hinton, G. *Learning Multiple Layers of Features from Tiny Images* (Citeseer, 2009).
27. Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K. & Madry, A. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems* 5014–5026 (NIPS, 2018).

Acknowledgements

X.H. and R.R. were supported in part by the NIH under award no. R01HL148248. We thank W. Lee, S. Mohan, M. Goldstein, A. Li, A.M. Puli, H. Singh, M. Sudarshan, and W. Whitney.

Author contributions

X.H. and R.R. designed the problem and performed all the experiments. All authors wrote the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

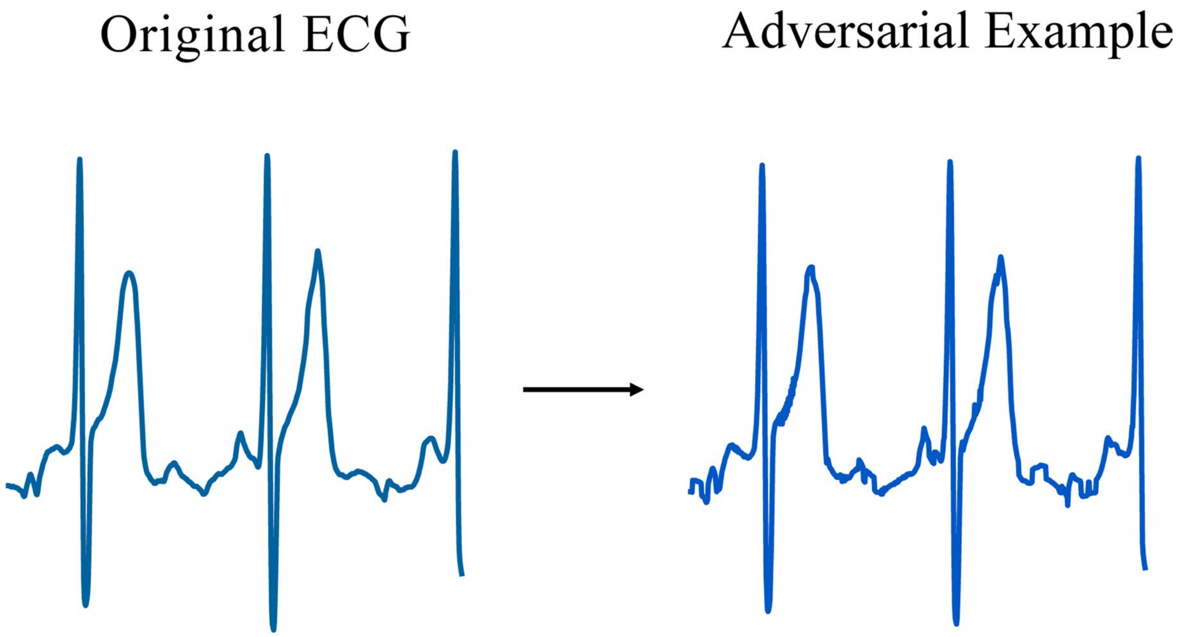
Extended data is available for this paper at https://doi.org/10.1038/s41591-020-0791-x.

Supplementary information is available for this paper at https://doi.org/10.1038/s41591-020-0791-x.

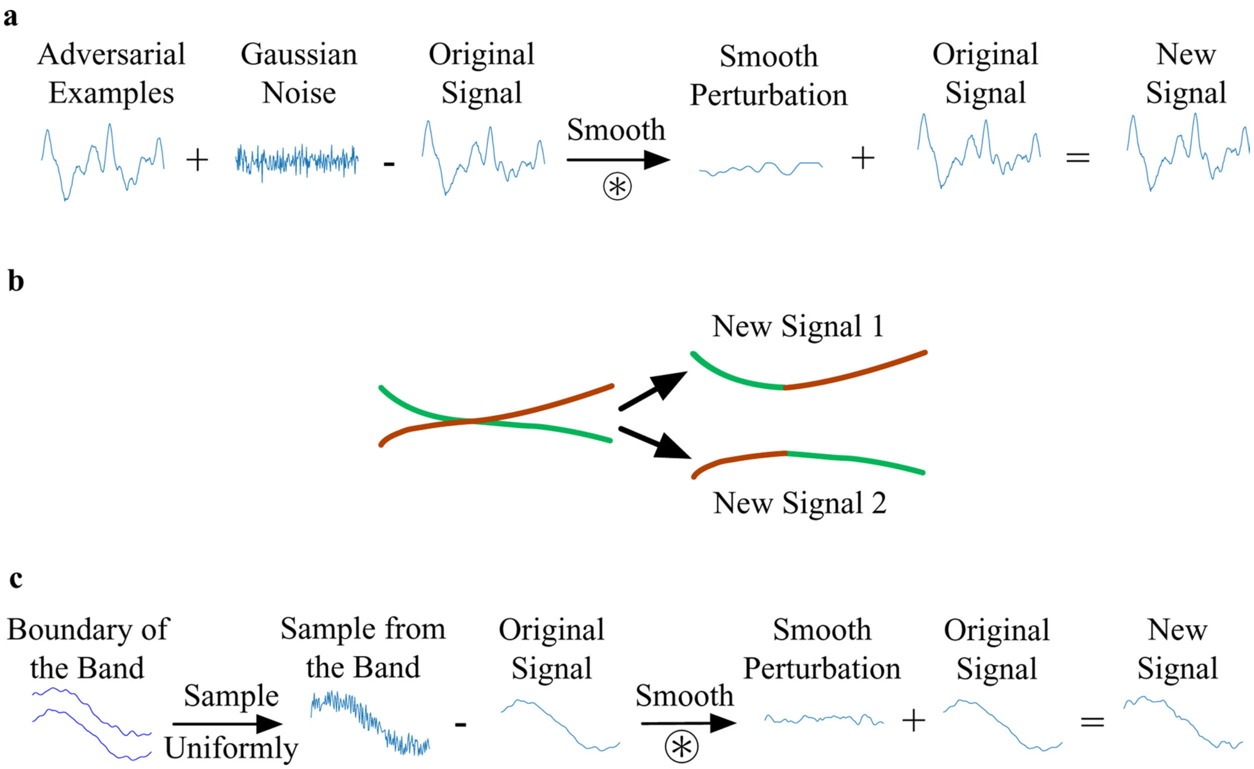
Correspondence and requests for materials should be addressed to X.H. or R.R.

Peer review information Michael Basson was the primary editor on this article and managed its editorial process and peer review in collaboration with the rest of the editorial team.

Reprints and permissions information is available at www.nature.com/reprints.



Extended Data Fig. 1 | An adversarial example created by Projected Gradient Descent (PGD) method. This adversarial example contains square waves that are physiologically implausible.



Extended Data Fig. 2 | Demonstration of three procedures to show the existence of the adversarial examples. a, We add a small amount of Gaussian noise to the adversarial example and smooth it to create a new signal. **b**, For intersected signals, we concatenate the left half of one signal with the right half of the other to create a new one. **c**, We sample uniformly from the band and smooth to create a new signal.

nature research

Corresponding author(s): Xintian Han, Rajesh Ranganath

Last updated by author(s): Feb 4, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a | Confirmed

☐ ☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement

☒ ☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly

☒ ☐ The statistical test(s) used AND whether they are one- or two-sided
Only common tests should be described solely by name; describe more complex techniques in the Methods section.

☒ ☐ A description of all covariates tested

☒ ☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons

☒ ☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)

☒ ☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted
Give P values as exact values whenever suitable.

☒ ☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings

☒ ☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes

☒ ☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The data is publicly available at (https://physionet.org/challenge/2017/}{https://physionet.org/challenge/2017/). We use python 3.6 (https://www.python.org/downloads/release/python-360/) to process the data. The processing code can be found at https://github.com/XintianHan/ADV_ECG/blob/master/utils/create_data.py.
Data analysis	We use python 3.6 (https://www.python.org/downloads/release/python-360/) and pytorch 1.0 (https://pytorch.org/)) to analysis the data. The analysis code can be found at https://github.com/XintianHan/ADV_ECG.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The dataset can be accessed from https://physionet.org/challenge/2017/

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	We used the public available dataset (https://physionet.org/challenge/2017/) contained 8,528 single-lead ECG recordings lasting from 9s to about 60s, including 5,076 Normal, 758 AF, 2,415 Other, and 279 Noise examples. 90% of the data set was used for training and 10% was used for testing.
Data exclusions	No data was excluded.
Replication	Fig. 1a, b were generated for 50 AF signals and 124 normal sinus rhythms. Fig. 3 was generated twice. Extended Data Fig. 1 was generated for 40 examples. We got similar results for the examples we generated.
Randomization	We separate the train and test dataset randomly.
Blinding	The investigators were blinded for train/test split. It was done randomly by the python program.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging