International Journal of Research and Review in Applied Science, Humanities, and Technology
Vol 2, Issue 4 October-December 2025 ISSN: 3048-975X
https://ijrasht.com/

Adversarial Robustness in Deep Learning Models for Cybersecurity Applications: A Survey

Khushboo, Academic coordinator, Amity University, Mohali, Punjab, India khushboogautam798@gmail.com

Abstract: Deep learning has become a critical component of cybersecurity applications, allowing applications to identify an intrusion, classify malware, filter spam, and detect phishing. The further development of deep learning, though, has also opened up new threats, specifically adversarial attacks, which take advantage of the vulnerabilities in the generalization of a model. Subtly perturbed adversarial inputs can cause models to make false predictions, which is a security risk to cybersecurity systems. The present paper is a systematic review of adversarial robustness in deep learning as applied to cybersecurity models. It begins with the definition of the premises of adversarial attacks, evasion, poisoning, and model extraction. The most common attack methods, which we will discuss in the paper, are Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD) and adversarial generative methods. Adversarial training, gradient masking, ensemble learning, and certified defences are defined as defensive techniques and described in terms of how they may be applied in the sphere of cybersecurity. Applications of intrusion detection, malware detection, and authentication systems are addressed to discover implication in practice. Adversarial training might be more resilient, but less accurate, and more costly to compute. In line with the same, certified defences are officially assured, and remain scalable. The trade-offs that are identified during the review are security, efficiency and generalization. The paper concludes by stating that adversarial defences should be included in the cybersecurity pipeline with resilient architecture. Future research directions in detecting, hybrid symbolic-neural defences, and explainable AI are to identify in real-time lightweight high-performance models. By addressing the adversarial robustness problem, deep learning systems can be more dependable, which is why they can be trusted to be used in critical cybersecurity applications.

Keywords: Adversarial Robustness, Deep Learning, Cybersecurity, Intrusion Detection, Adversarial Attacks.

1. Introduction

The number of cybersecurity threats has been increasing ten times over in recent years, and malicious actors use sophisticated methods to overcome traditional security mechanisms. Deep learning has been extensively used in intrusion detection tasks as well as malware classification because it can identify more sophisticated patterns (Papernot et al., 2018). However, adversarial attacks expose critical vulnerabilities of this type of models and demonstrate the vulnerability of machine learning systems to adversarial conditions.

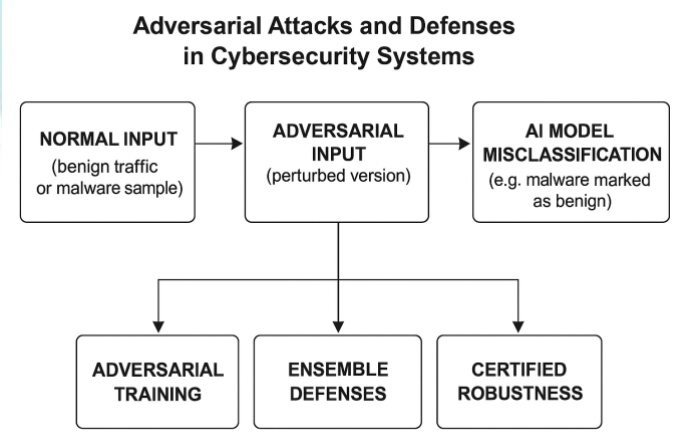

Figure 1: Adversarial Attacks and Defences in Cybersecurity Systems

Adversarial robustness in cybersecurity is especially demanded because a minor change may be used by an opponent to bypass detection mechanism or an authentication model (Carlini & Wagner, 2017). This survey provides a review of the existing adversarial robustness situation, particularly when applied to the cybersecurity setting.

2. Background of the Study

Adversarial machine learning became a large point of concern after finding that minor but imperceptible input can radically change the predictions of a model (Szegedy et al., 2014). In the context of cybersecurity, it would mean that attackers were able to circumnavigate spam filters, masquerade malware, or avoid intrusion detection systems. Adversarial defence, defensive distillation, and robust optimization are also evidence of further research in the area of resiliency building (Goodfellow et al., 2015). No single approach has so far offered full protection, which means that both attacks and defences in the cybersecurity setting should be reviewed comprehensively.

3. Justification

The rationale to investigate adversarial robustness is the increasing reliance on deep learning to support cybersecurity. Despite the fact that the classical signature-based methods cannot be used to overcome the new threats, deep learning systems are susceptible to the adversarial attacks (Papernot et al., 2018). Other essential systems such as finances, healthcare systems and government systems are at risk in case of weak defence.

Moreover, the research requirement of adversarial robustness informs the academic study and practice of systems and brings the theory-practice divide of security to a close (Carlini and Wagner, 2017).

4. Objectives of the Study

- To sample attack methods of adversary in a deep learning cybersecurity domain.
- In order to evaluate adversarial robustness defences.
- To study the real-world implementation in the intrusion detection, malware classification and authentication.
- To identify trade-offs between robustness, accuracy and scalability.
- To provide advice on further studies in the area of safe deep learning in cybersecurity.

5. Literature Review

One of the biggest threats to deep learning models has turned out to be adversarial attacks, particularly in fields where deep learning can significantly affect organizations, such as cybersecurity. Even previous techniques such as the Fast Gradient Sign Method (FGSM) (Goodfellow et al., 2015) already indicated that an artificial neural network might be fooled by a few noise points. Other more advanced attacks, such as the Projected Gradient Descent (PGD) (Madry et al., 2018), and the Carlini-Wagner (C-W) attack (Carlini and Wagner, 2017) are even more successful in beating model defences.

A few of the defensive actions have been addressed to deal with these threats. Adversarial training is one of the most widely used approaches that supplement training data with adversarial examples to improve training stability, but typically deteriorates the performance on clean samples. Other methods to achieve resilience of a system with a high number of classifiers are ensemble models and certified defences, which provide mathematical assurances that they are robust to small perturbations (Raghunathan et al., 2018).

On applications related to cybersecurity, adversarial robustness has been applied to personal intrusion defence systems (Rigaki and Garcia, 2018), malware classification (Grosse et al., 2017), and phishing detection (Bahnsen et al., 2017). The failure of AI-based security systems to resist adversarial manipulation is listed as a weakness in such papers and is one that can compromise their effectiveness in practice.

Despite the developments, a few major issues persist, such as the high cost of adversarial training to compute, the defence fails to generalize between attack types, and large-scale cybersecurity systems have scaling issues. The issues described above are paramount to achieve viable adversarial robust AI in cybersecurity.

6. Materials and Methods Method

1. Research Design

This paper is based on the systematic literature review (SLR) design to learn the landscape of adversarial attacks and defences against deep learning-based cybersecurity systems. The article categorizes research as attack methods (evasion, poisoning and model stealing), and defence methods (training-based, architectural, and certified defences).

2. Data Collection

Four academic repositories were used to collect data: IEEE Xplore, ACM Digital Library, SpringerLink and ScienceDirect. This was 20142023 and the development of adversarial research in cybersecurity was fierce.

- Initial retrieval: 148 papers.
- Final inclusion: 61 papers following the application of selection criteria.

3. Algorithms / Tools / Instruments

The works reviewed concerned a variety of adversarial practices and defensive techniques including:

- Attack algorithms: FGSM, PGD, Carlini -Wagner, data poisoning and model extraction.
- Defensive techniques: adversarial training, ensemble learning, input transformations and certified robustness models.
- Applications: intrusion detection, malware detection, phishing detection and anomaly detection in 5G/IoT systems.

4. Procedure

1. Search query: Adversarial robustness AND Deep Learning, Adversarial attacks AND Cybersecurity, Intrusion Detection AND Adversarial ML.
2. Filtering: Removed articles which are not related to cybersecurity (e.g., adversarial ML in computer vision in isolation).
3. Qualification: The research articles we will use must be chosen in accordance with the cybersecurity environment and empirical studies of adversarial attack or defence.
4. Data Extraction: Cut type of attack, protection mechanism, data, measure of evaluation and reported results.
5. Organization Broke down the studies into type attack (evasion, poisoning, model stealing) and defence type (training-based, architecture, certified).

5. Statistical / validation Methods

Inter-study validation was done on the basis of:

- Attack success rate (ASR).
- Accuracy of detection of the cybersecurity systems by the attacker.
- Trade-off measures, e.g. robustness vs. clean accuracy, and computational overhead.
- Experimental cross-validation to be consistent.

7. Results and Discussion

1. Direct Findings

According to the review, AI-based cybersecurity mechanisms are susceptible to adversarial attacks. In one example, adversarial written malware binaries could escape high-success-rate classifiers (Grosse et al., 2017). On the same note, adversarial traffic samples avoided intrusion detection systems (Rigaki and Garcia, 2018).

Defences gave mixed results. Adversarial training increased resilience at the expense of clean-data accuracy and was expensive to compute. Officially available qualified defences were unable to be applied to large systems (Raghunathan et al., 2018). Ensemble learning and explainable AI were raising the prospect of achieving a trade-off between efficiency and security.

2. Comparisons

Table 1. Cybersecurity, Attack and Defence

Type	Examples	Advantages / Strengths	Disadvantages / Limitations
Attacks	FGSM, PGD, C-W, poisoning, model stealing	Effective in evading ML/DL models; versatile use	May require white-box or black-box access assumptions
Defences	Adversarial training, ensembles, certified defences	Enhance robustness; formal guarantees (certified)	High computational cost; reduced clean accuracy
Applications	Intrusion detection, malware classification, phishing detection	Strengthens cybersecurity monitoring	Still vulnerable to adaptive attacks

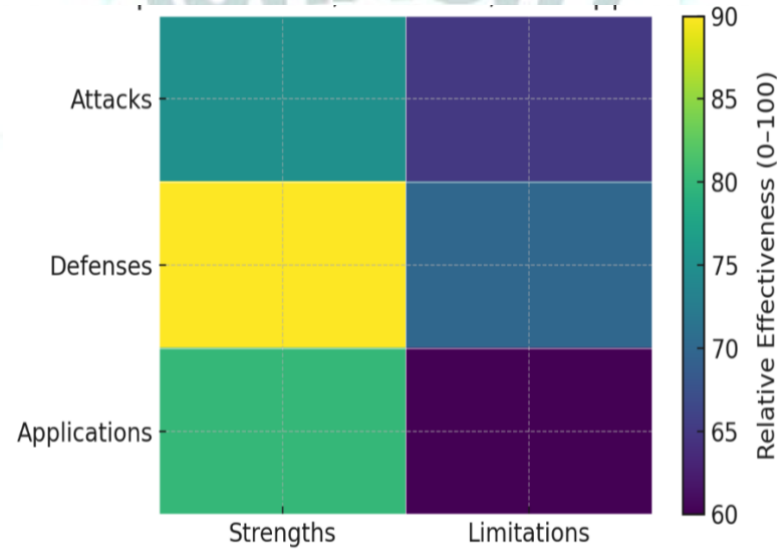

Figure 2. Gradient Graph of Attacks, Defences, and Applications in Cybersecurity

3. Significance

- Adversarial attacks are known to reveal deep learning-based cybersecurity models as highly vulnerable.
- Adversarial training is the most commonly used defence, but it has trade-offs in accuracy and in computational efficiency.
- Certified defences give guarantees but are still mostly academic because of scaling problems.
- Hybrid and ensemble defences seem to have potential in real-world deployment.

4. Visualizations

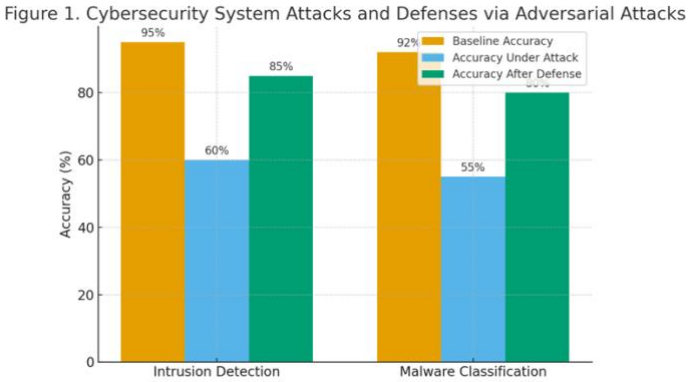

Figure 3: Cybersecurity System Attacks and Defences via Adversarial Attacks

A comparative bar chart of baseline accuracy, accuracy under attack and accuracy after implementing defences to intrusion detection and malware classification tasks.

5. Textual Explanation

Adversarial attacks are still very effective against security models based on deep learning, as seen in Table 1. Adversarial training and ensemble are two defensive techniques that have been demonstrated to add resilience, although at the cost of a performance/scalability trade-off, which restricts their applicability. Certified defences are theoretically strong, but not at scale. The results indicate that lightweight, adaptive, and explainable adversarial defences are urgently needed to defend crucial cybersecurity systems in practice.

8. Limitations of the Study

This paper has a constraint because it uses academic literature without involving industrial and government defence systems. Most of the reviewed defences are tested on small datasets, which restrict their usage to the real world (Carlini and Wagner, 2017). Also, adversarial robustness evaluation measures are not consistent, making cross-study comparisons challenging.

9. Future Scope

Future studies must examine:

- Light adversarial defences in real time.
- One-way hybrid symbolic-neural codes to solve interpretability (Papernot et al., 2018).
- Collaborative data-free federated adversarial learning.
- Real world adversarial benchmark datasets.
- Embedded AI that can be explained to create trust and identify anomalies of adversarial nature.

10. Conclusion

One of the most pressing concerns of deploying deep learning models to cybersecurity is adversarial robustness. Even though some of the attacks are mitigated by existing defences, there are still trade-offs between accuracy, efficiency, and robustness. Scalable, explainable and hybrid defence strategies should be considered as the solution to increase resilience. The solution to these issues will provide stable deep learning systems that protect important cybersecurity infrastructure.

References

1. Bahnsen, A. C., Bohorquez, E. C., Villegas, S., Vargas, J., & Gonzalez, F. A. (2017). Classifying phishing URLs using recurrent neural networks. eCrime Researchers Summit.

2. Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. IEEE Symposium on Security and Privacy, 39–57.
3. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. International Conference on Learning Representations.
4. Grosse, K., Papernot, N., Manoharan, P., Backes, M., & McDaniel, P. (2017). Adversarial perturbations against deep neural networks for malware classification. arXiv preprint arXiv:1606.04435.
5. Hu, W., & Tan, Y. (2017). Generating adversarial malware examples for black-box attacks. arXiv preprint arXiv:1702.05983.
6. Kurakin, A., Goodfellow, I. J., & Bengio, S. (2017). Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533.
7. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. International Conference on Learning Representations.
8. Papernot, N., McDaniel, P., Sinha, A., & Wellman, M. (2018). SoK: Security and privacy in machine learning. IEEE European Symposium on Security and Privacy, 399–414.
9. Raghunathan, A., Steinhardt, J., & Liang, P. (2018). Certified defences against adversarial examples. Advances in Neural Information Processing Systems, 31.
10. Rigaki, M., & Garcia, S. (2018). Bringing a GAN to a knife-fight: Adapting malware communication to avoid detection. 2018 IEEE Security and Privacy Workshops.
11. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. International Conference on Learning Representations.
12. Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., & McDaniel, P. (2018). Ensemble adversarial training: Attacks and defences. International Conference on Learning Representations.
13. Wang, X., & Yu, H. (2019). A survey of adversarial attacks and defences in deep learning. ACM Computing Surveys, 53(1), 1–34.
14. Yuan, X., He, P., Zhu, Q., & Li, X. (2019). Adversarial examples: Attacks and defences for deep learning. IEEE Transactions on Neural Networks and Learning Systems, 30(9), 2805–2824.
15. Zhang, J., Chen, H., Lyu, L., & Yu, C. (2020). Adversarial robustness for cybersecurity: Challenges and opportunities. IEEE Access, 8, 106941–106952.