School of Information Technology and Engineering at the ADA University

School of Engineering and Applied Science at the George Washington University

PROTECTING MACHINE LEARNING MODELS AGAINST ADVERSARIAL ATTACKS

A Thesis
Presented to the Graduate Program of Computer Science and Data Analytics
of the School of Information Technology and Engineering
ADA University

In Partial Fulfillment
of the Requirements for the Degree
Master of Science in Computer Science and Data Analytics
ADA University

By
Lala Shahbandayeva

April, 2023

THESIS ACCEPTANCE

This Thesis by: Lala Shahbandayeva
Entitled: *Protecting Machine Learning Models against Adversarial Attacks*

has been approved as meeting the requirement for the Degree of Master of Science in Computer Science and Data Analytics of the School of Information Technology and Engineering, ADA University.

Approved:

(Adviser)	(Date)
(Program Director)	(Date)
(Dean)	(Date)

ACADEMIC INTEGRITY STATEMENT

“I affirm that this is my own work, I attributed where I used the work of others, I did not facilitate academic dishonesty for myself or others, and I used only authorized resources for my Thesis, per the ADA University Academic Integrity requirements. If I failed to comply with this statement, I understand consequences will follow my actions. Consequences may range from failing the course to expulsion from the program/university and may include a transcript notation.”

Lala Shahbandayeva		April 24, 2023
(Full Name)	(Signature)	(Date: DD.MM.YY)

ABSTRACT

Machine Learning and Deep Learning have been widely used in different domains and showed their performance in many applications such as fraud detection, speech recognition, etc. One of the domains in which Machine Learning and Deep Learning demonstrated their effectiveness is network intrusion detection systems. Considering the success of ML and DL in these domains, they have been actively used - the models trained with new algorithms and datasets, deployed, and actively used in decision-making. Even though they produced high-performance metrics, most recent studies proved that machine learning and deep learning algorithms are not robust and secure against adversarial inputs in the computer vision domain. These findings have introduced a new concern about the application of machine learning and deep learning in security-related domains such as network intrusion detection systems. As a case in point, an adversarial network traffic flow can cause network intrusion detection systems to classify attacks as benign. In this paper, we demonstrate the performance of adversarial attacks against network intrusion detection systems which are built using deep neural networks based on the results of experiments. Based on our findings, the application of the adversarial examples and the robustness of the deep learning-based network intrusion detection systems are discussed. Based on the results, adversarial training increases the model performance, but generates extra complexity due to the re-training process. The usage of less adversarial features or Long Short-Term Memory (LSTM) models can help to increase model robustness without the need for a re-training process.

Keywords: machine learning, deep learning, adversarial attacks, adversarial inputs, network intrusion detection system

TABLE OF CONTENTS

ACKNOWLEDGEMENTS vi
LIST OF FIGURES vii
LIST OF TABLES viii
ABBREVIATIONS ix

1 INTRODUCTION 10
1.1 DEFINITION OF THE PROBLEM 10
1.2 OBJECTIVE OF THE STUDY 10
1.3 SIGNIFICANCE OF PROBLEM 11
1.4 REVIEW OF THE SIGNIFICANT RESEARCH 12
1.4.1 GENERAL UNDERSTANDING OF MACHINE LEARNING 12
1.4.2 GENERAL UNDERSTANDING OF ADVERSARIAL ATTACKS 14
1.5 ASSUMPTIONS AND LIMITATIONS 17
2 LITERATURE REVIEW 18
3 METHODOLOGY 23
3.1 STUDY PROCEDURE 23
3.2 STUDY STRUCTURE AND DATASET 24
3.3 ETHICAL CONSIDERATIONS 25
4 RESEARCH RESULTS AND ANALYSIS OF RESULTS 26
4.1 RESULTS 26
4.2 ANALYSIS OF RESULTS 27
5 SUMMARY AND FUTURE WORK 29
BIBLIOGRAPHY 30

ACKNOWLEDGMENTS

Firstly, I would like to express my sincere gratitude to my supervisor, Dr. Hasanov for guiding me through this thesis. His support and insights helped me to get desired outputs in this research.

Secondly, I would like to thank Dr. Adamov for his comments and suggestions. He provided the necessary guidance and support to conduct the research effectively.

I would also like to thank Dr. Kaisler and Dr. Sipple for their feedback during Master thesis I.

I would like to acknowledge the funding support from the "State Program on improving the international competitiveness of the higher education system in the Republic of Azerbaijan for 2019-2023", and by BP and its Co-Venturers which has provided me the opportunity to follow my academic goals.

Lastly, I would like to thank my family for their endless support and encouragement.

LIST OF FIGURES

No	Figure Caption	Page
4.1	Adversarial Features for NSL-KDD dataset	26
4.2	Feature importance for UNSW-NB dataset	26
4.3	Adversarial Features for NSL-KDD dataset	26
4.4	Feature importance for UNSW-NB dataset	26

LIST OF TABLES

No	Figure Caption	Page
4.1	Accuracy of the models before adversarial attack	25
4.2	Accuracy of the models after adversarial attack	25
4.3	Accuracy of the models after adversarial training	25
4.4	Accuracy of the models after dropping more adversarial and less important features	27

LIST OF ABBREVIATIONS

Abbreviation	Explanation
MLaaS	Machine Learning as a Service
LSTM	Long Short-Term Memory
IDS	Intrusion Detection System
NIDS	Network Intrusion Detection System
DNN	Deep Neural Network
NLP	Natural Language Processing
FGSM	Fast Gradient Sign Method
BIM	Basic Iterative Method
PGD	Projected Gradient Descent
GNN	Graph Neural Network

1 INTRODUCTION

1.1 Definition of the Problem

In recent years, Deep Neural Networks have demonstrated great potential and become the vital substance of real-world applications. DNNs have been started to be widely used in many domains including image analysis, cybersecurity, malware detection, intrusion detection, etc. Based on the recent advancements in computational power, researchers successfully applied deep learning for training classifiers in security-related domains such as malware detection [1][2][3][4][5] and intrusion detection [6][7] and obtained promising results and high accuracies.

There are two kinds of intrusion detection systems: the first type of intrusion detection system is signature-based IDS, and the second type of intrusion detection system is anomaly-based IDS. Signature-based intrusion detection systems are based on the pre-built database of known attacks. With the signature-based intrusion detection systems, the system can determine if the incoming network traffic is an attack or not by comparing the traffic with the signature database. Although these types of intrusion detection systems can effectively detect known attacks and provide high detection rates, they cannot detect zero-day attacks. For this reason, researchers started to build anomaly-based intrusion detection systems to detect unknown attacks. Anomaly-based intrusion detection systems can determine if the incoming network traffic is an attack based on the behavior of the traffic instead of comparing the traffic to some database.

To increase the success of anomaly-based intrusion detection systems, researchers started to apply Machine Learning (ML) to the problem domain which includes applying traditionally supervised and unsupervised ML algorithms [8]. However, considering the fact that traditional ML algorithms are dependent on domain knowledge is one of the main causes why researchers started to apply Deep Neural Networks to anomaly-based network intrusion systems. Results have demonstrated that this type of intrusion detection system can provide promising results with high accuracy [9].

1.2 Objective of the Study

With the advancements in technology and wide usage of the internet, the number of network attacks has been increasing in recent years. Although security practitioners are improving their security techniques, attackers are also motivated to improve their attacks and generate more powerful attacks to evade the system. Considering this tendency, maintaining the security

of the systems has become an important issue. Although anomaly-based intrusion detection systems which are built using Machine Learning algorithms provide high accuracy to detect attacks, how robust and secure these intrusion detection systems remain an open question.

While building the Machine Learning based intrusion detection systems, the researchers mainly focused on increasing the accuracies or other metrics of the system in order to improve the performance of the model. However, in recent years, the robustness and security of these models started to become a crucial topic. According to recent studies, these systems are not secure against adversarial examples [3]. This shows that machine learning or deep learning-based systems are insufficient in critical areas such as cybersecurity. There are four types of attacks in adversarial machine learning. These are inference, poisoning, evasion, and extraction. In this study, we research adversarial examples (evasion attacks).

An adversarial attack is an input to the machine learning or deep learning system which is injected by the attacker in order to make the system make wrong decisions. There are two types of adversarial examples: white-box and black-box attacks. In the white-box setting, the attacker might have knowledge of the architecture of the machine learning or deep learning model. In black-box attacks, the attacker might not have knowledge about the architecture of the system that is being attacked.

With the start of applying Deep Neural Networks in anomaly-based intrusion detection systems, the lack of transparency of DNNs has become the major concern of researchers as to whether or not to apply deep neural networks in security-related applications such as network intrusion detection systems. In order to improve the lack of transparency of DNNs, new explanation methods have been started to build to explain the results of the model. The explanation of Deep Neural Networks has been done through forward and backward propagation [10][11][12][13][14][15]. Other techniques are also available to provide an interpretation of results in which the architecture of the model is unknown [16].

This paper examines the security and robustness of intrusion detection systems that are built using Deep Neural Networks, considering the significance of protecting intrusion detection systems against adversarial examples.

1.3 Significance of Problem

The robustness of the machine learning and deep learning models is based on how these systems are secure against threats or adversarial examples. There are several negative outcomes of insecure machine learning and deep learning models. Although the explanation methods help to avoid these outcomes in some domains, it is still difficult to interpret in other domains. As a case in point, with the advancements in interpretation techniques in DNNs, adversarial attacks in image analysis have become explainable. Nonetheless, these methods cannot be used in security-related applications such as malware classification or intrusion detection due to feature dependence.

Despite the successes of machine learning and deep learning systems, it has become obvious that these systems are not secure against adversarial attacks in image classification in which adversarial examples can cause the classifier to make wrong decisions. Considering the importance of security in security-critical domains such as network intrusion detection, applying machine learning and deep learning to network intrusion detection systems can result in crucial security issues. There is quite a number of research have been conducted on adversarial machine learning in image classification; however, the number of researches on this problem in network intrusion detection did not receive enough attention.

Intrusion detection systems detect attacks from the network flow either through signature-based or anomaly-based methods. Machine Learning and Deep Learning methods also propose advantages to detect zero-day attacks with the help of training the model with benign and attack traffic flows. However, since machine learning algorithms are domain-dependent, and DNNs are not explainable in security-related domains, the security of deep learning-based network intrusion detection systems remains an open research problem.

The application of machine learning and deep learning algorithms in network intrusion detection generated a significant security problem called adversarial machine learning. Adversarial examples have the potential to generate inputs that will eventually cause the network intrusion detection system to give wrong outputs which makes intrusion detection systems prone to more insecure attacks and hacks compared to traditional network intrusion detection systems.

In this research study, we examine adversarial machine learning from the network intrusion detection setting to demonstrate how sensitive deep neural networks are against adversarial examples. The outcomes of this study can be used by the cybersecurity teams to analyze how their machine learning-based network intrusion detection systems secure against adversarial attacks.

The significance of this study is to analyze adversarial attacks in the network intrusion detection setting and suggest defense methods. We conduct research on the effect of adversarial examples on machine learning and deep learning-based network intrusion detection systems. Several adversarial attacks have been generated and tested on the Fast Gradient Sign Method [17], Basic Iterative Method [18], etc. to demonstrate that adversarial examples can deceive network intrusion detection systems.

1.4 Review of the Significant Research

1.4.1 General Understanding of Machine Learning

Considering the fact that machine learning is actively used in industry after researchers used these techniques in the research and laboratories for the last decades [43]. Machine learning employs algorithms with the goal to automate the processes by building or developing models using datasets, minimizing the cost function to generalize the model, and deploying the model into production. Developing a machine learning or deep learning model mainly requires two stages called training the model and deploying the model into production [44].

1.4.1.1 Supervised Learning

With supervised learning, during the training phase, the machine learning model is taught with the dataset that consists of features and labels to differentiate the classes for classification (with discrete labels) and calculate the result for regression (with continuous labels) which aims to minimize the loss function. The main goal of supervised learning is to find the relation between features and labels. The later step is to validate/test the model performance before deploying the model into production to make decisions with new, real-world, unlabeled data. Supervised learning models have been applied in many domains including, but not limited to spam filtering, intrusion detection, object detection, etc. [45][46][47].

1.4.1.2 Semi-Supervised Learning

With semi-supervised learning, during the training phase, the machine learning model is taught with both labeled and unlabeled data, so that a small set of the inputs have outputs and for other inputs, the outputs are not available. As a result, the machine learning model is built using the combination of both labeled and unlabeled data, and when the unlabeled data is used appropriately, then the model can show very good performance results [48].

1.4.1.3 Unsupervised Learning

With unsupervised learning, the machine learning model is taught with unlabeled data. Therefore, it differs from the aforementioned machine learning methods based on not having any associated target values and the data pattern is identified by the machine learning model. In this type of learning, the machine learning model tries to rearrange data into classes which can also be used to get initial insights from the data before training the model [49]. Unsupervised machine learning is also actively used in the industry for a variety of tasks [50][51][52].

1.4.1.4 Reinforcement Learning

Reinforcement learning is another type of machine learning technique in which the data is based on the actions, rewards, and observations [53][54]. The reinforcement learning agent interacts with the environment and learns the environment based on the observations and rewards. It also has a real-world application as winning the champion of GO [55].

1.4.1.5 Federated Learning

Another machine learning approach is federated learning in which multiple clients can jointly build the machine learning model without having the data in a single centralized server. Only the aggregation of the model updates is performed by sending back the updates to a central location which helps to increase privacy and scalability [56].

1.4.1.5 Ensemble Learning

Another machine learning technique is called ensemble learning in which a combination of multiple machine learning models finds better performance by combining the performance of different models which also helps to reduce overfitting and improve overall accuracy [57].

1.4.2 General Understanding of Adversarial Attacks

A review of the significant literature on adversarial machine learning demonstrates that adversarial attacks can happen in training phase or deployment phase of building a machine learning model. For the training phase of the machine learning model, an attacker may control the part of the data including outputs as well as model parameters, code, etc. that is used for training to generate poison attacks. Apart from attacks generated during the training phase, there are other adversarial attack types that are generated during the deployment phase in which the machine learning model has already been trained and the attacker can create evasion attacks to cause integrity violations which can make machine learning model to give wrong predictions. Moreover, an attacker can also generate privacy attacks during this phase to obtain information about the machine learning model and the dataset.

One of the main types of the adversarial attacks is called poisoning attacks. This attack type can be generated during the training of the machine learning model. In this attack type, the attacker makes changes to training data with all types of machine learning techniques [58][59][60]. An attacker can also get a chance to control the machine learning model and the parameters of the model with federated learning [61].

Another main type of adversarial attack is a privacy attack. Similar to poisoning attacks, this attack type is also consisting of data and model privacy attacks depending on whether the attacker aims to attack the model or the training data. There are different types of poisoning attacks. As a case in point, data reconstruction attacks [62], membership inference attacks [63], etc. The main difference between poisoning attacks and privacy attacks is that privacy attacks are generated during the deployment phase.

The last main types of adversarial attacks are called evasion attacks. Similar to privacy attacks, evasion attacks are also generated during the deployment phase. An attacker can generate evasion attacks to modify testing data in order to generate adversarial samples similar to the data before the adversarial attack by using distance measures to change the predictions of the model [17][19][63][64]. Evasion attacks are later divided into two categories called white-box evasion attacks and black-box evasion attacks.

The division of adversarial attacks into white-box attacks and black-box attacks is based on what they know about the machine learning model. In white-box attacks, an attacker has the knowledge about the whole machine learning system which includes the training dataset, parameters of the model, structure of the model, algorithm, etc. The goal of analyzing the white-box attacks is to understand how much the machine learning systems are robust against adversarial examples and predict possible solutions. Compared to the white-box adversarial attacks, black-box adversarial attacks have little to no information about the architecture of the machine learning model.

Therefore, an attacker has no knowledge about how the machine learning model has been trained. Taking the real-world scenarios into account, black-box adversarial attacks are more practical and can happen more frequently than white-box adversarial attacks. However, for building secure and robust machine learning models, it is important to consider both white-box and black-box adversarial attacks. Adversarial samples show the transferability behavior where the adversarial samples that are generated through white-box or black-box adversarial attacks can be used to create adversarial inputs for other machine learning models [19][20][65][66].

1.4.2.1 White-box Evasion Attacks

One type of white-box evasion attack is the adversarial attack against the machine learning models which can be distinguished by the human eye. As a case in point, an adversarial image may mislead the face recognition system to evade and unlock the door. As a result, people who do not have permission can enter the building. Sharif et al. [67] proposed this adversarial attack by generating the inputs by adding extra inputs on the image which cause the facial recognition system to let unauthorized people. Apart from facial recognition systems, white-box evasion attacks can also be applied in another image recognition system where the attacker can mislead the machine learning system by adding noise to road signs [68].

1.4.2.2 Black-box Evasion Attacks

Compared to the white-box adversarial attacks, black-box attacks have a more complex nature and show results closer to the real-world scenarios because in the real-world attacker may not have in- formation about the machine learning architecture, model parameters, etc. As a case in point, in the real-world only available data that is presented to the attacker are inputs and outputs. Therefore, for black-box adversarial attacks, extracting the information about the machine learning model becomes the main motivation.

The attacker can use queries to get the output for the given input while interacting with machine learning or deep learning system. Considering the fact that today the machine learning as a service on cloud platforms like AWS, Azure, etc. where users can get the output of the model by querying, it becomes an important issue.

There are two types of black-box evasion attacks. The first type of black-box evasion attack is based on the logits of the machine learning model. During this type of attack, the attacker uses several optimization approaches [69][70][71][72][73] based on the logits.

The second type of black-box evasion attack is based on decision boundaries. Compared to the logits-based black-box evasion attacks, during this type of adversarial attack, adversarial do not get the information about the model scores, instead attackers can only receive information about the final prediction of the machine learning or deep learning system. One of the first examples of decision boundary-based black-box evasion attacks is Boundary Attack and HopSkipJumpAttack [74][75]. Recent significant reviews of the literature show that improved decision boundary-based black-box adversarial attacks have been proposed [76][77][78].

1.4.2.3 Poisoning Attacks

Apart from evasion attacks, attackers can also generate poisoning attacks against the machine learning model in which the main difference between evasion and poisoning attacks is that poisoning attacks happen in the training phase, while evasion attacks happen in the deployment phase of building machine learning or deep learning model. Poisoning attacks have been predominantly used in many domains such as network intrusion detection [79][80][81], image recognition [82][83][84], NLP [85][86][87], and others security domains [88][89][90].

By their nature, poisoning attacks are more powerful than evasion attacks. Apart from dividing poisoning attacks into white-box poisoning attacks [58][80][91] and black-box poisoning attacks [92], this attack type can also be divided into other types called availability, backdoor, and model poisoning attacks. The first type of the poisoning attack aims to reduce the overall performance of the machine learning model and have the same effect on all of the samples equally. On the other hand, the second type of poisoning attack aims to modify the specific subset of the samples, so that the machine learning model only makes incorrect predictions with specific inputs.

Real-world poisoning attacks have been applied in supervised, unsupervised [93][94], federated [95][96][97], etc. machine learning models. One of the examples of availability poisoning attacks in supervised machine learning models is clean-label poisoning attacks. When the attacker can only have training features, this type of attack can show better results when the labeling process is done by another software or system. In order to carry out clean-label poisoning attacks, attackers can add noise or modify the training data to reduce the performance of the machine learning model [98][99].

1.4.2.4 Privacy Attacks

Similar to evasion attacks, privacy attacks are generated by the attackers during the deployment phase. Privacy attacks are divided into several subcategories called membership inference attacks, memorization attacks, reconstruction attacks, etc. Without needing the information about the architecture of the machine learning or deep learning system, training dataset, model parameters, etc. an attacker can perform reengineering to obtain statistical information about the dataset. Since the attacker can reconstruct the model using statistical information, these types of attacks are called reconstruction attacks [100][101].

Another type of privacy attack is called a membership inference attack. If the attacker generates black-box adversarial membership attacks, the attacker can query the deployed machine learning or deep learning model [102][103][104][115][116]. On the other hand, the attacker can also generate white-box adversarial membership attacks [117][118][119], if the machine learning model parameters, structure, etc. are known by the attacker.

The other type of privacy attacks is called property inference attacks and model extraction attacks. With the first type of the attack, the attacker tries to obtain private information about individuals or groups by analyzing the publicly available information about people or groups [105][106][107][108][109][110]. With model extraction attacks, the attacker aims to obtain information about the structure of the machine learning or deep learning model, parameters

of the model, etc. [111][112][113][114].

1.5 Assumptions and Limitations

One of the main assumptions is that adversarial training can protect deep learning-based intrusion detection systems. Another assumption is that removing adversarial features can increase the robustness of the model against adversarial attacks. The last assumption is that with Long Short-Term Memory (LSTM) or complex models or in models with more data points adversarial attacks may be less successful.

The limitations include not generating adversarial adaptive attacks and performing these attacks in real-world Machine Learning as a Service (MLaaS) applications.

2 LITERATURE REVIEW

One of the very first studies that is conducted to discover that machine learning and deep learning models are not secure against adversarial attacks was proposed by Szegedy et al. [19]. In this approach, the adversarial examples have been generated using the box-constrained Limited memory approximation of the Broyden-Fletcher Goldfarb-Shanno (LBFGS) optimization algorithm in which changes made to the hand-written images cause the deep neural network to make wrong decisions. After this study, a series of other studies have been conducted on generating adversarial attacks and providing defenses against these attacks. Since LBFGS is computationally expensive, other techniques have been proposed. One of these techniques is called Fast Gradient Sign Method (FGSM) has been proposed by Goodfellow et al. [20]. It creates adversarial examples based on the gradient of the loss function which with the reference to the given images as input; thus, provides less computational power with the help of the backpropagation. This approach has been extended with the help of optimization by Kurakin et al. [21].

Another approach was proposed by Papernot et al. [22] in which forward propagation helps to generate adversaries. Apart from these adversarial attacks, another attack was proposed by Moosavi-Dezfooli et al. [23] to distinguish attacks based on the distance from the input image to the adversarial image. Other attack algorithms were proposed by Carlini and Wagner [24] which are based on the gradient and are much more powerful compared to the aforementioned algorithms. These three types of attacks are based on the logits and work with different distance measures.

Despite the fact that several adversarial attacks and defense mechanisms have been proposed [25][26][27], these attacks and the defense mechanisms against these attacks are mainly based on image classification, image recognition, and similar tasks using the widely used image datasets like MNIST, CIFAR10, etc. [28].

Although a huge amount of research on adversarial examples in computer vision tasks has been proposed, very little research on the application of adversarial attacks in security-related domains has been proposed. One of these studies was conducted by Grosse et al. [29] which is the one of first applications of adversarial attacks in security settings. The proposed method is based on the DNNs and has shown the approach of the attackers to evade the systems to perform malicious behaviors. In this research study, adversarial attacks have been used to exploit the malware detection system.

With the recent studies, it has become obvious that network intrusion detection systems that are built using deep neural networks provide a significant improvement over the traditional network intrusion detection systems with the help of less need for domain knowledge and labeled dataset of network traffic flow.

As a case in point, Javaid et al. [30], Shone et al. [31], Tang et al. [32], and Yin et al. [33] have proposed the usage of different machine learning or deep learning algorithms in the network intrusion detection setting. Although all of these studies have demonstrated effective results, there was little or no concern against the adversarial machine learning and deep learning and the potential security issues of the network intrusion detection systems that are built using deep neural networks. As a result, the applications of machine learning or deep learning in security domains such as network intrusion detection systems have been limited.

Recent research has demonstrated the significance of deep neural networks in network intrusion detection systems in several types of networks including, but not limited to Long short-term memory [34], DNNs to detect attacks in host and network levels [35][36], etc. However, very little attention has been put on the effect of adversarial examples against these methods.

Rigaki et al. [37] have conducted research based on the significance and potential risks of adversarial machine learning in network intrusion detection. The adversarial attacks (namely evasion attacks) have been generated against the network intrusion detection systems that are built using machine learning and analyzed the security of the machine learning classifiers using different performance scores after being attacked by different adversarial examples. In this study, they created the Fast Gradient Sign Method and Jacobian-based Saliency Map Attack against the intrusion detection systems which are built using machine learning classifiers such as Random Forest, Decision Tree, Support Vector Machine, and Multi-layer Perceptron. The generated attacks were white-box attacks, and thus, have had some knowledge about the architecture of the model. This research has proved that under adversarial attacks the performance of the machine learning-based intrusion detection system decreases rapidly.

Similar research has been conducted by other researchers as well. Wang et al. [38] have demonstrated how the accuracy of the machine learning-based network intrusion detection system built using the NSL-KDD dataset decreases.

Other studies have been done by researchers to demonstrate the difference and similarities between adversarial machine learning and adversarial deep learning. On one hand, Apruzzese et al. [39] and Martins et al. [40] studied the adversarial attacks against machine learning-based network intrusion detection systems. On the other hand, other researchers studied adversarial attacks against deep learning-based network intrusion detection systems.

Yang et al. [41] conducted research to examine the performance of adversarial attacks on a network intrusion detection system that is built using deep neural networks.

Guo et al. [42] proposed a new method to derive explanations for classification results for security applications under adversarial attacks.

Papernot et al. [65] proposed that because of the existing vulnerable inputs in the training dataset, the security of machine learning and deep learning models remains as an open research issue.

Considering the fact that researches on the protection of machine learning and deep learning systems against adversarial samples remain as an open research area [120[121][122], it is crucial to further analyze the performance of machine learning and deep learning models against new attacks.

While the security of machine learning and deep learning models is crucial, it is also important to keep the performance of the machine learning and deep learning models based on their data domains [123][124]. Thus, it is crucial to understand the effects of the adversarial inputs against the machine learning and deep learning model and the proposed defense methods to avoid overhead.

In order to avoid these vulnerabilities, one of the approaches is to use Generative Adversarial Networks (GANs). Generative Adversarial Networks consist of two sub-models. These sub-models are called generators and discriminators. In the generator, new adversarial inputs are generated and sent to the discriminator to distinguish them from real and fake inputs. Until the discriminator classifies the generated input as real input, this process continues. This helps to increase the general performance of the model against attacks [125][126].

Grosse et al. [127] proposed another method to avoid adversarial attacks in which the algorithm generates adversarial samples using data augmentation by inserting non-malicious data into the training dataset. Using this approach, machine learning or deep learning model becomes robust against adversarial examples.

Arpit et al. [128] proposed that due to the high degree of memorization, deep learning models are not secure against adversarial examples. Jo et al. [129] proposed that due to the transferability feature, deep learning models are not secure against the adversarial inputs.

Another approach to protecting machine learning and deep learning models is to decrease the information to the model by not feeding a part of the dataset to the model.

Sharif et al. [130] demonstrated the real-world effects of adversarial inputs in machine learning and deep learning models. Barreno et. al [131] proposed the idea to provide the protection against adversarial attacks for machine learning and deep learning models by using regularization.

Kolcz et. al [132] demonstrated the importance of the security of machine learning and deep learning models and proposed approaches to protect the models. These approaches consider weighting the features and averaging the model in several steps to allow protection against adversarial attacks, so that attackers will not be able to inject adversaries into the model.

Globerson et. al [133] proposed another defense technique to increase the robustness of machine learning and deep learning models against adversarial samples. This approach considers avoiding the extra weighting of the features with models which are prone against the elimination of the features to allow protection against adversarial attacks.

Considering all of the approaches, it is important to remember that without the possibility of the adversarial attacks, the aforementioned solutions can increase the complexity of the model and may decrease the performance of the machine learning or deep learning model [134][135][136].

Using hierarchical algorithms as a defense technique was also proposed to protect machine learning models against adversarial attacks [137]. Biggio et al. [138] proposed a new approach using random subset to protect machine learning models against adversarial attacks.

The aforementioned defense techniques consider protecting traditional machine learning models against adversarial inputs. However, it is important to consider the applications of deep learning in the domains such as intrusion detection systems. One technique is to use threshold values to capture the change in performance metrics due to the adversarial examples [139]. Dalvi et al. [140] proposed another technique is to perform adversarial classification.

Akhtar et al. [141] proposed new defense techniques in order to consider the security of deep learning models and protecting deep learning models against adversarial examples by using adversarial detection. This has been done by distinguishing whether the input is adversarial or non-adversarial, and if it is an adversarial example, by finding the actual output.

Ross et al. [142] proposed a defense technique by using regularization in order to protect the deep learning models against adversarial attacks by calculating the change in inputs and outputs. In this approach, the attacker cannot have the impact to the deep learning model when the change in adversarial input has less difference. While this approach is similar for most of the deep learning algorithms, it may increase the complexity of the model and training process. Moreover, this approach can still be vulnerable against strong adversarial examples.

Ross et al. [142] proposed another defense technique for image data by transforming the features. This transformation process can protect the deep learning model against adversarial examples.

Xie et al. [143] proposed randomization as a defense technique for deep learning models in order to protect them against the adversarial inputs. This technique provides the protection by applying randomization before the training process starts.

Xu et al. [144] proposed a new defense technique to provide the security for deep learning models in image classification tasks by using compression of the features. This technique aims to find the adversarial inputs. By making changes to the pixel color and then training the deep learning model with both of the features. After the training stage, the results of both of the models are compared to check whether a great change happened or not. If there is a huge change, it means that the feature can have high participation in adversarial examples. Although this approach is computationally simple, when a small change between features happens, it cannot detect all of the adversarial features.

Mohassel et al. [145] proposed a defense strategy against adversarial examples in order to protect deep learning models against adversarial attacks. This strategy achieves the robustness of the deep learning model by training the models in parallel.

A similar approach is proposed by Papernot et al. [146] in which protection for machine learning or deep learning model is achieved by providing knowledge sharing between the target models.

A similar approach is proposed by Shokri et al. [147] in which several deep learning models can be run in parallel. This strategy provides the security for the deep learning model with the help of distributed learning approach. While performing this approach no data is shared between the deep learning models. However, parameters of the deep learning models can be shared with each other. This strategy also helps to run the deep learning models in parallel.

Even though research on adversarial machine learning has mostly been related to computer vision tasks, the impacts of adversarial attacks in network security [148] context are equally or more crucial mainly in malware detection [149] and network intrusion [150] in which the application of machine learning and deep learning in these fields become more popular.

Rosenberg et al. [151] reviewed the adversarial attacks on security-related domains such as intrusion detection systems, cyber-physical systems, etc.

Liu et al. [152] analyzed the adversarial attacks and defense techniques of machine learning.

Qui et al. [153] researched on adversarial attacks on cloud security, malware detection, and intrusion detection.

Different from these researches, this paper only focuses on the network security in terms of network intrusion detection and applies several approaches to provide robust deep learning-based intrusion detection models against adversarial attacks.

Biggio et al. [154] demonstrated the research on adversarial machine learning in the applications of computer vision and cybersecurity. However, this paper did not analyze the adversarial machine learning in network security.

Duddu et al. [155] researched different researches on adversarial machine learning against malware classifiers.

Buczak et al. [156] reviewed the complexity and challenges of machine learning and deep learning-based intrusion detection systems. However, they did not research the adversarial attacks in this domain.

Although Zhang et al. [157] reviewed the adversarial attacks as a constraint in machine learning-based mobile and wireless networking, they did not discuss the limitations of machine learning and deep learning-based network security application.

3 METHODOLOGY

3.1 Study Procedure

3.1.1 Adversarial Training

ART (adversarial robustness toolbox) library has been used to generate the adversarial attacks. Based on the number of correctly predicted/number of all predictions, the performance metric was calculated to show how negatively adversarial attacks affect to the Network Intrusion Detection model. Additionally, those adversarial white-box evasion attacks have been used to show how positively adversarial training affects to the robustness of the network intrusion detection model.

Steps:

• A simple network intrusion detection model was built using neural networks. For the dataset, NSL-KDD and UNSW-NB15 datasets have been used.

• Several white-box adversarial evasion attacks have been generated using ART (adversarial-robustness-toolbox)

• The model was re-trained with the perturbed data and the model is attacked again to see if there are any improvements.

3.1.2 Detecting Adversarial Features

Some of the features are more vulnerable against the adversarial attacks than others. Having features which are prone to adversarial examples increases the risk of adversarial attacks to the model. In order to avoid this, adversarial features and feature importance graphs have been plotted. This approach aims to drop adversarial features which are less important for feature selection.

3.1.3 LSTM for Intrusion Detection

Apart from the simple deep learning-based intrusion detection model, a more complex model has been trained using LSTMs. This approach aims to discover whether simple models are more prone to adversarial attacks or not. Moreover, it aims to discover whether LSTM's ability to pass past information to the present makes the model more robust against adversarial examples.

3.2 Study Structure and Dataset

3.2.1 NSL-KDD Dataset

The NSL-KDD dataset) has been used for training the Network Intrusion Detection model. The dataset is publicly available online and free to use for academic purposes.

The dataset contains the records of the network traffic seen by a simple intrusion detection network that the traffic encountered by a real Intrusion Detection System. The dataset contains 43 features and labels per record, with 41 of them are features and the last two are labels (whether it is a normal or attack) and score (the severity of the traffic input itself).

Training data has 25973 samples (67343 normal samples, 58630 attack samples). Testing data has 22544 samples (9711 normal samples, 12833 attack samples).

4 types of network attacks are presented in the dataset:

1. DoS - Denial of service.
2. Probing - Surveillance and other probing attacks.
3. U2R - Unauthorized access to local super user.
4. R2L - Unauthorized access from a remote machine.

Available types of features and labels in the dataset are:

- Categorical (Features: 2, 3, 4, 42).
- 6 Binary (Features: 7, 12, 14, 20, 21, 22).
- 23 Discrete (Features: 8, 9, 15, 23–41, 43).
- 10 Continuous (Features: 1, 5, 6, 10, 11, 13, 16, 17, 18, 19).

3.2.2 UNSW-NB15 Dataset

UNSW-NB15 is a network intrusion dataset. The dataset contains raw network packets.

Training data has 175341 samples (56000 normal samples, 119341 attack samples). Testing data has 82332 samples (37000 normal samples, 45332 attack samples)

Types of network attacks presented in the dataset are:

1. DoS - Denial of Service.

2. Worms - A malware that replicates itself and spreads through a network by taking advantage of vulnerabilities.

3. Backdoors - Hidden entry points for attackers to access to a system or network while evading security measures.

4. Fuzzers - An attack type which produces unexpected inputs to discover vulnerabilities in the system.

5. Exploit - An attack that gets access to a software or hardware system without authorization.

6. Generic - An attack that does not fall into any categories.

7. Shellcode - An attack that injects malicious code to perform malicious activities.

8. Reconnaissance - An attack that aims to learn about the system to find entry points to the system.

9. Analysis - Network traffic that is not classified as an attack but is used for analysis.

Types of features and labels available in the dataset are:

- Categorical: attack cat, state, service, proto.
- Binary: is_sm_ips_ports, is_ftp_login.
- Numerical: all other features.

3.3 Ethical Considerations

While conducting this research no personally identifiable information was collected. For academic research purposes, free use of the UNSW-NB15 dataset is granted by the authors. Use or redistribution of the NSL-KDD dataset is granted by a citation to the NSL-KDD dataset and the paper [158]

4 RESEARCH RESULTS AND ANALYSIS OF RESULTS

4.1 Results

4.1.1 Adversarial Training

NSL-KDD	UNSW-NB
81.7%	88.1%

Table 4.1: Accuracy of the models before the adversarial attack

	NSL-KDD	UNSW-NB
FGSM	18.3%	12.0%
BIM	27.5%	43.3%
PGD	21.1%	11.1%
DeepFool	18.8%	34.5%

Table 4.2: Accuracy of the models after adversarial attacks

	NSL-KDD	UNSW-NB
FGSM	80.6%	88.2%
BIM	79.1%	86.8%
PGD	77.8%	88.7%
DeepFool	78.3%	87.9%

Table 4.3: Accuracy of the models after adversarial training

Adversarial attacks have been generated on both NSL-KDD and UNSW-NB15 models. Based on the number of correctly predicted/number of all predictions, the performance metric has been calculated to show how negatively adversarial attacks affect to the Network Intrusion Detection model. Additionally, those adversarial white-box evasion attacks have been used to show how positively adversarial training affects to the robustness of the network intrusion detection model.

4.1.2 Adversarial Features

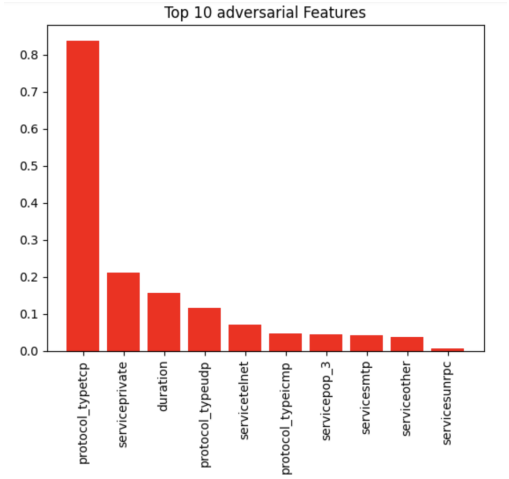

Figure 4.1: Adversarial Features for NSL- KDD dataset

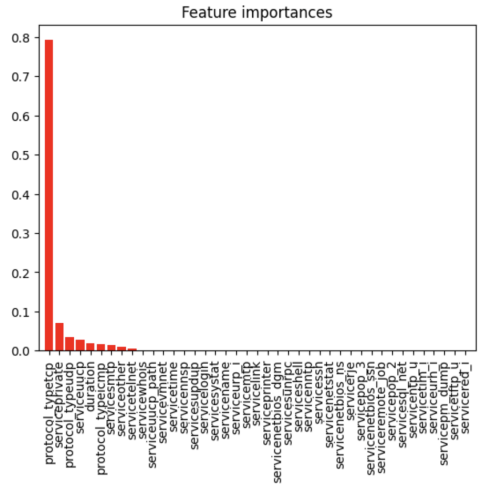

Figure 4.2: Feature importance for NSL- KDD dataset

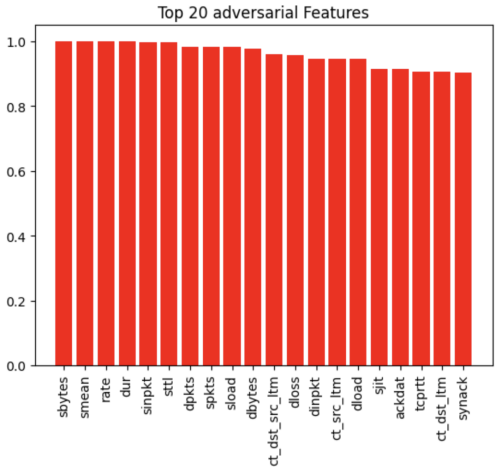

Figure 4.3: Adversarial Features for UNSW-NB15 dataset

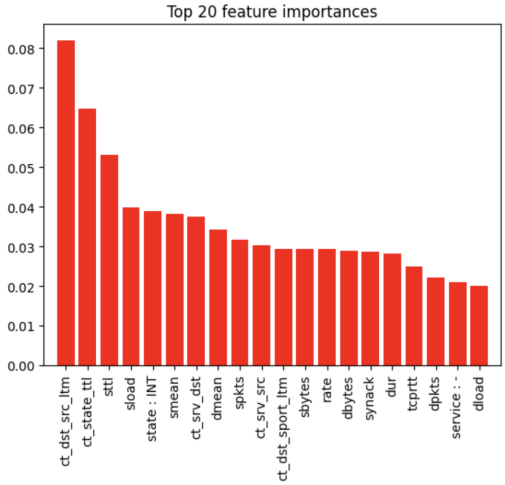

Figure 4.4: Feature importance for UNSW- NB15 dataset

4.2 Analysis of Results

Table 4.1 demonstrates the accuracy of the deep feed-forward models when there is no adversarial attack. The analysis of the data was performed by comparing the accuracies of the models before (Table 4.2) and after (Table 4.3) adversarial training. Moreover, by plotting adversarial features (Figure 4.1 and Figure 4.3) and important features (Figure 4.2 and Figure 4.4), the decision of removing some features has been made.

	NSL-KDD	UNSW-NB
FGSM	20.91%	13.2%
BIM	20.95%	35.6%
PGD	20.99%	35.6%
DeepFool	20.95%	10.1%

Table 4.4: Accuracy of the models after dropping more adversarial and less important features

Based on the results, performance improvements were experienced after re-training the feed-forward deep learning model with the perturbed data. However, this process requires re-training of the model which increases the complexity. Other than that, Table 4.4 demonstrates that after removing less important adversarial features, the performance of the feed-forward deep learning model has been increased for some attack types. This process is less computationally expensive than re-training the model. Lastly, analysis and comparison of the feed-forward deep learning model and LSTM deep learning model demonstrate that LSTMs are more robust against adversarial samples, due to passing past information to the present. Therefore, it is more difficult to generate adversarial attacks and bypass the LSTM-based intrusion detection model because it requires more samples and time to fool. Therefore, for attackers, it becomes computationally expensive to attack LSTMs.

5 Summary and Future Work

Nowadays, applications of machine learning and deep learning have been widely used in many domains including machine learning-based network intrusion detection systems. Related work on the security of deep learning-based intrusion detection systems against adversarial attacks is not deeply researched by researchers compared to the adversarial attacks in image classification tasks. In this paper, we discussed the impacts of adversarial attacks in Deep Learning based Network Intrusion Detection. Additionally, we proposed two other approaches which are less computationally expensive than adversarial training.

This paper leaves a space to conduct future research on this topic, especially on the impacts of adversarial attacks in Large Language Models (LLMs). Moreover, since we performed analysis with feed-forward neural networks and LSTMs, we also aim to continue our experiments with Graph Neural Networks and demonstrate how well GNNs perform against adversarial attacks. We believe that with the enough number of experiments, we can achieve novel security measurements in deep learning which can result in more secure deep learning models against adversaries.

Bibliography

[1] D. Arp, M. Spreitzenbarth, M. Hubner, H. Gascon, K. Rieck, and C. Siemens. "DREBIN: Effective and Explainable Detection of Android Malware in Your Pocket." In Proceedings of the 20th Network and Distributed System Security Symposium, (2014).

[2] G. E. Dahl, J. W. Stokes, L. Deng, and D. Yu. "Large-scale malware classification using random projections and neural networks." In Proceedings of the 38th International Conference on Acoustics, Speech and Signal Processing, (2013).

[3] K. Grosse, N. Papernot, P. Manoharan, M. Backes, and P. McDaniel. "Adversarial perturbations against deep neural networks for malware classification." arXiv preprint arXiv:1606.04435, (2016).

[4] J. Saxe and K. Berlin. "Deep neural network-based malware detection using two-dimensional binary program features." In Proceedings of the 10th International Conference on Malicious and Unwanted Software, (2015).

[5] Q. Wang, W. Guo, K. Zhang, A. G. Ororbia, X. Xing, X. Liu, and C. L. Giles. "Adversary resistant deep neural networks with an application to malware detection." In Proceedings of the 23rd International Conference on Knowledge Discovery and Data Mining, (2017).

[6] A. Javaid, Q. Niyaz, W. Sun, and M. Alam. "A deep learning approach for network intrusion detection system." In Proceedings of the 9th International Conference on Bio-inspired Information and Communications Technologies, (2016).

[7] T. A. Tang, L. Mhamdi, D. McLernon, S. A. Zaidi, and M. Ghogho. "Deep learning approach for network intrusion detection in software-defined networking." In Proceedings of the 12th International Conference on Wireless Networks and Mobile Communications, (2016).

[8] J. Kevric, S. Jukic, and A. Subasi. "An effective combining classifier approach using tree algorithms for network intrusion detection," Neural Computing and Applications, (2017).

[9] Z. Wang. "Deep learning-based intrusion detection with adversaries," IEEE Access, (2018).

[10] R. Fong, A. Vedaldi. "Interpretable Explanations of Black Boxes by Meaningful Perturbation." In Proceedings of the 16th International Conference on Computer Vision, (2017).

[11] J. Li, W. Monroe, and D. Jurafsky. "Understanding Neural Networks through Representation Erasure." arXiv preprint arXiv:1612.08220, (2016).

[12] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling. "Visualizing deep neural network decisions: Prediction difference analysis." In Proceedings of the 5th International Conference

on Learning Representations, (2017).

[13] T. Gehr, M. Mirman, D. D. Cohen, P. Tsankov, S. Chaudhuri, and M. Vechev. "AI2: Safety and Robustness Certification of Neural Networks with Abstract Interpretation." In Proceedings of the 39th IEEE Symposium on Security and Privacy, (2018).

[14] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra. "Grad-cam: Visual explanations from deep networks via gradient-based localization". arxiv:1610.02391 (2016)

[15] A. Shrikumar, P. Greenside, and A. Kundaje. "Learning Important Features Through Propagating Activation Differences". arXiv:1704.02685, (2017).

[16] S. M. Lundberg and S. Lee. "A unified approach to interpreting model predictions". arXiv:1705.07874, (2017)

[17] I. J. Goodfellow, J. Shlens, C. Szegedy. "Explaining and harnessing adversarial examples," arXiv preprint arXiv:1412.6572, (2014).

[18] A. Kurakin, I. Goodfellow, S. Bengio. "Adversarial examples in the physical world," arXiv preprint arXiv:1607.02533, (2016).

[19] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus. "Intriguing properties of neural networks," arXiv preprint arXiv:1312.6199, (2013).

[20] I. J. Goodfellow, J. Shlens, and C. Szegedy. "Explaining and harnessing adversarial examples." [Online]. Available: https://arxiv. org/abs/1412.6572, (2014).

[21] A. Kurakin, I. Goodfellow, and S. Bengio. "Adversarial examples in the physical world." [Online]. Available: https://arxiv. org/abs/1607.02533 (2016).

[22] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami. "The limitations of deep learning in adversarial settings," IEEE Symposium Security Privacy, (2015).

[23] S. M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard. "DeepFool: A simple and accurate method to fool deep neural networks," IEEE Computer Vision Pattern Recognition, (2016).

[24] N. Carlini and D. Wagner. "Towards evaluating the robustness of neural networks." In Proceedings of the 38th IEEE Symposium on Security and Privacy, (2017)

[25] I. J. Goodfellow, J. Shlens, and C. Szegedy. "Explaining and harnessing adversarial examples," arXiv preprint arXiv:1412.6572, (2014).

[26] A. Kurakin, I. Goodfellow, and S. Bengio. "Adversarial examples in the physical world," arXiv preprint arXiv:1607.02533, (2016).

[27] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. "Towards deep learning models resistant to adversarial attacks," arXiv preprint arXiv:1706.06083, (2017).

[28] Y. LeCun. "Backpropagation applied to handwritten zip code recognition," Neural Com-

puting, (1989).

[29] K. Grosse, N. Papernot, P. Manoharan, M. Backes and P. McDaniel. "Adversarial perturbations against deep neural networks for malware classification." arXiv preprint arXiv:1606.04435 (2016).

[30] A. Javaid, Q. Niyaz, W. Sun, and M. Alam. "A deep learning approach for network intrusion detection system," 9th EAI International Conference BioInspired Information Communication Technology, (2015).

[31] N. Shone, T. N. Ngoc, V. D. Phai, and Q. Shi. "A deep learning approach to network intrusion detection," IEEE Emerging Topics Computer Intelligence, (2018).

[32] T. A. Tang, L. Mhamdi, D. McLernon, S. A. R. Zaidi, and M. Ghogho. "Deep learning approach for network intrusion detection in software-defined networking," in Proceedings International Conference Wireless Networking Mobile Communication, (2016).

[33] C. Yin, Y. Zhu, J. Fei, and X. He. "A deep learning approach for intrusion detection using recurrent neural networks," IEEE Access, (2017).

[34] A. Diro and N. Chilamkurti. "Leveraging LSTM networks for attack detection in fog-to-things communications," IEEE Communications Magazine, (2018).

[35] R. Vinayakumar, M. Alazab, K. Soman, P. Poornachandran, S. Venkatraman. "Deep learning approach for intelligent intrusion detection system," IEEE Access (2019).

[36] Y. Zeng, H. Gu, W. Wei, and Y. Guo. "Deep-Full-Range: A deep learning based network encrypted traffic classification and intrusion detection framework," (2019).

[37] R. Maria. "Adversarial deep learning against intrusion detection classifiers." (2017).

[38] Z. Wang. "Deep learning-based intrusion detection with adversaries," IEEE Access (2018).

[39] G. Apruzzese, M. Colajanni, L. Ferretti and M. Marchetti. "Addressing adversarial attacks against security systems based on machine learning." 11th international conference on cyber conflict, (2019).

[40] N. Martins, J. M. Cruz, T. Cruz and P. H. Abreu. "Analyzing the footprint of classifiers in adversarial denial of service contexts." EPIA Conference on Artificial Intelligence, (2019).

[41] K. Yang, J. Liu, C. Zhang, and Y. Fang. "Adversarial examples against the deep learn- ing based network intrusion detection systems," IEEE Military Communications Conference, (2018).

[42] W. Guo, D. Mu, J. Xu, P. Su, G. Wang, X. Xing. "LEMNA: Explaining Deep Learning based Security Applications," Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, (2018).

[43] T. Mitchell, "Machine Learning." McGraw Hill

[44] K. P. Murphy, "Machine Learning: A Probabilistic Perspective." MIT Press, 2012.

[45] I. Sutskever, O. Vinyals, and Q. V. Le, “Sequence to sequence learning with neural networks." in Advances in Neural Information Processing Systems, 2014, pp. 3104–3112.

[46] H. Drucker, D. Wu, and V. N. Vapnik, "Support vector machines for spam categorization," IEEE Transactions on Neural Networks, vol. 10, no. 5, pp. 1048–1054, 1999.

[47] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks." in Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.

[48] K. Shailaja, B. Seetharamulu and M. A. Jabbar, "Machine Learning in Healthcare: A Review." 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2018, pp. 910-914, doi: 10.1109/ICECA.2018.8474918.

[49] D. Erhan, Y. Bengio, A. Courville, P.-A. Manzagol, P. Vincent, and S. Bengio, "Why does unsupervised pre-training help deep learning?" Journal of Machine Learning Research, vol. 11, pp. 625–660, 2010.

[50] A. Krizhevsky and G. Hinton, "Learning multiple layers of features from tiny images." 2009.

[51] A. K. Jain, M. N. Murty, and P. J. Flynn, "Data clustering: A review." ACM Computing Surveys, vol. 31, no. 3, pp. 264–323, 1999.

[52] Sabharwal, A., Selman, B. (2011). Book review. Artificial Intelligence, 175(5-6), 935–937. doi:10.1016/j.artint.2011.01.005

[53] J. Hu and M. P. Wellman, "Nash Q-learning for general-sum stochastic games." Journal of Machine Learning Research, vol. 4, pp. 1039– 1069, 2003.

[54] R. S. Sutton and A. G. Barto, "Reinforcement Learning: An Introduction." MIT Press, 1998.

[55] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, et al., "Mastering the game of Go with deep neural networks and tree search." Nature, vol. 529, no. 7587, pp. 484–489, 2016.

[56] Bharati, S. et al. (2022) "Federated learning: Applications, challenges and Future Directions." International Journal of Hybrid Intelligent Systems, 18(1-2), pp. 19–35. Available at: https://doi.org/10.3233/his-220006.

[57] I. D. Mienye and Y. Sun, "A Survey of Ensemble Learning: Concepts, Algorithms, Applications, and Prospects," in IEEE Access, vol. 10, pp. 99129-99149, 2022, doi: 10.1109/ACCESS.2022.3207287.

[58] B. Biggio, B. Nelson, and P. Laskov. "Poisoning attacks against support vector machines." In Proceedings of the 29th International Conference on International Conference on Machine Learning, ICML, 2012.

[59] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg. "BadNets: Evaluating backdooring attacks on deep neural networks." IEEE Access, 7:47230–47244, 2019.

[60] Y. Liu, S. Ma, Y. Aafer, W. Lee, J. Zhai, W. Wang, and X. Zhang. "Trojaning attack on neural networks." In NDSS. The Internet Society, 2018.

[61] P. Kairouz, H. Brendan McMahan, B. Avent, A. Bellet, M. Bennis, A. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings, R. D'Oliveira, H. Eichner, S. Rouayheb, D. Evans, J. Gardner, Z. Garrett, B. Ghazi, P. B. Gibbons, M. Gruteser, Z. Harchaoui, C. He, L. He, Z. Huo, B. Hutchinson, J. Hsu, M. Jaggi, T. Javidi, G. Joshi, M. Khodak, A. Korolova, F. Koushanfar, S. Koyejo, Y. Liu, P. Mittal, M. Mohri, R. Nock, A. Ozgur, R. Pagh, M. Raykova, H. Qi, D. Ramage, R. Raskar, D. Song, W. Song, S. U. Stich, Z. Sun, A. Suresh, P. Vepakomma, J. Wang, L. Xiong, Z. Xu, Q. Yang, F. Yu, H. Yu, S. Zhao. "Advances and open problems in Federated Learning". Available at: https://arxiv.org/abs/1912.04977v1.

[62] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. "Membership inference attacks against machine learning models." In 2017 IEEE Symposium on Security and Privacy (SP), pages 3–18. IEEE, 2017.

[63] I. Dinur and K. Nissim. "Revealing information while preserving privacy. In Proceedings of the 22nd ACM Symposium on Principles of Database Systems," PODS '03, pages 202–210. ACM, 2003.

[64] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Srndic, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. "Evasion attacks against machine learning at test time. In Joint European conference on machine learning and knowledge discovery in databases." pages 387–402. Springer, 2013.

[65] N. Papernot, P. D. McDaniel, and I. J. Goodfellow, "Transferability in machine learning: from phenomena to black-box attacks using adversarial samples." CoRR, vol. abs/1605.07277, 2016. [Online]. Available: http://arxiv.org/abs/1605.07277

[66] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks." [Online]. Available: http://arxiv.org/abs/1611.02770

[67] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter. "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition." In Proceedings of the 23rd ACM SIGSAC Conference on Computer and Communications Security, October 2016.

[68] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song. "Robust physical-world attacks on deep learning visual classification." In 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, pages 1625–1634. Computer Vision Foundation / IEEE Computer Society, 2018.

[69] A. Ilyas, L. Engstrom, and A. Madry. "Prior convictions: Black-box adversarial attacks with bandits and priors." In International Conference on Learning Representations, 2019.

[70] P. Chen, H. Zhang, Y. Sharma, J. Yi, and C. Hsieh. "Zoo: Zeroth order optimization based

black-box attacks to deep neural networks without training substitute models." In Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, AISec'17, page 15–26, New York, NY, USA, 2017. Association for Computing Machinery.

[71] N. Narodytska and S. Kasiviswanathan. "Simple black-box adversarial attacks on deep neural networks." In 2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pages 1310–1318, 2017.

[72] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin. "Black-box adversarial attacks with limited queries and information." In Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholm, Sweden, July 10-15, 2018, volume 80 of Proceedings of Machine Learning Research, pages 2142–2151. PMLR, 2018.

[73] S. Moon, G. An, and H. Oh Song. "Parsimonious black-box adversarial attacks via efficient combinatorial optimization." In International Conference on Machine Learning (ICML), 2019.

[74] W. Brendel, J. Rauber, and M. Bethge. "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models." In 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net, 2018.

[75] J. Chen, M. I. Jordan, and M. J. Wainwright. "HopSkipJumpAttack: A query-efficient decision-based attack." In 2020 IEEE Symposium on Security and Privacy, SP 2020, San Francisco, CA, USA, May 18-21, 2020, pages 1277–1294. IEEE, 2020.

[76] M. Cheng, S. Singh, P. H. Chen, P. Chen, S. Liu, and C. Hsieh. "Sign-opt: A query-efficient hard-label adversarial attack." In International Conference on Learning Representations, 2020.

[77] S. N. Shukla, A. Sahu, D. Willmott, and Z. Kolter. "Simple and efficient hard label black-box adversarial attacks in low query budget regimes." In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery Data Mining, KDD'21, page 1461–1469, New York, NY, USA, 2021. Association for Computing Machinery.

[78] M. Cheng, T. Le, P. Chen, H. Zhang, J. Yi, and C. Hsieh. "Query efficient hard-label black-box attack: An optimization-based approach." In 7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019. OpenReview.net, 2019.

[79] S. Venkatesan, H. Sikka, R. Izmailov, R. Chadha, A. Oprea, and M. J. De Lucia. "Poisoning attacks and data sanitization mitigations for machine learning models in network intrusion detection systems." In MILCOM, pages 874–879. IEEE, 2021.

[80] H. Xiao, B. Biggio, G. Brown, G. Fumera, C. Eckert, and F. Roli. "Is feature selection secure against training data poisoning?" In International Conference on Machine Learning, pages 1689–1698, 2015.

[81] G. Severi, J. Meyer, S. Coull, and A. Oprea. "Explanation guided back-door poisoning attacks against malware classifiers." In the 30th USENIX Security Symposium (USENIX Security 2021), 2021.

[82] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg. "BadNets: Evaluating backdooring attacks on deep neural networks." IEEE Access, 7:47230–47244, 2019.

[83] J. Geiping, L. H. Fowl, W. R. Huang, W. Czaja, G. Taylor, M. Moeller, and T. Goldstein. "Witches' brew: Industrial scale data poisoning via gradient matching." In International Conference on Learning Representations, 2021.

[84] A. Shafahi, W. R. Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitras, and T. Goldstein. "Poison frogs! Targeted clean-label poisoning attacks on neural networks." In Advances in Neural Information Processing Systems, pages 6103–6113, 2018.

[85] E. Wallace, T. Z. Zhao, S. Feng, and S. Singh. "Concealed data poisoning attacks on NLP models." In NAACL, 2021.

[86] S. Li, H. Liu, T. Dong, B. Zhao, M. Xue, H. Zhu, and J. Lu. "Hidden backdoors in human-centric language models." In CCS'21: 2021 ACM SIGSAC Conference on Computer and Communications Security, Virtual Event, Republic of Korea, November 15 - 19, 2021, pages 3123–3140. ACM, 2021.

[87] X. Chen, A. Salem, D. Chen, M. Backes, S. Ma, Q. Shen, Z. Wu, and Y. Zhang. "Badnl: Backdoor attacks against nlp models with semantic-preserving improvements." In Annual Computer Security Applications Conference, ACSAC '21, page 554–569, New York, NY, USA, 2021. Association for Computing Machinery.

[88] C. Sabottke, O. Suciu, and T. Dumitras. "Vulnerability disclosure in the age of social media: Exploiting Twitter for predicting real-world exploits." In 24th USENIX Security Symposium (USENIX Security 15), pages 1041–1056, Washington, D.C., August 2015. USENIX Association.

[89] R. Perdisci, D. Dagon, W. Lee, P. Fogla, and M. Sharif. "Misleading worm signature generators using deliberate noise injection." In 2006 IEEE Symposium on Security and Privacy (SP'06), Berkeley/Oakland, CA, 2006. IEEE.

[90] B. Nelson, M. Barreno, F. J. Chi, A. D. Joseph, B. I. P. Rubinstein, U. Saini, C. Sutton, and K. Xia. "Exploiting machine learning to subvert your spam filter." In First USENIX Workshop on Large-Scale Exploits and Emergent Threats (LEET 08), San Francisco, CA, 2008. USENIX Association.

[91] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, and B. Li. "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning." In 2018 IEEE Symposium on Security and Privacy (SP), pages 19–35, 2018.

[92] B. Biggio, B. Nelson, and P. Laskov. "Support vector machines under adversarial label noise." In Proceedings of the Asian Conference on Machine Learning, volume 20 of Proceedings of Machine Learning Research, pages 97-112, South Garden Hotels and Resorts, Taoyuan, Taiwan, 14–15 Nov 2011. PMLR.

[93] M. Kloft and Pavel Laskov. "Security analysis of online centroid anomaly detection." Journal of Machine Learning Research, 13(118):3681–3724, 2012.

[94] B. Biggio, K. Rieck, D. Ariu, C. Wressnegger, I. Corona, G. Giacinto, and F. Roli. "Poisoning behavioral malware clustering." In Proceedings of the 2014 Workshop on Artificial Intelligent and Security Workshop, AISec'14, page 27–36, New York, NY, USA, 2014. Association for Computing

[95] V. Shejwalkar, A. Houmansadr, P. Kairouz, and D. Ramage. "Back to the drawing board: A critical evaluation of poisoning attacks on production federated learning." In 43rd IEEE Symposium on Security and Privacy, SP 2022, San Francisco, CA, USA, May 22-26, 2022, pages 1354–1371. IEEE, 2022.

[96] M. Fang, X. Cao, J. Jia, and N. Zhenqiang Gong. "Local Model Poisoning Attacks to Byzantine-Robust Federated Learning." In USENIX Security, 2020.

[97] V. Shejwalkar and A. Houmansadr. "Manipulating the byzantine: Optimizing model poisoning attacks and defenses for federated learning." In NDSS, 2021.

[98] J. Feng, Q. Cai, and Z. Zhou. "Learning to confuse: Generating training time adversarial data with auto-encoder." Advances in Neural Information Processing Systems, volume 32. Curran Associates, Inc., 2019.

[99] L. Fowl, P. Chiang, M. Goldblum, J. Geiping, A. Bansal, W. Czaja, and T. Goldstein. "Preventing unauthorized use of proprietary data: Poisoning for secure dataset release," 2021.

[100] Y. Vorobeychik and B. Li, "Optimal randomized classification in adversarial settings." in 13th International Conference on Autonomous Agents and Multi-Agent Systems, 2014, pp. 485–492.

[101] N. Carlini, F. Tramer, E. Wallace, M. Jagielski, A. HerbertVoss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, A. Oprea, and C. Raffel. "Extracting training data from large language models." In 30th USENIX Security Symposium (USENIX Security 21), pages 2633–2650. USENIX Association, August 2021.

[102] S. Yeom, I. Giacomelli, M. Fredrikson, and S. Jha. "Privacy risk in machine learning: Analyzing the connection to overfitting." In IEEE Computer Security Foundations Symposium, CSF '18, pages 268–282, 2018. https://arxiv.org/ abs/1709.01604.

[103] N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramer. "Membership inference attacks from first principles." In 2022 IEEE Symposium on Security and Privacy (SP), pages 1519–1519, Los Alamitos, CA, USA, May 2022. IEEE Computer Society.

[104] R. Shokri, M. Stronati, C. Song, and V. Shmatikov. "Membership inference attacks against machine learning models." In 2017 IEEE Symposium on Security and Privacy (SP), pages 3–18. IEEE, 2017.

[105] A. Suri and D. Evans. "Formalizing and estimating distribution inference risks." Proceedings on Privacy Enhancing Technologies, 2022.

[106] H. Chaudhari, J. Abascal, A. Oprea, M. Jagielski, F. Tramer, and J. Ullman. "SNAP: Efficient extraction of private properties with poisoning." In 2023 IEEE Symposium on Security and Privacy (SP), 2023.

[107] K. Ganju, Q. Wang, W. Yang, C. A. Gunter, and N. Borisov. "Property inference attacks on fully connected neural networks using permutation invariant representations." In Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, CCS'18, page 619–633, New York, NY, USA, 2018. Association for Computing Machinery.

[108] S. Mahloujifar, E. Ghosh, and M. Chase. "Property inference from poisoning." In 2022 IEEE Symposium on Security and Privacy (SP), pages 1120–1137, 2022.

[109] G. Ateniese, L. V. Mancini, A. Spognardi, A. Villani, D. Vitali, and G. Felici. "Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers." International Journal Security and Networking, 10(3):137–150, September 2015.

[110] W. Zhang, S. Tople, and O. Ohrimenko. "Leakage of dataset properties in Multi-Party machine learning." In 30th USENIX Security Symposium (USENIX Security 21), pages 2687–2704. USENIX Association, August 2021.

[111] F. Tramer, F. Zhang, A. Juels, M. K. Reiter, and T. Ristenpart. "Stealing machine learning models via prediction APIs." In USENIX Security, 2016.

[112] V. Chandrasekaran, Kamalika Chaudhuri, Irene Giacomelli, Somesh Jha, and Songbai Yan. Exploring connections between active learning and model extraction. In Proceedings of the 29th USENIX Conference on Security Symposium, SEC'20, USA, 2020. USENIX Association.

[113] M. Jagielski, N. Carlini, D. Berthelot, A. Kurakin, and N. Papernot. "High accuracy and high fidelity extraction of neural networks." In Proceedings of the 29th USENIX Conference on Security Symposium, SEC'20, USA, 2020. USENIX Association.

[114] N. Carlini, M. Jagielski, and I. Mironov. "Cryptanalytic extraction of neural network models." In Advances in Cryptology–CRYPTO 2020, pages 189–218, Cham, 2020. Springer International Publishing.

[115] C. A. Choquette-Choo, F. Tramer, N. Carlini, and N. Papernot. "Label-only membership inference attacks." In Marina Meila and Tong Zhang, editors, Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research, pages 1964–1974. PMLR, 18–24 Jul 2021.

[116] J. Ye, A. Maddi, S. K. Murakonda, V. Bindschaedler, and R. Shokri. "Enhanced membership inference attacks against machine learning models." In Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security, CCS'22, page 3093–3106, New York, NY, USA, 2022. Association for Computing Machinery.

[117] A. Sablayrolles, M. Douze, C. Schmid, Y. Ollivier, and H. Jegou. "White-box vs black-box: Bayes optimal strategies for membership inference." In ICML, volume 97 of Proceedings of Machine Learning Research, pages 5558–5567. PMLR, 2019.

[118] K. Leino and M. Fredrikson. "Stolen Memories: Leveraging model memorization for calibrated white-box membership inference." In Proceedings of the 29th USENIX Conference on Security Symposium, SEC'20, USA, 2020. USENIX Association.

[119] M. Nasr, R. Shokri, and A. Houmansadr. "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning." In IEEE Symposium on Security and Privacy, pages 739–753. IEEE, 2019.

[120] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. Tygar, "Adversarial machine learning." In Proceedings 4th ACM Workshop Security and Artificial Intelligence, 2011.

[121] B. Biggio et al., "Security evaluation of support vector machines in adversarial environments." in Support Vector Machines Applications, 2014.

[122] N. Papernot, P. McDaniel, A. Sinha, and M. Wellman, "Towards the science of security and privacy in machine learning," arXiv preprint arXiv:1611.03814, 2016.

[124]N. Dalvi, P. Domingos, S. Sanghai and D. Verma, "Adversarial classification." In Proceedings, 10th ACM International Conference on Knowledge Discovery and Data Mining, 2004.

[125] M. Fredrikson, S. Jha, and T. Ristenpart, "Model inversion attacks that exploit confidence information and basic countermeasures," in Proceedings, 22nd ACM Conference Computer and Communications Security, 2015.

[125] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y Bengio. 2014. "Generative Adversarial Networks." arXiv:1406.2661 [stat.ML]

[126] W. Hu and Y. Tan. 2017. "Generating Adversarial Malware Examples for Black-Box Attacks Based on GAN." CoRR abs/1702.05983 (2017). arXiv:1702.05983

[127] K. Grosse, N. Papernot, P. Manoharan, M. Backes, and P. D. McDaniel. 2017. "Adversarial Examples for Malware Detection." In ESORICS.

[128] D. Arpit, S. Jastrze¸bski, N. Ballas, D. Krueger, E. Bengio, M. S. Kanwal, T. Maharaj, A. Fischer, A. Courville, Y. Bengio, and S. Lacoste-Julien, "A closer look at memorization in deep networks," in Proceedings of the 34th International Conference on Machine Learning, ser. Proceedings of Machine Learning Research, D. Precup and Y. W. Teh, Eds., vol. 70. International Convention Centre, Sydney, Australia: PMLR, 06–11 Aug 2017, pp. 233–242.

[129] J. Jo and Y. Bengio, "Measuring the tendency of CNNs to learn surface statistical regularities." CoRR, vol. abs/1711.11561, 2017. [Online]. Available: http://arxiv.org/abs/1711.11561

[130] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition." in 23rd ACM SIGSAC Conference on Computer and Communications Security, 2016, pp. 1528–1540.

[131] M. Barreno, B. Nelson, R. Sears, A. D. Joseph, and J. D. Tygar, "Can machine learning be secure?" in ACM Symposium on Information, Computer and Communications Security, 2006,

pp. 16–25.

[132] A. Kołcz and C. H. Teo, "Feature weighting for improved classifier robustness." In Proceedings 6th Conference in Email Anti-Spam, Jul. 2009, pp. 1–8.

[133] A. Globerson and S. T. Roweis, "Nightmare at test time: Robust learning by feature deletion." in Proceedings 23rd International Conference in Machine Learning (ICML), Jun. 2006, pp. 353–360.

[134] F. Zhang, P. P. Chan, B. Biggio, D. S. Yeung, and F. Roli, "Adversarial feature selection against evasion attacks," IEEE Transactions on Cybernetics, 2016.

[135] B. Biggio, B. Nelson and P. Laskov, "Support vector machines under adversarial label noise," in Asian Conference Machine Learning, 2011.

[136] H. S. Anderson, J. Woodbridge, and B. Filar, "DeepDGA: Adversarially-Tuned Domain Generation and Detection." In Proceedings 2016 ACM Workshop Artificial Intelligence and Security

[137] N. Šrndic´ and P. Laskov, "Detection of malicious PDF files based on hierarchical document structure." In Proceedings 20th Annual Network Distributed System Security Symposium, Feb. 2013, pp. 1–16.

[138] B. Biggio, G. Fumera, and F. Roli, "Multiple classifier systems for robust classifier design in adversarial environments." The International Journal of Machine Learning and Cybernetics, vol. 1, nos. 1–4, pp. 27–41, Dec. 2010.

[139] B. Nelson, M. Barreno, F. J. Chi, A. D. Joseph, B. I. P. Rubinstein, U. Saini, C. A. Sutton, J. D. Tygar, and K. Xia, "Exploiting machine learning to subvert your spam filter." In Proceedings USENIX Workshop on Large-Scale Exploits and Emergent Threats, Apr. 2008, pp. 1–9.

[140] N. N. Dalvi, P. M. Domingos, Mausam, S. K. Sanghai, and D. Verma, "Adversarial classification." In Proceedings 10th ACM SIGKDD International Conference in Knowledge Discovery and Data Mining, Aug. 2004, pp. 99–108.

[141] N. Akhtar and A. S. Mian, "Threat of adversarial attacks on deep learning in computer vision: A survey." IEEE Access, vol. 6, pp. 14410–14430, Jul. 2018.

[142] A. S. Ross and F. Doshi-Velez, "Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients." in Proceedings 32nd AAAI Conference on Artificial Intelligence, Feb. 2018, pp. 1660–1669

[143] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. L. Yuille, "Mitigating adversarial effects through randomization." In Proceedings 6th International Conference on Learning Representations, May 2018, pp. 17–32.

[144] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks." In Proceedings The Network and Distributed System Security Symposium, Feb. 2018, pp. 1–15.

[145] R. Shokri and V. Shmatikov, "Privacy-preserving deep learning." In Proceedings 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton) pp. 1310–1321.

[146] N. Papernot, M. Abadi, U. Erlingsson, I. J. Goodfellow, and K. Talwar, "Semi-supervised knowledge transfer for deep learning from private training data." In Proceedings 5th International Conference on Learning Representations, Apr. 2017, pp. 1–16.

[147] P. Mohassel and Y. Zhang, "SecureML: A system for scalable privacy-preserving machine learning," in Proceedings. IEEE Symposium on Security and Privacy (SP), May 2017, pp. 19–38.

[148] Y.Vorobeychik, and M.Kantarcioglu, "Adversarial machine learning, " Synthesis Lectures on Artificial Intelligence and Machine Learning, vol. 12, no. 3, pp. 1–169, 2018.

[149] X. Liu, Y. Lin, H. Li, and J. Zhang, "Adversarial examples: Attacks on machine learning- based malware visualization detection method," arXiv preprint arXiv:1808.01546, 2018.

[150] Z.Wang, "Deep learning-based intrusion detection with adversaries," IEEE Access, vol. 6, pp. 38367–38384, 2018.

[151] I. Rosenberg, A. Shabtai, Y. Elovici, and L. Rokach, "Adversarial machine learning attacks and defense methods in the cyber security domain," ACM Computing Surveys (CSUR), vol. 54, no. 5, pp. 1–36, 2021.

[152] Q. Liu, P. Li, W. Zhao, W. Cai, S. Yu, and V. C. Leung, "A survey on security threats and defensive techniques of machine learning: A data-driven view," IEEE Access, vol. 6, pp. 12103–12117, 2018.

[153] S. Qiu, Q. Liu, S. Zhou, and C. Wu, "Review of artificial intelligence adversarial attack and defense technologies," Applied Sciences, vol. 9, no. 5, p. 909, 2019.

[154] B. Biggio and F. Roli, "Wild patterns: Ten years after the rise of adversarial machine learning, " Pattern Recognition, vol. 84, pp. 317– 331, 2018.

[155] V. Duddu, "A survey of adversarial machine learning in cyber warfare, "Defence Science Journal, vol. 68, no. 4, pp. 356–366, 2018.

[156] A. L. Buczak and E. Guven, "A survey of data mining and machine learning methods for cyber security intrusion detection," IEEE Communications Surveys Tutorials, vol. 18, no. 2, pp. 1153-1176, 2015.

[157] W. E. Zhang, Q. Z. Sheng, A. Alhazmi, and C. Li, "Adversarial attacks on deep-learning models in natural language processing: A survey," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 11, no. 3, pp. 1–41, 2020.

[158] M. Tavallaee, E. Bagheri, W. Lu, and A. Ghorbani, "A Detailed Analysis of the KDD CUP 99 Data Set." Submitted to Second IEEE Symposium on Computational Intelligence for Security and Defense Applications (CISDA), 2009.