Projected Gradient Descent Adversarial Attack and Its Defense on a Fault Diagnosis System

Mustafa Sinasi Ayas*, Selen Ayas†, and Seddik M. Djouadi‡
*Department of Electrical and Electronics Engineering, Karadeniz Technical University, Trabzon, Turkey
†Department of Computer Engineering, Karadeniz Technical University, Trabzon, Turkey
‡Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN 37996, USA
msayas@ktu.edu.tr, selenguven@ktu.edu.tr, djouadi@eecs.utk.edu

2022 45th International Conference on Telecommunications and Signal Processing (TSP) | 978-1-6654-6948-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/TSP55681.2022.9851334

***Abstract*—Knowledge-based fault diagnosis methods have become more preferred as they do not need precise model and signal patterns required in model-based and signal-based diagnosis methods, respectively. Machine learning (ML) techniques provide notable results on fault diagnosis by mapping information from raw signals to health condition. However, their vulnerabilities against malicious attacks arises as in other industrial application employing ML methods. In this paper, first, a common white-box adversarial attack called *projected gradient descent* (PGD) adversarial attack is injected into a deep residual learning (DRL) network model, which decides health condition of a rolling bearing. Then, robustness of the DRL model is analyzed to examine the effect of the implemented adversarial machine learning (AML). After that, *adversarial training* technique is used to improve the robustness of the DRL model. The experimental results show that it is possible to implement AML with existing methods to force model to misclassification. Even for a quite perturbation, the average classification accuracy of the DRL model is decreased from 99.98% to 61.25%. The results also indicate that the *adversarial training* technique increases the robustness of the model.**

***Keywords*—Fault diagnosis system; adversarial machine learning; projected gradient descent; adversarial training**

I. Introduction

Achieving a desirable performance in an industrial process in addition to the reliability of the process is just as important from both safety and maintainability viewpoints. Therefore, fault diagnosis is an important research field of process engineering attracting remarkable interest from both academia and industry. Fault diagnosis of rolling bearings, major part of rotating machinery, is one of the attractive research area. Early detection of any fault occurred in a rolling bearing may prevent whole process to break down [1].

Knowledge-based fault diagnosis methods have become more preferred as they do not need precise model and signal patterns required in model-based and signal-based methods, respectively. In knowledge-based methods, massive data collected from long-term continuous monitoring are used to conduct a relation between raw measured data and status of the process [2]. Industry 4.0 steps up the data collection procedure and leads researchers to use the potential of massive data in fault diagnosis [3]. From rolling bearing fault point of view, vibration signals having nonlinear and uneven stability characteristics constitute this massive data. To handle the challenges on vibration signals, machine learning (ML) algorithms are commonly used throughout health condition monitoring process of the rolling bearing [1].

ML approaches provide satisfying results on bearing fault diagnosis by mapping information from vibration signals to health condition. However, the problem of increasing vulnerabilities against malicious attacks arises in such a fault diagnosis system as in other industrial application employing ML approaches [4]. The malicious attacks performed against ML-based systems are called as Adversarial Machine Learning (AML) attacks. The purpose of AML is to covertly force ML model to change its nominal decision by injecting slight perturbations. Thus, the performance of the model may decrease as the number of misclassifications increases. [5].

In our previous study [6], we proposed a deep residual learning (DRL) network model to monitor the health condition of a rolling bearing. The performance of the proposed diagnosis system was tested on a public dataset containing 10 different health condition with a comparison to the state-of-art diagnosis methods. We obtained a performance with an average accuracy of 99.98% outperforming existing methods.

In this study, our first aim is to investigate the robustness of our fault diagnosis system presented in [6]. For this purpose, a common white-box adversarial attack named *projected gradient descent* (PGD) [7] is injected into the DRL model in the diagnosis system. The PGD adversarial attack is widely considered the strongest "first-order adversarial" attack in the literature [8]. The second aim of the study is to analyze *adversarial training* [7] as a defense approach for PGD adversarial attack. The performance of the system under attack with and without defense mechanism is tested for 10 different health conditions. The results obtained show that 1) bearing fault diagnosis systems are vulnerable to AML, and 2) *adversarial training* performs a defense against PDG adversarial attack. However, the defense strategy needs to be improved.

Mustafa Sinasi Ayas and Selen Ayas would like to thank the Scientific and Technological Research Council of Turkey (TUBITAK) BIDEB 2219 program.

II. Related Work

A. Fault Diagnosis

A convolutional neural network (CNN) based technique inspired by LeNet-5 was proposed for fault diagnosis in [9], where vibration signals of a motor bearing taken from a public dataset were converted into 2D gray-level images. A deep convolutional TL network with a two-stage approach was presented in [10]. In one of the stages, the health condition of the bearing was classified employing 1D CNN, which learns discriminative features of raw signals. A deep adversarial domain adaptation model was presented as a fault diagnosis solution in [11]. A deep stack autoencoder and feature learning were combined to extract more valuable bearing fault features. In this work, pretty good health classification results were obtained compared to published traditional ML and deep learning (DL) methods. A CNN model fed by 2D gray-level images was presented in [1]. A signal-to-image mapping technique was used to convert vibration signals into 2D images. Ayas et al. [6] proposed a DRL based model built to learn end-to-end mapping between 2D images generated from vibration signals and health condition of motor bearing.

B. Adversarial Machine Learning

Szegedy et al. [12] was the first to mention the vulnerabilities in neural networks in 2013. They emphasized that the network can misclassify an image by injecting a scarcely perceivable perturbation, which is calculated by maximizing the network's prediction error. By injecting the computed perturbation into the legitimate image, a misleading image that the human eye cannot perceive can be obtained. Then, in 2015, Goodfellow et al. [13] presented the Fast Gradient Sign Method (FGSM) which can generate adversarial examples according to the gradient calculated efficiently utilizing backpropagation. DeepFool was proposed in [14], which tricks deep neural networks by producing minimal perturbation sufficient for misclassification. The Basic Iterative Method (BIM) an iterative version of the FGSM introduced was presented by Kurakin et al. [15]. The projected gradient descent (PGD) known as a different form of the BIM was presented as the most powerful "first-order adversarial" attack by Madry et al. [7]. The PGD uses first order information regarding the network.

Studies on AML applications have been growing up rapidly. In 2018, Suciu et al. [16] introduced poisoning attack against Android malware detection system having a linear classifier. In 2019, Kuppa et al. [17] launched a gray-box inference integrity attack against unsupervised anomaly detector and achieved a notable result. In 2020, Xu et al. [18] launched an inference integrity gray-box attack against malware classifier, using both n-strongest nodes and FGSM methods having the benefit of generating valid file evading detection. In 2021, Vakhshiteh et al. [19] presented a comprehensive review study on adversarial attacks performed against face recognition systems. The paper emphasizes that face recognition systems are vulnerable to AML even if input images look natural.

III. Projected Gradient Descent Adversarial Attack

PGD adversarial attack was originally proposed by Madry et al. [7] as the most strongest "first-order adversarial" attack. Actually PGD is a well-known optimization technique projecting gradients in a ball [20]. PGD is basically the same as BIM, iterative version of FGSM with small step size, except that PGD starts at a random point in the ball and perform random restarts. (1) is the iterative equation of the PGD [7].

$$x^{t+1} = \prod_{x+S} (x^t + \alpha \, sign(\nabla_x L(\theta, x, y))) \tag{1}$$

where $L()$ is a loss function, x is the input to the model whose parameters represented by θ, y is the target related to the x, $\prod_{x+S}$ is a projection operator with perturbation set $x + S$, and α is a gradient step size. The cross-entropy loss function is used in this study.

The y parameter in (1) emphasized, because adversarial attacks can trick the ML classifier either as *targeted* or *untargeted*. In the *targeted* case, the malicious attacker forces the classifier to produce a certain class of output. On the other hand, the classifier is expected to give a misclassification where the label of the output class is not important in the *untargeted* case.

IV. Defending Against AML: Adversarial Training

The idea behind defense mechanism against AML can be classified into three main groups [8]: 1)Target model can be modified to increase robustness of the model subject to adversarial examples, 2) Input of model can be modified to remove perturbations, and 3) Additional mechanism such as detectors can be added.

The widely used defense mechanisms are inspired from the first idea. Adversarial training, aiming to improve model robustness by utilizing adversarial samples during the training process, is the most commonly employed defense mechanism. Basically, adversarial training is formulated as a minimax problem [7]:

$$\underset{\theta}{min}\mathbb{E}_{(x,y)\sim D}[\underset{\delta\in S}{max}L(\theta, x+\delta, y)] \tag{2}$$

where S is set of perturbations, D is data distribution over (x, y) pairs. The inner *maximization* problem solves maximizing the loss function problem for the input x, which corresponds to potential adversarial sample. Conversely, the outer *minimization* problem aims to find model parameters θ which minimizes the loss value calculated with adversarial input x. As a result, the *maximization* problem corresponds to attacking a model, while the *minimization* problem is to train a robust model against AML.

V. Experimental Results and Discussion

A. Fault Diagnosis System

In our previous study [6], we proposed a DRL method for bearing fault diagnosis. The layer configuration of the

model is given in Table I. The DRL method was tested on a commonly utilized public dataset provided by Case Western Reserve University (CWRU) Bearing Data Center [21].

The dataset [21] contains vibration signals were collected using accelerometers attached to the CWRU bearing test stand for 10 different corresponding health conditions. While one of these health condition is normal case, the other 9 cases are different faulty conditions.

The vibration signals were converted to 2D images with the size of 64×64 in data preprocessing step. Some examples of 2D images acquired for each health condition are demonstrated in Fig. 1 within green dashed rectangle. For details of the conversion process, we refer to our previous study [6]. The images look like perturbation images injected during a general AML to manipulate the original data.

TABLE I: The layer configurations of the DRL model [6]

Layer Name		
conv		7x7, 64, s:2, p:3
pooling		3x3 max pooling, s:2, p:1
Block_A	conv1	3x3, 64, s:1, p:1
	conv2	3x3, 64, s:1, p:1
Block_A	conv1	3x3, 64, s:1, p:1
	conv2	3x3, 64, s:1, p:1
Block_B	conv1	3x3, 128, s:2, p:1
	conv2	3x3, 128, s:1, p:1
	conv2	-
pooling		8x8 average pooling, s:8, p:0
fully connected		128x10 fully connections

B. Implementation of PGD Adversarial Attack

In this section, the average classification accuracy of 10 different bearing health conditions is analyzed after PGD adversarial attack. It should be noted that the average accuracy was obtained as 99.98% by the DRL model before the attack [6]. The implemented adversarial attack is an *untargeted attack* that aims to force the DRL model to produce a misclassification result without specifying a custom output label.

First, the effects of ϵ and α on the *untargeted attack* performance is discussed. In the experiments, the number of steps is set to be 40, ϵ is set to 0.005 and increased to 0.02 with a 0.005 incremental step. Because our converted images are 8 bits, the increase of ϵ is chosen to be higher than $1/255$ to avoid dismissing all the information below $1/255$ of the dynamic range. Besides ϵ, the effect of α is also examined by adjusting its value between 0.0005 and 0.00125. The results acquired from the fault diagnosis system, as summarized in Table II, show that an increase in ϵ results in a dramatic decrease of accuracy. Even for a quite small ϵ value, i.e. $\epsilon = 0.005$, the classification accuracy of the DRL model in fault diagnosis system reduces to 61.25%. The produced perturbations and generated adversarial samples are visualized in Fig. 1. It is seen from the figure that the generated adversarial images resemble the clear images when viewed from the human eye. To measure the visual quality between adversarial and clear images, Peak Signal-to-Noise Ratio (PSNR), i.e., $PSNR = 10 \times log_{10}\left(\frac{MAX_I^2}{MSE}\right)$ is

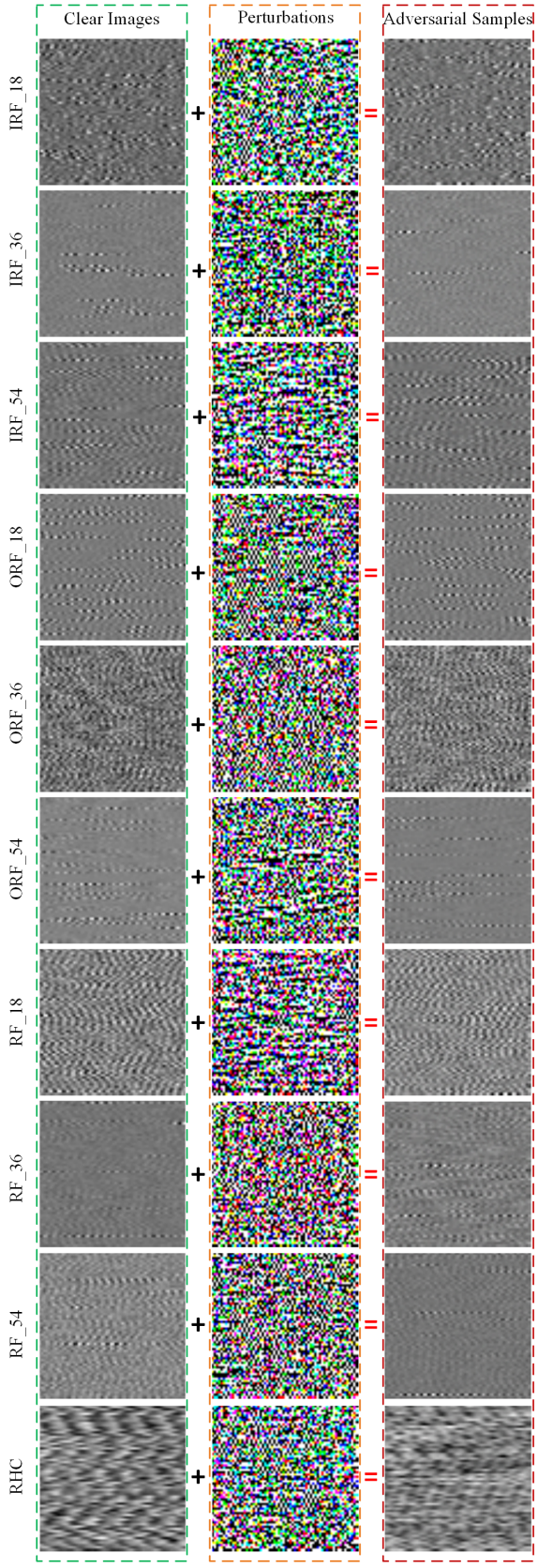

Fig. 1: Examples of PGD adversarial attacks to fool DRL model used in the fault diagnosis system with 10 different health condition classes.Left: clear images, Middle: adversarial perturbations generated by the PGD technique, Right: adversarial samples.

TABLE II: Average accuracy (%) for different ϵ and α

ϵ \ α	0.0005	0.00075	0.001	0.00125
0.005	61.25	61.19	61.38	61.25
0.01	23.00	21.25	20.56	20.50
0.015	3.38	1.69	1.18	1.06
0.02	0.50	0.38	0.31	0.19

TABLE III: Average accuracy (%) after adversarial training

ϵ \ α	0.0005	0.00075	0.001	0.00125
0.005	83.07	83.00	83.09	83.06
0.01	27.56	26.37	25.88	25.53
0.015	12.45	11.16	10.34	10.01
0.02	5.43	5.37	5.26	5.09

calculated where MSE is the mean squared error and MAX_I is the maximum possible value of image I. The average PSNR value is $40.28\ dB$ for all of the images in the test dataset. In addition, PSNR value higher than 30 dB can be regarded as visually identical images [22].

C. Defending against PGD Adversarial Attack

Training a neural network with a dataset including adversarial samples in addition to clear images, remarkably increases the robustness of the network against adversarial attacks [13]. In this context, the robustness of the DRL model in fault diagnosis system is re-examined after retraining the model using a new dataset containing both clear and adversarial samples. The adversarial samples are generated by the PGD adversarial attack with $\epsilon = 0.005$ and $\alpha = 0.0005$.

Table III reports the average classification accuracy after adversarial training. The achieved average accuracies for 10 different health conditions are calculated as 83.07%, 27.56%, 12.45%, 5.43% for $\epsilon = 0.005$, $\epsilon = 0.01$, $\epsilon = 0.015$, and $\epsilon = 0.02$, respectively, with $\alpha = 0.0005$. The results indicate that the robustness of the DRL model is improved and its classification performance is increased 35.6% for $\epsilon = 0.005$ and $\alpha = 0.0005$, when compared to Table II.

VI. Conclusion

This paper presents how AML effects fault diagnosis system by fooling its ML model deciding the health condition of the system. In this paper, the PGD adversarial attack is utilized to generate perturbations injected to clear samples to produce adversarial samples. The experimental results indicate that it is possible to implement AML with existing methods and even for fairly small perturbations, the classification accuracy of the model in the diagnosis system decreases resulting in misdiagnoses. Moreover, the generated adversarial samples look like the clear images when viewed from the human eye, that is, the implemented attack succeeds without being detected. The results also show that *adversarial training* can be used as a defense mechanism to increase the robustness of the DRL model against potential AML.

References

[1] J. Zhao, S. Yang, Q. Li, Y. Liu, X. Gu, and W. Liu, "A new bearing fault diagnosis method based on signal-to-image mapping and convolutional neural network," *Measurement*, vol. 176, p. 109088, 2021.

[2] F. Jia, Y. Lei, L. Guo, J. Lin, and S. Xing, "A neural network constructed by deep learning technique and its application to intelligent fault diagnosis of machines," *Neurocomputing*, vol. 272, pp. 619–628, 2018.

[3] X.-W. Chen and X. Lin, "Big data deep learning: challenges and perspectives," *IEEE access*, vol. 2, pp. 514–525, 2014.

[4] R. S. S. Kumar, M. Nyström, J. Lambert, A. Marshall, M. Goertzel, A. Comissoneru, M. Swann, and S. Xia, "Adversarial machine learning-industry perspectives," in *2020 IEEE Security and Privacy Workshops (SPW)*. IEEE, 2020, pp. 69–75.

[5] E. Anthi, L. Williams, M. Rhode, P. Burnap, and A. Wedgbury, "Adversarial attacks on machine learning cybersecurity defences in industrial control systems," *Journal of Information Security and Applications*, vol. 58, p. 102717, 2021.

[6] S. Ayas and M. S. Ayas, "A novel bearing fault diagnosis method using deep residual learning network," *Multimedia Tools and Applications*, pp. 1–17, 2022.

[7] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

[8] N. Akhtar, A. Mian, N. Kardan, and M. Shah, "Advances in adversarial attacks and defenses in computer vision: A survey," *IEEE Access*, vol. 9, pp. 155 161–155 196, 2021.

[9] L. Wen, X. Li, L. Gao, and Y. Zhang, "A new convolutional neural network-based data-driven fault diagnosis method," *IEEE Transactions on Industrial Electronics*, vol. 65, no. 7, pp. 5990–5998, 2017.

[10] L. Guo, Y. Lei, S. Xing, T. Yan, and N. Li, "Deep convolutional transfer learning network: A new method for intelligent fault diagnosis of machines with unlabeled data," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 9, pp. 7316–7325, 2018.

[11] Z.-H. Liu, B.-L. Lu, H.-L. Wei, L. Chen, X.-H. Li, and M. Rätsch, "Deep adversarial domain adaptation model for bearing fault diagnosis," *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, vol. 51, no. 7, pp. 4217–4226, 2019.

[12] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

[13] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[14] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2574–2582.

[15] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," *arXiv preprint arXiv:1611.01236*, 2016.

[16] O. Suciu, R. Marginean, Y. Kaya, H. Daume III, and T. Dumitras, "When does machine learning {FAIL}? generalized transferability for evasion and poisoning attacks," in *27th USENIX Security Symposium (USENIX Security 18)*, 2018, pp. 1299–1316.

[17] A. Kuppa, S. Grzonkowski, M. R. Asghar, and N.-A. Le-Khac, "Black box attacks on deep anomaly detectors," in *Proceedings of the 14th International Conference on Availability, Reliability and Security*, 2019, pp. 1–10.

[18] P. Xu, B. Kolosnjaji, C. Eckert, and A. Zarras, "Manis: Evading malware detection system on graph structure," in *Proceedings of the 35th Annual ACM Symposium on Applied Computing*, 2020, pp. 1688–1695.

[19] F. Vakhshiteh, A. Nickabadi, and R. Ramachandra, "Adversarial attacks against face recognition: A comprehensive study," *IEEE Access*, vol. 9, pp. 92 735–92 756, 2021.

[20] S. Bubeck *et al.*, "Convex optimization: Algorithms and complexity," *Foundations and Trends® in Machine Learning*, vol. 8, no. 3-4, pp. 231–357, 2015.

[21] Case western reserve university bearing data center website. [Online]. Available: https://engineering.case.edu/bearingdatacenter

[22] R. Huang and K. Sakurai, "A robust and compression-combined digital image encryption method based on compressive sensing," in *2011 Seventh international conference on intelligent information hiding and multimedia signal processing*. IEEE, 2011, pp. 105–108.