Journal of Information and Technology

Resistance to Fast Gradient Sign Method Using Block Switching Algorithm

James Kihara Mwangi, Dr. Jane Kuria& Dr. John Wandeto

ISSN: 2617-3573

Stratford Peer Reviewed Journals and Book Publishing
Journal of Information and Technology
Volume 8||Issue 1 ||Page 43-64||February |2024|
Email: info@stratfordjournals.org ISSN: 2617-3573

Resistance to Fast Gradient Sign Method Using Block Switching Algorithm

[1]*James Kihara Mwangi, [2]Dr. Jane Kuria& [3]Dr. John Wandeto

[1]Student, Department of Information Technology, Dedan Kimathi University of Technology

[2&3] Lecturer, Department of Information Technology, Dedan Kimathi University of Technology

***How to cite this article*: Mwangi, K, J., Kuria, J., & Wandeto, J. (2024). Resistance to Fast Gradient Sign Method Using Block Switching Algorithm. *Journal of Information and Technology*, 8(*1*), 43-64. https://doi.org/10.53819/81018102t7002**

Abstract

Traditional ways of protecting against the "Fast Gradient Sign Method" attack usually involve methods like altering the input data before processing, training systems to recognize harmful inputs, or identifying harmful inputs directly. However, these traditional methods have a number of shortcomings, including their limited success, vulnerability to more advanced attacks, difficulty in understanding how they work, and too much dependence on standard sets of data for testing. By creating a strong protective, the system against The Fast gradient Sign Technique, the objective of this study is to enhance the resilience of machine learning algorithms against adversarial attacks while improving their safety and dependability in the highest level of accuracy and performance. The study is guided by three objectives: to investigate the robustness of existing Deep Learning algorithms for defense against the Fast Gradient Sign Method; to implement the block-switching algorithm for defending against the Fast Gradient Sign Method; and to evaluate the performance metric of the block-switching algorithm for the protection of deep learning models against adversarial attacks. The study will consider three theories that underpin the block-switching algorithm including: Avalanche effect, Cryptographic Strength, and Probability theory. The research will use datasets from the Modified National Institute of Standards and Technology and the Canadian Institute for Advanced Research. It will select commonly used deep learning models for image classification, such as Residual Neural Network, Visual Geometry Groups, or Inception, for analysis. The study will employ the Fast Gradient Sign Method to create adversarial examples for each model within the chosen datasets. The researcher will then compare each Deep Learning model's performance on the adversarial dataset with the original dataset to see how resilient each one is against first gradient sign adversarial assaults. To evaluate these criteria including accuracy, precision, recall, and F1 score will be applied. The research will perform a sensitivity analysis on the parameters used in the Fast Gradient Sign Method attack generation to investigate how the attack strength and the number of iterations affect the model's robustness against adversarial attacks. To perform the sensitivity analysis, the researcher will use Python and a set of test data in the Tensor Flow library.

Stratford Peer Reviewed Journals and Book Publishing
Journal of Information and Technology
Volume 8||Issue 1 ||Page 43-64||February |2024|
Email: info@stratfordjournals.org ISSN: 2617-3573

Keywords: *Block-Switching Algorithm, Cryptographic Strength, Adversarial Attacks, Probability Theory, Encryption Security.*

1.1 Introduction

Recent progress in deep learning has markedly improved capabilities in fields like image and speech recognition, underscoring the critical need for robust security measures in these technologies due to the fast-paced advancements in artificial intelligence (Neil et al., 2020; Xie et al., 2020). A major issue is how susceptible machine learning models are to adversarial examples. These are inputs that have been intentionally altered to mislead models into making wrong choices (Goodfellow, Shlens & Szegedy, 2015). Such deceptive inputs can cause significant mistakes, underscoring the continuous struggle to maintain the safety and dependability of machine learning systems. Techniques such as the Gradient Sign Method have been identified as effective means for attackers to exploit these vulnerabilities (Kurakin, Goodfellow & Bengio, 2016), prompting research into defense mechanisms to protect against such attacks, particularly in high-stakes areas like computer vision and the Industrial Internet of Things (Hassan, 2021; Dunn, Moustafa & Turnbull, 2020).

To tackle the limitations of traditional machine learning methods, researchers have developed representation learning, which involves creating concise and informative representations of input data that capture its essential characteristics (Bengio, Courville, & Vincent, 2013). Convolutional neural networks (CNNs) have emerged as a key technique in this domain, offering versatility in processing both numerical and visual data for a variety of applications, including cybersecurity (Taheri, Salem, & Yuan, 2018). Alongside, visualization-based botnet identification methods have been introduced, utilizing visual data representations to enhance data feature understanding (Vinayakumar et al., 2020). These advanced approaches enable the prediction and learning of distinctive network traffic features, contributing to the identification of malicious activities in cybersecurity (Catak et al., 2021). The increasing occurrence of adversarial attacks, which use slightly altered data to trick deep learning models, presents serious risks to the trustworthiness and dependability of these systems. This situation underscores the critical need for strong defense strategies to protect against these vulnerabilities (Janiesch et al., 2021).

As artificial intelligence becomes more prevalent, adversarial attacks, particularly evasion and poisoning attacks, are expected to rise, challenging the integrity of AI systems (Liang et al., 2022). Evasion attacks manipulate input data to cause incorrect algorithm predictions, while poisoning attacks corrupt the training set to teach the algorithm wrong input-output relationships. Despite ongoing efforts to mitigate these threats, attackers now target the machine learning algorithms themselves, previously focusing mainly on network attack detection (Merenda, Porcaro & Lero, 2020). Various methods have been proposed to generate adversarial examples, with some based on evolutionary optimization techniques (Su, Vargas, & Sakurai, 2019). The Fast Gradient Sign Method (FGSM) is a notable gradient-based technique that creates adversarial examples by altering the gradient's sign (Goodfellow, Shlens, & Szegedy, 2015). Despite advancements in defensive strategies like adversarial training, defensive distillation, and input transformations, these have not been entirely effective in thwarting attacks, indicating the ongoing challenge in securing AI systems against sophisticated adversarial techniques (Vinayakumar et al., 2020; Taheri et al., 2020).

The present study presents a strong defense strategy for machine learning that focuses on the model's ability to stay accurate even when faced with unexpected or distorted inputs (Catak et al., 2021). To fight back against the gradient sign method often used in adversarial attacks, the study suggests using a block-switching algorithm, an idea taken from the field of cryptography. In cryptography, block-switching algorithms work by encrypting data in set-sized pieces, which makes them more secure and efficient than methods that encrypt data bit by bit, as explained by Hutter & Tunstall (2019) and further explored by Alekseev and Bozhko (2020). Such algorithms, including the well-known Advanced Encryption Standard (AES) and Blowfish, perform a series of operations like substitution, permutation, and modular arithmetic on plaintext blocks to produce ciphertext (Dolmatov & Baryshkov, 2020). The proposed block-switching approach in machine learning disrupts the attacker's ability to exploit gradient information by randomly reordering the input blocks, thereby complicating the extraction of meaningful data and enhancing the model's defense against adversarial attacks.

1.2 Statement of the Problem

Traditional defensive systems against the FGSM attack often rely on techniques such as input preprocessing, adversarial training, or detecting adversarial examples. However, these traditional defenses have several weaknesses such as limited effectiveness, vulnerability to adaptive attacks, lack of interpretability, and over-reliance on benchmark datasets (Liao et al., 2018; Guo et al., 2017). Although the FGSM (Fast Gradient Sign Method) attack is widely recognized, the exact reasons behind its effectiveness are not fully understood (Goodfellow et al., 2015; Athalye et al., 2018). This gap in knowledge hinders the development of effective defense methods. Further research is needed on how well defenses against FGSM attacks can be applied in different situations. A frequently used example of an adversarial attack involves an image of a panda. The left picture shows a picture of a panda that has not been modified at all. The intricate neural network which has undergone training on a particular dataset of images, has the ability to identify the image in question as a panda. It is important to note the model possesses enough confidence in its judgment. In this regard, it is estimated that there is a 58% chance that the picture depicts a panda. The middle image shows the best direction to change all pixels when computed the exact way we may change the image to cause, for instance, ConvNet to make a mistake. The outcome is a picture that human beings cannot distinguish from that initial panda when the image of the organized assault is multiplied by a small coefficient and added to the original panda. (Goodfellow, Bengio & Courville, 2016). According to human perception, there is no discernible distinction or dissimilarity between the image on the left and the one on the right.

Stratford Peer Reviewed Journals and Book Publishing
Journal of Information and Technology
Volume 8||Issue 1 ||Page 43-64||February |2024|
Email: info@stratfordjournals.org ISSN: 2617-3573

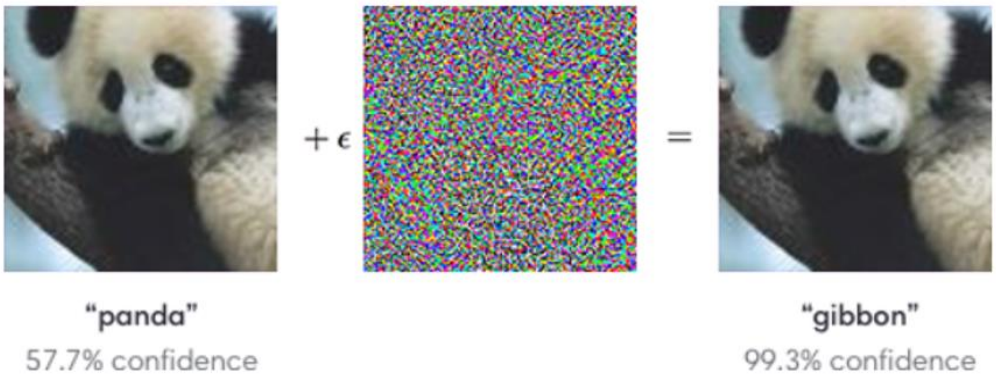

Figure 1: Adversarial Input

(Source: Goodfellow, Bengio & Courville, 2016)

1.3 Objectives of the Study

The study was guided by general and specific objectives.

1.3.1 General Objective

The objective of this study is to enhance the resilience of machine learning algorithms against adversarial attacks while improving their safety and dependability in the highest level of accuracy and performance.

1.3.2 Specific Objectives of the Study

i. To investigate the robustness of existing Deep Learning algorithms for defense against the Fast Gradient Sign Method.

ii. To implement the block-switching algorithm for defending against the Fast Gradient Sign Method.

iii. To evaluate the performance metric of the block-switching algorithm for the protection of deep learning models against adversarial attacks.

1.4 Research Questions

i. Can the robustness of existing Deep learning algorithms for defense against the Fast Gradient Sign Method investigated?
ii. Can the block-switching algorithm for defense against Fast Gradient Sign Method be implemented?
iii. Can the performance metric of the block-switching algorithm for the protection of deep learning models against adversarial attacks be evaluated?

1.4.1 Research Hypotheses

i. $H0_1$: There is no significant relationship between the robustness of existing Deep learning algorithms and defense against the Fast Gradient Sign Method.

ii. $H0_2$: There is no significant relationship between block-switching algorithm and defense against Fast Gradient Sign Method.
iii. $H0_3$: There is no significant relationship between the block-switching algorithm for protection of deep learning models against adversarial attack.

2.1: Conceptual Review

The study presents a review of concepts, such as adversarial attacks, whitebox and blackbox attacks and block switching methods.

2.1.1 Adversarial Attacks

Adversarial attacks pose a serious cybersecurity risk by altering data inputs to deceive machine learning models (MLMs) into making wrong choices, as pointed out by Carlini et al. (2018). With the growing use of machine learning, there's a heightened focus on the susceptibility of these systems to such attacks, jeopardizing their safety and reliability (Oprea, 2021). Machine learning systems depend on data and statistical patterns to perform tasks ranging from image recognition to language translation. However, this reliance makes them susceptible to adversarial manipulations that are subtle enough to evade human detection but can significantly skew the model's outputs, as discussed by Catak and Yayilgan (2021). These vulnerabilities emphasize the necessity for continuous research aimed at comprehending and mitigating adversarial threats. This is crucial to uphold the reliability and safety of machine learning applications.

Addressing adversarial attacks involves both detection and defense strategies. Detection is challenging due to the attackers' intent to make their manipulations imperceptible, making it hard to discern whether inaccuracies in model predictions are due to attacks or simple errors (Martinez et al., 2019). Some researchers have made progress in identifying data patterns indicative of attacks and enhancing model robustness against such manipulations (Sharif et al., 2017; Martinez et al., 2020). On the defense front, strategies include developing more attack-resistant algorithms, incorporating adversarial examples into training to improve model resilience, and employing ensemble models to complicate the attackers' efforts (Qui et al., 2021; Luo, 2020; Xie et al., 2020). These approaches aim to fortify machine learning models against adversarial influences, safeguarding their accuracy and reliability in various applications

2.1.2 Classification of Adversarial Attacks

Various adversarial examples can be categorized as follows

2.1.2.1 Whitebox vs. Blackbox Attacks

Cybersecurity professionals are constantly developing strategies to defend against various types of cyber-attacks, including whitebox and blackbox attacks. In a whitebox attack, the attacker has complete knowledge of the system's internal structure, including access to source code and design documents, enabling the identification and exploitation of obscure vulnerabilities (Athalye et al., 2018; Carlini & Wagner, 2017; Moosavi-Dezfooli et al., 2016). This thorough understanding empowers attackers to execute advanced attacks like code injection, involving the insertion of malicious code into an application to undermine its integrity (Goodfellow et al., 2015). In contrast, blackbox attacks occur when attackers lack knowledge about the internal workings of the system, resorting instead to trial-and-error techniques such as brute force attacks to identify vulnerabilities.

Despite requiring less skill, these attacks can be executed more quickly since they don't necessitate a deep understanding of the system's architecture (Dezfooli et al., 2016; Athalye et al., 2018; Su, Vargas & Sakurai, 2019). Each attack type presents unique challenges and requires tailored defensive strategies to ensure the security and integrity of the systems.

2.1.2.2 Imperceptible vs. Perceptible Attacks

Cyber attackers deploy a range of tactics to infiltrate systems and access sensitive data, distinguishing their methods based on visibility to users and systems. Imperceptible attacks, such as rootkits and Trojan horses, are stealthy operations designed to evade detection while granting attackers root-level access or delivering malicious payloads without the user's knowledge (Mingkang et al., 2021; Zhao, Liu & Larson, 2020). These attacks pose significant challenges due to their ability to remain hidden, making them particularly dangerous as they can silently compromise systems and exfiltrate data. In contrast, perceptible attacks like phishing and distributed denial of service (DDoS) assaults aim to be noticeable, either by directly engaging with the user to deceitfully obtain sensitive information or by overwhelming system resources to disrupt service, often triggering security alerts (Bai et al., 2022; Carlini & Wagner, 2018). The stark difference in the visibility of these attacks necessitates diverse defensive strategies, with imperceptible attacks requiring more sophisticated detection mechanisms to uncover hidden threats, while perceptible attacks demand robust preventive measures to block or mitigate the impact of the assault.

2.1.2.3 Targeted vs. Untargeted Attacks

In cybersecurity, the distinction between targeted and untargeted attacks highlights the varied strategies employed by cyber attackers. Targeted attacks are meticulously planned and executed against specific individuals or organizations, involving in-depth reconnaissance to exploit particular vulnerabilities. These attacks are characterized by their tailored nature and persistence, often conducted by well-resourced and organized groups aiming for precise objectives (Carlini & Wagner, 2018; Taori et al., 2018). Conversely, untargeted attacks, often referred to as "spray and pray" tactics, aim to exploit a broad range of systems by leveraging automated tools to scan for and exploit known vulnerabilities without specific targets in mind (Anthi et al., 2018). These attacks do not discriminate and instead cast a wide net in the hopes of finding and exploiting any available weaknesses. Preventive measures for targeted attacks include employee training to heighten awareness of security protocols, network segmentation to isolate critical systems and hinder lateral movement of attackers within the network, and strict access controls to restrict unauthorized access to sensitive data, all of which are crucial in mitigating the risks posed by these meticulously planned cyber incursions (Carlini & Wagner, 2018; Taori et al., 2018; Anthi et al., 2018).

2.1.3 Block Switching Method

The block Switching (BS) technique is a good optimization technique to train deep neural networks. The BS model is trained in two distinct levels. Individual sub-models that possess similar features within the same framework are trained separately, in the Fast phase using random weights (Ashutosh et al., 2021). Switching between various blocks, makes the network extract complex and important features to optimize the performance and the robustness of the proposed model, making it give an accurate prediction because random initialization and stochasticity in the

training process can lead to variations or unpredictability in the outcomes of MLMs (Xiao et al., 2020).

The sub-block is further subdivided into two segments after the Fast training, the lower module, and the upper module. Lower module is assembled into parallel blocks while the upper module is discarded of the block switching. In the second round, the entire block switching using the same set of training data, for the purpose to attain the best accuracy (Xiao et al., 2020).

2.1.4 FAST Gradient Sign Method (FGSM)

The Fast Gradient Sign Method (FGSM), pioneered by Goodfellow et al. (2015), stands out as a significant adversarial attack technique. It creates slight perturbations in input data, deceiving neural networks into reaching incorrect conclusions. By manipulating the gradient of the loss function, this method exposes the susceptibility of neural networks to minor alterations in input, stressing the importance of enhancing model robustness. However, FGSM's effectiveness is limited in scenarios where attackers cannot access the model's internals, such as in black-box attacks (Kurakin, Goodfellow, & Bengio, 2016; Carlini & Wagner, 2017). Despite various proposed defensive measures like adversarial training and defensive distillation, challenges remain in fully countering FGSM attacks (Vinayakumar et al., 2020; Taheri et al., 2020). Nonetheless, FGSM's ease of use and quick execution make it a valuable tool for both attacking and enhancing model defenses, suggesting its dual utility in cybersecurity (Goodfellow et al., 2015; Althaye et al., 2018).

2.2 Theoretical Review

The study considered three theories that underpin the block-switching algorithm including: Avalanche effect, Cryptographic Strength, and Probability theory.

2.2.1 Avalanche Effect

The Avalanche Effect, introduced by Horst Feistel in 1973, is fundamental in cryptography, emphasizing that minor changes in input should result in major alterations in output, a principle critical for the security of block ciphers (Shi, H., Deng & Guan, 2011). This effect enhances the difficulty for attackers to exploit vulnerabilities, as small errors or variations in input lead to unpredictable changes in the encrypted output, thereby securing the cipher against pattern analysis and attacks (Bhoge & Chatur, 2014). The block-switching algorithm capitalizes on this phenomenon, guaranteeing that even slight modifications in the plaintext lead to substantial and unforeseeable alterations in the ciphertext. This enhances the encryption's resilience against decryption efforts and data leakage (Echeverri, 2017; Ramanujam & Karuppiah, 2011). This research utilizes the Avalanche Effect to affirm the effectiveness of the block-switching algorithm, showcasing its ability to preserve encryption reliability and precision, even when confronted with minor input perturbations.

2.2.2 Cryptographic Strength

Cryptographic strength is a critical measure in cryptography, denoting the security level of an algorithm based on factors like key length, encryption rounds, and attack resistance, without being credited to any single individual (Preneel, 2000). The strength of the block-switching algorithm, for instance, hinges on these aspects, ensuring its effectiveness against various cryptographic

attacks and enhancing the accuracy of encryption outcomes (Afzal et al., 2000). Incorporating the Advanced Encryption Standard (AES) due to its robustness and widespread acceptance, the algorithm employs AES's symmetric key and variable key lengths for heightened security (Sulak et al., 2000; Ukrop, 2016). Furthermore, the adoption of Cipher Block Chaining (CBC) mode, which utilizes an Initialization Vector (IV) for encryption randomization, bolsters the algorithm against pattern detection and cipher vulnerabilities. The employment of multiple encryption rounds solidifies the algorithm's reliability, ensuring significant alterations in ciphertext output with any slight changes in the input, thereby underpinning the encryption process's accuracy and the cipher's overall cryptographic robustness (Afzal et al., 2020; Ukrop, 2016)..

2.2.3 Probability Theory

The foundational framework of probability theory, pivotal for analyzing the block-switching algorithm's effectiveness, traces back to the 17th-century correspondence between Blaise Pascal and Pierre de Fermat, later systematically expounded by Jacob Bernoulli in "Ars Conjectandi" (Grimmett & Stirzaker, 1992). This mathematical theory aids in evaluating the block-switching algorithm by estimating the likelihood of specific outcomes within the encryption process, particularly its capacity to generate ciphertext that appears random, thereby enhancing security by obscuring patterns in encrypted data (Sulak et al., 2010). Utilizing probability theory, one can assess the randomness of the ciphertext produced, calculating the probability of the output being indistinguishable from random data under given inputs and encryption settings (Li et al., 2012). This approach not only verifies the encryption's integrity but also gauges the encryption's resilience against cryptographic attacks by estimating the effort required for an adversary to compromise the system (Killmann & Schindler, 2001). Consequently, probability theory offers a robust mathematical basis for scrutinizing the block-switching algorithm's accuracy and security, ensuring its reliability in safeguarding data.

2.3 Empirical Literature Review

Adversarial attacks in machine learning have garnered significant attention due to their ability to exploit vulnerabilities in Machine Learning Models (MLMs) through carefully crafted perturbations, impacting various domains such as natural language processing, computer vision, and cybersecurity. These attacks are primarily categorized based on their reliance on the model's decision-making process or its comprehensive information, with techniques like Boundary Attack and Evolutionary Attack focusing on decision-based strategies, and others employing gradient-based methods to manipulate input data and induce errors (Qiu et al., 2019; Chen, Jordan & Wainwright, 2020; Dong et al., 2020; Mao et al., 2020; Mekala, Porter & Lindvall, 2020; Kwon et al., 2021). The sophistication of these attacks, especially in black-box scenarios where attackers have limited knowledge of the underlying model, underscores the urgent need for developing more resilient defense mechanisms to safeguard sensitive systems, particularly in areas like face recognition (Dong et al., 2019).

Research endeavors have focused on strengthening the resilience of deep neural networks against adversarial inputs. Khalil (2021) notably investigated the detection of first-order adversarial attacks using Siamese Neural Networks (SNNs), showcasing the model's exceptional performance across different datasets and attack strategies while maintaining generality. This progress in defending against adversarial threats underscores the effectiveness of specialized neural network

architectures in recognizing and mitigating malicious inputs, thereby enhancing the security and dependability of machine learning systems.

In parallel, innovative approaches like the one proposed by Taheri et al. (2020), which utilizes Generative Adversarial Networks (GANs) to generate new adversarial examples for model retraining, show promise in fortifying models against such attacks. This technique leverages the adversarial generation capabilities of GANs to iteratively improve model resilience, showcasing the effectiveness of adaptive retraining strategies in the ongoing battle against adversarial threats. Similarly, the research conducted by Chen (2020) on the resilience of deep neural networks and tree-based models introduces a novel method for learning robust trees, addressing the max-min saddle point issue in node splitting and suggesting efficient tree-building techniques that significantly enhance model robustness.

Emerging defensive strategies like the hierarchical random switching (HRS) and CRU-Net defense concept further illustrate the evolving landscape of adversarial defense mechanisms. HRS, with its multi-block randomization approach, demonstrates enhanced resilience against adaptive attacks and white-box misclassification techniques, improving the robustness-accuracy trade-off (Wang et al., 2020). CRU-Net, drawing inspiration from previous defensive models, employs residual learning and U-Net architectures to effectively map adversarial examples back to clean images, showcasing its efficacy in maintaining network robustness in computational vision applications (Ali et al., 2022). Additionally, the novel noise data enhancement framework (NDEF) introduced by Xie et al. (2021) incorporates random erasing to mitigate over-fitting to adversarial samples, offering a promising avenue for enhancing model defense against a broad spectrum of adversarial attacks. These advancements highlight the dynamic and multifaceted nature of research in adversarial defense, underscoring the critical importance of continuous innovation in securing machine learning models against evolving threats.

3.0 Research Methodology.

The study leveraged the extensive ImageNet dataset as its primary data source, providing a broad range of labeled images for comprehensive model training and evaluation. A selection of established deep learning models commonly employed in image classification tasks—specifically ResNet, VGG, and Inception—were chosen for their proven efficacy and widespread use within the research community. To assess the vulnerability of these models to adversarial threats, the Fast Gradient Sign Method (FGSM) was utilized to generate adversarial examples, introducing carefully crafted perturbations to the input images. In response to these adversarial challenges, a defensive architecture was devised, incorporating Convolutional Variational Auto-Encoders (CVAE), Block Switching (BS), and Grad-CAM, aimed at enhancing the models' resilience against such attacks. The robustness of the models was evaluated by comparing their performance on both the original and adversarially altered datasets, employing key metrics such as accuracy, precision, recall, and F1 score. This methodological approach facilitated a nuanced analysis of each model's resistance to FGSM-induced adversarial examples and the efficacy of the proposed defensive mechanisms. Additionally, a rigorous analysis delved into the models' sensitivities to adversarial parameters, leveraging TensorFlow and PyTorch libraries for implementation and validation. This comprehensive approach enabled a nuanced understanding of model vulnerabilities and the effectiveness of defensive strategies in mitigating adversarial threats.

4.0 Research Findings and Discussion

The study findings are presented per objective.

4.1 Experiment Results

Table 1 below represent the results obtained after performing the sensitivity analysis using Python and a set of test data in the TensorFlow library. The FGSM attack was utilized to generate adversarial examples with varying values of epsilon and iteration numbers.

Table 1: Summary of Sensitivity Analysis Results

Number of Images	Attack Model			Defense Model		
	Successful attack	Failed attack	Attack Model Accuracy (%)	Successful Defense	Failed Defense	Defense Model Accuracy (%)
80	74	6	91.89	71	9	90.96
80	75	5	93.33	67	13	83.96
160	150	6	93.33	134	26	83.96
240	225	15	93.33	201	39	83.96
322	299	23	92.31	283	39	89.81
392	363	29	92.01	344	48	89.63
397	368	29	92.12	347	50	89.16
401	372	29	92.20	349	52	88.65
402	373	29	92.23	350	52	88.69
412	383	29	92.43	357	55	88.12

The accuracy of the model is assessed using both the original test data and the generated adversarial examples. The process is repeated for different values of epsilon and number of iterations.

4.2 The Attack Model

Figure 2 shows the percentage of attack model accuracy as it changes with different iteration and number of images.

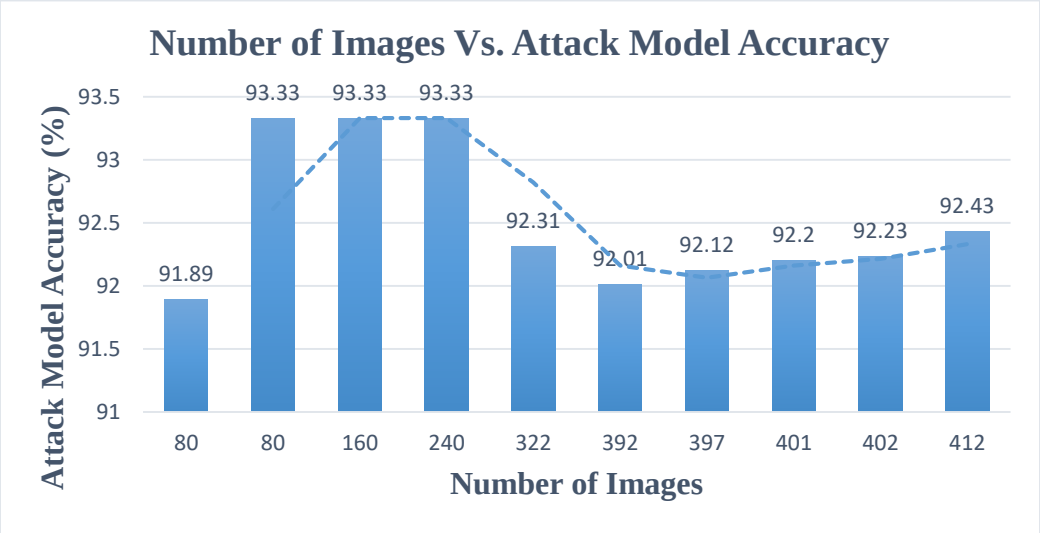

Figure 2: Number of images versus attack model accuracy

According to Figure 2, as the number of images rises from 80 to 240, the accuracy of the attack model increases from 91.89% to 93.33%. This indicates a positive correlation, where a greater number of images corresponds to higher accuracy in the attack model. Between 240 and 402 images, the accuracy of the attack model remains relatively consistent, hovering around 93.33% to 92.43%. This suggests that increasing the number of images beyond 240 does not lead to a significant improvement in accuracy. After 402 images, there is a slight decrease in accuracy, with the accuracy gradually dropping from 92.43% to 92.12%. This indicates a mild decrease in performance of the attack model as the number of images further increases. Between the data points, there are some fluctuations in accuracy even when the number of images remains constant. For example, there is variation in accuracy around 92.33% for some data points with 402 images. This suggests that factors other than the number of images may be influencing accuracy. Figure 3 below shows a similar trend when the successful attacks are plotted against the accuracy of the attack model.

Stratford Peer Reviewed Journals and Book Publishing
Journal of Information and Technology
Volume 8||Issue 1 ||Page 43-64||February |2024|
Email: info@stratfordjournals.org ISSN: 2617-3573

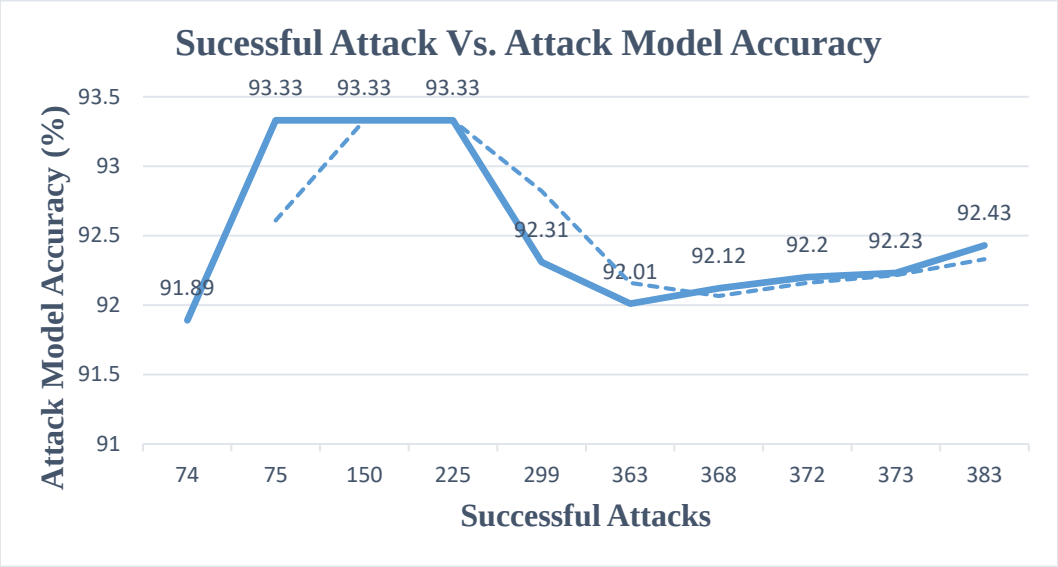

Figure 3: Successful Attacks vs. Attack Model Accuracy

From Figure 3 above, as the accuracy of the model increases from 91.89% to 93.33%, the number of successful attacks generally increases. This indicates a positive relationship between model accuracy and the number of successful attacks. Once attack model accuracy reaches 93.33%, the number of successful attacks stabilizes at 225. This suggests that beyond an accuracy of 93.33%, further improvements in accuracy do not result in a significant increase in successful attacks.

4.3 The Defense Model

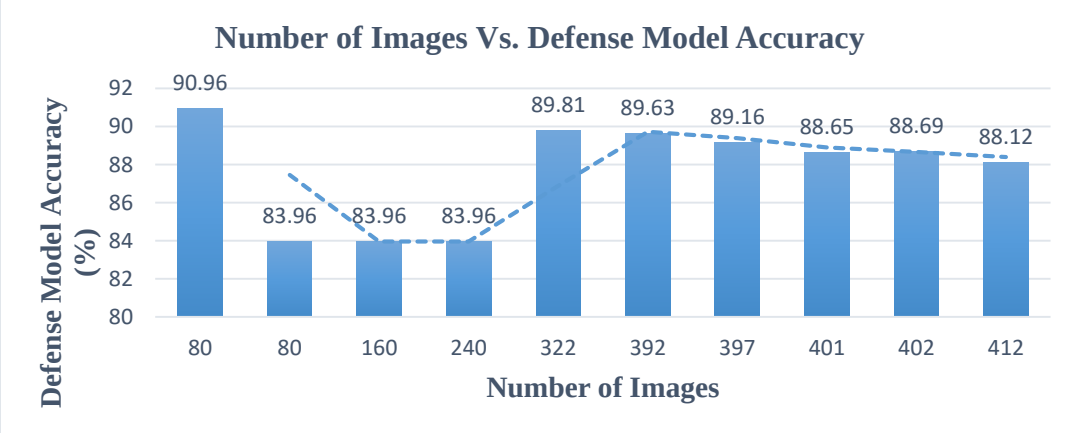

Figure 4: Number of Images vs. defense model accuracy

Figure 4 above shows that as the number of images increases from 80 to 412, the accuracy remains relatively stable around 83.96% to 90.96% for the defense model. There is a slight decrease in the accuracy of the defense model as the number of images increases from 80 to 412. This suggests that, in this dataset, having more images may not necessarily lead to higher defense model accuracy. There are fluctuations in accuracy even when the number of images remains the same.

Stratford Peer Reviewed Journals and Book Publishing
Journal of Information and Technology
Volume 8||Issue 1 ||Page 43-64||February |2024|
Email: info@stratfordjournals.org ISSN: 2617-3573

For instance, the model accuracy at 80 images is different for two data points (90.96% and 83.96%). This indicates that factors other than the number of images may be influencing accuracy in this case.

Figure 5 below shows the behavior of accuracy of the model for each instance of successful defense.

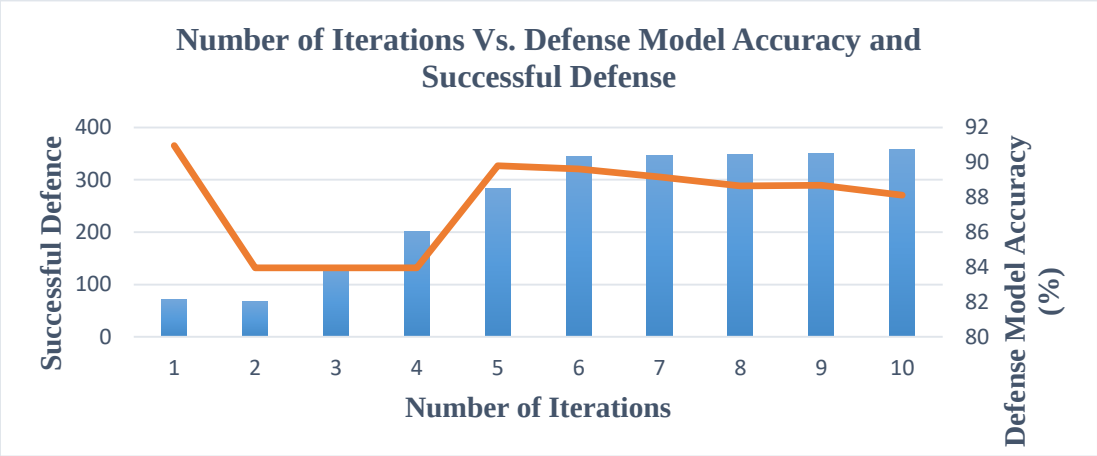

Figure 5: Number of Iterations Vs. Defense Model Accuracy and Successful Defense

The accuracy of the defense model remains constant at 83.96% for all data points. This indicates that the accuracy of the defense model is consistent and doesn't vary with changes in the number of successful defenses. The number of successful defense instances increases as you move down the dataset, going from 71 to 357. This suggests that the defense model is becoming more successful in defending against attacks.

4.4 Evaluation of the Models

The study models are evaluated per objective.

4.4.1 Successful Attacks versus Success Defense

The mean is a measure of central tendency. For successful attack, the mean is 268.20, and for successful defense, it is 250.30. On average, there were 268.20 successful attacks and 250.30 successful defenses, respectively. However, the median, which represents the middle value in a dataset when values are arranged from lowest to highest, differs. According to Manikandan (2011), the median for successful attacks is 331.00, while for successful defenses, it is 313.50. This suggests that in both cases, the median is higher than the mean, indicating a negatively skewed distribution.

The standard deviation measures the spread or dispersion of data points around the mean (El Omda & Sergent, 2023). A higher standard deviation indicates more variability in the data. For successful attack, the standard deviation is 127.240, and for successful defense, it is 121.092. This implies that there is some variability in the number of successful attacks and successful defenses, with a slightly lower standard deviation for successful defense. The variance is the square of the standard deviation and provides a measure of the spread of data. For successful attack, the variance is 16,189.956, and for Successful defense," it is 14,663.344.

Stratford Peer Reviewed Journals and Book Publishing
Journal of Information and Technology
Volume 8||Issue 1 ||Page 43-64||February |2024|
Email: info@stratfordjournals.org ISSN: 2617-3573

Skewness measures the asymmetry of the data distribution. A negative skewness value indicates a leftward (negatively) skewed distribution, meaning the tail on the left side of the distribution is longer or fatter than the right side (Senger, 2013). Both variables have negative skewness (Successful Attack: -0.715, Successful Defense: -0.679), confirming the leftward skew. The SPSS output table for the descriptive statistics is shown below;

Table 2: Descriptive Statistics

		Successful Attack	Successful Defense
N	**Valid**	10	10
	Missing	0	0
Mean		268.200	250.300
Median		331.000	313.500
Std. Deviation		127.240	121.092
Variance		16189.956	14663.344
Skewness		-.715	-.679
Std. Error of Skewness		.687	.687
Kurtosis		-1.337	-1.456
Std. Error of Kurtosis		1.334	1.334

A One-Sample T-Test was done to compare the means between successful attacks and successful defense and the results presented below.

Table 3: One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
Successful Attack	10	268.20	127.240	40.237
Successful Defense	10	250.30	121.092	38.293

The standard error of the mean for successful attack is 40.237 while for successful defense is 38.293. According to Lee et al. (2015), the SEM indicates the precision of the sample mean estimate. A higher SEM suggests that the sample mean may be less precise as an estimate of the population mean.

Stratford Peer Reviewed Journals and Book Publishing
Journal of Information and Technology
Volume 8||Issue 1 ||Page 43-64||February |2024|
Email: info@stratfordjournals.org ISSN: 2617-3573

Table 4: One-Sample Test

	Test Value = 0					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Successful Attack	6.666	9	.000	268.200	177.18	359.22
Successful Defense	6.536	9	.000	250.300	163.68	336.92

Both successful attack and successful defense have very low p-values (p = 0.000), indicating that the sample means for both variables are significantly different from the test value of 0. The 95% confidence intervals further support this conclusion, as they do not include 0. This suggests that both successful attack and successful defense are significantly greater than 0.

4.4.2 Attack versus Defense Accuracy

The results of descriptive statistical analysis using SPSS are presented below.

Table 5: Attach Vs. Defense Accuracy Descriptive Statistics

		Attack Accuracy	Defense Accuracy
N	Valid	10	10
	Missing	0	0
Mean		92.52	87.69
Median		92.27	88.67
Std. Deviation		.580	2.687
Variance		.336	7.217
Skewness		.783	-.707
Std. Error of Skewness		.687	.687
Kurtosis		-1.303	-1.256
Std. Error of Kurtosis		1.334	1.334

The mean for attack accuracy is 92.52, indicating that the average attack accuracy is 92.52%. The standard deviation for attack accuracy is 0.580, which quantifies the amount of variation or dispersion in the data (Lee et al., 2015). A lower standard deviation suggests less variability in attack accuracy. The variance is the square of the standard deviation and provides a measure of the spread of data. For attack accuracy, the variance is 0.336. Skewness measures the asymmetry of the data distribution. A positive skewness value, such as the one reported (0.783), indicates a rightward (positively) skewed distribution. This implies that the tail on the right side of the distribution is longer or thicker than the left side. Kurtosis, as described by Kim (2013), measures the "tailedness" of the data distribution. A negative value, like the reported -1.303, indicates a platykurtic distribution, meaning the distribution has thinner tails and is less peaked than a normal distribution. It's worth noting that the statistics provided for defense accuracy are similar to those for attack accuracy.

Stratford Peer Reviewed Journals and Book Publishing
Journal of Information and Technology
Volume 8||Issue 1 ||Page 43-64||February |2024|
Email: info@stratfordjournals.org ISSN: 2617-3573

A One-Sample T-Test was done to compare the means between successful attacks and successful defense and the results presented below.

Table 6: One-Sample Statistics: Attack Vs. Defense Accuracy

	N	Mean	Std. Deviation	Std. Error Mean (SEM)
Attack Accuracy	10	92.52	.580	.183
Defense Accuracy	10	87.69	2.687	.850

The standard error of the mean for attack accuracy is 0.183. According to Andrade (2020), a lower SEM suggests that the sample mean is a more precise estimate of the population mean. The SEM for defense accuracy is 0.850. The SEM for defense accuracy is higher than that of attack accuracy, indicating that the sample mean for defense accuracy is less precise as an estimate of the population mean.

Table 7: One-Sample Test: Attack Vs. Defense Accuracy

	Test Value = 0					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
Attack Accuracy	504.730	9	.000	92.518	92.10	92.93
Defense Accuracy	103.219	9	.000	87.690	85.77	89.61

Both attack accuracy and defense accuracy have very low p-values (p = 0.000), indicating that the sample means for both variables are significantly different from the test value of 0. The 95% confidence intervals further support this conclusion, as they do not include 0. This suggests that both attack and defense accuracy are significantly greater than 0.

4.4.3 Confusion Matrix

Table 8 shows the confusion matrix.

Table 8: Confusion Matrix

Aspect	Score
Accuracy	0.6748
Precision	0.873367
Recall	0.765664
F1 Score	0.815962

The achieved accuracy of 67.48% indicates that the model correctly predicts the class labels for approximately two-thirds of the instances. While moderate, this accuracy needs to be considered in the context of adversarial attacks, where model robustness is of primary concern. The high precision value of 87.34% is a notable strength. It signifies that when the model classifies an instance as adversarial, it is correct about 87.34% of the time. This is particularly crucial in applications where misclassifying an adversarial example as benign could have severe consequences. The recall of 76.56% implies that the model successfully identifies approximately

three-quarters of the actual positive instances, i.e., the adversarial examples in the dataset. While a respectable figure, the tradeoff with precision suggests there might be instances where actual adversarial examples are missed. The F1 score, being the harmonic mean of precision and recall, provides a balanced assessment. The achieved F1 score of 81.60% suggests a good compromise between precision and recall. It serves as an overall indicator of the model's performance against the Fast Gradient Sign Method (FGSM) attack.

The set of metrics collectively reveals the model's performance against adversarial attacks. The high precision reflects a strong capability to accurately classify adversarial instances, while the balance with recall offers a nuanced assessment of the model's overall effectiveness. The trade-off between precision and recall is a common challenge in machine learning, and it is apparent in this study. While precision is high, indicating confidence in the model's predictions, there is a trade-off with recall, suggesting a potential need for improvements in capturing a larger portion of actual adversarial instances. Overall, the achieved metrics collectively indicate a reasonable level of robustness against the FGSM attack. High precision is reassuring, especially in scenarios where false positives can have severe consequences. Further investigation into techniques that can enhance recall without sacrificing precision may be warranted. The sensitivity analysis on the parameters of the FGSM attack provides valuable insights into how the model responds to variations in attack strength and iterations. Understanding these sensitivities is crucial for refining and optimizing the defense mechanisms, including the block-switching algorithm

5.0 Summary of Study Findings

The analysis employed the FGSM attack to generate adversarial examples, with varying values of epsilon and iteration numbers. The primary focus is on the relationship between the number of images, attack and defense models, and their corresponding accuracy. The attack model's accuracy consistently increased as the number of images rose from 80 to 240, demonstrating a positive correlation. Figure 1 illustrates the relationship between the number of images and the accuracy of the attack model, reinforcing the observation that a higher number of images generally leads to improved accuracy. Successful attacks grew in number as the accuracy of the attack model increased, indicating a positive correlation between model accuracy and the number of successful attacks.

The defense model's accuracy remained relatively constant at 83.96% to 90.96% as the number of images increased from 80 to 412. This suggested that, in this dataset, an increase in the number of images did not necessarily lead to higher defense model accuracy. Figure 3 visually depicts the relationship between the number of images and defense model accuracy. The defense model's accuracy exhibited consistency at 83.96% across all data points, indicating that its performance did not significantly vary with changes in the number of successful defenses. Notably, the number of successful defense instances increased progressively, from 71 to 357, as we moved through the dataset. This signifies an enhancement in the defense model's effectiveness against attacks.

6.0 Conclusions

The study concludes that the number of images used in the dataset significantly impacts the accuracy of both the attack and defense models. Increasing the number of images beyond a certain point does not necessarily lead to a significant improvement in model accuracy; in some cases, it can even result in a mild decrease in performance. The accuracy of the defense model remains

relatively stable across different data points, indicating its consistency in handling attacks. Additionally, there is a positive relationship between the accuracy of the attack model and the number of successful attacks, up to a certain threshold. Moreover, the dataset exhibits certain variability and influences beyond the number of images, which can affect model accuracy.

7.0 Recommendations

The study recommends underscoring the importance of ensuring the integrity and robustness of machine learning models in cybersecurity. Policymakers should consider incorporating guidelines that encourage the evaluation and enhancement of both attack and defense models to adapt to evolving threats. Organizations utilizing machine learning models in their cybersecurity strategies should be aware of the critical role played by the quantity and quality of training data. Ensuring diverse and substantial datasets is essential for maintaining model accuracy.

The study further recommends exploring the block-switching algorithm as it offers a secured data encryption approach compared to traditional techniques. By using high-level encryption methods, organizations can effectively safeguard data against cyber-attacks and unauthorized access. Additionally, a strong defensive system enhances the reliability of an organization's cybersecurity measures. By establishing a system that's less susceptible to breaches or attacks, organizations can maintain the trust of their customers and stakeholders while minimizing the risk of harm. Consequently, numerous industries face requirements concerning data privacy and security. Through the implementation of a system utilizing the block-switching algorithm, organizations can ensure compliance with these regulations, avoiding penalties and legal liabilities. In the present era, cybersecurity holds increasing significance for businesses. Therefore, by adopting a system incorporating the block-switching algorithm, organizations raise their bars by showing commitment to cybersecurity and consumer data security.

REFERENCES

Afzal, S., Yousaf, M., Afzal, H., Alharbe, N., & Mufti, M. R. (2020). Cryptographic Strength Evaluation of Key Schedule Algorithms. Security and Communication Networks. https://doi.org/10.1155/2020/3189601

Alekseev, E., & Bozhko, A. (2020). Algorithms for switching between block-wise and arithmetic masking. https://eprint.iacr.org/2022/1624.pdf

Ali, K., Qureshi, A. N., Bhatti, M., Sohail, & Hijji, M. (2022). Defending Adversarial Examples by a Clipped Residual U-Net Model. Intelligent Automation & Soft Computing, DOI: 10.32604/iasc.2023.028810

Andrade, C. (2020). Understanding the Difference Between Standard Deviation and Standard Error of the Mean, and Knowing When to Use Which. Indian Journal of Psychological Medicine, 42(4), 409–410. https://doi.org/10.1177/0253717620933419

Anthi, E., Williams, L., Rhode, M., Burnap, P., & Wedgbury, A. (2021). Adversarial attacks on machine learning cybersecurity defences in Industrial Control Systems. Journal of Information Security and Applications. doi: https://doi.org/10.1016/j.jisa.2020.

Athalye, A., Engstrom, L., Ilyas, A., & Kwok, K. (2018). Synthesizing robust adversarial examples. In ICML.

Bai, J., Gao, K., Gong, D., Xia, S., Li, Z., & Liu W. (2022). Hardly Perceptible Trojan Attack against Neural Networks with Bit Flips https://arxiv.org/abs/2207.13417

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8), 1798-1828. https://doi.org/10.1109/TPAMI.2013.50

Bhoge, J. P., & Chatur, P. N. (2014). Avalanche Effect of AES Algorithm. International Journal of Computer Science and Information Technologies, 5(3), 3101-3103. https://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.659.9331&rep=rep1&type=pdf

Catak, F. O., & Yayilgan, S. Y. (2021). Deep Neural Network Based Malicious Network Activity Detection Under Adversarial Machine Learning Attacks. Communications in Computer and Information Science - Springer Science and Business Media Deutschland GmbH., 1382, 280-291. doi:https://doi.org/10.1007/978-3-030-71711-7_23

Chan, P. P. K., He, Z. M., Li, H., & Hsu, C. C. (2018). Data sanitization against adversarial label contamination based on data complexity. International Journal of Machine Learning and Cybernetics, 9(6), 1039–1052. doi:https://doi.org/10.1007/s13042-016-0629

Chen, J., Jordan, M.I., & Wainwright, M. J. (2020). HopSkipJumpAttack: a query-efficient decision-based attack. In 2020 IEEE symposium on security and privacy (pp. 1277-1294). IEEE

Deng, J., Berg, A. C., Fei-Fei, L., Hutchison, D., Kanade, T., Kittler, J., Kleinberg, J. M., Mattern, F., Mitchell, J. C., Naor, M., Nierstrasz, O., Pandu Rangan, C., Steffen, B., Sudan, M., Terzopoulos, D., Tygar, D., Vardi, M. Y., Weikum, G., & Li, K. (2010). What Does Classifying More Than 10,000 Image Categories Tell Us? In K. Daniilidis, P. Maragos, & N. Paragios (Eds.), Computer Vision – ECCV 2010 (Vol. 6315, pp. 71–84). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-15555-0_6

Dolmatov, V., Baryshkov, D. (2020). Block Cipher "Magma", RFC 8891. https://doi.org/10.17487/RFC8891

Dong, Y., Su, H., Wu, B., Li, Z., Liu, W., Zhang, T., & Zhu, J (2019). Efficient decision-based black-box adversarial attacks on face recognition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 7714-7722)

Dunn, C., Moustafa, N., & Turnbull, B. (2020, August 10). Robustness Evaluations of Sustainable Machine Learning Models against Data Poisoning Attacks in the Internet of Things. Sustainability, 12(16), 17. doi:http://dx.doi.org/10.3390/su12166434

Echeverri, C. (2017). Visualization of the Avalanche Effect in CT2. Doctoral dissertation, University of Mannheim. https://www.cryptool.org/assets/ctp/documents/BA_Echeverri.pdf

El Omda, S., & Sergent, S. R. (2023). Standard Deviation. In StatPearls. StatPearls Publishing. http://www.ncbi.nlm.nih.gov/books/NBK574574/

Goodfellow, I.J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. In Proceedings of the International Conference on Learning Representation ICLR, San Diego, CA, USA.

Grimmett, G. R., & Stirzaker, D. R. (1992). Probability and Random Processes, second edition. Oxford University Press.

Hassan, M. R. (2021, June 15). A Robust Deep-Learning-Enabled Trust-Boundary Protection for Adversarial Industrial IoT Environment. IEEE Internet of Things Journal, 8(12). doi:doi: 10.1109/JIOT.2020.3019225.

He, Z., Li, J., & Li, (2019). An Improved Block Switching Method for Image Compression. In Proceedings of the 2019 IEEE International Conference on Communications, Control, and Computing Technologies for Smart Grids, Chengdu, China.

Hsu, C.-C., Zhuang, Y.-X., and Lee, C.-Y. (2020). Deep fake image detection based on pairwise learning. Applied Sciences, 10(1), 370

Hu, S., & Cao, Y. (2018). A New Block-Switching Method for Video Compression. In Proceedings of the 2018 IEEE International Conference on Signal Processing, Communications, and Computing, Guangzhou, China, 1-5.

Hutter, M., Tunstall, M. (2019). Constant-time higher-order Boolean-to-arithmetic masking. Journal of Cryptographic Engineering, 9, 173–184. https://doi.org/10.1007/s13389-018-0191-z

Janiesch, C., Zschech, P., & Heinrich, K. (2021). Machine learning and deep learning. Electronic Markets, 31(3), 685–695. doi:https://doi.org/10.1007/s12525-021-00475-2

Killmann,W., & Schindler, W. (2001). AIS 31:A proposal for Functionality classes and evaluation methodology for true (physical)random number generators, Version 3.1[J].Bundesamt fiir Sicherheit in der Informationstechnik(BSI),

Kurakin, A., Goodfellow, I., & Bengio, S. (2016). Adversarial examples in the physical world. arXiv. arXiv:1607.02533.

Kuzlu, M., Fair, C., & Guler, O. (2021). Role of Artificial Intelligence in the Internet of Things (IoT) Cybersecurity. . Discover the Internet of Things, 1(1). doi:https://doi.org/10.1007/s43926-020-00001-4

Kwon, H., Kim, Y., Yoon, H., & Choi, D. (2021). Classification score approach for detecting adversarexamplesmple in deep nenetworkstwork. Multimed Tools App,l 80(7), 10339–10360.

Liao, Q., Zhong, Z., Zhang, Y., Xie, C., & Pu, S. (2018). Defense against adversarial attacks using high-level representation guided denoiser. In Proceedings of the European Conference on Computer Vision (ECCV), 489-504.

Liu, Y., Shi, X., & Chen, J. (2016). An Improved Block-Switching Method for H.264/AVC. Proceedings of the 2016 IEEE International Conference on Information and Automation, Ningbo, China, 791-796.

Liang, M., Chang, Z., Wan, Z., Gan, Y., Schlangen, E., & Šavija, B. (2022, January). Interpretable Ensemble-Machine-Learning models for predicting creep behavior of concrete. Cement and Concrete Composites, 125(104295), 17. doi:https://doi.org/10.1016/j.cemconcomp.2021.104295

Luo, Z. Z. (2020, July 13). Adversarial machine learning based partial-model attack in IoT. Proceedings of the 2nd ACM Workshop on Wireless Security and Machine Learning, 13-18. doi:https://doi.org/10.1145/3395352.3402619

Mao, C., Gupta, A., Nitin, V., Ray, B., Song, S., Yang, J., & Vondrick, C. (2020). Multitask learning strengthens adversarial robustness. In: European conference on computer vision, 16th European conference, GlasgAugust, August 23–28, in lecture notes in computer science, vol 12347. Springer, Cham, 158–174.

Manikandan, S. (2011). Measures of central tendency: Median and mode. Journal of Pharmacology & Pharmacotherapeutics, 2(3), 214–215. https://doi.org/10.4103/0976-500X.83300

Martinez, E. E. B., Oh, B., Li, F., & Luo, X. (2019). Evading Deep Neural Network and Random Forest Classifiers by Generating Adversarial Samples. In Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11358, 143-155. doi:https://doi.org/10.1007/978-3-030-18419-3_10

Merenda M, Porcaro C, & Lero D. (2020, April 29). Edge Machine Learning for AI-Enabled IoT Devices: A Review. Sensors (Basel), 20(9), 34. doi:doi: 10.3390/s20092533. PMID: 32365645; PMCID: PMC7273223

Mingkang Z., Tianlong, C., & Wang, Z. (2021). Sparse and imperceptible adversarial attack via a homotopy algorithm. arXiv preprint arXiv:2106.06027

Moosavi-Dezfooli, S.M., Fawzi, A., & Frossard, P. (2016). Deepfool: A simple and accurate method to fool deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2574–2582.

Neil, C., Greenewald, K.,, Lee, K., & Manso, G. F. (2020). The computational limits of deep learning initiative on the digital economy research brief. doi:https://doi.org/10.48550/arXiv.2007.05558

Oprea, A. (2021). Machine Learning Integrity and Privacy in Adversarial Environments. 1–2. https://doi.org/10.1145/3450569.3462164

Paje, R. E. J., Sison, A. M., & Medina, R. P. (2019). Multidimensional key RC6 algorithm, in Proceedings of the 3rd International Conference on Cryptography, Security and Privacy—ICCSP'19, pp. 33–38, Kuala Lumpur, Malaysia.

Preneel, B. (2000). "NESSIE project," in Encyclopedia of Cryptography and Security. Springer, Berlin, Germany.

Qiu, S., Liu, Q., Zhou, S., & Wu, C. (2019). Review of artificial intelligence adversarial attack and defense technologies. Appl Sci, 9(5), 909.

Ramanujam, S., & Karuppiah, M. (2011). Designing an algorithm with a high Avalanche Effect. IJCSNS International Journal of Computer Science and Network Security, 11(1), 106-111. http://paper.ijcsns.org/07_book/201101/20110116.pdf.

Sharif, M. Bhagavatula, S. Bauer, L., & Reiter, M. K. J. (2017). Adversarial generative nets: Neural network attacks on state-of-the-art face recognition. arXiv preprint arXiv:1801.00349, 2 (3).

Shi, H., Deng, Y., & Guan, Y. (2011). Analysis of the avalanche effect of the AES S box. In 2011 2nd International Conference on Artificial Intelligence, Management Science and Electronic Commerce (AIMSEC) (pp. 5425-5428). IEEE. https://doi.org/10.1109/AIMSEC.2011.6009935

Simion, E. (2015). The relevance of statistical tests in cryptography. IEEE Security & Privacy, 13 (1), 66–70.

Sulak, F., Doğanaksoy, A., Ege, B., et al. (2010). Evaluation of randomness test results for short sequences[C]//International Conference on Sequences and Their Applications. Springer, Berlin, Heidelberg, 309-319.

Taheri, S., Khormali, A., Salem, M., & Yuan, J. (2020). Developing a Robust Defensive System against Adversarial Examples Using Generative Adversarial Networks . Big Data Cogn. Comput., 4(2), 11. https://doi.org/10.3390/bdcc4020011

.

Taori, R., Kamsetty, A., Chu, B., & Vemuri, N. (2018). Targeted adversarial examples for black box audio systems. arXiv preprint arXiv:1805.07820.

Thacker, J. (2020). The Age of AI: Artificial Intelligence and the Future of Humanity. Zondervan.

Ukrop, M. (2016). Randomness analysis in authenticated encryption systems," Masarykovauniverzita, Fakultainformatiky, Brno, Czechia,, Ph.D. thesis.

Vinayakumar, R., Alazab, M., Srinivasan, S., Pham, Q.V., Padannayil, S.K., & Simran, K. (2020). A Visualized Botnet Detection System based Deep Learning for the Internet of Things Networks of Smart Cities. IEEE Trans. Ind. Appl.

Wang, X., Wang, S., Chen, P. U., Wang, Y., Kulis, B., Lin, X., & Chin, P. (2020). Protecting Neural Networks with Hierarchical Random Switching: Towards Better Robustness-Accuracy Trade-off for Stochastic Defenses. Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19).

Xie, Y., Li, Z., Shi, C., Liu, J., Chen, Y., & Yuan, B. (2021). Real-time, Robust and Adaptive Universal Adversarial Attacks Against Speaker Recognition Systems. Journal of Signal Processing Systems, 93(10), 1187–1200. https://doi.org/10.1007/s11265-020-01629-9.