Evaluating Network Intrusion Detection Models for Enterprise Security: Adversarial Vulnerability and Robustness Analysis

Vahid Heydari[1][a] and Kofi Nyarko[2][b]
[1]*Computer Science Department, Morgan State University, Baltimore, U.S.A.*
[2]*Electrical and Computer Engineering Department, Morgan State University, Baltimore, U.S.A.*
fi

Keywords: Adversarial Attacks, Machine Learning, Network Intrusion Detection Systems (NIDS), Cybersecurity, Model Robustness.

Abstract: Machine learning (ML) has become essential for securing enterprise information systems, particularly through its integration in Network Intrusion Detection Systems (NIDS) for monitoring and detecting suspicious activities. Although ML-based NIDS models demonstrate high accuracy in detecting known and novel threats, they remain vulnerable to adversarial attacks—small perturbations in network data that mislead the model into classifying malicious traffic as benign, posing serious risks to enterprise security. This study evaluates the adversarial robustness of two machine learning models—a Random Forest classifier and a Neural Network—trained on the UNSW-NB15 dataset, which represents complex, enterprise-relevant network traffic. We assessed the performance of both models on clean and adversarially perturbed test data, with adversarial samples generated via Projected Gradient Descent (PGD) across multiple epsilon values. Although both models achieved high accuracy on clean data, even minimal adversarial perturbations led to substantial declines in detection accuracy, with the Neural Network model showing a more pronounced degradation compared to the Random Forest. Higher perturbations reduced both models' performance to near-random levels, highlighting the particular susceptibility of Neural Networks to adversarial attacks. These findings emphasize the need for adversarial testing to ensure NIDS robustness within enterprise systems. We discuss strategies to improve NIDS resilience, including adversarial training, feature engineering, and model interpretability techniques, providing insights for developing robust NIDS capable of maintaining security in enterprise environments.

1 INTRODUCTION

With the exponential growth of digital networks and the rise of interconnected enterprise systems, the sophistication and frequency of cyberattacks have increased substantially. This makes robust network security a fundamental requirement for protecting enterprise information systems, which are often prime targets due to the sensitive data they manage. Network Intrusion Detection Systems (NIDS) play a crucial role in enterprise security by continuously monitoring network traffic for malicious activities, relying on patterns and anomalies to detect potential threats. In recent years, machine learning (ML) has become a transformative tool in the development of adaptive and efficient NIDS models capable of identifying both known and novel attack patterns (Sharafaldin et al., 2018; Lippmann et al., 2000; Buczak and Guven, 2016). Despite the strong performance of ML-based NIDS in controlled conditions, these models are vulnerable to adversarial attacks—strategic perturbations designed to deceive the model into misclassifying malicious traffic as benign, potentially exposing enterprise networks to undetected breaches (Goodfellow et al., 2015; Rigaki and Garcia, 2018).

A primary limitation of traditional NIDS evaluation is that it often assesses model performance solely on clean, unperturbed data. High accuracy on such data can create a misleading sense of reliability, suggesting that the model is resilient in dynamic, real-world settings. However, recent research has demonstrated that even minimal adversarial perturbations can severely compromise ML-based NIDS performance, undermining their effectiveness within live environments (Papernot et al., 2016; Kurakin et al., 2017). This vulnerability highlights the critical need to evaluate these models not only with tradi-

[a] https://orcid.org/0000-0002-6181-6826
[b] https://orcid.org/0000-0002-7481-5080

Heydari, V. and Nyarko, K.
Evaluating Network Intrusion Detection Models for Enterprise Security: Adversarial Vulnerability and Robustness Analysis.
DOI: 10.5220/0013204900003929
In *Proceedings of the 27th International Conference on Enterprise Information Systems (ICEIS 2025) - Volume 1*, pages 699-708
ISBN: 978-989-758-749-8; ISSN: 2184-4992

tional metrics, such as accuracy and the Area Under the Receiver Operating Characteristic Curve (AUC), but also for their robustness under adversarial conditions to ensure secure deployment within enterprise networks.

Adversarial attacks are particularly relevant to the cybersecurity domain because they exploit the model's weaknesses in a way that mimics real-world attack scenarios. For instance, attackers can manipulate traffic features to bypass detection while maintaining the functional integrity of their malicious activities. This capability poses a serious threat, as the model's inability to detect such adversarially modified samples can result in undetected breaches. The work of (Carlini and Wagner, 2017) demonstrated that adversarial attacks, even with minimal perturbations, could evade state-of-the-art NIDS, leading to significant drops in detection accuracy and, consequently, network security.

To investigate the impact of adversarial samples on NIDS, we selected the UNSW-NB15 dataset, a comprehensive dataset specifically designed for evaluating NIDS performance on modern attack types and diverse traffic features. This dataset contains both normal and attack traffic generated in a controlled environment using the IXIA PerfectStorm tool and includes a range of modern threats and benign network behaviors. Compared to older datasets, such as KDDCUP99, UNSW-NB15 better represents current network security challenges, making it suitable for evaluating adversarial robustness in NIDS models (Moustafa and Slay, 2015; Moustafa and Slay, 2016; Moustafa et al., 2019; Moustafa et al., 2017; Sarhan et al., 2021).

In this study, we evaluated the adversarial robustness of two machine learning models—a Random Forest classifier and a Neural Network—trained on the UNSW-NB15 dataset. We employed Projected Gradient Descent (PGD), a widely-used method for generating adversarial samples, to perturb the test data across different magnitudes. Both models were assessed on clean data as well as on adversarial samples generated with various levels of perturbation (epsilon values). While both models achieved high accuracy on clean data, our findings reveal that even small adversarial perturbations (epsilon = 0.01) significantly reduced detection accuracy, with the Neural Network demonstrating a more pronounced vulnerability compared to the Random Forest. These results underscore that traditional evaluation metrics alone do not fully capture a model's resilience to adversarial attacks.

The remainder of this paper is structured as follows. Section 2 reviews related work on ML-based NIDS and adversarial attacks. Section 3 presents our methodology, detailing the dataset, preprocessing steps, and adversarial sample generation. Section 4 presents the experimental results comparing both Random Forest and Neural Network models on clean and adversarial data, while Section 5 discusses the implications of our findings and future research directions. Finally, Section 6 concludes the paper.

2 RELATED WORK

2.1 Machine Learning Techniques for NIDS

ML techniques are widely employed in Network Intrusion Detection Systems NIDS due to their ability to detect malicious traffic patterns in network data. Common ML models applied in NIDS include Decision Trees, Random Forests, Support Vector Machines (SVMs), and Neural Networks. These models leverage features such as network flow, protocol types, and packet counts to classify traffic as benign or malicious.

Decision Trees. Decision Trees are popular in NIDS for their interpretability and efficiency in categorizing network traffic. They are effective in identifying threats by analyzing distinct behaviors within network features (Ullah et al., 2020; Khammas, 2020). However, Decision Trees are prone to overfitting, especially with high-dimensional data.

Random Forests. Random Forests, an ensemble learning technique, address the limitations of Decision Trees by generating multiple trees based on different subsets of the data, thus reducing overfitting and enhancing generalization. This makes Random Forests effective in detecting network attacks and well-suited for handling complex datasets with numerous features (Ullah et al., 2020; Khammas, 2020; Akhtar and Feng, 2022).

Support Vector Machines. Support Vector Machines (SVMs) are particularly useful for classifying data in high-dimensional feature spaces, often required in network traffic analysis (Ghouti and Imam, 2020; Arunkumar and Kumar, 2023). By learning optimal hyperplanes, SVMs can effectively separate various types of network traffic, making them suitable for detecting nuanced attack patterns.

Neural Networks. Neural Networks, particularly deep learning architectures, are powerful tools for NIDS due to their capacity to automatically extract relevant patterns from raw network data (Madani et al., 2022; Arivudainambi et al., 2019). Although they offer high performance, Neural Networks require substantial computational resources and often lack interpretability, which can limit their applicability in cybersecurity settings where model transparency is essential.

While these ML models perform well on clean data, recent studies indicate that traditional metrics such as accuracy and AUC may overestimate a model's effectiveness if adversarial robustness is not considered. Feature selection techniques, including Principal Component Analysis (PCA) and Correlation Analysis, are frequently employed to reduce data dimensionality and redundancy, potentially enhancing model performance. PCA, for example, emphasizes features that explain the most variance, while Correlation Analysis identifies and removes highly correlated features (Arivudainambi et al., 2019; Kok et al., 2019).

2.2 Adversarial Attacks on Machine Learning Models for NIDS

Adversarial attacks have emerged as a significant threat to ML-based NIDS models. These attacks introduce small, carefully crafted perturbations into input data, leading the model to misclassify malicious samples as benign, thereby exposing network security vulnerabilities.

One common adversarial attack method is the Fast Gradient Sign Method (FGSM), introduced by Goodfellow et al. (Goodfellow et al., 2015). FGSM generates adversarial samples by calculating the gradient of the model's loss concerning the input features and adding perturbations along the gradient direction. Though computationally efficient, FGSM is a single-step attack, limiting its ability to evade more robust defense mechanisms.

The PGD method, an iterative extension of FGSM, has become a standard for evaluating adversarial robustness (Madry et al., 2019; Chen and Hsieh, 2023). PGD recalculates the gradient in each iteration, incrementally adjusting the perturbations to move towards the decision boundary. This iterative approach allows PGD to exploit model weaknesses more effectively than FGSM, making it a preferred method for adversarial testing. Gressel et al. (Gressel et al., 2023) demonstrated PGD's effectiveness in bypassing ML models by applying controlled perturbations within epsilon boundaries to maintain sample plausibility. Shirazi et al. (Shirazi et al., 2019) also showed the potential of adversarial attacks to deceive phishing detection models, highlighting the importance of adversarial robustness in security applications.

2.3 Contributions of this Study

Our study advances current understanding of adversarial vulnerabilities in NIDS by offering a comparative analysis of two models—a Random Forest and a Neural Network—on the UNSW-NB15 dataset under PGD attacks. Key contributions include:

- **Model Comparison under Adversarial Conditions:** We assess how Random Forest and Neural Network models perform under adversarial perturbations, providing insights into the Neural Network's heightened vulnerability.
- **Realistic Perturbation Strategy with Customized Epsilon Values:** Our feature-specific epsilon calculation method tailors perturbations to each feature's range, ensuring that adversarial samples remain contextually valid.
- **Future Research Directions for NIDS Robustness:** We propose research avenues such as adversarial training, feature engineering, and interpretability methods to enhance model resilience against adversarial attacks.

By providing a focused analysis of PGD's impact and highlighting the Neural Network model's vulnerability, this study underscores the critical need for robust adversarial defenses in ML-based NIDS applications.

3 METHODOLOGY

3.1 Dataset Description

For this study, we utilized the *UNSW-NB15* dataset, a comprehensive dataset created specifically for evaluating NIDS in realistic enterprise and network security environments. The UNSW-NB15 dataset was generated using the IXIA PerfectStorm tool, which emulates complex, real-world network traffic by generating both normal and malicious activities. This tool enabled the capture of over 100 GB of data across two sessions, producing a dataset that reflects a wide range of normal and abnormal network behaviors commonly observed in modern infrastructures.

The dataset consists of 49 features characterizing various aspects of network traffic flows, including protocol type, flow duration, packet size, TCP

flags, source and destination IP addresses, and source and destination port numbers. These features were selected to represent both network layer and application layer characteristics, making the dataset suitable for training and evaluating NIDS models across multiple network attack scenarios. Additionally, the dataset includes contextual features that capture session-level information, aiding in the detection of complex, multi-stage attacks.

Each sample in the dataset is labeled as either *Normal* or one of nine attack categories, encompassing a wide spectrum of attack types commonly encountered in enterprise and network environments. These categories include:

- **Fuzzers.** Tools or techniques that automate random input generation to discover vulnerabilities.
- **Backdoors.** Techniques that provide attackers with unauthorized remote access to a system.
- **Denial of Service (DoS).** Attacks aimed at disrupting service availability by overwhelming resources.
- **Reconnaissance.** Methods used by attackers to gather information about a system or network.
- **Shellcode.** Payloads used for command execution, typically as part of an exploit.
- **Worms.** Self-replicating malware that spreads across networks.
- **Exploits, Analysis, and Generic Attacks.** Additional attack types targeting known software vulnerabilities and general malicious activities.

The UNSW-NB15 dataset addresses limitations found in older datasets like KDDCUP99 and NSL-KDD, which often focus on outdated or simplified attack vectors. By incorporating attack data from the Common Vulnerabilities and Exposures (CVE) database, UNSW-NB15 offers a more diverse and representative set of modern attack patterns, making it applicable to contemporary network security challenges. Given its extensive feature set and realistic representation of network traffic, UNSW-NB15 serves as a robust benchmark for testing NIDS model performance in complex, multi-faceted scenarios.

3.2 Data Preprocessing

Before training and testing the machine learning models, we applied the following preprocessing steps:

- **Categorical Features Encoding.** Categorical features such as *proto*, *service*, *state*, and *attack_cat* were transformed into numerical representations using Label Encoding to ensure that they could be effectively interpreted by the models.
- **Handling Missing and Infinite Values.** The dataset was examined for missing or infinite values. Missing values were replaced with the mean of the respective feature, while infinite values were clipped within valid feature ranges.
- **Train-Test Split.** The dataset was divided into training and testing sets with a stratified 70-30 split to ensure proportional representation of both normal and attack samples in each set.

3.3 Adversarial Sample Generation

We employed the *PGD* method to generate adversarial samples by perturbing the original test set. PGD is an iterative, gradient-based attack that modifies feature values to deceive the model into making incorrect predictions.

3.3.1 Feature-Specific Epsilon Calculation

To maintain realistic and proportional perturbations across features, we calculated ε as a fraction of each feature's range. For each feature f, epsilon was set as:

$$\varepsilon_f = \text{scaling factor} \times (\max(f) - \min(f))$$

We experimented with scaling factors of 0.01, 0.05, 0.1, and 0.15, corresponding to perturbations representing 1%, 5%, 10%, and 15% of the feature's range, respectively. This approach ensured that the adversarial samples remained plausible within the context of network traffic data.

3.3.2 PGD Attack

The PGD attack was iterated for 150 steps, incrementally modifying feature values within the calculated epsilon boundaries. All perturbed values were clipped within valid feature ranges to retain realistic traffic characteristics. The adversarial samples were subsequently evaluated on both the Random Forest and Neural Network models to assess the impact of adversarial perturbations.

3.4 Model Training and Evaluation

To assess adversarial robustness, we trained and evaluated both a *Random Forest Classifier* and a *Neural Network* model. This comparative approach allowed us to investigate differences in vulnerability between the two models.

- **Initial Evaluation on Clean Data.** Each model was first evaluated on the clean test set to establish a baseline performance. The evaluation metrics included accuracy, precision, recall, and AUC.
- **Evaluation on Adversarial Data.** For both models, adversarial test sets were generated using the different epsilon values. We then evaluated each model's performance on these adversarial samples to determine their robustness against varying levels of perturbation.

The results of these evaluations, including accuracy and AUC comparisons for each model and each level of epsilon, are presented in Section 4. The comparative analysis provides insights into the differential impact of adversarial attacks on Random Forest and Neural Network models, underscoring the critical need for robust adversarial defenses in ML-based NIDS.

4 RESULTS

In evaluating the performance of our Random Forest Classifier and Neural Network on both clean and adversarial test data, we report several key metrics: Accuracy, Precision, Recall, AUC, and Receiver Operating Characteristic (ROC) Curve.

Accuracy. measures the overall correctness of the model, representing the proportion of correct predictions among all predictions.

Precision. indicates the accuracy of positive predictions, specifically for attack samples, showing the proportion of true positives out of all predicted positives.

Recall. (or Sensitivity) measures the model's ability to correctly identify attack samples, representing the proportion of true positives among all actual positives.

AUC. represents the model's ability to differentiate between classes across various decision thresholds, where higher AUC values imply better discrimination between normal and attack samples.

ROC Curve. plots the True Positive Rate (sensitivity) against the False Positive Rate, showing the trade-off between sensitivity and specificity at different threshold levels. A higher curve (closer to the top left) indicates stronger classification performance.

4.1 Clean Data Results

On the clean test set, the Random Forest Classifier and Neural Network achieved the following metrics:

- **Random Forest:**
  - **Accuracy:** 0.87
  - **Precision:** 0.86
  - **Recall:** 0.88
  - **AUC:** 0.99
- **Neural Network:**
  - **Accuracy:** 0.79
  - **Precision:** 0.80
  - **Recall:** 0.75
  - **AUC:** 0.77

These results indicate that the Random Forest model outperformed the Neural Network model on clean data, achieving higher accuracy, precision, recall, and AUC.

4.2 Adversarial Data Results

Adversarial samples were generated using four different epsilon values ($\varepsilon = 0.01$, $\varepsilon = 0.05$, $\varepsilon = 0.1$, and $\varepsilon = 0.15$), each representing increasing levels of perturbation. Both models experienced performance degradation as epsilon increased, illustrating their vulnerability to adversarial attacks.

4.2.1 Epsilon = 0.01

- **Random Forest:**
  - **Accuracy:** 0.70
  - **AUC:** 0.77
- **Neural Network:**
  - **Accuracy:** 0.50
  - **AUC:** 0.56

At $\varepsilon = 0.01$, both models experienced a moderate drop in accuracy and AUC, with the Random Forest retaining more robustness compared to the Neural Network.

4.2.2 Epsilon = 0.05

- **Random Forest:**
  - **Accuracy:** 0.49
  - **AUC:** 0.55
- **Neural Network:**
  - **Accuracy:** 0.48
  - **AUC:** 0.52

At $\varepsilon = 0.05$, both models saw a significant drop in accuracy, with Random Forest slightly outperforming the Neural Network. The AUC values for both models indicate a weakened ability to distinguish between normal and attack samples.

4.2.3 Epsilon = 0.10

- **Random Forest:**
  - **Accuracy:** 0.47
  - **AUC:** 0.50
- **Neural Network:**
  - **Accuracy:** 0.47
  - **AUC:** 0.50

With $\varepsilon = 0.10$, both models approached random guessing levels of performance, with AUC values near 0.50. This result indicates that the adversarial perturbations have severely compromised the model's ability to discriminate between classes.

4.2.4 Epsilon = 0.15

- **Random Forest:**
  - **Accuracy:** 0.46
  - **AUC:** 0.50
- **Neural Network:**
  - **Accuracy:** 0.47
  - **AUC:** 0.50

At the highest perturbation level ($\varepsilon = 0.15$), both models performed near random.

4.3 Comparative Analysis

The results demonstrate that while the Random Forest model initially performed better on clean data, both models exhibited significant vulnerability to adversarial perturbations. The degradation in accuracy and AUC as epsilon increased underscores the susceptibility of ML-based NIDS to adversarial attacks.

These results indicate that the Random Forest model outperformed the Neural Network model on clean data, achieving higher accuracy, precision, recall, and AUC. Figure 1 presents a side-by-side comparison of accuracy for both models under clean and adversarial conditions, with Subfigure 1a showing the Random Forest model's performance and Subfigure 1b displaying the Neural Network's accuracy. This comparison emphasizes the Random Forest model's greater resilience on clean data.

Figure 2 further visualizes the ROC curves for each model across varying perturbation levels, capturing the effects of different epsilon values on classification robustness. Subfigure 2a presents the ROC curve for the Random Forest model, demonstrating high sensitivity and specificity on clean data, while Subfigure 2b displays the Neural Network's ROC curve, which highlights its lower robustness to adversarial perturbations. At lower epsilon values, the Random Forest model showed slightly higher resilience than the Neural Network, yet both models ultimately failed to maintain effectiveness under larger perturbations, with accuracy approaching random guessing at higher epsilon levels.

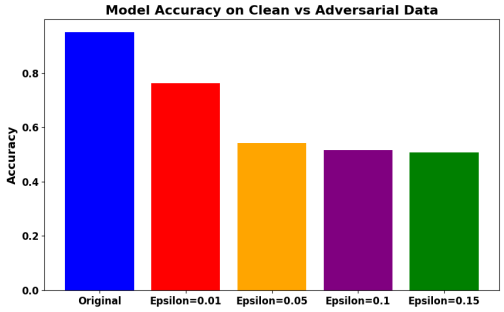
Together, Figures 1 and 2 underscore the need for robust adversarial defenses in NIDS, as both models, regardless of architecture, showed vulnerability to adversarial attacks. This analysis highlights the importance of adversarial robustness as a core design objective for effective NIDS development.

5 DISCUSSION

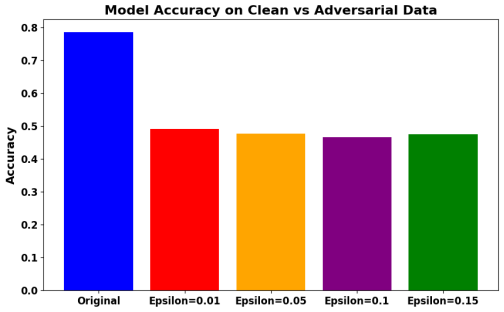
The results of our study underscore a significant and often overlooked challenge in network intrusion detection: high classification accuracy on clean data does not necessarily imply robustness against adversarial attacks. Using the UNSW-NB15 dataset, both our Random Forest classifier and Neural Network initially demonstrated strong performance on clean data, achieving accuracy and AUC values of 87% and 0.99, and 79% and 0.77, respectively. However, our analysis reveals that even small adversarial perturbations (as low as $\varepsilon = 0.01$) led to considerable performance degradation across both models. This finding highlights the susceptibility of machine learning models to adversarial attacks and suggests that robust defense mechanisms are essential for real-world deployment in cybersecurity applications.

5.1 Implications of Adversarial Vulnerability

The effectiveness of PGD attacks in degrading model performance—especially as the epsilon value increases—demonstrates a concerning vulnerability in NIDS to even minimal adversarial perturbations. With $\varepsilon = 0.01$ (representing a 1% perturbation relative to each feature's range), both models saw substantial drops in performance. For example, the Random Forest classifier's accuracy dropped from 87% to 70%, and the AUC decreased to 0.77. The Neural Network experienced an even sharper drop, with accu-

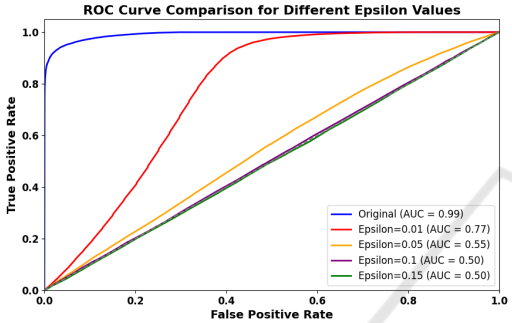


(a) Random Forest Model Accuracy

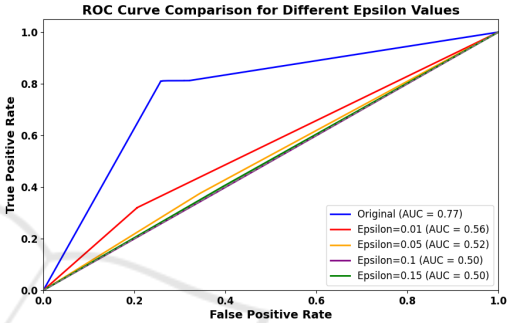


(b) Neural Network Model Accuracy

Figure 1: Model Accuracy on Clean vs. Adversarial Data for Random Forest and Neural Network Models.



(a) Random Forest Model ROC Curve



(b) Neural Network Model ROC Curve

Figure 2: ROC Curve Comparison for Random Forest and Neural Network Models.

racy falling to 50% and AUC to 0.56, highlighting that slight modifications in the input data can mislead both models. As epsilon values increased to 0.05 and beyond, performance deteriorated to near-random classification levels, demonstrating the models' inability to maintain reliable decision boundaries under adversarial conditions.

This vulnerability is particularly alarming in network security, where attackers could exploit these weaknesses by crafting low-visibility adversarial samples that evade detection while preserving the malicious functionality. For instance, subtle changes in network flow attributes like packet size or connection duration could allow attackers to bypass NIDS without altering the attack's objectives. This evasion could lead to undetected breaches in enterprise environments, where NIDS often serve as the first line of defense.

Our findings emphasize that high initial performance in ML-based NIDS, such as the Random Forest and Neural Network classifiers in our study, does not guarantee resilience under adversarial conditions. Without adversarial testing, deploying these models could create a false sense of security, leaving critical infrastructures exposed to sophisticated threats. Consequently, models used in security applications should be rigorously evaluated for both clean data performance and adversarial robustness to ensure their reliability in high-stakes environments.

5.2 Challenges in Ensuring Robustness

Despite the use of a comprehensive dataset like UNSW-NB15, our research highlights that achieving adversarial robustness in NIDS remains challenging. Many machine learning algorithms, including Random Forest and Neural Networks, are developed primarily to optimize accuracy on clean data without specific mechanisms to withstand adversarial attacks. This discrepancy means that traditional evaluation metrics, such as accuracy and AUC, may overestimate model performance in real-world threat scenarios where attackers may craft data to evade detection. Thus, developing models resilient to adversarial perturbations is crucial for applications in high-stakes settings, where attack tactics are constantly evolving.

Adversarial Training. Adversarial training, which incorporates adversarial examples into the training dataset, can improve robustness by helping models recognize and correctly classify perturbed samples.

However, this approach has significant drawbacks: it requires extensive computational resources, which may not be feasible for real-time NIDS applications, and it may not generalize well to new, unseen perturbations, leaving models vulnerable to novel attack strategies.

Robust Feature Engineering. Identifying and emphasizing features less susceptible to adversarial manipulation is another promising approach, but it is also constrained by the need for extensive domain knowledge to select robust features. Feature engineering might inadvertently eliminate useful information, potentially reducing overall model performance if essential features are removed due to their sensitivity to perturbations.

High-Dimensional and Complex Data Challenges. NIDS models also face the challenge of handling the high-dimensional and complex nature of network traffic data in adversarial settings. The rise of stealthy, low-footprint attacks—those with minimal, nearly undetectable modifications—presents additional difficulties. These attacks can subtly alter features to evade detection while retaining their functionality, requiring models that can discern such subtle differences without compromising clean data performance.

Adaptive Adversaries. Lastly, adaptive attackers who modify tactics based on observed defenses further complicate robustness efforts. A robust NIDS must withstand a broad range of perturbations while adapting to continuously evolving threat vectors. Static defense mechanisms may fall short in such scenarios, underscoring the need for models that dynamically adapt to adversarial strategies.

Addressing these challenges is crucial for developing robust NIDS models capable of maintaining high levels of security in adversarial environments. Potential solutions include dynamic model adaptation, ensemble methods, and enhanced feature engineering, which together could mitigate these challenges and enhance robustness.

5.3 Future Research Directions

Our findings point to several promising areas for future research to enhance NIDS robustness against adversarial attacks:

- **Adversarial Training.** Future work could explore adaptive adversarial training techniques tailored to evolving network attack patterns, focusing on dynamically generated adversarial examples across a range of epsilon values.
- **Defense-Guided Feature Engineering.** Future studies could develop feature engineering techniques that target sensitive features, reducing the model's reliance on easily manipulated inputs while maintaining relevance to network security.
- **Hybrid NIDS Models.** Exploring hybrid architectures that combine machine learning with traditional rule-based detection systems may offer resilience by adding an additional layer of filtering. For instance, ensemble methods that integrate deep learning and rule-based approaches might reduce the impact of adversarial perturbations.
- **Explainable AI and Model Interpretability.** Leveraging interpretability techniques to understand model decisions under adversarial conditions could provide insights into which features are most vulnerable, guiding future defense mechanisms and model design strategies.
- **Standardized Adversarial Testing Benchmarks.** Establishing standardized metrics and benchmarks for adversarial testing in NIDS models would enable researchers to make meaningful comparisons across studies, evaluating models on multiple dimensions that balance accuracy and robustness.

5.4 Limitations of the Study

While our research sheds light on adversarial vulnerability in network intrusion detection, certain limitations must be acknowledged. Our study used a Random Forest and a relatively simple Neural Network classifier, which, while effective, may differ in robustness compared to more advanced deep learning models. Additionally, the UNSW-NB15 dataset, though comprehensive, may not fully represent the range of adversarial strategies that could be encountered in modern network attacks. Future studies could expand upon our work by testing additional datasets, adversarial techniques, and ML models, contributing to a more holistic understanding of adversarial robustness in NIDS.

6 CONCLUSION

This study investigated the robustness of machine learning-based NIDS against adversarial attacks, specifically examining the impact of adversarial perturbations on models trained on the UNSW-NB15 dataset. By generating adversarial samples using

PGD with feature-specific epsilon calculations, we evaluated the resilience of both Random Forest and Neural Network classifiers. The results reveal that even high-performing models on clean data are significantly susceptible to adversarial attacks, underscoring a critical challenge for the deployment of ML-based NIDS in real-world environments.

Key findings include:

- **Vulnerability to Adversarial Attacks.** Both Random Forest and Neural Network models, despite achieving high accuracy on unperturbed data, showed notable performance degradation when evaluated on adversarial samples. This vulnerability highlights a substantial risk in cybersecurity, where attackers can exploit these weaknesses to bypass detection systems with minimal perturbations.
- **Impact of Adversarial Perturbation Scale.** As perturbation levels (epsilon values) increased, both models' accuracy and AUC dropped markedly, with the Neural Network showing greater sensitivity to smaller perturbations. This comparative analysis indicates that while model performance may vary by architecture, neither model proved resilient under adversarial conditions, emphasizing the need for adversarial testing and defense mechanisms.
- **Limitations of Retraining Strategies.** While retraining on adversarial samples is a promising approach, it often introduces new feature dependencies that could be exploited by adaptive attackers. This suggests that while adversarial retraining can improve robustness to some extent, it may not provide comprehensive protection against evolving threats.
- **Need for Continuous Adaptation and Evaluation.** Our study underscores the importance of ongoing evaluation and adaptation of ML models in cybersecurity, as static models are insufficient in the face of adaptive adversarial strategies. NIDS models must incorporate dynamic and robust defense techniques to maintain security in high-risk environments.

In summary, while machine learning models are essential for enhancing cybersecurity, their vulnerability to adversarial attacks remains a significant challenge. Future work should explore more adaptive and resilient approaches, including hybrid architectures, continuous adversarial training, and interpretability techniques, to bolster NIDS models against sophisticated and evolving adversarial tactics. This study serves as a call to action for the development of robust and secure NIDS models that can withstand adversarial manipulations while providing reliable protection within enterprise and critical infrastructure networks.

ACKNOWLEDGEMENTS

This work is supported in part by the Center for Equitable Artificial intelligence and Machine Learning Systems (CEAMLS) at Morgan State University. This paper benefited from the use of OpenAI's ChatGPT for language enhancement, including grammar corrections, rephrasing, and stylistic refinements. All AI-assisted content was subsequently reviewed and approved by the authors to ensure technical accuracy and clarity.

REFERENCES

Akhtar, M. S. and Feng, T. (2022). Malware analysis and detection using machine learning algorithms. *Symmetry*, 14(11):2304.

Arivudainambi, D., KA, V. K., Visu, P., et al. (2019). Malware traffic classification using principal component analysis and artificial neural network for extreme surveillance. *Computer Communications*, 147:50–57.

Arunkumar, M. and Kumar, K. A. (2023). Gosvm: Gannet optimization based support vector machine for malicious attack detection in cloud environment. *International Journal of Information Technology*, 15(3):1653–1660.

Buczak, A. L. and Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. *IEEE Communications Surveys & Tutorials*, 18(2):1153–1176.

Carlini, N. and Wagner, D. (2017). Adversarial examples are not easily detected: Bypassing ten detection methods.

Chen, P.-Y. and Hsieh, C.-J. (2023). Preface. In Chen, P.-Y. and Hsieh, C.-J., editors, *Adversarial Robustness for Machine Learning*, pages xiii–xiv. Academic Press.

Ghouti, L. and Imam, M. (2020). Malware classification using compact image features and multiclass support vector machines. *IET Information Security*, 14(4):419–429.

Goodfellow, I. J., Shlens, J., and Szegedy, C. (2015). Explaining and harnessing adversarial examples.

Gressel, G., Hegde, N., Sreekumar, A., Radhakrishnan, R., Harikumar, K., S., A., and Achuthan, K. (2023). Feature importance guided attack: A model agnostic adversarial attack.

Khammas, B. M. (2020). Ransomware detection using random forest technique. *ICT Express*, 6(4):325–331.

Kok, S., Abdullah, A., Jhanjhi, N., and Supramaniam, M. (2019). Ransomware, threat and detection techniques: A review. *Int. J. Comput. Sci. Netw. Secur*, 19(2):136.

Kurakin, A., Goodfellow, I., and Bengio, S. (2017). Adversarial machine learning at scale.

Lippmann, R., Haines, J. W., Fried, D. J., Korba, J., and Das, K. (2000). The 1999 darpa off-line intrusion detection evaluation. *Computer Networks*, 34(4):579–595. Recent Advances in Intrusion Detection Systems.

Madani, H., Ouerdi, N., Boumesaoud, A., and Azizi, A. (2022). Classification of ransomware using different types of neural networks. *Scientific Reports*, 12(1):4770.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. (2019). Towards deep learning models resistant to adversarial attacks.

Moustafa, N., Creech, G., and Slay, J. (2017). *Big Data Analytics for Intrusion Detection System: Statistical Decision-Making Using Finite Dirichlet Mixture Models*, page 127–156. Springer International Publishing.

Moustafa, N. and Slay, J. (2015). Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set). In *2015 Military Communications and Information Systems Conference (MilCIS)*, pages 1–6.

Moustafa, N. and Slay, J. (2016). The evaluation of network anomaly detection systems: Statistical analysis of the unsw-nb15 data set and the comparison with the kdd99 data set. *Information Security Journal: A Global Perspective*, 25(1–3):18–31.

Moustafa, N., Slay, J., and Creech, G. (2019). Novel geometric area analysis technique for anomaly detection using trapezoidal area estimation on large-scale networks. *IEEE Transactions on Big Data*, 5(4):481–494.

Papernot, N., McDaniel, P., Wu, X., Jha, S., and Swami, A. (2016). Distillation as a defense to adversarial perturbations against deep neural networks.

Rigaki, M. and Garcia, S. (2018). Bringing a gan to a knife-fight: Adapting malware communication to avoid detection. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 70–75.

Sarhan, M., Layeghy, S., Moustafa, N., and Portmann, M. (2021). *NetFlow Datasets for Machine Learning-Based Network Intrusion Detection Systems*, page 117–135. Springer International Publishing.

Sharafaldin, I., Lashkari, A. H., and Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. In *International Conference on Information Systems Security and Privacy*.

Shirazi, H., Bezawada, B., Ray, I., and Anderson, C. (2019). Adversarial sampling attacks against phishing detection. In *Database Security*.

Ullah, F., Javaid, Q., Salam, A., Ahmad, M., Sarwar, N., Shah, D., and Abrar, M. (2020). Modified decision tree technique for ransomware detection at runtime through api calls. *Scientific Programming*, 2020.