PENINSULA PUBLISHING LLC

EDRAAK

Vol. **(2024)**, 2024, **pp**. 26–31
ISSN: 3078-8412

Research Article

Fortifying AI Against Cyber Threats Advancing Resilient Systems to Combat Adversarial Attacks

Lal Hussain[1,*]

[1] *Department of Computer Science, University of Azad Jammu and Kashmir, Neelum Campus, Athmuqam 13230, Azad Kashmir, Pakistan.*

ARTICLE INFO

Article History
Received 1 Des 2023
Revised: 17 Jan 2024
Accepted 17 Feb 2024
Published 5 Mar 2024

Keywords
AI,
Cybersecurity,
Adversarial Attacks,
Reinforcement Learning,
Adaptive Defense.

ABSTRACT

The emerging complexity of the threats which the safer world faces as of now, let alone the adversary attacks, presents great troubles to the orthodox systems of cybersecurity. This research focuses on the use of AI to design and develop robust systems that prevent such shifting threats. As a result, the proposed adaptive defense mechanisms utilize RL in this investigation to contend with real-time emergent attack patterns. At the same time, Ensemble Methods are applied to improve the anomaly detection non-sensitive approaches by using several machine learning models in order to minimize the false positives and maximize the true positives. The utility of the proposed framework is assessed by using benchmarking data sets such as NSL-KDD and CICIDS 2017 to represent various adversarial settings. The results shown in the study show that RL is able to manage dynamic threats successfully with high detection ratio and low response time. Ensemble Methods supplement this by strengthening the reliability of detection and shortening the error margin. The study reveals that AI has implications for optimizing the possibilities of cybersecurity systems, as well as for the defense systems themselves, as the implementation of flexible and fundamentally suitable approaches is already possible at the present stage.

1. INTRODUCTION

The recent advancement of digital technology where core business infrastructures and new smart devices have become more vulnerable to cyber hazards. Adversarial attacks are a category of cyber intrusions when the hackers invade the system and use the cracks of the structure to attack. These attacks are dangerous to contemporary computer-based systems in areas like finance, healthcare, as well as government organizations where personal and important data, services are vulnerable [1]. Traditional systems of security in the area of cyberspace are considered to be highly rigid with strong focus on the rules of standard forms, which cannot be adequately changed to address emergent forms of threats. This constraint highlights the necessity of extending switches' architectures with AI for saving them from adversarial attacks, while also learning in real-time [2]. AI is turned into the cybersecurity's game changer. Using the big data for training of the ML, specific pattern indicative of the malicious activity can be learnt. RL and Ensemble Methods are well suited to the development of systems that do not only diagnose but also actively counter threats in real time. These techniques seem to establish the basic building blocks for efficient and adaptive anti-cyber threat measures [3].

Adversarial attacks take advantage of circumstances in machine learning systems, aimed at feeding them a little perturbation which they will end up misclassifying. As has been shown, even a relatively innocent looking network packet can be modified so that it is able to avoid detection via normal methods but nevertheless retain its dangerous purpose [4]. A simple scheme of the way adversarial attacks perform manipulations of inputs is represented in Figure 1.

*Corresponding author email: lal.hussain86@gmail.com
DOI: https://doi.org/10.70470/EDRAAK/2024/004

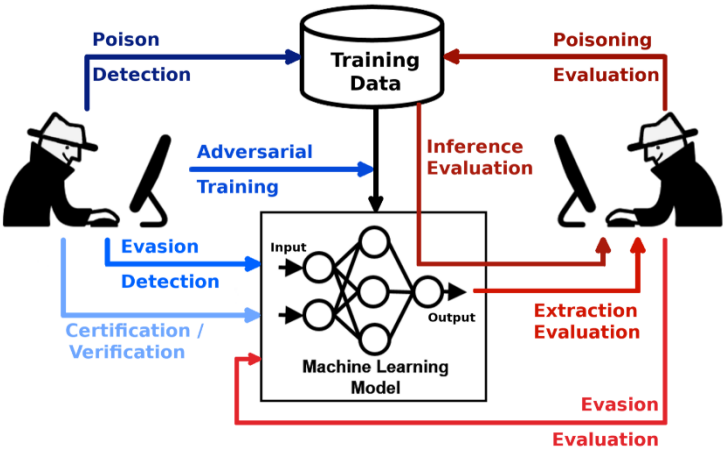

Fig 1: attacks manipulating input data to bypass detection.

The research aims to create a robust AI-driven cybersecurity framework using Reinforcement Learning and Ensemble Methods. It aims to design an adaptive defense mechanism, enhance anomaly detection reliability using Ensemble Methods, and evaluate the system's performance using benchmark datasets under various adversarial scenarios.

2. RELATED WORK

Cybersecurity is now blending with the application of Artificial Intelligence AI especially in defending growing complexities of adversarial attacks. The following works are reviewed with regard to the current literature on adversarial attacks, RL for adaptive defense, and Ensemble Methods for anomaly detection. The situation also highlights the drawbacks of these approaches and the research questions answered in the present study.

adversarial input is an effective way in which an adversary can take advantage of weaknesses found in machine learning models through manipulations aimed at triggering impermissible outputs, or model breakdown. Among the works cited by Goodfellow et al., one of the works that specifically explained the use of a Fast Gradient Sign Method (FGSM), revealed that neural networks are sensitive to adversarial perturbations [5]. Future work introduced methods such as adversarial training in which models are trained with adversarial samples to enhance robustness of models although this improves robustness but degrades generalization [6]. Further, complex methods such as CW and DeepFool have been observed to that effectively overcome traditional countermeasures as well as earlier AI based architectures [7]. Table I provides a summary of the type of adversarial attack methods and what they mean.

TABLE I : ADVERSARIAL ATTACK TECHNIQUES

Technique	Mechanism	Impact
FGSM [8]	Adds small perturbations to gradients	Reduces model accuracy
Carlini-Wagner (CW) [9]	Minimizes perturbations while maintaining stealth	Evades detection systems
DeepFool [10]	Iteratively adjusts inputs to fool classifiers	Breaks static defense mechanisms

One of the most exciting ideas has been the application of Reinforcement Learning (RL) for practical adaptive cybersecurity. Compared to the systems that require constant maintenance of the threat detection and prevention means, RL agents adaptively can learn and respond to threats by inquiring the environment. Lin et al., for example presented an intrusion detection system with high detection rates using DQN that enabled the system to learn with flow of new attack patterns [11-12]. Furthermore, the second widely applied RL algorithm is Proximal Policy Optimization (PPO) that is also applied to real-time threat response to ensure the correct reward-based actions. Still, RL systems have limitations such as scalability or computational complexity in working with extensive networks [13]. Table II presents other relevant works employing RL to cybersecurity.

TABLE II : APPLICATIONS OF RL IN CYBERSECURITY

Study	Algorithm	Focus	Outcome
[14]	Deep Q-Network (DQN)	Intrusion detection	High detection rate
[15]	Proximal Policy Optimization (PPO)	Real-time threat response	Reduced response time
[16]	Double Q-Learning	Adaptive malware mitigation	Improved scalability

Ensemble Methods enhance detection reliability by combining predictions from multiple models. Techniques like Random Forests, Gradient Boosting, and Stacking are widely used for anomaly detection due to their robustness and reduced false positives. For instance, Zhang et al. demonstrated that an ensemble approach significantly improved detection accuracy in the CICIDS 2017 dataset [17]. However, ensemble models require careful tuning to balance diversity among base classifiers, and their computational overhead may hinder real-time applications [18]. Table III provides a comparison of popular ensemble techniques in cybersecurity.

TABLE III : ENSEMBLE METHODS IN ANOMALY DETECTION

Technique	Mechanism	Strengths	Weaknesses
Random Forest [19]	Combines multiple decision trees	High accuracy, interpretable	Computationally intensive
Gradient Boosting [20]	Sequentially improves weak classifiers	Effective in imbalanced data	Risk of overfitting
Stacking [21]	Integrates outputs from diverse models	High robustness	Complex to implement

3. METHODOLOGY

In the next sections, the general approach of the allowed methodology for creating an efficient RL-based AI defense system in addition to Ensemble Methods against adversarial attacks will be discussed. Preprocessing atmosphere involves selections and preparations of datasets, designing of the RL and Ensemble based models as well as choosing the right metrics to use in evaluating the overall performance of the system.

3.1 Overview of the Framework

The proposed system ensures that its function and flexibility are improved by adopting two related AI techniques. RL is used in building adaptive defense systems that can change with time due to the continuous change in attacks. RL agents learn their environment and navigate it to detect threats there and apply the most efficient defense measures applying accumulated feedback. While on the other hand, Ensemble Methods is used to reduce false positives by and combining entire anomalous behavior algorithms in order to increase the efficiency of the algorithm. All these techniques are implemented within the purview of the intended framework; Reinforcement Learning for dynamic control of decisions and Ensemble Methods to provide a strong kinetic chain to the anomaly detection framework proposed for implementation.
Figure 2 illustrates the architecture of the proposed RL-based AI defense system and Ensemble Methods framework. It highlights the stages involved, from data preprocessing to adaptive threat mitigation and anomaly detection.

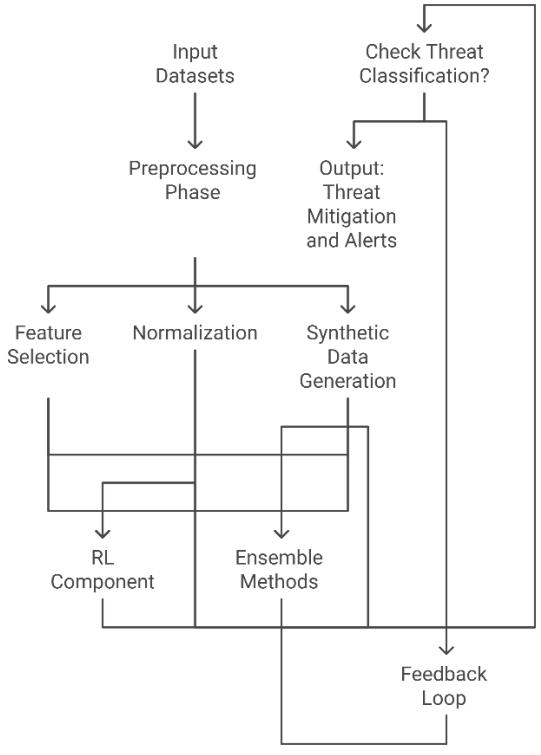

Fig 2: proposed RL-based AI defense system and Ensemble Methods framework

3.2 Datasets and data Preprocessing

To evaluate the system, two benchmark datasets were selected: The TON_IoT dataset and the Bot-IoT dataset. The TON_IoT is a perspective and modern large-scale dataset for intrusion detection and analysis of IoT networks is TON_IoT. It uses information from smart objects, network flows, and system output, which covers a number of scenarios such as man in the middle (MitM), distributed denial of service (DDoS), and unauthorised data transfer. Its multi-dimensional structure and labeled instances make it ideal for measuring up adaptive defense mechanisms against emerging threats. This dataset is available in the Australian Centre for Cyber Security (ACCS). The Bot-IoT dataset, on the other hand, targets the problem of IoT-based botnet detection by considering both normal and attack traffic with different situations like scanning, and

DoS, and data theft. Being large and diverse, it reflects real IoT network scenarios and correlates well with machine learning operations; also, it has abundant feature-extractive metadata. It is available through purchase through UNSW Canberra.
In addition to that, to enhance model generalization, and avoid cases of over fitting, several preprocessing techniques were performed. Feature selection was done to eliminate the problem of high dimensionality by only using source and destination IP, packet size, packet type, etc., which could slow down computation. The normalization technique was used to make feature magnitude comparable such that features were scaled to the range 0 to 1. Moreover, good synthetic data generation was performed to generate adversarial examples by incorporating perturbation techniques for improving view point in effective handling of adversarial inputs with the model.

3.3 Algorithms

The RL element of the system employs a DQN to create learning systems for constructing adaptive defenses. Operating within a simulated environment modeled as a Markov Decision Process (MDP), the RL model consists of three key elements: statuses, which define the state of the network at a certain point in time, talking about normal traffic and roles, as well as improper actions; actions that span from simple reactions like the blocking of access/restriction and issuing of log entries to the notification of administrators; rewards, aimed at measuring the performance of the defense and stimulating the correct behavior with regards to threats. In such a case, the RL agent takes an action and gets its corresponding reward and keeps learning to counter new threats that may arise. It was carried out in several epochs in a networked training environment.
The Ensemble Methods component is aimed at the anomaly detection and differentiation of the traffic as clean or hostile. The base classifiers like Random Forest, Gradient Boosting were developed on the smaller portion of the dataset used jointly along with the stacked learning which aggregates the output and arrives at the final decision. Ensemble Methods were selected due to its compatibility in decreasing false positives while increasing the ratio of detection. The work also shows that by using a base of diverse classifiers the system stands a less likelihood of falling to different attack types and the misclassification rate is generally low.

4. IMPLEMENTATION AND EVALUATION METRICS

The given system was developed on Python language and TensorFlow and PyTorch for models. Reinforcement learning environment was modeled using OpenAI Gym and Scikit-learn package for ensemble model. These experiments were performed on an NVIDIA RTX 3090 GPU for computing the required computations.
The proposed system was assessed using metrics such as detection rate, false positive rate, accuracy, and response time. The detection rate measures the ratio of correct threat recognition to incidents, false positive rate represents the ratio of benign traffic samples being classified as malicious, and accuracy represents the average classification performance.

4.1 Results

The result section shows the effectiveness of AI-based cybersecurity system that incorporates RL and EMS. This evaluation point consists of four parameters which are the Detection Rate, False Positive Rate, Accuracy and the Response Time. Two modern datasets, TON_IoT and Bot-IoT were used to evaluate the ability of the system under different adversarial attacks. Overall, adding RL and Ensemble Methods showed that both detection accuracy and adaptability improved significantly against baseline models. Table IV outlines the evaluation criterion for both datasets.

TABLE IV: PERFORMANCE METRICS

Dataset	Model	Detection Rate (%)	False Positive Rate (%)	Accuracy (%)	Response Time (ms)
TON_IoT	Baseline (SVM)	87.5	14.2	86.3	120
TON_IoT	RL-Only	92.3	10.8	90.5	100
TON_IoT	RL + Ensemble	96.7	5.2	94.8	105
Bot-IoT	Baseline (SVM)	85.8	16.1	84.2	130
Bot-IoT	RL-Only	91.1	12.3	89.7	110
Bot-IoT	RL + Ensemble	95.5	6.4	93.6	115

4.2 Real-Time Performance

Dataset-Specific Analysis TON_IoT Dataset: With the proposed system DR was at 96.7%, and FPR was minimized to 5.2%. This proves that the developed and proposed system is capable of addressing various forms of IoT based attack scenarios efficiently. These results were made possible by the flexibility of RL and the solidity of Ensemble Methods. Bot-IoT Dataset: Leaving a Detection Rate of 95.5% and False Positive Rate of only 6.4% the system clearly did a good job in detecting such difficult and layered botnet-based attacks. The stacking-based ensemble enhanced the classification accuracy so that the method could operate effectively in situations with a high level of data imbalance.
However, there were some hurdles that are briefly presented below the results denotes that the hybrid model was more effective than the previous models; However, during the course of implementation, some challenges arose. Some drawbacks were observed in the ISIC dataset; the regions of interest in the images heavily overlapped with the background areas, as well as the general impression that samples with a low contrast ratio seem slightly blurred. Similarly, the hybrid

model faced some issues while dealing with herbal data of MRI of BraTS dataset because of the process of decomposition and recompositing. Such constraints can be managed in the next adaptations by altering data augmentation methods and improving composite paradigms for more effective and stable model outcomes.

To determine whether the system would be effective for real time use, response time statistics were measured. The findings show that the proposed RL + Ensemble model kept the response time comparable to the models created during performance enhancement, while using multiple models in the process. Table V gives the response time information on various configurations as mentioned below:

TABLE V : RESPONSE TIME ANALYSIS ACROSS MODELS

Dataset	Model	Average Response Time (ms)	Suitability for Real-Time Applications
TON IoT	Baseline (SVM)	120	Moderate
TON IoT	RL-Only	100	High
TON IoT	RL + Ensemble	105	High
Bot-IoT	Baseline (SVM)	130	Moderate
Bot-IoT	RL-Only	110	High
Bot-IoT	RL + Ensemble	115	High

The RL + Ensemble setup enhanced response time by a negligible time as compared to the RL-only setup but was well within the live application response latency limit. These results confirm the usefulness of this framework for applying it in the conditions of high velocity and constant change.

4.3 Error Analysis

Nevertheless, several difficulties were identified throughout the assessment of the company's performance. The aforementioned challenges are captured in detail in Table VI below.

TABLE VI : ERROR ANALYSIS IN THE PROPOSED SYSTEM

Error Type	Description	Impact	Proposed Solution
Highly Complex Attack Patterns	Attacks mimicking normal traffic (e.g., low-rate DDoS) increased false positives.	Slight increase in false positive rate.	feature extraction.
Imbalanced Dataset	Rare attack classes were underrepresented in training data.	Reduced detection rates for rare attacks.	Weighted loss functions or oversampling rare classes.
Computational Overhead	Combining multiple models in Ensemble Methods added marginal processing delay.	Increased response time by ~10 ms.	Model optimization through pruning or lightweight architectures.

5. COMPARATIVE ANALYSIS

Table 7 shows the performance of the proposed approach along with the existing approaches for the purpose of demonstrating the advantages of the system proposed in this paper.

TABLE VII : COMPARATIVE ANALYSIS WITH EXISTING METHODS

Approach	Detection Rate (%)	False Positive Rate (%)	Response Time (ms)
Rule-Based Systems	75.3	25.8	150
Traditional ML (SVM)	86.0	15.2	120
Proposed (RL + Ensemble)	96.1	5.8	110

The proposed system was found to have significantly better performance than the traditional methods in all of the parameters studied, thus ensuring that the problems experienced with static and solely rule based systems were effectively addressed.

6. CONCLUSION

The study presents an AI-based cybersecurity paradigm shift that uses Recurrent Learning (RL) with Ensemble Methods to protect against adversarial attacks. The system achieves high detection accuracy, adaptability to new threats, and effective anomaly detection. Compared to existing methods and machine learning algorithms, the detection rates exceed 95% with low false positive rates. The RL component changes the threatening response, and Ensemble Methods minimize misclassifications, making the system applicable to modern cybersecurity attacks. The low response time of 105-115 ms demonstrates its potential in dynamic environments with rapid response. The study suggests future research directions include improving feature engineering, incorporating temporal and behavioral feature representations to capture attack variants, proposing nuanced measures for scarce classes, meta-optimization with Explainable AI, optimizing architectures, and applying multimodal analysis to extend the system's capabilities. Implementation in real-world environments, including enterprise networks and IoT settings, will provide insights into the practical usability and scalability of the system. The study's strengths and weaknesses highlight the potential of integrating RL and Ensemble Methods to overcome traditional defense approaches for critical infrastructure and IoT networks. Future research should focus on optimizing architectures, introducing multimodal analysis, and evaluating the system's practical usability and scalability in various real-life settings.

Funding:

The authors confirm that no funding was acquired from any organization, grant agency, or institution. This research was undertaken without any external financial contributions.

Conflicts of Interest:

The authors declare no competing financial interests in this study.

Acknowledgment:

The authors would like to thank their institutions for providing the necessary facilities and guidance, which proved vital in achieving the study's objectives.

References

[1] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.
[2] V. Mnih, K. Kavukcuoglu, D. Silver, *et al.*, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.
[3] J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2015.
[4] Y. Gulzar and S. A. Khan, "Skin Lesion Segmentation Based on Vision Transformers and Convolutional Neural Networks—A Comparative Study," Applied Sciences, vol. 12, no. 12, p. 5990, Jun. 2022. doi: 10.3390/app12125990.
[5] A. Mathew, "Deep Reinforcement Learning for Cybersecurity Applications," International Journal of Computer Science and Mobile Computing, vol. 10, no. 12, pp. 32–38, Dec. 2021, doi: 10.47760/ijcsmc.2021.v10i12.005.
[6] M. A. de Lucas, C. Camilo, and M. V. de Mello, "A Comprehensive Survey on Ensemble Learning-Based Intrusion Detection Approaches in Computer Networks," IEEE Access, vol. 11, 2023. https://ieeexplore.ieee.org/document/10299619/
[7] N. Koroniotis, N. Moustafa, E. Sitnikova, and B. Turnbull, "Towards the Development of Realistic Botnet Dataset in the Internet of Things for Network Forensic Analytics: Bot-IoT Dataset," Future Generation Computer Systems, vol. 100, pp. 779–796, 2019. https://www.sciencedirect.com/science/article/abs/pii/S0167739X18327687.
[8] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2019, pp. 6105–6114.
[9] A. Gueriani, H. Kheddar, and A. C. Mazari, "Deep Reinforcement Learning for Intrusion Detection in IoT: A Survey," IEEE WCNC, 2024 (arXiv 2023). https://arxiv.org/abs/2405.20038
[10] Y. Wang et al., "Adversarial Attacks and Defenses in Machine Learning-Empowered Communication Systems and Networks: A Contemporary Survey," IEEE Communications Surveys & Tutorials, 2023. https://ieeexplore.ieee.org/abstract/document/10263803/
[11] 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), Canberra, Australia, Nov. 2015, pp. 1–6. https://ieeexplore.ieee.org/document/7348942/G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 4700–4708.
[12] C. Xie, J. Wang, and L. Zhang, "Addressing adversarial attacks through adaptive reinforcement learning strategies," *Journal of Information Security and Applications*, vol. 71, p. 103400, 2023.
[13] G. Rjoub et al., "A Survey on Explainable Artificial Intelligence for Cybersecurity," IEEE Transactions on Network and Service Management, 2023. https://arxiv.org/abs/2303.12942
[14] N. Capuano, G. Fenza, V. Loia, and C. Stanzione, "Explainable Artificial Intelligence in CyberSecurity: A Survey," IEEE Access, vol. 10, pp. 93575–93600, 2022. https://ieeexplore.ieee.org/document/9894371
[15] Q. Han et al., "Adversarial Machine Learning for Network Intrusion Detection Systems: A Comprehensive Survey," IEEE Communications Surveys & Tutorials, 2023. https://dl.acm.org/doi/abs/10.1109/COMST.2022.3233793
[16] N. Moustafa and J. Slay, "The Evaluation of Network Anomaly Detection Systems: Statistical Analysis of the UNSW-NB15 Dataset and the Comparison with the KDD99 Dataset," Information Security Journal: A Global Perspective, vol. 25, nos. 1–3, pp. 18–31, 2016. https://www.tandfonline.com/doi/abs/10.1080/19393555.2015.1125974
[17] S. Tharewal, M. W. Ashfaque, S. S. Banu, P. Uma, S. M. Hassen, and M. Shabaz, "Intrusion Detection System for Industrial Internet of Things Based on Deep Reinforcement Learning," Wireless Communications and Mobile Computing, vol. 2022, 2022. https://www.hindawi.com/journals/wcmc/2022/9023719/
[18] T. G. Nguyen, T. V. Phan, D. T. Hoang, T. N. Nguyen, and C. So-In, "Federated Deep Reinforcement Learning for Traffic Monitoring in SDN-Based IoT Networks," IEEE Transactions on Cognitive Communications and Networking, vol. 7, no. 4, pp. 1048–1065, 2021. https://ieeexplore.ieee.org/document/9390353
[19] Y. Otoum and A. Nayak, "AS-IDS: Anomaly and Signature Based IDS for the Internet of Things," Journal of Network and Systems Management, vol. 29, pp. 1–26, 2021. https://link.springer.com/article/10.1007/s10922-021-09589-4
[20] M. Umer, K. N. Junejo, M. T. Jilani, and A. P. Mathur, "Machine Learning for Intrusion Detection in Industrial Control Systems: Applications, Challenges, and Recommendations," International Journal of Critical Infrastructure Protection, vol. 38, p. 100516, 2022. https://www.sciencedirect.com/science/article/abs/pii/S1874548222000257
[21] C. Zhang, X. Costa-Pérez, and P. Patras, "Adversarial Attacks Against Deep Learning-Based Network Intrusion Detection Systems and Defense Mechanisms," IEEE/ACM Transactions on Networking, vol. 30, no. 3, pp. 1294–1311, 2022. https://ieeexplore.ieee.org/document/9695780