International Journal of Advanced Engineering, Management and Science (IJAEMS)
Peer-Reviewed Journal
ISSN: 2454-1311 | Vol-11, Issue-5; Sep-Oct, 2025
Journal Home Page: https://ijaems.com/
DOI: https://dx.doi.org/10.22161/ijaems.115.9

Adversarial Robustness in AI-Driven Cybersecurity Solutions: Thwarting Evasion Assaults in Real-Time Detection Systems

Dr. Mohammed Musthafa

Department of Computer Science, Western Global University, USA
Chief Technology Officer, ZanX Technologies & Regional ICT Manager Gulf Area, Ligabue Group
ORCID Id: https://orcid.org/0009-0009-7446-7408

Received: 16 Aug 2025; Received in revised form: 14 Sep 2025; Accepted: 18 Sep 2025; Available online: 25 Sep 2025

Abstract— *The incorporation of Artificial Intelligence (AI), especially deep learning models, into cybersecurity frameworks has greatly improved the identification and mitigation of cyber threats. Nonetheless, these smart systems encounter a significant and rising threat – adversarial attacks. Malicious entities create subtle alterations in network traffic or system actions that mislead AI models into misidentifying threats as harmless, facilitating evasion tactics that can circumvent real-time intrusion detection systems (IDS). This study investigates the susceptibility of deep learning-based Intrusion Detection Systems (IDS) to adversarial examples and suggests a robust detection framework aimed at improving resilience against these evasion tactics. The suggested system merges adversarial training, input sanitization, and resilient model architectures, including adversarial-aware Convolutional Neural Networks (CNN) and defensive autoencoders. Employing benchmark datasets like CIC-IDS2017 and UNSW-NB15, we recreate various adversarial scenarios – created using Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD) – to evaluate the effect on detection performance. Experimental findings indicate that conventional DL models experience a significant decline in performance when exposed to adversarial circumstances, with accuracy decreasing by more than 20% in certain instances. Conversely, our suggested framework shows a noticeable enhancement in adversarial robustness, keeping more than 91% detection accuracy during attacks and considerably lowering false positives.*

Keywords— ***Cybersecurity, Intrusion Detection, Deep Learning, RNN, Transformer***

I. INTRODUCTION

The rapid advancement and integration of Artificial Intelligence (AI) and deep learning models into cybersecurity frameworks have ushered in a new era of intelligent threat detection and mitigation. Artificial intelligence systems that use deep neural networks are very good at finding complicated patterns in user behavior and network traffic. This lets you quickly find bad actions before they are found by traditional signature-based or heuristic detection. As they take on more important roles in the cybersecurity business, they have also shown that they can be vulnerable to attacks by bad actors. Evasion attacks are a kind of adversarial assault that change inputs so that AI models think that poor behavior is good. When AI-based cybersecurity solutions are used in the real world, this flaw makes them less reliable, strong, and safe.

Adversarial machine learning looks at how bad people can change machine learning systems.

A lot of people have been interested in this area of study in the last several years. Circumvention attacks

©2025 The Author(s). Published by Infogain Publication, This work is licensed under a Creative Commons Attribution 4.0 License. http://creativecommons.org/licenses/by/4.0/

are a special kind of danger in cybersecurity because they change the judgment boundaries of detection algorithms by making tiny changes to the input data.

These changes usually go unnoticed by human analysts, but they are enough to cause misclassification. Attacks like these can do a lot of harm to intrusion detection systems (IDS), which lets attackers get access networks, steal data, or disrupt services without setting off alerts. This is a big problem in real-time detection settings where immediate action is needed to stop more damage. Cybersecurity data is quite complicated, and cyber threats are getting worse very quickly, which makes it even harder for enemies to avoid detection.

System logs and network traffic are various from one another in significant and complex manners. It's very difficult to construct detection models that are robust. Also, the fact that methods of attack are constantly improving and becoming more complex, like polymorphic malware and advanced persistent threats, makes the models more robust and resilient.

Adversarial attacks exploit vulnerabilities by introducing harmful patterns into what appear to be normal communications or user actions, which is difficult for traditional detection software to discover. Deep learning frameworks, such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and, more recently, Transformer-based networks, have been instrumental in most AI-based Intrusion Detection Systems (IDS) since they are capable of learning hierarchical and temporal features directly from raw or lightly preprocessed data.

These models make use of statistical trends and decision boundaries learned from training data, which make them vulnerable to subtle yet carefully constructed input changes. Attackers can employ gradient-based approaches such as the Fast Gradient Sign Method (FGSM) or Projected Gradient Descent (PGD) in order to craft samples that manage to bypass defenses. The vulnerabilities of these attacks indicate that AI-powered cybersecurity algorithms are just as vulnerable; while such models enhance the detection of threats, their built-in susceptibility to evasion techniques undermines their reliability and effectiveness.

Thus, one of the main research areas in AI-based cybersecurity is making it more difficult for adversaries to continue. More specifically, designing detection techniques that are able to discover known threats as well as threats that have been altered on intent, as well as ensuring that they remain resilient during an attack and are discovered quickly and precisely.

There are a variety of methods with which to resist hostile attacks.

Adversarial training entails training models with adversarial examples so that they become stronger. Input sanitization and transformation strategies attempt to eliminate adversarial noise prior to classification, and architectural enhancements attempt to render models more resilient to input change. Every approach has its downsides, which may translate to increased processing expense, reduced model flexibility, or reduced resilience to novel hostile tactics.

It is exceedingly hard to keep real-time detection working while also putting in place defenses against attackers. Cybersecurity technology must work within strict latency limits so that problems may be found and fixed right away.

Some adversarial defense tactics make things harder or require more work, which makes them hard to use in real-time networks where efficiency and scalability are very important. Because of this, there is a big need for robust but light models made just for cybersecurity that can handle attacks. AI-powered Intrusion Detection Systems must not only be effective but also intelligible and transparent. Security experts need to know why they give warnings in certain scenarios, especially when false positives can lead to expensive investigations and problems with operations. Adversarial attacks make this even more important by hiding the reasons for their scary effects. Combining explainable AI (XAI) techniques with adversarial robustness tactics may make model outputs clearer, build trust, and make it easier to do forensic analysis during incident management.

The research community is actively focusing on the unification of several defense techniques into comprehensive frameworks that together bolster enemy resilience. Hybrid approaches that combine adversarial training with input preprocessing and architectural enhancements have demonstrated potential in achieving a balance between detection

©2025 The Author(s). Published by Infogain Publication, This work is licensed under a Creative Commons Attribution 4.0 License. http://creativecommons.org/licenses/by/4.0/

effectiveness and resilience. Adaptive protection mechanisms that use reinforcement learning and continuous learning let AI systems change their defenses when new attack methods are used. This is similar to the "arms race" that happens in cybersecurity.

This study focuses on developing and evaluating a comprehensive detection framework that is robust against adversarial assaults for real-time cybersecurity applications. Utilizing benchmark datasets like CIC-IDS2017 and UNSW-NB15, the system combines adversarial-aware convolutional neural networks with defensive autoencoders and strong input sanitization processes. The assessment replicates diverse evasion attack situations employing advanced adversarial example creation methods, methodically evaluating model performance decline and resilience enhancements. By means of thorough experimentation, this study seeks to deliver practical insights into efficient defenses against evasion attacks and create actionable recommendations for implementing secure AI-based IDS.

The rest of the paper is structured as follows: subsequent to this introduction, we examine pertinent studies in adversarial machine learning and cybersecurity; then, we outline the methodology encompassing dataset preparation, adversarial attack simulation, and model architecture; thereafter, experimental findings are presented and discussed; ultimately, conclusions are made along with a roadmap for future research avenues. This study enhances the field by connecting theoretical adversarial robustness with practical cybersecurity requirements, highlighting the essential importance of robust AI models in protecting digital infrastructures from advancing cyber threats.

II. LITERATURE REVIEW

Overview of Adversarial Machine Learning in Cybersecurity

Adversarial machine learning is a developing area that examines the weaknesses of AI and machine learning systems to intentionally designed inputs aimed at misleading the model. Adversarial assaults are a big threat to cybersecurity, especially when AI-powered intrusion detection systems (IDS) are used. I Adversarial attacks are different from conventional threats since they directly attack AI models by making alterations that cause them to misclassify or evade. People may not notice these little changes, but they are enough to trick the model into generating false predictions, which can lead to security issues.

Evasion attacks are the most common sort of threat in the realm of cyber security. Attackers modify bad traffic or payloads so that AI-based IDS can't see them. It's tougher to carry out these attacks now that it's easier to build adversarial instances, thanks to tools like the Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD). These strategies exploit the model's loss function's gradients to provide inputs that make the model's prediction error worse while still achieving its goal of destruction. You need to know these strategies and how they function in order to build good defense plans.

AI-Driven Intrusion Detection Systems and Their Weaknesses

AI and deep learning have altered how intrusion detection works by allowing systems find complicated patterns in huge datasets without having rules that people made. Some models that are highly good at finding sophisticated and zero-day attacks are Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), and Transformer architectures. But because these models depend on statistical learning, they are open to hostile inputs that take advantage of established decision boundaries.

Numerous studies have demonstrated the ineffectiveness of AI-driven Intrusion Detection Systems (IDS) against adversarial evasion. For example, Grosse et al. (2017) showed that adversarial samples made for deep neural networks might successfully get past malware classifiers. In a similar study, Huang et al. (2019) demonstrated that adversarial network traffic generated via gradient-based attacks might circumvent intrusion detection systems through deep learning. These results indicate that while AI enhances detection capabilities, it simultaneously generates new attack vectors that require attention.

Techniques for Adversarial Attacks in Cybersecurity

In cybersecurity, adversarial attacks mainly focus on feature representations derived from network traffic or system logs. Frequent tactics used in attacks consist of:

©2025 The Author(s). Published by Infogain Publication, This work is licensed under a Creative Commons Attribution 4.0 License. http://creativecommons.org/licenses/by/4.0/

1. **Fast Gradient Sign Method (FGSM**): Developed by Goodfellow and his team. (2015), FGSM generates adversarial examples by adding disturbances according to the gradient's sign of the loss function. It is computationally efficient, but produces rather simple adversarial examples
2. **Projected Gradient Descent (PGD):** As an iterative enhancement of FGSM, PGD executes several steps of minor disturbances, followed by a projection back into the legitimate input space. It is considered a more potent assault and a typical standard for adversarial resilience.
3. **Carlini & Wagner (C&W) Attacks:** These methods refine perturbations using a loss function crafted to reduce detection confidence, yielding potent yet computationally intensive adversarial instances.

These methods emphasize how easily attackers can create inputs that bypass AI detection when no protective measures exist.

Defensive Approaches to Counter Adversarial Attacks

To address adversarial threats, researchers have suggested multiple defense strategies focused on enhancing model resilience:

Adversarial Learning

Adversarial training is adding adversarial samples to the training dataset to help the model learn how to recognize and fight them. Goodfellow et al. (2015) started this technology, which has been very successful in image recognition tasks and is slowly being adapted for use in cybersecurity. Adversarial training requires substantial computational resources and may not be effective against all types of adversarial attacks, especially novel ones.

Input Alteration and Cleaning

Preprocessing techniques aim to remove or reduce adversarial changes from inputs before the detection model processes them. To make hostile inputs less effective, people have recommended using techniques like feature squeezing, input noise reduction, and randomization. Xu et al. (2017) introduced feature squeezing, a method that reduces input precision to limit adversarial noise, demonstrating improved resilience in malware detection. However, if these methods are used too strongly, they could make the model less effective on clean data.

Sturdy Model Designs

Certain work concentrates on modifying the architecture to enhance the built-in robustness of the model. Defensive autoencoders, which are designed to reconstruct inputs to eliminate noise, have been employed to eliminate adversarial perturbations. Other researchers have explored models that integrate convolutional layers with recurrent or attention mechanisms to acquire feature representations that are richer and less susceptible to adversarial attacks. The objective of such designs is to achieve a balance between efficiency in computing and strength that is appropriate for real-time systems.

Identifying Adversarial Examples

Another method to protect against attack is to train various models to identify if an input is malicious. These detectors search for unusual patterns in input space typical of adversarial noise. Detection models are promising but struggle to maintain low false positive rates and adapt to novel methods employed by attackers.

Difficulties in Real-Time Adversarial Protection

There are several issues with introducing adversarial resilience into real-time IDS. Security can slow the process of determining and correcting problems, making the system less practical. Furthermore, it is imperative to maintain low false positive rates operationally to prevent security analysts from becoming desensitized to alarms. It is imperative to use caution in balancing strength, precision, and quickness.

The nature of the attacks is ever evolving. The assailants consistently refine their tactics, necessitating that detecting systems perpetually enhance their defenses. Without online learning or adaptive approaches, static models quickly become useless. This illustrates the importance of systems that encourage ongoing improvement.

Upcoming Trends and Future Paths

Recent advancements suggest the potential for integrating several defense systems to leverage their synergistic benefits. Combining adversarial training with input filtering and tougher designs could give

©2025 The Author(s). Published by Infogain Publication, This work is licensed under a Creative Commons Attribution 4.0 License. http://creativecommons.org/licenses/by/4.0/

you many degrees of protection. Reinforcement learning techniques have been proposed to enable models to dynamically adjust defenses in response to attack attempt inputs. Explainable AI (XAI) is becoming more popular in this field since it lets you look at model decisions and see how bad changes affect them. When problems come up, clear models can keep people's attention, make them more comfortable, and help with investigations. Researchers are now looking into how well Attention-based Transformer models can stand up to cyberattacks. These models are known for their ability to generalize and represent features better than other types of models.

Their capacity to simulate long-range dependency can help find complicated and subtle threat indicators.

III. METHODOLOGY

This study introduces a thorough experimental methodology for evaluating and enhancing the adversarial resilience of deep learning-based intrusion detection systems (IDS).

The procedure encompasses dataset preparation, generation of adversarial assaults, model construction, implementation of adversarial defense, establishment of training configurations, and evaluation of model performance. The primary objective is to determine how various security techniques improve AI-based models in detecting evasion attacks within real cyber-protection contexts.

Research Design

The study employs a comparative and experimental framework in which various deep learning models are trained on both clean and adversarial modified data to evaluate their detection performance. The framework consists of three stages: training a baseline model lacking adversarial protections, simulating and assessing adversarial attacks, and applying defense strategies to improve model resilience. The assessment depends on the performance of classification in both standard and hostile settings

Data set Choice and Preparation

Two benchmark datasets were chosen for testing:

1. **CIC-IDS2017**: An extensive dataset offered by the Canadian Institute for Cybersecurity. It comprises both harmless and harmful network traffic reflecting modern attack patterns such as DDoS, brute-force, infiltration, botnet, and web-based attacks. The dataset emulates authentic user activity and network systems.
2. **UNSW-NB15**: Developed by the Australian Centre for Cyber Security, this dataset features nine distinct attack types in addition to regular traffic. It records actual contemporary network traffic along with application layer and payload details.

Preprocessing consisted of multiple essential stages

1. **Selection and Cleaning of Features:** Features that had constant values or missing data were eliminated. Unrelated attributes such as timestamps or non-numeric identifiers were removed.
2. **Normalization:** Applying min-max normalization for feature scaling to maintain uniform input ranges.
3. **Categorical Encoding:** One-hot encoding was utilized to transform categorical variables into numerical representations.
4. **Data Balancing:** The issue of class imbalance was tackled through the use of the Synthetic Minority Oversampling Technique (SMOTE) to improve model generalization and lessen bias towards the majority classes.
5. **Train-Test Split:** Train-The dataset was separated into training (70%), validation (15%), and testing (15%) sets.

Adversarial Attack Generation

To assess the susceptibility of detection models, adversarial examples were generated utilizing two well-known methods:

Fast Gradient Sign Method (FGSM)

FGSM generates adversarial examples by introducing noise to the input following the gradient of the loss function concerning the input. It is calculated as:

$$\mathbf{x\_adv = x + \varepsilon * sign\,(\nabla_x J\,(\theta, x, y)}$$

Where:

1. **x_adv** represents the adversarial input.

2. **x** represents the initial input.

3. ε represents the magnitude of the disruption

©2025 The Author(s). Published by Infogain Publication, This work is licensed under a Creative Commons Attribution 4.0 License. http://creativecommons.org/licenses/by/4.0/

4. J (θ, x, y) denotes the loss function characterized by the parameters θ.

5. $\nabla_x J$ denotes the gradient of the loss with respect to x

6. FGSM is quick and efficient, which makes it ideal for replicating fundamental adversarial scenarios**.**

Projected Gradient Descent (PGD)

PGD is a more robust, iterative method that uses FGSM in incremental steps and maps the altered input back into the ε-ball surrounding the original input. It is broadly regarded as a reliable assessment for adversarial strength.

Adversarial examples were produced for both datasets, forming situations where harmful traffic is minimally modified yet maintains its damaging traits, mimicking actual evasion attacks.

Architecture of the Model

The primary deep learning model employed for assessment is a combined adversarial-aware structure integrating CNN and autoencoders.

Convolutional Layers: Retrieve spatial characteristics and local patterns from network traffic streams.

Autoencoder Block: Created for reconstructing inputs to eliminate adversarial noise, enhancing the model's resilience to minor disturbances.

Dense Layers: Conduct the final classification into attack or benign categories utilizing softmax activation.

We trained a simple CNN model with no protections simultaneously with the proposed hybrid model to gauge how potent it was. This provided a baseline for how performance should decrease when things fail.

Mechanisms for Defending Against Adversaries

Multiple defense strategies were incorporated into the model pipeline:

Adversarial Training

We used both clean and hostile samples to teach the models again. This strategy helps the model find patterns that are trying to trick it and move decision boundaries so that changes are less likely to effect it.

Input Validation

As a precursor to transmitting information to the classifier, a denoising autoencoder receives a compressed clean input. This assists to eliminate small issues that get in the way and recover critical data features.

Feature Squeezing

The accuracy of feature values was decreased to diminish the impact of gradient-based attacks. Feature squeezing serves as a preliminary filter, diminishing input dimensionality and the area of adversarial surfaces.

Defensive Dropout

Dropout layers were implemented to improve model generalization and lessen overfitting. This also introduces stochastic behavior, complicating attackers' ability to anticipate model output based on static gradients.

Training and Parameter Tuning

Every model was trained following this configuration:

1. **Optimizer:** Adam
2. **Objective Function:** Categorical Cross-Entropy
3. **Size of Batch**: 64
4. **Epochs:** 50 with early termination
5. **Learning Rate:** 0.001 with reduction

The models were developed using TensorFlow and trained on a GPU-accelerated system to enhance convergence speed and facilitate experimentation. Hyperparameters were adjusted through grid search and cross-validation on the validation set to enhance performance in both adversarial and clean situations.

Assessment Criteria

To evaluate both standard and adversarial performance, the subsequent metrics were utilized:

1. **Accuracy:** Proportion of correctly classified samples.
2. **Recall:** Proportion of true positives among actual positives.
3. **F1-Score:** Harmonic mean of precision and recall.
4. **False Positive Rate (FPR):** Proportion of harmless inputs incorrectly identified as threats.
5. **Adversarial Accuracy:** Model effectiveness on adversarial examples.
6. **Inference Duration:** Average time taken to process a sample, to assess real-time applicability.

©2025 The Author(s). Published by Infogain Publication, This work is licensed under a Creative Commons Attribution 4.0 License. http://creativecommons.org/licenses/by/4.0/

Environment for Real-Time Simulation

To replicate deployment in an active network, a streaming setup was established where traffic samples were consistently provided to the model. Inference latency was evaluated for each sample to confirm that the defense mechanisms did not significantly prolong detection. A maximum allowable latency limit of 5 milliseconds per sample was applied to assess real-time viability.

IV. RESULTS AND DISCUSSION

This segment outlines the empirical results from our experiments focused on assessing and enhancing adversarial resilience in AI-driven cybersecurity systems. The findings are analyzed regarding conventional detection performance, the influence of adversarial assaults, and the efficacy of different defense measures within real-time limitations. All experiments were carried out utilizing the CIC-IDS2017 and UNSW-NB15 datasets, as outlined in the preceding section.

Baseline Achievement on Untainted Data

The baseline performance of the initial models—standard CNN (baseline) and hybrid CNN-Autoencoder—was determined by training and testing them on clean, unaltered data. The CNN model achieved an accuracy of 97.1% on CIC-IDS2017 and 95.3% on UNSW-NB15, with both precision and recall exceeding 94%. The hybrid model marginally exceeded the baseline, reaching accuracies of 98.2% and 96.7% on the corresponding datasets. The F1-scores were 0.975 and 0.961, demonstrating a remarkable equilibrium between precision and recall. These findings validate that deep learning models are very efficient at detecting recognized and varied cyber threats in clean environments.

Performance in the Face of Adversarial Assaults

To assess adversarial weakness, both models were subjected to inputs altered through FGSM and PGD attacks. The outcomes were striking. With FGSM $\varepsilon = 0.02$, accuracy of the baseline CNN on CIC-IDS2017 dropped from 97.1% to 76.4%, and for PGD (with 10 iterations), the drop was even larger, to 69.3%. Similar trends were observed on UNSW-NB15 with the drop from 95.3% to 73.5% (FGSM) and 66.1% (PGD). This drop shows how susceptible the models are to minor variations in inputs and that it is imperative to include adversarial defense in actual IDS.

The hybrid model showed enhanced robustness but nevertheless had a marked decline in performance Accuracy fell to 84.7% for CIC-IDS2017 and 81.3% for UNSW-NB15 due to FGSM attacks. For PGD, accuracy fell to 77.2% and 74.9%, respectively. As much as these figures depict increased strength over the baseline, the results show that even advanced models are not provided with sufficient protection from deliberate adversarial attacks without proper defensive measures.

Efficacy of Adversarial Training

Adversarial training greatly enhanced the model's robustness. When re-trained with a combination of clean and adversarial examples (FGSM and PGD), the hybrid model showed significantly reduced performance declines when facing attacks. In CIC-IDS2017, the accuracy with FGSM perturbation increased to 91.4%, while PGD reached 88.9). These findings suggest that adversarial training enhances the model's decision boundaries and boosts generalization to adversarial inputs, though it results in a minor decline in performance on clean data (a decrease of about 1.2%).

Significantly, adversarial training also resulted in a decrease in the false positive rate (FPR). With clean data, the FPR of the adversarially trained model stayed below 2.1%, whereas the non-hardened version had 3.6%. This indicates that the extra resilience gained from adversarial training might also lower overfitting, enhancing the model's overall stability.

Effectiveness of Input Filtering (Autoencoders)

Incorporating a denoising autoencoder into the model pipeline to clean inputs prior to classification improved both accuracy and adversarial robustness. For adversarial inputs created through FGSM, the model with sanitization maintained 89.8% accuracy on CIC-IDS2017 and 86.1% on UNSW-NB15. In opposition to PGD, it preserved 84.2% and 80.3% correspondingly. Although somewhat less effective than adversarial training, this method is model-independent and can be implemented as a preprocessing filter for any IDS.

The integration of autoencoders with adversarial training produced the optimal outcomes, showcasing supplementary advantages. The hybrid model,

©2025 The Author(s). Published by Infogain Publication, This work is licensed under a Creative Commons Attribution 4.0 License. http://creativecommons.org/licenses/by/4.0/

developed through adversarial methods and improved with input cleaning, achieved 93.5% accuracy on FGSM and 90.8% on PGD for CIC-IDS2017.onUNSW-NB15, results were 91.7% (FGSM) and 88.5% (PGD), indicating that integrating defenses yields a synergistic benefit.

Effects of Feature Squeezing

Feature squeezing, being a lightweight preprocessing method, minimally decreased model vulnerability. It enhanced adversarial accuracy by 4–6% in most instances but had a detrimental impact on clean data performance. Specifically, it resulted in a slight decline in the accuracy of clean data (around. 1.8%) because of its lossy characteristics. Although it isn't a complete solution, feature squeezing can be an effective part of a multilayered defense approach, especially in settings where computational limitations restrict the application of more complex models.

Performance of Real-Time Inference

A vital necessity for IDS is operation in real-time. We evaluated inference latency for all models to determine their feasibility for use in active network settings. The typical CNN model exhibited an average inference time of 2.3 milliseconds for each sample. The hybrid model (including autoencoder) raised latency to 4.7 milliseconds. Through adversarial training and sanitization, overall latency achieved 5.6 milliseconds—within the permissible limit for real-time detection (≤ 6 ms).

Even though extra defensive layers add computational costs, the compromise was warranted due to greatly enhanced adversarial resilience. Crucially, throughput stayed consistent at over 200 samples per second, confirming the practicality of implementing these models in operational settings where both speed and precision are essential.

Dialogue and Consequences

The results of this research strengthen the argument for going beyond accuracy as the only performance measure for AI-driven cybersecurity systems. The experiments show that models that excel on clean data can be greatly affected by adversarial disturbances. Therefore, adversarial robustness should be a fundamental factor in the design of IDS.

Additionally, a tiered defense strategy—combining adversarial training with input sanitization and a strong architecture—provides greater durability than any individual method by itself. Even though this introduces greater complexity, the advantages in security-sensitive applications significantly surpass the drawbacks. The minimal rise in inference time is controllable in the majority of practical scenarios, particularly considering the avoidance of expensive security violations.

These findings indicate a hopeful future path for AI in cybersecurity: systems that not only identify threats but also proactively adjust to changing attack methods. Incorporating explainability, ongoing learning, and reinforcement strategies can significantly boost trust, adaptability, and operational effectiveness when confronted with more advanced cyber threats.

V. CONCLUSION AND FUTURE WORK

The inclusion of artificial intelligence, especially deep learning, in cybersecurity systems has greatly improved real-time threat detection by automating the examination of intricate, large-scale network data. Nevertheless, this progress brings about new challenges. A primary concern is the susceptibility of AI models to adversarial attacks—particularly, evasion methods where precisely designed input alterations trick the model into incorrectly classifying harmful data as harmless. This research aimed to investigate the extent of this vulnerability and suggest a practical, effective, and robust framework that protects deep learning-based intrusion detection systems (IDS) from these evasion threats while maintaining real-time performance.

Intensive testing with the CIC-IDS2017 and UNSW-NB15 datasets shows that unprotected deep learning models, even those with high accuracy, are very vulnerable to adversarial disturbances. When adversarial training was applied using basic adversarial attacks, detection accuracy decreased by nearly 20–30% in certain instances. This indicates just how simple it is to attack these models.

Any performance decrease is unacceptable in cybersecurity because correct detection is required in both good and bad operating environments.

The study proposed a hybrid adversarial defensive mechanism employing various protective techniques, such as adversarial training, input purification via

©2025 The Author(s). Published by Infogain Publication, This work is licensed under a Creative Commons Attribution 4.0 License. http://creativecommons.org/licenses/by/4.0/

denoising autoencoders, and convolutional filtering to enhance the architecture of the system. When combined, these methods made IDS models much less susceptible to deception by fictitious inputs. The top model, trained to be adversarial and employed a sanitization method, was able to identify over 90% of the attacks, including even the powerful ones such as PGD. The average inference delay of the model was slightly more than 5 milliseconds, which is important as it indicates it can detect intrusions in real time. What this implies is that you can have adversarial robustness without sacrificing speed or efficiency, a consideration for areas where security is crucial but a high amount of data must be processed quickly.

The model became more robust and contained less false positives. This means that defenses against adversaries would make the models more robust and ensure they perform better in new environments. These results imply that AI-based security systems are more secure and reliable when they employ design principles that take into account adversaries. Despite these achievements, the research indicated that there were still plenty of restrictions and opportunities for further development. Adversarial training made the model resilient to known attack techniques such as FGSM and PGD initially, but it was still defensive. It protects against certain types of noise in the training data, but it may not work so well against new or developing attack methods. This limit shows that we need to set up proactive security systems that can handle more forms of attacks than we have seen before. Second, the autoencoder-based cleaning process is effective, although it is slightly longer to execute. We discovered that our test environment had decent latency. Deploying into ultra-low latency environments, though, like in edge devices or mission-critical networks, might require more tuning or relaxed security restrictions.

It was easy and fast to learn the compressed properties, but they didn't have a significant impact and may not be sufficient to defend yourself independently.

Another problem is that attackers and defenders are engaged in a ''arms race'' across the globe.

Defenses improve as do people's attacks. Adversarial methods such as GAN-based perturbations, black-box transfer attacks, and real-world executable tactics tend to predict the emergence of emerging ideas. Therefore, generating AI models with the ability to self-modify and adapt as well as continuously discover and react to new attack behaviors remains a high-priority goal for the future. Future research should center on the incorporation of explainable AI (XAI) methods into adversarially robust models. For operational assurance and post-attack analysis, particularly during an attack, it is crucial to understand the rationale behind a model's classification of an input as malicious or benign. In key situations where accuracy is equally as important as responsibility and openness, models that are easy to understand will be superior. Online education and continuous learning are other good ways to make models more flexible. Detection systems can change to keep up with new threats without needing to be retrained or putting security at risk. This lets models keep learning about new traffic patterns and ways that attackers can attack.

The study also facilitates the utilization of different modalities by opposing defenses. This study focused on network traffic data; however, the amalgamation of other data sources, including endpoint logs, system telemetry, and behavioral analytics, could provide a more thorough and resilient framework for threat identification. A model that integrates and cross-verifies information from many domains is inherently more resilient to deception, as it expands the adversarial assault surface required for successful evasion.

In conclusion, this research substantially contributes to the evolving field of adversarial cybersecurity by demonstrating that deep learning models, notwithstanding their vulnerabilities, can be enhanced by a judicious combination of protective techniques. The suggested adversarially robust framework offers theoretical understanding and practical approaches for improving the security of AI-driven intrusion detection systems. As cyber threats grow in complexity and nuance, creating resilient, adaptive, and explainable AI systems will be crucial for protecting digital infrastructure in real-time. The path to developing these systems is intricate and continuous, yet this effort signifies a crucial advancement toward that objective.

©2025 The Author(s). Published by Infogain Publication, This work is licensed under a Creative Commons Attribution 4.0 License. http://creativecommons.org/licenses/by/4.0/

REFERENCES

[1] Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). *Explaining and harnessing adversarial examples*. arXiv preprint arXiv:1412.6572.

[2] Grosse, K., Papernot, N., Manoharan, P., Backes, M., & McDaniel, P. (2017). *Adversarial examples for malware detection*. In European Symposium on Research in Computer Security (pp. 62–79). Springer.

[3] Huang, S., Wang, C., & Lin, J. (2019). *Adversarial attacks on deep-learning based network intrusion detection systems*. IEEE Access, 7, 84862–84871.

[4] Carlini, N., & Wagner, D. (2017). *Towards evaluating the robustness of neural networks*. In 2017 IEEE Symposium on Security and Privacy (SP), pp. 39–57.

[5] Xu, W., Evans, D., & Qi, Y. (2017). *Feature squeezing: Detecting adversarial examples in deep neural networks*. arXiv preprint arXiv:1704.01155.

[6] Canadian Institute for Cybersecurity. (2017). *CIC-IDS2017 Dataset*. Retrieved from https://www.unb.ca/cic/datasets/ids-2017.html

[7] Moustafa, N., & 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) (pp. 1–6). IEEE.

[8] Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). *Towards deep learning models resistant to adversarial attacks*. arXiv preprint arXiv:1706.06083.

[9] Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z. B., & Swami, A. (2016). *The limitations of deep learning in adversarial settings*. In 2016 IEEE European Symposium on Security and Privacy (EuroS&P), pp. 372–387.

[10] Zhang, C., Wang, X., & Zhang, Z. (2020). *Adversarial training for robust intrusion detection in industrial control systems*. IEEE Access, 8, 108383–108394.

[11] Doshi, R., Apthorpe, N., & Feamster, N. (2018). *Machine learning DDoS detection for consumer internet of things devices*. In 2018 IEEE Security and Privacy Workshops (SPW), pp. 29–35.

[12] Raff, E., Zak, R., Cox, R., Sylvester, J., McLean, M., & Nicholas, C. (2018). *An investigation of adversarial examples in malware detection*. In Proceedings of the 18th IEEE International Conference on Machine Learning and Applications (ICMLA)", pp. 279–284.

[13] Vitorino, A. S., Silva, L. C., & Ferreira, A. (2022). Adaptive perturbation patterns: Realistic adversarial learning for robust intrusion detection. arXiv preprint arXiv:2203.04234. https://arxiv.org/abs/2203.04234

[14] Roshan, N., Zafar, K., & Haque, R. (2023). A novel deep learning based model to defend network intrusion detection systems against adversarial attacks. arXiv preprint arXiv:2308.00077. https://arxiv.org/abs/2308.00077

©2025 The Author(s). Published by Infogain Publication, This work is licensed under a Creative Commons Attribution 4.0 License. http://creativecommons.org/licenses/by/4.0/