Evasion Attack on SVM in IDS Utilizing Machine Learning and Neural Networks

Doaa Samir Elsayed
Dept. of AI
Alexandria University
2103114
cds.doaaSamir2022@alexu.edu.eg

Clara Magdy Ghaly
Dept. of Healthcare
Alexandria University
20221424647
cds.klaramagdy24647@alexu.edu.eg

Suhaila Adel Ali
Dept. of Cybersecurity
Alexandria University
20221376543
cds.SohailaAdel06329@alexu.edu.eg

Basmala Akram Hegazy
Dept. of AI
Alexandria University
2103159
cds.Basmalaakram42693@alexu.edu.eg

Khaled Ahmed Mansour Ahmed
Dept. of AI
Alexandria University
20221320444
cds.KhaledAhmedMansour2022@alexu.edu.eg

Aml Ibrahim Mohamed
Dept. of Cybersecurity
Alexandria University
20221443988
cds.amlebrahim01887@alexu.edu.eg

Menna Allah Ahmed Saad
Dept. of Cybersecurity
Alexandria University
2106212
cds.MenatullahAhmed2022@alexu.edu.eg

***Abstract*—This paper talks about the types of Machine Learning and Neural Network and making the model gets attacked using Evasion Attack. The aim of this is to be able to detect any unusual activity in the network traffic according to the training dataset that we trained the model. We used dataset from kaggle called 'network intrusion'. This paper will take about IDS, the importance of IDS, advantages and Disadvantages. The paper will also talk about the types of machine learning and how it affected the dataset. The paper discuss and compare machine learning model and Neural Network model.**

I. Introduction

Machine Learning (ML) and Neural Networks (NN) have revolutionized the way we approach data analysis and decision-making processes. In this paper, we delve into the diverse landscape of ML and NN techniques, exploring their applications and implications in real-world scenarios. Our primary focus lies in understanding the vulnerabilities of these models, particularly in the context of network security, through the lens of Evasion Attacks.

The objective of this study is to analyze how ML and NN models can be targeted and compromised by malicious entities using Evasion Attacks. Specifically, we aim to develop a system capable of detecting anomalous activities within network traffic, leveraging a trained model based on a dataset sourced from Kaggle, titled 'network intrusion.'

Throughout this paper, we will shed light on Intrusion Detection Systems (IDS) and their pivotal role in safeguarding networks against unauthorized access and malicious activities. We will explore the significance of IDS, elucidating their advantages and disadvantages in contemporary cybersecurity paradigms.

Furthermore, our analysis will extend to the intricacies of different ML techniques and Neural Network (NN) architectures, delving into their impact on the efficacy and robustness of the IDS model. We will conduct a comparative study between traditional ML models and Neural Network (NN) models, evaluating their performance, scalability, and resilience against adversarial attacks.

This paper aims to help connect theories with real-world applications in using machine learning for Intrusion Detection Systems (IDS).

II. Methodology

A. Data Preprocessing and Challenges

This research paper handled Missing values in columns such as srcbytes, dstbytes, count, and srvcount were replaced with the mean value of each respective column to maintain data integrity.

This research paper handled Duplicate rows, if any, were identified and removed from the dataset to ensure the accuracy of the analysis.

It handled Structural Error, Data type conversion was performed to ensure consistency and accuracy in data representation. Columns such as srcbytes, dstbytes, count, and srvcount were converted to integers for better compatibility with the analysis.

It also handled Categorical Features by encoded using LabelEncoder to convert categorical data into numerical format, enabling compatibility with machine learning models.

It handled Outliers, Outliers were identified using Z-Score and removed from the dataset to prevent them from skewing the analysis results.

There was a challenge in the target column, the class 0 was much higher than class 1, therefore the model would be biased and the decision would not be accurate so there is way to handle The imbalance in the target variable (class) was addressed using the RandomOverSampler technique from the imbalanced-learn library, which resampled the dataset to create a more balanced distribution of classes.

And then it shows data normalization as Numerical features were normalized using MinMaxScaler tool to scale the data within a specific range, enhancing the performance of machine learning algorithms.

Then The preprocessed dataset was split into training and testing sets to facilitate model training and evaluation.

B. Machine Learning and Neural Network Models Selection

Model Selection: Various machine learning and neural network models were considered for intrusion detection, including but not limited to Decision Trees, Random Forest, Support Vector Machines (SVM), and Multilayer Perceptron (MLP) neural networks.

It shows training, each selected model was trained using the preprocessed dataset.

And doing early stopping will handle the overfitting in neural network. While training error decreases over time, validation error starts increasing again Thus, train for some time and return parameters with lowest validation error.Most commonly used form of regularization in deep learning. Effective, simple and computationally efficient form of regularization.Training time can be viewed as hyperparameter model selection problem.Efficient as a single training run tests all hyperparameters (unlike weight decay).

Model performance was evaluated using metrics such as accuracy, precision, recall, and F1-score to assess the effectiveness of the intrusion detection models in identifying network intrusions.

III. Machine Learning and Neural Network

A. Intrusion Detection System (IDS)

An Intrusion Detection System (IDS) is like a security guard for computer networks. Its main job is to monitor network traffic and look for any signs of unauthorized or suspicious activities. When it detects something unusual or potentially harmful, it raises an alarm or takes action to prevent further damage.

There are two main types of IDS:

(i) Signature-based IDS: This type compares network traffic patterns against a database of known attack signatures. If it finds a match, it flags it as an intrusion.

(ii) Anomaly-based IDS: This type learns what "normal" behavior looks like on the network and then flags any deviations from this baseline as potential intrusions. It's like noticing something unusual in a crowd and alerting security.

IDS is crucial for maintaining network security and preventing cyberattacks, as it helps to detect and respond to threats in real-time, reducing the risk of data breaches and system compromises.

IDS plays a crucial role in detecting potential threats and attacks early on. This early detection allows organizations to respond quickly and minimize damage.

By continuously monitoring network traffic, IDS helps identify suspicious activities that may go unnoticed by traditional security measures. This proactive approach improves overall security posture.

IDS can detect unauthorized access or malicious activities by insiders within the organization, providing protection against internal threats.

Many industries have regulatory requirements for security monitoring and incident response. Implementing IDS helps organizations meet these compliance standards.

Advantages of IDS:

1. It can detect threats in real-time, allowing for immediate response and mitigation measures.

2. IDS can automatically generate alerts when suspicious activities are detected, enabling security teams to respond promptly.

3. IDS systems often allow for the creation of custom rules and policies tailored to the specific security needs of an organization.

Disadvantages of IDS:

1. IDS may sometimes generate false positives, flagging normal activities as threats, which can lead to unnecessary alerts and manual intervention.

2. IDS systems may not detect all types of attacks, especially sophisticated and zero-day attacks that bypass known signatures or anomalies.

3. Depending on the scale and complexity of the network, IDS systems can be resource-intensive, requiring dedicated hardware and continuous monitoring.

B. Machine Learning Types

- Random Forest Classifier: Random Forest is an ensemble learning method that constructs a multitude of decision trees during training. It builds each tree using a random subset of the training data and a random subset of the features. Random Forest improves upon the single decision tree by reducing overfitting and increasing robustness. It's known for its high accuracy and ability to handle large datasets with high dimensionality.
- Decision Tree Classifier: Decision trees are a simple yet powerful model used for both classification and regression tasks. They work by partitioning the feature space into regions and assigning a label to each region. Decision trees are easy to interpret and understand, making them a popular choice for many applications. However, they can be prone to overfitting, especially when the tree is deep and complex.

- Support Vector Machine (SVM): SVM is a powerful supervised learning algorithm used for classification and regression tasks. It works by finding the hyperplane that best separates the classes in the feature space. SVM aims to maximize the margin between classes, making it robust to outliers. SVM can handle high-dimensional data efficiently and is effective in cases where the number of features is greater than the number of samples.
- Logistic Regression: logistic regression is a linear model used for binary classification tasks. It estimates the probability that a given input belongs to a certain class using the logistic function. Logistic regression is simple, fast, and interpretable. It's often used as a baseline model in classification tasks. However, logistic regression assumes a linear relationship between the features and the log-odds of the outcome, which may not always hold true.

C. Neural Network

Neural networks are a class of models inspired by the structure and function of the human brain. They consist of interconnected layers of neurons (nodes) that process and transform input data. Neural networks can learn complex patterns and relationships in data, making them highly flexible and powerful. Deep neural networks, in particular, with many hidden layers, have shown remarkable success in various domains, including image recognition, natural language processing, and reinforcement learning. However, training neural networks can be computationally intensive and requires a large amount of data. They can also be prone to overfitting, especially with insufficient training data or poor regularization.

D. Uses of Neural Network with IDS and challenges

Neural networks are increasingly used in Intrusion Detection Systems (IDS) due to several advantages they offer:

Complex Pattern Recognition: Neural networks excel at learning complex patterns and relationships within data. In the context of IDS, this capability allows them to detect subtle and sophisticated attack patterns that may not be easily discernible using traditional rule-based or signature-based detection methods.

Adaptability to Dynamic Environments: Neural networks are adaptive and can learn from new data, making them well-suited for dynamic and evolving network environments where attack techniques constantly change.

Feature Extraction: Neural networks can automatically extract relevant features from raw data, reducing the need for manual feature engineering. This is beneficial in IDS where extracting meaningful features from network traffic data can be challenging.

Anomaly Detection: Neural networks can be used for anomaly detection, which involves identifying deviations from normal behavior in network traffic. This is particularly useful for detecting previously unseen attacks or zero-day attacks.

Scalability: With advancements in hardware and software, neural networks can be implemented at scale to handle large volumes of network traffic data, making them scalable for enterprise-level IDS deployments.

False Positive Reduction: Neural networks can help reduce false positives by learning to distinguish between benign network activities and malicious behavior based on learned patterns and behaviors.

Continuous Learning: Neural networks can be designed for continuous learning, allowing them to adapt and improve over time as they encounter new types of attacks and network behaviors.

While neural networks offer these advantages, they also come with challenges such as the need for substantial computational resources, potential black-box nature (difficulty in interpreting internal workings), and the requirement for large and diverse training datasets to ensure effective learning and generalization. Despite these challenges, their ability to detect complex threats and adapt to changing environments makes them valuable components of modern IDS solutions.

E. cyber threats leveraging artificial intelligence (AI) techniques

- AI-Driven Phishing Attacks: AI can be used to enhance phishing attacks by creating highly targeted and convincing messages. Natural language processing (NLP) algorithms can generate phishing emails that mimic the writing style of trusted contacts, making them harder to detect.
- Poisoning Attacks: In poisoning attacks, adversaries inject malicious data into the training set of an AI system. This corrupts the learning process, leading to compromised performance or incorrect outputs during inference.
- Evasion Attacks: Evasion attacks aim to manipulate the input to an AI system in such a way that it fails to recognize malicious content. For example, modifying malware code to evade detection by antivirus software or altering images to bypass image recognition systems.
- Model Inversion Attacks: These attacks attempt to reverse-engineer the internal workings of an AI model by querying it with specific inputs and analyzing its outputs. This can reveal sensitive information or undermine the confidentiality of the model.
- Membership Inference Attacks: In membership inference attacks, adversaries attempt to determine whether a particular data point was used during the training of a machine learning model. This can compromise the privacy of individuals represented in the training data.
- Data Poisoning Attacks: Similar to poisoning attacks, data poisoning involves injecting malicious data into the training set. However, in data poisoning, the goal is not necessarily to compromise the model's performance but to manipulate its behavior in specific instances.
- Model Extraction Attacks: In model extraction attacks, adversaries attempt to extract the parameters or architecture of a target AI model by querying it with crafted inputs. This information can then be used to replicate

or reverse-engineer the model, leading to intellectual property theft or security breaches.

- Gradient Masking: Gradient masking attacks exploit vulnerabilities in the gradient information used during the training of deep learning models. By obscuring or manipulating gradient signals, adversaries can hinder the effectiveness of adversarial defenses.
- Backdoor Attacks: In backdoor attacks, adversaries insert hidden triggers or patterns into the training data or model parameters. These triggers can later be exploited to cause the model to behave maliciously when presented with specific inputs.
- Model Stealing Attacks: Model stealing attacks involve adversaries attempting to replicate or reconstruct a target model by querying it and using the responses to train a surrogate model. This can lead to intellectual property theft or unauthorized model replication.
- Generative Adversarial Networks (GANs) Attacks: Adversaries can use GANs to generate realistic but malicious data that can evade detection by AI-based security systems. GAN-generated content can be used for various malicious purposes, including phishing and malware propagation.

F. Evasion Attack on SVM

In the context of SVMs, evasion attacks refer to techniques employed by adversaries to manipulate input data in such a way that the SVM misclassifies it. The goal of the attacker is to find vulnerabilities in the SVM's decision boundary and exploit them to cause misclassifications.

Adversaries manipulate the features of input data points to push them across the decision boundary of the SVM. By making subtle modifications to the input data while ensuring that these modifications are imperceptible or minimally perceptible to humans, attackers aim to evade detection.

The ultimate aim of evasion attacks is to mislead the SVM classifier into making incorrect predictions. This could involve, for example, causing the SVM to classify a benign input as malicious or vice versa. By crafting adversarial examples, attackers can exploit vulnerabilities in the SVM's decision-making process.

G. Results

In Random Forest Classifier it achieved an accuracy of approximately 0.97. This indicates strong performance in classifying network data as normal or intrusive. Random Forests are known for their robustness and ability to handle large datasets with high dimensionality.

In Decision Tree Classifier it achieved an accuracy of approximately 0.924 While Decision Trees are interpretable and easy to understand, they may struggle with overfitting and may not generalize well to unseen data compared to ensemble methods like Random Forests.

In Support Vector Machine (SVM) it achieved an accuracy of approximately 0.93. SVMs are effective in high-dimensional spaces and are versatile due to the different kernel functions available. However, they may be sensitive to the choice of kernel and parameters. After using Evasion attack in this model, the accuracy decreased and became 0.494.

In Logistic Regression it achieved an accuracy of approximately 0.934. Logistic Regression is a simple yet powerful algorithm for binary classification tasks. It provides probabilities for outcomes and can be interpreted easily.

In Neural Network it achieved an accuracy of approximately 0.97. The neural network model, particularly the convolutional neural network (CNN), demonstrates competitive performance. CNNs are effective in capturing spatial patterns in data, making them suitable for tasks like image and sequence classification.

TABLE I
ACCURAY OF THE MODELS

Classifier	Accuracy
Random Forest	97%
Decision Tree	92.4%
Support Vector Machine (SVM)	93%
Logistic Regression	93.4%
Neural Network	97%

TABLE II
IMPACT OF EVASION ATTACK ON SVM ACCURACY

Before Attack	After Attack
0.93	0.494

H. Future Direction

In the future, Deep Learning (DL) model, particularly Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), can learn complex patterns and features from network traffic data. This leads to more accurate detection of intrusions and anomalies compared to traditional rule-based or signature-based IDS.

CONCLUSION

The paper discussed the importance of Machine Learning and Neural Network, It also referred to some of the challenges that was faced during the process. The paper point out how IDS was used in Machine Learning and Neural Network.

It also compares different Machine Learning types with its accuracy and it shows the effect on the SVM using Evasion Attack. The paper mentions different types of attacks that can be used in AI.

REFERENCES

[1] Support Vector Machine and Random Forest Modeling for Intrusion Detection System (IDS) Md. Al Mehedi Hasan1, Mohammed Nasser2, Biprodip Pal1, Shamim Ahmad3