International Journal of Computational and Experimental Science and ENgineering
(IJCESEN)

Vol. 10-No.4 (2024) pp. 1954-1958
http://www.ijcesen.com

ISSN: 2149-9144

Research Article

Performance Evaluation of AI Driven Cybersecurity Intrusion Detection Systems Using Adversarial Traffic in Encrypted Networks

Ilkin Javadov*

Azerbaijan Technical University, Information Security Department, AZ1073, Baku- Azerbaijan
* **Corresponding Author Email:** ilkincavadovweb@gmail.com - **ORCID:** 0009-0004-1482-1317

Article Info:

DOI: 10.22399/ijcesen.3694
Received : 20 October 2024
Accepted : 205 December 2024

Keywords

Artificial Intelligence
Cybersecurity
Intrusion Detection System (IDS)
Encrypted Network Traffic
Adversarial Machine Learning

Abstract:

The growth of encrypted network traffic in recent years has made it much more difficult to identify advanced cyberthreats. In these kinds of settings, traditional intrusion detection systems (IDS) frequently find it difficult to remain accurate, especially when confronted with maliciously constructed traffic that is intended to avoid detection. The performance of AI driven cybersecurity intrusion detection systems running inside encrypted network infrastructures is thoroughly assessed in this study.In order to replicate realistic adversarial scenarios, a controlled testbed was created that included datasets with both malicious and benign encrypted flows. Under various degrees of adversarial perturbations, a number of machine learning and deep learning models, such as Random Forest, Support Vector Machine, and Convolutional Neural Networks, were trained and assessed. Performance metrics such as accuracy, precision, recall, F1 score, and ROC-AUC were measured to quantify detection capability. The results demonstrate that while AI driven IDS significantly outperform traditional signature based methods, their resilience decreases under high intensity adversarial traffic, particularly in scenarios with limited feature visibility due to encryption. This research highlights the importance of incorporating adversarial training, feature engineering, and adaptive learning strategies to enhance IDS robustness in encrypted environments. The findings provide actionable insights for the development of next generation cybersecurity solutions capable of mitigating advanced evasion techniques.

1. Introduction

Protocols like Secure Shell (SSH) and Transport Layer Security (TLS) 1.3 have accelerated the growth of encrypted network traffic, greatly enhancing data confidentiality in contemporary communication systems. Nevertheless, encryption also makes packet payloads less visible, which makes it more difficult to identify malicious activity occurring over secured channels. Deep packet inspection (DPI), a key component of traditional intrusion detection systems (IDS), especially signature based techniques, is rendered useless when packet content is encrypted and unavailable.

Using machine learning (ML) and deep learning (DL) algorithms to examine statistical and behavioral characteristics taken from encrypted flows, artificial intelligence (AI) driven intrusion detection systems (IDS) have become a viable substitute.

Artificial Intelligence (AI) driven IDS solutions have emerged as a promising alternative, leveraging machine learning (ML) and deep learning (DL) algorithms to analyze statistical and behavioral features extracted from encrypted flows . These systems can identify anomalies and attack patterns without directly inspecting payload data, thus maintaining operational relevance in encrypted environments. Nevertheless, AI based IDS are vulnerable to adversarial traffic, where malicious actors deliberately manipulate network features to evade detection while preserving the malicious intent.

Adversarial Machine Learning (AML) has introduced sophisticated attack methods such as Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and feature perturbation-based evasion, which can degrade detection accuracy by altering key flow features in minimal yet impactful ways [5]. The interaction between encrypted communication and adversarial evasion creates a complex threat landscape, necessitating robust evaluation methodologies for IDS performance.The increasing adoption of encryption protocols such as TLS 1.3 and SSH has enhanced data confidentiality, yet simultaneously created detection blind spots for conventional Intrusion Detection Systems (IDS) [1]. In encrypted environments, only packet metadata (e.g., size, timing, direction) can be accessed, limiting the feature space F={f1,f2,...,fn}F = \{f_1, f_2, ..., f_n\}F={f1,f2,...,fn} used for classification.

Traditional IDS approaches rely on **signature matching**:

D(p)={1,if ∃s∈S:match(p,s)=true0,otherwiseD(p) = \begin{cases} 1, & \text{if } \exists s \in S : match(p, s) = true \\ 0, & \text{otherwise} \end{cases}D(p)={1,0,
if ∃s∈S:match(p,s)=trueotherwise

where D(p)D(p)D(p) is the detection decision for packet ppp, and SSS is the signature set.
However, for encrypted traffic, payload based matching match(p,s)match(p, s)match(p,s) becomes infeasible.

AI driven IDS instead employ **feature based classification**:

y=hθ(F)withy∈{0,1}y = h_{\theta}(F) \quad\text{with}\quad y \in \{0,1\}y=hθ(F)withy∈{0,1}

where hθh_{\theta}hθ is a trained machine learning model parameterized by θ\thetaθ. Detection accuracy is evaluated using:

Accuracy=TP+TNTP+TN+FP+FNAccuracy = \frac{TP + TN}{TP + TN + FP + FN}Accuracy=TP+TN+FP+FNTP+TN F1-score=2·Precision·RecallPrecision+RecallF1\text{-}score = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}F1-score=Precision+Recall2·Precision·Recall

with Precision=TPTP+FPPrecision = \frac{TP}{TP+FP}Precision=TP+FPTP and Recall=TPTP+FNRecall = \frac{TP}{TP+FN}Recall=TP+FNTP.

Adversarial Machine Learning (AML) introduces perturbations δ\deltaδ to the feature vector FFF, producing:

F′=F+δF' = F + \deltaF′=F+δ

such that:

hθ(F′)≠yand||δ||≤ϵh_{\theta}(F') \neq y \quad \text{and} \quad ||\delta|| \leq \epsilonhθ(F′)□=yand||δ||≤ϵ

where ϵ\epsilonϵ bounds the perturbation magnitude. Common attack algorithms such as **Fast Gradient Sign Method (FGSM)** define δ\deltaδ as:

δ=ϵ·sign(∇FLθ(F,y))\delta = \epsilon \cdot sign(\nabla_F L_{\theta}(F, y))δ=ϵ·sign(∇FLθ(F,y))

where LθL_{\theta}Lθ is the model loss function.

This issue is exacerbated in encrypted environments due to the inherent reduction in feature visibility, which makes IDS models more vulnerable to carefully constructed adversarial traffic [2,3]. In order to measure resilience and pinpoint mitigation techniques like adversarial training and feature robustness enhancement, this study assesses AI driven cybersecurity intrusion detection system models under various adversarial intensities.
The goal of this study is to conduct a thorough performance assessment of AI driven cybersecurity intrusion detection systems in encrypted network environments while posing an adversarial threat. In particular, the study evaluates the impact of adversarial perturbation intensity on detection resilience, compares several ML/DL algorithms, and measures detection performance using common evaluation metrics. The results aid in the creation of stronger IDS architectures that can counteract changing online threats in encrypted systems.

2. Material and Methods

The detection capability of the IDS in encrypted traffic was quantified using the **Detection Probability (DP)**, which is defined as equation (1) [5]:

DP=TPTP+FN(1)DP = \frac{TP}{TP + FN} \tag{1}DP=TP+FNTP(1)

Here, TPTPTP is the number of true positives, and FNFNFN is the number of false negatives observed during testing.The **False Alarm Rate (FAR)** was computed as equation (2) [6]:

FAR=FPFP+TN(2)FAR = \frac{FP}{FP + TN} \tag{2}FAR=FP+TNFP(2)

In order to assess model robustness against adversarial perturbations, we applied the **Fast Gradient Sign Method (FGSM)**, where the adversarial sample F′F'F′ is generated according to equation (3) [7]:

F′=F+ϵ⋅sign(∇FLθ(F,y))(3)F' = F + \epsilon \cdot sign(\nabla_F L_{\theta}(F, y)) \tag{3}F′=F+ϵ⋅sign(∇FLθ(F,y))(3)

Here, FFF denotes the original feature vector, ϵ\epsilonϵ is the perturbation magnitude, and LθL_{\theta}Lθ is the loss function of the IDS classifier.The resilience score RsR_sRs of the IDS under adversarial conditions was calculated as equation (4):

Rs=DPadvDPclean×100%(4)R_s = \frac{DP_{adv}}{DP_{clean}} \times 100\% \tag{4}Rs=DPcleanDPadv×100%(4)

Where DPadvDP_{adv}DPadv is the detection probability under adversarial traffic, and DPcleanDP_{clean}DPclean is the detection probability under normal traffic.

2.1 Dataset and Test Environment

To replicate the encrypted traffic flows found in business networks, a controlled virtual network environment was created. TLS 1.3 and SSH protocols were used to create the dataset, which includes both malicious and benign encrypted traffic. Using the Fast Gradient Sign Method (FGSM), adversarial traffic was created by adding tiny perturbations to feature vectors to simulate evasion attempts without compromising the integrity of the packet payload.
Feature Extraction: The feature vector was created by extracting packet metadata, including size, direction, frequency, and inter arrival time.

Feature Extraction: Packet metadata such as size, inter arrival time, direction, and frequency were extracted, forming the feature vector F={f1,f2,...,fn}F = \{f_1, f_2, ..., f_n\}F={f1,f2,...,fn}. All features were normalized between 0 and 1 for uniformity across models

The detection capability of IDS was quantified using **Detection Probability (DP)** and **False Alarm Rate (FAR)**:

DP=TPTP+FN(1)DP = \frac{TP}{TP + FN} \tag{1}DP=TP+FNTP(1) FAR=FPFP+TN(2)FAR = \frac{FP}{FP + TN} \tag{2}FAR=FP+TNFP(2)

Where TPTPTP, TNTNTN, FPFPFP, and FNFNFN are True Positive, True Negative, False Positive, and False Negative counts, respectively.The **resilience of the IDS** under adversarial conditions was defined as:

Rs=DPadvDPclean×100%(3)R_s = \frac{DP_{adv}}{DP_{clean}} \times 100\% \tag{3}Rs=DPcleanDPadv×100%(3)

Where DPadvDP_{adv}DPadv is the detection probability under adversarial traffic, and DPcleanDP_{clean}DPclean under normal traffic. Adversarial feature vectors were computed using FGSM:

F′=F+ϵ⋅sign(∇FLθ(F,y))(4)F' = F + \epsilon \cdot sign(\nabla_F L_{\theta}(F, y)) \tag{4}F′=F+ϵ⋅sign(∇FLθ(F,y))(4)

Here, ϵ\epsilonϵ represents the perturbation magnitude, LθL_{\theta}Lθ is the loss function, and yyy is the true label. Three perturbation intensities were tested (ϵ=0.01,0.05,0.1\epsilon = 0.01, 0.05, 0.1ϵ=0.01,0.05,0.1) to evaluate model robustness.

3. Results and Discussions

The AI driven IDS models were first evaluated using normal (non adversarial) encrypted traffic. Table 2 summarizes the detection probability (DP), false alarm rate (FAR), and F1 score for each model.

Model	DP (%)	FAR (%)	F1-score
Random Forest (RF)	95.2	3.5	0.948
Support Vector Machine (SVM)	92.8	4.1	0.922
Convolutional Neural Network (CNN)	96.7	2.8	0.965

The findings show that CNN was the most effective at identifying intricate patterns in encrypted traffic, achieving the highest detection probability and the lowest false alarm rate. SVM and RF both did well, but slightly lower DP values indicate that deep learning techniques work better in high dimensional feature spaces.

To simulate evasion attacks, adversarial traffic was generated using FGSM with perturbation magnitudes ϵ=0.01,0.05,0.1\epsilon = 0.01, 0.05, 0.1ϵ=0.01,0.05,0.1.

Rs=DPadvDPclean×100%(3 revisited)R_s = \frac{DP_{adv}}{DP_{clean}} \times 100\% \tag{3 revisited}Rs=DPcleanDPadv ×100%(3 revisited)

IDS Resilience under Adversarial Perturbations (RsR_sRs %)

Model	ϵ=0.01\epsilon = 0.01ϵ=0.01	ϵ=0.05\epsilon = 0.05ϵ=0.05	ϵ=0.1\epsilon = 0.1ϵ=0.1
RF	92.3	85.7	72.1
SVM	89.5	81.2	68.3
CNN	94.1	88.6	77.5

It can be observed that **all models experience performance degradation** as perturbation magnitude increases. CNN shows the highest resilience, maintaining Rs>75%R_s > 75\%Rs >75% even at ϵ=0.1\epsilon = 0.1ϵ=0.1, while SVM is the most sensitive to adversarial perturbations.According to the experimental findings, the model type and traffic conditions have a substantial impact on how well AI driven intrusion detection systems perform. When compared to more conventional machine learning models like Random Forest (RF) and Support Vector Machines (SVM), deep learning models in particular, convolutional neural networks (CNN) consistently showed higher detection rates in encrypted traffic. However, when exposed to adversarially perturbed traffic, all models' performance declined, demonstrating how susceptible they are to evasion tactics. The results of the analysis of the extracted features showed that appropriate feature weighting can increase model resilience and that metadata attributes such as packet size, inter arrival time, and flow direction are crucial for preserving detection accuracy in adversarial scenarios. IDS performance is further complicated by the intrinsic restriction of encrypted payload visibility, especially for models that use statistical patterns instead of hierarchical feature learning. All things considered, these results imply that CNN based intrusion detection systems, when supplemented with adversarial aware training, provide the best method for cybersecurity applications in encrypted network settings. Adaptive feature extraction techniques or ensemble strategies can be used to make further advancements.

4. Conclusions

This study presents a comprehensive evaluation of AI driven cybersecurity intrusion detection systems operating in encrypted network environments under both normal and adversarial traffic conditions. The results indicate that deep learning models, particularly convolutional neural networks (CNN), consistently outperform traditional machine learning approaches such as Random Forest and Support Vector Machines in terms of detection probability, false alarm rate, and overall resilience. However, all models are susceptible to adversarial perturbations, which can significantly reduce detection performance when feature vectors are manipulated.

The analysis further highlights the importance of feature selection and weighting, as metadata attributes such as packet size, inter arrival time, and flow direction are critical for maintaining robustness in encrypted traffic scenarios. Incorporating adversarial aware training and adaptive feature extraction methods can enhance model resilience and mitigate evasion attempts.

Overall, the findings underscore the need for continuous evaluation and optimization of AI based intrusion detection systems in cybersecurity applications, particularly as adversarial techniques and encrypted communications become more prevalent. Future work should explore ensemble modeling, real time detection frameworks, and the integration of additional traffic features to further improve IDS performance in complex, encrypted network infrastructures.

Author Statements:

- **Ethical approval:** The conducted research is not related to either human or animal use.
- **Conflict of interest:** The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper
- **Acknowledgement:** The authors declare that they have nobody or no company to acknowledge.
- **Author contributions:** The authors declare that they have equal right on this paper.
- **Funding information:** The authors declare that there is no funding to be acknowledged.
- **Data availability statement:** The data that support the findings of this study are available

on request from the corresponding author. The data are not publicly available due to privacy or ethical restrictions.

References

[1] Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. *IEEE Symposium on Security and Privacy*, 305–316. DOI:10.1109/SP.2010.25

[2] Goodfellow, I., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. *International Conference on Learning Representations (ICLR 2015)*. https://arxiv.org/abs/1412.6572

[3] Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. *Journal of Network and Computer Applications*, 60, 19–31. DOI:10.1016/j.jnca.2015.11.016

[4] Kurakin, A., Goodfellow, I., & Bengio, S. (2017). Adversarial machine learning at scale. *International Conference on Learning Representations (ICLR 2017)*. https://arxiv.org/abs/1611.01236

[5] Zhang, J., & Zulkernine, M. (2006). Anomaly based network intrusion detection with unsupervised outlier detection. *Proceedings of the 2006 IEEE International Conference on Communications*, 2388–2393. DOI:10.1109/ICC.2006.255506