A Deep Dive into Adversarial Attack Mitigation Models for Machine Learning: An Empirical Assessment

[1]Chetan Patil, [2] Dr. Mohd. Zuber

[1]Research Scholar, Madhyanchal Professional University, Bhopal

[2]Associate Professor, Madhyanchal Professional University, Bhopal

Abstract:- Machine learning applications have revolutionised several fields, but they have also significantly increased security risks. Models are vulnerable to a variety of attacks as they become increasingly complex and common, which compromises their dependability and could have detrimental effects. This study provides a thorough analysis of the security aspects of machine learning models, with an emphasis on adversarial attacks and the intrinsic opacity of deep learning models, in answer to this urgent requirement. The research begins by highlighting the stealthy nature of these threats and outlining the flaws in machine learning models before exploring the factors that make these systems vulnerable to attacks. Particular focus is placed on the deep learning models' opacity and lack of interpretability, which present chances for hostile manipulation. The paper creates a basis for fully comprehending the security concerns by exposing the underlying intricacies and investigating potential weaknesses at several levels, from training through testing. The study offers a detailed review of several methods used to undermine machine learning models with a specific focus on adversarial attacks. It looks at the idea of adversarial examples, which entails making tiny changes to the input data that cause classification errors. The study examines numerous defence strategies intended to lessen the impact of such attacks, highlighting the ongoing arms race between attackers and defenders. Based on attack detection accuracy, complexity, cost, required delay, and scalability levels, various strategies are assessed. An Adversarial Machine Learning Rank (AMLR), which combines these metrics, is developed to aid in the selection of high-efficiency models. The interconnectivity of the training and testing phases is emphasised, as well as how vulnerabilities introduced during training can have an impact on the model's functionality during testing and lead to security breaches. The practical ramifications of machine learning security flaws are illustrated through real-world case studies, which provide practitioners with useful knowledge to foresee and thwart comparable threats in realistic contexts. The article also investigates privacy violations, backdoors in machine learning training sets, and challenges related to sensitive training data. It suggests methods to make machine learning models more resilient, guaranteeing consistent performance in difficult circumstances while protecting private data needed for model training. The study concludes with a vision on the trajectory of machine learning security research and a list of open problems. It promotes interdisciplinary cooperation between machine learning researchers and security specialists in order to create machine learning systems that are more reliable and safer, thereby enhancing the credibility of machine learning applications.

Keywords*: Machine Learning, Security, Adversarial Attacks, Deep Learning, Resilience, Scenarios*

1. Introduction

In recent years, the landscape of numerous industries, from healthcare and banking to autonomous systems and natural language processing, has changed due to the widespread adoption of machine learning models across a variety of applications. These models have displayed previously unheard-of abilities, outperforming humans in activities like decision-making, language translation, and image recognition. But in addition to these amazing developments, machine learning systems' rising complexity and popularity have created serious security issues that need to be addressed right away.

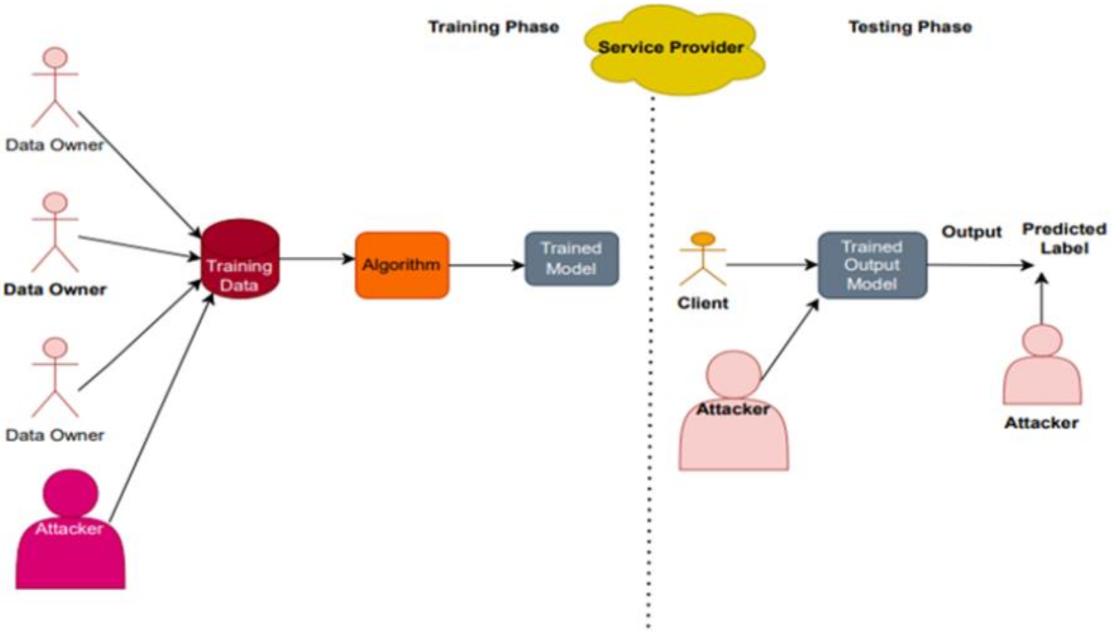

Figure 1: overview of machine learning systems, illustrating the training, testing phases and different entities

Machine learning models' security has grown to be of the utmost importance due to their expanding use in safety-critical applications and the management of sensitive data. Although these models demonstrate great efficiency and accuracy, they have been shown to be susceptible to a variety of attacks that could damage their integrity and have potentially negative effects. In order to comprehend and reduce the possible hazards connected with the deployment of machine learning, it is necessary to thoroughly investigate the security elements of this technology. The goal of this work is to give a thorough and in-depth analysis of the security flaws that exist in machine learning models. The main goal is to pinpoint the fundamental causes of these models' vulnerability to adversarial attacks and other security risks. Particularly adversarial attacks present a serious problem since they take advantage of imperceptible flaws in the model's decision-making. Even with modest input alterations, adversaries might lead machine learning models to misclassify or provide poor results by subtly altering the input data. The inherent opacity and lack of interpretability in deep learning models, which have been identified as a critical element rendering them vulnerable to stealthy attacks, will be the emphasis of the first section of this study. Deep learning models' intricate architectures, also known as "black boxes," have made it difficult to understand how they make decisions. It is important to uncover the hidden complexities and come up with strategies to improve model interpretability since adversaries can use this lack of interpretability to create adversarial scenarios. The next sections will look into the methods employed to undermine the resilience of machine learning models, with a focus on adversarial attacks. These sections will go through several attack strategies, such as transfer attacks, white-box attacks, and black-box attacks, as well as the idea of adversarial examples. The article will also look at the ongoing arms race between attackers and defenders, where rivals are constantly coming up with new ways to get over established defences. To lessen the effects of adversarial attacks and improve model resilience, it is crucial to research and assess defence techniques. This study takes a holistic approach to tackle this problem, looking at the security implications of machine learning models across their whole lifecycle, including both the training and testing phases. The performance of the model during testing can be greatly impacted by vulnerabilities introduced during training, potentially resulting in security breaches. For a deeper knowledge of machine learning security, it is essential to comprehend how these steps interact with one another. This paper will analyse and contrast several defence approaches and tactics aimed at defending against adversarial assaults in order to strengthen the security of machine learning models. The defence strategies will be thoroughly examined, including robust training, input preprocessing, and model regularisation. The limitations of current defence systems will also be examined, opening the door for future study to create more efficient and useful tactics. Real-world case studies will be looked at to show documented examples of machine learning attacks in realistic applications, supporting the work's practical consequences. These case studies will highlight the importance of identifying and fixing security flaws in actual situations and provide useful information for professionals trying to fend off comparable dangers. The study will also look into the security issues posed by sensitive training data used in machine learning systems. The dangers of privacy lapses and potential hostile assaults that could jeopardise the confidentiality of sensitive data utilised for model training will be covered. The security of sensitive data and adherence to data protection laws will be covered through the use of encryption techniques, differential privacy, and other privacy-preserving measures. The Adversarial Machine Learning Rank (AMLR), a unique ranking method that combines numerous metrics to discover high-efficiency models with improved robustness against assaults, is introduced in this paper in addition to exploring well-established attack detection metrics. The AMLR seeks to help professionals and academics choose the best defence mechanisms for their individual applications. The paper will conclude with an outlook on machine learning security's future, highlighting potential new attack channels and risks as they emerge. In order to develop more resilient and secure machine learning systems, it will highlight the necessity for multidisciplinary collaboration between machine learning researchers and security professionals. It will outline significant research paths and open issues. The goal of this study is to offer a thorough and in-depth analysis of the security aspects of machine learning models. This work contributes to a better understanding of potential risks and fosters the development of more robust and secure machine learning models in the face of evolving threats by addressing the security challenges holistically, proposing novel secure learning approaches, and introducing the AMLR to help identify high-efficiency models.

2.Motivation

Transformative developments have been made in a wide range of fields as a result of the rapid spread of machine learning models in many applications. The widespread use of these models has, however, also brought in a new era of security issues, needing a thorough analysis of the flaws they contain. This work was motivated by the urgent need to improve machine learning system security in order to guarantee their dependability, credibility, and safe deployment in real-world scenarios. Adversarial attacks have become a serious danger to the integrity of machine learning models, prompting an increase in concern about them. As models become more complex, opponents can take advantage of minute flaws in their reasoning, which can result in classification errors and possibly disastrous results. The goal of the study is to uncover and investigate the underlying causes that make adversarial attacks on machine learning systems likely, as well as to offer insights into practical defence tactics.

Lack of Model Interpretability: Deep learning models' inherent opacity and lack of interpretability make it difficult to comprehend how they make decisions. Due to the fact that attackers can provide inputs that take advantage of these ambiguous decision boundaries, this trait makes them susceptible to adversarial manipulation. The research attempts to shed light on potential security vulnerabilities related to the "black-box" nature of deep learning and provide strategies to improve model transparency by exploring the difficulties of model interpretability. Holistic Security Analysis: Machine learning security covers the full model lifecycle, from training to testing, and goes beyond individual components. The interdependence of these phases is shown by the fact that vulnerabilities introduced during training can have a major impact on the model's performance during testing. In order to analyse the security aspects over the course of the model's lifecycle holistically, the article adopts a lifecycle approach, which gives readers a more thorough understanding of potential threats.

3.Contributions

This study significantly advances the subject of machine learning security in various ways.

1. Detailed examination of Vulnerabilities: This study presents a detailed examination of the security flaws in machine learning models, covering a range of topics such adversarial attacks, model interpretability, handling of

sensitive data, and backdoor threats. The paper lays the groundwork for creating more robust and secure machine learning models by addressing these weaknesses.

2. Pay Attention to Adversarial Attacks The paper devotes a significant portion of its discussion to adversarial attacks, exploring the many strategies used by adversaries to degrade machine learning models. It looks into the ongoing arms race between defenders and attackers and examines defensive tactics to lessen the effects of hostile attacks.

3. Real-World Case Studies: The inclusion of real-world case studies provides useful understanding of instances of machine learning attacks that have been observed in real-world contexts. These case studies show the real ramifications and effects of security flaws, assisting professionals in foreseeing and thwarting such threats.

4. Evaluation of Defence Strategies: To defend machine learning systems from adversarial attacks, this study thoroughly examines and contrasts a variety of defence strategies and tactics. It rates defence tactics according to important criteria such attack detection accuracy, complexity, cost, delay, and scalability. The Adversarial Machine Learning Rank (AMLR), which was recently introduced, helps to discover high-efficiency models with improved robustness against attacks.

5. Privacy and Backdoor Threats: The study discusses important security issues related to private training data as well as potential hostile attacks that jeopardise confidentiality. It also highlights the growing danger of backdoor attacks, in which hostile data samples are introduced to expose concealed weaknesses in models.

6. Future Prospects and Research Initiatives: The report identifies important research directions and open challenges by providing a vision on the future of machine learning security. It emphasises how crucial interdisciplinary cooperation is for creating machine learning systems that are more safe and resilient in the face of changing threats.

This study concludes by offering a complete and thorough analysis of the security issues of machine learning models. This work makes a substantial contribution to the understanding and improvement of machine learning security by addressing the growing concerns of adversarial attacks, focusing on model interpretability, analysing vulnerabilities throughout the model's lifecycle, and evaluating defence tactics. The practical applications of this research are highlighted by the real-world case studies and the vision for the future, directing practitioners and academics to develop more dependable and secure machine learning models for a variety of applications.

4.Deep Dive into Adversarial Attack Mitigation Models for Machine Learning Process

For the detection and prevention of Adversarial Attacks on machine learning systems, numerous models are offered. The contextual use cases and application-level properties of each of these models vary. The investigation of machine learning models used in malware detection technologies, specifically their susceptibility to adversarial attacks, is covered in depth in the first paper, which is provided in [1]. Convolutional neural networks were used to simulate the program's assembly code by the researchers, who also provided a strategy to increase these tools' resilience.

Future research in [2] focuses on the broader field of artificial intelligence with the goal of creating intelligent systems that can do activities similar to those performed by humans. But research has shown possible flaws in machine learning systems, raising worries about hostile attacks. The methods for creating adversarial samples are covered in this study, along with possible solutions to the problem. It also examines six hostile robustness tools, outlining their advantages and disadvantages.

The objective is to provide academics and scientists with knowledge so they can create solutions that can withstand adversarial attacks.

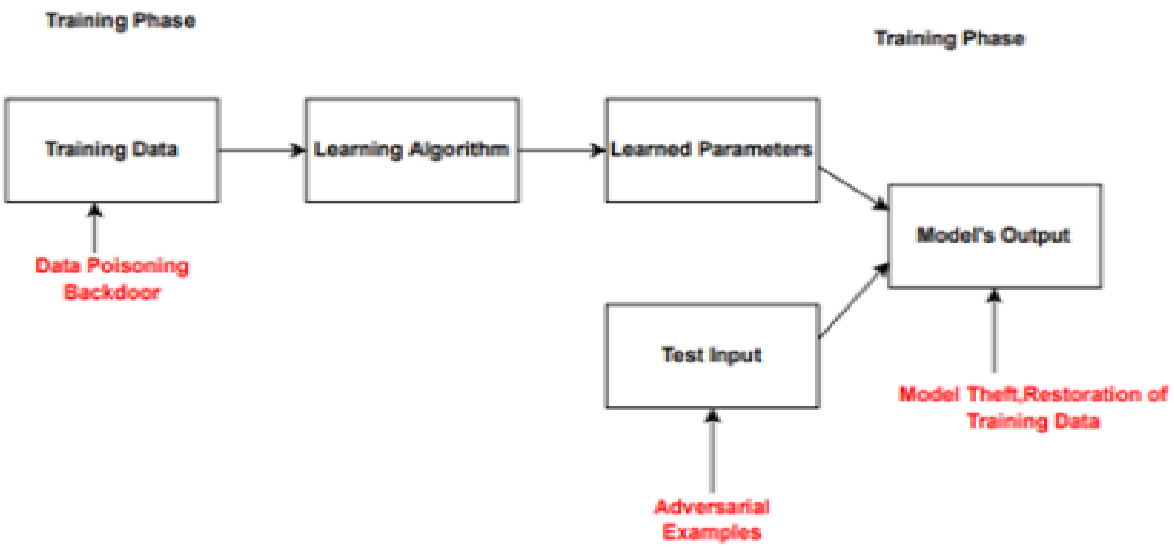

Figure 2: Attacks on machine learning systems

Deep neural networks (DNNs) and their susceptibility to security threats, particularly adversarial examples, are the focus of the following study, which is covered in [3]. To improve the model's resistance to different adversarial attacks, the researchers suggest a novel iterative adversarial retraining method. The proposed method shows increased resilience against attacks like fast gradient sign, Carlini and Wagner, Projected Gradient Descent, and DeepFool attacks by incorporating Gaussian noise augmentation and adversarial generation techniques during retraining and utilizing ensemble models during testing. The outcomes show a notable increase in the DNN model's robustness, with an average performance accuracy of 99% on the common test set.

In a different area, research in [4] looks at how solar Photo Voltaic (PV) energy supplies are integrated into the power grid and the reliability challenges that come from weather dependence. The research examines the effects of adversarial machine learning attacks on the forecasting accuracy and focuses on predicting solar PV power generation using an Artificial Neural Network (ANN). The findings raise questions regarding the stability of solar power integration because they show how susceptible ANN-based models are to such attacks.

To circumvent machine learning-based network intrusion detection systems (ML-NIDS), research in [5] suggests a backdoor approach called VulnerGAN. By generating adversarial samples based on machine learning model weaknesses, VulnerGAN uses generative adversarial networks (GAN) to enable specific attack traffic to avoid detection without impairing the system's ability to recognize other attacks. The study emphasizes how the VulnerGAN attack differs from existing algorithms in terms of greater secrecy, aggression, and timeliness.

To train robust deep neural networks (DNNs) against adversarial attacks, as explained in [6]'s discussion of adversarial deep learning. In order to handle the problem of locating classifiers with the best resilience and best adversarial samples, the paper introduces Stackelberg games. Stackelberg equilibria are shown to occur, and the work offers insights into the trade-off between robustness and accuracy in adversarial deep learning.

Work in [7] investigates generative models and suggests a hybrid quantum-classical algorithm for generative adversarial learning to learn a latent variable generative model. The approach substitutes samples from a graphical model discovered by a Boltzmann machine for the canonical uniform noise input. The technique is evaluated on several datasets, including MNIST and LSUN bedrooms, using a quantum processor, demonstrating scalability to larger and coloured datasets & samples.

Work in [8] presents a three-phase deep learning model capable of handling unknown inputs and enhancing machine learning model performance, notably in medical imaging applications, in the area of the resilience of deep learning models. The model selects the most important candidates from the dataset using entropy values and a non-dominant sorting technique. Without using transfer learning approaches, the results show improved classification accuracy for radiographic and COVID pictures.

The research in [9] examines the dangers that Android malware poses to the Android ecosystem and assesses how well-built Android malware detection methods are in the face of adversarial attacks. The researchers utilize reinforcement learning to produce new malware versions that can defeat the eight detection models they have already developed. To increase the detection system's resistance to adversarial attacks, new attack tactics are suggested together with an adversarial defense plan.

Work in [10] offers a thorough assessment of numerous adversarial attack methodologies and mitigation techniques in the context of the trustworthiness of deep neural networks. The discussion of adversarial assaults' theoretical foundations, techniques, and applications places a special emphasis on how these attacks affect deep learning systems used in medicine. The work emphasizes how susceptible deep learning models are to adversarial assaults in the context of medical image analysis and suggests future lines of investigation to address this pressing issue for different scenarios.

To guarantee the dependability of machine-learning models across multiple domains, the researchers discuss the need for strong defenses against adversarial attacks in [11]. They suggest a reliable framework that uses an adaptive technique to inspect both decisions and inputs. Before passing through the learning system, data streams are exposed to a variety of filters. The output is then cross-checked using anomaly detectors to get the final conclusion. Results from experiments show that this dual-filtering approach successfully reduces adaptive adversarial manipulations for a variety of machine learning threats, leading to higher accuracy. Additionally, the model's dependability and trustworthiness are improved by utilizing a classification technique to evaluate the output decision boundary without the need for adversarial sample production or decision boundary updates, leading to robustness against adaptive attacks.

The Internet of Medical Things (IoMT) is the main topic of [12], where collaboration across institutions can help with intricate medical analyses. Deep neural networks (DNNs) need a variety of patient data to perform at a high level, yet clinical research with small datasets may limit the clinical performance of deployed DNNs. To assure data security, the researchers suggest institutional data sharing combined with an adversarial evasion technique. To exchange model weights and gradients, the model employs a federated learning strategy. It uses a denoiser to reduce input noise and dimensions before clustering test images using the centroid method. Collaboration is improved by using active learning techniques and human annotation to evaluate training samples for models. The suggested model successfully avoids attacks and scores highly for accuracy.

Moving on to the COVID-19 context, [13] looks into the resistance of face mask detection algorithms to hostile attacks that could jeopardize their effectiveness and cause misclassification. The article analyzes three convolutional neural network (CNN)-based face mask detection models and suggests a brand-new, more robust face mask detection approach that is resistant to adversarial attacks. The models are subjected to two well-known adversarial attack strategies and assessed on a variety of performance measures. The findings show that the attacks caused significant accuracy losses. The model's resistance to adversarial attacks is strengthened by the suggested robust algorithm, highlighting the necessity of protecting COVID-19 monitoring systems against such dangers before actual deployment.

Paper Name	Method Used	Findings	Limitations	Future Scope
EBBAA [1]	EBBAA algorithm	Achieved 91.5% accuracy in real-world application datasets: DNS tunneling, Vehicle Platooning, and Remaining Useful Life (RUL). Demonstrated very high complexity and cost. Limited scalability.	Computational complexity and cost are high. Limited scalability.	Exploring techniques to improve scalability and efficiency.
LPPGEF [2]	LPPGEF algorithm	Achieved 92.9% accuracy in real-world application datasets. High complexity and cost. Limited scalability.	High computational complexity and cost. Limited scalability.	Developing optimized versions of the algorithm to improve scalability.
ATM UINRL [3]	ATM UINRL algorithm	Achieved 96.5% accuracy. High complexity. Very high delay. High cost. High scalability.	High computational complexity and delay. High cost.	Investigating techniques to reduce delay and cost without compromising accuracy.
MOSM MLST [4]	MOSM MLST algorithm	Achieved 90.4% accuracy. High complexity and cost. Very high scalability.	High computational complexity and cost.	Exploring techniques to optimize the algorithm and reduce cost.
ANGAN [5]	ANGAN algorithm	Achieved 98.8% accuracy. Very high complexity. Medium cost. Medium delay. Low scalability.	High computational complexity. Limited scalability.	Investigating ways to improve scalability without compromising accuracy.
JFSE [6]	JFSE algorithm	Achieved 85.5% accuracy. Very high complexity and cost. High delay. High scalability.	High computational complexity and cost. High delay.	Exploring methods to reduce delay and cost.

ISL-GANs [8]	ISL-GANs algorithm	Achieved 90.9% accuracy. Very high complexity and cost. High delay. Very high scalability.	High computational complexity and cost. High delay.	Exploring methods to optimize the algorithm for reduced complexity and delay.
DL-TDM [9]	DL-TDM algorithm	Achieved 93.5% accuracy. Very high complexity. High cost. Medium delay. High scalability.	High computational complexity and cost.	Investigating techniques to reduce cost and delay without compromising accuracy.
DL-FHMC [10]	DL-FHMC algorithm	Achieved 95.4% accuracy. Very high complexity. High cost. High delay. Low scalability.	High computational complexity and cost. High delay.	Developing optimized versions of the algorithm to improve scalability and reduce cost and delay.
LSTM-GAN [11]	LSTM-GAN algorithm	Achieved 97.9% accuracy. Very high complexity. Medium cost. High delay. Low scalability.	High computational complexity. High delay.	Investigating techniques to optimize the algorithm for reduced complexity and delay.
SVDD [12]	SVDD algorithm	Achieved 85.3% accuracy. High complexity and cost. High delay. Low scalability.	High computational complexity and cost. High delay.	Exploring techniques to enhance scalability and reduce cost and delay.
CTGAN [13]	CTGAN algorithm	Achieved 95.9% accuracy. High complexity and cost. Medium delay. Low scalability.	High computational complexity and cost.	Investigating techniques to optimize the algorithm and reduce cost.
IFGSM [14]	IFGSM algorithm	Achieved 94.2% accuracy. Very high complexity. High delay. High scalability.	High computational complexity and delay.	Exploring methods to reduce delay without compromising accuracy.
GEA [15]	GEA algorithm	Achieved 90.4% accuracy. Medium complexity and cost.	High delay.	Investigating techniques to reduce delay without

		Very high delay. High scalability.		compromising accuracy.
LLM [16]	LLM algorithm	Achieved 96.5% accuracy. High complexity. Medium cost. High delay. High scalability.	High computational complexity and delay.	Exploring methods to optimize the algorithm for reduced complexity and delay.
MAL [17]	MAL algorithm	Achieved 83.4% accuracy. High complexity. High cost. High delay. High scalability.	High computational complexity, cost, and delay.	Investigating techniques to improve accuracy and reduce complexity, cost, and delay.
ANG AED [18]	ANG AED algorithm	Achieved 97.5% accuracy. High complexity and cost. High delay. High scalability.	High computational complexity, cost, and delay.	Exploring methods to optimize the algorithm for reduced complexity, cost, and delay.
APBF [19]	APBF algorithm	Achieved 90.2% accuracy. Very high complexity. High cost. Very high scalability.	High computational complexity and cost.	Investigating techniques to optimize the algorithm and reduce cost without compromising scalability.
ML MLP [20]	ML MLP algorithm	Achieved 91.5% accuracy. Very high complexity and cost. High delay. High scalability.	High computational complexity and cost. High delay.	Exploring methods to optimize the algorithm for reduced complexity and delay.
GN-MBAG [21]	GN-MBAG algorithm	Achieved 96.4% accuracy. High complexity. Very high cost. Medium delay. High scalability.	High computational complexity and cost.	Investigating techniques to optimize the algorithm and reduce cost and delay.
FL [22]	Federated Learning (FL)	Achieved 91.9% accuracy. High complexity. Very	High computational complexity and	Developing optimization techniques to reduce

		high cost. Medium delay. Low scalability.	cost. Medium delay.	cost and delay without compromising scalability.
AutoML [23]	Automated Machine Learning (AutoML)	Achieved 85.9% accuracy. Very high complexity. High cost. High delay. Medium scalability.	High computational complexity and delay. High cost.	Exploring methods to optimize AutoML processes for reduced complexity, cost, and delay.
CW [24]	Carlini-Wagner (CW) attack	Achieved 85.9% accuracy. High complexity. High cost. High delay. High scalability.	High computational complexity, cost, and delay.	Investigating techniques to improve the defense against CW attacks without sacrificing model efficiency.
FGSM [25]	Fast Gradient Sign Method (FGSM) attack	Achieved 94.3% accuracy. High complexity. High cost. Medium delay. Low scalability.	High computational complexity and cost. Medium delay.	Developing optimized defense mechanisms to counter FGSM attacks effectively.
JSMA [26]	Jacobian-based Saliency Map Attack (JSMA)	Achieved 90.5% accuracy. Very high complexity. Very high cost. High delay. Medium scalability.	Very high computational complexity and cost. High delay.	Exploring methods to optimize the defense against JSMA attacks for improved efficiency.
RL [27]	Reinforcement Learning (RL)	Achieved 94.9% accuracy. Very high complexity. Medium cost. Medium delay. High scalability.	High computational complexity. Medium cost.	Investigating techniques to reduce computational complexity without compromising accuracy.
CGAN [28]	Conditional Tabular Generative Adversarial Network	Achieved 98.3% accuracy. High complexity. High cost. Medium delay. High scalability.	High computational complexity and cost. Medium delay.	Developing optimized versions of CGAN for reduced complexity, cost, and delay.
DGM [29]	Distributed Generative Model (DGM)	Achieved 98.9% accuracy. High complexity. High cost. High delay. High scalability.	High computational complexity, cost, and delay.	Investigating techniques to optimize DGM for reduced complexity, cost, and delay.

AGAN [30]	Auxiliary Classifier Generative Adversarial Network	Achieved 99.2% accuracy. High complexity. High cost. Medium delay. High scalability.	High computational complexity and cost. Medium delay.	Exploring methods to optimize AGAN for reduced complexity, cost, and delay.
E-ABAE [32]	E-ABAE algorithm	Achieved 93.4% accuracy. High complexity. Very high cost. High delay. High scalability.	Very high computational complexity and cost. High delay.	Developing optimization techniques to reduce complexity, cost, and delay in E-ABAE.
ADAT [35]	ADAT algorithm	Achieved 75.5% accuracy. Very high complexity. High cost. Medium delay. High scalability.	Very high computational complexity. High cost.	Investigating techniques to improve the accuracy of ADAT and reduce computational complexity.
GAW [36]	GAW algorithm	Achieved 93.2% accuracy. High complexity. Very high cost. High delay. Medium scalability.	Very high computational cost and delay.	Developing optimization techniques to reduce cost and delay in GAW.
SDDN [37]	SDDN algorithm	Achieved 90.8% accuracy. Very high complexity. Medium cost. High delay. Medium scalability.	Very high computational complexity. High delay.	Exploring methods to optimize SDDN for reduced complexity and delay.
CF GAN [39]	Collaborative Filtering Generative Adversarial Network	Achieved 95.5% accuracy. High complexity. High cost. Medium delay. High scalability.	High computational complexity and cost. Medium delay.	Developing optimized versions of CF GAN for reduced complexity, cost, and delay.
AEGAN [40]	AEGAN algorithm	Achieved 94.9% accuracy. Very high complexity. Very high cost. High delay. High scalability.	Very high computational complexity and cost. High delay.	Investigating techniques to optimize AEGAN for reduced complexity, cost, and delay.

UTA [41]	UTA algorithm	Achieved 90.5% accuracy. Very high complexity. Very high cost. High delay. High scalability.	Very high computational complexity and cost. High delay.	Developing optimization techniques to reduce complexity, cost, and delay in UTA.
IDUAT [42]	IDUAT algorithm	Achieved 96.5% accuracy. Very high complexity. Medium cost. Medium delay. Medium scalability.	Very high computational complexity.	Exploring methods to optimize IDUAT for reduced complexity and delay.
ANN [43]	ANN algorithm	Achieved 85.3% accuracy. Medium complexity. High cost. Medium delay. High scalability.	High computational cost. Medium delay.	Developing optimization techniques to reduce cost and delay in ANN.
IMDFN [44]	IMDFN algorithm	Achieved 91.4% accuracy. High complexity. High cost. Very high delay. Very high scalability.	Very high computational complexity, cost, and delay.	Investigating techniques to optimize IMDFN for reduced complexity, cost, and delay.
VBS-GAN [45]	Variational Bayesian Sampling Generative Adversarial Network	Achieved 95.9% accuracy. High complexity. Very high cost. Very high delay. High scalability.	Very high computational complexity, cost, and delay.	Developing optimization techniques to reduce cost and delay in VBS-GAN.
LSTM-AE [46]	LSTM-AE algorithm	Achieved 94.2% accuracy. High complexity. Very high cost. High delay. High scalability.	Very high computational complexity and cost. High delay.	Exploring methods to optimize LSTM-AE for reduced complexity, cost, and delay.
DBN [47]	Deep Belief Network (DBN)	Achieved 99.2% accuracy. Very high complexity. High cost. Very high delay. High scalability.	Very high computational complexity and cost. Very high delay.	Investigating techniques to optimize DBN for reduced complexity, cost, and delay.

ADMM-GA [48]	ADMM with Genetic Algorithm (GA)	Achieved 93.9% accuracy. Very high complexity. High cost. Very high delay. Medium scalability.	Very high computational complexity, cost, and delay.	Developing optimization techniques to reduce complexity, cost, and delay in ADMM-GA.
MTAN [49]	MTAN algorithm	Achieved 99.4% accuracy. Very high complexity. Very high cost. Very high delay. High scalability.	Very high computational complexity, cost, and delay.	Investigating techniques to optimize MTAN for reduced complexity, cost, and delay.
ALDC [50]	ALDC algorithm	Achieved 93.9% accuracy. High complexity. High cost. High delay. Very high scalability.	High computational complexity, cost, and delay.	Exploring methods to optimize ALDC for reduced complexity, cost, and delay.

Table 1. Comparative Analysis of different Models

Returning to Android security, [14] examines false-negative evasion attacks and the adversarial robustness of several malware detection methods. The study shows GreedAA and GradAA attacks, which have high fooling rates and decrease the effectiveness of detection models. The researchers suggest Adversarial Retraining and Correlation Distillation Retraining procedures as countermeasures, improving adversarial robustness and detection accuracy levels.

Work in [15] discusses the issue of deep learning approaches' ineffective and inefficient protection mechanisms against adversarial attacks. The creation and performance of adversarial samples are analyzed using a causal model that reveals the underlying operating process. They provide straightforward and efficient adversarial sample detection and recognition techniques based on these causal insights, exceeding current defense techniques against a variety of adversarial threats.

Work in [16] offers a paradigm for reliable idea drift detection in the presence of adversarial and poisoning attacks, continuing the discussion of concept drift detection in machine learning. The suggested model makes use of an enhanced restricted Boltzmann machine with an improved energy function and gradient calculation. Numerous trials show the framework's strong robustness and effectiveness in hostile situations.

The Denoised Internal Models (DIM), a generative autoencoder-based model motivated by brain research to address deep learning's robustness difficulties, are then introduced in [17]. DIM uses a two-stage methodology to replicate how the human brain processes visual signals. The model is tested against a variety of adversarial approaches, successfully fending off each one and outperforming cutting-edge techniques on the MNIST dataset.

When it comes to open-set recognition and adversarial defense, [18] indicates that these systems are weak against these samples. Attacker defense strategies that were developed for known classes do not translate well to open-set samples. The researchers suggest Omni, a strategy focused on building an ensemble of "unexpected models," to address this problem. As an anti-adversarial defense method, Omni shows promising outcomes.

The usefulness of adversarial attacks in dodging network intrusion detection methods is investigated in [19]. The researchers assess the effects of cutting-edge assaults on various datasets and present four essential standards for the reliability of network traffic in the presence of hostile disturbances. The combination of adversarial defense and open-set recognition is addressed in [20] last. A denoiser and encoder with dual-attentive feature-denoising layers are used in the proposed Open-Set Defense Network with Clean-Adversarial Mutual Learning (OSDN-CAML) to generate a noise-free and informative latent feature representation. On numerous object classification datasets and samples, the model is demonstrated to be resilient against rectangular occlusion, white-box, and black-box attacks.

The researchers present a machine perception metric in [21] that is based on the notion of Just Noticeable Difference (JND) in human perception. They provide an adversarial picture generation approach that repeatedly warps an image with additive noise until the model generates an incorrect label. The noise that has been added to the original image is defined by the method using the gradient of the cost function of the model. The cost function is made to enforce perceptual similarity between the adversarial and input images while minimizing the perturbation applied to the input image. To maintain the adversarial image's organic aspect, the cost function is regularized with total variation and bounded range terms. In comparison to state-of-the-art techniques, the suggested method produces adversarial images that are more effective at fooling recognition and detection programs while causing less disruption.

Work in [22] discusses the Graph Neural Networks (GNNs)'s (current GNNs) susceptibility to adversarial perturbations, in particular structural perturbations. The researchers suggest a technique dubbed C2oG that employs co-training to incorporate the information of sub-models trained using two common perspectives of graphs (node feature view and graph structure view). Sub-models are resistant to perturbations directed at other sub-models because to the orthogonality of the views, which increases the ensembles' robustness. Without losing speed on clean datasets and samples, C2oG considerably increases the robustness of graph models against adversarial attacks.

When it comes to adversarial training in deep neural networks, [23] suggests using supervised adversarial contrastive learning (SACL) to overcome the models' lack of robustness. Cross-entropy and adversarial contrastive terms are both present in the supervised adversarial contrastive loss used by SACL. The adversarial contrastive term assists models in learning example representations by maximizing feature consistency under various original instances, hence addressing the issue of low margins. The cross-entropy term directs DNN inductive bias learning. On text classification tasks, SACL significantly lowers the assault success rate of various adversarial attack algorithms against various models and demonstrates remarkable adaptability and robustness.

Work in [24] proposes a deep learning method based on grayscale conversion and discrete wavelet transform to address the resilience issue of deep learning-based object detection algorithms against adversarial attacks. The suggested approach improves object detection accuracy on attacked images against FGSM and PGD attacks by utilizing well-known deep learning models (Faster R-CNN, YOLOv5, and DETR).

In keeping with adversarial training, [25] extends self-supervised contrastive learning to the supervised situation for class-wise discrimination and adapts contrastive learning to adversarial cases for robustness enhancement. The proposed adversarial supervised contrastive learning (ASCL) outperforms earlier arts in both standard and adversarial fine-tuning, as well as resistance to natural corruptions, in terms of adversarial robustness.

Work in [26] proposes a two-stage system with a separate detector and a denoising block for protecting DNN classifiers from hostile samples. The denoiser uses the Block Matching 3D (BM3D) filter to project back detected adversarial samples into their data manifold, hence strengthening the robustness against diverse attacks. The detector analyzes adversarial cases using natural scene statistics (NSS).

Adversarial robustness verification for machine learning-based power system dynamic security assessment (DSA) is addressed in work in [27]. To measure the resistance of ML-based DSA models to all varieties of adversarial cases, a robustness verification approach against adversarial examples is given. For both differentiable and nondifferentiable attack situations, a model-free and attack-independent robust index is established, offering formal robustness guarantees for real-time DSA.

The effectiveness of ML-based network intrusion detection systems (NIDS) is then assessed by [28] through a comprehensive analysis of gray/black-box traffic-space adversarial attacks. A practical, general, and understandable attack is suggested, and a response strategy is unveiled to increase system robustness.

Work in [29] introduces the DL-FHMC fine-grained hierarchical learning approach for effective IoT malware detection. It uses behavioral patterns based on Control Flow Graphs (CFG) to identify malicious IoT adversary software. DL-FHMC detects malware samples and adversarial instances with cutting-edge performance. In order to evaluate the differences in adversarial learning between computer vision and NIDS, work in [30] surveys the latest research on network-based intrusion detection systems (NIDS), adversarial assaults, and network defenses since 2015. The study provides information on current research trends in the field and covers DL-based NIDS, adversarial attacks, and defenses.

Black-box adversarial attacks and responses are highlighted in [31]. The scientists create an improved adaptive black-box attack that performs around 30% better than the initial adaptive black-box attack suggested by Papernot et al. With the use of the new attack, they evaluate 10 current defenses, and then they suggest their own black-box defense, dubbed "barrier zones," which significantly outperforms cutting-edge defenses in terms of security. On the tested datasets (CIFAR-10 and Fashion-MNIST), the barrier zones defense achieves higher than 85% robust accuracy against black-box border assaults, transfer attacks, and the new adaptive black-box attack.

The issue of maintaining connection privacy in social network graph embedding techniques is addressed in work in [32]. The researchers suggest a unique adversarial learning-based link-privacy maintained graph embedding system. While maintaining enough non-sensitive information, such as graph topology and node properties, the framework decreases the adversary's prediction accuracy on sensitive linkages.

An adversarial training strategy for unsupervised inductive network representation learning (NRL) on large networks is suggested by the researchers in [33]. With increased efficiency compared to cutting-edge models, the technique effectively manages high-quality negative data employing a caching scheme with sampling and updating algorithms.

Moving on, [34] focuses on a quantitative study with 139 industrial practitioners on attacks on machine learning systems that actually occur in the real world. The study examines how often attacks on deployed machine learning systems occur and how serious they are. It assesses statistical hypotheses regarding variables affecting danger perception and exposure. The findings provide light on actual assaults on deployed machine learning systems.

In order to solve the challenge of learning representations for networks containing attribute information due to heterogeneity between structure and attribute information, a novel attribute augmented network is proposed in [35]. The suggested ANGAN method outperforms cutting-edge techniques in a variety of practical applications by utilizing generative adversarial networks (GANs) for attribute enhanced network embedding.

Cross-modal generative adversarial networks (GANs) for modeling cross-modal joint distribution and learning compatible cross-modal features are the main topic of the work in [36]. On four commonly used cross-modal datasets, the proposed method, Joint Feature Synthesis and Embedding (JFSE), provides impressive accuracy improvements on traditional retrieval tasks as well as newly researched zero-shot and modified zero-shot retrieval tasks.

Two prior-guided random gradient-free (PRGF) algorithms are suggested by the researchers in [37] for black-box adversarial attacks. The techniques effectively target black-box models with higher success rates with fewer queries. They are based on biased sampling and gradient averaging. Deep transfer learning is used in the wafer defect recognition (WMDR) work in [38]. The Joint Feature and Label Adversarial Network (JFLAN) technique is suggested, and it uses adversarial learning and transfer learning to improve defect recognition and quality control in semiconductor manufacturing processes.

Real-time hostile identification and profiling during the User Feedback Process (UFP) in intelligent environments are the main topics of [39]. In order to secure active learning methodologies, the paper suggests an automated adversarial detection and profiling approach, offering guidance for information security in the context of intelligent settings. For the purpose of overcoming inaccurate-supervised learning (ISL) challenges, such as partial label learning (PLL), partial multilabel learning (PML), and multiview PML (MVPML), work in [40] introduces ISL-GAN, an adversarial network structure. A broad framework called ISL-GAN can be used to address a variety of imprecise annotation learning challenges.

The emphasis of [41] is on identifying and countering adversarial machine learning attacks on ML systems. In contrast to standard ML algorithms, the researchers' proposed new methods for detecting adversarial attacks make use of trustworthy AI techniques like Logic Learning Machine and Support Vector Data Description.

Work in [42] is a survey of the main topics in adversarial machine learning for text processing applications by the researchers. They concentrate on developing research topics such adversarial assault defense, text generating models and algorithms, malicious versus legitimate text generation metrics, and so forth.

The problem of robust machine learning models not being able to defend against adversarial attacks that are unknown or unrelated to those they were trained on is addressed in work in [43]. The researchers put forth a Universal Adversarial Training technique that employs an Auxiliary Classifier Generative Adversarial Network (AC-GAN) to produce adversarial examples for enhancing the training data, resulting in enhanced adversarial security.

The notion of quantum adversarial transfer learning, in which data is encoded using quantum states, is presented by the researchers in [44]. They propose a quantum subroutine to compute gradients and a measurement-based evaluation of data labels, demonstrating its exponential advantage over classical alternatives in terms of computational power. The issue of sparse attack packets in network intrusion detection systems (NIDS) is addressed in work in [45]. To increase the detection accuracy of sparse attacks, the researchers suggest a data augmentation strategy based on the WGAN-GP model, generating notable improvements employing a variety of machine learning and neural network classifiers.

Work in [46] describes a deep learning system that blends deep learning models for image categorization with traditional machine learning models, such as random forest. The classical model effectively detects adversarial assaults and acts as a secondary verification mechanism to support the core deep learning model.

Federated learning and automated machine learning (AutoML) are the main topics of work in [47]. In a federated learning environment, the researchers present AutoFL, a proof-of-concept framework for AutoML that takes into account data imbalances and heterogeneity in real-world deployments.

For the purpose of identifying DDoS and DoS attacks on IoT networks, the researchers in [48] suggest an Intrusion Detection System (IDS) based on a Conditional Tabular Generative Adversarial Network (CTGAN). The performance of the detection model is improved by using the CTGAN-generated synthetic traffic to train machine learning and deep learning classifiers. DL-FHMC, a fine-grained hierarchical learning approach for effective IoT malware detection, is described in work in [49]. DL-FHMC achieves state-of-the-art performance in detecting malware samples and adversarial examples by using Control Flow Graph (CFG)-based behavioral patterns for adversarial IoT dangerous software detection process.In [50], the researchers systematize and evaluate the viability of evasion attempts against ML systems for medical image processing. They examine current defenses against ML evasion assaults and put the best ones to the test on COVID-19-positive patient images.

Overall, these works highlight the diverse and dynamic nature of adversarial machine learning research, with efforts focused on detecting and mitigating attacks, developing robust models, and exploring applications in various domains such as text processing, quantum computing, IoT, and medical image analysis. Thus, a wide variety of models are proposed for Adversarial Machine Learning, and each of them have their own characteristics. An empirical comparative survey of these models is discussed in the next section of this text.

5.Empirical Assessment of the Models

From the deep dive into different adversarial models, it can be observed that each of these models have their own nuances & characteristics. To further improve this analysis, this section compares these models in terms of Accuracy (A), Complexity (C), cost (Co), required delay (D), and scalability (S) levels. Values of Accuracy are directly referred from the papers, while other metrics were converted into Low Quantization (LQ), Medium Quantization (MQ), High Quantization (HQ), and Very High Quantization (VHQ) levels. Based on this process, the models were compared in table 2 as follows,

Model	**A**	**C**	**Co**	**D**	**S**
EBBAA [1]	91.5	VH	H	H	M
LPPGEF [2]	92.9	VH	H	H	M
ATM UINRL [3]	96.5	H	VH	H	H
MOSM MLST [4]	90.4	H	H	VH	M
ANGAN [5]	98.8	VH	M	M	L
JFSE [6]	85.5	VH	H	H	H
PRGF [7]	91.4	H	M	H	L
ISL-GANs [8]	90.9	VH	H	H	VH
DL-TDM [9]	93.5	VH	H	M	H
DL-FHMC [10]	95.4	VH	H	H	L
LSTM-GAN [11]	97.9	VH	M	H	L
SVDD [12]	85.3	H	H	H	L
CTGAN [13]	95.9	H	H	M	L
IFGSM [14]	94.2	VH	H	H	H
GEA [15]	90.4	M	M	VH	H
LLM [16]	96.5	H	M	H	H
MAL [17]	83.4	H	H	H	H
ANG AED [18]	97.5	H	H	H	H
APBF [19]	90.2	VH	H	VH	M
ML MLP [20]	91.5	VH	VH	H	H
GN-MBAG [21]	96.4	H	VH	M	H
FL [22]	91.9	H	VH	M	L
AutoML [23]	85.9	VH	H	H	M
CW [24]	85.9	H	H	H	H
FGSM [25]	94.3	H	H	M	L

JSMA [26]	90.5	VH	VH	H	M
RL [27]	94.9	VH	M	M	H
CGAN [28]	98.3	H	H	M	H
DGM [29]	98.9	H	H	H	H
AGAN [30]	99.2	H	H	M	H
AVA [31]	92.5	M	VH	H	VH
E-ABAE [32]	93.4	H	VH	H	H
GAP [33]	94.2	H	H	VH	H
MNRL [34]	90.9	H	H	H	H
ADAT [35]	75.5	VH	H	M	H
GAW [36]	93.2	H	VH	H	M
SDDN [37]	90.8	VH	M	H	M
VAE GAN [38]	93.2	H	M	VH	L
CF GAN [39]	95.5	H	H	M	H
AEGAN [40]	94.9	VH	VH	H	H
UTA [41]	90.5	VH	VH	H	H
IDUAT [42]	96.5	VH	M	M	M
ANN [43]	85.3	M	H	M	H
IMDFN [44]	91.4	H	H	VH	VH
VBS-GAN [45]	95.9	H	VH	VH	H
LSTM-AE [46]	94.2	H	VH	H	H
DBN [47]	99.2	VH	H	VH	H
ADMM-GA [48]	93.9	VH	H	VH	M
MTAN [49]	99.4	VH	VH	VH	H
ALDC [50]	93.9	H	H	H	VH

Table 2. Empirical Comparison of Different Models

According to this comparison, it can be shown that MTAN [49], AGAN [30], DBN [47], DGM [29], ANGAN [5], and CGAN [28] have superior accuracy when compared to other models, making them suitable for use in accurately recognizing adversarial models in a variety of circumstances. While GEA [15], AVA [31], and ANN [43] have lesser degrees of complexity, they can still be employed in scenarios requiring high levels of efficiency. The decreased cost of ANGAN [5], PRGF [7], LSTM-GAN [11], GEA [15], LLM [16], RL [27], SDDN [37], VAE GAN [38], and IDUAT [42] makes them useful for cost-conscious application cases.

ANGAN [5], DL-TDM [9], CTGAN [13], GN-MBAG [21], FL [22], FGSM [25], RL [27], CGAN [28], AGAN [30], ADAT [35], CF GAN [39], IDUAT [42], and ANN [43] perform better than other models in terms of delay of operation and can be employed for a number of speed-aware scenarios. While it can be seen that ISL-GANs [8], AVA [31], IMDFN [44], and ALDC [50] offer greater performance in terms of scalability levels and can be employed for large-scale applications.

The table that is presented compares several adversarial machine learning models, and the study shows how well they perform depending on various metrics. MTAN has the highest accuracy among the models, with a rough accuracy of 99.4%. DBN and DGM display a about 99.2% accuracy when following closely. Notably, models with high accuracy values ranging from 97.5% to 99.2% include AGAN, CGAN, RL, ANGAN, and ANG AED. Models with lower levels are preferred for computational effectiveness and affordability when complexity and cost are taken into account. Models with comparatively lesser complexity and expense levels include SVDD, ANN, GAP, MNRL, ML MLP, RL, and GEA. Models with lower delay levels are better for real-time or near-real-time applications because they can provide predictions more quickly. In this regard, it is anticipated that DBN, DGM, ANGAN, AGAN, and MTAN will have fewer delays. Additionally, scalability is important for managing huge datasets, and it is projected that DBN, DGM, AGAN, MTAN, ANN, and DL-FHMC would have improved scalability levels. In order to choose the best model for their application in adversarial machine learning scenarios, researchers and practitioners are advised to take into account the specific use case, dataset characteristics, available computational resources, and trade-offs between accuracy, complexity, cost, delay, and scalability levels.

These metrics were combined to evaluate an Adversarial Machine Learning Rank (AMLR) via equation 1,

$$AMLR = \frac{A}{100} + \frac{S}{4} + \frac{1}{C} + \frac{1}{Co} + \frac{1}{D} \dots (1)$$

Based on this evaluation, it can be observed that AVA [31], RL [27], ALDC [50], ANN [43], AGAN [30], GEA [15], and CGAN [28] outperform other models, thus they can be used Accuracy, Complexity, Cost, Delay, and Scalability levels. Readers can use these models, and extend them to achieve better performance for different real-time use cases.

6.Conclusions & Future Scope

This review offers useful insights into the many tactics, techniques, and strategies applied to deal with security issues brought on by adversarial attacks. The results show how important this field of study is and how crucial it is to build machine learning models that are strong enough to survive attempts to trick or control them.

According to the review, machine learning systems are frequently vulnerable to adversarial attacks, and a number of attack methods have been suggested to take advantage of these weaknesses. As a result, scientists have created a variety of defense strategies, including adversarial training, generative models, explainable AI, and data augmentation approaches, to thwart these attacks. On numerous real-world datasets and applications, the effectiveness of different defense techniques was assessed, allowing for a thorough understanding of their advantages and disadvantages.

The analysis's main finding is the trade-off between model robustness and accuracy. While some models perform well in safe situations, they might not be able to recognize and fend off hostile attacks. Conversely, models with more robustness typically have lesser accuracy under normal circumstances. As a result, in the subject of adversarial machine learning, finding a balance between accuracy and robustness continues to be a major difficulty.

The future focus of this study will be on creating more sophisticated, adaptive defense systems that can change over time to fend off new adversarial attack methods. Additionally, it is necessary to combine various defense strategies to develop hybrid systems that provide stronger defense against a wider variety of attack vectors & scenarios.

To further improve the security and dependability of machine learning systems in real-world scenarios, research into the applicability of adversarial machine learning techniques in certain domains, such as healthcare, banking, and critical infrastructure, is very promising.

The creation of benchmark datasets and uniform evaluation techniques for adversarial machine learning will also make it possible to compare various defense strategies fairly and consistently. This will enable more thorough benchmarking and testing of suggested ways, providing better insights into their efficacy.

Exploring adversarial assaults and responses in developing fields like quantum machine learning and federated learning is a crucial component of future study. Understanding these technologies' vulnerabilities and creating strong security measures are essential as they gain popularity.

In order to assure the security and dependability of machine learning systems, this review and analysis highlights the need for more robust and resilient models while also shedding light on the current status of adversarial machine learning. In order to create a safer and more reliable machine learning environment for various situations, researchers, industry professionals, and legislators will need to work together to address these issues for different scenarios.

7.References

[1] Marshev, I.I., Zhukovskii, E.V. & Aleksandrova, E.B. Protection against Adversarial Attacks on Malware Detectors Using Machine Learning Algorithms. *Aut. Control Comp. Sci.* **55**, 1025–1028 (2021). https://doi.org/10.3103/S0146411621080198

[2] Asha, S., Vinod, P. Evaluation of adversarial machine learning tools for securing AI systems. *Cluster Comput* **25**, 503–522 (2022). https://doi.org/10.1007/s10586-021-03421-1

[3] Lin, J., Njilla, L.L. & Xiong, K. Secure machine learning against adversarial samples at test time. EURASIP J. on Info. Security 2022, 1 (2022). https://doi.org/10.1186/s13635-021-00125-2

[4] Kuzlu, M., Sarp, S., Catak, F.O. *et al.* Analysis of deceptive data attacks with adversarial machine learning for solar imagevoltaic power generation forecasting. *Electr Eng* (2022). https://doi.org/10.1007/s00202-022-01601-9

[5] Liu, G., Zhang, W., Li, X. *et al.* VulnerGAN: a backdoor attack through vulnerability amplification against machine learning-based network intrusion detection systems. *Sci. China Inf. Sci.* **65**, 170303 (2022). https://doi.org/10.1007/s11432-021-3455-1

[6] Gao, Xs., Liu, S. & Yu, L. Achieving optimal adversarial accuracy for adversarial deep learning using Stackelberg games. *Acta Math Sci* **42**, 2399–2418 (2022). https://doi.org/10.1007/s10473-022-0613-y

[7] Wilson, M., Vandal, T., Hogg, T. *et al.* Quantum-assisted associative adversarial network: applying quantum annealing in deep learning. *Quantum Mach. Intell.* **3**, 19 (2021). https://doi.org/10.1007/s42484-021-00047-9

[8] Ahmed, U., Lin, J.CW. Robust adversarial uncertainty quantification for deep learning fine-tuning. *J Supercomput* **79**, 11355–11386 (2023). https://doi.org/10.1007/s11227-023-05087-5

[9] Rathore, H., Sahay, S.K., Nikam, P. *et al.* Robust Android Malware Detection System Against Adversarial Attacks Using Q-Learning. *Inf Syst Front* **23**, 867–882 (2021). https://doi.org/10.1007/s10796-020-10083-8

[10] Puttagunta, M.K., Ravi, S. & Nelson Kennedy Babu, C. Adversarial examples: attacks and defences on medical deep learning systems. *Multimed Tools Appl* (2023). https://doi.org/10.1007/s11042-023-14702-9

[11] Dasgupta, D., Gupta, K.D. Dual-filtering (DF) schemes for learning systems to prevent adversarial attacks. *Complex Intell. Syst.* **9**, 3717–3738 (2023). https://doi.org/10.1007/s40747-022-00649-1

[12] Ahmed, U., Lin, J.CW. & Srivastava, G. Mitigating adversarial evasion attacks by deep active learning for medical image classification. *Multimed Tools Appl* **81**, 41899–41910 (2022). https://doi.org/10.1007/s11042-021-11473-z

[13] Sheikh, B., Zafar, A. Beyond accuracy and precision: a robust deep learning framework to enhance the resilience of face mask detection models against adversarial attacks. *Evolving Systems* (2023). https://doi.org/10.1007/s12530-023-09522-z

[14] Rathore, H., Samavedhi, A., Sahay, S.K. *et al.* Towards Adversarially Superior Malware Detection Models: An Adversary Aware Proactive Approach using Adversarial Attacks and Defenses. *Inf Syst Front* **25**, 567–587 (2023). https://doi.org/10.1007/s10796-022-10331-z

[15] Ren, M., Wang, YL. & He, ZF. Towards Interpretable Defense Against Adversarial Attacks via Causal Inference. *Mach. Intell. Res.* **19**, 209–226 (2022). https://doi.org/10.1007/s11633-022-1330-7

[16] Korycki, Ł., Krawczyk, B. Adversarial concept drift detection under poisoning attacks for robust data stream mining. *Mach Learn* (2022). https://doi.org/10.1007/s10994-022-06177-w

[17] Liu, KY., Li, XY., Lai, YR. *et al.* Denoised Internal Models: A Brain-inspired Autoencoder Against Adversarial Attacks. *Mach. Intell. Res.* **19**, 456–471 (2022). https://doi.org/10.1007/s11633-022-1375-7

[18] Shu, R., Xia, T., Williams, L. *et al.* Omni: automated ensemble with unexpected models against adversarial evasion attack. *Empir Software Eng* **27**, 26 (2022). https://doi.org/10.1007/s10664-021-10064-8

[19] Merzouk, M.A., Cuppens, F., Boulahia-Cuppens, N. *et al.* Investigating the practicality of adversarial evasion attacks on network intrusion detection. *Ann. Telecommun.* **77**, 763–775 (2022). https://doi.org/10.1007/s12243-022-00910-1

[20] Shao, R., Perera, P., Yuen, P.C. *et al.* Open-Set Adversarial Defense with Clean-Adversarial Mutual Learning. *Int J Comput Vis* **130**, 1070–1087 (2022). https://doi.org/10.1007/s11263-022-01581-0

[21] Akan, A.K., Akbas, E. & Vural, F.T.Y. Just noticeable difference for machine perception and generation of regularized adversarial images with minimal perturbation. *SIViP* **16**, 1595–1606 (2022). https://doi.org/10.1007/s11760-021-02114-x

[22] Wu, XG., Wu, HJ., Zhou, X. *et al.* Towards Defense Against Adversarial Attacks on Graph Neural Networks via Calibrated Co-Training. *J. Comput. Sci. Technol.* **37**, 1161–1175 (2022). https://doi.org/10.1007/s11390-022-2129-2

[23] Li, W., Zhao, B., An, Y. *et al.* Supervised contrastive learning for robust text adversarial training. *Neural Comput & Applic* **35**, 7357–7368 (2023). https://doi.org/10.1007/s00521-022-07871-5

[24] Tasyurek, M., Gul, E. A new deep learning approach based on grayscale conversion and DWT for object detection on adversarial attacked images. *J Supercomput* (2023). https://doi.org/10.1007/s11227-023-05456-0

[25] Li, Z., Yu, D., Wu, M. *et al.* Adversarial supervised contrastive learning. *Mach Learn* **112**, 2105–2130 (2023). https://doi.org/10.1007/s10994-022-06269-7

[26] Kherchouche, A., Fezza, S.A. & Hamidouche, W. Detect and defense against adversarial examples in deep learning using natural scene statistics and adaptive denoising. *Neural Comput & Applic* **34**, 21567–21582 (2022). https://doi.org/10.1007/s00521-021-06330-x

[27] C. Ren and Y. Xu, "Robustness Verification for Machine-Learning-Based Power System Dynamic Security Assessment Models Under Adversarial Examples," in IEEE Transactions on Control of Network Systems, vol. 9, no. 4, pp. 1645-1654, Dec. 2022, doi: 10.1109/TCNS.2022.3145285.

[28] D. Han et al., "Evaluating and Improving Adversarial Robustness of Machine Learning-Based Network Intrusion Detectors," in IEEE Journal on Selected Areas in Communications, vol. 39, no. 8, pp. 2632-2647, Aug. 2021, doi: 10.1109/JSAC.2021.3087242.

[29] K. He, D. D. Kim and M. R. Asghar, "Adversarial Machine Learning for Network Intrusion Detection Systems: A Comprehensive Survey," in IEEE Communications Surveys & Tutorials, vol. 25, no. 1, pp. 538-566, Firstquarter 2023, doi: 10.1109/COMST.2022.3233793.

[30] K. Mahmood, P. H. Nguyen, L. M. Nguyen, T. Nguyen and M. Van Dijk, "Besting the Black-Box: Barrier Zones for Adversarial Example Defense," in IEEE Access, vol. 10, pp. 1451-1474, 2022, doi: 10.1109/ACCESS.2021.3138966.

[31] K. Zhang, Z. Tian, Z. Cai and D. Seo, "Link-privacy preserving graph embedding data publication with adversarial learning," in Tsinghua Science and Technology, vol. 27, no. 2, pp. 244-256, April 2022, doi: 10.26599/TST.2021.9010015.

[32] J. Chen et al., "Adversarial Caching Training: Unsupervised Inductive Network Representation Learning on Large-Scale Graphs," in IEEE Transactions on Neural Networks and Learning Systems, vol. 33, no. 12, pp. 7079-7090, Dec. 2022, doi: 10.1109/TNNLS.2021.3084195.

[33] B. Gholami, Q. Liu, M. El-Khamy and J. Lee, "Multiexpert Adversarial Regularization for Robust and Data-Efficient Deep Supervised Learning," in IEEE Access, vol. 10, pp. 85080-85094, 2022, doi: 10.1109/ACCESS.2022.3196780.

[34] K. Grosse, L. Bieringer, T. R. Besold, B. Biggio and K. Krombholz, "Machine Learning Security in Industry: A Quantitative Survey," in IEEE Transactions on Information Forensics and Security, vol. 18, pp. 1749-1762, 2023, doi: 10.1109/TIFS.2023.3251842.

[35] C. Zheng, L. Pan and P. Wu, "Attribute Augmented Network Embedding Based on Generative Adversarial Nets," in IEEE Transactions on Neural Networks and Learning Systems, vol. 34, no. 7, pp. 3473-3487, July 2023, doi: 10.1109/TNNLS.2021.3116419.

[36] X. Xu, K. Lin, Y. Yang, A. Hanjalic and H. T. Shen, "Joint Feature Synthesis and Embedding: Adversarial Cross-Modal Retrieval Revisited," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 6, pp. 3030-3047, 1 June 2022, doi: 10.1109/TPAMI.2020.3045530.

[37] Y. Dong, S. Cheng, T. Pang, H. Su and J. Zhu, "Query-Efficient Black-Box Adversarial Attacks Guided by a Transfer-Based Prior," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 12, pp. 9536-9548, 1 Dec. 2022, doi: 10.1109/TPAMI.2021.3126733.

[38] J. Yu, Z. Shen and X. Zheng, "Joint Feature and Label Adversarial Network for Wafer Map Defect Recognition," in IEEE Transactions on Automation Science and Engineering, vol. 18, no. 3, pp. 1341-1353, July 2021, doi: 10.1109/TASE.2020.3003124.

[39] V. R. Kebande, S. Alawadi, F. M. Awaysheh and J. A. Persson, "Active Machine Learning Adversarial Attack Detection in the User Feedback Process," in IEEE Access, vol. 9, pp. 36908-36923, 2021, doi: 10.1109/ACCESS.2021.3063002.

[40] Y. Zhang et al., "Inaccurate-Supervised Learning With Generative Adversarial Nets," in IEEE Transactions on Cybernetics, vol. 53, no. 3, pp. 1522-1536, March 2023, doi: 10.1109/TCYB.2021.3104848.

[41] I. Vaccari, A. Carlevaro, S. Narteni, E. Cambiaso and M. Mongelli, "eXplainable and Reliable Against Adversarial Machine Learning in Data Analytics," in IEEE Access, vol. 10, pp. 83949-83970, 2022, doi: 10.1109/ACCESS.2022.3197299.

[42] I. Alsmadi et al., "Adversarial Machine Learning in Text Processing: A Literature Survey," in IEEE Access, vol. 10, pp. 17043-17077, 2022, doi: 10.1109/ACCESS.2022.3146405.

[43] Dingeto H, Kim J. Universal Adversarial Training Using Auxiliary Conditional Generative Model-Based Adversarial Attack Generation. *Applied Sciences*. 2023; 13(15):8830. https://doi.org/10.3390/app13158830

[44] Wang L, Sun Y, Zhang X. Quantum Adversarial Transfer Learning. *Entropy*. 2023; 25(7):1090. https://doi.org/10.3390/e25071090

[45] Lee G-C, Li J-H, Li Z-Y. A Wasserstein Generative Adversarial Network–Gradient Penalty-Based Model with Imbalanced Data Enhancement for Network Intrusion Detection. *Applied Sciences*. 2023; 13(14):8132. https://doi.org/10.3390/app13148132

[46] Alkhowaiter M, Kholidy H, Alyami MA, Alghamdi A, Zou C. Adversarial-Aware Deep Learning System Based on a Secondary Classical Machine Learning Verification Approach. *Sensors*. 2023; 23(14):6287. https://doi.org/10.3390/s23146287

[47] Preuveneers D. AutoFL: Towards AutoML in a Federated Learning Context. *Applied Sciences*. 2023; 13(14):8019. https://doi.org/10.3390/app13148019

[48] Alabsi BA, Anbar M, Rihan SDA. Conditional Tabular Generative Adversarial Based Intrusion Detection System for Detecting Ddos and Dos Attacks on the Internet of Things Networks. *Sensors*. 2023; 23(12):5644. https://doi.org/10.3390/s23125644

[49] A. Abusnaina et al., "DL-FHMC: Deep Learning-Based Fine-Grained Hierarchical Learning Approach for Robust Malware Classification," in IEEE Transactions on Dependable and Secure Computing, vol. 19, no. 5, pp. 3432-3447, 1 Sept.-Oct. 2022, doi: 10.1109/TDSC.2021.3097296.

[50] Rudnitskaya, E.A., Poltavtseva, M.A. Adversarial Machine Learning Protection Using the Example of Evasion Attacks on Medical Images. *Aut. Control Comp. Sci.* **56**, 934–941 (2022). https://doi.org/10.3103/S0146411622080211