Hard-Label Black-Box Adversarial Attack on Deep Electrocardiogram Classifier

Jonathan Lam*
University of California, Los Angeles
Los Angeles, California
jlam7@g.ucla.edu

Pengrui Quan*
University of California, Los Angeles
Los Angeles, California
prquan@g.ucla.edu

Jiamin Xu*
University of California, Los Angeles
Los Angeles, California
jiaminxu1019@g.ucla.edu

Jeya Vikranth Jeyakumar
University of California, Los Angeles
Los Angeles, California
vikranth94@g.ucla.edu

Mani Srivastava
University of California, Los Angeles
Los Angeles, California
mbs@ucla.edu

ABSTRACT

Through aiding the process of diagnosing cardiovascular diseases (CVD) such as arrhythmia, electrocardiograms (ECGs) have progressively improved prospects for an automated diagnosis system in modern healthcare. Recent years have seen the promising applications of deep neural networks (DNNs) in analyzing ECG data, even outperforming cardiovascular experts in identifying certain rhythm irregularities. However, DNNs have shown to be susceptible to adversarial attacks, which intentionally compromise the models by adding perturbations to the inputs. This concept is also applicable to DNN-based ECG classifiers and the prior works generate these adversarial attacks in a white-box setting where the model details are exposed to the attackers. However, the black-box condition, where the classification model's architecture and parameters are unknown to the attackers, remains mostly unexplored. Thus, we aim to fool ECG classifiers in the black-box and hard-label setting where given an input, only the final predicted category is visible to the attacker. Our attack on the DNN classification model for the *PhysioNet Computing in Cardiology Challenge 2017* [12] database produced ECG data sets mostly indistinguishable from the white-box version of an adversarial attack on this same database. Our results demonstrate that we can effectively generate the adversarial ECG inputs in this black-box setting, which raises significant concerns regarding the potential applications of DNN-based ECG classifiers in security-critical systems.

CCS CONCEPTS

• **Theory of computation → Adversary models**; • **Computing methodologies → Machine learning**; • **Hardware →** *Sensors and actuators.*

KEYWORDS

deep neural network, adversarial attack

ACM Reference Format:
Jonathan Lam, Pengrui Quan, Jiamin Xu, Jeya Vikranth Jeyakumar, and Mani Srivastava. 2020. Hard-Label Black-Box Adversarial Attack on Deep Electrocardiogram Classifier. In *The 1st ACM International Workshop on Security and Safety for Intelligent Cyber-Physical Systems (SecICPS '20), November 16–19, 2020, Virtual Event, Japan.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3417312.3431827

1 INTRODUCTION

The World Health Organization has reported that cardiovascular diseases (CVDs) are the leading cause of death, taking an estimated 17.9 million lives each year (31% of all deaths worldwide) [2]. In certain cases, the irregularities may occur sporadically in a patient's daily life. Detecting these irregularities within a patient's heartbeat allows them to receive potentially life-saving treatment that prevents this issue from manifesting into a more severe condition, such as a heart attack or stroke. The main abnormality we focus upon is arrhythmia, a type of CVD where there is an irregularity in the rate or rhythm of the heartbeat and is usually diagnosed by analyzing recordings of the electrical activity of the heart (ECGs). Unfortunately, a study on cardiology in the U.S. [24] predicted as much as an 18% increase in demand for cardiologists over the next 10 years. With such a limited number available, it's impractical to have cardiologists manually analyze all ECG data.

As the pervasiveness of machine learning and neural networks continues to expand into all walks of life, (from social media advertising to stock market trends to speech recognition) there have been numerous developments in their applications towards the automation of healthcare practices. Google[1] has recently developed a machine-learning algorithm to identify cancerous tumors in mammograms, while researchers have utilized DNNs to identify types of skin cancer from images. Furthermore, recent works have shown that DNNs can accurately recognize and classify different types of arrhythmia [23, 28] from ECG signals. As the automation of diagnosing arrhythmia and other CVDs becoming a much more realistic endeavor with the assistance of DNNs, regions lacking a sufficient number of cardiologists would greatly benefit from this.

*The three authors contributed equally to this research.

SecICPS '20, November 16–19, 2020, Virtual Event, Japan

ACM ISBN 978-1-4503-8133-8/20/11.
https://doi.org/10.1145/3417312.3431827

[1] https://research.google/teams/health/

Companies like Tricog [1] have already adopted these machine-learning based methods to accomplish these healthcare goals in over 12 countries, including China, Singapore, India, and Malaysia.

Adversarial Attacks: However, it has been observed that DNNs are susceptible to *adversarial attacks*, where inputs are intentionally designed to cause a misclassification. For instance, previous research has demonstrated that while a well-trained image classifier exceeds the abilities of the average human in recognizing objects, it can be fooled by adding almost imperceptible perturbations to its inputs. [6, 8, 16, 32]. This phenomenon is also encountered in time-series data, with prominent examples including speech signals [3, 25] and ECG data [9, 17, 18]. Such adversarial attacks pose a serious threat to medical deep-learning systems, ultimately hindering the deployment of DNNs for such applications. Without proper defense mechanisms against adversarial attacks, the risk of mistreatment or medical fraud occurring would be significantly higher if these susceptible diagnosis systems were to replace certain tasks normally completed by experienced doctors or cardiologists. [14].

Despite the looming concern of such consequential effects from the use of recently developed ECG diagnostic systems, these time-domain based attacks have seldom been explored relative to the image and speech domains. Previous works have only shown that ECG classifiers can be fooled by adding imperceptible perturbations [9, 18] in the white-box setting, where attacks receive far more assistance and information with regards to the given architecture and parameters of the DNNs. In this paper, we focus on attacking ECG classifiers in a more realistic hard-label and black-box setting where the attacker only has access to the final top-1 category of the deep-learning model especially.

Details of the Study: We conduct our study using the *PhysioNet Computing in Cardiology Challenge 2017* which contains single-lead ECG signals of various classifications. Unlike prior works [9, 18] which functioned under white-box conditions, our attack performs under the black-box and hard-label setting. We demonstrate that our algorithm can still produce imperceptible data sets that effectively fool a modern ECG classifier at a rate on par with the white-box attack. To improve the perceptual similarity of the adversarial ECG data set, we propose an approach to smoothing the adversarial signal while performing random searches during the attack (Fig. 2). Our experiments demonstrate that the perceptual similarity improved significantly. As a result, our attack indicates that a deep-learning based ECG diagnostic system can be compromised without knowledge of any information regarding the model.

Key Contributions: This paper's contributions can be summarized as follows:

- We adversarially attack ECG classifiers in the black-box hard-label setting. This is conducted in a computationally efficient manner with a fairly minimal number of queries.
- We propose a smoothing step in our attack framework to improve the human perceptual quality of the adversarial ECG signals. In addition, we demonstrate that the proposed attack can perform comparably with an attack in the white-box setting in creating imperceptible, adversarial ECG data sets in the time domain.
- We determine the success rate of creating imperceptible signals via qualified physicians, showing that the adversarially generated ECG data sets and their respective original data sets are consistently categorized the same.

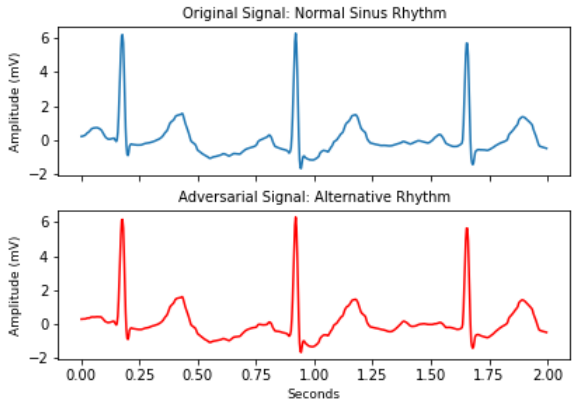

Figure 1: Example of Successful Black-Box Attack in Time Domain

2 BACKGROUND

In recent years, various strategies have been proposed to fool deep neural networks. The attacks can be classified into two categories: targeted and untargeted. In this paper, we focus on the targeted version where the perturbed input must be categorized as a chosen class. Furthermore, based on knowledge of the classifying model, the attacks can be further divided into two classes: white-box attacks and black-box. White-box attacks assume complete knowledge of the trained models, including its architecture and parameters. Conversely, black-box attacks assume the attacker has no knowledge of the information mentioned above and is simply acting as a standard user. This setting is far more practical when attacking online machine-learning services like Google Cloud AI[2] and AWS Machine Learning[3].

Problem Formulation. The following is a formulation of the targeted attack problem. Consider the benign ECG signal $\mathbf{x}^{org} \in \mathbb{R}^N$. A C-class ECG classifier produces the corresponding top-1 prediction $y \in \{1, ..., C\}$. An attacker generally aims to minimize a distance metric between $\mathbf{x}^{adv}$ and $\mathbf{x}^{org}$ with the misclassification constraint satisfied:

$$\underset{\mathbf{x}^{adv} \in \mathbb{R}^N}{\text{minimize}} \quad D(\mathbf{x}^{adv}, \mathbf{x}^{org}) \tag{1a}$$

$$\text{subject to} \quad f(\mathbf{x}^{adv}) = t \tag{1b}$$

The L_p-distance is commonly used as the objective metric, as it computes element-wise distances. By minimizing the L_p-distance, the attacker seeks to make the adversarial inputs inconspicuous from the human perspective.

Adversarial Attacks: White-Box Setting. Equation (1) can be combined and simplified as follows:

$$\underset{\mathbf{x}^{adv} \in \mathbb{R}^N}{\text{minimize}} \; D(\mathbf{x}^{adv}, \mathbf{x}^{org}) + \lambda F(\mathbf{x}^{adv}) \tag{2}$$

[2]https://cloud.google.com/products/ai
[3]https://aws.amazon.com/cn/machine-learning/

$F(\cdot)$ is an objective function, mapping the input to a non-negative number. $F(\mathbf{x}^{adv}) = 0$ if and only if $f(\mathbf{x}^{adv}) = t$. An example of a potential $F(\cdot)$ is:

$$F(\mathbf{x}^{adv}) = \max(\max_{i \neq t} g(\mathbf{x}^{adv})_i - g(\mathbf{x}^{adv})_t, 0) \tag{3}$$

where $g(\cdot)_i$ is the logit output corresponding to the i^{th} class. Hence, in the white-box setting, since the attackers have access to all information regarding the model, the above problem is differentiable and can be solved via gradient descent.

2.1 Adversarial Attacks on ECG Data

[9, 18] have shown that deep-learning based ECG classifiers can be fooled by imperceptible perturbation added to benign ECG signals. They typically make use of the gradient-based white-box attack and further apply the Gaussian low-pass filters to smooth the injected noise for better imperceptibility by humans. [9] proposes to use the local variance to generate the imperceivable perturbations for ECG signals. Given $\boldsymbol{\delta} = \mathbf{x}^{adv} - \mathbf{x}^{org}$ and $Var(\cdot)$ referring to variance calculation, the similarity metric called smoothness is defined as follows:

$$d_i = \delta_i - \delta_{i-1},\ i = 2, ..., n \tag{4}$$

$$d_{smooth}(\boldsymbol{\delta}) \stackrel{\text{def}}{:=} Var(\mathbf{d}) \tag{5}$$

Smoothness metric d_{smooth} quantifies the variation of the difference between the benign signal and the adversarial signal by computing the variance of $\mathbf{d}$. We will also adopt such a metric in evaluating our attack performance. Additionally, [9] also considers the transformation process in their attack scheme and applies the Expectation Over Transformation (EOT) [5] to generate the adversarial inputs that are robust against potential transformation functions. However, these previous approaches are only viable within the white-box setting, where the machine learning model is fully exposed to the attacker. In this setting, the optimal perturbation direction can be pointed to by the gradient of the victim model as computed by back-propagation. But in real-world machine learning applications, such an assumption is unrealistic.

2.2 Black-box Hard-Label Adversarial Attack

In hard-label attacks, the details of the model are not revealed and the attacker can only query the model for the corresponding hard-label decision instead of the probability of all its outputs. [6, 10, 11, 19] attack DNN-based image classifiers in the extreme cases, where only the model classification with the highest confidence is given to the attackers. Among these, [6] works toward the decision boundary, traveling along it to produce adversarial images using imperceptible perturbations.

3 METHODOLOGY

To generate adversarial data sets, we choose a decision-based boundary attack [6], which performs random searches near the decision boundary. In addition, we also noticed that directly minimizing the L_2 distance in the image domain isn't optimal for attacking time-series data, which produces highly visible jaggedness. Thus, we adopted a smoothing strategy to generate adversarial perturbations with a better perceptual quality.

3.1 Boundary Attack: Black-Box Setting

In the hard-label black-box setting, an attacker cannot utilize the back-propagated gradient to generate adversarial perturbations for the original input data. Instead, the boundary attack initializes an ECG signal in the target class as the adversarial ECG signal which will be updated through subsequent iterations. The random searches ensure that successful samples of $\mathbf{x}^{adv}$ stay within the adversarial region and also reduce the L_2 distance from $\mathbf{x}^{org}$.

Algorithm 1 Boundary Attack

1: **Input:** original ECG $\mathbf{x}^{org}$, ECG $\mathbf{x}^{target}$ in the target class t, hard-label black-box classifier $f(\mathbf{x}) : \mathbb{R}^N \rightarrow \{1, 2, ..., C\}$
2: **Output:** adversarial ECG $\mathbf{x}^{adv}$ s.t. $\|(\mathbf{x}^{org} - \mathbf{x}^{adv})\|_2$ is minimized.
3: Initial step size γ and β. Let $\mathbf{x}^1 = \mathbf{x}^{target}$
4: **for** $i = 1 : N_0$ **do**
5: Generate random noise $\boldsymbol{\eta} \in \mathbb{R}^N$ and project it such that $\langle \boldsymbol{\eta}, \mathbf{x}^{org} - \mathbf{x}^i \rangle = 0$:

$$\mathbf{e} \stackrel{\text{def}}{:=} \frac{\mathbf{x}^{org} - \mathbf{x}^i}{\|\mathbf{x}^{org} - \mathbf{x}^i\|_2} \tag{6}$$

$$\boldsymbol{\eta} \leftarrow \boldsymbol{\eta} - \langle \boldsymbol{\eta}, \mathbf{e} \rangle \mathbf{e} \tag{7}$$

6: i) Perform orthogonal step:

$$\mathbf{x}_o^{i+1} = \mathbf{x}^{org} + \frac{1}{\sqrt{1+\gamma^2}} (\gamma \frac{\|(\mathbf{x}^{org} - \mathbf{x}^i)\|_2}{\|\boldsymbol{\eta}\|_2} \boldsymbol{\eta} - (\mathbf{x}^{org} - \mathbf{x}^i)) \tag{8}$$

7: ii) Perform step towards original ECG data:

$$\mathbf{x}^{i+1} = \mathbf{x}_o^{i+1} + \beta(\mathbf{x}^{org} - \mathbf{x}_o^{i+1}) \tag{9}$$

8: **if** $\mathbf{x}^{i+1}$ is not adversarial **then**
9: $\mathbf{x}^{i+1} = \mathbf{x}^i$
10: Increase γ and β if the attack success rate is too high. Otherwise, decrease them.
11: **return** $\mathbf{x}^{i+1}$

In our initial iterations, the adversarial EGG signal seeks the decision boundary separating adversarial and non-adversarial inputs. For future iterations near the boundary, the attack samples noise $\boldsymbol{\eta}$ and projects $\mathbf{x}^i + \boldsymbol{\eta}$ onto the sphere centered at $\mathbf{x}^{org}$ with radius $D(\mathbf{x}^{org}, \mathbf{x}^i)$ (Eqn. 8). After the initial orthogonal step, it steps towards $\mathbf{x}^{org}$ with size $\beta(\mathbf{x}^{org} - \mathbf{x}^i)$, (Eqn. 9). The iteration is complete when an introduction of noise decreases the L_2 distance and the signal remains adversarial, or it has exhausted its maximum number of attempts. We refer our readers to [6] for a detailed explanation.

3.2 Improving the Smoothness

The algorithm 1 mentioned above is designed to generate imperceptible perturbations in the image domain, but is not as effective in doing so for time-series data. Thus, we apply a low-pass Hanning filter onto our noise, removing a great deal of jaggedness from the ECG signals created. This results in a smoother, far more realistic signal as demonstrated in Figure 2.

$$\boldsymbol{\eta} \leftarrow \boldsymbol{\eta} \circledast \boldsymbol{h} - \langle \boldsymbol{\eta} \circledast \boldsymbol{h}, \mathbf{e} \rangle \mathbf{e} \tag{10}$$

$\boldsymbol{h} \in \mathbb{R}^t$ is a discrete Hanning filter with window size equal to t. By replacing Eqn. (7) with Eqn. (10) in Algorithm 1, we are able to

generate adversarial ECG signals without significantly disruptive noise.

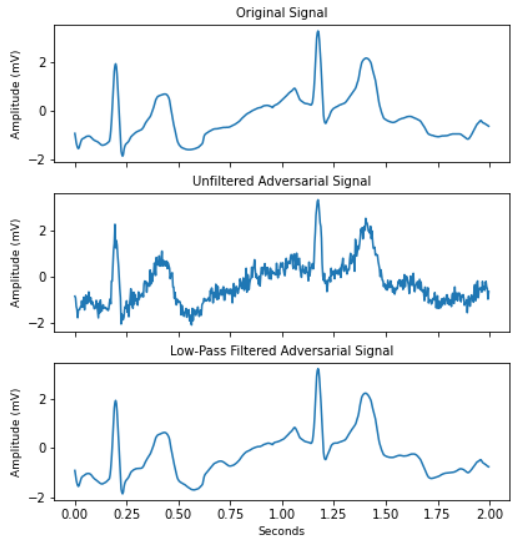

Figure 2: Effects of Low-Pass Hanning Filter

4 IMPLEMENTATION

4.1 PhysioNet Computing in Cardiology Challenge 2017 Database

In this database [12], there are over 8500 single short ECG lead recordings. The lengths range from 9 seconds to 60 seconds, with a median length of 30 seconds. The recordings were sampled at a rate of 300 Hz. There were four different classifications of rhythms, normal sinus(N), atrial fibrillation(AF), alternative (O), and too noisy to be classified (~). An alternative rhythm is any other kind of arrhythmia which isn't AF. As noted in [9], the relevant cases occur when adversarially attacking a normal sinus rhythm to be classified as arrhythmia (AF & alternative rhythms), and vice versa.

4.2 Deep Neural Network Arrhythmia Classifier

For this experiment, we use a trained model which was submitted specifically for the PhysioNet Computing in Cardiology Challenge 2017 by [4]. The model performed with an accuracy of 79%, on par with current state-of-the-art ECG signal classifying models. It was based on a previous model from [26], which used a 34-layer Residual Network to classify single-lead ECG signals into 14 different classes. Unlike its predecessor, the victim model for our experiment only has four different classifications: AF, N, O, ~. This specific model [4] is open-source, a main factor in our decision to attack it.

4.3 Processing & Attacking the ECG Data

We convert all 8,528 .mat files from the *PhysioNet Computing in Cardiology Challenge 2017* database into NumPy arrays and separate them based on their respective classifications. We then process the data exactly as the attack in the white-box setting did [9], by taking the first 9000 values and normalizing the array afterwards. If there are less than 9000 values, zeros are padded. Similar to [9], we only attack original and target ECG signals which are correctly classified by the model, avoiding the possibility of initial misclassifications. Finally, for each of the 12 combinations of two distinct labels, one of which being designated as the original class and the other as the target class, we randomly create 10 pairs to be inputted into both our boundary attack algorithm and the white-box attack.[4]

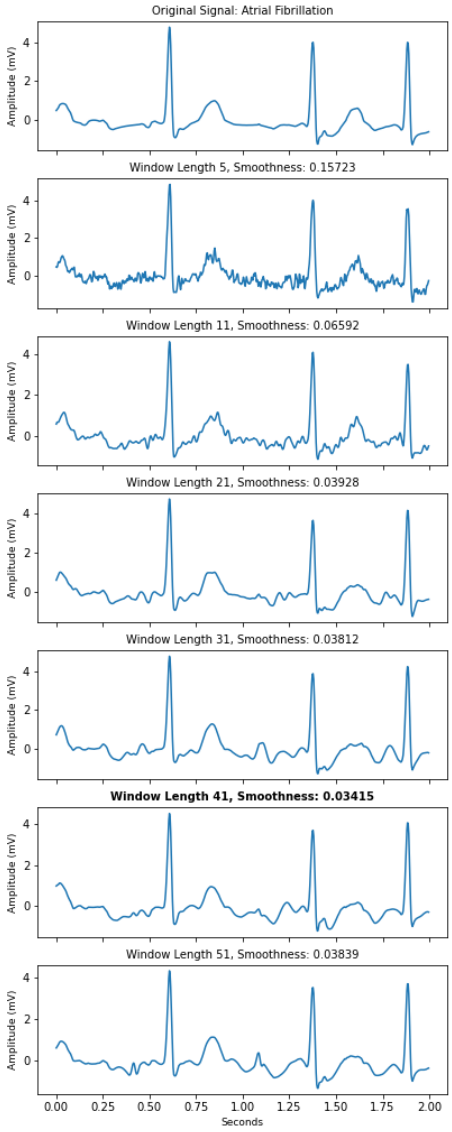

Figure 3: Comparing Various Window Lengths

4.4 Determining the Window Length

We initially attacked all 120 pairs with a window length of 5, which proved ineffective. Using the 10 worst performing pairs based upon

[4]Our code is available at https://github.com/nesl/Black-box-ECG-attack

the smoothness metric (5), we increment the size of the window and repeat this. We sought to not oversimplify the signal, while maintaining the major characteristics such as the spikes and smaller bumps before and after the beat. The adversarial signals in Figure 3 are intended to classify as normal sinus rhythms, while appearing as the original signal's example of an AF rhythm. As the length of the window increases, the adversarial signal appears visually cleaner due to smoother perturbations. The average smoothness decreases until the window length was 41, then increases for greater lengths, indicating that we have reached a relative minimum and our desired value for the Hanning filter.

5 RESULTS

Metrics Based Evaluation: Based on the experiment (1), the performance of our adversarial attack under black-box conditions better minimizes the quantitative metrics when generating adversarial ECG signals. In Table 1, we compare the average L_2 and smoothness(5), both metrics which quantify differences between generated adversarial ECG signals and the original ECG signals. Also, we report our average number of queries, which can be used to measure the effectiveness of our black-box boundary attack when compared against future black-box style attacks. As seen in Table 1, given sufficient queries, the adversarial attack in the black-box setting actually outperforms the white-box attack, [9] as it reduces the L_2 distance by 61.9% and improves the smoothness metric by 16.2%.

Table 1: Quantitative Comparisons w/ Various Metrics

	Average L_2	Average Smoothness	Average Queries
ECG-adv[9]	28.4897	0.02035	/
Boundary Attack (ours)	**10.8635**	**0.01705**	**31564**

Perception Based Evaluation: We had two physicians, both of whom achieved a 100% accuracy rate in classifying our original ECG signals, qualitatively determine the classifications of our adversarial signals. From these classifications, we determined whether our attacks succeeded or failed. A successful attack is defined as follows: a physician or cardiologist classifies the adversarial and original ECG signals as the same, but when the adversarial signal is inputted into the model, it remains in the target signal's class. Table 2 consists of a confusion matrix, where the original ECG signal classification is labeled on the vertical axis, while the target classification is labeled on the horizontal axis. The diagonal of empty cells for each confusion matrix is because there were no original and target pairs created with identical classes, as an adversarial attack wouldn't be relevant. As seen in Table 2, our black-box attack succeeded 98.3% of the time, while the ECG-adv[9] success rate was 100%, as determined by the two physicians.

6 DISCUSSION & FUTURE WORK

Discussion: The process of adversarially attacking an ECG classifier, versus say an image classifier, varies significantly due to how the trained human eye determines their effectiveness. For classifying images, the human eye typically looks for consistent similarities across its entirety. This favors a metric such as the minimization of the $\mathbf{L}_2$ distance, which helps create an adversarial image with less discrepancies across the board. However, with something like ECG

Table 2: Success Rate Confusion Matrices

	Boundary Attack (ours)				ECG-adv[9]			
	Target Classification							
	AF	N	O	~	AF	N	O	~
AF	/	100%	100%	100%	/	100%	100%	100%
N	100%	/	100%	100%	100%	/	100%	100%
O	100%	100%	/	80%	100%	100%	/	100%
~	100%	100%	100%	/	100%	100%	100%	/
Overall	**98.3%**				**100%**			

data, there are certain identifiable features that a human can pick out when classifying them, without needing to look too deeply at every individual discrepancy. DNN-based ECG classifiers accounting too much for the change between each data point can be compromised, as shown by the effectiveness of our adversarial attacks, which render key characteristics identifiable to the trained human eye while still having enough room to create discrepancies via noise to throw off the model. Here, imperceptibility should still occur as long as the data sets are categorized in the same class, regardless if they look mildly different as a whole. While this is something the experienced human eye can adapt to, the DNN often becomes distracted and thereby fooled. The following in-depth examples demonstrate this methodology enabling a successful attack.

For context, the key features in every ECG signal to look for and analyze are the rate in beats per minute (bpm), the rhythm/regularity, the P waves, the PR Interval (PRI), the QRS Complex, the QT interval, the ST segment, and the T waves, as shown in Figure 4.

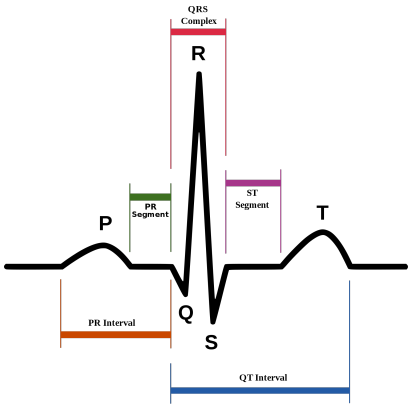

Figure 4: Features of a Normal ECG Signal

For the four key pairings of original and target ECG signals ($N \rightarrow AF$, $AF \rightarrow N$, $O \rightarrow N$, & $N \rightarrow O$), we retain certain defining features to ensure the trained human eye maintains an identical classification, while incorporating others to an extent in order to fool the ECG classifying model.

For instance, in Figure 5, one can see that in a normal sinus rhythm there is a regular rhythm rate of 60-100 bpm, each QRS complex is preceded by a normal P wave, the normal P wave is in the correct orientation, the PR interval remains constant, and the QRS complexes are less than 100 ms wide [7]. To the trained human eye this signal is normal due to the presence of the features mentioned above. But with the addition of noise creating a wavier baseline,

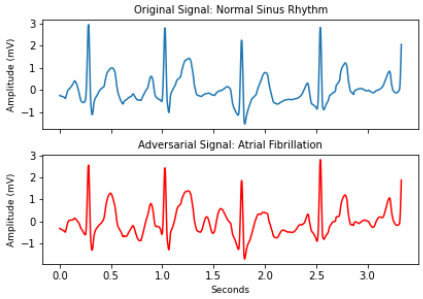

Figure 5: N incorrectly classified as AF after the attack (N → AF)

the model was fooled into classifying this as atrial fibrillation.

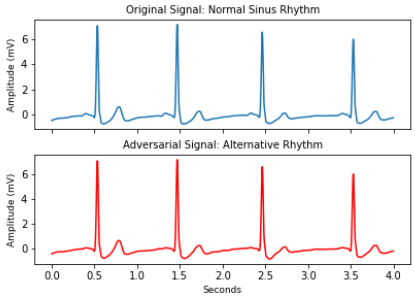

Figure 6: N incorrectly classified as O after the attack (N → O)

In Figure 6, the standard features of a normal sinus rhythm remain, but our attack has rendered the P wave far more difficult to identify, which caused the model to misclassify this as an alternative rhythm. However, since the trained human eye can still notice the smaller P wave occurring at regular intervals while appearing similar in size & shape, it was still classified as normal.

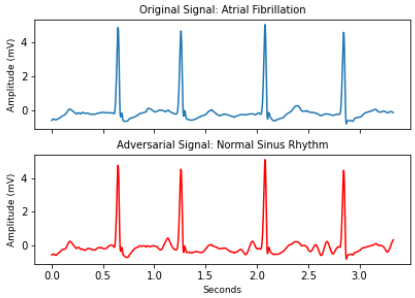

Figure 7: AF incorrectly classified as N after the attack (AF → N)

In Figure 7, there are clear signs of atrial fibrillation [21], including the absence of P waves; an erratic, wavy baseline; and an irregular rate of QRS complexes. However, while our adversarial image maintains these features, it also adds some waviness to the baseline that is typical in AF, but leads the model to mistake them for P waves in a normal sinus rhythm.

Finally, in Figure 8, there is a noticeably abnormal P wave in the 2nd beat, followed by a normal QRS complex. This strongly indicates Premature Atrial Complex (PAC) [22] which falls under the

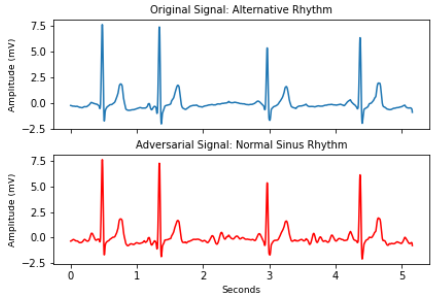

Figure 8: O incorrectly classified as N after the attack (O → N)

category of an alternative rhythm. Our attack alters the abnormal P wave to be considered similar to the others, leading the model to classified it as sinus arrhythmia. This is a type of normal sinus rhythm, except with variations in the time between successive P waves and an irregular rate of QRS complexes. The trained human eye can still pick out the abnormal P wave and recognize this feature that identifies PAC.

Overall, our adversarial decision-based boundary attack performed very comparably with the attack in the white box-setting, demonstrating the legitimacy of security threats to ECG diagnostic systems.

Future Work: The primary direction will be researching on similarity metrics that correlate with the trained human perception in order to better quantify the assessment of adversarial ECG examples. This is per the fact that qualitative classifications like ECG data are highly reliant on particular patterns and features the curved lines present [13, 15], a property that conventional metrics such as L_p distance cannot entirely capture. A secondary direction is the development of methods that can detect these adversarial ECG inputs, as the means of defending against the application of low-pass filters that improve the attack's perceptual smoothness and similarity [20, 30] is an area still largely unexplored. Finally, as mentioned in the discussion, certain features in the adversarial inputs are likely the main factors in compromising the security of DNNs when classifying ECG signals. We aim to explore methods to advance the existing explainability approaches [27, 31] and uncertainty quantifications [29] of DNNs to help combat these adversarial attacks on ECG data sets.

ACKNOWLEDGMENTS

The research presented in this paper is supported in part by the U.S. Army Research Laboratory and the U.K. Ministry of Defence under Agreement Number W911NF-16-3-0001 and the National Science Foundation (NSF) under awards # CNS-1705135 and CNS-1822935, and by the National Institutes of Health (NIH) award # P41EB028242 for the mDOT Center. The views and conclusions contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the NSF, the NIH, the U.S. Army Research Laboratory, the U.S. Government, the U.K. Ministry of Defence or the U.K. Government. The U.S. and U.K. Governments are authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

REFERENCES

[1] 2008. Tricog. https://www.tricog.com/. [Online; accessed 10-September-2020].

[2] World Health Organization 2018. [n.d.]. *World Health Organization. 2018. Cardiovascular disease is the leading global killer.* https://www.who.int/health-topics/cardiovascular-diseases/#tab=tab_1

[3] Moustafa Alzantot, Bharathan Balaji, and Mani Srivastava. 2018. Did you hear that? adversarial examples against automatic speech recognition. *arXiv preprint arXiv:1801.00554* (2018).

[4] Fernando Andreotti, Oliver Carr, Marco AF Pimentel, Adam Mahdi, and Maarten De Vos. 2017. Comparing feature-based classifiers and convolutional neural networks to detect arrhythmia from short segments of ECG. In *2017 Computing in Cardiology (CinC)*. IEEE, 1–4.

[5] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. 2018. Synthesizing robust adversarial examples. In *International conference on machine learning*. PMLR, 284–293.

[6] Wieland Brendel, Jonas Rauber, and Matthias Bethge. 2017. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248* (2017).

[7] Dr. Ed Burns. [n.d.]. *Normal Sinus Rhythm (NSR) Overview.* https://litfl.com/normal-sinus-rhythm-ecg-library/

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

[9] Huangxun Chen, Chenyu Huang, Qian Zhang, and Wei Wang. [n.d.]. ECGadv: Generating Adversarial Electrocardiogram to Misguide Arrhythmia Classification System.

[10] Minhao Cheng, Thong Le, Pin-Yu Chen, Jinfeng Yi, Huan Zhang, and Cho-Jui Hsieh. 2018. Query-efficient hard-label black-box attack: An optimization-based approach. *arXiv preprint arXiv:1807.04457* (2018).

[11] Minhao Cheng, Simranjit Singh, Pin-Yu Chen, Sijia Liu, and Cho-Jui Hsieh. 2019. Sign-OPT: A Query-Efficient Hard-label Adversarial Attack. *arXiv preprint arXiv:1909.10773* (2019).

[12] Gari D Clifford, Chengyu Liu, Benjamin Moody, H Lehman Li-wei, Ikaro Silva, Qiao Li, AE Johnson, and Roger G Mark. 2017. AF Classification from a short single lead ECG recording: the PhysioNet/Computing in Cardiology Challenge 2017. In *2017 Computing in Cardiology (CinC)*. IEEE, 1–4.

[13] Philipp Eichmann and Emanuel Zgraggen. 2015. Evaluating subjective accuracy in time series pattern-matching using human-annotated rankings. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*. 28–37.

[14] Samuel G Finlayson, Hyung Won Chung, Isaac S Kohane, and Andrew L Beam. 2018. Adversarial attacks against medical deep learning systems. *arXiv preprint arXiv:1804.05296* (2018).

[15] Anna Gogolou, Theophanis Tsandilas, Themis Palpanas, and Anastasia Bezerianos. 2018. Comparing similarity perception in time series visualizations. *IEEE transactions on visualization and computer graphics* 25, 1 (2018), 523–533.

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

[17] Xintian Han, Yuxuan Hu, Luca Foschini, Larry Chinitz, Lior Jankelson, and Rajesh Ranganath. 2019. Adversarial Examples for Electrocardiograms. *arXiv preprint arXiv:1905.05163* (2019).

[18] Xintian Han, Yuxuan Hu, Luca Foschini, Larry Chinitz, Lior Jankelson, and Rajesh Ranganath. 2020. Deep learning models for electrocardiograms are susceptible to adversarial attack. *Nature Medicine* (2020), 1–4.

[19] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. 2018. Black-box adversarial attacks with limited queries and information. *arXiv preprint arXiv:1804.08598* (2018).

[20] Muzammal Naseer, Salman Khan, and Fatih Porikli. 2019. Local gradients smoothing: Defense against localized adversarial attacks. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 1300–1307.

[21] Dr Chris Nickson. [n.d.]. *Atrial Fibrillation.* https://litfl.com/atrial-fibrillation/

[22] Dr Chris Nickson. [n.d.]. *Premature Atrial Complex (PAC).* https://litfl.com/premature-atrial-complex-pac/

[23] Bahareh Pourbabaee, Mehrsan Javan Roshtkhari, and Khashayar Khorasani. 2018. Deep convolutional neural networks and learning ECG features for screening paroxysmal atrial fibrillation patients. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 48, 12 (2018), 2095–2104.

[24] PYA. [n.d.]. *PYA COMPENSATION STUDY: SPOTLIGHT ON CARDIOLOGY.* http://www.pyapc.com/wp-content/uploads/PYA-Compensation-Study-Spotlight-on-Cardiology.pdf

[25] Yao Qin, Nicholas Carlini, Garrison Cottrell, Ian Goodfellow, and Colin Raffel. 2019. Imperceptible, robust, and targeted adversarial examples for automatic speech recognition. In *International Conference on Machine Learning*. PMLR, 5231–5240.

[26] Pranav Rajpurkar, Awni Y Hannun, Masoumeh Haghpanahi, Codie Bourn, and Andrew Y Ng. 2017. Cardiologist-level arrhythmia detection with convolutional neural networks. *arXiv preprint arXiv:1707.01836* (2017).

[27] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. 1135–1144.

[28] Giovanna Sannino and Giuseppe De Pietro. 2018. A deep learning approach for ECG-based heartbeat classification for arrhythmia detection. *Future Generation Computer Systems* 86 (2018), 446–455.

[29] Murat Sensoy, Lance Kaplan, and Melih Kandemir. 2018. Evidential deep learning to quantify classification uncertainty. In *Advances in Neural Information Processing Systems*. 3179–3189.

[30] Yash Sharma, Gavin Weiguang Ding, and Marcus Brubaker. 2019. On the effectiveness of low frequency perturbations. *arXiv preprint arXiv:1903.00073* (2019).

[31] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. 2013. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034* (2013).

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