PEOPLE'S DEMOCRATIC REPUBLIC OF ALGERIA
الجمهـــوريـــــة الجـزائـريـــــة الديـمـقـراطيـــــة الشـعـبيـــــة
MINISTRY OF HIGHER EDUCATION AND SCIENTIFIC RESEARCH
وزارة التعليـم العالـــي والبحـــث العلمـي
HIGHER SCHOOL OF COMPUTER SCIENCE AND DIGITAL TECHNOLOGIES-BEJAIA
المـــدرسة العـــليا في علـــوم وتكنـولوجيـات الاعـــلام الالـــي بجـــاية

Dissertation Submitted to the Department Of Computer Science in Partial Fulfillment of the Requirements for Master's Degree in Computer Science

Specialty: Artificial Intelligence and Data Sciences

Submitted By:

Ahmed Yacine Bouchouareb

Theme

Enhancing Adversarial Robustness in Machine Learning: Techniques and Evaluations

Supervised by :

Mrs. Radia Kassa ESTIN
Pr. Pierre-Martin Tardif *University of Sherbrooke*
Dr. Jordan Félicien Masakuna *University of Sherbrooke*
Mr. DJeff Kanda Nkashama *University of Sherbrooke*

Members of jury:

▪ **Dr.** HARFOUCHE Lynda	President	ESTIN
▪ **Dr.** DAOUDI Meroua	Examiner	ESTIN
▪ **Mrs.** DJENNANE Lynda	Examiner	ESTIN
▪ **Mrs.** KHELOUF Hanane	Examiner	ESTIN
▪ **Mrs.** KASSA Radia	Supervisor	ESTIN

Academic year: 2023/2024

Abstract

This master's report aims to provide a comprehensive review of the literature on the robustness of machine learning models against adversarial attacks. The primary objectives are to explore existing methodologies, highlight key research findings, and identify gaps in current knowledge. The report examines autoencoder-based approaches for detecting adversarial examples as well as other defensive techniques such as adversarial training and regularization techniques. Various adversarial crafting methods, such as Fast Gradient Sign Method (FGSM)[10] and Projected Gradient Descent (PGD)[17], are analyzed in depth. The insights gained will serve as a solid foundation for the development of more robust models in future research.

Keywords— Machine Learning, Adversarial Examples, Robustness, Autoencoders, FGSM, PGD, Anomaly Detection, Adversarial Attacks

Résumé

Ce mémoire de master vise à fournir une revue complète de la littérature sur la robustesse des modèles d'apprentissage automatique face aux attaques adversariales. Les objectifs principaux sont d'explorer les méthodologies existantes, de mettre en avant les résultats de recherche majeurs et d'identifier les lacunes dans les connaissances actuelles. Le rapport examine les approches basées sur les autoencodeurs pour détecter les exemples adversariaux ainsi que d'autres techniques défensives telles que l'entraînement adversarial et les techniques de régularisation. Diverses méthodes de génération d'attaques adversariales, comme FGSM[10] et PGD[17], sont analysées de manière approfondie. Les connaissances acquises serviront de base solide pour le développement de modèles plus robustes dans les recherches futures.

Mots-clés— Apprentissage automatique, Exemples adversariaux, Robustesse, Autoencodeurs, FGSM, PGD, Détection d'anomalies, Attaques adversariales

الملخص

يهدف هذا البحث إلى تقديم مراجعة شاملة للأدبيات المتعلقة بصلابة نماذج التعلم الآلي في مواجهة الهجمات الخصمية. تتمثل الأهداف الرئيسية في استكشاف المنهجيات الحالية، وتسليط الضوء على النتائج البحثية الرئيسية، وتحديد الثغرات في المعرفة الحالية. يستعرض التقرير الأساليب المعتمدة على الشبكات التلقائية لاكتشاف الأمثلة الخصمية، بالإضافة إلى تقنيات دفاعية أخرى مثل التدريب الخصمي وتقنيات التنظيم. كما يتم تحليل طرق إنشاء الهجمات الخصمية المختلفة، مثل *FGSM* و *PGD* بعمق. وستوفر المعرفة المكتسبة أساسًا قويًا لتطوير نماذج أكثر متانة في الأبحاث المستقبلية.

الكلمات المفتاحية: التعلم الآلي، الأمثلة العدائية،الترميز التلقائي، *PGD* ،*FGSM*، اكتشاف الشذوذ، الهجمات العدائية.

Contents

List of Figures

List of Acronyms

AD Anomaly Detection

AEs Autoencoders

AI Artificial Intelligence

AL Adversarial Learning

C&W Carlini & Wagner

CNNs Convolutional Neural Networks

DL Deep Learning

DNN Deep Neural Networks

DP Data Poisoning

ECG Electrocardiogram

FGSM Fast Gradient Sign Method

FL Federated Learning

GANs Generative Adversarial Networks

MAE Mean Absolute Error

ML Machine Learning

MSE Mean Squared Error

PGD Projected Gradient Descent

RE Reverse Engineering

RL Reinforcement Learning

RNNs Recurrent Neural Networks

SGD Stochastic Gradient Descent

SimCLR Simple Framework for Contrastive Learning of Visual Representations

TTE Test-Time Evasion

CHAPTER 1

Introduction

1.1 Context and Importance

Machine learning (ML) has become integral to a wide range of applications, from image recognition and natural language processing to autonomous driving and healthcare diagnostics [16, 26]. As these models are increasingly deployed in critical and high-stakes environments, ensuring their robustness—especially against adversarial attacks—has become a significant concern [10, 28].

Adversarial attacks involve deliberately crafting inputs that are designed to deceive machine learning models, often leading to incorrect or even dangerous predictions [4]. These attacks highlight vulnerabilities in machine learning systems, especially in models that perform well under standard conditions but fail under adversarial manipulation. Given the potential consequences of such failures, enhancing the robustness of machine learning models is not just an academic challenge but a practical necessity [17].

1.2 Objectives of the Report

The goals of this master's report are to provide a global overview of the field of adversarial machine learning, with a specific focus on exploring various adversarial attack strategies and the defensive techniques developed to counter them. Specifically, the objectives are to:

- Gain a comprehensive understanding of the landscape of adversarial attacks and their implications for machine learning models.
- Explore and evaluate the effectiveness of different defense mechanisms proposed in the literature to mitigate these attacks.
- Identify gaps in the existing research and suggest potential directions for future exploration in the field of adversarial robustness.

In addition to the literature review, this report discusses the practical application and testing of a method that considers the reconstruction loss as a vector for detecting adversarial examples, as proposed in a recent study[30]. The efficacy of this approach is evaluated in terms of its ability to differentiate between normal, anomalous, and adversarial data points.

1.3 Structure of the Report

The report is organized as follows:

- **Chapter 2: Basics of Machine Learning and Deep Learning** – This chapter provides a foundational overview of machine learning and deep learning concepts, necessary for understanding the subsequent chapters.
- **Chapter 3: Robustness in Machine Learning** – This chapter discusses the concept of robustness in machine learning, including challenges and the importance of creating models that can withstand adversarial conditions.

- **Chapter 4: Adversarial Examples** – This chapter focuses on adversarial examples, detailing various crafting methods such as the FGSM[10] and the PGD[17], as well as their impact on machine learning models.
- **Chapter 5: Literature Review** – This chapter presents a comprehensive review of key papers, organized by themes such as adversarial defenses and robustness techniques, while identifying gaps and suggesting future research directions.
- **Chapter 6: Conclusion** – This chapter summarizes the main findings of the report, discusses the broader implications, and suggests directions for future research.

1.4 Scope and Limitations

This report primarily focuses on supervised learning models and their robustness against adversarial attacks. While some discussions may touch upon unsupervised learning these areas are not the primary focus. Additionally, this review is limited to adversarial defense techniques that have been widely studied in the literature, excluding some emerging methods due to time constraints.

The dataset discussed in this report, NSL-KDD, is a widely used dataset for network intrusion detection. [29]. The review also emphasizes methods that are computationally feasible for practical deployment, leaving out some resource-intensive techniques that may not be applicable in real-world scenarios.

CHAPTER 2

Basics of Machine Learning and Deep Learning

2.1 Introduction

Machine Learning (ML) and Deep Learning (DL) are critical components of Artificial Intelligence (AI), enabling systems to learn from data and make predictions. This chapter introduces the fundamental concepts of ML, including its main types: supervised, unsupervised, and reinforcement learning. It also covers deep learning, focusing on neural network architectures like Convolutional Neural Networks (CNNs),Recurrent Neural Networks (RNNs), and Autoencoders (AEs).

We will explore how these models are trained and evaluated, discussing key concepts like loss functions and optimization algorithms. The chapter concludes by addressing common challenges in ML, such as overfitting, data quality, and adversarial examples, which are crucial for understanding the need for robust and reliable AI models.

2.2 Machine Learning

ML is a branch of AI that focuses on developing algorithms that can learn from and make predictions based on data. Unlike traditional programming, where rules are explicitly programmed, machine learning models identify patterns in data to make decisions with minimal human intervention [2].

2.2.1 Types of Machine Learning

Machine learning can be broadly categorized into three types:

- **Supervised Learning**: In supervised learning, models are trained on labeled data, where each training example is paired with an output label. The goal is to learn a mapping from inputs to outputs based on the provided examples [20].
- **Unsupervised Learning**: Unsupervised learning deals with unlabeled data. The model attempts to uncover hidden patterns or structures in the input data, such as clustering similar data points or reducing dimensionality [11].
- **Reinforcement Learning (RL)**: In RL, an agent learns to make decisions by interacting with an environment. The agent receives rewards or penalties based on its actions, and the goal is to learn a policy that maximizes cumulative rewards over time [27].

Fig2.1 shows the types of ML.

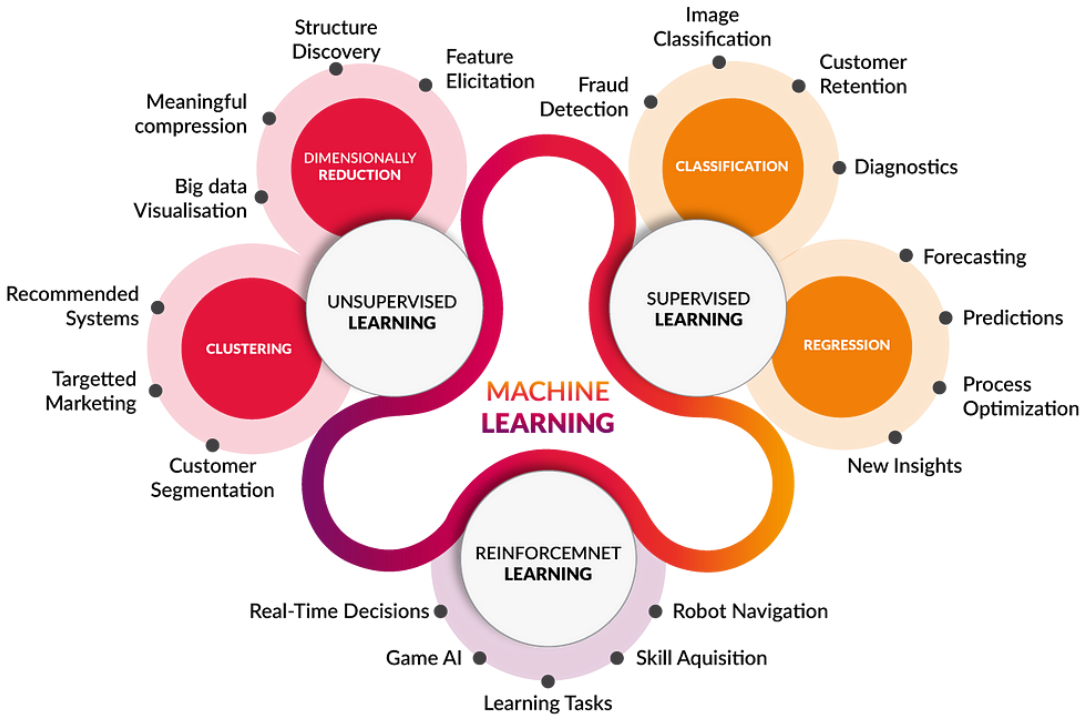

Figure 2.1: Types of Machine Learning source

2.3 Deep Learning

DL is a subset of machine learning that utilizes neural networks with many layers (hence "deep"). These networks, known as deep neural networks , are particularly effective at learning hierarchical representations of data [16].

2.3.1 Neural Networks

A neural network is a computational model inspired by the human brain, consisting of layers of nodes (neurons). Each neuron receives inputs, processes them, and passes the output to the next layer. The network learns by adjusting the weights of connections between neurons based on the error in predictions [9].

2.3.2 Key Architectures in Deep Learning

Several key neural network architectures have been developed for different types of tasks:

- **Convolutional Neural Networks (CNNs)**: Primarily used for image and spatial data, CNNs are designed to automatically and adaptively learn spatial hierarchies of features [15]. The figure 2.2 shows the architecture of a CNN.

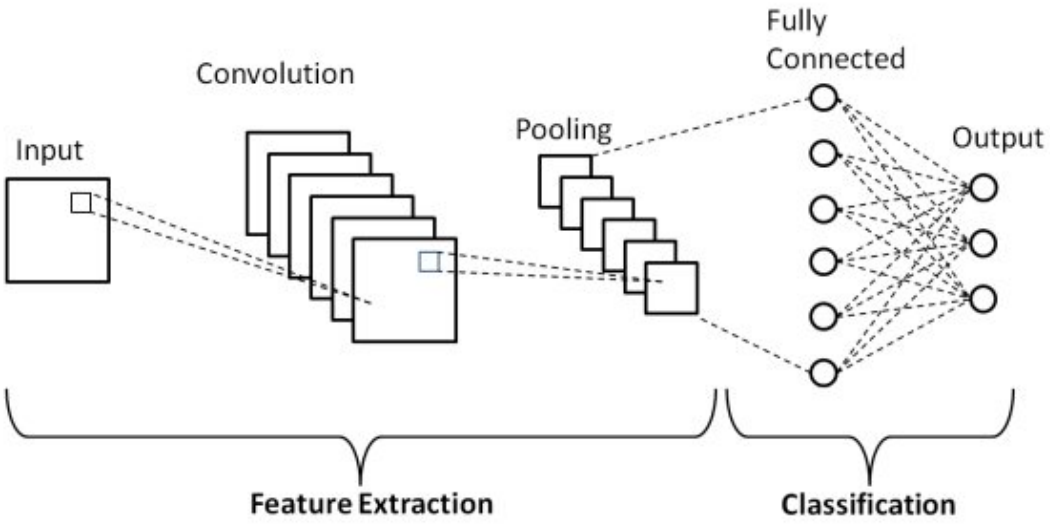

Figure 2.2: Convolutional Neural Network source

- **Recurrent Neural Networks (RNNs)**: These are designed for sequential data, such as time series or natural language, where the order of data points is important. RNNs maintain a hidden state that captures information from previous time steps [13].

  The figure 2.3 shows the difference between the architecture of an RNN and a feed forward neural network.

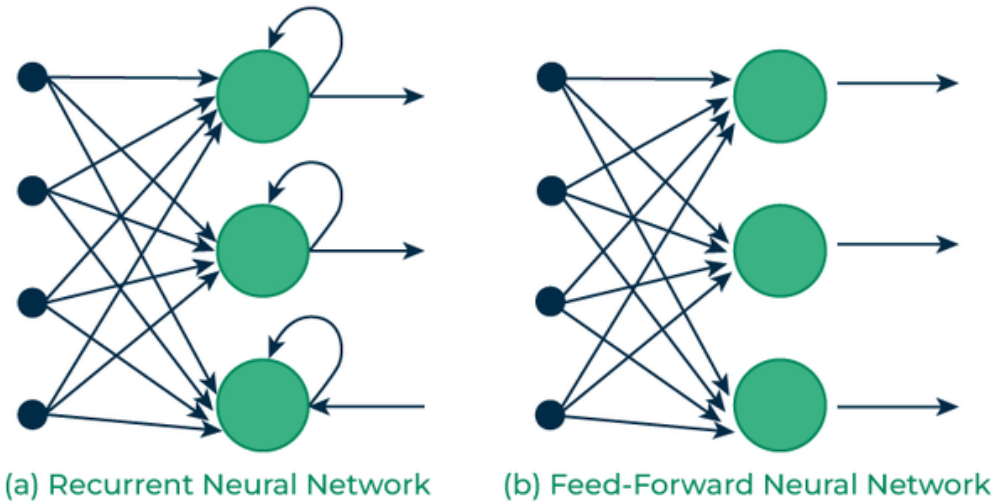

Figure 2.3: Recurrent Neural Network source

- **Autoencoders (AEs)**: AEs are unsupervised learning models designed for dimensionality reduction and feature learning. They consist of an encoder that compresses the input data and a decoder that attempts to reconstruct the original data from this compressed representation [12].

  This figure 2.4 shows the architecture of an autoencoder.

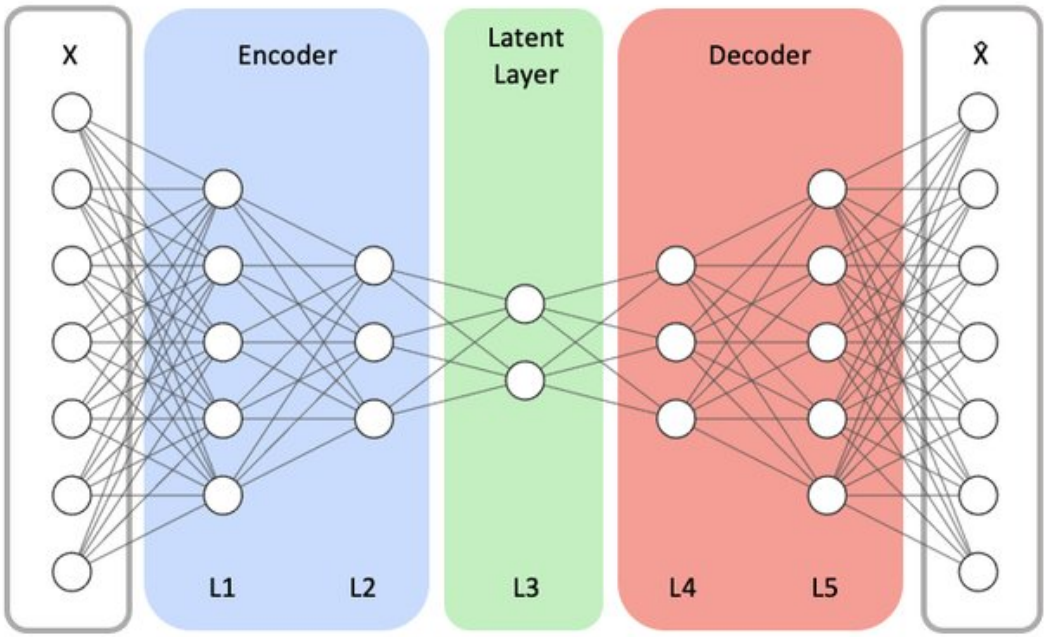

Figure 2.4: Autoencoder Architecture source

2.4 Training and Evaluation of Models

Training a machine learning model involves finding the optimal set of parameters (weights) that minimize the error on the training data. This process typically involves the following steps:

- **Loss Function**: The loss function quantifies the error between the model's predictions and the actual target values. Common loss functions include Mean Squared Error (MSE) for regression tasks and Cross-Entropy Loss for classification tasks [21].
- **Optimization Algorithm**: Optimization algorithms, such as Stochastic Gradient Descent (SGD) and Adam, are used to minimize the loss function by iteratively updating the model's parameters [14].
- **Model Evaluation**: After training, models are evaluated on a separate test set to assess their generalization performance. Common evaluation metrics include accuracy, precision, recall, and F1-score for classification, and R-squared and Mean Absolute Error (MAE) for regression [24].

2.5 Challenges in Machine Learning

While machine learning has achieved significant success, it also faces several challenges:

- **Overfitting**: Overfitting occurs when a model learns the training data too well, including noise and outliers, leading to poor generalization to new data [11].

- **Data Quality**: The quality of the training data significantly impacts the performance of machine learning models. Poor data quality, including mislabeled data, missing values, and imbalanced classes, can lead to biased or inaccurate models [8].
- **Adversarial Examples**: As discussed in subsequent chapters, adversarial examples represent a significant challenge to the robustness of machine learning models. These are carefully crafted inputs designed to fool the model into making incorrect predictions [28].

2.6 Conclusion

This chapter provided a brief overview of the fundamental concepts in machine learning and deep learning. Understanding these basics is crucial for appreciating the challenges posed by adversarial examples and the importance of robustness in machine learning models, topics that will be explored in greater depth in the following chapters.

CHAPTER 3

Robustness in Machine Learning

3.1 Introduction

Robustness in machine learning refers to the ability of a model to maintain its performance when exposed to perturbations or adversarial conditions. In the context of adversarial attacks, robustness is critical to ensuring that machine learning models can resist and correctly classify inputs that have been intentionally manipulated. This chapter explores the concept of robustness, the challenges in achieving it, and the significance of robust models in real-world applications.

3.2 Challenges in Achieving Robustness

Developing robust machine learning models is challenging due to several factors:

3.2.1 Model Sensitivity to Small Perturbations

Machine learning models, particularly deep neural networks, are known to be sensitive to small perturbations in the input data. This sensitivity can lead to significant changes in the model's output, making it vulnerable to adversarial attacks. Research has shown that even minor modifications, often imperceptible to humans, can cause a model to misclassify an input with high confidence [28].

3.2.2 The Trade-Off Between Robustness and Accuracy

There is often a trade-off between a model's robustness and its accuracy on clean data. Techniques designed to improve robustness, such as adversarial training, can lead to a decrease in accuracy on non-adversarial examples. Balancing this trade-off is a key challenge in the development of robust machine learning models [31].

3.2.3 Generalization of Robustness Across Different Attacks

Most defense mechanisms are designed to protect against specific types of adversarial attacks. However, these defenses may not generalize well to other types of attacks, limiting their effectiveness. Ensuring that a model is robust across a wide range of adversarial scenarios is an ongoing research challenge [3].

3.3 Importance of Robustness in Real-World Applications

Robustness is especially important in applications where the consequences of model failure can be severe, such as in healthcare, autonomous driving, and security systems. In these domains, ensuring that models can resist adversarial attacks is crucial for maintaining safety and reliability.

3.3.1 Healthcare

In healthcare, machine learning models are increasingly used for diagnostic and treatment recommendations. A lack of robustness could lead to incorrect diagnoses or treatment plans, potentially endangering patients' lives. Therefore, robust models are essential to ensure accurate and reliable medical decision-making [7].

3.3.2 Autonomous Driving

Autonomous vehicles rely on machine learning models for tasks such as object detection and navigation. Adversarial attacks on these models could lead to catastrophic outcomes, such as collisions or navigation errors. Robustness is critical in ensuring the safety and reliability of autonomous driving systems [6].

3.3.3 Security Systems

In security applications, machine learning models are used for tasks such as facial recognition and intrusion detection. Adversarial attacks on these systems could lead to security breaches or unauthorized access, highlighting the need for robust models in this domain [1].

3.4 Recent Advances in Enhancing Robustness

Recent research has led to several advances in enhancing the robustness of machine learning models. Some of the key techniques include:

3.4.1 Adversarial Training

Adversarial training, which involves training the model on adversarial examples, is one of the most widely used methods for improving robustness. By exposing the model to adversarial conditions during training, it learns to classify these inputs correctly [17].

3.4.2 Regularization Techniques

Regularization techniques, such as weight regularization and dropout, have been shown to improve robustness by reducing the model's sensitivity to input perturbations. These techniques help prevent overfitting and improve the model's generalization capabilities [9].

3.4.3 Defensive Distillation

Defensive distillation is a technique where the model is trained to predict the probabilities output by a teacher model that has been trained on adversarial examples. This process makes the student model more resistant to adversarial attacks [23].

The figure 3.1 illustrates the concept of defensive distillation.

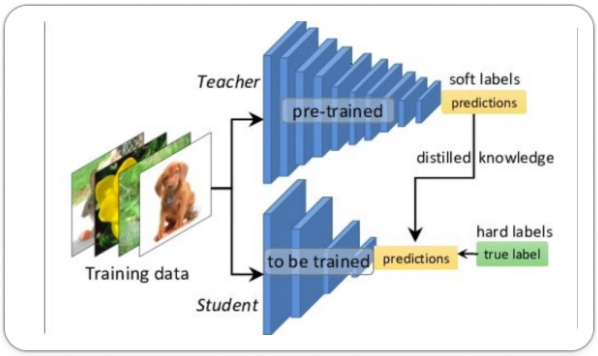

Figure 3.1: Defensive Distillation source minute 1:18

3.5 Conclusion

This chapter discussed the concept of robustness in machine learning, the challenges in achieving it, and its importance in real-world applications. While significant progress has been made, achieving robust models that can generalize across different types of adversarial attacks remains a key area of ongoing research.

CHAPTER 4

Adversarial Examples

4.1 Introduction

Adversarial examples refer to inputs to machine learning models that have been intentionally perturbed to cause the model to make incorrect predictions. These perturbations are often imperceptible to humans but can lead to significant misclassifications by the model. This chapter explores the concept of adversarial examples, the techniques used to craft them, and their impact on the performance of machine learning models.

4.1.1 Categories of Adversarial Attacks

Adversarial attacks pose significant challenges to the robustness and security of machine learning models. According to the survey[32], adversarial attacks can be categorized into several types based on their goals and methodologies:

- **Evasion Attacks**: These attacks aim to manipulate input data to evade detection by the model. Examples include the FGSM[10] and PGD[17].
- **Poisoning Attacks**: Involve injecting malicious data into the training set, causing the model to learn incorrect patterns. This type of attack can degrade the model's performance over time.
- **Exploratory Attacks**: Focus on probing the model to gather information about its structure or parameters, which can then be exploited for further attacks.

4.2 Definition and Characteristics of Adversarial Examples

Adversarial examples are crafted by introducing small, carefully designed perturbations to the input data, which are specifically aimed at deceiving machine learning models. These perturbations are typically designed to be minimal in magnitude, making them difficult to detect by humans[28, 10].

4.3 Crafting Methods

There are several methods to generate adversarial examples, each with varying levels of complexity and effectiveness.

4.3.1 Fast Gradient Sign Method (FGSM)

The FGSM is one of the most widely used techniques for crafting adversarial examples. It leverages the gradients of the loss function with respect to the input data to create a perturbation that maximizes the model's loss.[10]

4.3.2 Projected Gradient Descent (PGD)

PGD is an iterative variant of FGSM, where the perturbation is repeatedly refined through multiple steps of gradient ascent, with each step constrained to stay within a specified perturbation budget.[17]

4.3.3 Carlini & Wagner (C&W) Attack

The C&W attack is a powerful optimization-based method for generating adversarial examples. It formulates the problem as an optimization problem, where the objective is to find the smallest possible perturbation that misclassifies the input while maintaining the perturbation below a certain threshold.[4]

4.4 Effects of Adversarial examples

Adversarial examples have a profound impact on the performance of machine learning models. These perturbed inputs, while often imperceptible to human observers, can lead to significant misclassifications and degrade model accuracy. The effects of adversarial examples are particularly pronounced in high-stakes applications such as medical diagnostics, where accurate predictions are crucial.

4.4.1 Impact on Model Performance

The introduction of adversarial examples can lead to several detrimental effects on machine learning models:

- **Increased Error Rates:** Adversarial examples can significantly increase the error rates of models, leading to poor classification performance.
- **Reduced Generalization:** Models that are susceptible to adversarial attacks often generalize poorly to new, unseen data.
- **Misclassification:** Adversarial attacks can cause models to make incorrect predictions, which can have severe consequences in critical applications.

To illustrate the impact of adversarial examples and the effectiveness of different defense mechanisms, we compare the performance of models under PGD attacks with and without the application of defense techniques.

The figure 4.1 shows the effects of PGD attacks on model performance.

Table 4: Comparison of methods under PGD attacks in situation I

		Accuracy (Performance Drop)	f_1 score (Performance Drop)
	No Defense	0.3906±0.0695 (54.80%±8.08%)	0.2558±0.0555 (66.94%±7.16%)
Baselines	JR	0.7515±0.0272 (12.91%±3.14%)	0.6595±0.0276 (14.96%±3.00%)
	NSR	0.8316±0.0118 (3.88%±0.98%)	0.7112±0.0232 (6.92%±1.92%)
	DD	0.7562±0.0109 (12.76%±1.09%)	0.6330±0.0117 (17.73%±1.06%)
	AT	0.8535±0.0030 (1.22%±0.45%)	0.7469±0.0153 (2.49%±2.19%)
Proposed	Init-CardioDefense	0.8485±0.0045 (2.53%±0.76%)	0.7527±0.0114 (4.59%±1.83%)
	Dist-CardioDefense	0.8551±0.0064 (1.33%±0.43%)	0.7572±0.0104 (2.15%±0.54%)
	CardioDefense	**0.8612±0.0050 (0.89%±0.26%)**	**0.7709±0.0114 (-0.67%±1.77%)**

Figure 4.1: PGD's effects on performance [25]

4.5 Conclusion

This chapter has explored the intricacies of adversarial examples, providing a comprehensive overview of the methods used to craft these deceptive inputs and their potential impact on machine learning models. By categorizing adversarial attacks into evasion, poisoning, and exploratory types, the chapter highlights the diverse strategies employed by attackers and the corresponding challenges in defending against them. The detailed examination of crafting methods such as the FGSM, PGD and the C&W attack further underscores the sophistication of these techniques and their effectiveness in compromising model performance. As adversarial threats continue to evolve, it becomes increasingly important to develop robust countermeasures to safeguard the integrity and accuracy of machine learning systems.

CHAPTER 5

Literature Review

5.1 Introduction

The security of ML models has become a critical concern as these models are increasingly deployed across diverse sectors, including healthcare, finance, autonomous vehicles, and social media. Adversarial attacks, which involve subtle manipulations of input data to deceive ML models, pose a significant threat to the integrity and reliability of these systems. This chapter provides a comprehensive review of the current state of ML security, focusing on the challenges posed by adversarial attacks and the corresponding defense mechanisms. Drawing on insights from two recent surveys, as well as a selection of key papers, the review covers various attack strategies, including the FGSM and PGD, and evaluates the effectiveness of existing defensive techniques such as adversarial training, autoencoder-based Anomaly Detection (AD), and the novel concept of vector reconstruction error. The chapter concludes by identifying gaps in the current research and proposing areas for future investigation, particularly in the context of enhancing the robustness of ML models against sophisticated adversarial threats.

5.2 Comprehensive Overview of ML Security and Adversarial Learning (AL)

The security landscape in ML is rapidly evolving as the deployment of ML models expands across various domains, including medical, military, automotive, and social networking. Despite their widespread success, these models are increasingly vulnerable to a range of adversarial attacks that threaten their integrity and reliability. Two comprehensive surveys provide a detailed examination of these security concerns, offering insights into both the nature of the threats and the defenses developed to counter them.

The first survey systematically analyzes the security issues associated with ML, covering the entire spectrum from training to deployment. It categorizes security threats into five major types: training set poisoning, backdoors in the training data, adversarial example attacks, model theft, and the recovery of sensitive training data. The survey emphasizes that these attacks are not just theoretical but have been demonstrated in real-world conditions, where their stealthy nature is particularly concerning due to the unexplained behavior of deep learning models. Additionally, the survey outlines several security evaluation methods and suggests future research directions aimed at strengthening the resilience of ML systems [32].

The figure 5.1 illustrates the categories of attacks in ML.

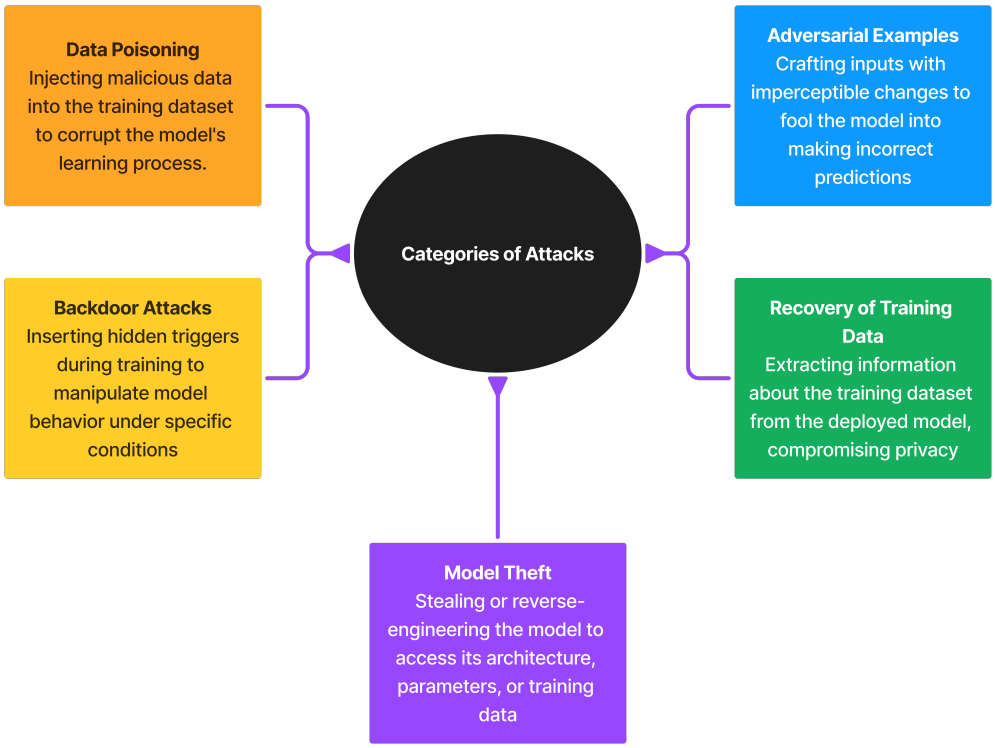

Figure 5.1: categories of attacks in ML

The second survey narrows its focus to AL, specifically targeting the vulnerabilities of Deep Neural Networks (DNN) classifiers. It provides a taxonomy of attacks, such as Test-Time Evasion (TTE), Data Poisoning (DP), backdoor DP, and Reverse Engineering (RE), and critically assesses the corresponding defense mechanisms. The survey distinguishes between robust classification and AD as defense strategies and evaluates the effectiveness of various approaches under different attack scenarios. It challenges conventional wisdom in AL, particularly in areas like the relationship between attack strength and success, and the assumptions regarding an attacker's knowledge of the ground truth. Moreover, the survey delves into novel issues, such as the susceptibility of query-based reverse engineering to AD defenses and the challenges of detecting attacks that aim to embed false content rather than alter classification decisions [18].

The figure 5.2 illustrates the phases of attacks in ML.

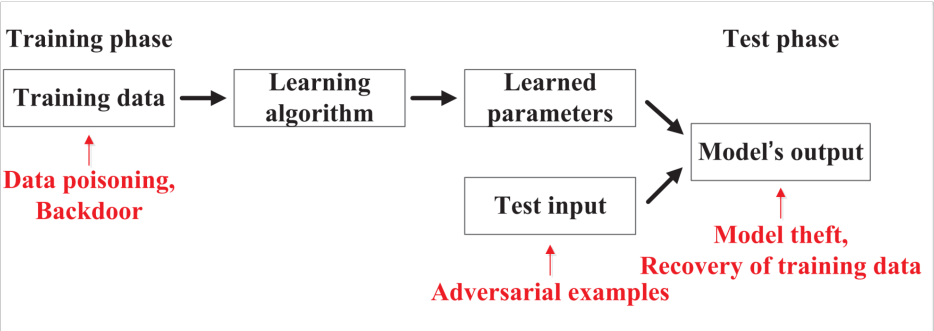

Figure 5.2: Attacks phases [32]

Together, these surveys provide a comprehensive overview of the current state of ML security, highlighting both the breadth of potential threats and the depth of defensive strategies. They underscore the importance of rigorous security evaluations and the need for ongoing research to address unresolved challenges in the field. By integrating insights from these works, this report builds a foundation for understanding the complexities of defending ML models in adversarial environments.

5.3 Attacking Techniques

5.3.1 Fast Gradient Sign Method (FGSM)

FGSM is one of the earliest and most widely used methods for generating adversarial examples. It works by adding perturbations to the input data in the direction of the gradient of the loss function with respect to the input, scaled by a small factor. This method is efficient and easy to implement but is relatively straightforward to defend against using techniques like adversarial training [10].

$$\text{Adversarial Example: } x_{adv} = x + \epsilon \cdot \text{sign}(\nabla_x J(\theta, x, y)) \tag{5.1}$$

5.3.2 Projected Gradient Descent (PGD)

PGD is an iterative extension of FGSM and is considered a more powerful attack. In each iteration, the adversarial perturbation is updated by taking a small step in the direction of the gradient, followed by a projection step to ensure the perturbation remains within a specified norm ball. PGD is recognized as one of the most effective methods for crafting adversarial examples and is often used as a benchmark in evaluating model robustness [17]. The Projected Gradient Descent attack can be formulated as:

$$x_{t+1} = \text{Proj}_{\mathcal{B}_\epsilon(x)} \left(x_t + \alpha \cdot \text{sign} \left(\nabla_x \mathcal{L}(f_\theta(x_t), y) \right) \right)$$

Where:

- x_t is the adversarial example at step t.
- α is the step size.
- $\nabla_x \mathcal{L}(f_\theta(x_t), y)$ is the gradient of the loss function $\mathcal{L}$ with respect to the input x_t.
- $\text{sign}(\cdot)$ is the sign function.
- $\text{Proj}_{\mathcal{B}_\epsilon(x)}(\cdot)$ projects the perturbed example back onto the ϵ-ball around the original input x.

5.3.3 A Robust Analysis of Adversarial Attacks on Federated Learning (FL) Environments

FL has emerged as a prominent branch of AI, particularly with the proliferation of mobile computing and IoT technologies. Unlike traditional centralized learning paradigms, FL relies on a distributed computing approach, where multiple decentralized devices contribute to the learning process. While this methodology enhances privacy by keeping data localized, it also introduces significant security challenges, as many participating devices operate outside the protective scope of a centralized system.

This paper provides a robust analysis of the security vulnerabilities inherent in federated learning environments, focusing on various adversarial attacks that can compromise the integrity and effectiveness of FL models. Key threats include data leakage, communication issues, poisoning attacks, and system manipulation through backdoor mechanisms. These attacks are categorized based on their modus operandi, offering a detailed examination of poisoning and inferencing attacks, among others.

The study not only reviews the different types of attacks but also evaluates the effectiveness of existing defense strategies designed to mitigate these risks in federated environments. By systematically analyzing the challenges posed by adversarial examples in FL, this paper contributes to a deeper understanding of the security issues faced in Federated ML and proposes potential solutions to enhance the robustness of these systems [22].

5.3.4 A Comprehensive Overview of Generative Adversarial Networks (GANs) and Their Applications

GANs have become a pivotal innovation in AI, particularly in generating high-quality synthetic data. By leveraging two neural networks in a zero-sum game framework—where one network generates data and the other attempts to discern real data from synthetic—GANs excel in producing sharp and discrete outputs.

This survey offers an extensive review of GAN architectures, their prevalent variants, and a broad range of applications across numerous sectors. GANs have been widely used in computer vision tasks such as image processing, video generation, and prediction. Additionally, they have found applications in scientific fields like protein engineering, astronomical data analysis, remote sensing image

dehazing, and crystal structure synthesis. Beyond these, GANs have also made significant inroads into finance, marketing, fashion design, sports, and music.

My interest in this survey was driven by the role GANs play in the creation of adversarial examples. GAN-based techniques have been increasingly used to generate adversarial examples that can deceive ML models, a crucial aspect of studying model robustness. Understanding the underlying mechanics and applications of GANs provided essential insights into how these networks can be harnessed for adversarial purposes, as well as how they might be defended against. The survey thoroughly covers the theoretical foundations of GANs, the various variants developed, and the metrics used for evaluation, making it a valuable resource for comprehending the potential and risks associated with GAN-generated adversarial examples [5].

The figure 5.3 illustrates the basic architecture of a GAN.

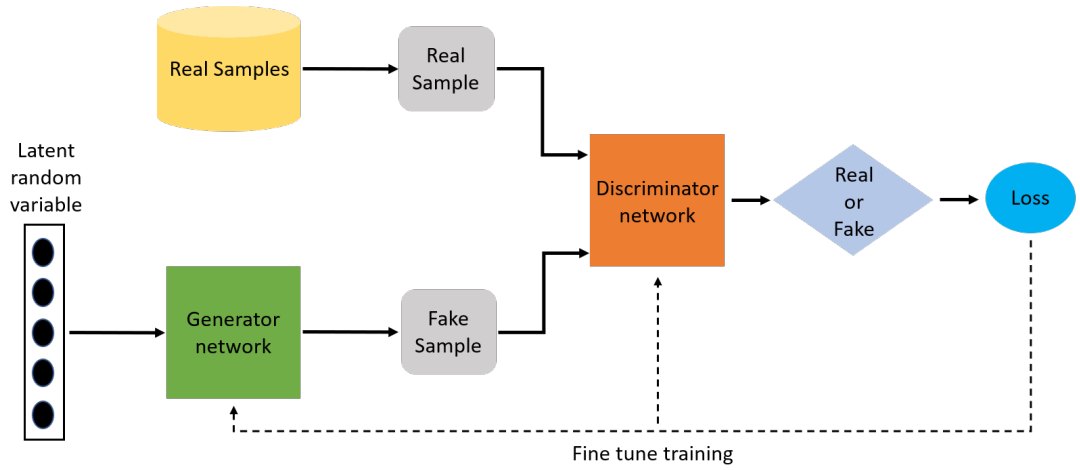

Figure 5.3: Basic GAN architecture [5]

5.4 Defensive Techniques

5.4.1 Adversarial Training

AL is one of the most widely used techniques to enhance model robustness. It involves augmenting the training data with adversarial examples, allowing the model to learn how to correctly classify perturbed inputs. This method has been shown to improve resistance against specific attacks, such as FGSM and PGD, but may require substantial computational resources and can still be vulnerable to more sophisticated attacks.

5.4.2 Training Distillation

Training distillation, or adversarial distillation, is another defense mechanism that enhances robustness. In this technique, a "teacher" model is trained with adversarial examples, and the knowledge from this model is then distilled into a "student" model, which is smaller and potentially more robust. This approach has been applied successfully in various domains, including Electrocardiogram (ECG) classification, as discussed in the paper *Cardio Defense: Defending Against Adversar-*

ial Attack in ECG Classification with Adversarial Distillation Training. Zhang et al. [25] proposes an innovative approach for defending ECG classification models. The technique involves training a model using a mixture of normal and adversarial examples. This process distills the knowledge of adversarial robustness into the model, improving its ability to correctly classify both normal and adversarially perturbed ECG signals. The study demonstrates that this technique enhances the model's resistance to adversarial attacks without significantly compromising its accuracy on clean data.

5.4.3 AD using Autoencoders

AEs have emerged as effective tools for AD due to their capability to learn efficient data representations. This paper proposes a novel method for AD using autoencoders by considering the reconstruction loss as a vector rather than a scalar. By analyzing the vector reconstruction error, the model can capture more nuanced differences between normal and anomalous data points. Liu et al. [30] introduces a method that enhances traditional autoencoder-based AD. The authors report that this approach improves the detection accuracy for various anomalies, providing a more granular view of the reconstruction process and highlighting the potential for enhanced AD in complex datasets.

5.4.4 Classification of Adversarial Attacks

Recent advances in adversarial defense have introduced various techniques to mitigate the impact of adversarial attacks. Mazda Moayeri and Soheil Feizi [19] propose a method that leverages self-supervised learning for adversarial detection. Their approach, known as SimCat, uses embeddings from a Simple Framework for Contrastive Learning of Visual Representations (SimCLR) encoder to classify and detect various adversarial attacks. This approach aligns with the trend towards embedding-based methods in AD and offers valuable insights into efficient adversarial detection.

5.5 Gaps in Knowledge

While significant progress has been made in the fields of adversarial attack detection and AD, several gaps remain that this study aims to address:

1. **Limited Focus on Adversarial Attacks in AD**: Traditional AD techniques often concentrate on identifying deviations from normal behavior without explicitly considering the impact of adversarial attacks. Existing methods typically classify data into normal or anomalous categories but do not differentiate between anomalies caused by natural deviations and those induced by adversarial manipulations. This limitation can reduce the effectiveness of AD systems in environments where adversarial threats are prevalent.

2. **Integration of Attack Classification**: Few studies have integrated the classification of data based on the type of attack [19] alongside traditional

AD. There is a need for methodologies that not only identify anomalous data but also categorize it by the type of adversarial attack, providing a more detailed understanding of the threats. By classifying data into normal, anomalous, or attacked categories, and further identifying the specific attack techniques used (e.g., FGSM, PGD), models can offer more robust protection against a variety of adversarial strategies.

5.6 Theoretical Framework

Autoencoders in AD: Autoencoders are a type of neural network used primarily for unsupervised learning of data representations. They are structured with an encoder, which compresses the input into a lower-dimensional latent representation, and a decoder, which reconstructs the input from this representation. In AD, autoencoders are trained on normal data, allowing them to effectively reconstruct similar inputs. Anomalies are identified by their higher reconstruction error, as they deviate from the learned patterns of normal data.

Adversarial Attack Detection: Adversarial attacks exploit vulnerabilities in ML models by introducing small, carefully crafted perturbations to input data, resulting in incorrect outputs. Techniques like the FGSM[10] and PGD[17] are commonly used to generate adversarial examples. These methods leverage the model's gradients to create inputs that appear normal to human observers but cause significant errors in the model's predictions. Detecting such attacks involves identifying the patterns and features that these adversarial perturbations introduce.

Vector Reconstruction Error: Traditional approaches to AD using autoencoders focus on scalar reconstruction loss, which measures the overall difference between input and output. This approach, however, may not capture the full extent of deviations, particularly in the context of adversarial attacks. By utilizing vector reconstruction error[30], each element of the reconstruction loss is considered, providing a more detailed analysis of how the input and output differ. This technique offers enhanced sensitivity in distinguishing between normal, anomalous, and adversarially perturbed data, making it a valuable tool for improving AD and adversarial defense.

Integration in This Study: In this research, the theoretical concepts of autoencoders, adversarial attack detection, and vector reconstruction error are integrated to benchmark ML models' ability to detect adversarial attacks. Autoencoders are used to generate reconstruction errors from the dataset, while adversarial attacks are applied using FGSM and PGD methods. The vector reconstruction error approach is tested to differentiate between normal, anomalous, and adversarially attacked data, providing a comprehensive framework for evaluating model performance in the presence of adversarial threats.

5.7 Conclusion

The literature review highlights the growing complexity of adversarial attacks and the corresponding challenges in defending ML models. While significant advancements have been made in understanding and mitigating these threats, the review

also uncovers several critical gaps in the current research. Notably, traditional AD methods often fail to differentiate between natural deviations and those induced by adversarial manipulations. Moreover, the integration of attack classification into AD remains an underexplored area. Addressing these gaps will require the development of more sophisticated detection mechanisms and the exploration of novel approaches to adversarial defense. Future research should focus on enhancing the robustness of ML models, particularly in the context of complex and evolving adversarial strategies.

CHAPTER 6

Conclusion

This research aimed to assess the effectiveness of machine learning models in detecting adversarial attacks, with a particular focus on the dynamics of autoencoder reconstruction loss and the potential of a novel feature engineering technique. The key findings of this study are as follows:

- **Feature Engineering Technique**: The approach of using reconstruction error as a vector did not result in a significant enhancement in the detection performance of the models. Although theoretically promising, the empirical results revealed that this method did not offer additional discriminative power when compared to the traditional scalar representation.
- **Adversarial Attack Detection**: The study emphasized the considerable challenge in detecting PGD attacks, in contrast to the FGSM. The models consistently achieved higher accuracy, recall, precision, and F1 scores for FGSM attacks, while struggling with the more complex and iterative nature of PGD attacks.
- **Model Performance Analysis**: The detailed examination using boxplots and confusion matrices provided a clear understanding of the strengths and limitations of various candidate models. The results highlight the need for more sophisticated detection mechanisms to effectively counter advanced adversarial attacks.

6.1 Key Insights

- **Detection Model Robustness**: The difficulty in detecting PGD attacks underscores the critical need for more robust and advanced adversarial detection techniques. The current models and feature engineering approaches may be insufficient to manage the complexities of sophisticated attack strategies.
- **Future Research Directions**: The findings suggest several promising avenues for future research, including the development of advanced feature engineering techniques, the integration of adversarial training, comprehensive evaluations across diverse datasets and attack types, and the exploration of hybrid detection approaches.

6.2 Practical Implications

- **Application in Practice**: The insights from this study are particularly relevant for practitioners aiming to enhance the resilience of anomaly detection systems against adversarial attacks. These findings can guide the development of more effective defense mechanisms to counter sophisticated adversarial techniques.
- **Research and Development**: The study underscores the ongoing need for research into adversarial detection, advocating for the exploration of innovative methodologies and the refinement of existing approaches to improve detection accuracy and robustness.

In conclusion, while the novel feature engineering technique of using reconstruction error as a vector did not yield significant improvements, this study provides valuable insights into the challenges and potential directions for enhancing adversarial detection. Addressing these challenges will require sustained research efforts and the development of more advanced and robust detection methods to effectively counter sophisticated adversarial attacks.

Bibliography

[1] Battista Biggio et al. 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.

[2] Christopher M Bishop. *Pattern recognition and machine learning*, volume 4. Springer, 2006.

[3] Nicholas Carlini et al. Evaluating and understanding the robustness of adversarial logit pairing. *arXiv preprint arXiv:1909.07414*, 2019.

[4] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57, 2017.

[5] Ankan Dash, Junyi Ye, and Guiling Wang. A review of generative adversarial networks (gans) and its applications in a wide variety of disciplines: From medical to remote sensing. *IEEE Access*, 12:18330–18357, 2024.

[6] Kevin Eykholt et al. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1625–1634, 2018.

[7] Samuel G Finlayson et al. Adversarial attacks on medical machine learning. *Science*, 363(6433):1287–1289, 2019.

[8] Salvador García, Julián Luengo, and Francisco Herrera. *Data Preprocessing in Data Mining*, volume 72. 01 2015.

[9] Ian Goodfellow et al. *Deep learning*. MIT press, 2016.

[10] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples, 2015.

[11] Trevor Hastie et al. *The elements of statistical learning: data mining, inference, and prediction*. Springer, 2009.

[12] Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[13] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[15] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In F. Pereira, C.J. Burges, L. Bottou, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 25. Curran Associates, Inc., 2012.

[16] Yann LeCun, Y. Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436–44, 05 2015.

[17] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks, 2019.

[18] David Miller, Zhen Xiang, and George Kesidis. Adversarial learning targeting deep neural network classification: A comprehensive review of defenses against attacks this article provides a contemporary survey of adversarial learning (al), focused particularly on defenses against attacks on deep neural network classifiers. *Proceedings of the IEEE*, PP:1–1, 02 2020.

[19] Mazda Moayeri and Soheil Feizi. Sample efficient detection and classification of adversarial attacks via self-supervised embeddings, 2021.

[20] Mehryar Mohri et al. *Foundations of machine learning*. MIT press, 2018.

[21] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[22] Akarsh K. Nair, Ebin Deni Raj, and Jayakrushna Sahoo. A robust analysis of adversarial attacks on federated learning environments. *Comput. Stand. Interfaces*, 86:103723, 2023.

[23] Nicolas Papernot et al. Distillation as a defense to adversarial perturbations against deep neural networks. In *2016 IEEE Symposium on Security and Privacy (SP)*, pages 582–597. IEEE, 2016.

[24] David MW Powers. Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*, 2011.

[25] Jiahao Shao, Shijia Geng, Zhaoji Fu, Weilun Xu, Tong Liu, and Shenda Hong. Cardiodefense: Defending against adversarial attack in ecg classification with adversarial distillation training. *Biomedical Signal Processing and Control*, 91:105922, 2024.

[26] David Silver, Aja Huang, Christopher Maddison, Arthur Guez, Laurent Sifre, George Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, Sander Dieleman, Dominik Grewe, John Nham, Nal Kalchbrenner, Ilya Sutskever, Timothy Lillicrap, Madeleine Leach, Koray Kavukcuoglu, Thore Graepel, and Demis Hassabis. Mastering the game of go with deep neural networks and tree search. *Nature*, 529:484–489, 01 2016.

[27] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[28] Christian Szegedy et al. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

[29] Mahbod Tavallaee et al. A detailed analysis of the kdd cup 99 data set. *2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications*, pages 1–6, 2009.

[30] Hasan Torabi, Seyedeh Leili Mirtaheri, and Sergio Greco. Practical autoencoder based anomaly detection by using vector reconstruction error. *Cybersecurity*, 6(1):1, 2023.

[31] Dimitris Tsipras et al. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

[32] Mingfu Xue, Chengxiang Yuan, Heyi Wu, Yushu Zhang, and Weiqiang Liu. Machine learning security: Threats, countermeasures, and evaluations. *IEEE Access*, 8:74720–74742, 2020.