A Study on the Impact of Adversarial Attacks on Machine Learning-Based Intrusion Detection Systems in Smart Grids

Pradhap R
Department of Information Technology
Sri Lanka Institute of Information Technology
Sri lanka
Student ID: IT22337108; Email:IT22337108@sliit.lk

***Abstract*—** Machine learning systems are increasingly being applied in critical areas such as healthcare, finance, and cybersecurity. However, alongside this growth, several vulnerabilities in these models have also been exposed. One significant area of concern is adversarial attacks, where small, carefully designed changes to input data can cause machine learning models to make incorrect predictions. This paper provides a detailed review of recent research in adversarial machine learning, focusing on different types of attacks including evasion, poisoning, and inference attacks. I also discuss the existing defense mechanisms developed to counter these attacks, highlighting their strengths and limitations. Additionally, this review identifies gaps in current research, particularly in the practical application and evaluation of defenses in real-world scenarios. Finally, I outline possible directions for future work in this field, aiming to improve the security and robustness of machine learning systems against adversarial threats.

Keywords— Adversarial Machine Learning, Evasion Attacks, Poisoning Attacks, Model Robustness, Machine Learning Security, Adversarial Defenses

I. INTRODUCTION

In recent years, the rapid development of Artificial Intelligence (AI) and Machine Learning (ML) has transformed various industries and become an essential part of everyday life. From applications like image recognition, speech processing, and medical diagnosis to advanced systems such as autonomous vehicles and smart grids, machine learning has proven to be a powerful tool capable of solving complex problems. As the volume of data increases and computational resources become more accessible, ML systems continue to evolve and deliver impressive results.

However, alongside these advancements, new security challenges have also emerged. One of the most critical issues in this domain is the vulnerability of machine learning models to adversarial attacks. Adversarial attacks involve intentionally manipulating the input data in subtle ways to mislead ML models into making incorrect predictions. These changes are often imperceptible to the human eye but can cause even highly accurate models to produce faulty outputs. For example, an image of a stop sign could be slightly altered so that a self-driving car's vision system misclassifies it as a speed limit sign, posing serious safety risks.

Adversarial attacks can occur in different stages of an ML system's lifecycle, including during the training phase (known as poisoning attacks) or at deployment (evasion attacks). Attackers might also attempt to extract sensitive information from models through model extraction attacks. These threats are particularly concerning in security-sensitive environments such as healthcare systems, financial institutions, and autonomous vehicles, where accurate and reliable ML predictions are crucial.

The study of adversarial machine learning (AML) focuses on understanding these attack techniques and developing effective defense mechanisms. While various strategies, such as adversarial training and defensive distillation, have been proposed to improve model robustness, there is still no foolproof method to counter all types of adversarial examples. The dynamic nature of these attacks and the complexity of ML models make it challenging to anticipate and defend against every possible scenario.

This paper explores the concept of adversarial attacks in machine learning, examining how they work, their potential consequences, and the current state of defense mechanisms. By highlighting the importance of securing ML systems, this research aims to contribute to the growing body of knowledge in making AI applications safer and more trustworthy in the face of evolving cybersecurity threats.

II. PRISMA-BASED LITERATURE REVIEW METHODOLOGY

To ensure my approach was clear, organized, and unbiased, I followed a structured literature review method inspired by the PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) framework. This method helped me identify, select, and analyze relevant studies while keeping the process transparent.

The focus of my study was on adversarial attacks in deep learning systems. Specifically, I looked at the types of attacks, how they are executed, and the defense mechanisms used to protect machine learning models. The research process followed these steps:

Identification

I started by conducting a thorough search using academic databases such as IEEE Xplore, SpringerLink, and Google Scholar. I used keywords like adversarial attacks, deep learning security, poisoning attacks, evasion attacks, FGSM, PGD, Carlini & Wagner attack, and adversarial defenses to gather a wide range of research papers, technical reports, and articles related to the topic.

Screening

In this step, I removed duplicate records and non-peer-reviewed materials such as blogs, whitepapers, and opinion pieces. I also excluded studies that didn’t focus specifically

on adversarial attacks in deep learning or lacked technical depth.

Eligibility
Next, I reviewed the abstracts and full texts of the remaining papers to ensure they were aligned with the scope of my research. Only studies that clearly discussed different types of adversarial attacks, their implementation, and defense techniques in machine learning systems were included.

Inclusion
After the selection process, I chose a set of high-quality, relevant research papers for in-depth analysis. These studies covered popular adversarial attack methods like the Fast Gradient Sign Method (FGSM), Projected Gradient Descent (PGD), and Carlini & Wagner (C&W) attacks. I also looked into various defense mechanisms and security challenges in deep learning models.

Experimental-Focus
In addition to the literature review, I also analyzed the datasets and models used in adversarial attack experiments. Given the limited availability of public datasets for certain protocols (like IEC 60870-5-104), I emphasized the need for new datasets to support future research in this area.

Moreover, I explored how attacks are launched in different scenarios, such as during the training phase (poisoning attacks) or inference phase (evasion attacks). I also examined the level of access an attacker has to a target model—whether full knowledge (white-box), partial knowledge (gray-box), or no knowledge (black-box) to understand the different threat models involved.

III. Understanding Deep Learning Models and Adversarial Attacks

A. Convolutional Neural Networks (CNN)

In the field of deep learning, Convolutional Neural Networks (CNNs) are a specialized type of neural network commonly used for tasks in computer vision, such as image recognition. CNNs take input in the form of 2D structures and work by creating feature maps through layers of processing, a process known as sub-sampling. These networks are designed to function similarly to the human brain, enabling them to preserve key details of the input data, even when there are small distortions or changes. One of the important techniques used in CNNs is max-pooling, which reduces the dimensionality of the data, making it more manageable while still retaining the essential features.

A key aspect of CNNs is their ability to extract features from data, making them extremely useful for preprocessing tasks in image recognition. However, researchers have found that CNNs can be vulnerable to attacks, particularly during the training phase. These attacks typically involve making small changes to the input data, such as altering a single pixel, modifying a few pixels by a small amount, or combining both approaches. These slight modifications can cause the network to misclassify the data, leading to a significant drop in its performance.

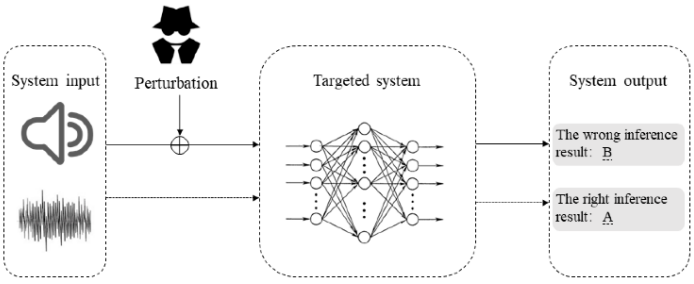
For example, an attacker might change just one pixel in an image, and the CNN might incorrectly identify the image as something entirely different, even though the change is almost imperceptible to the human eye. Such attacks have been shown to reduce the model's confidence in its predictions by up to 84%, highlighting the vulnerability of CNNs to these types of adversarial manipulations. Researchers are continually exploring ways to defend against such attacks to make CNN models more robust and reliable. [1] [2]

B. Recurrent Neural Networks (RNNs)

Recurrent Neural Networks (RNNs) are a type of neural network designed to handle sequential data, such as text or speech. Unlike traditional neural networks, RNNs have a unique feature: they maintain a state known as Long Short-Term Memory (LSTM). This allows RNNs to process data in a flexible way, making them suitable for tasks like handwriting recognition or speech recognition. The key advantage of RNNs is their ability to process large amounts of sequential data, as they not only connect neurons across different layers but also allow neurons within the same layer to be connected.

However, RNNs can be vulnerable to adversarial attacks, especially when applied to sequence data. These attacks involve manipulating the input data to deceive the network into making incorrect predictions or classifications. In one example, attackers can craft adversarial sequences specifically designed to fool the network into misclassifying or incorrectly processing the data. This can be particularly harmful in applications like fake reviews or spam detection, where the RNN could generate misleading content or identify legitimate content as spam.

For instance, attackers could generate fake reviews using an RNN-based model, making them indistinguishable from real reviews. These fake reviews are inexpensive to produce and can scale easily, making them harder to detect. The attack can be further refined by adjusting the rate at which reviews are generated, avoiding patterns that might make them detectable by traditional systems. To counter these threats, researchers are developing new approaches to defend against these types of attacks. One such approach focuses on the vulnerabilities of RNNs, particularly the information loss that occurs during the training process, which can be exploited to create adversarial sequences.



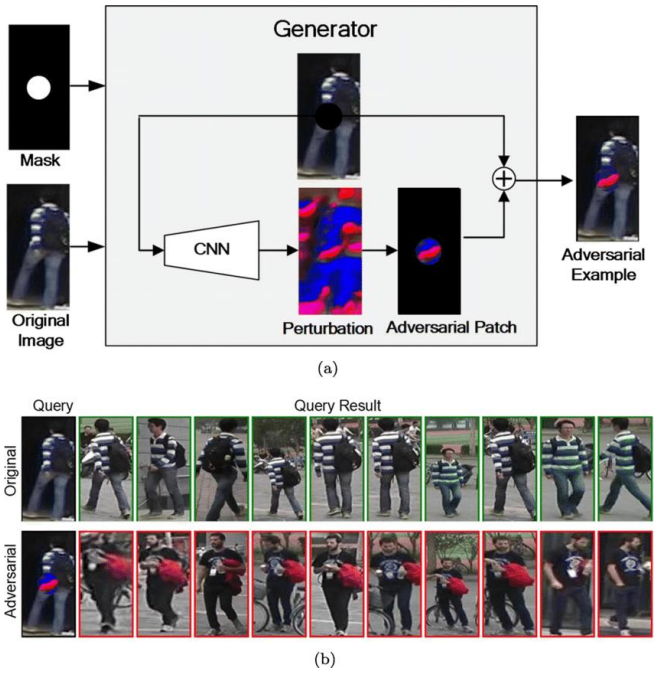
C. *Generative Adversarial Networks*

Generative Adversarial Networks (GANs) are a type of deep learning model widely used for generating new data by learning the underlying distribution of a given set of training examples. The purpose of a GAN is to study the training data and then generate new examples that match the same probability distribution. This makes GANs very effective for tasks like image generation. They are particularly popular in privacy-sensitive applications such as medical diagnoses and facial recognition, where protecting personal information is crucial.

However, GANs can also be used for malicious purposes, such as in attacks that aim to exploit the privacy of deep learning models. One example is a Model Inversion Attack (MIA), which was explored by researchers like Fredrikson et al. (2015). In an MIA, an attacker gains access to a trained deep learning model and uses it to extract private information about the individuals in the dataset, such as personal medical records. This is a harmful scenario, especially in sensitive fields like healthcare. [3]

In particular, GANs can be used for generative model inversion attacks, as proposed by Zhang et al. (2019). Their method involves two steps: First, a GAN is trained on publicly available data to learn how to generate realistic images that resemble private, sensitive data. Then, the trained GAN is used to reverse-engineer private data from a deep learning model. For example, if the model is trained to recognize facial features, an attacker could use GANs to generate a realistic image of someone's face, even without access to their actual image.

While GANs can also be used to generate adversarial examples to improve the robustness of defensive models, the potential for their use in privacy attacks is a significant concern. These attacks highlight the need for stronger security measures in deep learning systems, especially those involved in sensitive areas like medical diagnoses or biometric recognition. [4]



IV. Basics of Adversarial Machine Learning

Adversarial Machine Learning (AML) focuses on these attacks, particularly targeting classification algorithms. Classification algorithms are designed to recognize patterns, such as identifying objects in images or categorizing videos. These models go through a training phase, where they are fed a large dataset of labeled images. During training, the model adjusts its internal parameters to correctly classify the images based on their pixel values. After training, the model is tested with new images it hasn't seen before, and it is expected to classify them accurately.

Humans can classify images based on features like shape, size, color, and orientation. However, machine learning models rely on pixel values to make predictions. If an attacker subtly alters these pixel values, it can trick the model into misclassifying an image, even though the changes are often imperceptible to the human eye. This vulnerability is what adversarial attacks exploit, often by introducing small amounts of noise that degrade the model's performance.

A. *Influence of Attack*

- **Causative Attack:** This type of attack happens during the training phase. Attackers tamper with the training data to mislead the model. This is also known as a poisoning attack. For example, an attacker might add mislabeled images to the training set, causing the model to learn incorrect patterns.
- **Exploratory Attack:** These attacks occur after the model has been trained. Attackers exploit the adversarial space to force the model into making incorrect predictions. An example would be feeding a carefully crafted adversarial image to the trained model, causing it to misclassify the image. [5]

B. *Specificity of Attack*

- **Targeted Attack:** In a targeted attack, the goal is to manipulate the model to produce a specific, incorrect output. For instance, an attacker might alter an image of a cat to make the model classify it as a dog, even though it is clearly a cat.

- **Indiscriminate Attack:** This type of attack aims to cause random misclassifications without targeting a specific output. The attacker doesn't control what the model predicts, they just want to disrupt its performance. [5]

C. *Security Violation*

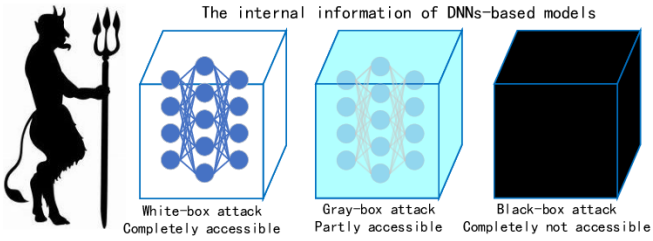
- **Integrity Attack:** The aim of an integrity attack is to lower the model's true positive rate, making it less effective at detecting or classifying objects correctly. For example, an attacker might modify data to reduce the accuracy of a spam filter, causing it to miss many spam messages.

- **Availability Attack:** These attacks aim to disable or disrupt the system entirely. A common example is a denial-of-service attack, where the attacker floods the system with input, causing it to crash or become unresponsive. [5]

Adversarial attacks are a significant challenge for machine learning models, as they can severely impact performance and security. Whether manipulating the training data or introducing small, subtle changes to input data, attackers can exploit these vulnerabilities to degrade a model's accuracy and reliability. Understanding the different types of attacks and their impacts is crucial for developing more robust and secure machine learning systems.

V. Types of Adversarial Machine Learning (AML)

Adversarial attacks are classified into white-box, gray-box, and black-box attacks, depending on how much information the attacker has about the target machine learning model. [5] [6] [7] [8]



Types of Adversarial Machine Learning Attacks		
Attack Type	**Attacker's Knowledge**	**Example**
White-Box Attack	Full knowledge of the model, including its architecture, parameters, hyperparameters, and training data.	An insider with access to a facial recognition system manipulates images to bypass security.
Black-Box Attack	No or very limited knowledge of the model. The attacker can only observe outputs for given inputs.	An external attacker tests different images on a public image classifier API to learn its behavior.
Gray-Box Attack	Partial knowledge of the model, such as knowing the algorithm type or system architecture, but not the trained parameters or dataset.	A developer who knows the target system uses a CNN model but lacks access to its training data or settings.

VI. How Poisoning and Evasion Attacks Compromise ML Models

1. Poisoning Attacks

Poisoning attacks happen during the **training phase** of a machine learning model. In this type of attack, the attacker intentionally inserts misleading or harmful data into the training data set. The goal is to corrupt the learning process so that the model behaves incorrectly when it is later used on new, unseen data. [9] [1] [10] [11]

How it works
The attacker adds specially crafted, incorrect data points to the training set. These poisoned data points are designed in a way that causes the model to learn wrong patterns, reducing its overall accuracy or causing it to make specific errors.

Example:
Imagine a spam detection system trained with email data. An attacker could add emails containing spam content but label them as 'not spam' in the training dataset. As a result, the model might learn to classify actual spam emails as safe messages once it's deployed.

2. Evasion Attacks

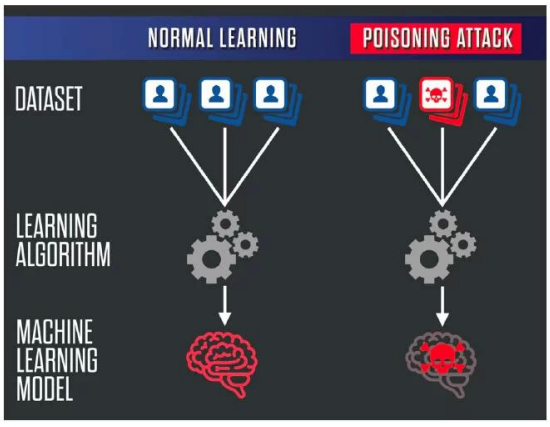
Evasion attacks happen during the **inference phase**, after the model has already been trained and deployed. In this type of attack, the attacker slightly changes the input data in a way that causes the model to make incorrect predictions, without the changes being noticeable to a human observer. [1] [10] [9] [12] [13]

How it works
The attacker adds small, carefully calculated changes (called *perturbations*) to the input data. These changes are often invisible to the human eye but are enough to trick the model into making wrong decisions.
Evasion attacks can be further divided into:

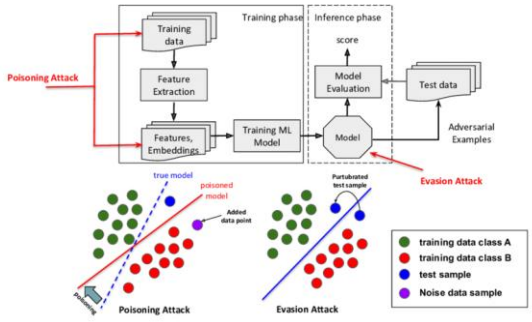
- **Targeted attacks:** Where the attacker forces the model to predict a specific incorrect output. *Example:* Changing an image of a "dog" so the model classifies it as a "cat."
- **Untargeted attacks:** Where the attacker's goal is simply to cause any incorrect prediction, without caring what the wrong output is. *Example:* Modifying an image so the model predicts anything other than the correct label.

Example:
An attacker slightly modifies an image of a stop sign by adding tiny noise that humans can't notice. But when the modified image is fed into a self-driving car's model, it might incorrectly identify it as a speed limit sign, leading to dangerous consequences.

Training Data Poisoning (Backdoor Poisoning)



A schematic representation of the distinction between evasion attacks and poisoning attacks.



VII. Addressing Adversarial Attacks in Smart Grid Intrusion Detection Systems

Adversarial attacks in machine learning (ML) models used for Intrusion Detection Systems (IDS) in smart grid (SG) communications, specifically for the IEC 60870-5-104 protocol. Adversarial attacks are deliberate attempts to manipulate or mislead machine learning models, causing them to perform incorrectly or fail. Here's how this study contributes to countering these challenges. [9] [14]

The lack of realistic datasets for IEC 60870-5-104 systems has been a major barrier to enhancing security, as it hampers the development and testing of effective IDS models. This research addresses that gap by providing a clean, labeled dataset from the IEC 60870-5-104 communication protocol, containing detailed information from the protocol's header. This dataset enables the training of machine learning models specifically tailored to detect cyberattacks.
By using real-world data, these models can be trained to identify threats more accurately, making them less susceptible to adversarial attacks, where attackers attempt to manipulate or deceive the system with subtle changes to the data.

This study presents a new hierarchical intrusion detection approach for IEC 60870-5-104-based systems. It is designed to detect and classify up to 11 different types of cyberattacks while accurately distinguishing between legitimate and malicious network traffic. Additionally, the approach incorporates an attack family analysis, enabling faster and more precise responses to emerging cyber threats.

By distinguishing between various attack types, this method enhances the system's ability to resist adversarial manipulation. The hierarchical structure allows for more efficient detection, even when attackers craft adversarial examples intended to mislead the system. As a result, this approach strengthens the model's resilience against advanced adversarial tactics. [15]

Enhancing the IDS's Robustness Against Adversarial Attacks

A key contribution of this research is testing the IDS against adversarial machine learning attacks, such as:

- **Fast Gradient Sign Method (FGSM)**

This attack creates small changes to the input data in the direction of the gradient of the loss function. These changes are designed to mislead the model into making incorrect predictions. For example, if a model is trained to identify malicious network traffic, FGSM could slightly adjust the data in a way that causes the model to misclassify malicious traffic as legitimate, even though it's still harmful. [15] [16] [17]

- **Projected Gradient Descent (PGD)**

PGD is an iterative version of FGSM, where small perturbations are repeatedly added to the data in a series of steps. This makes it harder for the model to defend against the attack because the changes accumulate over time. For example, a malicious attacker might use PGD to make subtle, incremental changes to network traffic that eventually lead to a misclassification, despite the IDS being trained to recognize these attacks.

- **Carlini and Wagner (C&W)**

This attack method aims to find the smallest possible perturbation that will cause the model to make an incorrect decision. The goal is to make the change so subtle that it's hard for humans to notice, but the model still misclassifies the input. For example, an attacker might use C&W to manipulate a traffic pattern in a way that's almost identical to normal traffic, but the IDS incorrectly flags it as safe. [15]

By testing the IDS against well-known adversarial attack methods like FGSM, PGD, and C&W, this research ensures that the IDS can maintain a high detection rate even when adversarial techniques are used to alter the input data. This means the IDS is not only effective under normal conditions but also robust in detecting attacks that attempt to evade the system by subtly manipulating the data.

VIII. Conclusion

This research has shown that while deep learning models like CNNs, RNNs, and GANs have greatly improved the performance of intrusion detection systems, they remain highly vulnerable to adversarial attacks. Techniques such as FGSM, PGD, and C&W can easily fool even well-trained models by introducing small, unnoticeable changes to the input data. However, most of the existing studies mainly focus on image classification tasks and few have addressed adversarial threats in critical infrastructure systems like smart grids.

Therefore, there is a clear gap in literature when it comes to securing machine learning-based intrusion detection systems against adversarial attacks in industrial control systems (ICS) using protocols like IEC 60870-5-104. Most of the current defense mechanisms are either too computationally expensive or reduce model accuracy in normal conditions. But without reliable defenses, these systems remain an easy target for cyberattacks.

Another issue is that many publicly available datasets are outdated, incomplete, or not designed with adversarial scenarios in mind. Hence, it becomes difficult to properly

evaluate how different models behave under such attacks. More realistic and up-to-date data sets are urgently needed.

In conclusion, while progress has been made in adversarial machine learning, the research community must focus more on real-world cybersecurity applications. Future work should aim to develop lightweight, effective defense strategies and richer datasets that reflect the challenges faced by critical infrastructure systems today.

IX. REFERENCES

[1] S. A. ,. Y. N. Z. Muhammad Maaz Irfan, Towards Deep Learning: A Review On Adversarial Attacks, Islamabad, Pakistan: International Conference on Artificial Intelligence (ICAI), 2021.

[2] D. C. M. Arjun Thangaraju, "EXPLORING ADVERSARIAL ATTACKS AND DEFENSES IN DEEP LEARNING," 2022 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT), 2022.

[3] J. Z. F. Y. X. L. X. P. T. L. a. B. H. Zhanke Zhou, Model Inversion Attacks: A Survey of Approaches and Countermeasures, New York, NY, USA: ACM, 2024.

[4] G. F. Y. L. J. H. B. L. a. Q. Y. Ye Peng, "Detecting Adversarial Examples for NetworkIntrusion Detection System with GAN," International Conference on Computing, Power, and Communication Technologies (IC2PCT), 2023.

[5] P. H. M. Sagar Kamat, Comprehending and Detecting Vulnerabilities using Adversarial Machine Learning Attacks, International Conference on Artificial Intelligence and Signal Processing (AISP), 2022.

[6] X. X. Y. X. SICONG ZHANG, "A Brute-Force Black-Box Method to Attack Machine Learning-Based Systems in Cybersecurity," 2020.

[7] I. S. T. P. V. Ayush Sinha, "Machine Learning Vulnerabilities at the Edge: A Study of Multi-stage Gray-box Attacks," nternational Conference on Computing Communication and Networking Technologies (ICCCNT) , 2024.

[8] A. A. Huda Ali Alatwi, "Adversarial Black-Box Attacks Against Network Intrusion Detection Systems: A Survey," 021 IEEE World AI IoT Congress (AIIoT), 2021.

[9] 2. A. G. F. A. 1st Mohammed Alkurdi, "Adversarial Attack Resilient ML-Assisted Hardware," 2024 IEEE Computer Society Annual Symposium on VLSI (ISVLSI).

[10] P. R. S. S. S. N. S. K. Tanmay Singh, "Adversarial Attacks and Defences of Various Artificial Intelligent Models," International Conference On Smart Technologies For Smart Nation (SmartTechCon), 2023.

[11] "A Comprehensive Analysis of Poisoning Attack and Defence Strategies in Machine Learning Techniques," 2024 International Conference on Computing, Power, and Communication Technologies (IC2PCT), 2024.

[12] D. M. G. Lourdu Mahimai Doss P, "Transferability of Evasion Attacks in Machine Learning Models," International Conference on Intelligent Systems, 2023.

[13] D. M. G. Lourdu Mahimai Doss, "Evasion and Poison attacks on Logistic Regression-based Machine Learning Classification Model," Eighth International Conference on Science Technology Engineering and Mathematics (ICONSTEM), 2023.

[14] A. Z. ,. B. U. H. Khushnaseeb Roshan, "A Novel Deep Learning based Model to Defend Network Intrusion Detection System against Adversarial Attacks," International Conference on Computing for Sustainable Global Development (INDIACom), 2023.

[15] A. A. (. I. A. (. M. I. S. A.-K. (. M. I. M. Q. (. M. I. HADIR TERYAK, "Double-Edged Defense: Thwarting Cyber Attacks and Adversarial Machine Learning in IEC 60870-5-104 Smart Grids," Associate Editor Thomas I. Strasser., 2023.

[16] N.-T. P. M. Nelson Makau Mutua, "Realistic Adversarial Attacks on Smart Grid Intrusion Detection Systems," 2023.

[17] T. G. A. J.-A. G. Yusuf Ademola Sodiq, "Adversarial-aware Machine Learning for Enhancing," 2024 25th International Middle East Power System Conference (MEPCON), 2024.

[18] M. C. F. M. Giovanni Apruzzese, "Addressing Adversarial Attacks Against Security Systems Based on Machine Learning," 11th International Conference on Cyber Conflict, 2019.

[19] S. NESS, "Adversarial Attack Detection in Smart Grids Using Deep Learning Architectures," University of Vienna, Austria, 2024.

[20] Y. S. a. Y. E. Sagduyu, "Evasion and Causative Attacks with Adversarial," Intelligent Automation, Inc., Rockville, MD 20855, USA, 2017.

[21] Y. W. G. C. X. J. S. Minghao Jiang1, "PST: a More Practical Adversarial Learning-based Defense Against Website Fingerprinting," 2020 IEEE Global Communications Conference, 2020.