Chapter 9
Adversarial Attacks on Deep Learning–Based Intrusion Detection Systems on Challenges and Countermeasures

A. Jeyaram
https://orcid.org/0009-0005-2415-4930
Bharath Institute of Higher Education and Research, India

A. Muthukumaravel
Bharath Institute of Higher Education and Research, India

ABSTRACT

DL-based IDS are crucial for detecting and reducing security risks in network settings. Nevertheless, these systems are susceptible to adversarial assaults that exploit model flaws. This research presents a defensive architecture that combines dynamic retraining, ensemble learning, real-time monitoring, and model interpretability to improve the resilience of IDS. Over subsequent years, the performance assessment reveals significant improvements in accuracy (from 0.85 to 0.94), precision (from 0.78 to 0.91), recall (from 0.89 to 0.95), and F1 score (from 0.83 to 0.93). Significantly, there was a reduction in false positive rates from 0.12 to 0.06 and a drop in false negative rates from 0.11 to 0.05. Analysis of feature significance serves to identify crucial aspects that influence the predictions made by a model, hence improving its interpretability. The suggested structure facilitates the ability of IDS to adjust and react to evolving threats efficiently.

DOI: 10.4018/979-8-3373-4672-4.ch009

INTRODUCTION

The growing dependence on IDS based on deep learning has greatly improved the security stance of diverse networks and systems. These systems use AI to identify and effectively address various security risks, including but not limited to malware, phishing attacks, network intrusions, and data breaches. Nevertheless, despite their efficacy, IDS based on DL are not impervious to adversary manipulation (Khushnaseeb and Zafar, 2023). Adversaries may exploit the vulnerabilities present in these systems by creating malicious inputs designed to avoid detection, compromising the effectiveness of their security measures. The integrity and dependability of contemporary cybersecurity infrastructures are significantly challenged by adversarial assaults targeting DL-based IDS. These assaults use various methods, including evasion attacks and poisoning attacks, to mislead IDS into incorrectly categorizing or disregarding hostile activity (Syed Agha Hassnainmohsan et al., 2023). Evasion attacks include altering input data to elude detection while poisoning attacks specifically target the training process by introducing malevolent samples into the training dataset (Boye, 2024). The potential ramifications of effective adversarial assaults on IDS may be significant, resulting in unnoticed security breaches, data exfiltration, and system penetration (Dhinakaran et al., 2023).

DL-based IDS are susceptible to adversarial assaults due to their dependence on intricate neural network structures. However, these designs can capture complex patterns and correlations in data that are inherently vulnerable to manipulation by adversaries. Adversaries leverage the non-linear characteristics of DL models to manipulate input data in an undetectable manner by human observers, but may result in substantial alterations to model predictions (Afnanalotaibi, 2023). The advent of adversarial machine learning as an area of academic inquiry has brought attention to the susceptibilities of DL models and the possible hazards linked to their use in applications that need high levels of security (Priscila et al., 2023). Researchers have established the viability of adversarial attacks on diverse deep learning-based systems, such as image classifiers, natural language processing models, and IDS (Sayedelahl, 2024). The occurrence of these assaults has emphasized the need for strong defensive measures to protect against ever-changing cyber threats (Minu et al., 2023). To tackle the difficulties presented by adversarial assaults on DL-based IDS, a comprehensive and multifaceted strategy is necessary (Nageswaraguptha et al., 2022).

First and foremost, it is necessary to get a more profound understanding of the inherent weaknesses in DL models and their consequences for security applications (Devi and Rajasekaran, 2023). By identifying and analyzing the inherent faults in these models, researchers may create designs that are more durable and less vulnerable to hostile manipulation (Harshadsathaye et al., 2022). In addition, establishing efficient