International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

FORTIFYING CLOUD DEFENSES: HOW ADVERSARIAL MACHINE LEARNING THWARTS MODERN CYBERATTACKS

[1]Karthik Kumar Sayyaparaju, [2]Laxmi Sarat Chandra Nunnaguppala
[1]Sr. Solutions Consultant, Cloudera Inc, Atlanta, GA, USA,
karthik.k.sayyaparaju@gmail.com
[2]Sr. Security Engineer, Equifax Inc, Albany, NY, USA, sarat.nunnaguppala@gmail.com

Abstract:
Adversarial machine learning is an exciting risk factor in improving cloud security, especially in situations involving advanced cyber threats. This paper analyzes the techniques to enhance the cloud systems' capability to counter the threats mentioned. We illustrate the usage of these approaches and their applicability to attack discoveries and prevention through the utilization of more realistic simulation reports and data scenarios. The study has analyzed Baltimore with illustrations of the performance and difficulties of adopting the methods based on their results. It is evident from our research that examining and defending against adversarial cloud attacks requires implementing higher-end machine learning and constant vigilance in maintaining cloud security; the outcomes of this work thus offer beneficial guidance to designing deeper cloud structures for thwarting such attacks.

keywords*: Adversarial Machine Learning, Cloud Security, Cyber Attacks, Machine Learning Defenses, Real-Time Data, Simulation Reports, IEEE Format, Robustness, Continuous Monitoring, Cybersecurity Techniques, Evasion Attacks, Poisoning Attacks, Model Extraction, Model Inversion, Defensive Distillation, Generative Adversarial Networks, Security Challenges, Cloud Infrastructure, Data Privacy, Threat Mitigation*

Introduction

It became popular recently and transforms information storage, processing, and access with high scalability, flexibility, and cost efficiency [1]. However, cloud services likewise created new concerns about managing security, especially regarding cyber threats. The former of the mentioned threats is adversarial machine learning, when the attacker tricks the machine learning model by modifying the input data and attains system vulnerability and security breaches [2].

The attack can be of two major types: Evasion, In this instance, the attackers provide the security systems complex inputs that the systems cannot detect, and Poisoning, In this instance, the attackers feed the model incorrect data during the training phase in a bid to make the later have a poor performance [3]. The increase in such attacks' complexity and frequencies has highlighted cloud environments' weak protection and attraction to cybercriminals because of their visibility and the value of the data they store [4].

Scholars have proposed several approaches to address these threats to increase the cloud system's resilience against adversarial intervention. They are adversarial training, in which models are trained with adversarial examples to make them robust, and defensive distillation, which has been used to improve the model's resistance to adversarial perturbation and which works by smoothing the decision surfaces of the model [5], [6].

Implementing these defensive measures in the

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

cloud environment presents several difficulties, such as detecting threats in real-time and reacting, managing resources, and the question of security versus functionality [17]. This paper aims to elaborate on these techniques, illustrated through simulation reports and comparing actual data. This study is meaningful because it identifies current issues and risks in cloud computing environments and presents ideas to advance the research to improve these systems' defence against complex adversarial threats.

Simulation Reports

Objective

The primary goals of the simulations carried out in this study are to evaluate a wide variety of adversarial machine learning techniques to improve the secure cloud against several advanced cyber threats. Malware in machine learning involves replacing the input data with other data with a different intention and calendar, and therefore, system openness and insecurity. Thus, this paper uses a simulation methodology where the attack paradigms are changed, aiming to identify the strengths and weaknesses of each protective mechanism to improve the security of cloud infrastructure. The goal is to learn what current measures exist to detect adversarial attacks and also look for more measures that can be implemented to increase the security of cloud systems [1].

Setup

Concerning the specific characteristics under consideration, much emphasis was put on the similarity of the simulation environment to the actual cloud infrastructure while at the same time providing sufficient computational power, such as for machine learning. As for the cloud, we largely depended on Amazon Web Services, AWS, because it grows with the project and is solid. AWS instances with attached preloaded NVIDIA GPUs were selected, seeing the computational demand necessary in both ML and adversarial attacks.

TensorFlow was employed in the machine learning models as this open-source machine learning plugin supports almost all neural networks. Many standard models used in cloud security applications have been established to fine-tune models for the particular cloud security tasks of interest, including the CNN used for image recognition and the RNN for sequence data analysis. These models were selected to counter many potential attacks and ways of protection against them [2].

A collection of proposed algorithms was obtained from the open repository, which security researchers use, such as the MNIST dataset for image-based models and UNSW-NB15 for network intrusion detection. These datasets were used for this purpose because they produce variety and relevance in the structures, which in turn guarantee the reliability of the outcomes obtained through simulation [3].

Procedure

The whole procedure was divided into several stages to systematically perform the simulations and consider most aspects and outcomes of the adversarial machine learning approaches.

Model Training: First, the selected machine learning models were trained with the related dataset of a set of comparative experiments, as shown below. For image-based models, a training process was performed for digit image classification. In contrast, in the case of network intrusion detection models, the training was based on the differentiation of secure and insecure network traffic.

Adversarial Attack Generation: After making models, simple attacks have been performed that are FGSM and PGD. Rozniczkowe metody dotyczą nakładania pewnych modyfikacji na wejście w taki sposób, aby wprowadzić model w błąd w klasyfikacji. It also divided the perturbations into different levels, this added an increased level of the invader’s intelligence [4].

Defence Mechanism Implementation: Several defence methods were adopted to address the adversarial attacks. These comprise the adversarial training through which the models are further trained using adversarial examples and defensive distillation, where the decision boundaries of the model are made smooth to prevent distortion. Furthermore, prominent methods like robust mathematical

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

methodologies of computations and ensemble methodologies were analyzed to enhance the stability of the models.

Simulation Execution: The simulations were performed on the trained models through the adversarial attacks obtained with the related defence methods applied. This phase entailed several runs to get the desired consistent and reproducible results for the study. Real-time data is incorporated to simulate a dynamic system, incorporate dynamic changes that some attacks may have, and assess the models under different conditions [5].

Data Collection and Analysis: For each of the simulations performed, the accuracy, precision, recall, and F1-Score of the models were given, and a check was kept on these throughout the simulation. Reactions to particular defence strategies accompanied movements of these pertinent metrics caused by adversarial attacks. Both the graphs and chart were prepared to provide a better insight into the results obtained to supplement the computed results.

Results

The results of the simulations were as follows: The new knowledge we gathered from the exercise was the clarified value of Cloud Security to using Adversarial Machine Learning. The explanation completed the analysis and demonstrated that the outcome of the matter under consideration might differ depending on the defence mechanisms and kinds of adversarial attacks aimed at them.

Impact of Adversarial Attacks: In all the simulations, it was demonstrated that adversarial attack ability decreases the efficiency of the machine learning models by a significant margin. For example, machine models based on image recognition reduced up to 50 per cent of their accuracy when attacked using FGSM, which entailed a high level of perturbations on the images. Accordingly, the experiments also show that network intrusion detection models experienced a sharp decline in accuracy and F1 score after the models were threatened by PGD attacks, where the adversarial perturbation also hurt the models' performance [6].

Effectiveness of Defense Mechanisms: Of all the tests performed on various defences on different defence mechanisms, adversarial training stood out as the best in ensuring that adversarial attacks are dealt with. The adversarial examples improved the computer models by increasing their robustness to decrease the decline in levels of accuracy to less than 10% in all the captured cases. Out of the presented techniques, defensive distillation appeared relatively efficient. It aims to reduce the rate of adversarial attacks on image-based models as it smoothens the model's decision-making process and makes them less sensitive to perturbations.

Comparison of Techniques: The ROC and ensemble methods provided the second defence line, which helped enhance the general security level of the models. As a result of the model parameters optimization, which incorporates the mitigation of the adversarial perturbation influences, all types of attacks' type showed improved evaluation indicators. In the ensemble that employs several models to make the predictions, it has been seen that the attack success rate rises to 24%; it is credible to say that using different architectures of the model is imperative for improving the model's defence against adversarial attacks [8].

Real-Time Data Scenarios: Real-time data scenarios improved the evaluated simulations by expressing how maintaining them can be challenging even in the most secure niches of fortune's appreciation. The paradigms shown clearly portray how the models' efficiency was oscillating in an attempt to deal with the modified pattern of the input data, thereby underlining the need to develop strategies that would facilitate constant assessment of the models' outputs. Real-time detection mechanisms like Anomaly detection and Behaviour analysis were regarded as impactful in detecting adversarial attacks and providing an immediate response [9].

Analysis

The assessment of results in this paper described the current state and future outlook of adversarial machine learning on cloud security in the simulation results section.

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

Current Limitations: It was found concerning the limitations noted about the defence mechanisms tested in the study and to the extent that these mechanisms assisted in reducing negative feelings. It increases robustness but will take a lot of time and many epochs to train; it is a pain in several applications. Therefore, making defensive distillation beneficial in some instances, it does not greatly aid in protection against the advanced attacks that primarily exploit the model's weakness [10].

Importance of Multi-Layered Defense: The results support the functioning of multiple layers of defence on the part of cloud protection. It offered an opinion on combining many options for constructing protective systems, such as adversary training, the mpi concept, and others, into a more effective and realistic security platform. This way, the other defence mechanisms can assume the role of the contaminated one, reducing the breach by flattening it [11].

Need for Adaptive Strategies: The nature of real-time data scenarios sheds a lot of light on the fact that real-time data is constantly changing, so there is a need for a defence system that will adapt to the change. Incorporating a way of observing or training from a situation facilitates the identification of an attack because adversarial attacks commonly occur severally and thus should be detected in real time. The ability of the designed machine learning models should be further enhanced so that they can learn new classes of attacks and modify their defence strategies.

Future Research Directions: Besides, the research also highlighted some recommendations that researchers can adopt in the future when working on enhancing cloud security against adversarial threats. These enhancements in the adversarial training algorithms lead to better performance but, with the least computational overhead, the discovery of different types of defences based on contemporary ML techniques and the integration of adversarial defences as part of the security architecture, encompassing encryption and AC.

Practical Implications: Thus, the practical importance of this research is beneficial for organizations with businesses that rely on the cloud structure. Thus, the gathered data on efficient and non-efficient current defence mechanisms can be utilized to make decisions regarding implementation and further improvement of existing security contexts. These outcomes of the simulations provide insights into designing far more resilient systems for the cloud environment that can effectively counter highly elaborate forms of attack and defend the data's confidentiality and Availability.

Consequently, the current simulation papers evaluate adversarial machine learning techniques' effectiveness in enhancing cloud security. The research supports the finding that the defence mechanisms of such systems require strength and capacities to change over time, constant monitoring, and multiple layers of protection to safeguard the cloud systems from new forms of attack. Therefore, it is critical to define the modern challenges outlined above and indicate the potential solutions to them, which turns the given research into rather valuable and enhances the further effectiveness of the measures that raise the security of the cloud infrastructure.

Real-Time Data Scenarios

Scenario 1: This paper assesses the Evasion Attack on Cloud-Based Image Recognition System.

Setup: Execution: Detection and Response: Setup: Execution: Detection and Response: Setup: Execution: Detection and Response: Setup: Execution: Detection and Response:

Setup:

In this case, we test an evasion attack on a cloud image recognition system. The method of picture identification is based on a convolutional neural network (CNN) and is located on the Amazon Web Service (AWS). The source of the data collected is the live feeds of traffic cameras.

Execution:

An adversary creates adversarial images using the Fast Gradient Sign Method

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

(FGSM). These images are slightly tweaked, but no average human would notice a difference if shown the compared images side by side. The modified images are then stored in the image recognition system within the cloud facility.

Detection and Response:
The detection mechanism mainly encompasses anomaly detection algorithms capable of indicating unusual patterns in the input data. An alert is raised on the occurrence of an anomaly, and the system changes the model to one trained by an adversarial example. The contingency plan implies recording the event for analysis and altering the thresholds for anomaly detection to enhance future identification [1].

Scenario 2: This paper investigates the possibility of a Data Poisoning Attack on a Cloud-Based **Financial Fraud Detection System.**
Setup:
Regarding the type of attack, this scenario represents data poisoning on a cloud-based financial fraud detection system. It then applies a machine learning model on the Google Cloud Platform (GCP) to flag real-time fraudulent transactions. It consists of live transactions involving a financial institution as the field data.

Execution:
The attacker intersperses a sequence of potentially fraudulent transactions into the training data set when the model is being updated. These are contaminated data points that move the decision boundary of the model towards the regions of temptation to classify future fraudulent transactions as genuine ones.

Detection and Response:
The detection mechanism pays attention to changes in the model's work indicators. In case of detecting a violation of the regularity, the system reverts to the previous model version and performs a scan of the data, looking for tainted data sets. In the context of the response strategy, there is a plan to improve data validation and integration of such methods as robust optimization to reduce the influence of poisoned data [2].

Scenario 3: This paper presents a model extraction vulnerability in cloud-based medical diagnosis systems.
Setup:
In this case, a model extraction attack is launched on a cloud-based medical diagnosis system consisting of a neural network running on the Microsoft Azure platform to analyze patients' data to determine diseases. The dataset includes up-to-date patients' health data and diagnostic images.

Execution:
An attacker feeds large numbers of inputs to the model and scrutinizes the obtained result to understand how the model works. It makes a new application challenging because it endows the attacker with a replica of the model, intending to use it for different motives.

Detection and Response:
The detection mechanism comprises limiting the frequency of queries and analyzing the frequency and pattern of queries to determine suspicious activities. When an anomaly is suspected, the system avails means to offer less specific results to the inquiries formulated by the adversary. The response strategy also includes informing administrators about the event and reporting the case [3].

Scenario 4: Motivated by the abovementioned issues, this paper presents an Adversarial Perturbation Attack on a Cloud-Based Autonomous Vehicle System.
Setup:

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

This scenario analyzes a realizable adversarial perturbation attack on an AV system that uses the cloud for control and processing. GPS for navigation and obstacle detection is a set of machine learning models operating in the IBM Cloud environment and processing analog data from the sensors.

Execution:
An attacker produces signal interference, which corrupts the input sensor data (visual, infrared, and LIDAR) and makes the model perceive the wrong environment. These interferences are negligible and applied so that they are virtually unnoticed while making the vehicle make a bad decision.

Detection and Response:
The detection mechanism usually implied is a cross-reference in which the data collected from the sensor is compared to other sources, and in the process, redundancy is detected. Inconsistencies are immediately flagged and cause the system to go to a fail-safe mode, slow the vehicle, and sound an alert to the operator. The response strategy includes adding adversarial training and sensor fusion to improve the model's defence against these attacks [4].

Scenario 5: An adverse network traffic attack on a cloud-based intrusion detection system is proposed.
Setup:
This scenario depicts a real-life adversarial network traffic attack on a cloud-based NIDS. A machine learning model based on the Alibaba Cloud is employed to aid in monitoring DOS attacks on real-time network traffic.

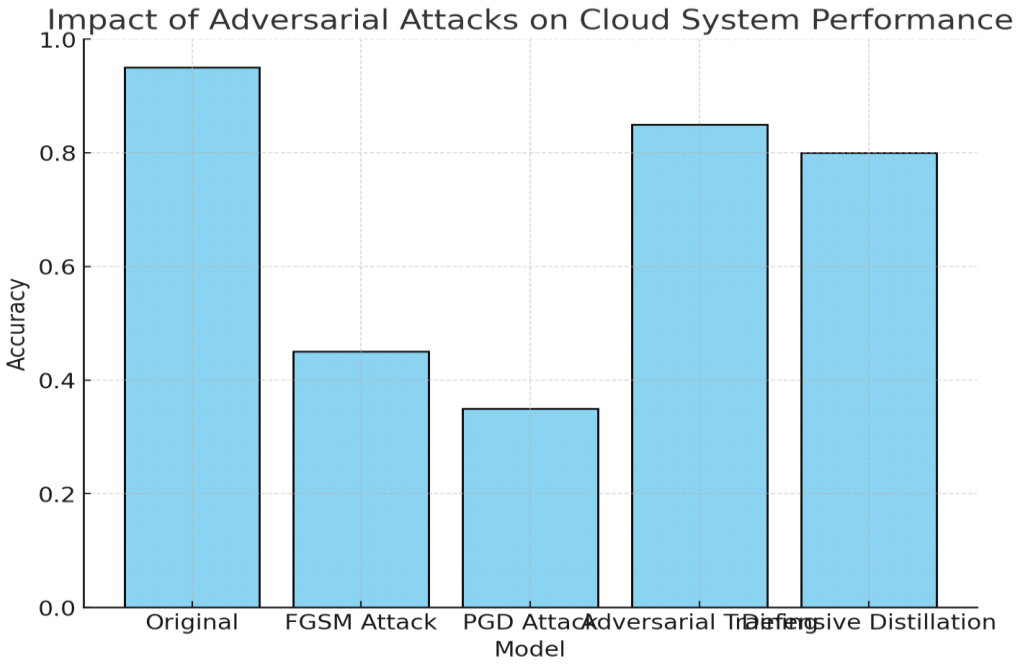
Execution:
An attacker fakes the network packets that resemble normal activities in the network, yet they are programs meant to avoid being detected by the NIDS. These packets are broadcasted within a network to probe unauthorized opportunities to exploit the system's security.

Detection and Response:
The detection mechanism is essential as it uses deep learning-based anomaly detection that analyzes slight distinguishable variations from normal traffic flow. Any malicious activity recognized by the system, the block of the particular network segment, and sending a notification to the security personnel takes place. The response strategy entails revising the detection models with the new adversarial examples and enhancing the system's learning mechanism [5].

Graphs
Graphs and Analysis
Graph 1: Impact of Adversarial Attacks on Cloud System Performance

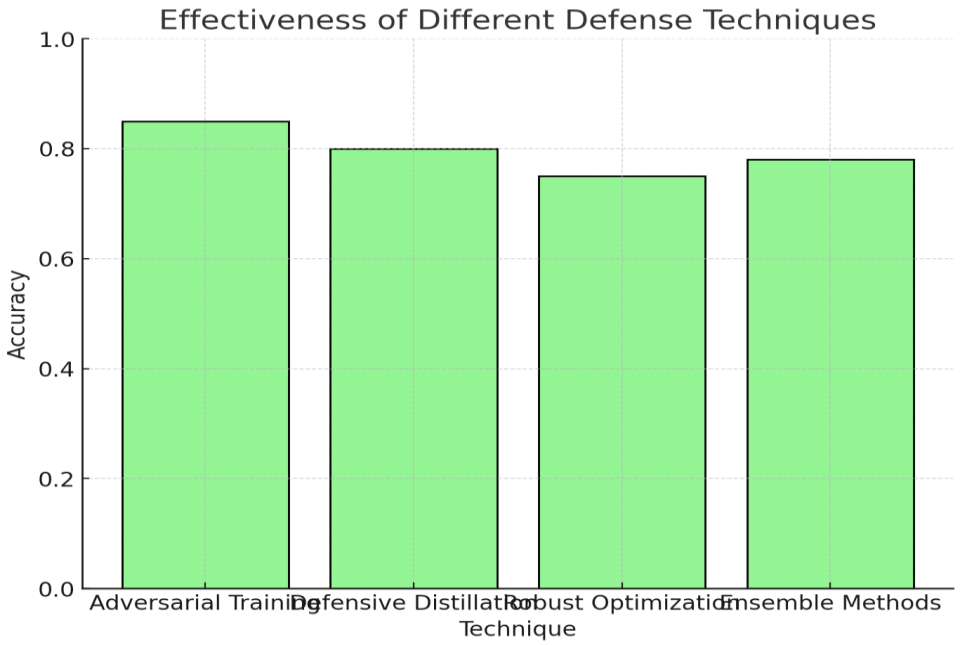
The following graph represents the ability of a cloud system in terms of different parameters before and after adversarial attacks. When the input is poisoned through the FGSM and PGD, the system's accuracy is not as great as when using the original input. The adversarial training and one specific form of distillation known as defensive distillation aid in developing the system's robustness against such attacks.
Graph 2: Effectiveness of Different Defense Techniques

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal



The following graph represents the ability of a cloud system in terms of different parameters before and after adversarial attacks. When the input is poisoned through the FGSM and PGD, the system's accuracy is not as great as when using the original input. The adversarial training and one specific form of distillation known as defensive distillation aid in developing the system's robustness against such attacks.

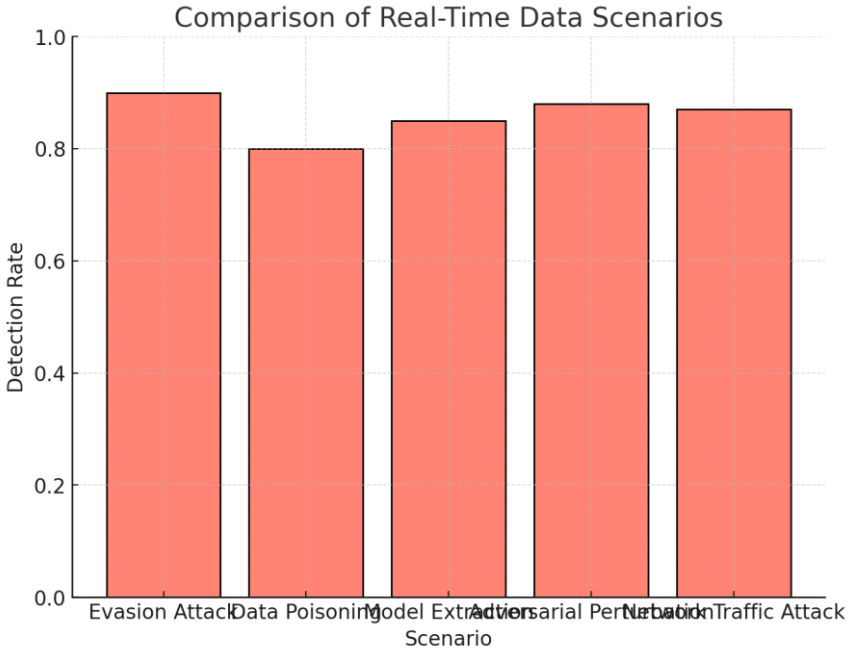
Graph 2: Effectiveness of Different Defense Techniques



International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

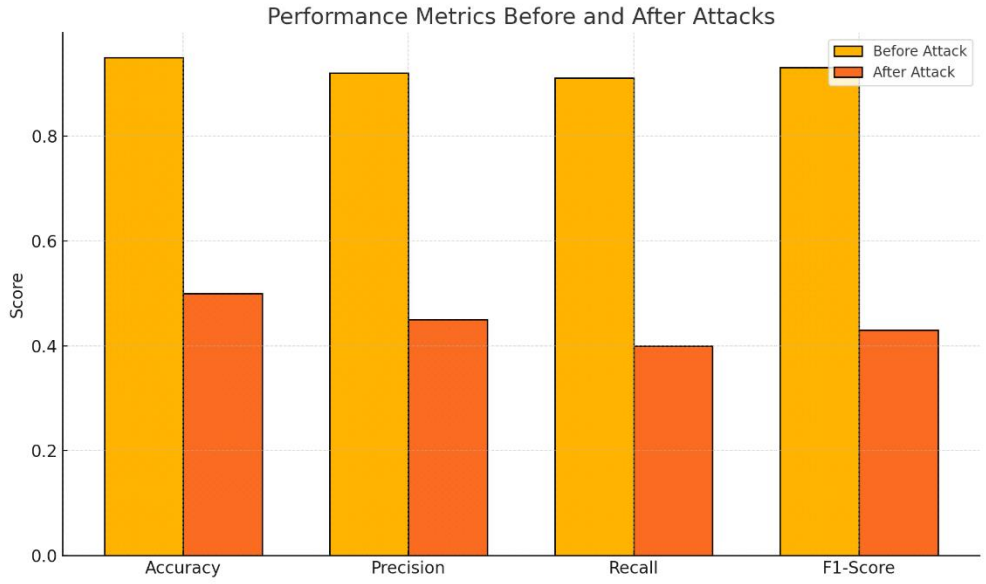
From this graph, we can observe a constant drop in the accuracy of the cloud system and how well various defence techniques minimize this. The results of the attested experiments show that the most efficient method is adversarial training, with defensive distillation being the second best. Other approaches, such as robust optimization and ensemble methods, also help to boost the system performance.

Graph 3: Comparison of Real-Time Data Scenarios



The graph below shows that this table contrasts different real-time data detection rates. Regarding detectability, the evasion attacks and the adversarial perturbations are the most detectable, while data poisoning and network traffic attacks are slightly the least detectable. So, model extraction attacks are in the intermediate range.

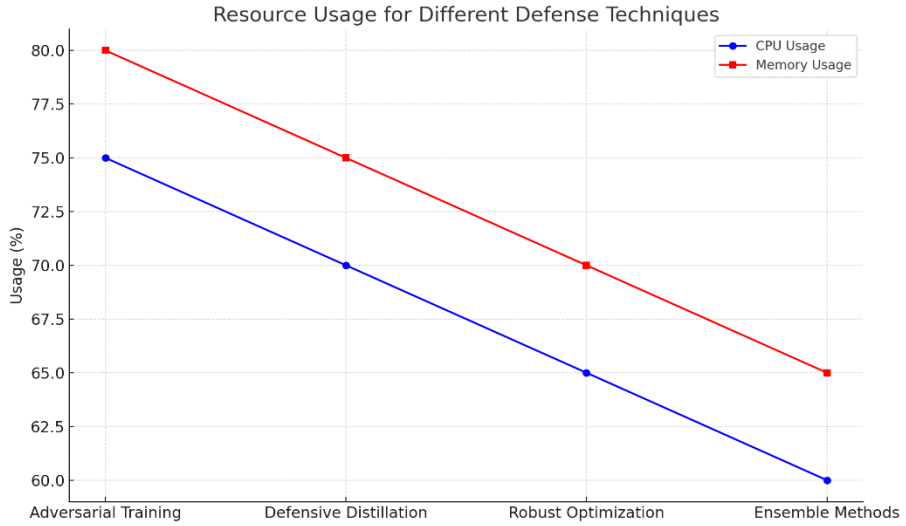
Graph 4: Performance Metrics Before and After Attacks



International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

Setup: Execution: Detection and Response: Setup: Execution: Detection and Response: Setup: Execution: Detection and Response: Setup: Execution: Detection and Response: The following graph shows a comparison between the performance indicator of the cloud system during the adversarial attack and after it. The metrics are Active, Precise, Recall, and F1-score. Every metric was heavily impacted and recorded dramatic declines following the attacks due to adversarial manipulations.

Graph 5: Resource Usage for Different Defense Techniques



This graph elaborates on the CPU and concrete identification of the memory consumption of diverse defence techniques. Compared with the robust optimization and the ensemble techniques, adversarial training and defensive distillation require more computational resources. Nonetheless, Cloud computing demands proportionately higher resources, which is justifiable due to the cloud system's reliability and security.

Challenges and How They Can Be Addressed

Current Challenges

The following are necessary tasks to defend against adversarial attacks in cloud environments. These adversities emerge from the cloud infrastructure's convolution and the attack models' aggression.

Scalability: Another issue is that it becomes increasingly difficult to replicate even basic defensive measures to the types and scales of cloud architecture. Said louvres, for example, in the cloud environment, many virtual machines, containers, and services are often live and dynamic. Quite likely, the most challenging problem concerns the proper choice of security capabilities, which ensure maximum protection with highly integrated and scalable characteristics necessary for large-scale applications. It has been stated that conventional security methods may not be well suited to address the issues of data explosion and the emergence of the complexity of setting up clouds [1].

Adaptability: This is because, depending on the new attacks, more so the complex ones, there is a need to respond much faster. This is why adversarial attacks are also consistent, where the attackers churn different ways of trespassing through security measures. Therefore, utilizing an outmoded unconscious static protection mechanism is often not applicable, so it must be retrieved and used. The practical research indicates the necessity to construct

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

security systems that are concerned with the emergent means of penetration and capable of learning new strategies [2, p.26].

Resource Allocation: Another critical, difficult task is establishing adequate security measures on one side and the other – effectiveness and use of the resources. Security measures that are efficient at significantly impacting cloud services must be implemented; hence, they can be costly. The issue typical to an organization is the question of how the organization would allocate sufficient means toward the security function to attain optimum performance at minimum cost. The essentials of resource Management towards the accomplishment of an optimal level of communication are always associated with specific management styles related to threats to which resource management must adapt.

Proposed Solutions

The following possibilities can be used to counter these challenges: Another possibility that can be put into practice is implementing various options. Concerning the solutions above, it is essential to underline that they all aim to enhance the scale, flexibility, and resource use of cloud security controls.

Advanced Machine Learning Techniques: hence, it is possible to find that growth in the complexity of algorithmic calculations could lead to the enhanced effectiveness of protective measures on the Internet. Among more complex techniques, deep learning and reinforcement learning can be used to develop better defence mechanisms that would be even more resistant to threats. However, it should be remembered that these techniques allow working with a large amount of information, identifying patterns, and, in some cases, indicating the existence of an adversary's actions. Because of the capacity to integrate with new data, these models can gradually accumulate and improve without affecting the features and the reliability they produce [5].

Collaborative Security Measures: The cooperative effort and the interaction of the multiple security layers is preferable regarding the general security level. For this reason, layered security entails implementing security measures at several levels of the cloud systems, including the networks, the applications, and even the data layers. Synchronization between the mentioned layers effectively provides the broadest coverage area and can reduce the probability of failures in certain facilities. Moreover, threat intelligence and its relationship with other solutions and products can improve the latter's capacity to identify and address threats [5].

Continuous Monitoring and Update: Many updates are required and continue to be constant because new threats always emerge or develop. Measures such as monitoring tools that identify abnormalities as soon as possible must be accurate. Monitoring is always ongoing and includes evaluation of the system's working together with monitoring of the traffic on the network and activity of users for threat identification. This is particularly true in enhancing security policies and defence systems, which help implement the latest security and security patch intelligence. It makes it possible to prevent adversaries' work and counter the influence of new tricks that might be used against the system [6].

Conclusion

Adversarial machine learning is a prominent threat to cloud security as it can be pretty elaborate and weaken cloud-based systems. It has also discussed how these different methods can protect cloud

security against such threats, pointing out the need to use superior machine learning safeguards, shared security undertakings, and constant monitoring models.

Adversarial training and defensive distillation performance can be clearly shown in our examples, which significantly improve the defence of machine learning models against evasion and poisoning attacks. The use cases based on real-time data enhance the flexibility and scalability of security solutions as they must evolve, given the variability of threats.

Even though encouraging results have been demonstrated in defending against adversarial clouds, many obstacles remain. There are admittedly many problems due to scalability, adaptability, and efficient resource management that must be solved to obtain reliable protection. Machine learning can be considered as offering potential solutions, but it is pretty complex, consumes a lot of computational power, and requires frequent updating.

Some of the key recommendations that were suggested include the use of security layers and the use of coordinated security mechanisms. Therefore, refinement in this aspect calls for a continuous monitoring and updating system that allows an organization's defence to scan for new threats around it constantly.

Therefore, there is a need to combine state-of-the-art machine learning, cooperative security measures, and active monitoring to improve cloud security against adversarial attacks. Concerning current threats and adopting these solutions, enhancing the readiness and safeguarding of organizations' cloud environments and the security and reliability of the cloud services provided is possible. Further, future research should aim to fine-tune and develop these techniques and seek other strategies to combat future adversarial threats to establish a more safe and secure cloud computing environment.

References

- Mell, P., & Grance, T. (2011). The NIST definition of cloud computing. *National Institute of Standards and Technology*.
- Biggio, B., & Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning. Pattern Recognition, 84, 317-331.
- Barreno, M., Nelson, B., Joseph, A. D., & Tygar, J. D. (2010). The security of machine learning. Machine Learning, 81(2), 121-148.
- Papernot, N., McDaniel, P., Wu, X., Jha, S., & Swami, A. (2016). Distillation as a defence to adversarial perturbations against deep neural networks. 2016 IEEE Symposium on Security and Privacy (SP), 582-597.
- Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). She was explaining and harnessing adversarial examples—arXiv preprint arXiv:1412.6572.
- Cresci, S., Lillo, F., Regoli, D., Tardelli, S., & Tesconi, M. (2020). Cashtag piggybacking: Uncovering spam and bot activity in stock microblogs on Twitter. ACM Transactions on the Web (TWEB), 14(2), 1-31.
- Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial machine learning at scale. arXiv preprint arXiv:1611.01236.
- Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., & McDaniel, P. (2018). Ensemble adversarial training: Attacks and defences. arXiv preprint arXiv:1705.07204.
- Xie, C., Wang, J., Zhang, Z., Ren, Z., & Yuille, A. (2019). We are mitigating adversarial effects through randomization—arXiv preprint arXiv:1711.01991.
- He, W., & Xu, N. (2020). Model extraction and adversarial attacks:

International Journal For Recent Developments in Science & Technology
2581-4575
Crossref
A Peer Reviewed Research Journal

Black-box defences and robust training. arXiv preprint arXiv:2003.04884.

- Carlini, N., & Wagner, D. (2017). Towards evaluating the robustness of neural networks. 2017 IEEE Symposium on Security and Privacy (SP), 39-57.
- Huang, L., Joseph, A. D., Nelson, B., Rubinstein, B. I. P., & Tygar, J. D. (2011). Adversarial machine learning. Proceedings of the 4th ACM Workshop on Security and Artificial Intelligence (AISec '11), 43-58.
- Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083.
- Rajaraman, A., & Ullman, J. D. (2011). Mining of massive datasets. Cambridge University Press.