Accurate and adversarially robust classification of medical images and ECG time-series with gradient-free trained sign activation neural networks

1st Zhibo Yang
Department of Computer Science
New Jersey Institute of Technology
Newark, USA
zy328@njit.edu

2nd Yanan Yang
Department of Computer Science
New Jersey Institute of Technology
Newark, USA
mx42@njit.edu

3th Yunzhe Xue
Department of Computer Science
New Jersey Institute of Technology
Newark, USA
yx277@njit.edu

4th Frank Y. Shih
Department of Computer Science
New Jersey Institute of Technology
Newark, USA
frank.y.shih@njit.edu

5th Justin Ady
Division of Vascular and Endovascular Surgery
Robert Wood Johnson University Hospital
Newark, USA
Jwa60@rwjms.rutgers.edu

6th Usman Roshan
Department of Computer Science
New Jersey Institute of Technology
Newark, USA
usman@njit.edu

***Abstract*—Adversarial attacks in medical AI imaging systems can lead to misdiagnosis and insurance fraud as recently highlighted by Finlayson et. al. in *Science* 2019. They can also be carried out on widely used ECG time-series data as shown in Han et. al. in *Nature Medicine* 2020. At the heart of adversarial attacks are imperceptible distortions that are visually and statistically undetectable but cause the machine learning model to misclassify data. Recent empirical studies have shown that a gradient-free trained sign activation neural network ensemble model requires a larger distortion than state of the art models. We apply them on medical data in this study as a potential solution to detect and deter adversarial attacks. We show on chest X-ray and histopathology images, and on two ECG datasets that this model requires a greater distortion to be fooled than full-precision, binary, and convolutional neural networks, and random forests. We show that adversaries targeting the gradient-free sign networks are visually distinguishable from the original data and thus likely to be detected by human inspection. Since the sign network distortions are higher we expect an automated method could be developed to detect and deter attacks in advance. Our work here is a significant step towards safe and secure medical machine learning.**

***Index Terms*—histopathology, X-ray, ECG, adversarial attack, robust classification, gradient-free trained sign activation neural networks**

I. Introduction

While machine learning holds great promise for accurate and automated medical diagnosis it can be fooled with imperceptible distortions known as adversarial attacks [1]–[6]. As a result attackers can trick models into misdiagnosis which can lead to insurance fraud, or fool models into producing incorrect results in large-scale clinical studies and tilt the conclusion in their favor. One way to defend against such attacks is to use models that require a high distortion to make the input adversarial. In this scenario it is more likely to detect and stop such an attack in advance.

Gradient-free trained sign activation networks have recently shown great promise in defending against adversarial attacks [7]–[10]. These networks are trained with a stochastic coordinate descent algorithm [8]. Their minimum distortion to make the input adversarial (also called the adversarial distortion) has been empirically estimated to be higher than other state of the art models [9]. This means an image has to undergo considerable modification (that may be detectable in advance) before it can fool a model.

In this paper we study the adversarial distortion of gradient-free trained sign networks and other state of the art models when attacking chest X-ray and histopathology image, and ECG time-series data. We find that the gradient free trained sign networks have a higher estimated distortion than all other models on both image and ECG time-series data. We also show that adversaries targeting the gradient free sign networks are visually distinguishable from the original data thus making them detectable in advance. Our work is a first step towards secure and robust medical machine learning systems.

II. Methods

A. Gradient-free trained sign activation neural networks

Sign activation networks can be trained with a recently proposed gradient-free stochastic coordinate descent algorithm [8]–[10]. In order to understand this training algorithm for a single hidden layer network we first show it for a simple linear classifier. Suppose we are given binary class data $x_i \in R^d$ and $y_i \in \{-1, +1\}$ for $i = 0...n-1$. We wish to determine a linear classifier $w \in R^d, w_0 \in R$ that minimizes the empirical risk for a given loss function $\sum_i L(w, w_0, x_i, y_i)$. At the high level our approach is simple: start with a random solution $w_i \in N(0,1), w_0 \in N(0,1)$ for $i = 0...d-1$ and iteratively make incremental changes that improve the risk. In

each iteration we select a random set of features (coordinates) from w called F. For each feature $w_i \in F$ we add/subtract a learning rate η and then determine the w_0 that optimizes the risk. We consider all possible values of $w_0 = \frac{w^T x_i + w^T x_{i+1}}{2}$ for $i = 0...n-2$ and select the one that minimizes the loss. To avoid local minima in our search we consider a random sample of the training data in each iteration. We set this to 75% of the training data in image and text data experiments and 25% in the ECG data.

We call the above search *stochastic coordinate descent* abbreviated by SCD. In order to train a single hidden layer network we apply SCD to the final node and then a randomly selected hidden node in each iteration of the algorithm. We can train sign activation networks with and without binary weights using our SCD training procedure above. In the case of binary weights we don't need a learning rate. We apply parallelism and several heuristics in practice to speed up the real runtimes.

B. Data

We obtained a multiclass dataset of histopathology images of colorectal cancer from https://zenodo.org/record/53169#.X5TZOJNJFE7 [11] and a binary dataset of chest X-ray images (pneumonia vs normal) from Kaggle https://www.kaggle.com/paultimothymooney/chest-xray-pneumonia. The chest X-ray images were resized to 96×96 and downsampled to have 1,584 images per class. The colorectal images are tissue tiles of dimensions 150×150 that we also resize to 96×96. The colorectal cancer dataset has 8 classes each containing 625 images from which we extract two classes which give the highest test accuracy (these are the Mucosa and Adipose tissue images from class 6 and 7 respectively). We then randomly divide each of the two datasets into an 80:20 train test split.

We also obtained two ECG time-series datasets from Kaggle https://www.kaggle.com/shayanfazeli/heartbeat. The first one is the PTB Diagnostic ECG dataset [12], [13] and the second is the MIT-BIH Arrhythmia dataset [14]. The PTB Diagnostic (PTBD) dataset is two classes of normal and abnormal heartbeat readings. The MIT dataset contains five classes of different heartbeat rhythms with a total of 87554 points in the train and 21892 in the test. Both datasets have 187 heartbeat readings (features) per datapoint. We divide the PTBDB dataset randomly into a single 80:20 train test split (yielding 13096 train and 1456 test points) and the MIT-BIH dataset comes with separate train and test datasets.

C. Methods compared

We train two types of sign activation networks with our algorithm: (1) SCD01: 01-loss in the final node, and (2) SCDCE: cross-entropy loss in the final node. Since sign activation is non-convex and our training starts from a different random initialization we run it a 100 times and output the majority vote.

We include in our methods the full-precision sigmoid activation counterpart denoted as MLP, two convolutional neural networks LeNet [15] and ResNet18 [16], binary neural networks [17] (implemented in the Larq library [18]), and random forest [19]. For each model we train it a 100 times starting from different random initialization except for ResNet18 which we train 10 times. We then output the majority vote of each model.

For the ECG time series data we use an ensemble of 10 convolutional neural networks (CNN) with 1D convolutional kernels. Each of our CNNs has the following structure: 64 1x16 Conv1D kernels $\rightarrow$ MaxPool 1x4 $\rightarrow$ 128 1x16 Conv1D kernels $\rightarrow$ MaxPool 1x4 $\rightarrow$ 256 1x16 Conv1D kernels $\rightarrow$ MaxPool 1x2 $\rightarrow$ FullyConnected $\rightarrow$ Output.

The code for sign activation networks is freely available from from https://github.com/zero-one-loss/scd_github. We implemented all models in Python, numpy, and Pytorch [20].

III. Results

A. Chest X-ray and colorectal cancer histopathology images

We first evaluate the clean test accuracy of all our models. In Table I we show the accuracies of all models on the validation data. We see that the convolutional networks ResNet18 and LeNet lead in the chest X-ray dataset but the other methods are not too far behind.

TABLE I
AVERAGE ACCURACY OF VALIDATION DATA IN CHEST X-RAY AND HISTOPATHOLOGY IMAGES

	Chest X-ray	Histopathology
SCD01	90.7%	99.6%
SCDCE	91.3%	99.6%
MLP	88.7%	100%
LeNet	92.6%	99.6%
ResNet18	94.3%	99.6%
Random forest	84.7%	100%

We then run the HopSkipJump boundary based black box attack [21] to determine the adversarial distortion of five randomly selected images from the chest X-ray and five from the colorectal cancer histopathology validation datasets. This is an estimate of the minimum distortion required to make an image adversarial: the larger the value the more robust the model is since a large distortion is also likely to be detected in advance. Finding the exact minimum distortion is in fact an NP-hard problem as shown for ReLu activated neural networks [22], [23] and tree ensemble classifiers [24]. Even approximating the minimum distortion for ReLu activated neural networks is NP-hard [25]. In previous work the distortions reported by HopSkipJump have been shown to be lower (tighter and more accurate) than other boundary attack methods [9], [21].

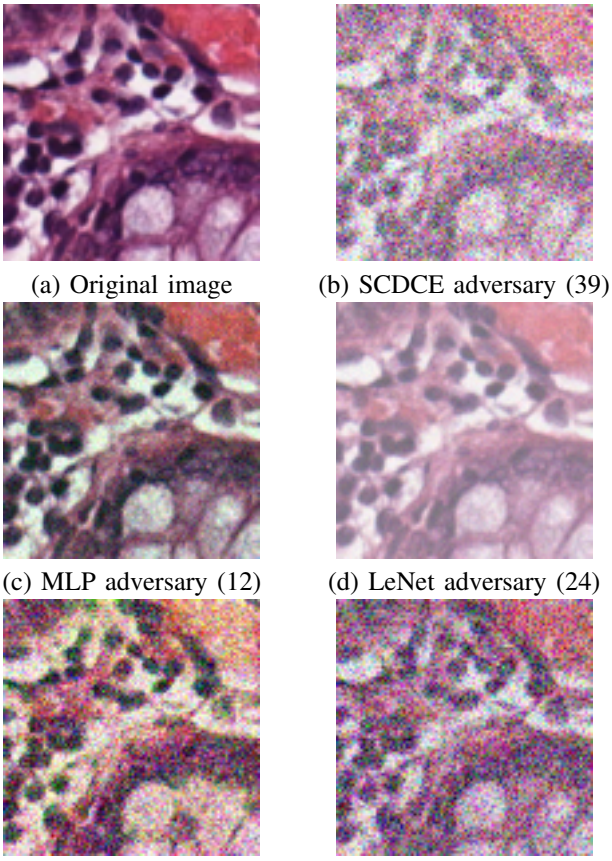
We use the HopSkipJump implementation in the IBM Adversarial Robustness Toolkit (ART) [26]. In order to obtain as accurate an estimate of the adversarial distortion as possible we run HopSkipJump 10 times each with the same fixed initial image (that is misclassified by all models) and maximum iterations of 100 and report the minimum value. We use a fixed image because otherwise random initial points are not misclassified by some of our models. For a single datapoint this typically takes several hours to finish and thus we are able to report the distortion of only five random images per dataset.

In Table II we see the adversarial distortions of five random test images each from the chest X-ray and histopathology datasets and their averages. The gradient free trained sign network SCDCE has the highest distortion on each of the two datasets. In the chest X-ray dataset we see that MLP is the second best after SCDCE and in the colorectal dataset ResNet18 follows the SCD models. When we average across the two datasets the distortions of the SCD models are even higher with SCDCE taking the lead and twice better than MLP, LeNet, and ResNet18.

TABLE II
MINIMUM ESTIMATED L_2 ADVERSARIAL DISTORTION OF 5 RANDOM IMAGES EACH FROM CHEST X-RAY AND COLORECTAL CANCER VALIDATION DATASETS AS GIVEN BY HOPSKIPJUMP WHEN ATTACKING THE DIFFERENT MODELS.

	Chest X-ray					
	SCD01	SCDCE	MLP	LeNet	Res18	RF
Image 0	15.3	18.4	14.8	1	4.3	18.5
Image 1	12.1	16.4	15.1	0.5	2.9	11.3
Image 2	12.8	17.6	14.5	4.2	0.6	9.2
Image 3	10	7.7	10.7	0.3	0.1	12
Image 4	11.3	14	12.9	4.1	0.4	2.4
Average	12.3	14.8	13.6	3.2	0.5	10.7
	Colorectal histopathology					
	SCD01	SCDCE	MLP	LeNet	Res18	RF
Image 0	28.3	41	9.9	29	31.6	19.9
Image 1	4.4	6.3	2.8	7	6.2	3.9
Image 2	35.8	36.1	9.9	36.8	39.8	30.4
Image 3	30	38.6	12	24.1	19.1	28.7
Image 4	17.2	26.5	7.7	17.1	19	13.4
Average	24.1	29.7	8.5	22.8	23.1	19.2
Combined Average	17.7	22.3	11	13	11.8	15

To get a visual feel for the distortions we plot the original and adversarial images of Image 3 (shown above in Table II) from the colorectal dataset. In Figure 1 we see that all adversarial images have a high distortion with SCDCE having the highest. As a result the SCDCE adversary also looks the dottiest compared to others and can easily be spotted as abnormal and potentially adversarial.



(a) Original image (b) SCDCE adversary (39) (c) MLP adversary (12) (d) LeNet adversary (24) (e) ResNet18 adversary (19) (f) RF adversary (29)

Fig. 1. Original and adversarial images produced by HopSkipJump attack. Also shown is their minimum estimated L_2 distance to the original image (adversarial distortion).

B. ECG time series

As we did for the images above, we firsts compare the clean test accuracies of all methods in Table III. As above the CNN leads in accuracy but the gradient-free sign networks, MLP, and random forest are not far behind.

TABLE III
AVERAGE ACCURACY OF VALIDATION DATA IN PTBD AND MIT-BIH ECG DATASETS

	PTBD	MIH-BIH
SCD01	91.1%	96.4%
SCDCE	93.3%	96.6%
BNN	80.4%	86.3%
MLP	96.1%	97.1%
CNN	99.6%	99.9%
Random forest	97.6%	97.5%

We picked 37 random datapoints from the PTBD test dataset and 21 from the MIT-BIH test dataset and attacked all models on these points with HopSkipJump. We attack each point 10 times each time starting from an initial pool size of 1000 random points and 100 iterations, and report the minimum. As we saw above in the case of images, the gradient free sign network model has the highest distortion on both datasets individually and in the combined average. In the MIT-BIH dataset SCDCE leads in distortion by almost twice the value of the next best BNN.

TABLE IV
AVERAGE MINIMUM ESTIMATED L_2 ADVERSARIAL DISTORTION OF 37 AND 21 RANDOM TEST POINTS FROM PTBD AND MIT-BIH DATASETS RESPECTIVELY AS GIVEN BY HOPSKIPJUMP.

	PTBD	MIH-BIH	Combined average
SCD01	.18	.35	.27
SCDCE	.1	.39	.25
BNN	.14	.24	.19
MLP	.08	.15	.12
CNN	.1	.22	.16
Random forest	.14	.27	.21

In Figure 2 we see the ECG readings of a single random test datapoint. The larger distortion of SCD01 and SCDCE clearly shows more spikes than the original and can spotted as abnormal and adversarial. In comparison the other models

look similar to the original. We can see that RF's distortion comes mainly from one unusual spike in the beginning but the rest is smooth and similar to the original.

IV. Conclusion

We present for the first time a model that is robust to adversarial attacks in chest X-ray and histopathology images, and also in ECG time series data. We show that the gradient free trained sign networks require a greater distortion in order to fool the model and thus are likely to be detected in advance. While more research is required to show distortion on a larger cohort and to create methods that can detect high distortion attacks in advance, our work here is a significant step towards robust medical machine learning models.

References

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

[2] Huangxun Chen, Chenyu Huang, Qianyi Huang, Qian Zhang, and Wei Wang. Ecgadv: Generating adversarial electrocardiogram to misguide arrhythmia classification system. In *AAAI*, pages 3446–3453, 2020.

[3] Samuel G Finlayson, John D Bowers, Joichi Ito, Jonathan L Zittrain, Andrew L Beam, and Isaac S Kohane. Adversarial attacks on medical machine learning. *Science*, 363(6433):1287–1289, 2019.

[4] Suzanne C Wetstein, Cristina González-Gonzalo, Gerda Bortsova, Bart Liefers, Florian Dubost, Ioannis Katramados, Laurens Hogeweg, Bram van Ginneken, Josien PW Pluim, Marleen de Bruijne, et al. Adversarial attack vulnerability of medical image analysis systems: Unexplored factors. *arXiv preprint arXiv:2006.06356*, 2020.

[5] Adnan Qayyum, Junaid Qadir, Muhammad Bilal, and Ala Al-Fuqaha. Secure and robust machine learning for healthcare: A survey. *arXiv preprint arXiv:2001.08103*, 2020.

[6] Maoqiang Wu, Xinyue Zhang, Jiahao Ding, Hien Nguyen, Rong Yu, Miao Pan, and Stephen T. Wong. Evaluation of inference attack models for deep learning on medical data. *arXiv preprint arXiv:2011.00177*, 2020.

[7] Yunzhe Xue, Meiyan Xie, and Usman Roshan. Defending against substitute model black box adversarial attacks with the 01 loss. *arXiv preprint arXiv:2009.09803*, 2020.

[8] Yunzhe Xue, Meiyan Xie, and Usman Roshan. On the transferability of adversarial examples between convex and 01 loss models. In *IEEE International Conference on Machine Learning and Applications*, 2020.

[9] Yunzhe Xue, Meiyan Xie, and Usman Roshan. Towards adversarial robustness with 01 loss neural networks. In *IEEE International Conference on Machine Learning and Applications*, 2020.

[10] Meiyan Xie, Yunzhe Xue, and Usman Roshan. Stochastic coordinate descent for 0/1 loss and its sensitivity to adversarial attacks. In *Proceedings of 18th IEEE International Conference on Machine Learning and Applications - ICMLA 2019*, page to appear, 2019.

[11] Jakob Nikolas Kather, Cleo-Aron Weis, Francesco Bianconi, Susanne M Melchers, Lothar R Schad, Timo Gaiser, Alexander Marx, and Frank Gerrit Zöllner. Multi-class texture analysis in colorectal cancer histology. *Scientific reports*, 6:27988, 2016.

[12] R Bousseljot, D Kreiseler, and A Schnabel. Nutzung der ekg-signaldatenbank cardiodat der ptb über das internet. *Biomedizinische Technik/Biomedical Engineering*, 40(s1):317–318, 1995.

[13] Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220, 2000.

[14] George B Moody and Roger G Mark. The impact of the mit-bih arrhythmia database. *IEEE Engineering in Medicine and Biology Magazine*, 20(3):45–50, 2001.

[15] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[17] Zechun Liu, Baoyuan Wu, Wenhan Luo, Xin Yang, Wei Liu, and Kwang-Ting Cheng. Bi-real net: Enhancing the performance of 1-bit cnns with improved representational capability and advanced training algorithm. In *Proceedings of the European conference on computer vision (ECCV)*, pages 722–737, 2018.

[18] Lukas Geiger and Plumerai Team. Larq: An open-source library for training binarized neural networks. *Journal of Open Source Software*, 5(45):1746, 2020.

[19] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[20] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. dÁlché Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[21] Jianbo Chen, Michael I Jordan, and Martin J Wainwright. Hopskipjumpattack: A query-efficient decision-based attack. *arXiv preprint arXiv:1904.02144*, 3, 2019.

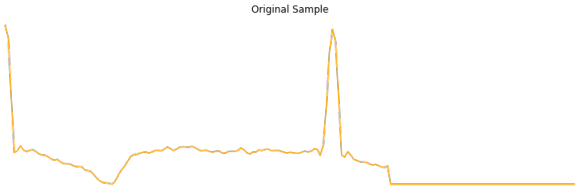
[22] Guy Katz, Clark Barrett, David L Dill, Kyle Julian, and Mykel J Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. In *International Conference on Computer Aided Verification*, pages 97–117. Springer, 2017.

[23] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifiable distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2, 2017.

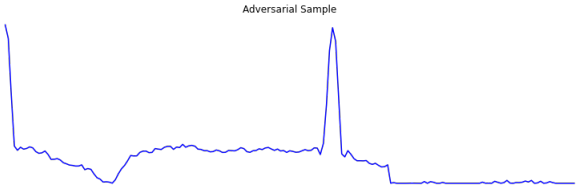
[24] Alex Kantchelian, J Doug Tygar, and Anthony Joseph. Evasion and hardening of tree ensemble classifiers. In *International Conference on Machine Learning*, pages 2387–2396, 2016.

[25] Tsui-Wei Weng, Huan Zhang, Hongge Chen, Zhao Song, Cho-Jui Hsieh, Duane Boning, Inderjit S Dhillon, and Luca Daniel. Towards fast computation of certified robustness for relu networks. *arXiv preprint arXiv:1804.09699*, 2018.

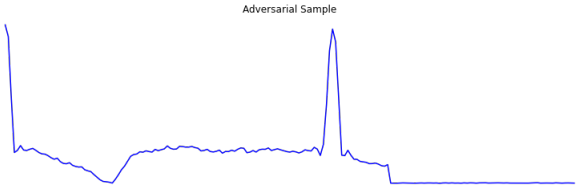
[26] Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, et al. Adversarial robustness toolbox v1. 0.0. *arXiv preprint arXiv:1807.01069*, 2018.



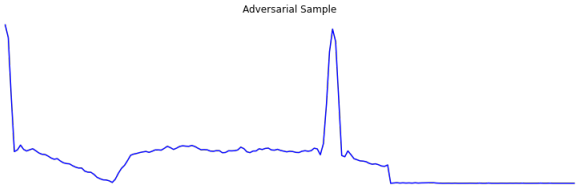
(a) Original image



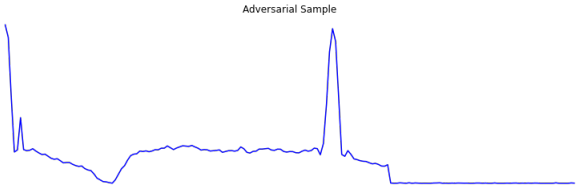
(b) MLP adversary (.08)



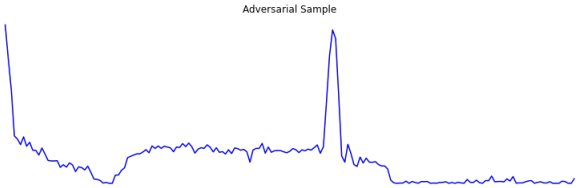
(b) BNN adversary (.05)



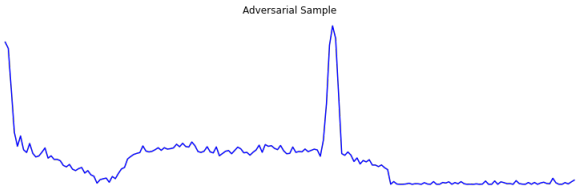
(c) CNN adversary (.04)



(d) RF adversary (.17)



(e) SCD01 adversary (.31)



(f) SCDCE adversary (.28)

Fig. 2. Original and adversarial images produced by HopSkipJump attack. Also shown is their minimum estimate L_2 distance to the original image (adversarial distortion).