International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys.13, No.1 (May-23)

http://dx.doi.org/10.12785/ijcds/1301105

Defenses for Adversarial attacks in Network Intrusion Detection System – A Survey

N. Dhinakaran[1] and S. Anto[2]

[1]*Research Scholar, School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India*
[2]*Associate Professor, Department of Computational Intelligence, School of Computer Science and Engineering, Vellore Institute of Technology, Vellore, India*

Received 27 Jun. 2022, Revised 30 Apr. 2023, Accepted 14 May. 2023, Published 30 May. 2023

Abstract: In computer security, machine learning has a greater impact in recent years. Ranging from spam filtering, malware analysis, and traffic analysis to network security the usage of machine learning algorithms are manifold. In the area of network security, machine learning techniques are used especially in developing intrusion detection systems. There are basically two kinds of intrusion detection systems - host intrusion detection systems and network intrusion detection systems. Even though machine learning techniques have greatly improved the efficiency of the intrusion detection systems, they are vulnerable to adversarial attacks which are designed and launched by adaptive adversaries who know the working principles of machine learning models. In recent years adversarial machine learning has gained attention in the domain of machine learning in which attackers exploit the inherent fallacies in the assumptions made in the machine learning models. In the domain of network security especially in intrusion detection systems, the significant role of adversarial machine learning has not been addressed in detail. This survey examines different types of defenses deployed to mitigate the impact of adversarial attacks. Their effectiveness in dealing with attacks is analysed and their limitations are discussed.

Keywords: Adversarial Machine Learning, Intrusion Detection System, Machine Learning, Adversarial Samples, Network Security

1. INTRODUCTION

An Intrusion Detection System (IDS) is a surveillance system that looks for malicious activity and also generates alert messages when it is detected. Different kinds of intrusion detection systems are used to secure the applications and information. They are, host intrusion detection systems, network intrusion detection systems, signature based, anomaly based and hybrid. Host intrusion detection system is installed in the personal computer and observes the inflow and outflow of traffic and looks for any abnormal activity through analysis. Network intrusion detection system is a system designed to operate at the network level. It is designed and installed such a way to monitor network traffic from all network and personal computer systems. It also monitors the traffic which goes into the devices in the network. In signature based intrusion detection system, the system has a database of the signature of all the known attacks and this signature is used to detect attacks. But it cannot detect attacks which are not in the signature database. A benchmark for normal network behaviour is analysed and set in anomaly based intrusion detection systems. If any network behaviour is not in line with this normal benchmark then that behaviour is deemed to be malicious. To notify this attack an alarm will be raised by the intrusion detection system. The blend of signature and anomaly based intrusion detection system are known as hybrid intrusion detection system.

Adversarial machine learning is the study of vulnerabilities exists in the machine learning and deep learning based intrusion detection models. In this, attackers exploit the inherent problems in the assumptions made in the machine learning models. The main objective of these attacks is to influence the machine learning model so that they misclassify them. That is, a malignant attack will be classified as normal, thus enabling the attacker to bypass the peripheral security measures [1]. Attackers can take advantage of the gap between the data distribution fitted by a machine learning model and the theoretical data distribution space, known as adversarial space, to fool machine learning algorithms. Adversarial attack is launched using the following steps. Consider a data point x belonging to class C. An adversarial attacker changes the x to new data point x’ by adding small perturbation so that x’ is incorrectly classified by the classifier as something other than class C.

There are two phases in a machine learning model as the training phase and the inference phase. The adversarial attacks can happen in either phase. The methods used by malicious actors to launch adversarial attacks can be

E-mail address: dhinakaran.n2020@vitstudent.ac.in, anto.s@vit.ac.in

classified into two based on the phase of the attack, namely data poisoning and model poisoning. In a data poisoning attack, the attacker uses incorrect labels of data samples to train the model, causing the model would produce wrong labels at the test time. In the model poisoning attack, the attacker generates adversarial samples from clean samples by adding imperceptible perturbations. Then this adversarial sample is used with testing algorithm to make the model generate false label.

Machine learning based intrusion detection systems suffer from following issues: large data, novel attacks, false alarm rate, unbalanced data set, response time and adversarial attacks [2]. Among the problems mentioned, adversarial attacks endanger the security of the machine learning model itself.

The machine learning models used in network security applications are under attack from adversarial samples. These adversarial samples fool the classifier to misclassify attack traffic as normal/benign traffic. It poses grave threat on the security of the classifier. There is a need to know about the vulnerabilities of machine learning based intrusion detection systems. Defense tactics for machine learning-based intrusion detection systems are still a relatively new area of study. This is the motivation for doing an in-depth survey that will provide directions for future research. This work differs from the previous survey [3], which covered limited defense strategies and did not provide a classification, the latest mitigation strategies are examined in detail.

A. *Attacks*

The adversarial samples are generated using algorithms such as Fast Gradient Sign Method (FGSM)[4], Projected Gradient Descent (PGD) [5], zeroth order optimization (ZOO) [6], Jacobian-based Saliency Map Attack (JSMA) [7], NES [8], Boundary Attack [9], Pointwise[10], HopSkipJumpAttack[11], Carlini-Wagner (CW) [12], Opt-Attack[13], Adaptive SMOTE (A-SMOTE)[14], IDSGAN uses WGAN[15] to generated adversarial samples for network intrusion detection system domain, Hydra for adversarial evasion attacks [16], Mutual Information based Adversarial Attack [17] [18], DoSWGAN[19], Elastic Net Method (ENM) [20], FlowMerge[21], Anti-Intrusiond Detection AutoEncoder (AIDAE)[22], Generative Adversarial Active Learning (Gen-AAL)[23], Polymorphic DDoS attacks using GANs[24], Brute-force Black-box Method[25], Universal Adversarial Sample Generator (U-ASG)[26], Constraint-Iteration Fast Gradient Sign Method (CIFGSM)[27], Attack-GAN[28], TANTRA[29], Selective and Iterative Gradient Sign Method (SIGSM) [30], Wasserstein Generative Adversarial Networks with Gradient Penalty (GP-WGAN) [31], Randomized Rounding Approach (rFGSMS), Deterministic Approach (dFGSMS), DeepFool [32], Multi-Step Bit Gradient Ascent (BGAS) [33], Basic Iterative Method (BIM) [34], Bit Coordinate Ascent (BCAS)[33], Momentum Iterative Fast Gradient Sign Method (MI-FGSM) [35]. These are used in network intrusion domain.

These attacks are classified into white box, black box and grey box attacks. In white box attacks the attacker knows the weights, gradients, parameters used in the model. In black box attack the attacker has no knowledge about the training process used, features used, values of the parameters and the model gradients. In grey box attacks the attacker does not have the full knowledge of the model; he only has partial knowledge about the classifier.

L-BFGS [36], One-pixel [37], Universal perturbations [38], UPSET [39], ANGRI [39], ATNs [40], Houdini [41] attacks are used exclusively in computer vision domain.

Gradient based adversarial attacks are generated by box-constrained L-BFGS, FGSM, JSMA, Carlini-Wagner (CW), DeepFool, PGD, elastic net adversarial methods. These methods use the gradients of the input features rather than the loss of the cost function to create adversarial samples. ZOO utilizes the confidence score predicted by the classifier to generate adversarial attacks.

2. DEFENSES

Adversarial attacks significantly reduce the accuracy of intrusion detection system classifiers. Improving the accuracy of classifiers in the midst of the presence of adversarial samples is the work of defense mechanisms.

The survey examines various defense mechanisms used to mitigate the effects of adversarial attacks. The scope of this study relates only to the defense mechanisms proposed exclusively for the area of network intrusion detection systems.

The defense mechanisms are divided into proactive and reactive mechanisms. A proactive defense strategy anticipates the nature of attacks and is prepared to overcome them in advance. The reactive defense mechanism addresses the problem of adversarial attacks as they arise.

A. *Proactive*

In min-max optimization [33], the authors want to know how resilient deep learning-based intrusion detection systems are to adversarial attacks generated using the Max approach to maximize the loss. They used Fast Gradient Sign Method, FGSMS with Multiple Step, Multi-Step Bit Gradient Ascent (BGA), Bit Coordinate Ascent (BCA) to create adversarial attacks. They have proposed adversarial training and Principal Component Analysis (PCA) based dimension reduction as defence mechanism using UNSW-NB 15 [70] dataset. They found that adversarial sample generation techniques such as BCA and BGA, developed for binary domains can also be used for continuous domains and BGA showed great effect in bypassing the classifiers among all the attacks. Their proposed defence technique made the Deep Neural Network (DNN) based IDS more reliable. DNN has an accuracy of 92% under normal

TABLE I. Accuracy in % of GAN based [42] approach

Model	Original	After Attack	After GAN Adversarial Training	Improvement
DNN	89.12	56.55	84.31	27.76
RF	86.12	56.63	81.31	24.68
LR	87.6	56.15	86.64	30.49
NB	69.64	56.55	82.83	26.28
DT	84.86	60.32	83.22	22.9
KNN	85.37	56.55	79.31	22.76
SVM	88.49	43.44	84.31	40.87
GB	87.6	65.38	82.97	17.59

circumstances and BCA based adversarial attack completely evaded it. DNN combined with PCA and adversarial training combined, the evasion rate falls from the range of 24.2-15.7 to the range of 5.0-2.9. The results of this investigation shows that DNN which used the dFGSM adversarial samples for adversarial training performed better under attack from Fast Gradient Sign Method, FGSMS with Multiple Step, Multi-Step Bit Gradient Ascent (BGA) adversarial samples with the exception of BCA samples. Here, the methodology used is Adversarial Training along with PCA, which entails retraining the classifier. The proposed defence mechanism can only defend against known attacks and fails against unknown adversarial attacks. This paper considered only black-box and white-box Attacks

In Generative Adversarial Network (GAN) based [42] technique the authors wished to find the effect of adversarial attacks generated using GANs and proposed a defence technique using GAN to protect against adversarial perturbations. The following models DNN, RF, Logistic Regression(LR), Naïve Bayes (NB), DT, KNN, SVM, and gradient boosting (GB) are tested for their robustness. Adversarial attacks which can evade detection by machine learning based IDS are crafted using GAN. These attacks were created by only modifying the content features of the data instances thus preserving the functional behaviour of the network traffic. Here probe attacks were passed on as normal traffic by fooling the classifier. The general defence methodology used here is adversarial training. Here the adversarial training based defence technique is used to increase the resilience of the classifiers this needs the retraining of the classifiers with augmented training data. Models trained with GAN based adversarial samples show better resilience when attacked with adversarial samples. On an average there is a 26% improvement in accuracy with SVM achieving the highest improvement. Work needs to be done to evaluate the performance for other types of attacks like r2l, u2r, dos. The attack method used here is black box attack. KDD99 [71] dataset is used. In Table **??** the effectiveness of the proposed mechanism is reported.

In robust self-protection [43], only network flows corresponding to legitimate traffic and Denial of Service (DoS) / Distributed Denial of Service (DDoS) attacks from CIC-IDS2017 dataset are used in this assessment. A multilayer perceptron (MLP) is used as the classifier. Here the authors have designed and provided a protection mechanism against DDoS attacks in Software-Defined Networking (SDN) environment. They used FGSM attack technique to generate adversarial flows of DDoS attacks such as Hulk, Slowloris to bypass the MLP classifier and trick into thinking it was normal traffic flow. FGSM is successful in fooling the classifier. As a defence mechanism, adversarial retraining of the model is proposed and implemented. The FGSM generated adversarial samples of DDoS attack samples were added with original training data and the model is retrained with this combined data. This newly trained model is able to detect the adversarial attacks with improved accuracy and without much delay. This improvement in performance is shown by less system load and quick server response time.

In reconstruction from partial observation (RePO) [44], the dataset used is CIC-IDS2017. The model used here is denoising Auto Encoder with masks to block out certain input features during training and testing. This is tested in Software Defined Network (SDN) environment. The authors used Reconstruction from Partial Observation (RePO) with denoising Auto Encoder to improve the detection performance of network intrusion detection system (NIDS) by up to 29% in normal scenarios and by upto 45% in an adversarial scenario. It is a white box attack scenario. Without using mitigation NIDS detection rate dropped by 70% for packet-based IDS and 68% for flow-based IDS. The adversarial attacks are generated according to the principles set out in [45]

In ensemble adversarial training [46] CSE-CICIDS2018 dataset is used. From this dataset 250,000 anomalous traffic flows are randomly selected to convert them into adversarial samples. For this purpose the following white box attack approaches are used Fast Gradient Sign Method [4], Iterative Attack (I-FGSM) [34] and Momentum Iterative Fast Gradient Sign Method (MI-FGSM) [72]. The extended dataset is used to rebuild the models such as Multilayer Perceptron (MLP) [73], Convolutional Neural Network (CNN) [74], and CNN with Long Short-Term Memory (LSTM) layers, i.e., C-LSTM [75] and the result is examined. This approach increased the models' resilience, and the results showed that these models were able to identify hard-to-detect anomalous malicious traffic.

TABLE II. Defense mechanisms and their performance

No	Method	Type of Defense	ML/DL Model	Attacks	Dataset	Network	Performance
1	Min-max optimization [33] (proactive)	Adversarial Training	DNN	PGD, FGSM, BGAS, BCAS	UNSW-NB 15	Traditional-Wired	evasion rate falls from the range of 24.2-15.7 to the range of 5.0-2.9
2	GAN based [42](proactive)	Adversarial Training	DNN, LR, SVM, KNN, NB,RF,DT, and GB	GAN based	KDD99	Traditional - Wired	Accuracy improved on an average of 26%.SVM had the highest improvement
3	Robust self-protection [43](proactive)	Adversarial Training	MLP	FGSM	CICIDS2017	Software Defined Network (SDN)	adversarial Hulk attack is stopped in 10s
4	Reconstruction from Partial Observation (RePO) [44](proactive)	Adversarial Training	denoising autoencoders	Hashemi [45]	CICIDS2017	SDN	49% improvement in detection performance
5	Ensemble Adversarial Training [46](proactive)	Adversarial Training	MLP, CNN, C-LSTM, Ensembling	FGSM, Iterative-FGSM and Momentum Iterative-FGSM.	CSE-CIC-IDS2018	Traditional - Wired	Attack Success Ratio drops to 6.7% for multi-class classification and 5.78% for binary classification
6	Min-Max Formulation [47](proactive)	Adversarial Training	ANN,CNN, and RNN	FGSM, BIM, PGD, Carlini and Wagner (CW) and Deepfool.	UNSW-NB 15 and NSL-KDD	Traditional - Wired	Improved detection rate of the models under investigation
7	DEF-IDS [48](proactive)	Adversarial Training	DNN	Multiclass GAN, FGSM, DeepFool, JSMA, and BIM	CSE-CIC-IDS2018	Traditional - Wired	Showed improved performance with 97.9% acccuracy
8	Adversarial Deep Learning [49] (proactive)	Adversarial Training	GAN	GAN based	CICDDoS 2019	SDN	GAN performed better with accuracy of 94.38% than CNN, MLP and LSTM
9	GCNN and Data Augmentation [50] (proactive)	Adversarial Training	GCNN	FGSM	Australian Defence Force Academy Linux Dataset (ADFA-LD)	Traditional - Wired	24.3% improvement in accuracy

TABLE II. Defense mechanisms and their performance (continued)

No	Method	Type of Defense	ML/DL Model	Attacks	Dataset	Network	Performance
10	Deep Adversarial Learning [51] (proactive)	Adversarial Training	LR, SVM and DNN	The Poisson-Gamma Joint Probabilistic Model method, and Deep Generative Neural Networks	KDD Cup 99	Traditional - Wired	Better F1 score
11	RNN-ADV [52] (proactive)	Adversarial Training	RNN	JSMA	NSL KDD	Traditional - Wired	13.44% improvement in F1 score
12	Robust DNN [53] (proactive)	Adversarial Training	DNNs	FGSM, BIM, PGD	NSL KDD	Traditional - Wired	Improved the accuracy of the DNN
13	Robust Random Forest [54] (proactive)	Adversarial Training	Random Forest	JSMA	power system dataset	Industrial Control Systems	21% improvement in F1 score
14	Adversarially Trained RF [55] (proactive)	Adversarial Training	Random Forest	FGSM, JSMA, Deepfool and CW	NSL-KDD and CICIDS2017	Traditional - Wired	Better performance with 99% accuracy
15	Hardened ML Classifier [56] (proactive)	Adversarial Training	J48 Decision Tree, RF, Bayesian Network, and SVM	Rule Based Attack Sample Generations	IoT smart home dataset	IoT	Performance improvement in F1 score
16	Deep neural network-based detection model [57] (proactive)	Adversarial Training	DNN	FGSM	KDDCUP99	Traditional - Wired	F1 value improved to 0.996858
17	Hardening Cyber Detectors [58] (proactive)	Defensive Distillation	Random Forest	Features like number of outgoing(Src) or incoming(Dst) bytes, duration of the flows, total number of transmitted packets are modified as group	CTU-13	Traditional - Wired	46% improvement in recall
18	Recursive Feature Elimination Based (RFE) [59] (proactive)	Feature Removal based	ML binary classifier	FGSM	CICIDS2017	Traditional - Wired	Improves robustness

TABLE II. Defense mechanisms and their performance (continued)

No	Method	Type of Defense	ML/DL Model	Attacks	Dataset	Network	Performance
19	Model Voting Ensembling [46] (proactive)	Ensemble Learning	MLP, CNN and C-LSTM	NES, Boundary, HopSkipJumpAttack, Pointwise, and Opt-Attack	CSE-CIC-IDS2018	Traditional - Wired	Attack Success Ratio is close to 0%
20	AppCon [60] (proactive)	Ensemble Learning	RF, MLP, DT, AB, “Wide and Deep” (WnD)	small perturbations in the values of flow-based features combinations	CTU-13	Traditional - Wired	Blocks 75% of evasion attacks
21	Detect and Reject [61] (proactive)	Defence against Transferability	Random Forest	FGSM, PGD	NSLKDD	Traditional - Wired	Improved the performance with better accuracy
22	Adversarial Query Detection [46] (reactive)	Adversarial Query Detection	deep similarity encoder	NES, Boundary, HopSkipJump Attack, Pointwise, and Opt-Attack	CSE-CIC-IDS2018	Traditional - Wired	Significant reduction in Attack Success Rate
23	RNN-ADV [62] (proactive)	Finding Optimal Weights	Random Neural Network with artificial bee colony (ABC) algorithm	JSMA	NSL-KDD	Traditional - Wired	17% improvement F1 score when compare with DNN
24	Neural Activation Based [63] (proactive)	Neural Activation	ANN, Adaboost, RF, SVM, KNN	FGSM, BIM, CW and PGD	CICIDS2017	Traditional - Wired	RF and KNN perform better with a recall value of 0.99
25	Adversarial Sample Detector [64] (reactive)	Adversarial Sample Detection	Bidirectional Generative Adversarial Network and DNN	FGSM, PGD, MI-FGSM	NSL-KDD	Traditional - Wired	26.46% improvement in accuracy
26	FGMD [65] (proactive)	Feature Grouping	LSTM	Modifying values of feature and their related features	IoTID[66], MedBIoT [67]	IoT	Better performance with 98% accuracy
27	Data Transformation [68] (proactive)	Defense against Data Poisoning	RF, MLP, KNN	Altering values of the features duration, exchanged_bytes, total_packets	Netflow based data	Traditional - Wired	55% percent reduction in Attack Severity in RF
28	MANDA [69] (reactive)	Manifold based	ML	FGSM, BIM, CW	NSL KDD	Traditional - Wired	98.41% true-positive rate (TPR) when under CW attack

In min-max formulation [47] the authors have crafted adversarial examples using FGSM, DeepFool, PGD, Basic Iterative Method (BIM) and CW by inner maximization. These adversarial attacks have significantly reduced the accuracy of Artificial Neural Networks (ANN), Convolution Neural Network (CNN), and Recurrent Neural Network (RNN) models. These adversarial samples are then added with training data to create an augmented dataset. This expanded dataset is then used in the evaluation of ANN, CNN, and RNN for the NSL-KDD [76] dataset and UNSW-NB 15. This adversarial retraining approach based on min-max formulation, improved the performance of the models for adversarial samples.

In DEF-IDS [48] CSE-CIC-IDS2018 dataset is used. This defense mechanism consists of two parts. In the first part, Multiclass GAN is used to generate samples which mimic the original data. The Multiclass GAN in turn is a combination of Auxiliary Classifier GAN (AC-GAN) [77] and Semi-Supervised GAN (SGAN) [78]. The second part is multi-source adversarial retraining which combines the adversarial examples crafted using FGSM, DeepFool, JSMA, and BIM with the original dataset. This retrained DNN on ensemble of adversarial training mechanism showed superior performance with an accuracy of 0.979.

In adversarial deep learning [49] method, the authors have utilized GAN for DDoS identification in an SDN scenario. They have employed adversarial training approach as defense mechanism. In order to add adversarial data to the original dataset, GAN is used as an adversarial sample generator. Adversarial-trained GAN performed better than MLP, CNN, LSTM classifiers for CICDDoS 2019 [79] dataset with an accuracy of 94.38

In Gated Convolutional Neural Network (GCNN) and data augmentation [50] method the authors first tested the robustness of the GCNN against adversarial perturbations performed using the FGSM method. The GCNN is trained using the Australian Defence Force Academy Linux Dataset (ADFA-LD) [80]. The accuracy of GCNN for clean samples is 97.5%. This dropped to 35.4% under FGSM attack. Then the authors used an adversarial training method to increase the security of GCNN. In this defence method, the adversarial samples generated by FGSM are added with original clean samples and then the GCNN is trained using this augmented dataset. This process increased the accuracy of the GCNN from 35.4% to 60.7%. This defence procedure suffers from disadvantages such as the need to retrain the classifier with adversarial samples, since only FGSM samples are used for the adversarial training the model becomes ineffective against other adversarial samples other FGSM, need to evaluate the model in terms of the training time and testing time taken.

In deep adversarial learning [51] the authors proposed the data augmentation method to primarily address the data shortage and class distribution imbalances in the NIDS dataset such as KDD Cup 99. To address this problem, intrusion data is generated using the poisson-gamma joint probabilistic method and Deep Generative Neural Networks. The extended dataset is used for the training the IDSes which are based on Logistic Regression, Support vector machines (SVMs) and DNN. This method improved the detection of low frequency attacks such as R2L. This method requires retraining the classifier with newly generated data set. The model becomes ineffective against adversarial samples other than those for which it was trained for. The authors did not explicitly assess the impact of adversarial samples on the accuracy of the model.

In Random Neural Network-ADV [52] method the authors have used adversarial retraining approach to improve the robustness of random neural network. NSL KDD dataset is used for the experiments. The clean dataset is used to train the random neural network model initially and the F1 Score of 96.61 is obtained for DoS traffic. But the same F1 score drops to 24.15 when the model is attacked with adversarial samples. The adversarial samples are generated using JSMA technique. Then training dataset is improved with the adversarial samples of JSMA. Random neural network is trained using the improved dataset to learn the attack patterns of JSMA. Random Neural Network-ADV improves the F1 score from 24.15 to 37.59 for DoS traffic when tested with adversarial data. The solution outlined in this paper request the need for retraining of the classifier with adversarial samples. Since only JSMA attack samples are used for Adversarial training, the model will become ineffective against adversarial attacks other than JSMA. The authors have not evaluated the training time and testing time taken.

In Robust DNN [53] the authors study the performance of deep learning based IDS under adversarial machine learning scenario. They studied the performance adversarial training of deep learning model as a solution to problem of adversarial machine learning in IDS. Under normal circumstances the DNN based IDS model performs with 99.61% accuracy. The adversarial samples were generated using FGSM, BIM and PGD and they reduce the accuracy of the classifier to 14.13%, 8.85% and 8.85% respectively. Though the newly trained model showed improved performance in detecting adversarial samples, its accuracy dropped when presented with stronger attacks. NSL KDD dataset is used for these above experiments. The problems found in this paper are the need for retraining of the classifier with adversarial samples, the problem of only PGD samples are being used for Adversarial training. The model will become ineffective against adversarial attacks other than PGD, found slight decrease in detector accuracy on unattacked network traffic and did not evaluate the training time and testing time taken.

In robust random forest [54] the authors tested the performance of IDS under adversarial environment in Industrial

Control Systems and proposed solution for the problem. They used power system data set for this testing. Under clean sample the Random Forest based IDS has shown to be performing well with an F1 score of .62. JSMA technique is used to generate adversarial samples which have successfully fooled the classifier and the F1 score reduced to 0.55 from 0.62. To improve the performance adversarial training approach is used. In this solution JSMA generated samples are labelled with original labels and supplemented with original clean samples. This improved data set is used for training the model. The newly trained model learned the attack pattern well. After this proposed defence mechanism the model performed well in identifying adversarial samples and the F1 score improved from 0.55 to 0.76. This has few problems and they are need for retraining of the classifier with adversarial samples, only JSMA attack samples are used for Adversarial training, the model will become ineffective against adversarial attacks other than JSMA and did not evaluate the training time and testing time taken.

In adversarially trained Random Forest [55] method adversarial retraining approach is used to improve the robustness of classifier based on Random Forest. Training dataset is improved with the adversarial samples of FGSM, Deep Fool, CW. NSL KDD dataset is used in this experiment. Random Forest is trained using the improved dataset to learn the attack patterns of CW and Area Under the Curve (AUC) of the RF has improved from 0.62 to 0.9944. The CW attack lowers the AUC of the RF classifier to .62 from 0.9884. The following are the challenges in the proposed method, need for retraining of the classifier with adversarial samples, the model will become ineffective against adversarial attacks other than what it is trained on, and did not evaluate the training time and testing time taken.

In hardened machine learning classifier [56] method the authors used adversarial retraining approach is for hardening of classifiers to defend against adversarial attacks. Training dataset is enhanced with the adversarial examples. Home Internet of things (IoT) network dataset is used for experiments. J48 Decision Tree, Random Forest, Bayesian Network, and SVM classifiers are trained using the enhanced dataset to learn the adversarial attack patterns. Rule based attack sample generations technique is used to generate adversarial samples. Adversarial sample reduces F1 score from 99.9 to 79.8 for random forest. Proposed method increases the F1 score of random forest classifier from 79.8 to 99.9. The following are the issues in the proposed method: need for retraining of the classifier with adversarial samples, the model will become ineffective against adversarial attacks other than what it is already trained with, and did not evaluate the training time and testing time.

In deep neural network-based detection model [57] the authors proposed the adversarial retraining approach to improve the resilience of classifiers to adversarial samples. Training dataset is enhanced with the adversarial samples of FGSM. KDD99 dataset is used. Deep neural network-based detection model is trained using the enhanced dataset to learn the FGSM adversarial samples. FGSM technique reduces the F1 value of a model from 0.997379 to 0.176636. Proposed method increases F1 value to 0.996858. This approach has following issues need for retraining of the classifier with adversarial samples, only FGSM samples are used for Adversarial training, the model will become ineffective against adversarial attacks other than FGSM, and did not evaluate the training time and testing time taken.

In hardening cyber detectors [58] the authors proposed a model to make Random Forest more resilient to adversarial perturbations by training the Random Forest using probability labels instead of hard class labels . CTU-13 [81] dataset is used in this experiment. As a first step, probability labels from hard class labels are generated, and then a supervised model trained with the generated probability labels to perform the cyber detection is deployed. The adversarial sample reduces recall from 0.9684 to 0.2573. In the proposed method recall has improved from 0.2573 to 0.5152. The problems of this method are training time of the proposed method has increased from 32.3 s to 87.0s, need for training a new model, increase in false positives rates of the baseline performance of the "hardened" classifier, and distillation becomes ineffective against adversarial samples of CW technique.

In recursive feature elimination based (RFE) [59], first the largest absolute difference under FGSM method for attack is calculated and then those features are eliminated. Principle Component Analysis (PCA), t-Distributed Stochastic Neighbourhood Embedding (t-SNE), Unified Manifold and Projection (UMAP), and parallel co-ordinate plots are used to analyze and visualize the CICIDS2017 dataset. Recursive Feature Elimination Based (RFE) [82] is used to remove the features from the dataset. This feature removal strategy enabled the model to be robust against adversarial attacks.

In model voting ensembling [46] Multilayer Perceptron (MLP) [73], Convolutional Neural Network (CNN) [74], and CNN with Long Short-Term Memory layers (C-LSTM)[75] models are used together to decide whether a flow is normal or anomalous. In this method a flow is classified as normal if only if all the three classifier predicted it as normal, otherwise the classification result would be anomalous. This approach is able to defend against Bot, DoS attack-SlowHTTPTest, DoS attack-Hulk, DoS attack-GoldenEye, DoS attack-Slowloris, FTP-BruteForce, SSHBruteforce, DDoS attack-LOIC-UDP and DDoS attack-LOICHTTP adversarial samples with Adversarial Success Rate close to 0

In AppCon [60] Application Constraints (AppCon) method is proposed by the authors to counter effects of the adversarially crafted samples. This is a model agnostic approach. Random Forest (RF), Multi-Layer Perceptron (MLP), Decision Tree (DT), AdaBoost (AB), and "Wide and Deep" (WnD) are used as baseline models to compare with. CTU-13 dataset is used. This method restricts the range of attacks that can be launched by the attackers. Then there are ensembles of detectors and each detector takes the application traffic flow from a particular web application. The attacks are generated by changing the values in the combination of flow-based features. This method thwarts over 75% of evasion attacks without comprising performance in normal scenario.

In Detect & Reject [61] the authors first studied the impact of the transferability of the adversarial attacks on the classifiers and then the performance of ensemble IDS using SVM, DT, Logistic Regression (LR),, RF, and Linear Discriminant Analysis (LDA) using the majority voting rule. FGSM and PGD are used to construct adversarial attacks from NSLKDD data. They used the Adversarial Robustness Toolbox (ART) for the attack algorithms. The authors used DNN to create adversarial samples. These adversarial samples greatly reduced the accuracy of the DNN, SVM, LR, DT and LDA. RF showed better resilience than other models. Ensemble IDS, which comprises SVM, DT, LR, RF and LDA is fooled by the adversarial samples generated using DNN. This shows the effectiveness of the transferrable nature of the attack samples. The proposed Detect and Reject defence mechanism where in which each of the five models are retrained with the original training data and also the adversarial data labelled as "adversarial". Now the model has to predict three classes "normal", "attack" and "adversarial". After this tweak in the training phase the RF shown to be performing better with greater accuracy than DT, LR, SVM and LDA.

Random Neural Network is used in this RNN-ADV [62] approach. Ant Bee Colony optimization algorithm is used to find the optimal weights for Random Neural Network. Then this trained RNN-ADV model is tested against adversarial samples generated using JSMA technique. This approach shows better performance than DNNs for normal traffic under adversarial attack. F1 score of normal traffic using RNN-ADV is 52.6% but DNN's F1 score is 35.69% when presented JSMA attack samples. Time to train the model is longer in this approach

In neural activation based [63] method ANN model is used for training and testing the IDS based on CICIDS2017 dataset [83]. The adversarial samples were created using FGSM, BIM, CW and PGD methods. Then the neural activations of the ANN for the test data of the CICIDS2017 dataset and also the neural activations of the adversarial samples of the four attack algorithms are harvested. Then these neural activations are used to train the following models ANN, Adaboost, RandomForest, SVM, Nearest Neighbor to detect adversarial attacks. Then the trained models are tested against adversarial samples and RF and KNN showed better performance with a recall score of 0.99.

In order to protect against data poisoning attacks data transformation [68] is proposed as defense mechanism. In this dataset(X) is used to train the IDS classifier and the dataset is located in a database server. T is an invertible function with the property as shown in Equation 1.

$$T^{-1}(T(k)) = T^{-1}(k') = k, \forall k \in k \tag{1}$$

T is used to transform each data sample x into x'. This change is done by multiplying the values of flow_duration by value d, and dividing the values of its flow_exchanged_bytes by value b. Attack Severity is calculated using Equation 2.

$$Attack\ Severity = 1 - \frac{Recall\ (after\ the\ attack)}{Recall\ (before\ the\ attack)} \tag{2}$$

This method significantly decreases the effects of a poisoning attack. Attack Severity reduces from 0.7016 to 0.1587 in Random Forest. The challenges in using this method are increase in retraining time because of inverse transformation operation and it does not defend against testing time attacks. The experiments are conducted using Netflow [84] based data.

B. Reactive

In adversarial query detection [46] the similarity in successive queries to the IDS by the attacker using the black box attack is exploited to identify attacks. A deep similarity encoder is deployed for this purpose. If a particular Internet Protocol (IP) address has this suspicious flow then that can be construed as part of an active attack and the IP address can be blacklisted. This approach has significantly reduced the Adversarial Success Rate of adversarial samples.

In adversarial sample detector (ASD) [64], Bidirectional Generative Adversarial Network is used as defence mechanism to develop ASD. The generator module takes in the normal data and learns the distribution of the normal data. Then the test data, test data's reconstructed sample and discriminator feature matching error are used to calculate the reconstruction error. If the reconstruction error is high then that test instance can be construed as adversarial example. Adversarial examples are not fed to the classifier based on DNN only non adversarial examples are fed. Accuracy of the model is reduced to 0.164, 0.46, .30 from 0.76 respectively by the following attacks FGSM, PGD, MI-FGSM. The ASD method improves the accuracy of DNN when attacked with FGSM and PGD. ASD does not improve accuracy in the event MI-FGSM attacks.

MANDA [69] is a MANifold and Decision boundary-based adversarial sample detection system. NSL KDD dataset is used here for evaluation. Most often, adversarial samples are closer to the decision line to reduce magnitude of the change made to the original sample, and although they are mischaracterised into different classes, they are usually closer to their original cluster of samples. These two properties of the adversarial samples are used to develop a system to detect adversarial samples. FGSM, BIM, CW attacks are used to evaluate the system and MANDA has shown to be resilient with 98.41% true-positive rate and 5% false-positive rate.

In Table II the comparative analysis of the surveyed defense mechanisms are listed.

3. DISCUSSION

From the survey following insights are drawn. There are two broad categories of defense mechanisms proactive and reactive. In proactive defense mechanism the classifiers are trained in advance to detect the original class of the adversarial samples and they are adversarial training, defensive distillation, feature removal based, ensemble learning, defence against transferability, finding optimal weights, neural activation, feature grouping, defense against data poisoning. In reactive defense mechanism the method used is to detect and identify the adversarial samples and they are adversarial query detection, adversarial sample detection, manifold based.

Most of the defense mechanisms, 16 out of 28, are based on adversarial training. Adversarial training improves the classifier's performance in adversarial scenarios, but can only provide security against known attacks. It does not recognize attacks for which it has not been trained. It requires retraining the model with known attacks, which takes some time. Only one method is based on defensive distillation. Both adversarial training and defensive distillation mechanisms are shown to be not resilient against CW attack. Most of the defense mechanisms are proactive and only three are reactive defense mechanisms. There are only three methods [46] [64] [69] proposed in network intrusion detection domain to detect adversarial samples. Little research is done on the network domain specific adversarial attack generation methods. GAN based [42] approach and FGMD [65] are the only ones to consider domain specific constraints while generating adversarial attacks. So there is a need to develop stronger defenses against adversarial attacks in network intrusion detection system domain.

4. CONCLUSIONS AND FUTURE WORK

This survey examined in detail various defense strategies to protect the intrusion detection classifiers. This work has categorized these strategies according to the underlying methodologies used to harden the intrusion detection classifiers. The adversarial attacks have reduced the accuracy of classifiers to 8-43%. The mitigation mechanisms have increased the accuracy to 79-99%. This performance improvement is only achieved for known attacks. Developing a stronger defense mechanism to detect known and unknown adversarial attacks is a promising research topic.

References

[1] C. Chio and D. Freeman, *Machine learning and security: Protecting systems with data and algorithms.* " O'Reilly Media, Inc.", 2018.

[2] M. Aljanabi, M. A. Ismail, and A. H. Ali, "Intrusion detection systems, issues, challenges, and needs," *International Journal of Computational Intelligence Systems*, vol. 14, pp. 560–571, 2021.

[3] I. Moisejevs, "Adversarial attacks and defenses in intrusion detection systems : A survey," *International Journal of Artificial Intelligence and Expert Systems (IJAE)*, vol. 8, 2019.

[4] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2015.

[5] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018.

[6] P. Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C. J. Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *AISec 2017 - Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, co-located with CCS 2017*, 2017.

[7] N. Papernot, P. Mcdaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *Proceedings - 2016 IEEE European Symposium on Security and Privacy, EURO S and P 2016*, 2016.

[8] A. Eyas, L. Engstrom, A. Athalye, and J. Lin, "Black-box adversarial attacks with limited queries and information," in *35th International Conference on Machine Learning, ICML 2018*, vol. 5, 2018.

[9] W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," in *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings*, 2018.

[10] L. Schott, J. Rauber, M. Bethge, and W. Brendel, "Towards the first adversarially robust neural network model on mnist," in *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[11] J. Chen, M. I. Jordan, and M. J. Wainwright, "Hopskipjumpattack: A query-efficient decision-based attack," in *Proceedings - IEEE Symposium on Security and Privacy*, vol. 2020-May, 2020.

[12] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proceedings - IEEE Symposium on Security and Privacy*, 2017.

[13] M. Cheng, H. Zhang, C. J. Hsieh, T. Le, P. Y. Chen, and J. Yi, "Query-efficient hard-label black-box attack: An optimization-based approach," in *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[14] P. Li, W. Zhao, Q. Liu, X. Liu, and L. Yu, "Poisoning machine learning based wireless idss via stealing learning model," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 10874 LNCS, 2018.

[15] Z. Lin, Y. Shi, and Z. Xue, "Idsgan: Generative adversarial networks forÂ attack generation against intrusion detection," in *Advances in Knowledge Discovery and Data Mining*, J. Gama, T. Li, Y. Yu,

E. Chen, Y. Zheng, and F. Teng, Eds. Cham: Springer International Publishing, 2022, pp. 79–91.

[16] J. Aiken and S. Scott-Hayward, "Investigating adversarial attacks against network intrusion detection systems in sdns," in *IEEE Conference on Network Function Virtualization and Software Defined Networks, NFV-SDN 2019 - Proceedings*, 2019.

[17] M. Usama, A. Qayyum, J. Qadir, and A. Al-Fuqaha, "Black-box adversarial machine learning attack on network traffic classification," in *2019 15th International Wireless Communications and Mobile Computing Conference, IWCMC 2019*, 2019.

[18] M. Usama, J. Qadir, A. Al-Fuqaha, and M. Hamdi, "The adversarial machine learning conundrum: Can the insecurity of ml become the achilles' heel of cognitive networks?" *IEEE Network*, vol. 34, 2020.

[19] Q. Yan, M. Wang, W. Huang, X. Luo, and F. R. Yu, "Automatically synthesizing dos attack traces using generative adversarial networks," *International Journal of Machine Learning and Cybernetics*, vol. 10, 2019.

[20] P. Y. Chen, Y. Sharma, H. Zhang, J. Yi, and C. J. Hsieh, "Ead: Elastic-net attacks to deep neural networks via adversarial examples," in *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, 2018.

[21] A. Abusnaina, A. Khormali, D. H. Nyang, M. Yuksel, and A. Mohaisen, "Examining the robustness of learning-based ddos detection in software defined networks," in *2019 IEEE Conference on Dependable and Secure Computing, DSC 2019 - Proceedings*, 2019.

[22] J. Chen, D. Wu, Y. Zhao, N. Sharma, M. Blumenstein, and S. Yu, "Fooling intrusion detection systems using adversarially autoencoder," *Digital Communications and Networks*, vol. 7, 2021.

[23] D. Shu, N. O. Leslie, C. A. Kamhoua, and C. S. Tucker, "Generative adversarial attacks against intrusion detection systems using active learning," in *WiseML 2020 - Proceedings of the 2nd ACM Workshop on Wireless Security and Machine Learning*, 2020.

[24] R. Chauhan and S. S. Heydari, "Polymorphic adversarial ddos attack on ids using gan," in *2020 International Symposium on Networks, Computers and Communications, ISNCC 2020*, 2020.

[25] S. Zhang, X. Xie, and Y. Xu, "A brute-force black-box method to attack machine learning-based systems in cybersecurity," *IEEE Access*, vol. 8, 2020.

[26] C. Yang, L. Zhou, H. Wen, and Y. Wu, "U-asg: A universal method to perform adversarial attack on autoencoder based network anomaly detection systems," in *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications Workshops, INFOCOM WKSHPS 2020*, 2020.

[27] Y. Wang, Y. Wang, E. Tong, W. Niu, and J. Liu, "A c-ifgsm based adversarial approach for deep learning based intrusion detection," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 12519 LNCS, 2020.

[28] Q. Cheng, S. Zhou, Y. Shen, D. Kong, and C. Wu, "Packet-level adversarial network traffic crafting using sequence generative adversarial networks," *CoRR*, vol. abs/2103.04794, 2021. [Online]. Available: https://arxiv.org/abs/2103.04794

[29] Y. Sharon, D. Berend, Y. Liu, A. Shabtai, and Y. Elovici, "Tantra: Timing-based adversarial network traffic reshaping attack," *IEEE Transactions on Information Forensics and Security*, vol. 17, pp. 3225–3237, 2022.

[30] Á. L. P. Gómez, L. F. Maimó, A. H. Celdrán, F. J. G. Clemente, and F. Cleary, "Crafting adversarial samples for anomaly detectors in industrial control systems," in *Procedia Computer Science*, vol. 184, 2021.

[31] C. S. Shieh, T. T. Nguyen, W. W. Lin, Y. L. Huang, M. F. Horng, T. F. Lee, and D. Miu, "Detection of adversarial ddos attacks using generative adversarial networks with dual discriminators," *Symmetry*, vol. 14, 2022.

[32] 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 Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, 2016.

[33] R. A. Khamis, M. O. Shafiq, and A. Matrawy, "Investigating resistance of deep learning-based ids against adversaries using min-max optimization," in *IEEE International Conference on Communications*, vol. 2020-June, 2020.

[34] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," in *5th International Conference on Learning Representations, ICLR 2017 - Workshop Track Proceedings*, 2019.

[35] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2018.

[36] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, 2014.

[37] J. Su, D. V. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Transactions on Evolutionary Computation*, vol. 23, 2019.

[38] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7 2017, pp. 1765–1773.

[39] S. Das and P. N. Suganthan, "Differential evolution: A survey of the state-of-the-art," *IEEE Transactions on Evolutionary Computation*, vol. 15, 2011.

[40] S. Baluja and I. Fischer, "Learning to attack: Adversarial transformation networks," in *32nd AAAI Conference on Artificial Intelligence, AAAI 2018*, 2018.

[41] M. Cisse, Y. Adi, N. Neverova, and J. Keshet, "Houdini: Fooling deep structured visual and speech recognition models with adversarial examples," in *Advances in Neural Information Processing Systems*, vol. 2017-December, 2017.

[42] M. Usama, M. Asim, S. Latif, J. Qadir, and Ala-Al-Fuqaha, "Generative adversarial networks for launching and thwarting adversarial attacks on network intrusion detection systems," in *2019 15th International Wireless Communications and Mobile Computing Conference, IWCMC 2019*, 2019.

[43] C. Benzaid, M. Boukhalfa, and T. Taleb, "Robust self-protection

against application-layer (d)dos attacks in sdn environment," in *IEEE Wireless Communications and Networking Conference, WCNC*, vol. 2020-May, 2020.

[44] M. J. Hashemi and E. Keller, "Enhancing robustness against adversarial examples in network intrusion detection systems," in *2020 IEEE Conference on Network Function Virtualization and Software Defined Networks, NFV-SDN 2020 - Proceedings*, 2020.

[45] M. J. Hashemi, G. Cusack, and E. Keller, "Towards evaluation of nidss in adversarial setting," in *Big-DAMA 2019 - Proceedings of the 3rd ACM CoNEXT Workshop on Big DAta, Machine Learning and Artificial Intelligence for Data Communication Networks, Part of CoNEXT 2019*, 2019.

[46] C. Zhang, X. Costa-Perez, and P. Patras, "Tiki-taka: Attacking and defending deep learning-based intrusion detection systems," in *CCSW 2020 - Proceedings of the 2020 ACM SIGSAC Conference on Cloud Computing Security Workshop*, 2020.

[47] R. A. Khamis and A. Matrawy, "Evaluation of adversarial training on different types of neural networks in deep learning-based idss," in *2020 International Symposium on Networks, Computers and Communications, ISNCC 2020*, 2020.

[48] J. Wang, J. Pan, I. Alqerm, and Y. Liu, "Def-ids: An ensemble defense mechanism against adversarial attacks for deep learning-based network intrusion detection," in *Proceedings - International Conference on Computer Communications and Networks, ICCCN*, vol. 2021-July, 2021.

[49] M. P. Novaes, L. F. Carvalho, J. Lloret, and M. L. ProenÃ§a, "Adversarial deep learning approach detection and defense against ddos attacks in sdn environments," *Future Generation Computer Systems*, vol. 125, 2021.

[50] Y. Wang, S. lv, J. Liu, X. Chang, and J. Wang, "On the combination of data augmentation method and gated convolution model for building effective and robust intrusion detection," *Cybersecurity*, vol. 3, 2020.

[51] H. Zhang, X. Yu, P. Ren, C. Luo, and G. Min, "Deep adversarial learning in intrusion detection: A data augmentation enhanced framework," *CoRR*, vol. abs/1901.07949, 2019. [Online]. Available: http://arxiv.org/abs/1901.07949

[52] A. U. H. Qureshi, H. Larijani, M. Yousefi, A. Adeel, and N. Mtetwa, "An adversarial approach for intrusion detection systems using jacobian saliency map attacks (jsma) algorithm," *Computers*, vol. 9, 2020.

[53] I. Debicha, T. Debatty, J.-M. Dricot, and W. Mees, "Adversarial training for deep learning-based intrusion detection systems," *CoRR*, vol. abs/2104.09852, 2021. [Online]. Available: https://arxiv.org/abs/2104.09852

[54] 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, 2021.

[55] N. Martins, J. M. Cruz, T. Cruz, and P. H. Abreu, "Analyzing the footprint of classifiers in adversarial denial of service contexts," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 11805 LNAI, 2019.

[56] 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," *Computers and Security*, vol. 108, 2021.

[57] F. O. Catak and S. Y. Yayilgan, "Deep neural network based malicious network activity detection under adversarial machine learning attacks," in *Communications in Computer and Information Science*, vol. 1382, 2021.

[58] G. Apruzzese, M. Andreolini, M. Colajanni, and M. Marchetti, "Hardening random forest cyber detectors against adversarial attacks," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 4, 2020.

[59] A. Mccarthy, P. Andriotis, E. Ghadafi, and P. Legg, "Feature vulnerability and robustness assessment against adversarial machine learning attacks," in *2021 International Conference on Cyber Situational Awareness, Data Analytics and Assessment, CyberSA 2021*, 2021.

[60] G. Apruzzese, M. Andreolini, M. Marchetti, V. G. Colacino, and G. Russo, "Appcon: Mitigating evasion attacks to ml cyber detectors," *Symmetry*, vol. 12, 2020.

[61] I. Debicha, T. Debatty, J. M. Dricot, W. Mees, and T. Kenaza, "Detect reject for transferability of black-box adversarial attacks against network intrusion detection systems," in *Communications in Computer and Information Science*, vol. 1487 CCIS, 2021.

[62] A. U. H. Qureshi, H. Larijani, N. Mtetwa, M. Yousefi, and A. Javed, "An adversarial attack detection paradigm with swarm optimization," in *Proceedings of the International Joint Conference on Neural Networks*, 2020.

[63] M. Pawlicki, M. ChoraÅ, and R. Kozik, "Defending network intrusion detection systems against adversarial evasion attacks," *Future Generation Computer Systems*, vol. 110, 2020.

[64] Y. Peng, G. Fu, Y. Luo, J. Hu, B. Li, and Q. Yan, "Detecting adversarial examples for network intrusion detection system with gan," in *Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS*, vol. 2020-October, 2020.

[65] H. Jiang, J. Lin, and H. Kang, "Fgmd: A robust detector against adversarial attacks in the iot network," *Future Generation Computer Systems*, vol. 132, 2022.

[66] H. Kang, D. H. Ahn, G. M. Lee, J. D. Yoo, K. H. Park, and H. K. Kim, "Iot network intrusion dataset," 2019. [Online]. Available: https://dx.doi.org/10.21227/q70p-q449

[67] A. Guerra-Manzanares, J. Medina-Galindo, H. Bahsi, and S. NÃµmm, "Medbiot: Generation of an iot botnet dataset in a medium-sized iot network," in *ICISSP 2020 - Proceedings of the 6th International Conference on Information Systems Security and Privacy*, 2020.

[68] G. Apruzzese, M. Colajanni, L. Ferretti, and M. Marchetti, "Addressing adversarial attacks against security systems based on machine learning," in *International Conference on Cyber Conflict, CYCON*, vol. 2019-May, 2019.

[69] N. Wang, Y. Chen, Y. Xiao, Y. Hu, W. Lou, and T. Hou, "Manda: On adversarial example detection for network intrusion detection

system," *IEEE Transactions on Dependable and Secure Computing*, 2022.

[70] N. Moustafa and J. Slay, "Unsw-nb15: A comprehensive data set for network intrusion detection systems (unsw-nb15 network data set)," in *2015 Military Communications and Information Systems Conference, MilCIS 2015 - Proceedings*, 2015.

[71] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the kdd cup 99 data set," in *IEEE Symposium on Computational Intelligence for Security and Defense Applications, CISDA 2009*, 2009.

[72] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9185–9193.

[73] O. Faker and E. Dogdu, "Intrusion detection using big data and deep learning techniques," in *ACMSE 2019 - Proceedings of the 2019 ACM Southeast Conference*, 2019.

[74] Y. Zhang, X. Chen, D. Guo, M. Song, Y. Teng, and X. Wang, "Pccn: Parallel cross convolutional neural network for abnormal network traffic flows detection in multi-class imbalanced network traffic flows," *IEEE Access*, vol. 7, 2019.

[75] Y. Zhang, X. Chen, L. Jin, X. Wang, and D. Guo, "Network intrusion detection: Based on deep hierarchical network and original flow data," *IEEE Access*, vol. 7, 2019.

[76] B. Ingre and A. Yadav, "Performance analysis of nsl-kdd dataset using ann," in *International Conference on Signal Processing and Communication Engineering Systems - Proceedings of SPACES 2015, in Association with IEEE*, 2015.

[77] A. Odena, C. Olah, and J. Shlens, "Conditional image synthesis with auxiliary classifier gans," in *34th International Conference on Machine Learning, ICML 2017*, vol. 6, 2017.

[78] A. Odena, "Semi-supervised learning with generative adversarial networks," *arXiv preprint arXiv:1606.01583*, 2016.

[79] I. Sharafaldin, A. H. Lashkari, S. Hakak, and A. A. Ghorbani, "Developing realistic distributed denial of service (ddos) attack dataset and taxonomy," in *Proceedings - International Carnahan Conference on Security Technology*, vol. 2019-October, 2019.

[80] G. Creech and J. Hu, "Generation of a new ids test dataset: Time to retire the kdd collection," in *IEEE Wireless Communications and Networking Conference, WCNC*, 2013.

[81] S. GarcÃa, M. Grill, J. Stiborek, and A. Zunino, "An empirical comparison of botnet detection methods," *Computers and Security*, vol. 45, 2014.

[82] S. Ustebay, Z. Turgut, and M. A. Aydin, "Intrusion detection system with recursive feature elimination by using random forest and deep learning classifier," in *International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism, IBIGDELFT 2018 - Proceedings*, 2019.

[83] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in *ICISSP 2018 - Proceedings of the 4th International Conference on Information Systems Security and Privacy*, vol. 2018-January, 2018.

[84] C. Systems, "Cisco ios netflow version 9 flow-record format," *White Paper*, 2011. [Online]. Available: https://www.cisco.com/en/US/technologies/tk648/tk362/technologies_white_paper09186a00800a3db9.html

N. Dhinakaran N. Dhinakaran is doing his PhD in the School of Computer Science and Engineering at Vellore Institute of Technology, Vellore, India. He has eleven years of teaching experience. He completed his ME in the year 2007. His areas of interest include networking, machine learning, and network security. He has published various papers in reputed journals.

S. Anto S Anto is an Associate Professor in the Department of Computational Intelligence in Vellore Institute of Technology, Vellore, India. He has more than 20 years of teaching experience in the field of engineering education. His area of research are artificial intelligence, fog computing and classifier parameter optimization. His focus is on applying cognitive computing to improve the effectiveness of medical expert systems. He has published three patents and many research articles in refereed journals. He is an active member of IEEE.