International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

Analysis and Classification of Adversarial Machine Learning Attacks Against Machine Learning-Based Network Intrusion Detection Systems

Mr.Y.H.S.S. Phaneedra[1] , Polisetty Nikhitha Sowmya[2], Kolla Triveni[3], Garaga Naveen Kumar[4], Kadali Nikitha Sri Satya Gayatri[5]
[1]Assistant Professor , [2, 3, 4, 5]B.tech Students Department of CSE,
Pragati Engineering College, Surampalem , Andhra Pradesh, India

Abstract- Network Intrusion Detection Systems (NIDS) play a critical role in modern cybersecurity infrastructures by monitoring network traffic and identifying suspicious or malicious activities. In recent years, machine learning techniques have significantly improved the performance of intrusion detection systems by enabling automated traffic analysis and anomaly detection. However, the integration of machine learning into security systems also introduces new vulnerabilities that can be exploited by attackers. One such threat is adversarial machine learning, where malicious actors manipulate training or testing data to deceive machine learning models and degrade their performance. This study presents a comprehensive analysis of adversarial machine learning attacks targeting network intrusion detection systems. The work explores how adversarial samples are generated by introducing small perturbations into original datasets, which results in incorrect predictions by the intrusion detection model. Furthermore, the paper classifies adversarial attacks based on several criteria, including attacker knowledge level, misclassification objectives, affected learning phase, and the intended security violation. Understanding these attack strategies is essential for designing more robust and secure intrusion detection systems capable of defending against adversarial manipulation.

Keywords – Adversarial Machine Learning, Network Intrusion Detection Systems (NIDS), Cybersecurity, Poisoning Attack, Evasion Attack, Machine Learning Security, Adversarial Samples.

I. INTRODUCTION

With the rapid development of digital communication networks and emerging technologies such as cloud computing, Internet of Things (IoT), and future 6G communication systems, the amount of data generated and exchanged across networks has increased significantly. This rapid expansion of network infrastructure has also introduced new cybersecurity challenges, as attackers continuously attempt to exploit system vulnerabilities and compromise sensitive information. As a result, protecting network systems from malicious activities has become a major concern in modern cybersecurity environments [5], [6].

Network Intrusion Detection Systems (NIDS) play a crucial role in identifying suspicious network behaviour and protecting systems against cyberattacks. These systems monitor network traffic and analyse patterns to detect anomalies or malicious activities. Traditionally, intrusion detection systems relied on signature-based detection methods, which identify attacks by comparing network traffic with previously known attack patterns. Although signature-based systems are effective in detecting known threats, they often fail to identify new or unknown attack patterns, commonly referred to as zero-day attacks [4], [5].

To overcome these limitations, machine learning techniques have been widely integrated into intrusion detection systems. Machine learning models can analyse large volumes of network traffic data and learn patterns associated with normal and malicious behaviour. By applying supervised and unsupervised learning methods, these models can identify anomalies and detect previously unseen attacks more efficiently. Popular machine learning algorithms used in intrusion detection systems include Decision Trees, Support Vector Machines (SVM), Random Forest, Artificial Neural Networks, and Logistic Regression [5], [6].

Despite the advantages of machine learning-based intrusion detection systems, these models also introduce new security vulnerabilities. One of the most significant threats is Adversarial Machine Learning (AML), where attackers intentionally manipulate input data to mislead machine learning models. In adversarial attacks, small perturbations are

added to the original dataset to generate adversarial samples that cause the model to produce incorrect predictions. As a result, malicious network traffic may be incorrectly classified as legitimate, allowing attackers to bypass the security mechanisms of the intrusion detection system [1], [2].

Adversarial attacks can occur during different stages of the machine learning lifecycle, including the training phase and the testing phase. During the training phase, attackers may inject malicious data into the training dataset to influence the learning process, a technique known as a poisoning attack. In contrast, during the testing phase, attackers modify input samples to evade detection without altering the trained model, which is referred to as an evasion attack. These attacks can significantly degrade the performance and reliability of intrusion detection systems [7], [8].

Therefore, understanding the characteristics and classification of adversarial machine learning attacks is essential for designing robust and secure intrusion detection systems. This study focuses on analysing adversarial attacks targeting machine learning-based network intrusion detection systems and provides a classification of these attacks based on different criteria such as attacker knowledge level, misclassification objectives, affected operational phase, and intended security violation. Such analysis can help researchers and cybersecurity professionals develop more resilient defence mechanisms against adversarial threats [9], [10].

II. LITERATURE SURVEY

Researchers have proposed several approaches to enhance the performance of Network Intrusion Detection Systems (NIDS) using machine learning techniques. With the increasing complexity of cyber threats and the rapid growth of network infrastructures, intelligent intrusion detection systems have become essential for identifying malicious network behavior and preventing unauthorized access. However, recent studies also highlight the vulnerabilities of machine learning models when exposed to adversarial attacks, which can significantly reduce the reliability of these systems [1], [2].

Mahfouz et al. conducted a comparative analysis of various machine learning classifiers for intrusion detection systems. Their study evaluated algorithms such as Decision Trees, Support Vector Machines (SVM), Random Forest, and Artificial Neural Networks for detecting network attacks. The results demonstrated that machine learning models can effectively improve the accuracy of intrusion detection systems by identifying abnormal network traffic patterns and detecting malicious activities within network environments [4].

Ahmad et al. presented a systematic study on machine learning and deep learning approaches used in network intrusion detection. Their research explored different classification techniques and highlighted the advantages of deep learning models in handling large-scale network traffic data. The study concluded that advanced learning algorithms can significantly improve detection accuracy and enable the identification of previously unknown cyber threats [5].

Jmila and Khedher investigated the impact of adversarial machine learning attacks on network intrusion detection systems. Their work analysed how attackers can manipulate input data to mislead machine learning models. The study demonstrated that even small perturbations introduced into network datasets can significantly degrade the performance of intrusion detection systems by causing incorrect classifications and increasing system vulnerability to cyberattacks [7].

Ibitoye et al. provided a comprehensive survey on adversarial attacks targeting machine learning models used in cybersecurity applications. The authors discussed different attack strategies and explained how adversarial samples can be generated to exploit weaknesses in machine learning models. Their research emphasized the importance of developing robust defence mechanisms to protect intrusion detection systems from adversarial manipulation and ensure the reliability of AI-based cybersecurity solutions [1].

Sharma and Chen conducted a systematic study on adversarial attacks against machine learning-based intrusion detection systems. Their work analysed various types of adversarial attacks and demonstrated how these attacks can increase the false positive and false negative rates of detection systems. The study also highlighted the need for improved security mechanisms capable of defending machine learning models from adversarial threats in network environments [8].

From the above studies, it is evident that machine learning techniques have significantly improved the performance of intrusion detection systems in identifying network attacks and abnormal network behaviour. However, the integration of machine learning also introduces new vulnerabilities that can be exploited through adversarial attacks. Therefore, further research is required to understand the characteristics of adversarial machine learning attacks and develop effective defence strategies for protecting network intrusion detection systems from such threats [2], [9], [10].

III. SYSTEM ANALYSIS

Existing System

Traditional Network Intrusion Detection Systems (NIDS) are designed to monitor network traffic and identify suspicious

activities that may indicate potential cyberattacks. These systems generally operate using two primary approaches: signature-based detection and anomaly-based detection. Signature-based systems detect attacks by comparing network traffic patterns with previously known attack signatures stored in a database. While this approach is effective for identifying known threats, it is unable to detect new or previously unseen attacks that do not match existing signatures. These unknown threats are often referred to as zero-day attacks, which pose a significant challenge for traditional security systems [4], [5].

To address these limitations, machine learning-based intrusion detection systems have been developed. These systems utilize machine learning algorithms to analyse network traffic data and classify activities as normal or malicious. Commonly used algorithms include Decision Trees, Support Vector Machines (SVM), Random Forest, Naïve Bayes, Logistic Regression, Artificial Neural Networks, and k-Nearest Neighbors. These models are trained using labelled datasets containing both benign and malicious network traffic samples, enabling them to identify patterns associated with cyberattacks and abnormal network behaviour [5], [6].

Furthermore, ensemble learning approaches have been introduced to enhance the performance of intrusion detection systems. Algorithms such as Random Forest and Gradient Boosting combine multiple classification models to improve prediction accuracy and reduce the risk of overfitting. These ensemble techniques have demonstrated improved detection performance in various network security applications [5].

However, despite the advantages of machine learning-based intrusion detection systems, these models are vulnerable to Adversarial Machine Learning (AML) attacks. In adversarial attacks, attackers intentionally manipulate input data by introducing small perturbations that cause machine learning models to produce incorrect predictions. As a result, malicious network traffic can be misclassified as legitimate traffic, allowing attackers to bypass security mechanisms and compromise network systems [1], [2].

Recent studies have shown that adversarial attacks can significantly degrade the performance of machine learning-based intrusion detection systems. These attacks exploit weaknesses in the learning models and increase the false positive and false negative rates, thereby reducing the reliability of cybersecurity defence mechanisms. Understanding these vulnerabilities is therefore essential for designing more secure intrusion detection systems [7], [8].

LIMITATIONS OF EXISTING SYSTEM

- Despite the improvements introduced by machine learning-based intrusion detection systems, several challenges remain when deploying these systems in real-world cybersecurity environments.
- One of the primary challenges is the vulnerability of machine learning models to adversarial attacks. Attackers can generate adversarial samples by introducing small perturbations to input data, which may cause the intrusion detection system to misclassify malicious traffic as legitimate network activity [1], [7].
- Another limitation is the inability of traditional detection systems to effectively identify sophisticated or zero-day attacks. Signature-based detection mechanisms rely on previously known attack patterns and therefore cannot detect new or unknown threats that have not been previously recorded [5].
- High computational requirements also present challenges in large-scale network environments. Machine learning algorithms often require significant computational resources for training and analysing large volumes of network traffic data.
- Model interpretability issues represent another challenge. Many advanced machine learning and deep learning models operate as complex black-box systems, making it difficult for cybersecurity analysts to understand how the model reaches specific classification decisions [8].
- The performance of intrusion detection systems is also highly dependent on the quality and diversity of training datasets. Incomplete or biased datasets may lead to inaccurate predictions and reduced detection capability.
- Additionally, machine learning models may produce false positives and false negatives, where legitimate traffic is incorrectly classified as malicious or malicious traffic remains undetected. Such errors reduce the overall reliability of intrusion detection systems.
- Finally, scalability challenges arise due to the rapid growth of network traffic and data volume. Existing intrusion detection systems may struggle to efficiently process large datasets generated by modern network infrastructures.

Proposed System

The proposed system introduces a comprehensive analysis framework for studying adversarial machine learning attacks against network intrusion detection systems. The objective of this framework is to understand how adversarial samples can be generated and used to manipulate machine learning models employed in NIDS.

In the proposed approach, network traffic data is first collected and pre-processed to remove noise, redundant information, and irrelevant features. Data preprocessing techniques are applied to ensure that the dataset is suitable for machine learning analysis. Feature extraction methods are then used to obtain relevant network characteristics that represent the behaviour of network traffic.

These extracted features are used to train machine learning classifiers capable of distinguishing between normal network activity and malicious attacks. The trained models analyse network patterns and identify anomalies that may indicate potential cyber threats.

The framework also investigates how adversarial samples can be generated by introducing small perturbations into the original dataset. These perturbations are designed to deceive machine learning models and force them to produce incorrect predictions. The study categorizes adversarial attacks based on several criteria, including:

- Attacker knowledge level (white-box, black-box, gray-box)
- Misclassification objectives
- Operational phase affected (training or testing)
- Type of security violation

By analysing the characteristics and classification of adversarial machine learning attacks, the proposed framework provides insights into potential vulnerabilities within machine learning-based intrusion detection systems.

This analysis can help researchers and cybersecurity professionals design more robust intrusion detection models capable of resisting adversarial manipulation and improving overall network security [1], [2], [9], [10].

IV.SYSTEM DESIGN

System Architecture

Below diagram depicts the whole system architecture.

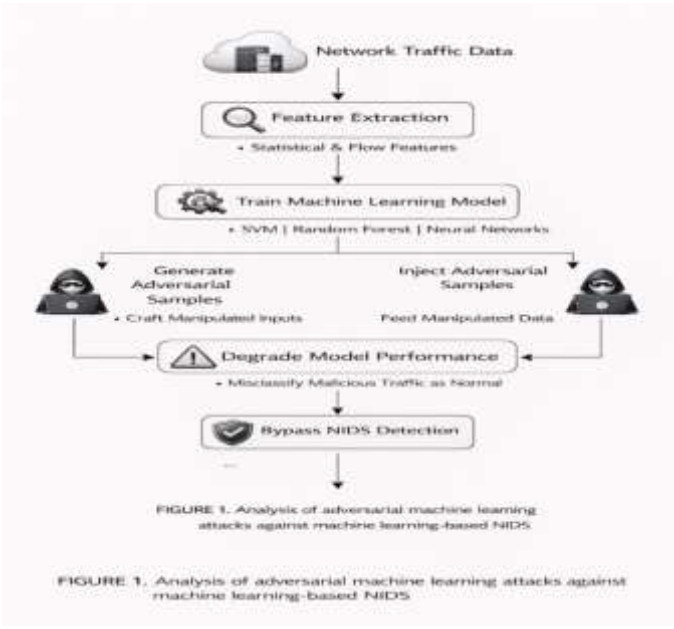

Fig 1. Methodology followed for proposed model

V. SYSTEM IMPLEMENTATION MODULES

This section describes the implementation modules of the proposed framework for analysing adversarial machine learning attacks on Network Intrusion Detection Systems (NIDS). The framework follows a modular architecture consisting of network traffic data acquisition, preprocessing, feature extraction, machine learning model training, adversarial attack simulation, and performance evaluation. This structured design improves the reliability and effectiveness of intrusion detection systems while enabling comprehensive analysis of adversarial threats in network environments.

Network Traffic Data Collection Module

The Network Traffic Data Collection Module is responsible for acquiring network traffic datasets used to train and evaluate intrusion detection models. The datasets are collected from publicly available cybersecurity repositories and simulated network environments. These datasets contain both normal network traffic samples and malicious traffic samples representing different types of cyberattacks such as denial-of-service attacks, probing attacks, and unauthorized access attempts.

The collected network traffic data includes multiple attributes such as packet size, protocol type, connection duration, source and destination addresses, and traffic flow characteristics. These attributes represent the behavioural patterns of network communication and are essential for identifying abnormal activities within the network. The raw network traffic data is stored in a structured format and forwarded to the preprocessing module for further processing.

Data Preprocessing Module

The Data Preprocessing Module prepares the collected network traffic data for machine learning analysis. Raw network datasets often contain missing values, redundant records, and noisy data that may negatively affect model performance. Therefore, preprocessing techniques are applied to improve dataset quality.

The preprocessing stage includes the following steps:

1. **Data Cleaning**
2. Invalid and redundant network traffic records are removed to ensure data consistency.
3. **Data Normalization**
4. Feature values are normalized to maintain consistent ranges across different attributes, allowing machine learning algorithms to process the data more effectively.
5. **Data Transformation**

Categorical network attributes such as protocol types are converted into numerical representations that can be used as input for machine learning models.

These preprocessing steps improve data quality and enhance the reliability of intrusion detection models [5], [6].

Feature Extraction and Feature Engineering Module

The Feature Extraction Module identifies relevant characteristics of network traffic that contribute to detecting malicious activities. Network traffic datasets typically contain numerous attributes, but not all features are equally useful for intrusion detection. In this module, important features such as packet size, connection duration, protocol type, traffic flow statistics, and network behaviour indicators are extracted and analysed.

Feature engineering techniques are applied to transform raw network traffic data into meaningful representations that capture patterns associated with cyberattacks. Feature selection techniques are also applied to remove redundant or irrelevant attributes. Reducing the number of features helps decrease computational complexity and improves the efficiency of machine learning models.

Machine Learning Training Module

The Machine Learning Training Module builds classification models capable of distinguishing between normal network traffic and malicious network activities. The processed dataset is divided into training and testing subsets to evaluate the performance of different machine learning algorithms.

Several machine learning algorithms are implemented and evaluated, including:

• Support Vector Machine (SVM)
• Decision Tree
• Random Forest
• Naïve Bayes
• Artificial Neural Networks

Each model is trained using historical network traffic data. During the training process, the models learn patterns associated with normal and malicious network behaviour. These learned patterns enable the models to detect intrusion attempts when analysing new network traffic samples.

The trained models are then tested using unseen data to evaluate their detection capability and classification accuracy.

Adversarial Sample Generation and Attack Simulation Module

The Adversarial Attack Simulation Module analyses how adversarial machine learning attacks can manipulate intrusion detection systems. In this module, adversarial samples are generated by introducing small perturbations into the original dataset.

These perturbations modify network traffic features in a way that appears legitimate but causes machine learning models to produce incorrect predictions. As a result, malicious traffic may be misclassified as normal network activity, allowing attackers to bypass intrusion detection mechanisms.

The generated adversarial samples are used to simulate real-world attack scenarios and evaluate the robustness of machine learning-based intrusion detection systems. This analysis helps identify potential vulnerabilities within the detection models.

Model Evaluation and Security Analysis Module

The Model Evaluation Module assesses the performance of the intrusion detection models and analyses their vulnerability to adversarial attacks. Several evaluation metrics are used to measure the effectiveness of the detection system.

The evaluation metrics include:

- Accuracy – Measures the overall correctness of predictions.
- Precision – Indicates how many detected intrusions are actually malicious.
- Recall – Measures the system's ability to detect actual attacks.
- F1-Score – Provides a balanced measure of precision and recall.
- ROC–AUC Score – Evaluates the model's ability to distinguish between normal and malicious traffic.

In addition to evaluating model accuracy, this module also analyses how adversarial attacks influence intrusion detection performance. By comparing model performance before and after adversarial attack simulation, the framework provides insights into the robustness of machine learning models against adversarial threats.

This analysis helps researchers and cybersecurity professionals design more secure and resilient intrusion detection systems capable of resisting adversarial manipulation [1], [7], [8].

VI. RESULTS AND DISCUSSION

This section presents the experimental results and performance evaluation of the proposed framework for analysing adversarial machine learning attacks against network intrusion detection systems. Experiments were conducted using network traffic datasets containing both normal and malicious activities. Machine learning models were trained using extracted network features to classify network traffic patterns. The evaluation focuses on comparing model performance and analysing how adversarial attacks affect intrusion detection accuracy.

Accuracy Comparison of Machine Learning Models

Several machine learning algorithms were evaluated to determine their effectiveness in detecting malicious network activities. The evaluated models include Support Vector Machine (SVM), Decision Tree, Random Forest, Naïve

Bayes, and Artificial Neural Networks. Model performance was measured using metrics such as accuracy, precision, recall, and F1-score.

Table 1. Performance Comparison of Intrusion Detection Models

Model	Accuracy (%)	Precision	Recall	F1-Score
Decision Tree	86.2	0.85	0.84	0.84
Naïve Bayes	84.5	0.83	0.82	0.82
Support Vector Machine	89.4	0.88	0.87	0.87
Artificial Neural Network	91.3	0.90	0.89	0.89
Random Forest	93.1	0.92	0.91	0.91

From the comparison results, the Random Forest model achieved the highest classification accuracy of 93.1%, outperforming other machine learning algorithms. The improved performance of Random Forest can be attributed to its ensemble learning structure, which combines multiple decision trees to improve prediction stability and reduce overfitting [5], [6].

ROC Curve Analysis

The Receiver Operating Characteristic (ROC) curve is used to evaluate the performance of classification models by measuring the relationship between the True Positive Rate (TPR) and False Positive Rate (FPR) across different classification thresholds. The Area Under the Curve (ROC–AUC) metric provides a comprehensive measure of the model's ability to distinguish between normal and malicious network traffic.

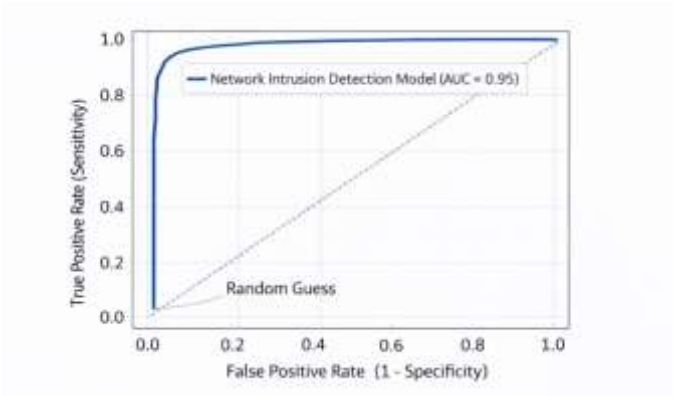

Fig 2. ROC Curve for Network Intrusion Detection Model

In the conducted experiments, the Random Forest model achieved a ROC–AUC score of 0.95, indicating strong capability in distinguishing between normal network behaviour and cyberattack traffic. A ROC curve that approaches the top-left corner of the graph represents higher detection sensitivity and lower false alarm rates.

The ROC analysis confirms that machine learning-based intrusion detection systems can effectively identify malicious network activities under normal operating conditions.

Adversarial Attack Impact Analysis

To further analyse the robustness of machine learning-based intrusion detection systems, adversarial samples were introduced into the network traffic dataset. These adversarial examples were generated by applying small perturbations to the original network traffic features while preserving the overall structure of the data.

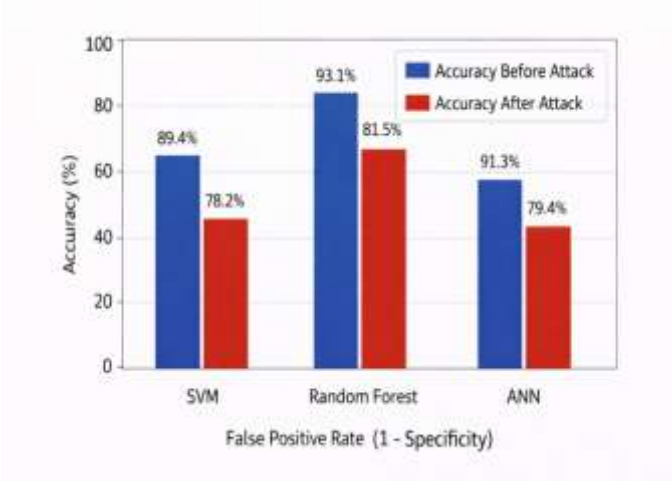

Fig 3. Impact of Adversarial Samples on Intrusion Detection Performance

The experimental results reveal that adversarial samples significantly influence the prediction behaviour of machine learning models. Even minor modifications to network traffic attributes can cause the intrusion detection system to misclassify malicious traffic as legitimate activity. This behaviour demonstrates how adversarial attacks exploit vulnerabilities in machine learning models.

The analysis also shows that adversarial attacks increase the false positive and false negative rates of intrusion detection systems. As a result, some malicious activities remain undetected, which may allow attackers to bypass network security mechanisms.

The evaluation of adversarial attack impact provides important insights into the limitations of current machine learning-based intrusion detection systems. Understanding these vulnerabilities can help researchers develop more robust defense strategies and adversarially resilient machine learning models for improving cybersecurity systems [1], [7], [8].

VII. CONCLUSION AND FUTURE WORK

This study presented an analytical framework for examining adversarial machine learning attacks targeting machine learning-based Network Intrusion Detection Systems (NIDS). The research investigated how adversarial samples can be generated by introducing small perturbations into network traffic datasets and how these adversarial inputs can manipulate machine learning models used for intrusion detection.

The experimental analysis demonstrated that machine learning techniques significantly improve the capability of intrusion detection systems to identify malicious network activities. Algorithms such as Support Vector Machines, Decision Trees, Random Forest, and Artificial Neural Networks can effectively classify network traffic patterns and detect cyber threats under normal operating conditions. However, the study also revealed that these machine learning models are vulnerable to adversarial manipulation, where carefully crafted adversarial samples can cause the detection models to produce incorrect classifications. As a result, malicious network traffic may be incorrectly labeled as legitimate traffic, allowing attackers to bypass network security mechanisms and compromise system integrity [1], [7].

The results highlight the importance of understanding adversarial attack strategies in order to develop more secure and reliable intrusion detection systems. By analysing the impact of adversarial samples on model performance, the proposed framework provides insights into potential vulnerabilities in machine learning-based cybersecurity systems.

Future research can focus on developing robust defense mechanisms to improve the resilience of intrusion detection models against adversarial attacks. Techniques such as adversarial training, defensive distillation, and anomaly detection mechanisms can be explored to strengthen the security of machine learning models used in network monitoring. Additionally, integrating advanced deep learning architectures, real-time traffic monitoring systems, and adaptive security mechanisms may further enhance the ability of intrusion detection systems to detect and prevent sophisticated cyberattacks in modern network environments [2], [8], [10].

REFERENCES

1. O. Ibitoye, R. Abou-Khamis, M. El Shehaby, A. Matrawy, and M. O. Shafiq, “The threat of adversarial attacks against machine learning in network security: A survey,” Journal of Electronics and Electrical Engineering, vol. 4, no. 1, pp. 16–59, 2025.
2. N. Akhtar, A. Mian, N. Kardan, and M. Shah, “Advances in adversarial attacks and defenses in computer vision: A survey,” IEEE Access, vol. 9, pp. 155161–155196, 2021.
3. K. Eykholt et al., “Robust physical-world attacks on deep learning visual classification,” in Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR), Salt Lake City, UT, USA, Jun. 2018, pp. 1625–1634.
4. A. Mahfouz, D. Venugopal, and S. Shiva, “Comparative analysis of machine learning classifiers for network intrusion detection,” in Proc. International Congress on Information and Communication Technology (ICICT), London, U.K., Aug. 2019.
5. Z. Ahmad, A. S. Khan, C. W. Shiang, J. Abdullah, and F. Ahmad, “Network intrusion detection system: A systematic study of machine learning and deep learning approaches,” Transactions on Emerging Telecommunications Technologies, vol. 32, no. 1, 2021.
6. M. Al Lail, A. Garcia, and S. Olivo, “Machine learning for network intrusion detection—A comparative study,” Future Internet, vol. 15, no. 7, art. no. 243, 2023.
7. H. Jmila and M. I. Khedher, “Adversarial machine learning for network intrusion detection: A comparative study,” Computer Networks, vol. 214, art. no. 109073, 2022.
8. S. Sharma and Z. Chen, “A systematic study of adversarial attacks against network intrusion detection systems,” Electronics, vol. 13, no. 24, art. no. 5030, 2024.
9. E. Tabassi, K. J. Burns, M. Hadjimichael, A. D. Molina-Markham, and J. T. Sexton, “A taxonomy and terminology of adversarial machine learning,” NIST Interagency Report (NIST IR), 2019, pp. 1–29.
10. H. Khazane, M. Ridouani, F. Salahdine, and N. Kaabouch, “A holistic review of machine learning adversarial attacks in IoT networks,” Future Internet, vol. 16, no. 1, art. no. 32, 2024.