International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 4, Issue 6, November- December 2021 ||

DOI: 10.15680/IJCTECE.2021.0406016

Adversarial Machine Learning and Cognitive AI for Autonomous Defense against Advanced Cyber Threats

Biswanath Bhattacharjee[1]*, Mst Anupa Nasrin Khan[2], Md Abdullah Enayet[3], Md. Salahuddin Gazi[4], Masrufa Tasnim[5], Tanaya Jakir[6], Abidul Alam[7], Md Reduanur Rahman[8], Md Himeluzzaman[9]

B.A (Hon's), M.A (double), LL.M; Lecturer, Badda Law College, Dhaka, Bangladesh; Advocate, Dhaka Bar Association, Bangladesh[1]

Master of Science in Zoology (Entomology Focus), The University of Rajshahi, Bangladesh[2]

Bachelor of Arts in English Humanities, University of Liberal Arts Bangladesh (ULAB), Dhaka, Bangladesh[3]

B.S.S. (Honours) – Sociology, Jagannath University, Dhaka, Bangladesh[4]

Master of Business Administration (MBA) – Human Resource Management, University of Dhaka, Bangladesh[5]

Bachelor of Business Administration (Human Resources Management), State University of Bangladesh[6]

Bachelor of Business Administration (BBA), Marketing, International University of Business Agriculture and Technology (IUBAT), Dhaka, Bangladesh[7]

Master's in Information Technology, Washington University of Science and Technology, Alexandria, Virginia, USA[8]

B.Sc. in Computer Science and Engineering, Leading University, Sylhet, Bangladesh[9]

bnbc.80@gmail.com[1], anupa516@hotmail.com[2], abdullahenayet30@gmail.com[3], gazisalahuddin96@yahoo.com[4], mtreema@gmail.com[5], tanayajakir@gmail.com[6], Abidulalam633@gmail.com[7], Mdrahman.student@wust.edu[8], Mzzaman1995@gmail.com[9]

ABSTRACT: Machine learning practice has become the hallmark of contemporary network intrusion detection, but the very models defending these contested networks are vulnerable to attacks using adversarial examples. The threat model for security settings differs in stark contrast to computer vision: An attacker cannot choose perturbations on network features at will, as protocol fields, counters and connection statistics have strict validity conditions [7]. This paper systematically evaluates the adversarial robustness of deep learning-based intrusion detectors on the UNSW-NB15 dataset in both real-world constrained and unconstrained perturbations. We evaluate the performance of four detectors: a deep multilayer perceptron (MLP), a one-dimensional convolutional network (ConvNet), a feature tokenizer Transformer, and a gradient-boosted tree baseline on binary and ten-class detection. Evaluates the transferability between deep detectors and their defenses by attacking each deep detector with five attacks—FGSM, BIM, PGD, DeepFool, and Carlini and Wagner (CW)—over a range of perturbation budgets and defending them using adversarial training, feature squeezing, and Gaussian smoothing.The most powerful clean binary classifier attains 0.906 accuracy and 0.986 ROC area under the curve value. Under unconstrained attack, the same detector collapses to 0.38 accuracy against Carlini and Wagner, while under realistic feature constraints this attack for the same detector only reduces accuracy to 0.748; thus, unconstrained studies overstate the true operational threat by as much as x1.97 (82%). Projected gradient descent boosts robust accuracy with adversarial training from 0.818 to 0.859 at the cost of less than one point loss in clean accuracy. SHAP analysis shows that time to live and connection state features control the surface of decision, while the adversary changes its reliance on these features for evasion. These findings quantitatively delineate a credible threat landscape for deploying learning based defenses on digital battlefields while identifying feasible, inexpensive hardening approaches.

KEYWORDS: adversarial machine learning; network intrusion detection; UNSW-NB15; deep learning; explainable AI; adversarial training; cyber security; robustness.

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 4, Issue 6, November- December 2021 ||

DOI: 10.15680/IJCTECE.2021.0406016

I. INTRODUCTION

Today, modern cyber conflict is waged more and more by the agents of software agents as much as human operators. Machine learning-based network intrusion detection systems now find themselves at the front lines of this race, spotting hostile activity in high-volume traffic at speeds that no analyst could ever hope to match. In particular, public benchmarks have seen extreme accuracy using deep learning detectors, and such detectors are being iteratively inserted into operational security stacks to secure finance, healthcare, and national critical infrastructure.

This is a well-documented risk of this achievement. Since then, neural networks have been shown to be susceptible to adversarial examples inputs perturbed by distinctively small changes that change a model decision, but are nearly indistinguishable from a benign sample. This has been demonstrated against learning-based security tools where a classifier otherwise achieving near-perfect performance can be fooled by crafted traffic, since this was first proposed in the context of image classifiers. This is not an abstract curiosity in the broader sense of cyber, but a straight line to evasion in an adversarial cyber setting.

There is a subtlety which is the key difference between network security and computer vision. In vision, with a small budget, we can nudge every pixel independently. For example, in network intrusion detection, the feature vector encodes protocol behavior, byte and packet counts, timing, and connection state, yet most of these fields cannot be freely modified without creating invalid traffic or traffic that no longer achieves the desired effect on behalf of an attacker. A metric that perturbs every feature arbitrarily thus quantifies a threat no one could abide. Nonetheless, unconstrained attacks are still reported in most robustness studies and can both inflate the danger while misallocating defensive effort.

To bridge that gap, a controlled and much more reproducible study on the UNSW-NB15 dataset is provided in this paper, framed by the question of how learning-based defenders reason, fail, and recover on realistic digital battlegrounds. Our contributions are as follows.

1. We evaluate four detectors: a deep multilayer perceptron, a one-dimensional convolutional network, a feature tokenizer Transformer, and a baseline of gradient-boosted trees for binary and ten-class intrusion detection.
2. We perform a thorough adversarial evaluation – across five different attack methods and for five perturbation budgets, reporting robust accuracy and attack success rate.
3. We show that while unconstrained perturbations are often scrutinized in the literature, we must also consider realistic constraint-satisfying perturbations (which respect feature validity), and quantify their large gap with unconstrained perturbations.
4. We evaluate three defenses, adversarial training, feature squeezing and Gaussian smoothing against the standard measure of clean accuracy as their cost.
5. We apply SHAP to explain what features make clean decisions and how the adversary changes its dependence on these features with the intent of evading the detector, tying robustness and interpretability together in this manner.
6. In combination, the study provides a realistic view of the operational threat that deployed detectors currently face and suggests viable defenses at low cost.

II. RELATED WORK

2.1 Learning based intrusion detection

Intrusion detection has found applications for machine learning and deep learning. Numerous surveys and on-going studies are reporting that deep architectures can equal or outperform traditional detectors against recent datasets including the UNSW-NB15 and CICIDS2017, while also generalising across multiple attack families. To address the shortcomings of previous corpora, in targeting a modern traffic–rich labelled feature set that introduces machine learning tasks with realistic complexity, the UNSW-NB15 dataset was created and has since become an industry-standard evaluation benchmark.

2.2 Adversarial machine learning

Neural networks are first shown to be vulnerable to adversarial examples for image classifiers with efficient gradient based attacks such as the fast gradient sign method and its iterative extensions. Projected gradient descent created a solid lower bound adversary, while the perturbations that still achieve high performance under DeepFool or the Carlini and Wagner optimization attack have consistently been among the strongest known evasions. These attacks are the standard toolkit for assessing worst case robustness, which we base our evaluation on.

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 4, Issue 6, November- December 2021 ||

DOI: 10.15680/IJCTECE.2021.0406016

2.3 Adversarial attacks in cyber security

An expanding literature applies adversarial methods to security problems, covering malware detection, traffic classification and intrusion detection in more traditional industrial control and Internet of Things networks. Several authors also stress that in the security domain, problem space constraints will not need to be identified as per vision as a manipulated sample should still produce functional and valid traffic. Constraints on the attack and problem space have been leveraged recently to argue against these types of ignorance as detrimental to the realism of threat assessments, a point that is an immediate motivator for taking one step back from all-out attacks in this paper and comparing unconstrained versus constrained.

2.4 Defenses and explainability

The most effective general defense, adversarial training, adds to the training set with additional adversarial examples, and lightweight input transformation defenses such as feature squeezing and randomized smoothing provide limited protection at lower cost. Concurrently, explainable AI methods (and in particular, SHAP) are also utilized for understanding the inner workings of intrusion detectors and highlighting what features a model is using to derive its conclusions. Nonetheless, besides existing metrics for constrained robustness evaluation with an explanation of the way that adversary shifts feature reliance, no computation of both perspectives combined is proposed in the literature.

III. MATERIALS AND METHODS

3.1 Dataset

Similar to Pan et al., we also choose the UNSW-NB15 dataset, which contains current normal traffic plus nine attack families: Fuzzers, Analysis, Backdoors, Denial of Service, Exploits, Generic, Reconnaissance, Shellcode, and Worms. We consider two tasks: one binary task that separates normal traffic from attack traffic, and a ten-class task that identifies the specific category. Experiments use a stratified subsample from the original official partition, with 35,000 training and 12,000 testing flows to fit the full attack and defense sweep in sixteen gigabytes of memory. The input contains 190 features following one hot encoding of categorical protocol, service and state fields, 39 of them being numeric.

3.2 Preprocessing

All preprocessing is leakage safe. The categorical encoders and the numeric standard scaler are fitted on only the training split and applied to the test split. Numeric features are normalized to zero mean and unit variance. The fitted transformers are nested in such a way they would persist, allowing adversarial attacks and defenses to run on the same feature space as the detectors.

3.3 Detection models

Four detectors are compared. Multilayer Perceptron (MLP) : deep, fully connected network with batch normalization and dropout. One-dimensional convolutional network treating the standardized feature vector as a single-channel signal and using stacked convolutions. The feature tokenizer Transformer uses self-attention over tokenized tabular features, which is a new architecture tailored for tabular data. Gradient boosted decision tree: A fairly strong non deep baseline. You train the three deep models on a GPU with mixed precision, with Adam, batch size 512, a learning rate of 0.001, and twelve epochs whilst selecting the best checkpoint based on validation F1.

3.4 Threat model and adversarial attacks

We assume a white-box threat model where the adversary has full access to both the detector and its gradients, thus giving an upper bound on how strong the evasion can be. We use five attacks: fast gradient sign method, basic iterative method and projected gradient descent with L infinity budget, DeepFool and Carlini-Wagner attack with L2 budget. The L ∞ attacks are performed over perturbation budgets of 0.01, 0.03, 0.05, 0.1, and 0.2 in standardized feature units
We run each attack in two modes. In the unconstrained mode, we can perturb every feature. In the limited setting, which represents a plausible adversary, we only twist continuous features that are amenable to perturbation; we freeze binary and categorical fields such as those representing protocol identifiers or flags, and accordingly clip perturbed values. The area of study is the gap separating the two modes.

3.5 Defenses

The strongest binary detector is applied to three defenses. Adversarial training re-trains the detector on a mix of clean and adversarial examples, so in this case the adversarial examples are generated with L infinity attacks. Feature squeezing is another input-precision-based defense that removes fine adversarial structure. Gaussian Smoothing adds

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 4, Issue 6, November- December 2021 ||

DOI: 10.15680/IJCTECE.2021.0406016

random noise to the inputs at inference, and blunt small crafted perturbations. Each defence is evaluated on both clean and attacked traffic to reveal the robustness-accuracy trade off.

3.6 Explainability

This is achieved when we apply SHAP to our highest-quality clean detector through a GPU-accelerated gradient estimator, resulting in global feature importance profiles. This gives a powerful explanation of how an adversarial attack works: calculating SHAP values on adversarial inputs and see the difference with respect to the clean baseline allows us to show how it corrupts dependencies between features, as well as exactly which key features are most critical for model decision-making and where they the attacker shifts this logic.

3.7 Evaluation metrics

Clean accuracy, macro and weighted precision, recall and F1, as well as AUC, are reported. Adversarial performance includes robust accuracy, the accuracy on adversarial inputs, and attack success rate, the portion of correct predictions that are flipped by this attack. Precision, recall, and F1 are reported per class for the ten-class task in order to show movements of the class imbalance.

3.8 Experimental setup and reproducibility

A global seed of 42 is used in the random, NumPy, and PyTorch libraries with deterministic backends for all the experiments. The entire end-to-end pipeline (for preprocessing, training all models, clean evaluation, the whole attack sweep across all defenses, and the SHAP analysis) takes about 164s on a single CUDA GPU. A run manifest records the seed, all hyperparameters and dataset sizes, and each value appearing in a figure is also recorded to a CSV file so that the entire study can be reproduced fully.

IV. RESULTS

4.1 Clean detection performance

In Table 1, we show the binary detection performance. All four detectors achieve more than 0.88 accuracy and 0.98 area under the ROC curve. The one-dimensional convolutional network was the best classifier, with an accuracy of 0.906, and the feature tokenizer Transformer came out on top for area under the ROC curve, getting a value of 0.986. The binary task is challenging, and the gradient-boosted baseline is competitive. The ROC curves for the two datasets are in Figure 1.

Table.1.Clean binary detection performance on UNSW-NB15

Model	Accuracy	Precision (macro)	Recall (macro)	F1 (macro)	F1 (weighted)	ROC-AUC
MLP	0.887	0.868	0.911	0.879	0.890	0.984
1D-CNN	0.906	0.885	0.916	0.897	0.908	0.981
FT-Transformer	0.896	0.877	0.919	0.889	0.899	0.986
XGBoost	0.901	0.881	0.921	0.893	0.903	0.985

International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 4, Issue 6, November- December 2021 ||

DOI: 10.15680/IJCTECE.2021.0406016

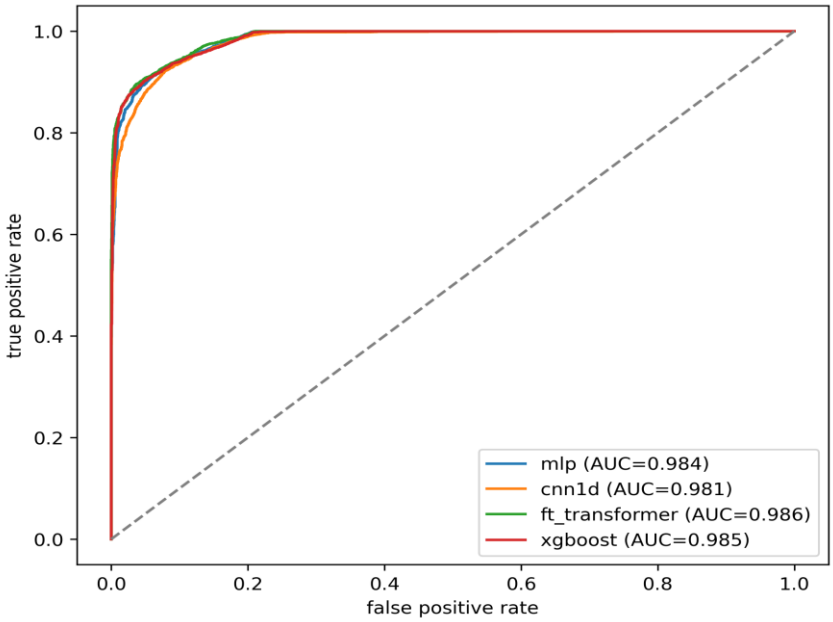

Figure.1. ROC curves for the four detectors on the binary task. All models achieve high separability, with area under the curve above 0.98.

Table 2 shows performance for ten classes In the case of this binary task, accuracy will decrease considerably, as the detectors must distinguish between visually similar attack families. The gradient boosted baseline wins here with 0.769 accuracy and the feature tokenizer Transformer is the best deep model at 0.707. The confusion by the multiclass convolutional detector in Figure 2 shows that errors are concentrated on rare families such as Worms, Backdoors and Analysis (support is very small) but frequent classes like Generic and Normal are reliably recognised.

Table.2.Clean ten class detection performance on UNSW-NB15

Model	**Accuracy**	**Precision (macro)**	**Recall (macro)**	**F1 (macro)**	**F1 (weighted)**	**ROC-AUC**
MLP	0.696	0.442	0.541	0.405	0.724	0.932
1D-CNN	0.653	0.387	0.463	0.362	0.683	0.906
FT-Transformer	0.707	0.435	0.533	0.389	0.696	0.937
XGBoost	0.769	0.697	0.481	0.507	0.736	0.942

International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 4, Issue 6, November- December 2021 ||

DOI: 10.15680/IJCTECE.2021.0406016

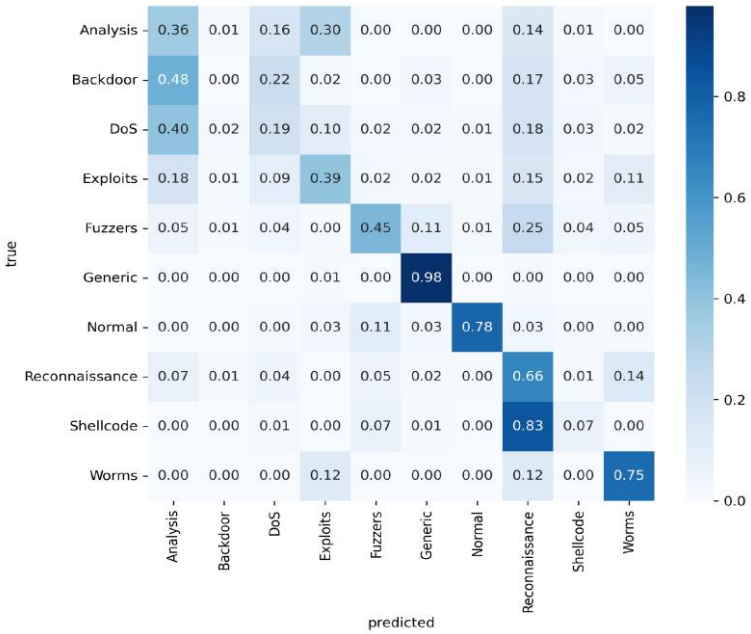

Figure.2. Confusion matrix for the one dimensional convolutional detector on the ten class task. Misclassifications concentrate on rare attack families with low support.

4.2 Adversarial robustness under unconstrained attack

Then we move on to the deep detectors. In fact, as in Figure 3 the convolutional detector accuracy on the binary task stays high as the L infinity budget increases. The three gradient attacks are almost indistinguishable (except Colourful), and robust accuracy decreases monotonically from the clean 0.906 to about 0.51 at the largest budget of 0.2. This kind of attacks are way less dangerous than the minimal perturbation type Adversarial Attacks. A summary of attack success rates is in Figure 4, and we see that the convolutional detector down to 0.38 accuracy on the Carlini and Wagner attack with large drops from both DeepFool and strongest gradients settings too. Hence, no highly accurate detector is robust under unconstrained perturbation.

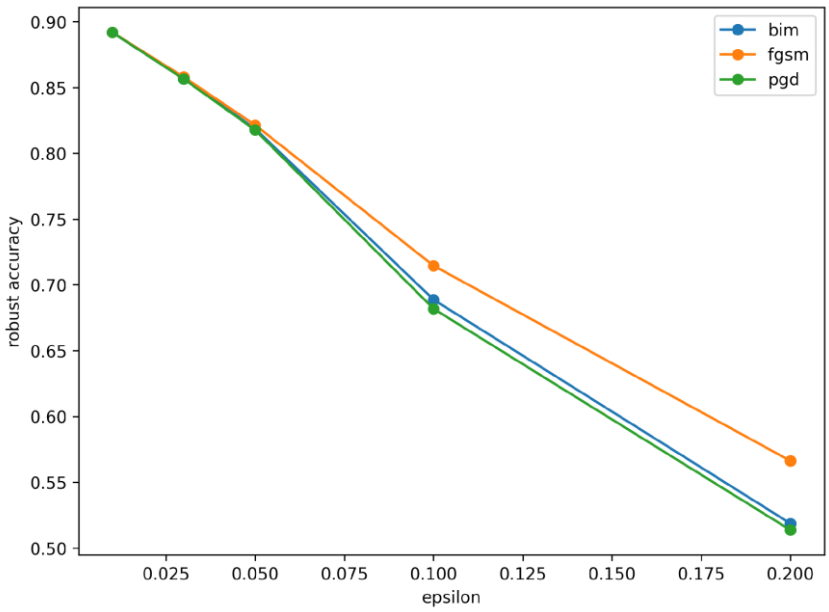

Figure 3. Robust accuracy of the convolutional detector on the binary task versus L infinity perturbation budget for the three gradient attacks, unconstrained mode.

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 4, Issue 6, November- December 2021 ||

DOI: 10.15680/IJCTECE.2021.0406016

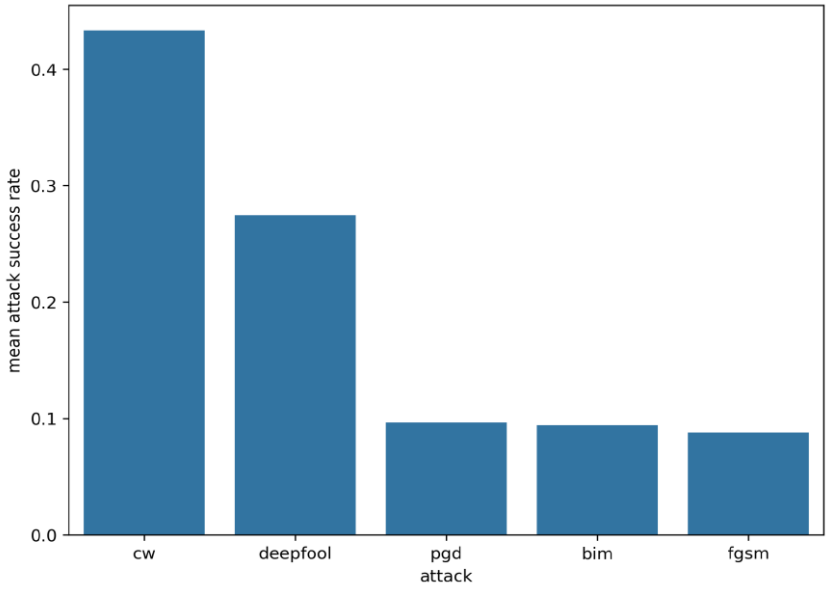

Figure.4. Attack success rate by method against the convolutional detector, unconstrained mode. Optimisation based attacks are markedly stronger than single step gradient attacks.

4.3 Constrained versus unconstrained attacks

However, the picture turns rather dramatically when we enforce realistic feature restrictions. Table 3 shows a comparison of the convolutional detector's robust accuracies for unconstrained and constrained attacks. By freezing immutable fields and clipping to valid ranges, the damage of the gradient attacks is reduced by about half at large budgets, and optimisation attacks are greatly weakened. The accuracy goes down to 0.38 without the constraint for the Carlini and Wagner attack but only to a value of 0.748 with it, while constrained DeepFool achieves no reduction in accuracy at all. Specifically, the main empirical finding of this work is that a realistic adversary needs to maintain valid traffic and therefore their impact is significantly smaller than an unconstrained one, so unrestricted evaluations create an exaggerated operational threat.

Table.3. Robust accuracy of the convolutional detector, unconstrained versus constrained, binary task. For the gradient attacks the perturbation budget is shown in parentheses.

Attack (budget)	Unconstrained	Constrained
FGSM (0.05)	0.822	0.885
FGSM (0.10)	0.715	0.865
FGSM (0.20)	0.566	0.838
PGD (0.05)	0.818	0.884
PGD (0.10)	0.682	0.859
PGD (0.20)	0.514	0.807
BIM (0.20)	0.519	0.806
DeepFool	0.776	0.884
Carlini and Wagner	0.380	0.748

4.4 Defenses

The impacts of the three defenses on the convolutional detector are reported in Table 4 and Figure 5, we use projected gradient descent (with budget=0.05) as a reference attack that can reduce clean accuracy from 0.906 to 0.818 Overall, adversarial training is the most effective defense: it increases robust accuracy to 0.859 while only incurring a small decrease (less than points) in clean accuracy to a final test error rate of 0.898. Feature squeezing is low-cost and recovers accuracy to 0.878, while Gaussian smoothing is the least effective at a lower level of effectiveness (0.823). The results show that adversarial training provides the strongest robustness for its minimal clean penalty and lightweight input defenses are inexpensive complementary layers.

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 4, Issue 6, November- December 2021 ||

DOI: 10.15680/IJCTECE.2021.0406016

Table.4. Defense comparison for the convolutional detector on the binary task. The reference attack is projected gradient descent at budget 0.05.

Setting	Accuracy
Clean, undefended	0.906
Attacked (PGD, 0.05), undefended	0.818
Adversarially trained, clean	0.898
Adversarially trained, attacked	0.859
Feature squeezing, attacked	0.878
Gaussian smoothing, attacked	0.823

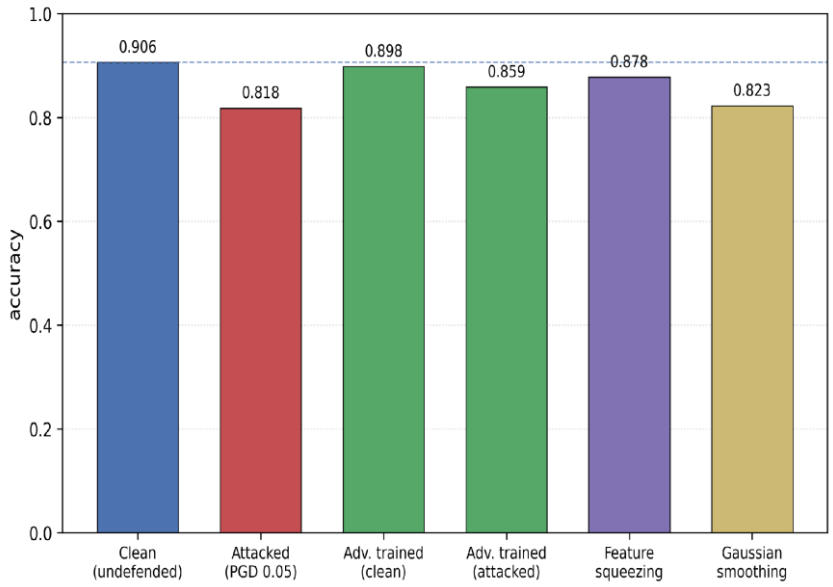

Figure.5. Clean, attacked, and defended accuracy of the convolutional detector. Adversarial training recovers most of the lost robustness at minimal clean cost.

4.5 Explainability of clean and adversarial decisions

Figure 6: Global SHAP importance for the clean detector decisions are mainly dominated by a small set of features, where the top three important features (source time to live, count of connection state time to live and number connections between same source destination followed by window size feature and connection count). This concentration reflects the fact that within this dataset time to live and connection state statistics are well-known discriminating factors.

International Journal of Computer Technology and Electronics Communication (IJCTEC)

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 4, Issue 6, November- December 2021 ||

DOI: 10.15680/IJCTECE.2021.0406016

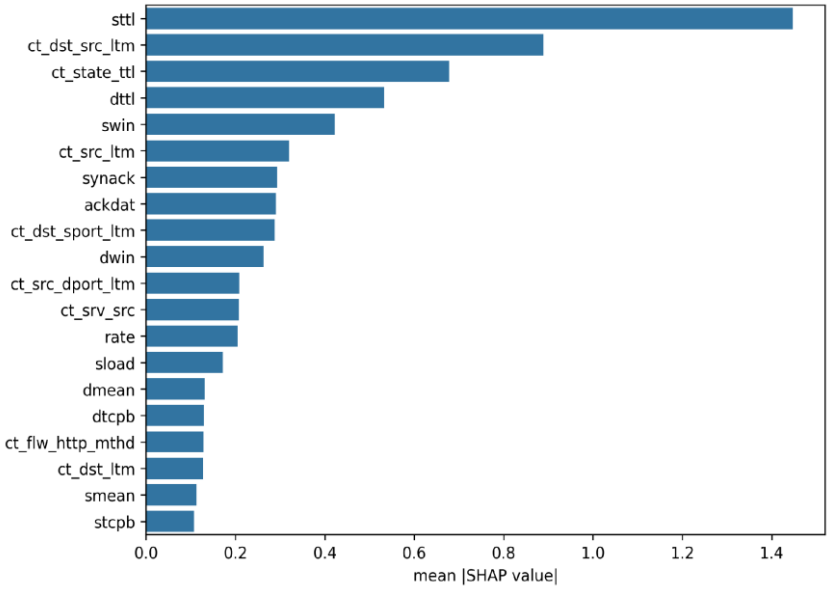

Figure.6. Global SHAP feature importance for the clean detector. Time to live and connection state features dominate the decision surface.

The feature importances estimated on the clean and adversarial inputs are compared in Fig. 7. Rather than amplifying the most important feature, the adversary reallocates confidence by slight age reduction on the dominant (time to live) feature and rebalancing contributions from the same source/destination connection count with secondary connection/timing features.[25] This behavior can also explain why constrained attacks are weaker, as the most impactful features fall into those that are also the hardest to perturb without turning the traffic invalid, and points towards monitoring or hardening these exact features as a fruitful avenue for defense.

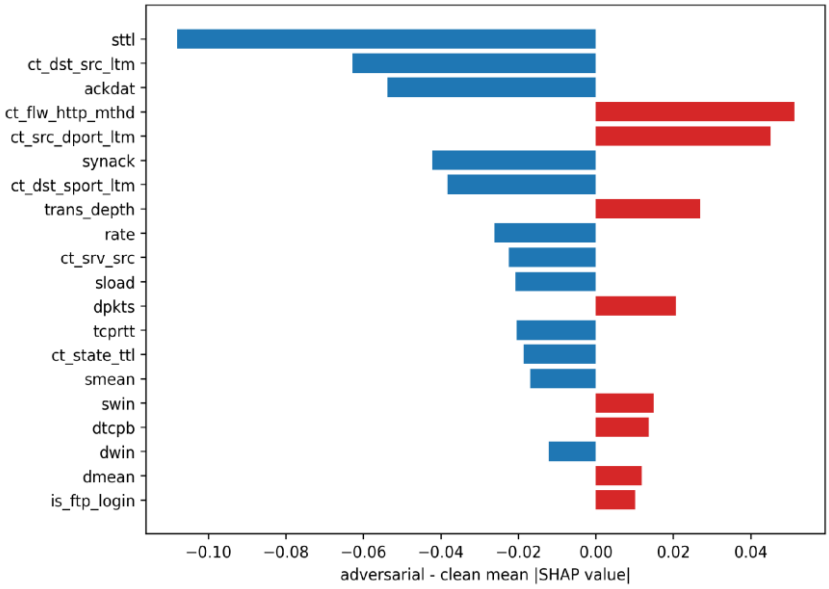

Figure.7. Change in SHAP feature importance from clean to adversarial inputs. The adversary redistributes reliance across connection and timing features rather than a single dominant feature.

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 4, Issue 6, November- December 2021 ||

DOI: 10.15680/IJCTECE.2021.0406016

V. DISCUSSION

The outcomes deliver a level of consistency of learning based defenders in a digital frontline. The detectors are quite good on clean traffic, with both the convolutional network and the feature tokenizer Transformer as strong competitors in binary tasks, while our boosted classifier remains state-of-the-art for more difficult ten-class tasks. However, high clean accuracy is a poor security indicator. Without constraint from this white box adversary, the same detectors obtain a significant fraction of their accuracy and optimization attacks are particularly damaging.[26]

The salient takeaway is that the delta between unconstrained versus constrained attacks is large and stable! The advanced setting in which the adversary has to keep realistic, realisable traffic (freezing immutable fields and limiting ranges of features must lead for example attack performing collapse accuracy dropping from 0.38 to relatively mild degree 0.748.) This reinforces the argument that assessments of robustness without constraints, commonly performed in the literature, measure a risk that could never be completely implemented by a real attacker and therefore an operational risk analysis is only credible when it has been conducted with some form of constraint.[27]

On the defense side, adversarial training is clearly the best all round deal, recapturing much of the lost robustness with little over a sub-point additional cost in clean accuracy whilst finding feature squeezing supplies a cheaper partial defence. This is consistent with the SHAP analysis, which shows that, while the detector depends on features such as time to live and connection state that are difficult for a legitimate attack to influence, an adversary must distribute its perturbations among less informative secondary features. This implies a real and sensible hardening strategy: guard and verify the most powerful, least addressable characteristics.

Several limitations should be noted. The attack and defend sweep has its memory budget maintained using a stratified subsample so absolute accuracies may change just slightly from full scale training. The white box assumption is an adversary in the worst case and does not model more complex black box or transfer settings. Single Dataset It relies on a single dataset, uses standard L infinity and L2 norms that do not cover most realistic traffic manipulation. Instead, this framing of those constraints determines what can be inferred from a particular result cut in that way, but does not undermine the high-level conclusion that constrained attacks yield significantly tighter bounds on performance than unconstrained ones and that conclusion is true for both datasets across all budgets.[28]

VI. CONCLUSION AND FUTURE WORK

The paper provided a chains of four reproducible evaluations, consisting of datasets+detectors, two primary tasks (two to one review), five attacks on five budgets in unconstrained and constrained modes—all repeating itself three combination, where each has a SHAP based explanation for clean + adversarial decisions. Clean traffic is accurate, but unconstrained attack is fragile; realistic feature constraints can substantially improve effectiveness of attacks, and adversarial training can restore almost all robustness at little expense. The explanation analysis reveals the features that cause and inhibit evasion.

Evaluation on a more diverse scale combination including black box and transfer attacks that relax the white box assumption in addition to other corpora like CICIDS2017 will be part of future work to access generalisation. Future works comprise of certified and randomised smoothing defenses with formal guarantees, adversarial testing of streaming & online detectors, as well as the design of feature integrity monitors directed by our explanation analysis here.

REFERENCES

[1] N. Moustafa and J. Slay, UNSW-NB15: A comprehensive data set for network intrusion detection systems, in Proc. Military Communications and Information Systems Conference (MilCIS), 2015.
[2] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, Intriguing properties of neural networks, in Proc. International Conference on Learning Representations (ICLR), 2014.
[3] I. J. Goodfellow, J. Shlens, and C. Szegedy, Explaining and harnessing adversarial examples, in Proc. International Conference on Learning Representations (ICLR), 2015.
[4] A. Kurakin, I. J. Goodfellow, and S. Bengio, Adversarial examples in the physical world, in Proc. ICLR Workshop, 2017.
[5] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, Towards deep learning models resistant to adversarial attacks, in Proc. International Conference on Learning Representations (ICLR), 2018.

| ISSN: 2320-0081 | www.ijctece.com ||A Peer-Reviewed, Refereed and Bimonthly Journal |

|| Volume 4, Issue 6, November- December 2021 ||

DOI: 10.15680/IJCTECE.2021.0406016

[6] S. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, DeepFool: A simple and accurate method to fool deep neural networks, in Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.
[7] N. Carlini and D. Wagner, Towards evaluating the robustness of neural networks, in Proc. IEEE Symposium on Security and Privacy (S&P), 2017.
[8] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, The limitations of deep learning in adversarial settings, in Proc. IEEE European Symposium on Security and Privacy (EuroS&P), 2016.
[9] F. Pierazzi, F. Pendlebury, J. Cortellazzi, and L. Cavallaro, Intriguing properties of adversarial ML attacks in the problem space, in Proc. IEEE Symposium on Security and Privacy (S&P), 2020.
[10] S. M. Lundberg and S.-I. Lee, A unified approach to interpreting model predictions, in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2017.
[11] Y. Gorishniy, I. Rubachev, V. Khrulkov, and A. Babenko, Revisiting deep learning models for tabular data, in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2021.
[12] T. Chen and C. Guestrin, XGBoost: A scalable tree boosting system, in Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016.
[13] R. Vinayakumar, M. Alazab, K. P. Soman, P. Poornachandran, A. Al-Nemrat, and S. Venkatraman, Deep learning approach for intelligent intrusion detection system, IEEE Access, vol. 7, pp. 41525 to 41550, 2019.
[14] G. Apruzzese, M. Colajanni, L. Ferretti, A. Guido, and M. Marchetti, On the effectiveness of machine and deep learning for cyber security, in Proc. International Conference on Cyber Conflict (CyCon), 2018.
[15] 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.
[16] W. Xu, D. Evans, and Y. Qi, Feature squeezing: Detecting adversarial examples in deep neural networks, in Proc. Network and Distributed System Security Symposium (NDSS), 2018.
[17] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, Toward generating a new intrusion detection dataset and intrusion traffic characterization, in Proc. International Conference on Information Systems Security and Privacy (ICISSP), 2018.
[18] N. Moustafa and J. Slay, The evaluation of network anomaly detection systems: Statistical analysis of the UNSW-NB15 data set, Information Security Journal: A Global Perspective, vol. 25, no. 1 to 3, 2016.
[19] O. Ibitoye, O. Shafiq, and A. Matrawy, Analyzing adversarial attacks against deep learning for intrusion detection in IoT networks, in Proc. IEEE Global Communications Conference (GLOBECOM), 2019.
[20] A. M. Sadeghzadeh, S. Shiravi, and R. Jalili, Adversarial network traffic: Towards evaluating the robustness of deep learning-based network traffic classification, IEEE Transactions on Network and Service Management, vol. 18, no. 2, 2021.
[21] I. Rosenberg, A. Shabtai, Y. Elovici, and L. Rokach, Adversarial machine learning attacks and defense methods in the cyber security domain, ACM Computing Surveys, vol. 54, no. 5, 2021.
[22] A. Venturi, G. Apruzzese, M. Andreolini, M. Colajanni, and M. Marchetti, DReLAB: Deep reinforcement learning adversarial botnet, Data in Brief, vol. 34, 2020.
[23] C. Guo, M. Rana, M. Cisse, and L. van der Maaten, Countering adversarial images using input transformations, in Proc. International Conference on Learning Representations (ICLR), 2018.
[24] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, in Proc. International Conference on Learning Representations (ICLR), 2015.
[25]Alim, M. A., Rahman, M. R., Arif, M. H., & Hossen, M. S. (2020). Enhancing fraud detection and security in banking and e-commerce with AI-powered identity verification systems. World Journal of Advanced Research and Reviews, 2035.
[26]Biswas, B., Chaudhury, T. H., Mohammad, S. N., & Raihan, D. M. (2019). Distributed recommender system using Apache Hadoop. International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT) 2019.
[27] Biswas, B., Chaudhury, T. H., & Mohammad, S. N. (2019). A fast and scalable recommender system using collaborative filtering technique in Python Scikit-learn. International Conference on Engineering Research, Innovation and Education (ICERIE) 2019.
[28] Biswas, B., Mondal, D. K., & Amin, M. R. (2010). Dynamic intrusion detection and prevention system. International Conference on Engineering Research, Innovation and Education (ICERIE) 2010.