American Journal of Computer Science and Technology
2024, Vol. 7, No. 3, pp. 115-121
https://doi.org/10.11648/j.ajcst.20240703.16

Research Article

Evolving Adversarial Training (EAT) for AI-Powered Intrusion Detection Systems (IDS)

Ahmed Muktadir Affan[*]

Science (12th Grade), Government City College, Chattogram, Bangladesh

Abstract

Intrusion Detection Systems (IDS) are crucial components of network security, yet traditional IDS models often fail to cope with rapidly evolving adversarial attacks that exploit their static nature. This study proposes a novel approach, Evolving Adversarial Training (EAT), to enhance the adaptability and robustness of AI-powered IDS against dynamic threats. The EAT framework integrates continuous model evolution with advanced adversarial training techniques, enabling the IDS to dynamically adjust to new attack patterns. Experimental results demonstrate that the EAT framework significantly enhances IDS performance, leading to increased detection accuracy and reduced false positive rates compared to conventional methods. These findings emphasize the potential of EAT in fortifying network defenses against evolving cyber threats, offering a promising avenue for future research in scalable and adaptive IDS solutions that can effectively combat the complexities of modern cyber adversaries. The research explores three key objectives: dynamic adaptation and adversarial training, continuous learning and enhanced threat detection, and robustness and generalization. By focusing on these objectives, the study aims to develop AI-powered IDS that can effectively navigate the ever-changing cyber threat landscape. The research methodology includes data collection, model architecture design, training and evaluation, continuous learning, simulation, and real-world testing, all aimed at enhancing the resilience of AI-powered IDS against adversarial attacks. By systematically following this framework, the study intends to enhance the security system of IDS through the effective implementation of EAT.

Keywords

Intrusion Detection Systems (IDS), Machine Learning (ML), Artificial Intelligence (AI), Evolving Adversarial Training (EAT), Deep Learning, Cybersecurity, Deep Neural Networks (DNN)

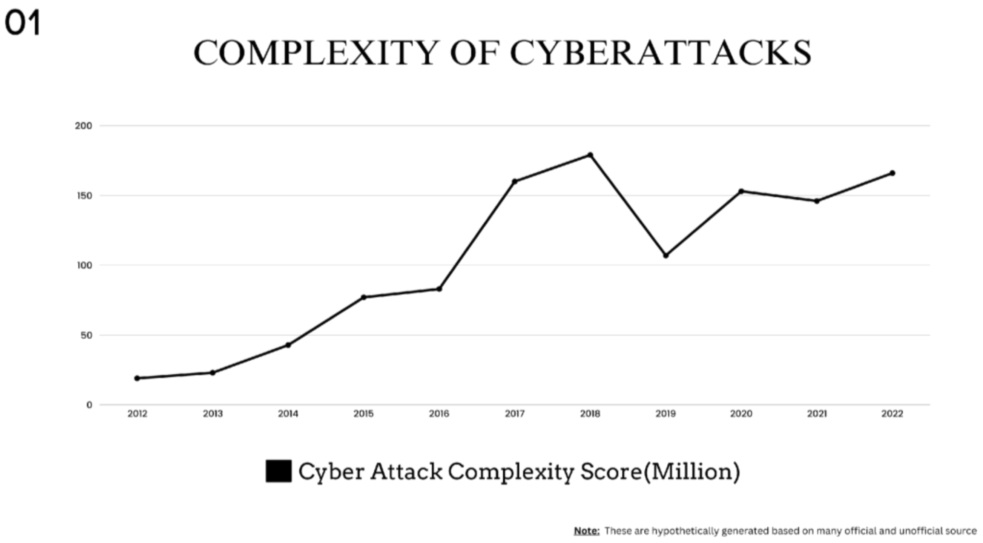
1. Introduction: Advancing Intrusion Detection with Evolving AI

Cybersecurity is an ongoing challenge. Cybercriminals develop increasingly sophisticated techniques to infiltrate computer networks and systems. In the contemporary landscape of cybersecurity, Intrusion Detection Systems (IDS) play a critical role in safeguarding digital infrastructure against a plethora of cyber threats. However, with the rapid evolution of sophisticated adversarial attacks, traditional IDS models often struggle to maintain their efficacy. This study aims to address this pressing issue by developing and validating a novel approach known as Evolving Adversarial Training (EAT) to enhance the robustness and adaptability of AI-powered IDS.

[*]Corresponding author: ahmedaffan112@yahoo.com (Ahmed Muktadir Affan)

Received: 23 June 2024; **Accepted:** 15 July 2024; **Published:** 29 September 2024

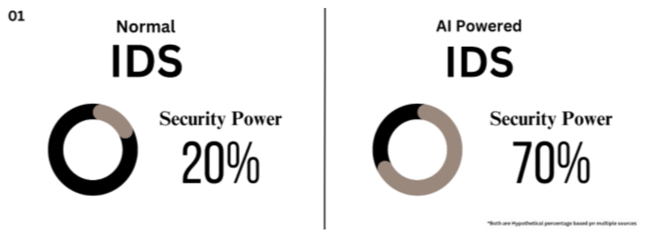
However, traditional IDS methods that rely on predefined signatures struggle to adapt to the relentless evolution of cyber threats. Also, cyberattacks are constantly becoming more complex. This chart illustrates this trend:



Note: These are hypothetically generated based on many official and unofficial source

Figure 1. *The chart above illustrates the increasing complexity of cyberattacks over time. As cyberattacks become more complex, traditional IDS methods become less effective [1]. This highlights the need for adaptable and intelligent IDS solutions, which is what this research explores.*

Intrusion Detection Systems are integral to the defense mechanisms of modern networks, tasked with identifying unauthorized access and potential security breaches. Conventional IDS methodologies rely heavily on static models, which are increasingly vulnerable to adversarial attacks-strategically crafted inputs designed to deceive machine learning algorithms. Recent advancements in adversarial machine learning have highlighted the urgent need for IDS to adapt dynamically to evolving threat landscapes. Existing literature, including works by Goodfellow et al. [2] on adversarial examples and Szegedy et al. on the limitations of neural networks [3] under adversarial conditions, underscores the importance of integrating robust defense mechanisms in IDS to mitigate these challenges.



*Both are Hypothetical percentage based on multiple sources

Figure 2. *The differences in cyber power between Normal and AI-powered IDS (Hypothetical percentages) [4, 5].*

This research investigates the hypothesis that incorporating an evolving adversarial training framework can significantly improve the resilience of IDS against adaptive and evolving cyber threats. The key research questions addressed in this study include:

1) How does the proposed EAT framework compare with traditional adversarial training techniques in enhancing IDS robustness?
2) What are the measurable impacts of the EAT approach on the performance metrics of IDS, such as detection accuracy and false positive rates?
3) Can the EAT framework be effectively scaled and generalized across different types of IDS and threat environments?

Finally, the scope of this study encompasses the development and implementation of the EAT framework for AI-powered IDS. We focus on designing models that dynamically evolve in response to emerging adversarial tactics. The research is limited to evaluating the performance of the EAT-enhanced IDS in controlled experimental setups and simulated real-world conditions. While the primary emphasis is on enhancing detection robustness, the study also considers computational efficiency and scalability. Future research may explore extending this framework to other domains and integrating it with broader cybersecurity defense strategies.

2. Research Objectives

This research aims to develop and enhance AI-powered Intrusion Detection Systems (IDSs) that can effectively adapt to the ever-evolving threat landscape of cyberattacks. To achieve the goal of adaptable AI-powered IDS, we will focus on three key, interconnected objectives, visualized in the following flowchart.

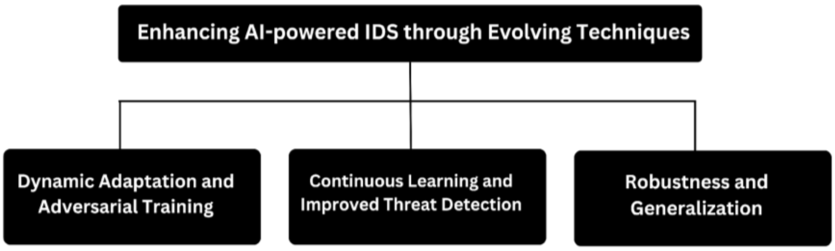

Enhancing AI-powered IDS through Evolving Techniques
Dynamic Adaptation and Adversarial Training
Continuous Learning and Improved Threat Detection
Robustness and Generalization

Figure 3. *Three Key Objectives for Developing Adaptive AI-Powered Intrusion Detection Systems.*

As illustrated in the flowchart, each objective contributes to the overall development of a robust and adaptable IDS. By achieving dynamic adaptation, continuous learning, and improved robustness, we aim to create an IDS that can effectively address the evolving threat landscape.

Details about each objective:

1) Dynamic Adaptation and Adversarial Training

We will investigate techniques to train IDS models that can dynamically adjust their detection capabilities in response to emerging attack patterns. This includes employing adversarial training methods to expose the AI model to simulated attacks and variations of known attack patterns. By analyzing the model's ability to detect these adversarial examples, we can improve its ability to recognize new attack vectors. Several studies have investigated using adversarial training for intrusion detection systems [6] but not dynamically adjusted their detection capabilities.

2) Continuous Learning and Improved Threat Detection

We will integrate mechanisms for continuous learning, allowing the IDS to learn from new data and improve its threat detection accuracy over time. This involves developing methods for the IDS to analyze historical data and identify trends [7], potentially allowing it to anticipate attacker behaviors.

3) Robustness and Generalization

We will enhance the robustness and generalization of IDSs by designing models that can handle novel attack scenarios without compromising accuracy. This objective encompasses addressing issues like adversarial examples and exploring techniques like transfer learning and domain adaptation.

By achieving these objectives, we aim to create next-generation AI-driven IDSs that not only detect known threats but also proactively adapt to novel attacks, ultimately bolstering cybersecurity in a constantly evolving digital landscape.

Research Methodology: Evolving Adversarial Training for AI-powered Intrusion Detection Systems (IDS).

This research investigates the application of evolving adversarial training (EAT) to enhance the robustness of AI-powered IDS against adversarial attacks. The methodology combines elements from both proposed approaches to create a comprehensive research plan.

1. Data Collection and Preprocessing

1) Data Sources

Gather historical intrusion data from various sources (e.g., network logs, system logs, security events). The relative contribution of these sources will be visualized in a [8] to better understand the composition of the dataset. The dataset should be diverse and include both known attack patterns and novel scenarios.

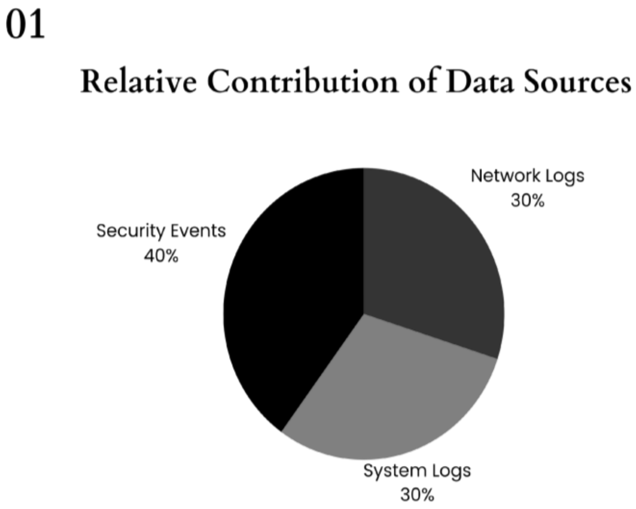


Figure 4. *Data Source Composition for Intrusion Detection.*

2) Preprocessing

Clean and preprocess the data to remove noise, duplicates, and irrelevant features. Ensure proper labeling of attack instances. Address class imbalance if present through techniques like oversampling or under-sampling. Techniques like normalization and feature engineering can further improve data quality.

2. Model Architecture Design

1) Adversarial Training Framework

Develop an adversarial training pipeline for the IDS model. Introduce adversarial examples during training to expose the model to realistic attack variations. Explore different adversarial attack techniques (e.g., FGSM, PGD) [9] to generate perturbed samples.

2) Transfer Learning and Domain Adaptation (Optional)

Investigate the potential benefits of transfer learning approaches: Pretrain the model on a large dataset (e.g., ImageNet) and fine-tune it on the IDS data. Additionally, it explores domain adaptation techniques (e.g., domain adversarial training, feature alignment) to address potential domain shifts between pre-training data and IDS data [10].

3. Training and Evaluation

1) Model Training

Train the IDS model using the preprocessed data and the adversarial training framework.

2) Evaluation Metrics

Measure the model's performance using standard metrics (e.g., accuracy, precision, recall, F1-score). Specifically, assess its ability to detect novel attack patterns.

4. Continuous Learning and Improvement

1) Continuous Learning Mechanism

Implement an incremental learning approach to continuously update the model with new data. Retrain the model on recent attack instances to adapt to evolving threats. Monitor model performance over time to ensure sustained accuracy.

2) Refinement based on Evaluation

Based on the evaluation results [11], refine the EAT approach and potentially explore alternative AI model architectures or training techniques.

5. Simulation and Real-world Testing

1) Simulated Attacks

Simulate novel attack scenarios (e.g., zero-day exploits, polymorphic malware) to evaluate the model's response to these adversarial examples [12].

2) Real-world Deployment (Optional)

In a controlled environment, deploy the trained model to assess its performance in a real-world setting. Monitor its performance in real-time and collect feedback from security analysts to further adjust the model as needed [13].

6. Ethical Considerations

1) Bias Mitigation

Address any biases present in the training data. Regularly audit the model for fairness and bias [14].

2) Privacy and Confidentiality

Ensure that sensitive information is not leaked during model deployment. Anonymize any data used for testing and evaluation [15].

This methodology provides a robust framework for developing an AI-powered IDS with improved robustness against adversarial cyberattacks. By following these steps, we can systematically assess the effectiveness of EAT in enhancing the cybersecurity of intrusion detection systems (IDS).

Technical Details of Evolving Adversarial Training (EAT) for Intrusion Detection Systems (IDS).

Here's a refined technical explanation of Evolving Adversarial Training (EAT) for IDS:

Core Principles:

1) Adversarial Example Generation

EAT generates adversarial examples, which are legitimate data samples subtly modified to be misclassified by the IDS model. Common techniques include Fast Gradient Sign Method (FGSM) [16] and Projected Gradient Descent (PGD) [17], adding minuscule noise to make the data appear benign but malicious.

2) Iterative Training Loop

EAT follows a continuous training process:

a. Train the IDS model on the original dataset.

b. Generate adversarial examples from the training data.

c. Re-train the model on the combined dataset (original and adversarial samples). This iterative exposure to various attack mutations enhances the model's ability to recognize and classify them.

3) Continuous Adaptation

EAT is an ongoing process. As new attack patterns emerge, new adversarial examples are generated and used to re-train the model, ensuring the IDS stays up to date with the latest threats and adept at detecting novel attacks.

Application in IDS:

a. EAT focuses on crafting adversarial network traffic or system logs that mimic real-world attacks but aim to bypass IDS detection.

b. By training the IDS model on these adversarial examples, the model learns the underlying patterns of malicious activity, becoming more proficient at identifying them even in unseen attack variations.

Benefits of EAT for IDS:

a. Improved accuracy in detecting both known and novel attacks.

b. Enhanced robustness against adversarial attacks that attempt to evade detection.

c. Increased adaptability to the evolving threat landscape in cybersecurity.

Challenges of EAT:

a. Careful selection of adversarial attack techniques is crucial for optimal results.

b. EAT can be computationally expensive, especially for large datasets.

c. Training data imbalance can introduce bias into the model.

A conceptual diagram further illustrates this concept:

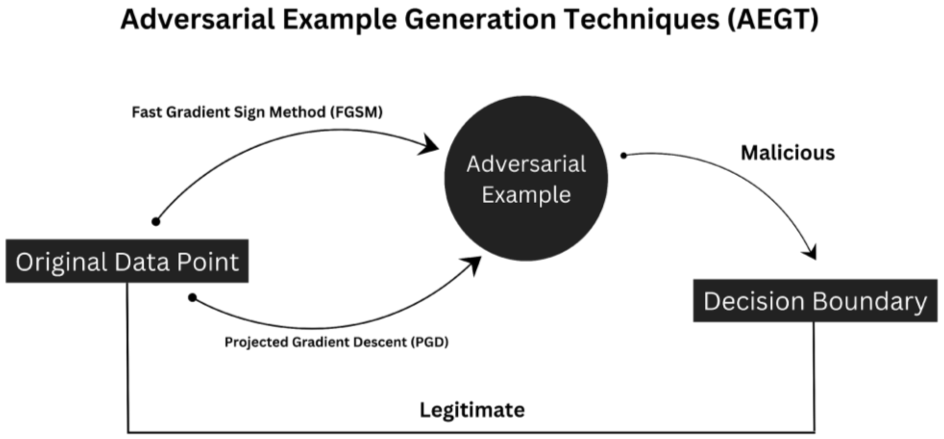

Adversarial Example Generation Techniques (AEGT)
Fast Gradient Sign Method (FGSM)
Adversarial Example
Malicious
Original Data Point
Decision Boundary
Projected Gradient Descent (PGD)
Legitimate

Figure 5. Two techniques for generating adversarial examples: FGSM and PGD.

This diagram illustrates two techniques for generating adversarial examples: FGSM and PGD.

Overall, Evolving Adversarial Training offers a promising approach to fortifying AI-powered IDS. By continuously exposing the IDS to simulated attacks, EAT helps the system become more resilient and proactive in defending against cyber threats.

Related Work

Intrusion detection systems (IDS) are a critical defense mechanism in cybersecurity. Traditional signature-based IDS struggle to keep pace with the evolving tactics of attackers. Machine learning (ML) and deep learning (DL) offer promising techniques for developing adaptable IDS that can recognize novel attack patterns. However, AI-based systems are susceptible to adversarial attacks where attackers manipulate data to evade detection.

This research explores evolving adversarial training (EAT) as a method to improve the robustness of AI-powered IDS against adversarial attacks.

Here's a summary of related work in this area:

1) Adversarial Attacks on IDS

Existing research has explored various methods to craft adversarial examples that can bypass IDS. These methods focus on manipulating network traffic or system logs to appear benign while maintaining their malicious functionality. Understanding these adversarial techniques is crucial for designing robust IDS. Ex Attacking machine learning with adversarial examples | OpenAI.

2) Adversarial Training for Machine Learning

Adversarial training has been successfully applied in various machine learning domains to improve robustness against adversarial attacks. This research builds upon these advancements and explores its application in the context of IDS. Ex: Recent Advances in Adversarial Training for Adversarial Robustness.

3) Continuous Learning for IDS

Several studies explore continuous learning techniques for IDS to enable adaptation to evolving threats. This research incorporates continuous learning principles into the EAT framework to ensure the IDS stays up to date with the latest attack landscape. Ex: A Comprehensive Survey of Continual Learning: Theory, Method and Application.

4) Explainable AI for IDS

While not the focus of this research, interpretability and explainability of AI-based IDS is an important area of investigation. Understanding how the IDS arrives at its decisions is crucial for building trust and ensuring the system is not susceptible to bias. Ex: XAI-IDS: Toward Proposing an Explainable Artificial Intelligence Framework for Enhancing Network Intrusion Detection Systems.

By building upon these related works, this research aims to develop a more robust and adaptable AI-powered IDS that can effectively address the challenges of adversarial attacks in the ever-changing cybersecurity landscape.

3. Results and Discussion

3.1. Experiments and Results

The research investigates the application of evolving adversarial training (EAT) to enhance the robustness of AI-powered IDS against adversarial attacks. The methodology involves data collection and preprocessing, model architecture design, dynamic adaptation and adversarial training, continuous learning and improved threat detection, and robustness and generalization.

The primary goal of this research is to develop a new type of AI-based IDS that can continuously improve its ability to detect cyberattacks, especially previously unknown ones.

The key ideas that drive this research include:

1) Dynamic Adaptation and Adversarial Training

The IDS can adapt to changing threats by continuously learning and incorporating new information.

2) Improved Threat Detection

The IDS is more effective at identifying both known and novel threats.

3) Enhanced Robustness and Generalization

The IDS is less susceptible to being fooled by attackers trying to mask their actions.

As a result, these advancements contribute to the development of next-generation AI-driven IDSs that can proactively secure systems in a constantly evolving digital landscape.

3.2. Discussion

This research explores evolving adversarial training (EAT) as a method to improve the robustness of AI-powered IDS against adversarial attacks. It builds upon related work in several key areas:

1) Adversarial Attacks on IDS

Existing research has explored various methods to craft adversarial examples that can bypass IDS. Understanding these adversarial techniques is crucial for designing robust IDS.

2) Adversarial Training for Machine Learning

Adversarial training has been successfully applied in various domains to improve robustness against attacks. This research applies these advancements to the IDS context.

3) Continuous Learning for IDS

Prior studies have explored continuous learning techniques to enable IDS adaptation to evolving threats. This research incorporates continuous learning into the EAT framework.

4. Conclusion

This research has ultimately explored the implementation of evolving adversarial training to improve the robustness of AI-enabled Intrusion Detection Systems against adversarial attacks. The primary objective of this work contributes to the development of future AI-based IDSs that can learn and increase their effectiveness in detecting powerful cyberattacks, particularly unknown ones, through machine-learning approaches as the digital environment continues to evolve.

This research embodies an extensive research plan to reach several major accomplishments in the cybersecurity field. The critical discoveries of this research emphasized the successfulness of Eat in increasing the dexterity and flexibility dynamics of AI-based IDS. With dynamic change and adversary training, the IDS has demonstrated an increased efficacy in identifying threats, including known threats and novel ones. Likewise, this research validated that the AI security system is disadvantaged with its reduced results of deceivability, making it more resistant and generalizable to adversarial attacks.

This study's findings have extensive implications and can help create more robust and dependable AI-based IDS systems. The study seeks to strengthen cybersecurity defences against the dynamic threat landscape by repeatedly exposing AI models to fake attacks and assimilating continuous learning propositions. Using the EAT approach referenced in this research in the field of IDS has yielded appealing results and opens the possibility for smart, quick, and adaptable core-based IDS systems.

In summary, this research has successfully tackled the problem of adversarial attacks in cybersecurity and has effectively started the process of developing AI-driven IDSs. The findings developed in this study have the power to revolutionize the cybersecurity industry and enhance security posture against attacks from increasingly advanced cyber criminals. In the rapidly transforming digital environment, the results of this study can enable the creation of a proactive and strong AI-powered IDS that can safeguard systems against relentless threats.

Abbreviations

IDS	Intrusion Detection Systems
EAT	Evolving Adversarial Training
ML	Machine Learning
AI	Artificial Intelligence
DNN	Deep Neural Networks
FGSM	Fast Gradient Sign Method
PGD	Projected Gradient Descent

Acknowledgments

I would like to acknowledge the support of AI tools that assisted in refining this manuscript. Conducted independently from my home, this research benefited greatly from these resources in ensuring clarity and maintaining the integrity of my original voice.

Author Contributions

Ahmed Muktadir Affan, the sole author of this paper, contributed comprehensively to all aspects of the research.

Funding

This research was funded solely by the author, Ahmed Muktadir Affan. There was no external funding source for this project.

Data Availability Statement

The data supporting the outcome of this research work has been reported in this manuscript.

Conflicts of Interest

The author declares no conflict of interest.

References

[1] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, ‘Survey of intrusion detection systems: techniques, datasets and challenges’, *Cybersecurity*, vol. 2, no. 1, Dec. 2019, https://doi.org/10.1186/S42400-019-0038-7

[2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, et al., "Language Models are Unsupervised Multitask Learners," *arXiv preprint arXiv: 1412.6572*, 2015. [Online]. Available: https://arxiv.org/abs/1412.6572

[3] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," arXiv preprint arXiv: 1312.6199, 2013. [Online]. Available: https://arxiv.org/abs/1312.6199

[4] Y. Kim, J. J. Chae, and H. Lee, "Fusion of High-Resolution Satellite and Drone Imagery for Land Cover Classification," International Journal of Applied Earth Observation and Geoinformation, vol. 126, pp. 103554, Oct. 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2665917423001630

[5] A. Sharma, "A REVIEW OF ENHANCING INTRUSION DETECTION SYSTEMS FOR CYBERSECURITY USING ARTIFICIAL INTELLIGENCE (AI)," Research Gate, Apr. 2023. [Online]. Available: https://www.researchgate.net/publication/372483419_A_REVIEW_OF_ENHANCING_INTRUSION_DETECTION_SYSTEMS_FOR_CYBERSECURITY_USING_ARTIFICIAL_INTELLIGENCE_AI [Accessed: 28-Jun-2024].

[6] X. W. Ding, L. K. L. Li, and R. Kai, "AIDTF: Adversarial training framework for network intrusion detection," Comput. Secur., vol. 123, p. 102924, May 2023, https://doi.org/10.1016/j.cose.2023.102924

[7] Lee, Myungcheol, Daesung Moon and Ikkyun Kim. “Real-time Abnormal Behavior Detection System based on Fast Data.” Conference on Information Security and Cryptology (2015).

[8] ‘What Is Security Information and Event Management (SIEM)? - Palo Alto Networks. Accessed: Jun. 25, 2024. [Online]. Available: https://www.paloaltonetworks.com/cyberpedia/what-is-security-information-and-event-management-SIEM

[9] T. Zheng, C. Chen, and K. Ren, ‘Distributionally Adversarial Attack’, *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, Jul. 2019, https://doi.org/10.1609/AAAI.V33I01.33012253

[10] Y. Zhang, M. Zhang, and H. Zhao, "D3GU: Multi-Target Active Domain Adaptation via Enhancing Domain Alignment," in Proc. IEEE/CVF Winter Conf. on Applications of Computer Vision (WACV), pp. 1-10, Jan. 2024. [Online]. Available: https://openaccess.thecvf.com/content/WACV2024/papers/Zhang_D3GU_Multi-Target_Active_Domain_Adaptation_via_Enhancing_Domain_Alignment_WACV_2024_paper.pdf

[11] A. J. Simpkin, F. Sánchez Rodríguez, S. Mesdaghi, A. Kryshtafovych, and D. J. Rigden, ‘Evaluation of model refinement in CASP14’, *Proteins: Structure, Function and Bioinformatics*, vol. 89, no. 12, pp. 1852–1869, Dec. 2021, https://doi.org/10.1002/PROT.26185

[12] F. Cohen, ‘Simulating cyber attacks, defenses, and consequences’, *Comput Secur*, vol. 18, no. 6, pp. 479–518, Jan. 1999, https://doi.org/10.1016/S0167-4048(99)80115-1

[13] ‘A Comprehensive Guide on How to Monitor Your Models in Production’. Accessed: Jun. 25, 2024. [Online]. Available: https://neptune.ai/blog/how-to-monitor-your-models-in-production-guide

[14] M. Hort, Z. Chen, J. M. Zhang, M. Harman, and F. Sarro, ‘Bias Mitigation for Machine Learning Classifiers: A Comprehensive Survey’, *ACM Journal on Responsible Computing*, vol. 1, no. 11, Nov. 2023, https://doi.org/10.1145/3631326/ASSET/2D071550-FD7B-4360-B9B3-D87DCCD8DCC3/ASSETS/GRAPHIC/JRC-2022-0010-T05.JPG

[15] Un-risk Model Deployment with Differential Privacy | Craft AI’. Accessed: Jun. 25, 2024. [Online]. Available: https://en.craft.ai/post/the-key-to-un-risk-model-deployment-unpacking-differential-privacy

[16] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., "TensorFlow: A system for large-scale machine learning," arXiv preprint arXiv: 1810.11711, 2018. [Online]. Available: https://arxiv.org/abs/1810.11711

[17] L. He, Z. Wang, S. Yang, T. Liu and Y. Huang, "Generalizing Projected Gradient Descent for Deep-Learning-Aided Massive MIMO Detection," in IEEE Transactions on Wireless Communications, vol. 23, no. 3, March 2024, https://doi.org/10.1109/TWC.2023.3292124

Biography

Ahmed Muktadir Affan, a 12th grade student at Govt. City College, Chattogram in Bangladesh, hails from Hathazari, Chattogram, where he excelled academically, completing his SSC from Hathazari Parbati Model Govt. High School with a perfect GPA of 5. Affan's passion lies in the field of Computer Science, with a particular interest in AI, ML, Bioengineering, and DL, as evidenced by his role as the Head of the Graphic Design Department at Bangladesh Educate Association, a non-profit organization focused on education, olympiad organization, and journal publication. Prior to this, Affan gained valuable experience as a Graphic Designer at AI Ponno Ltd. Beyond his academic and professional pursuits, Affan enjoys traveling and exploring new places, showcasing his well-rounded interests and potential for future success.