Could Min-Max Optimization Be A General Defense Against Adversarial Attacks?

Rana Abou Khamis and Ashraf Matrawy
School of Information Technology, Carleton University, Ottawa, Canada.
{rana.aboukhamis,ashraf.matrawy}@carleton.ca

***Abstract*— Adversarial learning based on Min-Max formulations has been broadly employed in deep neural networks (DNNs) as an effective defense approach against adversarial attacks. Motivated by the level of resistance achieved by adversarial trained models against a single type of adversarial attack, in this paper we investigate if utilizing Min-Max formulation in various deep learning-based Intrusion Detection System (IDS) architectures may be considered an optimized defense against different types of state-of-the-art adversarial attacks. To investigate this, we generate adversarial samples using multiple attack methods using two benchmark IDS datasets, UNSW-NB 15 and NSL-KDD. Then, we conduct comprehensive experiments on adversarial trained models, including convolutional neural networks (CNN) and recurrent neural networks (RNN) architectures. Our results demonstrate that the adversarial IDS models that were trained against one type of attack show robustness against different adversarial attacks that could reach up to 40% higher accuracy than IDS models trained by adversarial-free (baseline) datasets. Finally, we demonstrate that training models with Carlini and Wagner (CW) adversarial samples in CNN leads to better robustness against other adversarial attacks.**

Keywords: Deep Neural Networks, Intrusion Detection System, Adversarial Samples, Adversarial Learning.

I. Introduction

While deep learning-based IDS aims to effectively classify benign and malignant inputs, adversarial samples often expose blind spots in the inputs. Adversarial samples have been intentionally designed to target a model to degrade performance and result in incorrect decisions with high confidence. This wrong output can be done by adding a calculated perturbation to the input. With this pressing problem, several researchers [1] [2] proposed several adversarial defense methods in the past few years, and many types of research in computer vision, malware detection, and IDS [3] [4] [5] show that augmenting crafted inputs during the training time significantly robust models against adversarial attacks and fortified them.

In our previous work [6], we conducted a comprehensive experimental study and analysis of Min-Max optimization that combines both attack and defense in deep learning-based intrusion detection systems on three popular deep neural network architectures, including artificial neural networks (ANN), Convolutional Neural Networks (CNN) and one of the Recurrent Neural Networks (RNN) architecture called Long Short-term Memory (LSTM) and two IDS benchmark datasets, UNSW-NB15 and NSL-KDD, in an adversarial environment. We provided a performance comparison between several state-of-the-art adversarial attacks, such as the Fast Gradient Sign Method (FGSM) [7], and more powerful attacks with multi-steps like the Basic Iterative Method (BIM) [8], Projected Gradient Descent (PGD) [3], Carlini and Wagner (CW) [9], and Deepfool [10]. The result in [6] provides evidence that the IDS framework proposed in [11] can reliably solve the optimization problem in deep learning-based intrusion detection systems trained by a single type of adversarial attack. Motivated by the degree of robustness we achieved in [6], we make the following **contributions** in this paper. First, we demonstrate the power of Min-Max formulation and how the pre-trained CNN and RNN models by a single type of adversarial samples achieved better robustness against different types and unknown adversarial samples and beat the accuracy of the baseline IDS models trained by an adversarial-free dataset. Secondly, in the CNN model, we show that the defense technique based on the Min-Max formulation works better when it is adversarially-trained with Carlini and Wanger samples in comparison with other types of adversarial samples. Overall, the investigation in this paper extends the experiments in [6] using CNN and RNN. We investigate whether the IDS framework

979-8-3503-7099-7/24/$31.00 ©2024 IEEE

proposed in [11], which is based on the Min-Max formulation in deep learning-based IDS, can be considered a general defense against various types of adversarial attacks and attacks unknown to the model. We note that this paper is based on the content of the first author's thesis [12]. The rest of the paper follows. In section II, we review related work on adversarial training, followed by the experiment methodology in section III. Section IV presents experiment results and analysis. We will summarize our findings and outline the future work in section V.

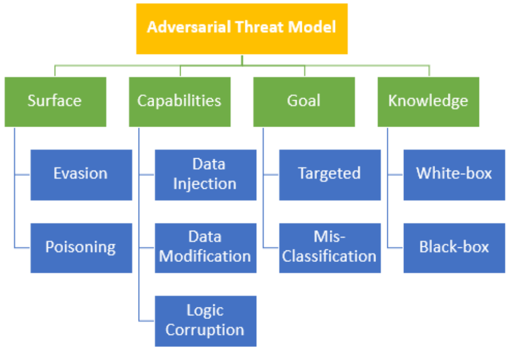

Fig. 1: Adversarial Attack Taxonomy

II. Related work

Adversarial samples in DNN have been studied in various fields, and research concentrates on two main points: (1) generating capable adversarial samples that can attack and deceive a model with small perturbations; and (2) training and defending models against adversarial samples. We reviewed some previous works that have studied adversarial attacks and proposed defense techniques, and we find, as shown in Table I, that many researchers have widely studied and developed adversarial attack methods in computer vision and image recognition, whereas adversarial attacks and defense in IDS are still in the early stages of research and have significant potential for further studies. Besides, Table I presents some research that uses adversarial training defense techniques based on Min-Max formulation. We show that most of the work in Table I is in image classification. Finally, we classified some IDS research in Table II based on the adversarial attack taxonomy presented in Figure 1. We design our adversarial attack taxonomy in Figure 1 based on the two surveys of adversarial attacks [30] and [31], where each adversarial attack can be classified under this classification. The thorough study of related work in adversarial attack and defense encourages us to study whether Min-Max optimization can represent a general defense and minimize the risk of adversarial attacks for the adversarially-trained model by one type of adversarial sample.

III. Methodology

The experimental approach we use in this paper is similar to our previous work [6], including model architecture and hyperparameters for the learning algorithm to build the CNN and RNN over UNSW-NB15 and NSL-KDD datasets. The CNN network architecture we built for this work has a combination of three types of layers: convolution, pooling, and fully connected layers. The CNN architecture has three Conv1D that extract features from the packet flow. It also has two Max-Pooling1D that apply dimensionality reduction to reduce the inputs' size and decrease computation time. We use four fully connected layers with the ReLU activation function and one output layer with a softmax activation layer to classify the input into one of the categories: benign or attack. As for RNN models, we use two LTSM layers with sigmoid and one fully connected layer with ReLU. We completed 10 epochs with 32 batch sizes for both architectures.

We conduct our experiments and train all deep learning models on a large dataset divided between benign and attack flows. Our NSL-KDD [32] training set consists of approximately 100,778 flows and 25,195 records for testing. While we use UNSW-NB15 [33] 1,17478 records for training and 57,863 for testing with the same preprocessing techniques used in [6] to remove outliers, normalize features, and apply feature selection processes to avoid overfitting and enhance the performance of the models in terms of prediction accuracy. We consider several assumptions regarding our attack threat model. The attack is an evasion attack, in which an attacker has access to the IDS model during prediction time, causing the model decision to be misclassified, taking into account a complete knowledge of the targeted model to perform a white-box attack.

Evaluation Metric: For the evaluation metric,

TABLE I: Overview of Adversarial Training and Min-Max optimization in relevant work.

Paper	Dataset	Application	Algorithm	Attack Method
[6]	UNSW-NB15,NSL-KDD	IDS	ANN,CNN,RNN	FGSM,PGD,CW,Deepfool
[11]	UNSW-NB15	IDS	ANN	FGSM, BGA, BCA
[13]	CIFAR-10,ImageNet	Image Classification	Wide-ResNet	FGSM,PGD,Deepfool
[14]	CIFAR-10	Image Classification	Wide-ResNet	PGD, CW
[15]	CIFAR-10/100,Imagenet	Image Classification	Resnet, WideRenset	PGD
[16]	MNIST,CIFAR-10	Image Classification	MLP, All-CNN, LeNet, etc	PGD
[17]	MNIST,CIFAR-10	Image Classification	CNN	PGD, CW
[2]	MNIST,CIFAR-10,ImageNet	Image Classification	MLP, LeNet, ConvNet, ResNet, etc	Deepfool, FGS
[4]	Binary file	Malware Detection	ANN	FGSM, PGD
[18]	MNIST,CIFAR-10	Image Classification	ANN	FGSM, BIM
[19]	MNIST	Image Classification	NN	PGD
[20]	MNIST	Image Classification	CNN	FGSM,PGD
[21]	MNIST,CIFAR10	Image Classification	DDN-Rony,TRADES, CNN	PGD
[5]	MNIST	Image Classification	CNN	FGSM,PGD
[3]	MNIST,CIFAR-10	Image Classification	CNN	FGSM,PGD,CW

TABLE II: Classification of relevant works in Adversarial Attacks based on Figure 1

Paper	Application	Algorithms	Datasets	Surface	Capabilities	Goals	Knowledge
[11]	IDS	ANN	UNSW-NB15	Evasion	Data injection	Targeted	WhiteBox
[6]	IDS	ANN,CNN,RNN	NSL-KDD,UNSW-NB15	Evasion	Data injection	Targeted	WhiteBox
[22]	IDS	FNN,SNN	BoT-IoT	Evasion	Data injection	Targeted	WhiteBox
[23]	IDS	GAN	KDD99	Evasion	Data injection	Targeted	BlackBox
[24]	IDS	DNN	NSL-KDD	Evasion	Data modification	Targeted	WhiteBox
[25]	IDS	DNN	NSL-KDD	Evasion	Data modification	Targeted	White Box
[26]	IDS	NB,LR,ST, SVM	ASNM-NPBO	Evasion	Data modification	Targeted	WhiteBox
[27]	IDS	SVM	HTTP Traffic Attack	Poison	Data Injection	Targeted	WhiteBox
[28]	IDS	SVDD	HTTP Traffic	Poison	Data modification	Targeted	White Box
[29]	IDS	Supervised ML	IoT smarthome dataset	Evasion	Data modification	Targeted	white-box

we use the prediction accuracy (AC) that represents the total number of correctly classified samples from both benign and attacking samples among all predicted samples to evaluate the performance of the trained models in the adversarial environment. The adversarial training process based on the Min-Max approach is formalized in *Algorithm I* introduced in [11]. The algorithm describes the IDS framework solution based on the Min-Max formulation.

First, we train the four baseline models with adversarial-free datasets for both architectures. We achieve 96% for both CNN models: CNN-KDD trained using the KDD dataset and CNN-UNSW trained using UNSW. Similarly, we achieve 96% for both RNN models: RNN-KDD trained using the KDD dataset and RNN-UNSW trained using UNSW. Then, we generate strong adversarial samples from FGSM [7], PGD [3], CW [9], and Deepfool [10] that maximize the loss on UNSW and NSL-KDD using the inner maximizer. We test the four baseline models, CNN-UNSW, CNN-KDD, RNN-UNSW, and RNN-KDD, against the set of attacks we consider in this paper. Models accuracy significantly decreases, as shown in Tables III, IV, V, and VI. Thereafter, we retrain the CNN and RNN models using the Min-Max approach formulated in *Algorithm I* by a single type of adversarial sample. We retrained 16 adversarial IDS models using two datasets to investigate the Min-Max approach in more detail. We trained eight adversarially-trained CNN models, eight adversarially-trained RNN models, and four baseline models for both CNN and RNN.

We then evaluate the robustness of adversarial models that are retrained using *Algorithm I* against a single type of attack using a Min-Max approach against different and unknown adversarial attacks.

IV. Experiments

In this work, we utilize the inner maximizer in IDS framework presented in [11] to generate adversarial samples from different methods to attack CNN and RNN models trained with a single type of attack to evaluate their robustness and compare the results with the four CNN and RNN baseline IDSs: CNN-UNSW, CNN-KDD, RNN-UNSW, and RNN-KDD. Selecting CNN and RNN architectures and not ANN, because, in our previous work in [6], we studied the Min-Max on multiple adversarial attacks using ANN architecture. We perform the experiments in this work using only four attacks, including FGSM, PGD, CW, and Deepfool due our observation in [6] that BIM and PGD effectiveness are nearly similar in most of the models. In our analysis, we measured the accuracy before and after applying Min-Max formulation. We also compared the impact of adversarial samples generated by inner-maximization on re-trained IDS and evaluate their robustness against multiple adversarial attacks during testing phase. The tables IV,III,V,& VI present the accuracy of the trained models, with each column representing a trained models with a single type of adversarial using the Min-Max method. The row labeled ”adversarial attack methods” represents the techniques employed by the inner maximizer to create adversarial samples.

A. Performance of Adversarial Trained CNN IDSs

In this experiment, we test the CNN adversarially-trained models with a single type of perturbation by other adversarial attacks: FGSM, CW, PGD, and Deepfool, as shown in Figures 2 and 3. The result values of the two tables III and IV summarize the prediction accuracy of five trained CNN IDSs, including the two baseline CNN models.

TABLE III: **Prediction Accuracy Results for CNN-based IDS on UNSW-NB15**

-	Adversarial Attack Methods			
Model	FGSM	CW	PGD	Deepfool
CNN Baseline	**71.31**	**66.2**	**46.45**	**66.94**
FGSM model	**94.51**	68.82	72.76	50.42
CW model	98.69	**88.86**	71.6	78.05
PGD model	73.77	69.52	**95.47**	58.54
Deepfool model	69.07	64.66	53.9	**92.05**

TABLE IV: **Prediction Accuracy Results for CNN-based IDS on NSL-KDD**

-	Adversarial Attack Methods			
Model	FGSM	CW	PGD	Deepfool
CNN Baseline	**32.48**	**23.37**	**18.71**	**12.63**
FGSM model	**95.69**	66.05	51.17	70.29
CW model	87.38	**92.67**	72.04	77.96
PGD model	55.36	58.43	**95.52**	59.13
Deepfool model	66.16	49.52	43.25	**94.31**

We observe that adversarial-free models (baselines) have low accuracy compared to the other adversarial-trained CNN models. The results are expected because we train the baseline models with adversarial-free datasets, as seen in the bold values in Tables III and IV. We are also not surprised that all training methods are relatively robust to adversarial samples from the same attack methods, as seen in the bold and underlined values in Table III and Table IV.

Figures 2 and 3 show the information in the tables in a graphical representation. The baseline model’s accuracy decreased when attacked with FGSM, PGD, CW, and Deepfool samples generated by the inner maximizer. By retraining the CNN models using Min-Max, the prediction accuracy increased for all models in comparison to the baseline models while being attacked.

It is worth noticing that CW-CNN models in both UNSW and NSL-KDD achieved higher than 70% against all adversarial samples. We underlined the CW model in both Tables III and IV. The FGSM model has the second highest accuracy, where the average of all accuracy is almost 70%. PGD and Deepfool achieve 67% and 63% as an average accuracy. Examining the CNN-based IDS in both UNSW-NB15 and NSL-KDD, the Deepfool models are more resilient against FGSM and CW. However, Deepfool models are less resilient against PGD attacks that have the highest impact, achieving 53.9% and 43.25 against Deepfool models in UNSW-NBB and NSL-KDD, respectively.

B. Performance of Adversarial Trained RNN IDSs

Similar to adversarial-trained CNN-based IDSs, we noticed all RNN models are more resilient to some adversarial samples than other adversarial samples.

TABLE V: **Prediction Accuracy Results for RNN-based IDS on UNSW-NB15**

-	Adversarial Attack Methods			
Model	FGSM	CW	PGD	Deepfool
RNN Baseline	**44.74**	**43.94**	**43.35**	**10.36**
FGSM model	**88.83**	73.53	55.59	62.19
CW model	69.55	**87.09**	63.37	67.04
PGD model	63.67	68.22	**95.23**	69.13
Deepfool model	59.4	54.7	25.66	**94.75**

TABLE VI: **Prediction Accuracy Results for RNN-based IDS on NSL-KDD**

-	Adversarial Attack Methods			
Model	FGSM	CW	PGD	Deepfool
RNN Baseline	**46.92**	**5.76**	**18.07**	**31.59**
FGSM model	**95.53**	71.30	60.85	54.93
CW model	90.95	**95.21**	62.99	60.78
PGD model	61.33	67.51	**95.52**	78.22
Deepfool model	65.58	57.55	72.66	**95.32**

The predictions made by the four adversarial trained models, which include the RNN IDS baselines (RNN-UNSW and RNN-KDD), are shown in Figures 4 and 5. As expected, we noticed that the adversarial-free models (baseline) have low accuracy compared to the other adversarial-trained

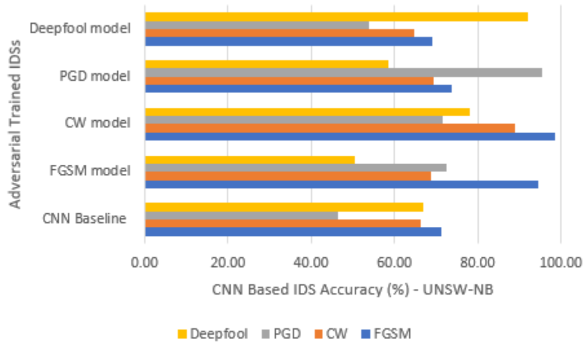

Fig. 2: Evaluation of multiple attacks against adversarially-trained CNN models for UNSW-NB15

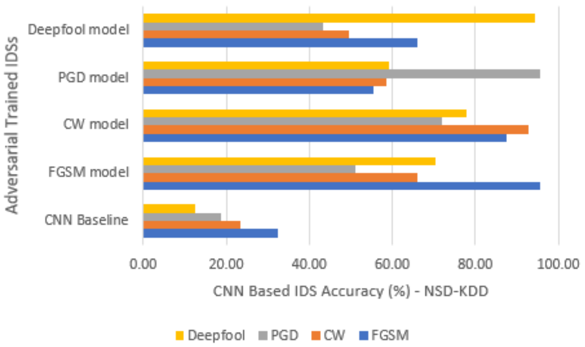

Fig. 3: Evaluation of multiple attacks against adversarially-trained CNN models for NSL-KDD

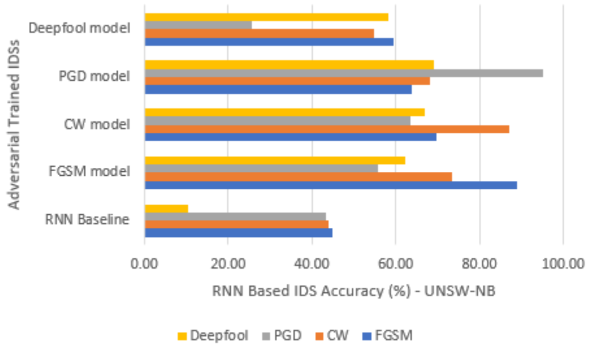

Fig. 4: Evaluation of multiple attacks against adversarially-trained RNN models for UNSW-NB15

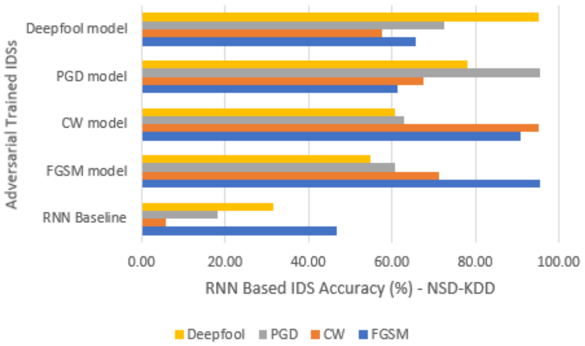

Fig. 5: Evaluation of multiple attacks against adversarially-trained RNN models for NSL-KDD

RNN models. We are also not surprised that all training methods are relatively robust to adversarial samples from the same adversarial attack methods, as shown in the bold and underlined values in Tables VI and V. For example, the baseline accuracy of the RNN IDS model of UNSW decreased to 44% via FGSM samples. By retraining the RNN IDS of UNSW with FGSM samples using Min-Max, the accuracy of the prediction increased to 88.83%. If we take a closer look at both Tables V and VI, we observe that the FGSM model is less robust to Deepfool attacks, where accuracy achieves 54.93% in RNN-KDD and 62.19% in RNN-UNSW. This accuracy is consistent with what has been found in CNN-UNSW, where accuracy in the FGSM model achieves 50.42% against deepfool attacks. Similarly, we observe that CW-RNN models of both UNSW and NSL-KDD achieved high accuracy against all other adversarial attacks, where accuracy against all attacks achieved higher than 60%.

V. Discussion and Conclusion

We study the robustness of models that are adversarially-trained with a single type of adversarial sample in an environment where multiple adversarial attacks could happen. Through the comprehensive comparison and analysis of the experimental results, we can draw the following conclusions about the robustness of the adversarial-trained IDSs:

- All Deepfool models have the lowest robustness against all four adversarial attack methods. More specifically, the Deepfool model's robustness in RNN-UNSW was the least compared to other models.
- CW models have the best robustness against all other adversarial attack methods. More specifically, the accuracy of CNN-UNSW and CNN-KDD, when trained by CW samples, outperforms other adversarial-trained models. We noticed that some adversarial IDS models are more robust than other adversarial samples.

The CW model's accuracy is 98.69% against FGSM samples, complying with CW sample accuracy that achieved only 88.86% against new CW samples, although the CW model was trained by CW samples, not FGSM samples. We speculate that this might be because the FGSM attack is considered a fast attack, not robust, and generally has less effectiveness among all adversarial models.

- We demonstrate that training CNN models with CW adversarial samples improves model performance against other adversarial samples.
- FGSM and PGD models have the same robustness against all adversarial attack methods in both datasets in CNN and RNN.

The overall accuracy did not decrease more than 50% of the initial accuracy of the baseline models. However, following the results of this paper, although the Min-Max training approach shows reasonable robustness in the face of multiple adversarial attacks generated by the inner-maximizer, we did not obtain the best possible performance and robustness for all CNN and RNN models. In addition, the models in this paper did not outperform the performance of models trained by a single type of adversarial attack and attacked by the same type [6]. As a result, finding a general defense and solution for IDS based on deep learning is still challenging. Future work may consider more measurement metrics like precision rate, specificity, recall rate, and evasion rate to evaluate the performance and compare various models' robustness and the effectiveness of adversarial attacks.

VI. ACKNOWLEDGEMENT

This work was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC) through the NSERC Discovery Grant program.

REFERENCES

[1] Y. Zeng *et al.*, "A deep learning based network encrypted traffic classification and intrusion detection framework," *IEEE Access*, vol. 7, pp. 45182–45190, 2019.

[2] Z. Yan *et al.*, "Deep defense: Training dnns with improved adversarial robustness," in *Advances in Neural Information Processing Systems*, pp. 419–428, 2018.

[3] A. Madry *et al.*, "Towards deep learning models resistant to adversarial attacks," *ICLR*, 2017.

[4] A. Al-Dujaili *et al.*, "Adversarial deep learning for robust detection of binary encoded malware," in *2018 IEEE SPW*, pp. 76–82, IEEE, 2018.

[5] E. Wong *et al.*, "Provable defenses against adversarial examples via the convex outer adversarial polytope," in *International Conference on Machine Learning*, pp. 5286–5295, 2018.

[6] R. Abou Khamis and A. Matrawy, "Evaluation of adversarial training on different types of neural networks in deep learning-based idss," in *2020 ISNCC*, pp. 1–6, IEEE, 2020.

[7] I. J. Goodfellow *et al.*, "Explaining and harnessing adversarial examples," *ICLR*, 2014.

[8] A. Kurakin *et al.*, "Adversarial examples in the physical world," *CoRR*, 2016.

[9] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57, IEEE, 2017.

[10] S.-M. Moosavi-Dezfooli *et al.*, "Deepfool: a simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE conference on CVPR*, pp. 2574–2582, 2016.

[11] R. Abou Khamis *et al.*, "Investigating resistance of deep learning-based ids against adversaries using min-max optimization," *IEEE ICC 20*, 2020.

[12] R. Abou Khamis, "Evaluating adversarial learning on different types of deep learning-based intrusion detection systems using min-max optimization," Master's thesis, Carleton Univeristy, 2020.

[13] A. Shafahi *et al.*, "Universal adversarial training.," in *AAAI*, pp. 5636–5643, 2020.

[14] M. Cheng *et al.*, "Cat: Customized adversarial training for improved robustness," 2020.

[15] Y. Balaji *et al.*, "Instance adaptive adversarial training: Improved accuracy tradeoffs in neural nets," *arXiv preprint arXiv:1910.08051*, 2019.

[16] J. Wang *et al.*, "Towards a unified min-max framework for adversarial exploration and robustness," 2019.

[17] H. Zhang *et al.*, "The limitations of adversarial training and the blind-spot attack," in *International Conference on Learning Representations*, 2018.

[18] U. Shaham *et al.*, "Understanding adversarial training: Increasing local stability of supervised models through robust optimization," *Neurocomputing*, vol. 307, pp. 195–204, 2018.

[19] A. Raghunathan *et al.*, "Certified defenses against adversarial examples," in *International Conference on Learning Representations*, 2018.

[20] F. Tramèr *et al.*, "Ensemble adversarial training: Attacks and defenses," in *6thICLR 2018-Conference Track Proceedings*, 2018.

[21] G. W. Ding *et al.*, "Max-margin adversarial (mma) training: Direct input space margin maximization through adversarial training," *arXiv preprint arXiv:1812.02637*, 2018.

[22] O. Ibitoye *et al.*, "Analyzing adversarial attacks against deep learning for intrusion detection in iot networks," *IEEE GlobalCom*, 2019.

[23] M. Usama *et al.*, "Generative adversarial networks for launching and thwarting adversarial attacks on network intrusion detection systems," in *2019 15th International Wireless Communications & Mobile Computing Conference (IWCMC)*, pp. 78–83, IEEE, 2019.

[24] Y. Peng *et al.*, "Evaluating deep learning based network intrusion detection system in adversarial environment," in *2019 IEEE 9th ICEIEC*, pp. 61–66, IEEE, 2019.

[25] Z. Wang, "Deep learning-based intrusion detection with adversaries," *IEEE Access*, vol. 6, pp. 38367–38384, 2018.

[26] I. Homoliak *et al.*, "Improving network intrusion detection classifiers by non-payload-based exploit-independent obfuscations: An adversarial approach," *ICST Trans. Security Safety*, 2018.

[27] B. Biggio *et al.*, "Security evaluation of pattern classifiers under attack," *IEEE transactions on knowledge and data engineering*, vol. 26, no. 4, pp. 984–996, 2013.

[28] M. Kloft and P. Laskov, "Online anomaly detection under adversarial impact," in *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics*, pp. 405–412, 2010.

[29] E. Anthi, L. Williams, A. Javed, and P. Burnap, "Hardening machine learning denial of service (dos) defences against adversarial attacks in iot smart home networks," 2021.

[30] O. Ibitoye *et al.*, "The threat of adversarial attacks on machine learning in network security–a survey," *arXiv preprint arXiv:1911.02621*, 2019.

[31] A. Chakraborty *et al.*, "Adversarial attacks and defences: A survey," *arXiv preprint arXiv:1810.00069*, 2018.

[32] M. Tavallaee *et al.*, "A detailed analysis of the kdd cup 99 data set," in *2009 IEEE symposium on CISDA*, pp. 1–6, IEEE, 2009.

[33] N. Moustafa *et al.*, "Unsw-nb15: a comprehensive data set for network intrusion detection systems," in *2015 MilCIS*, pp. 1–6, IEEE, 2015.