Analyzing Adversarial Attacks Against Deep Learning for Intrusion Detection in IoT Networks

Olakunle Ibitoye, Omair Shafiq and Ashraf Matrawy
School of Information Technology
Carleton University, Ottawa, Canada
Email: {Kunle.Ibitoye, Omair.Shafiq, Ashraf.Matrawy}@carleton.ca

***Abstract*— Adversarial attacks have been widely studied in the field of computer vision but their impact on network security applications remains an area of open research. As IoT, 5G and AI continue to converge to realize the promise of the fourth industrial revolution (Industry 4.0), security incidents and events on IoT networks have increased. Deep learning techniques are being applied to detect and mitigate many of such security threats against IoT networks. Feed-forward Neural Networks (FNN) have been widely used for classifying intrusion attacks in IoT networks. In this paper, we consider a variant of the FNN known as the Self-normalizing Neural Network (SNN) and compare its performance with the FNN for classifying intrusion attacks in an IoT network. Our analysis is performed using the BoT-IoT dataset from the Cyber Range Lab of the center of UNSW Canberra Cyber. In our experimental results, the FNN outperforms the SNN for intrusion detection in IoT networks based on multiple performance metrics such as accuracy, precision, and recall as well as multi-classification metrics such as Cohen Cappas score. However, when tested for adversarial robustness, the SNN demonstrates better resilience against the adversarial samples from the IoT dataset, presenting a promising future in the quest for safer and more secure deep learning in IoT networks.**

Index Terms Intrusion Detection, Adversarial samples, Feed-forward Neural Networks (FNN), Resilience, Self-normalizing Neural Networks (SNN), Internet of things (IoT).

I. Introduction

As the Internet of Things (IoT) emerges and expands over the next several years, the security risks in IoT will increase. There will be bigger rewards for successful IoT breaches and hence greater incentive and motivation for attackers to find new and novel ways to compromise IoT systems. Traditional methods and techniques for protecting against cyber threats in the traditional internet will prove inadequate in protecting against the unique security vulnerabilities that would be expected in the internet of things [1]. Hence security researchers and professionals would need to evaluate existing processes and improve upon them to create more efficient security solutions to address the security vulnerabilities in the emerging Internet of Things.

Managing security challenges in any network involves three broad strategies namely prevention, detection and mitigation. Successful security solutions for IoT networks will need to adopt all three measures. For the scope of this paper, we focus on Intrusion Detection Systems (IDS) and consider deep learning based IDS for detecting and classifying network traffic within an IoT enviroment.

Deep learning based IDS have an advantage over conventional anomaly based IDS because they help overcome the challenge of proper feature selections [2]. However, two major challenges of deep learning in security applications are the lack of transparency of the deep learning models [3], and the vulnerability of the deep learning models to adversarial attacks [4]. For the scope of this study, we focus on adversarial vulnerability of the deep learning models.

An adversarial attack occurs when an adversarial example is fed as an input to a machine learning model. An adversarial example is an instance of the input in which some feature has been intentionally perturbed with the intention of confusing a machine learning model to produce a wrong prediction. Szegedy et al. [4] demonstrated how a deep learning model for image recognition could be confused into making wrong predictions by introducing a tiny perturbation to the image. Other researchers [5] [6] have also proved that adversarial attacks are equally effective against deep learning models in network security applications such as malware detection and intrusion detection systems.

Klambauer et al. [7] proposed the Self-normalizing Neural Networks (SNN) which is a variant of the FNN that uses a Scaled Exponential Linear Unit (SELU) activation function.

Our Contributions in this paper are as follows: For our **first contribution**, this is to the best of our knowledge, the first study to demonstrate the effects of adversarial samples on a deep learning based Intrusion Detection System (IDS) within the context of an IoT network. **For our second contribution** we provide a comparison between the performance of an IDS implemented with two different deep learning models - a Self-normalizing Neural Network (SNN) and a typical Feed-forward Neural Network (FNN) within the context of an IoT network. In our **third contribution**, we demonstrate that while the IDS implemented with FNN performs better than the SNN based IDS with regards to performance metrics such as accuracy, precision and recall, the SNN based IDS is however more robust to adversarial samples. In our **fourth and final contribution**, we analyze the effects of feature normalization on the adversarial robustness of deep learning based IDS in IoT. This is the first study to the best of our knowledge to demonstrate that normalization of input features in a deep learning-based IDS adversely impacts the ability of the deep learning model to resist adversarial attacks.

978-1-7281-0962-6/19/$31.00 ©2019 IEEE

II. Related Work

While previous research [6] have utilized deep learning techniques for intrusion detection in traditional networks, in this study, we extend this research area by specifically applying deep learning for intrusion detection in the context of IoT. We then demonstrate that deep learning models used for intrusion detection in IoT can be confused with adversarial samples.

Koroniotis et al. [8] in the original paper that described the IoT dataset that we used for our experiments implemented LSTM, SVM and RNN machine learning techniques to analyze the IoT dataset but they did not evaluate the adversarial robustness of their machine learning models in their study. Additionally, their study only carried out binary classification on the dataset and the prediction output of the machine learning models was classified as either attack or normal traffic. We note that the usefulness of such studies prevails for network forensic analysis use cases where it is essential to classify the output into the various categories of attacks.

Hodo et al. [9] analyzed the threat of intrusion detection against IoT using a very limited dataset sample of 2313 training samples, 496 validation samples and 496 test samples. The dataset also contains only DDoS/DoS traffic and normal traffic. In a realistic IoT environment, we expect a larger network traffic dataset with hundreds of thousands or millions of records with a more heterogeneous attack profile on the network. We used a dataset containing over 3.6 million records and a more heterogeneous attack profile consisting of 5 target labels.

Zheng [6] implemented several adversarial attack algorithms against a deep learning based intrusion detection system in a traditional network using multi-layer perceptron Feed-forward Neural Network and compared the results from the various adversarial attacks. The author demonstrated that the deep learning based IDS classifier using the FNN was adversely impacted by the adversarial samples. However, the NSL-KDD dataset that was used was generated over a decade ago and may not represent the type of network attack traffic that would be expected in today's IoT networks. Warzynski et al. [10] also evaluated the NSL-KDD dataset by training a FNN to classify the network packets, and then tested the resilience of their model to adversarial examples. The dataset used in their experiment may not represent a typical IoT network traffic.

Based on our literature review and study of related work, we discovered that no researcher has evaluated the resilience of Self-normalizing Neural Networks (SNN) to adversarial examples for deep learning based IDS in IoT networks. Hence our study is novel and offers a useful contribution in understanding the security of machine learning and artificial intelligence in IoT.

III. Problem Definition and Proposed Study

Zheng [6] demonstrated that a deep learning based IDS that could correctly identify DoS attacks with an accuracy of 93% could have its performance degraded to as low as 24% with adversarial samples. In this study, our objective is to investigate how the performance accuracy of a deep learning-based IDS for IoT could be improved in the presence of adversarial samples.

Based on our literature review and to the best of our knowledge, no study has associated normalization in a deep learning model with the adversarial resilience of the model. Since a Self-normalizing Neural Network (SNN) maintains the stability of the network during the gradient descent process through internal normalization of the parameters [7], we seek to determine how normalization impacts the adversarial resilience. Currently, the performance of SNN in the context of intrusion detection in IoT is not known as well as its resilience to adversarial samples. Since no study, to the best of our knowledge, has tested the resilience of SNN for intrusion detection in IoT, we propose to carry out our study to close out this gap.

IV. Experimental Approach

In carrying out this study, we implement deep learning based intrusion detection systems for an IoT dataset. We then test the resilience of the deep learning IDSes to adversarial samples. To demonstrate this, we create our own adversarial samples from the IoT dataset which we used in training the deep learning models. The methods used in crafting the adversarial samples for this study are the Fast Gradient Sign Method (FGSM) [11], Basic Iteration Method (BIM)[12], and the Projected Gradient Descent[13]. Our study considers the following **assumptions**. The attacks are evasion attacks which are launched during the prediction phase of the deep learning model. Also, a complete knowledge of the deep learning model is assumed, hence they are white box attack. In this study, we do not target any specific prediction outcome, rather we seek to confuse the deep learning classifier to make a mistake and produce a misclassification. Hence a reliability attack. Our expected outcome is to degrade the performance of deep learning classifier, as measured by various performance metrics.

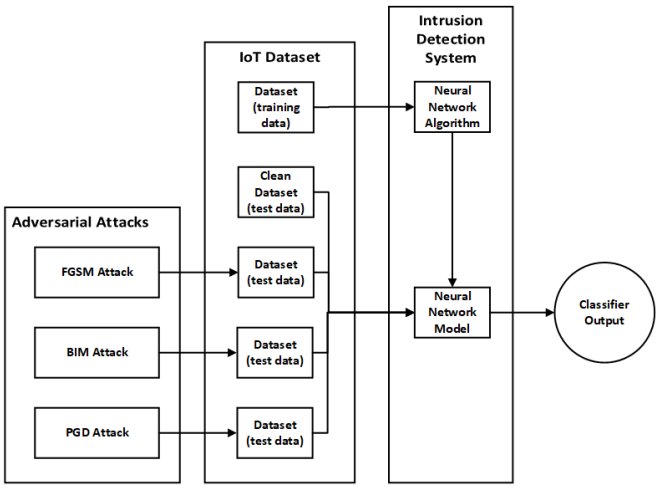

Fig. 1: Solution Overview Architecture

A. Development platform and tools

We develop our deep learning code in python language with jupyter notebook hosted in Google Colaboratory. Colaboratory is an interactive environment provided by Google for writing and executing code in python and other languages [14]. Colaboratory offers advanced GPU features, is hosted in the cloud, and comes with several pre-installed deep learning frameworks and libraries that accelerate the task of building machine learning models.

B. Dataset

For our dataset, we use the BoT-IoT dataset [8] provided from the Cyber Range Lab of The center of UNSW Canberra Cyber. This dataset provides a realistic representation of an IoT network since it was created in a dedicated IoT environment, and contains adequate number of records with heterogeneous network profiles.

The BoT IoT dataset consists over **72 million records** of network activity in a simulated IoT environment. A scaled-down version of the dataset comprising of approximately **3.6 million records** is also available and was used for our study. A selection of the 10 best features have been provided in the original dataset and were also used for this study. [8].

The training and test dataset have 5 output classes each which reflect the normal traffic and the 4 types of attacks which were carried out against the IoT network.

TABLE I: BoT-IoT Dataset Features

Feature	Description
pkSeqID	Row Identifier
Stime	Record start time
Seq	Argus sequence number
Mean	Average duration of aggregated records
Stddev	Standard deviation of aggregated records
Min	Minimum duration of aggregated records
Max	Maximum duration of aggregated records
Srate	Source-to-destination packets per second
Drate	Destination-to-source packets per second
N_IN_Conn_P_SrcIP	Total Number of packets per source IP
N_IN_Conn_P_DstIP	Total Number of packets per Destination IP

TABLE II: BoT-IoT Dataset Target Classes

Target Label	Training Samples	Test Samples
DDoS	1541315	385309
DoS	13201485	330112
Reconnaisance	72919	18163
Normal	370	107
Theft	65	14

C. Building the FNN and SNN deep learning based IDS

We implement two IDSes for our IoT dataset. The first IDS is implemented using a Feed Forward Artificial Neural Network (FNN) as shown in Fig. 2 while the second IDS is implemented using a Self-normalizing Neural Network (SNN) as shown in Fig. 3. In each Neural Network model design, we create 3 hidden layers and 16 neurons for each layer, giving us a total of 48 neurons in the hidden layers.

The intuition behind SNN is to keep the mean and the variance as close to 0 and 1 respectively throughout each layer of the neural network. As shown in Algorithm 1, for the SNN, we use a Scaled Exponential Linear Unit (SeLU) activation function while for the FNN we use a Rectifier Linear Unit(ReLU). The FNN uses a basic dropout layer to prevent overfitting and ensure better stability in the network during the learning phases while the SNN uses an AlphaDropout layer to retain the mean and variance at 0 and 1 respectively.

For initializing the weights, we select Glorot Uniform initializer [15] for the FNN while we use a Lecun Uniform Initializer [16] for the SNN.

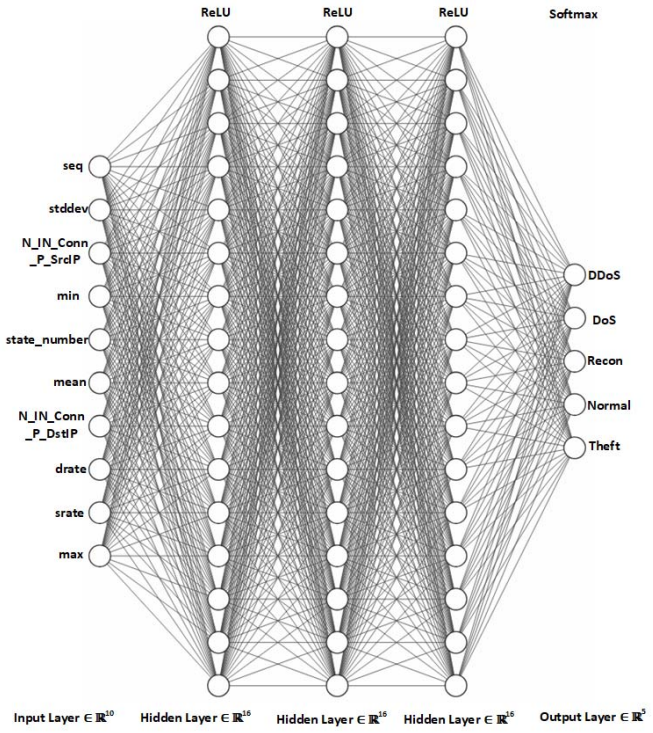

Fig. 2: FNN Architecture

D. Generating The Adversarial Samples

We generate our adversarial samples using the Adversarial Robustness Toolbox (ART) [17] framework which is provided by IBM and is made available for public use.

The first method we use in generating the adversarial examples for the IoT dataset is the Fast Gradient Sign Method (FGSM). This method performs a one step gradient update along the direction of the sign of gradient for every input in the dataset. [11]. The second method is the Basic Iteration method (BIM) which runs a finer optimization of the FGSM with minimal smaller changes for multiple iterations [12]. In each iteration, the each feature of the input values is clipped to avoid too large a change on each feature. The third method is the Projected Gradient Descent (PGD) which

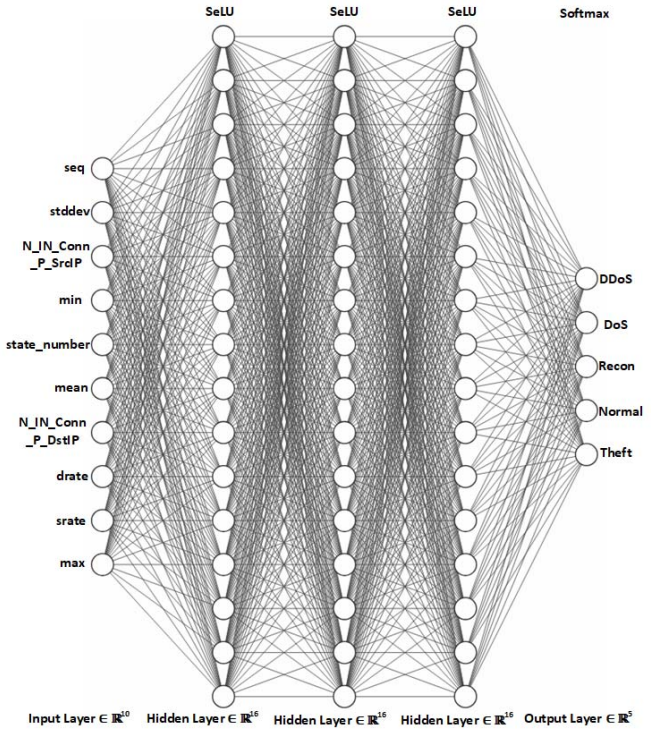

Fig. 3: SNN Architecture

is also a variation of the FGSM attack but omits the random start feature of the FGSM [13]. All three methods are model dependent methods[18] and rely on the model gradient.

V. Results & Evaluation

Our first result in subsection (A) below illustrates the impact of adversarial samples on a deep learning based IDS implemented using a FNN for the IoT dataset used in this paper. In our second result in subsection (B), we provide a performance comparison between the SNN and the FNN IDSes. In our third result in subsection (C), we compare the adversarial resilience of both the FNN and the SNN IDSes. Our final evaluation in subsection (D) shows the effect of feature normalization on deep learning based IDS using the IoT dataset.

A. Effect of Adversarial Samples on Deep learning based IDS in IoT networks

In our first experiment, we demonstrate that the FNN deep learning based IDS was significantly degraded by the adversarial samples generated from the IoT dataset. After training the IDS model, we achieve an initial accuracy of 95.1%. We then evaluate the performance of the IDS once again using the three adversarial sample datasets that were created in the previous section. The prediction accuracy of the FNN IDS is reduced from 95.1% to 24% from the FGSM adversarial samples. We repeat the experiment with the BIM and PGD adversarial samples and achieve accuracies of 18% and 31% respectively as shown in Fig. 4.

Algorithm 1: Adversarial Testing for FNN and SNN

```
for Each neural network FNN, SNN do
    initialize number of hidden layers L, weights w;
    Add input layer, activation layer, dropout layer
    for i in L - 1 do
        Add Dense layer
        if ANN model then
            Add ReLU activation layer;
            Add Dropout Layer;
        else
            Add SeLU activation Layer;
            Add AlphaDropout Layer;
        end
    end
    Add output layer, softmax activation layer, dropout
     layer
    while no of epochs not complete do
        compute training and validation loss;
    end
    Evaluate model performance
    for each adversarial attack method do
        Craft adversarial samples x'
        Evaluate model performance with adversarial
         samples x'
    end
end
```

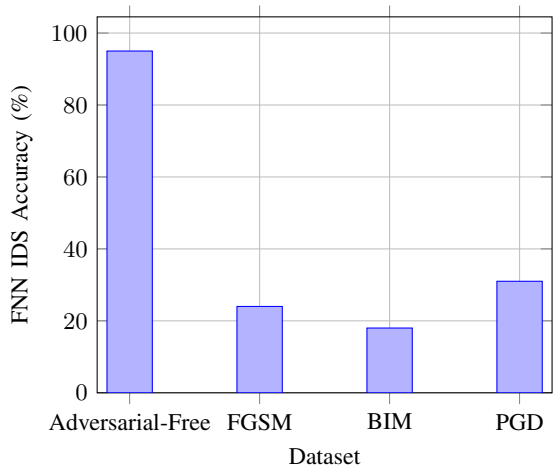

Fig. 4: Effect of Adversarial Samples on FNN IDS

B. Performance Comparison of FNN and SNN IDS using the adversarial-free IoT dataset

In our second experiment, we compare both the FNN and SNN IDSes. The SNN IDS underperforms the FNN IDS based on several performance metrics as shown in Fig. 5. For classification metrics namely precision, recall and F1-score, the FNN IDS consistently outperforms the SNN IDS over multiple experiment runs. For additional multiclassification metrics such Copen Cappa Score and MC Coefficient, the FNN IDS outperforms the SNN IDS as shown in Fig. 5.

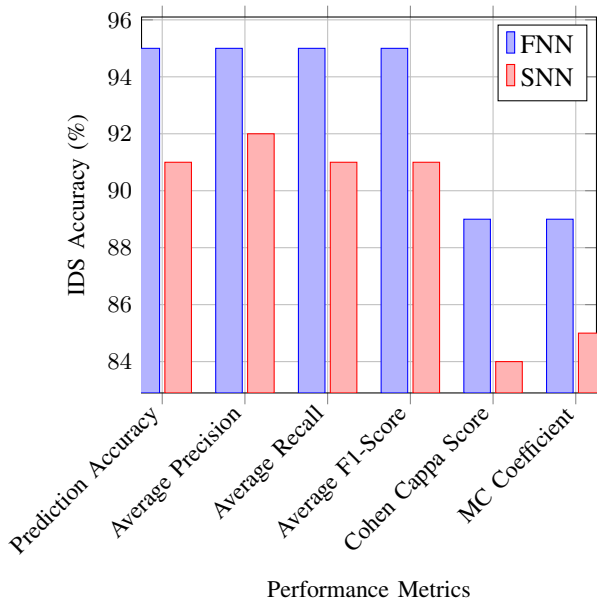

Fig. 5: Comparison of FNN and SNN IDSes

C. Comparison of Adversarial Resilience of FNN IDS and SNN IDS

Both the FNN IDS and SNN IDS performance on the IoT dataset were degraded by the adversarial samples which we created. We however observe that the SNN IDS is more resilient to the adversarial attacks than the FNN IDS as shown in Fig. 6.

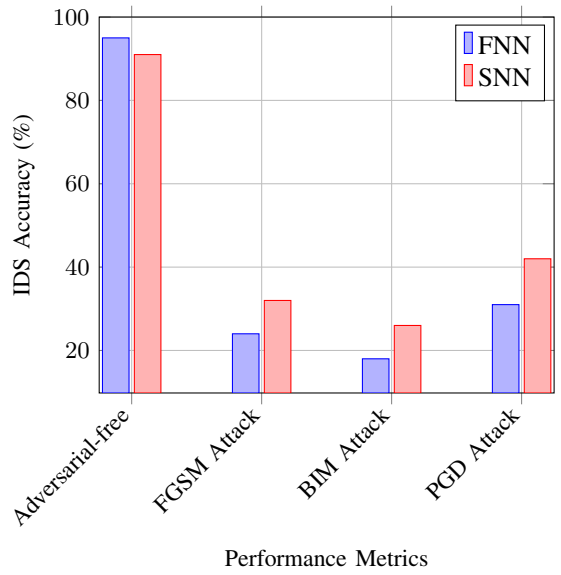

Fig. 6: Adversarial Resilience of FNN and SNN IDS models

D. Effect of Feature Normalization on a Deep Learning based IDS for IoT

In our final experiment, we refrain from carrying out feature normalization on the IoT dataset. As shown in Fig. 7 & 8, both IDSes have a significantly lower prediction accuracy on the adversarial-free dataset when the input features are not normalized. However, their resilience to adversarial samples is improved.

Fig. 9 compares the effect of feature normalization on various classification metrics for both FNN and SNN based IDSes using the adversarial-free dataset. The results indicate that both IDSes yield more accurate results on the adversarial-free dataset when the input features are normalized.

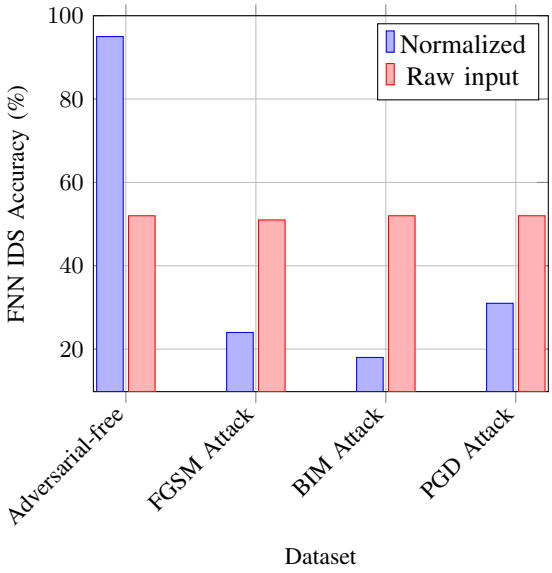

Fig. 7: Effect of Feature Normalization on Deep learning based IDS using FNN

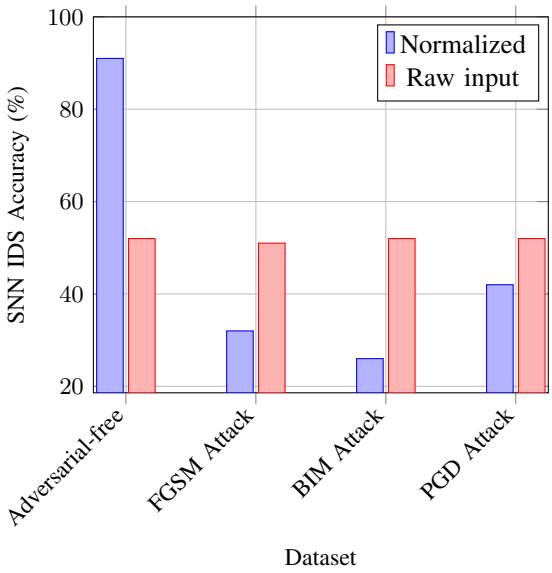

Fig. 8: Effect of Feature Normalization on Deep learning based IDS using SNN

VI. CONCLUSION

We created two deep learning based IDSes for an IoT dataset using two types of neural network models - a FNN

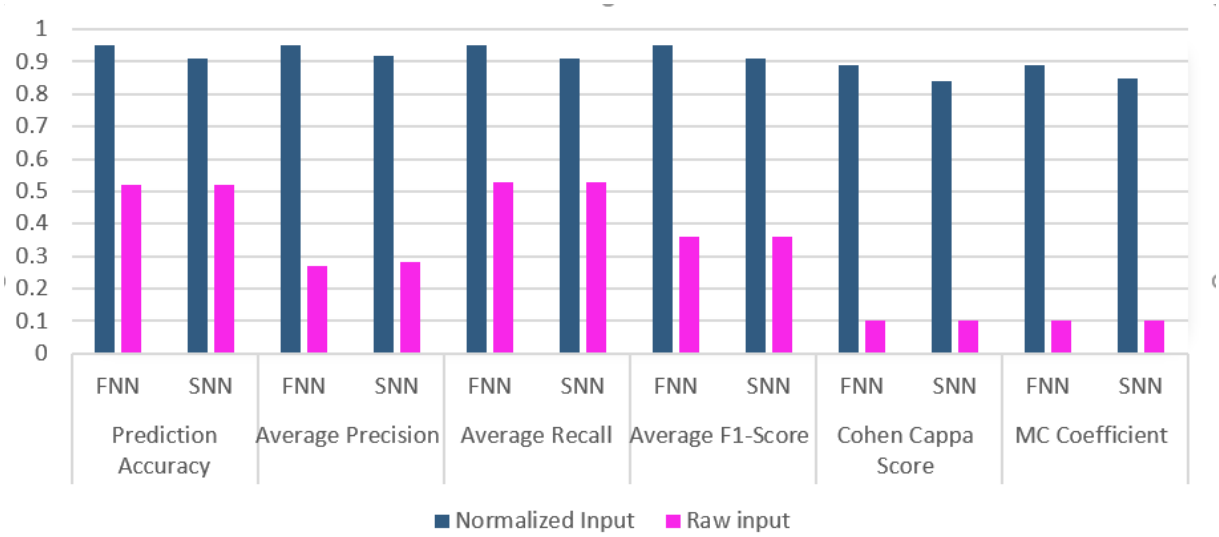

Fig. 9: Effect of Feature Normalization in Deep Learning Based IDS in IoT Networks

and a SNN - and observed that both models were impacted differently by the adversarial samples. Our experiments with the IoT dataset show that the self-normalizing feature of the SNN makes it more resilient to gradient based adversarial samples.

Our results further show that feature normalization of the IoT dataset negatively affects the adversarial resilience of the deep learning based IDSes. When the input features are normalized, both IDSes have better performance metrics, but they are more vulnerable to adversarial samples.

From our experiments, the SNN IDS on average had a 9% higher performance accuracy than the FNN IDS when subjected to adversarial samples. However, the SNN IDS performance accuracy under adversarial attacks was still below 50% and could not be regarded as a suitable defence against adversarial attacks on deep learning based IDSes in real life applications.

Our future research work would seek to improve the performance of the deep learning-based IDSes under adversarial attacks. We would also like to investigate why the self-normalizing properties of the SNN makes the SNN IDS for the IoT dataset more resilient to adversarial samples.

ACKNOWLEDGEMENT

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

REFERENCES

[1] R. Roman, P. Najera, and J. Lopez, "Securing the internet of things," *IEEE Computer, Vol 44*, no. 9, pp. 51–58, 2011.

[2] A. Javaid, Q. Niyaz, W. Sun, and M. Alam, "A deep learning approach for network intrusion detection system," in *Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (formerly BIONETICS)*, pp. 21–26, ICST, 2016.

[3] W. Guo, D. Mu, J. Xu, P. Su, G. Wang, and X. Xing, "Lemna: Explaining deep learning based security applications," in *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pp. 364–379, ACM, 2018.

[4] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. J. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *ICLR*, vol. abs/1312.6199, 2014.

[5] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, pp. 372–387, IEEE, 2016.

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

[7] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," in *Advances in neural information processing systems*, pp. 971–980, 2017.

[8] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the development of realistic botnet dataset in the internet of things for network forensic analytics: Bot-iot dataset," *arXiv preprint arXiv:1811.00701*, 2018.

[9] E. Hodo, X. Bellekens, A. Hamilton, P.-L. Dubouilh, E. Iorkyase, C. Tachtatzis, and R. Atkinson, "Threat analysis of iot networks using artificial neural network intrusion detection system," in *2016 International Symposium on Networks, Computers and Communications (ISNCC)*, pp. 1–6, IEEE, 2016.

[10] A. Warzyński and G. Kołaczek, "Intrusion detection systems vulnerability on adversarial examples," in *2018 Innovations in Intelligent Systems and Applications (INISTA)*, pp. 1–4, IEEE, 2018.

[11] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *CoRR*, vol. abs/1412.6572, 2015.

[12] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," *ICLR*, 2017b.

[13] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations*, 2018.

[14] Google, "Colaboratory: Frequently asked questions," 2018. [Online; accessed 29-Mar-2019].

[15] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256, 2010.

[16] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, "Efficient backprop," in *Neural networks: Tricks of the trade*, pp. 9–48, Springer, 2012.

[17] M.-I. Nicolae, M. Sinn, M. N. Tran, A. Rawat, M. Wistuba, V. Zantedeschi, I. M. Molloy, and B. Edwards, "Adversarial robustness toolbox v0. 2.2," *arXiv preprint arXiv:1807.01069*, 2018.

[18] Z. Gong, W. Wang, and W.-S. Ku, "Adversarial and clean data are not twins," *arXiv preprint arXiv:1704.04960*, 2017.