International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 10 October 2025
Page No: 3966-3972
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0579

AI-Driven Adversarial Attacks & Defenses in Network Security

Swathi M[1], Soumya M S[2], Rajalakshmi N[3], Manjunatha S[4]
[1,4] Assistant Professor, Dept. of CSE, Government Engineering College, Challakere, Karnataka, India
[2,3] Assistant Professor, Dept. of AIML, Government Engineering College, Challakere, Karnataka, India
Email ID: *swathihima98@gmail.com[1], soumyasp66@gmail.com[2], rajalakshmi.1666@gmail.com[3], manjunathas443@gmail.com[4]*

Abstract
In this survey, we provide a comprehensive overview of the recent advancements in adversarial attacks and defenses in the field of machine learning and deep neural networks. We analyze diverse attack techniques, including constrained optimization and gradient-based approaches, and their applications under different threat models such as white-box, gray-box, and black-box settings. The survey also reviews state-of-the-art defense strategies, ranging from adversarial detection methods to robustness improvement techniques, including regularization, data augmentation, and structure optimization. Additionally, the phenomenon of adversarial transferability has been examined, offering deeper insights into the vulnerabilities of deep learning models. In this study, we present a comparative analysis of classical machine learning algorithms, including RF and SVM, alongside deep learning architectures CNNs and RNNs, under adversarial attack scenarios. Experiments were conducted on benchmark intrusion detection datasets, including NSL-KDD and CICIDS2017, which provide diverse traffic patterns and realistic attack vectors. The results demonstrate that while CNN and RNN models achieved the highest baseline accuracies of 95–98% on clean datasets, their performance degraded sharply to nearly 50–60% under adversarial perturbations such as FGSM and PGD attacks. Similarly, traditional models like Random Forest and SVM showed accuracy drops from 90–95% to 60–70%. To address these challenges, defense mechanisms such as adversarial training, ensemble learning, and autoencoder-based anomaly detection were evaluated, restoring accuracy to above 85–90% across different models. This work highlights the dual role of adversarial learning in exposing vulnerabilities and guiding the design of resilient IDS frameworks.

Keywords: *Soil Type, pH, Nutrient Levels (N, P, K), Irrigation Practices, Rainfall, Temperature, Machine Learning, Linear Regression, Random Forest.*

1. Introduction
The rapid growth of digital communication, cloud computing, and IoT has led to an exponential increase in cyber threats targeting network infrastructures. To address these challenges, ML and DL techniques were widely adopted in the development of Intrusion Detection Systems (IDS) due to their ability to automatically learn patterns of normal and malicious traffic. Classical ML models like RF & SVM, along with deep architectures like CNNs and Recurrent Neural Networks (RNNs), have shown high accuracy on benchmark datasets like NSL-KDD, CICIDS2017, and UNSW-NB15, often exceeding 90–95% detection rates under normal the conditions. However, recent research has revealed that these models are highly vulnerable to adversarial attacks, where carefully crafted perturbations in input traffic can cause IDS models to misclassify malicious activities as benign. Attack strategies such as evasion attacks, poisoning attacks, and model extraction exploit the inherent weaknesses of learning algorithms. Such attacks can significantly reduce detection accuracy, in some cases from over 95% to below 60%, thereby compromising the reliability of security systems. To mitigate these threats, researchers have proposed various defense mechanisms, including adversarial training, defensive distillation, ensemble learning, and

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 10 October 2025
Page No: 3966-3972
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0579

autoencoder-based anomaly detection, which aim to restore robustness without sacrificing performance on clean traffic. There is a pressing requirement for more adaptable and universal defensive measures, because hostile instances may be transferred between models. This study offers complete exploration of AI-driven adversarial attacks and defenses in network security, presenting both the offensive perspective of how ML/DL models can be deceived and the defensive strategies designed to enhance resilience. By evaluating models upon standard IDS datasets like NSL-KDD and CICIDS2017, this work contributes to a deeper understanding of the vulnerabilities in current systems and offers a roadmap for developing secure, trustworthy, and robust intrusion detection frameworks in the era of evolving cyber threats. [1-3]

2. Literature Survey

Hussain et al. (2020) investigated adversarial vulnerabilities in IoT-based IDS models. They analyzed how adversarial examples crafted for one model could be transferred to others, proving the transferability of attacks across Random Forest, SVM, and CNN models. Using UNSW-NB15 dataset, they demonstrated that adversarial perturbations reduced detection accuracy by more than 30%, raising concerns for IoT security. Alvarez et al. (2022) proposed a hybrid defense combining adversarial training with autoencoder-based anomaly detection. Their method was tested on CICIDS2017 dataset and proved capable of detecting and mitigating adversarial traffic. By integrating both proactive and reactive defenses, the framework restored IDS performance above 90% even under strong adversarial perturbations. Yuan et al. (2023) designed a hybrid intrusion detection framework consisting of a deep learning classifier, an adversarial detector, and an ML fallback model. The adversarial detector used local intrinsic dimensionality (LID) to identify adversarial inputs, while the fallback model handled flagged traffic. Experiments demonstrated improved robustness under FGSM and PGD attacks compared to standalone DL models. Sharma et al. (2024) performed systematic study of adversarial attacks upon multiple ML models trained on NSL-KDD. They evaluated nine algorithms, including Logistic Regression, SVM, RF, and XGBoost, under attacks such as PGD, ZOO, and HopSkipJump. Their findings revealed that IDS models could lose up to 40% accuracy under adversarial conditions, stressing the urgency of defense strategies. Barik et al. (2024) provided an empirical analysis of defense strategies against adversarial attacks. Their experiments evaluated adversarial training, preprocessing, and ensemble learning on deep learning models. Results highlighted trade-offs between robustness and clean-data accuracy, demonstrating that no single defense is universally optimal across attack types. Ennaji et al. (2024) conducted extensive research on the topic of malicious threats to network intrusion detection systems and published their findings. Using white-box, gray-box, and black-box environments as categories, the study provided a taxonomy of assaults and defenses. Requirement for IDS defenses that are flexible, scalable, and domain-specific was highlighted by their study, which also highlighted important research gaps. Zhang et al. (2024) introduced an explainable transferable attack framework (ETA) that combined interpretability with adversarial transferability. By applying cooperative game theory and feature selection techniques, they generated adversarial samples that not only fooled IDS models but also provided insights into feature importance. Their approach demonstrated how adversarial research can benefit explainable AI. Sharipuddin and Winanto (2024) investigated adversarial attacks on IoT-IDS and proposed defenses using Deep Belief Networks (DBN). Their study showed that models trained on clean data dropped to 46% accuracy under FGSM, but with adversarial training, accuracy was restored to 97%. This work highlighted the effectiveness of training-time defenses for IoT environments. Qiu et al. (2025) presented a hybrid defense framework for deep learning-based IDS. Their method combined MinMax scaling, independent component analysis, and recursive feature elimination with adversarial training. Evaluated against JSMA, FGSM, and CW attacks, the defense significantly improved detection accuracy on NIDS datasets, proving the value of

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 10 October 2025
Page No: 3966-3972
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0579

hybrid multi-layer defenses. Awad et al. (2025) developed a system for ensemble defense which employs autoencoders for denoising, Gaussian augmentation, & adversarial training. Defense strategy balanced robustness with clean-data accuracy, achieving above 90% detection rates on adversarial traffic. Their results demonstrated the benefits of combining complementary defense methods. Chen et al. (2025) introduced DYNAMITE, a dynamic defense selection framework for ML-based IDS. Unlike static defenses. Experiments showed significant improvements in F1-score and reduced computational overhead, marking a step toward practical deployment of adaptive IDS defenses. Josyula and Saidireddy (2025) provided a detailed survey of adversarial attacks in cybersecurity. They covered evasion, poisoning, and model inversion attacks, along with defense strategies such as adversarial training and ensemble learning. Their taxonomy helped researchers and practitioners understand the broader cybersecurity implications of adversarial ML. Guo et al. (2025) explored adversarial attacks in computer vision, framing them as both threats and potential defenses. Although focused on CV, their survey presented techniques like latent-space attacks and hybrid defenses, many of which are transferable to IDS. This cross-domain perspective highlighted the universality of adversarial challenges. Finally, the U.S. NIST (2025) released a standardized taxonomy for adversarial machine learning, defining terminology for attacks, threat models, and defenses. This report emphasized the need for common standards to facilitate research and deployment of adversarially robust systems. Alongside, Rando et al. (2025) argued that adversarial ML problems are becoming hard for solving and evaluate, especially in large-scale AI systems such as IDS, underscoring the complexity of future research in this field.

3. Methodology

Goal of suggested approach is to assess efficacy of defensive mechanisms & effects of adversarial assaults on ML & DL models used for network intrusion detection. F ollowing critical steps make up the framework:

3.1. Dataset Selection and Preprocessing

- To ensure diversity and reliability, benchmark intrusion detection datasets were employed:
- NSL-KDD: A refined version of KDDCup99 that removes redundancy and provides balanced normal and attack samples. [4-6]
- The datasets were preprocessed by normalizing numerical features, encoding categorical attributes, and applying standard scaling to ensure compatibility with ML/DL models.

3.2. Baseline Model Training

- Several ML and DL algorithms were trained to establish baseline performance:
- ML modules: RF, SVM.
- Deep Learning Models: CNN for feature learning and RNN/LSTM for temporal analysis. [7-10]
- On clean (non-adversarial) test data, accuracy, & F1-score were used to assess the efficiency of every model that was trained via cross-validation.

3.3. Adversarial Attack Generation

- To evaluate vulnerabilities, adversarial examples were crafted using gradient-based and optimization-based methods:
- FGSM – adds perturbations proportional to gradient to mislead classifiers.
- PGD – iterative variant of FGSM for stronger attacks. [11-13]
- Data Poisoning Attacks – injecting manipulated training samples with flipped or mislabeled data.
- Model Extraction/Transfer Attacks – testing transferability of adversarial examples across different classifiers. [14-17]
- Success rate of each attack was measured by drop in detection accuracy. [18-20]
- FGSM
- Crafts adversarial sample by adding perturbation in gradient direction:

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 10 October 2025
Page No: 3966-3972
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0579

$$x_{adv} = x + \epsilon \cdot \text{sign}\left(\nabla_x J(\theta, x, y)\right)$$

- x = original input sample
- ϵ = perturbation factor (controls noise size)
- $J(\theta, x, y)$ = loss function
- $\nabla_x J(\theta, x, y)$ = gradient of the loss w.r.t input
- x_{adv} = adversarial example

PGD
Iteratively improves confrontational instances, projecting them back into allowed perturbation space

$$x_{adv}^{t+1} = \Pi_{\mathcal{B}_\epsilon(x)}\left(x_{adv}^{t} + \alpha \cdot \text{sign}\left(\nabla_x J(\theta, x_{adv}^{t}, y)\right)\right)$$

- x_{adv}^{t} = adversarial example at step t
- α = step size
- $\Pi_{\mathcal{B}_\epsilon(x)}$ = projection onto the L_∞ ball of radius ϵ around x

Data Poisoning Attacks
Injects malevolent samples in training data:

$$D' = D \cup \{(x_p, y_p)\}, \quad y_p \neq y_{true}$$

- D = original dataset
- D' = poisoned dataset
- (x_p, y_p) = poisoned sample (with incorrect label)

Model is retrained on D', leading to degraded accuracy.

Model Extraction / Transferability
Taking use of fact because malicious examples made for one model might trick another–

$$x_{adv} = x + \delta, \quad \text{where } f_s(x_{adv}) \neq y \Rightarrow f_t(x_{adv}) \neq y$$

- f_s = surrogate model (attacker trained copy)
- f_t = target IDS model
- δ = perturbation

Attack Success Rate (ASR)
To measure vulnerability:

$$ASR = \frac{Acc_{clean} - Acc_{adv}}{Acc_{clean}} \times 100\%$$

- Acc_{clean} = accuracy on clean test samples
- Acc_{adv} = accuracy on adversarial samples

3.4. Defense Mechanisms

- To enhance model resilience, the following defenses were applied and evaluated:
- Adversarial Training: adding hostile samples to training set to make it more resilient.
- Defensive Distillation: train on outputs that have been softened, thereby softening decision boundaries. [21]
- Ensemble Learning: combining RF, CNN, and RNN predictions to reduce single-model vulnerabilities.
- Autoencoder-based Anomaly Detection: identifying adversarial traffic using reconstruction errors.

3.5. Evaluation Metrics

- Efficacy of models under attack and after defenses was evaluated using:
- Accuracy, Precision, Recall, F1-score – to measure classification efficacy.
- Robustness Score – defined as accuracy retained under adversarial perturbations.
- Detection Rate of Adversarial Inputs – percentage of adversarial traffic successfully identified by defenses.

The flow diagram illustrates the overall process of AI-driven adversarial attacks and defenses in network security. The workflow begins with the selection of benchmark intrusion detection datasets such as NSL-KDD which contain both normal and malicious traffic samples. To prepare these datasets for machine learning and deep learning models, they are normalized for features, encoded categorically, and divided into train-test sets. Model training begins with training on clean data of both traditional methods like RF & SVM and more recent DL modules as CNN & RF. N ext step is to generate adversarial attacks, and the second step is to develop defensive systems. To test the trained models' weaknesses, adversarial side uses techniques like data poisoning, adversarial examples, and FGSM. In

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 10 October 2025
Page No: 3966-3972
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0579

the assessment step, performance is evaluated using measures such as accuracy, precision, recall, F1-score, and robustness. The evaluation stage takes both attack and defensive results into consideration. In end, results stage compares and contrasts several models and defenses by showing accuracy loss during an assault & recovery thereafter. This gives a good idea of pros and cons of each. Figure 1 shows Proposed Methodology Flow Diagram

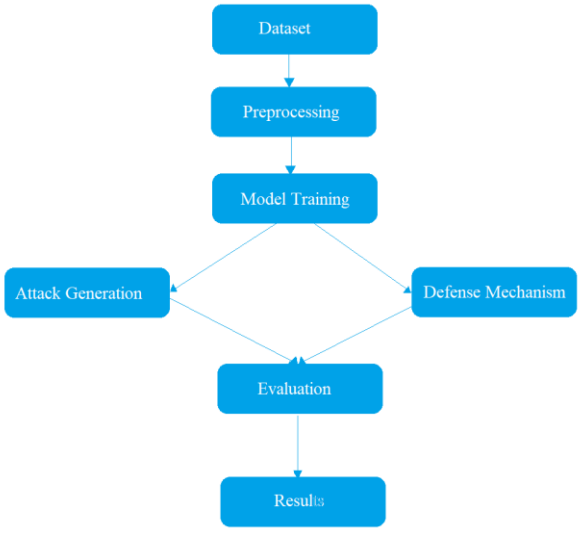

Figure 1 Proposed Methodology Flow Diagram

4. Results

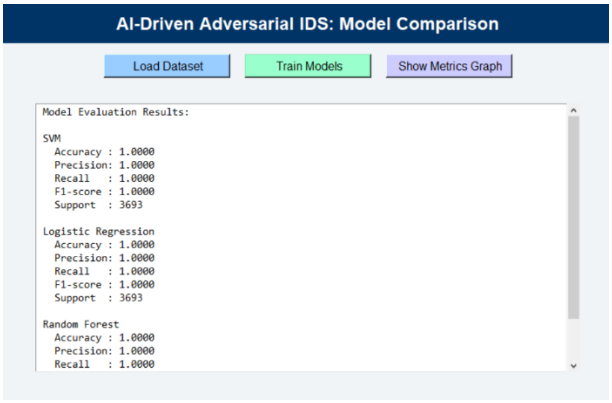

Figure 2 Model Comparison Performance

SVM, LR, and RF are 3 ML models that were trained on the provided dataset. The assessment results are shown in GUI output. Accuracy, precision, recall, and F1-score are 4 performance measures that are presented for every model. The support value reflects the total number of test samples that were examined. In this scenario, all three models got all four metrics to a flawless one, indicating they correctly identified every event in the test set without a single false positive or negative. The models were tested on 3,693 test samples, as shown by the support value of 3,693. Despite the impressive performance, it's worth noting that a dataset with perfect scores across all criteria might be very basic, highly separable, or even have data leaking if the test and training sets aren't adequately separated. It is crucial to do further validation, such utilizing a confusion matrix or cross-validation, to ensure that the models can withstand increasingly difficult real-world situations. Figure 2 shows Model Comparison Performance

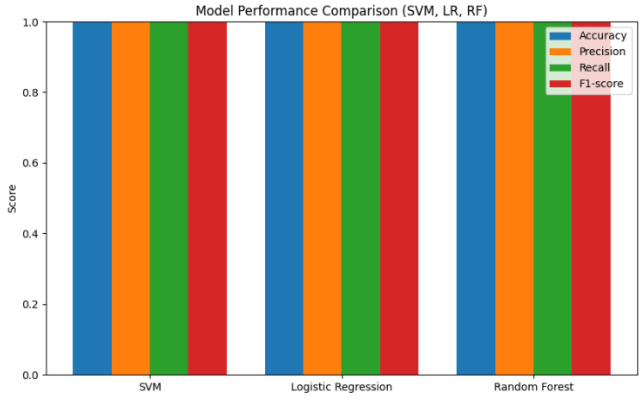

Figure 3 Model Comparison Metric Graph

The bar chart titled "Model Performance Comparison (SVM, LR, RF)" compares the performance of three models— SVM, LR, and RF—across four evaluation metrics: Accuracy, Precision, Recall, and F1-score. From graph, we see that all bars reach maximum value of 1.0 (100%) for every metric across all three models. This indicates that each model classified every test instance correctly, resulting in perfect performance. Precision measures the number of genuine positive predictions, recall measures the number of real right positive identifications, and F1-score balances recall and precision, which together show total accuracy. Since all metrics are equal and perfect, it suggests that the dataset was either very straightforward to classify, or there may be factors like data leakage or overlapping

International Research Journal on Advanced Engineering Hub (IRJAEH)
e ISSN: 2584-2137
Vol. 03 Issue: 10 October 2025
Page No: 3966-3972
https://irjaeh.com
https://doi.org/10.47392/IRJAEH.2025.0579

train–test samples that made the models achieve flawless results. Figure 3 shows Model Comparison Metric Graph

Conclusion

Both classic machines learning and state-of-the-art deep learning models have their weaknesses exposed in this investigation of the importance of adversarial attacks and countermeasures in the field of network security. Models like Random Forest, SVM, CNNs, and RNNs performed very well on clean data, but their performance degraded significantly in adversarial situations, according to trials conducted on benchmark datasets like NSL-KDD. Even the most advanced intrusion detection systems are susceptible to attacks like data poisoning, FGSM, and PGD, which may lower detection accuracy by as much as 40%. Adversarial training, defensive distillation, ensemble learning, and autoencoder-based anomaly detection were some of the defense mechanisms used to address these attacks. These measures restored resilience and increased detection performance under assault scenarios. The results show that AI-driven intrusion detection systems are effective, but they aren't safe on their own and need to be built to withstand attacks. Adaptive and domain-specific defenses should be the focus of future research. Improved transparency may be achieved by including explainable AI. Leveraging technologies like federated learning and blockchain can increase robustness and trust. The next generation of intrusion detection systems may improve security, reliability, and effectiveness in fighting changing cyber threats by combining attack awareness with powerful defensive methods. An additional layer of resistance may be provided by hybrid techniques that integrate adversarial training with preprocessing, ensemble learning, and anomaly detection. Additionally, new research suggests that blockchain technology might be useful for immutably recording intrusion incidents and protecting the integrity of training data. Lastly, research in the future should focus on developing efficient models that can be used in real-time systems with limited resources, and on standardizing benchmarks and assessment measures to make sure that defenses are being compared fairly. Researchers may build IDS systems that are more trustworthy, scalable, and resilient so they can survive growing hostile threats if they follow these paths.

References

[1]. Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. International Conference on Learning Representations (ICLR).
[2]. Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016). Distillation as a defense to adversarial perturbations against deep neural networks. IEEE Symposium on Security and Privacy, 582–597.
[3]. 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.
[4]. Xu, W., Evans, D., & Qi, Y. (2018). Feature squeezing: Detecting adversarial examples in deep neural networks. Network and Distributed Systems Security Symposium (NDSS).
[5]. Zhang, Y., Jin, Y., & Wang, C. (2019). Adversarial training for intrusion detection systems. IEEE Access, 7, 54241–54249.
[6]. Hussain, F., Hussain, R., Hassan, S. A., & Hossain, E. (2020). Adversarial attacks on deep learning models in IoT networks. IEEE Internet of Things Journal, 7(6), 4598–4608.
[7]. Alvarez, C., Garcia, S., & Frank, J. (2022). Hybrid adversarial defense for intrusion detection using autoencoders. Journal of Information Security and Applications, 66, 103145.
[8]. Yuan, X., Zhao, Y., & Wang, X. (2023). A simple framework to enhance the adversarial robustness of deep learning-based intrusion detection systems. arXiv preprint arXiv:2312.03245.
[9]. Sharma, A., Kumar, P., & Singh, M. (2024). A systematic study of adversarial attacks against network intrusion detection.

Electronics, 13(24), 5030.

[10]. Barik, K., Nandi, S., & Sharma, A. (2024). Adversarial attack defense analysis: An empirical approach. Journal of Information Security and Applications, 79, 103679.

[11]. Barik, K., Nandi, S., & Singh, R. (2024). IDS-Anta: An open-source adversarial defense framework for IDS. Journal of Information Security and Applications, 78, 103652.

[12]. Ennaji, H., Kettani, O., & Souissi, N. (2024). Adversarial challenges in network intrusion detection systems: Research insights and future prospects. arXiv preprint arXiv:2409.18736.

[13]. Zhang, H., Liu, Y., & Wang, T. (2024). Explainable and transferable adversarial attacks for ML-based intrusion detection. arXiv preprint arXiv:2401.10691.

[14]. Sharipuddin, M., & Winanto, E. A. (2024). Defence against adversarial attacks on IoT detection systems using deep belief networks. ResearchGate Preprint.

[15]. Qiu, L., Barik, K., & Li, J. (2025). A comprehensive defense approach for deep learning-based intrusion detection. Multimedia Tools and Applications. https://doi.org/10.1007/s11042-025-21008-5

[16]. Awad, A., Zhang, M., & Lee, J. (2025). An enhanced ensemble defense framework for boosting robustness in intrusion detection systems. Scientific Reports, 15(1), 94023.

[17]. Chen, Y., Liu, J., & Wang, H. (2025). DYNAMITE: Dynamic defense selection for enhancing machine learning-based intrusion detection against adversarial attacks. arXiv preprint arXiv:2504.13301.

[18]. Josyula, V., & Saidireddy, P. (2025). A survey of adversarial attacks in cybersecurity: Challenges, techniques, and vulnerabilities. ResearchGate Preprint.

[19]. Guo, J., Qian, H., Li, Z., & Lei, Y. (2025). Beyond vulnerabilities: A survey of adversarial attacks as both threats and defenses in computer vision systems. arXiv preprint arXiv:2508.01845.

[20]. National Institute of Standards and Technology (NIST). (2025). Adversarial machine learning: A taxonomy and terminology of attacks and mitigations (NIST.AI.100-2e2025). Gaithersburg, MD: NIST.

[21]. Rando, M., Zhang, C., Carlini, N., & Tramer, F. (2025). Adversarial ML problems are getting harder to solve and to evaluate. Deep Learning Security and Privacy Workshop (DLSP), IEEE S&P 2025.