Adversarial Attacks on Intel and MobileODT Cervical Cancer Screening

Nikolaos Bakalis

Advisor: Dr. Raj Shukla

Anglia Ruskin University

Acknowledgements

I would like to express my deepest gratitude to my advisor, Dr. Raj Shukla, for his invaluable guidance, support, and mentorship throughout the course of this research. Your insights and wisdom have been immensely helpful.

I am also grateful to Anglia Ruskin University for providing the resources and environment to conduct this research.

Lastly, I would like to thank my family and friends for their unwavering support and encouragement throughout this academic endeavor.

Contents

Abstract

This thesis explores the vulnerabilities and robustness of machine learning models, particularly in the context of adversarial attacks. Utilizing the Cervical Cancer Screening Dataset, various adversarial attack techniques such as FGSM, RPA, GNA, and BIM were implemented to assess their impact on machine learning models. The study also delves into defense mechanisms, focusing on ensemble learning as a promising strategy. The effectiveness of these attacks and defenses was evaluated through a comparative analysis of model accuracy on clean and attacked data. This research contributes to the ongoing efforts to secure machine learning systems against adversarial threats.

Introduction

The Rise of Machine Learning

In the last few decades, machine learning has emerged as a transformative technology, profoundly impacting a wide array of disciplines (Brookings, 2018). It has revolutionized healthcare by aiding in the early detection of diseases and personalized treatment plans (PMC, 2020). In finance, machine learning algorithms are employed for risk assessment, fraud detection, and automated trading. The entertainment industry leverages recommendation systems to personalize user experiences, while in the realm of cybersecurity, machine learning models are used to identify threats, anomalies, and vulnerabilities in real-time (McKinsey, 2018).

As machine learning algorithms have become increasingly sophisticated, they have opened up new avenues for innovation and efficiency (McKinsey, 2022). However, this sophistication is a double-edged sword. Advanced machine learning models are not only capable of solving complex problems but are also susceptible to equally complex forms of deception and manipulation (Belfer Center, 2021). This has led to new challenges in maintaining the integrity and reliability of these systems.

This conundrum has given birth to a new interdisciplinary field known as adversarial machine learning, situated at the crucial intersection of machine learning and cybersecurity (Belfer Center, 2021). Adversarial machine learning aims to understand the vulnerabilities inherent in machine learning models and to develop strategies for defending against malicious attacks (McKinsey, 2022). These attacks often involve subtle, carefully crafted perturbations to the input data, designed to mislead machine learning algorithms without detection (Belfer Center, 2021).

Adversarial Attacks and Defenses

One of the critical application areas of machine learning is in defending against adversarial attacks (Arxiv, 2023) (Belfer Center, 2021). These attacks often involve subtle, carefully crafted perturbations to the input data, designed to mislead machine learning algorithms without detection (IEEExplore, 2022). Traditional methods often fall short in the face of these advanced, adaptive forms of attacks. Therefore, the field of adversarial machine learning aims to improve the robustness and adaptability of machine learning systems by understanding their vulnerabilities and developing strategies for defense (Arxiv, 2023).

The Purpose of This Thesis

1. **In-depth Analysis of Adversarial Attacks:** This thesis provides a comprehensive exploration of the nature and types of adversarial attacks that machine learning models, particularly those used in malware detection, are susceptible to (Arxiv, 2023).
2. **Cervical Screening Dataset as a Case Study:** Utilizing the Cervical Cancer Screening Dataset, this thesis serves as a practical examination of how machine learning models can be susceptible to adversarial attacks in the healthcare domain. It offers insights

into the impact of such attacks on medical diagnostics and underscores the need for robust defense mechanisms in healthcare-related machine learning applications (MDPI, 2023).

3. **Critical Evaluation of Existing Defense Mechanisms:** The work critically assesses the limitations and vulnerabilities of current defense mechanisms against adversarial attacks in the context of malware detection systems (Arxiv, 2023).
4. **Introduction of Ensemble Learning as a Defense Strategy:** The thesis introduces and explores the application of ensemble learning as a potential defense mechanism against adversarial attacks, specifically in malware detection systems (Arxiv, 2023).
5. **Empirical Assessment of Ensemble Learning Framework:** The thesis includes a detailed examination and empirical assessment of an ensemble learning framework designed for adversarial malware defense, evaluating its effectiveness and limitations (Arxiv, 2023).
6. **Contribution to Cybersecurity Infrastructure:** By focusing on the vulnerabilities and defense mechanisms in machine learning models used in cybersecurity, the thesis aims to enhance the reliability and integrity of malware detection systems (MDPI, 2023).
7. **Guidance for Future Research:** The findings and insights from this research serve as a foundational resource that could guide future work in enhancing machine learning-based cybersecurity systems against adversarial threats (Arxiv, 2023).

Promotion of Robust and Resilient Systems: Ultimately, the thesis aims to contribute to the development of more robust and resilient cybersecurity infrastructure by shedding light on the vulnerabilities and potential defense mechanisms against adversarial attacks in machine learning models.

Ethics

Literature Review

The Concern for Robust Machine Learning: Adversarial machine learning has become an increasingly important field of study, particularly in the context of AI Security and Trustworthy AI (OpenAI, 2017). It aims to develop and scrutinize machine learning algorithms that are designed to be resilient against manipulative attacks. The focus is not just on building robust algorithms, but also on understanding the vulnerabilities that make existing systems susceptible to adversarial inputs (Arxiv, 2023) (Belfer Center, 2021).

Understanding Adversarial Manipulation: The concept of adversarial manipulation involves introducing slight, often imperceptible, changes to the input data that can lead machine learning models astray, causing them to make incorrect predictions or classifications (PMC, 2022). These deceptive changes are usually not random; they are finely-tuned manipulations created with a keen understanding of the model's internal workings.

This kind of manipulation exploits the features that the machine learning model has been trained to recognize, essentially turning the model's intelligence against itself (PMC, 2022).

The Complexity of Model Exploitation

Adversarial manipulations are not just simple tweaks to the data; they are calculated moves that take advantage of the model's learned behavior. Typically, these manipulations are created with detailed knowledge of the model's architecture and the features it uses for making predictions (Arxiv, 2023). This makes defending against such attacks particularly challenging, as the manipulations are tailored to exploit the specific machine learning model in question (PMC, 2022).

Adversarial Machine Learning

The advent of machine learning has brought about significant transformations across a multitude of sectors including healthcare, where it aids in diagnostics and treatment plans (PMC, 2020); finance, where it optimizes trading algorithms and identifies fraudulent activities (McKinsey, 2022); entertainment, where it powers recommendation engines; and not least, cybersecurity, where it helps in the identification and mitigation of threats in real-time. However, as these machine learning systems have grown more advanced and intricate, the techniques to deceive or manipulate them have also evolved in complexity. This phenomenon has led to the birth of a new interdisciplinary domain called adversarial machine learning, which resides at the crucial juncture of machine learning and AI Security and Trustworthy AI.

Adversarial machine learning focuses on the creation, evaluation, and fortification of machine learning algorithms that are designed to be resilient to deceptive and malicious manipulations. The term "adversarial manipulation" generally encompasses slight, carefully engineered alterations made to the input data that feed into machine learning models. The objective of these alterations is to trick the model into making incorrect predictions or misclassifications. Interestingly, these manipulations are not arbitrary; they are made by entities with a deep understanding of how the model functions and are fine-tuned to exploit the specific features that the model has learned to recognize. In essence, these manipulations weaponize the model's own learned intelligence against it.

The origins of adversarial machine learning can be traced back to the nascent phases of the machine learning discipline. Early researchers and practitioners found that their models, although highly effective on their training and validation sets, often faltered when exposed to data that were even slightly modified or contained noise. This was a wake-up call for the machine learning community, as it unveiled a significant shortcoming in these models: they excelled at identifying underlying patterns and correlations in their training data but struggled to generalize this understanding to new, subtly different, or altered data.

Over the years, adversarial machine learning has matured from being a specialized, somewhat obscure area of study to becoming a discipline of significant academic and practical importance. This shift is particularly noticeable in fields where the integrity,

reliability, and robustness of machine learning models are of utmost importance, such as in autonomous vehicles, healthcare systems, and financial algorithms. As machine learning becomes increasingly integrated into the fabric of daily life and critical systems, the urgency to comprehend, thwart, and mitigate adversarial attacks escalates correspondingly. In light of this, the core objective of this thesis is to make a meaningful contribution to the rapidly expanding body of knowledge in adversarial machine learning. Specifically, this work aims to investigate and develop innovative strategies for defending machine learning systems against adversarial attacks, with a focus on applications in malware detection.

Malware Detection Systems

Malware detection systems serve as pivotal elements in the architecture of cybersecurity, designed with the primary aim of detecting, isolating, and ultimately neutralizing malicious software, also known as malware (Emerald, 2023) (ScienceDirect, 2022). These systems are indispensable in protecting an array of digital assets—ranging from individual personal computers and mobile devices to extensive enterprise networks and cloud infrastructures—against a multitude of cyber threats (McKinsey, 2019). These threats take various forms, including but not limited to, viruses that corrupt or delete data, worms that self-replicate across networks, trojans that provide backdoor access to systems, ransomware that encrypts files demanding payment for their release, and spyware that clandestinely monitors user activity.

In the early days of cybersecurity, malware detection systems largely depended on signature-based detection mechanisms (ScienceDirect, 2022). This approach involved scanning the code within potentially harmful programs and comparing it against a predefined database of known malware signatures or patterns. While this method was quite effective in identifying and neutralizing malware that had already been cataloged, it had inherent limitations. Specifically, signature-based methods were often ineffectual against new strains of malware or versions that had been modified to evade detection. To overcome these shortcomings, the cybersecurity community turned towards heuristic-based detection methods. Unlike their signature-based counterparts, heuristic methods do not seek exact code matches. Instead, they analyze the general behavior, characteristics, and code patterns of a program to make an informed judgment on its malicious intent.

The integration of machine learning into malware detection represents a significant paradigm shift and has imbued these systems with enhanced capabilities (Emerald, 2023). Machine learning algorithms are trained to identify complex patterns and correlations within expansive datasets, which empowers them to detect a broader range of malware types (ScienceDirect, 2022). These intelligent systems are not confined to recognizing only known malware but are also proficient at identifying new or modified versions by understanding the underlying behaviors and characteristics that are common to malicious software. By leveraging machine learning, modern malware detection systems can adapt and evolve, thereby staying one step ahead of cybercriminals who continuously develop sophisticated forms of malware (McKinsey, 2019).

The Nature of Adversarial Attacks

Adversarial attacks in the context of machine learning are not random disruptions but deliberate, meticulously planned manipulations of the input data aimed at deceiving machine learning algorithms (Arxiv, 2023). These attacks take advantage of the vulnerabilities inherently present in the learning and decision-making processes of the models. The alterations made to the input data are generally subtle and meticulously crafted, making them almost imperceptible to human observers. Despite their subtlety, these slight changes can have a disproportionately large effect on the model's predictions or classifications, leading it to make grave errors (NCBI, 2020).

The motivations and objectives behind adversarial attacks can vary, leading to different types of attacks based on their intent. Exploratory attacks are those that focus on exploiting the existing vulnerabilities in a trained model. The primary aim is often to mislead the model into making incorrect classifications, which could be anything from mislabeling an image to incorrectly identifying a benign file as malicious. Causative attacks are more insidious, as they aim to corrupt the learning process itself. These attacks typically involve introducing a bias or skew in the training data, which can lead the model to make systematic errors when making predictions or classifications in the future (NCBI, 2020).

The level of knowledge an attacker possesses about the targeted model can also serve as a basis for categorizing adversarial attacks. In a white-box attack scenario, the attacker has comprehensive knowledge of the model's architecture, its parameters, and even the data used for training. This extensive knowledge allows for more targeted and effective attacks. Conversely, in a black-box attack scenario, the attacker has no insight into the model's internal mechanisms and can only interact with its inputs and outputs. Despite this limitation, black-box attacks can still be highly effective, particularly if the attacker has a good understanding of machine learning algorithms and their general vulnerabilities.

The threat posed by adversarial attacks is particularly acute in applications where the stakes are high and the margin for error is low (Europarl, 2020). This includes critical systems like cybersecurity measures protecting sensitive data, autonomous driving systems where errors could result in life-threatening situations, and healthcare systems where incorrect diagnoses could have severe consequences. Understanding the various forms, techniques, and impacts of adversarial attacks is crucial for developing robust defense mechanisms. In alignment with this need, this thesis aims to provide an in-depth exploration of the complex landscape of adversarial attacks. It will dissect their types, methodologies, and impacts, focusing specifically on how they compromise machine learning models employed in the realm of malware detection.

How Adversarial Attacks Work

Adversarial attacks operate by taking advantage of the inherent weaknesses in machine learning models (NCBI, 2021) (Arxiv, 2022). One such vulnerability lies in the high-dimensional, non-linear decision boundaries that these models employ to segregate and classify data. While these decision boundaries are generally effective for the majority of

standard inputs, they are susceptible to manipulation when faced with specially crafted adversarial inputs. These boundaries are formulated based on the model's training data, and although they are highly effective under regular conditions, they can be deceived to misclassify inputs under adversarial conditions.

The life cycle of an adversarial attack is often a multi-step process, each with its own complexities and nuances (Arxiv, 2022):

1. Understanding the Model: The first step in executing an adversarial attack involves gaining an understanding of the targeted machine learning model. The depth of this understanding can vary depending on the type of attack. In a white-box attack, the attacker has full visibility into the model's architecture, the parameters used, and even the training data. In contrast, a black-box attack only provides the attacker with access to the model's inputs and outputs, requiring them to make educated guesses or employ trial and error to exploit the model's vulnerabilities (Europarl, 2020).
2. Creating Adversarial Examples: Once the attacker has sufficient knowledge of the model, they proceed to create what are known as adversarial examples. These are strategically altered versions of legitimate inputs that the model is expected to handle. The changes made to these inputs are usually subtle and almost imperceptible to a human observer. However, these slight modifications are potent enough to trick the model into making an incorrect classification. The alterations are generally computed using the gradient of the model's loss function with respect to the input data. This gradient serves as a guide, showing how tiny changes to the input can result in significant shifts in the model's output, thereby allowing the attacker to create targeted and effective adversarial examples (ScienceDirect, 2021).
3. Deploying the Attack: The final step involves feeding these crafted adversarial examples into the targeted machine learning model. Because these examples are very close to legitimate inputs in the model's input space, the model is often deceived into misclassifying them as genuine, thereby falling into the trap set by the attacker (NCBI, 2021).

The effectiveness of adversarial attacks cannot be understated. They have the potential to deceive even the most well-trained and sophisticated machine learning models, inducing them to make critical errors (Europarl, 2020). This poses a formidable challenge to the deployment of machine learning algorithms in applications where security is paramount. Given these risks, the development and implementation of robust defensive mechanisms against adversarial attacks have become an area of high priority in the machine learning and cybersecurity communities (Europarl, 2019).

Types and Techniques

Adversarial attacks can be organized into various categories based on a range of differentiating criteria (Arxiv, 2023). These criteria could include how much the attacker knows about the machine learning model they are targeting, which stage of the machine learning process the attack is aimed at, and the specific methods or algorithms used to create

the adversarial examples. By understanding these classifications, we can gain a clearer picture of the potential vulnerabilities that machine learning models may face (Europarl, 2019) (NCBI, 2021).

1. Based on the Attacker's Knowledge: The level of information an attacker has about a machine learning model can significantly influence the effectiveness of the adversarial attack. In a white-box attack scenario, the attacker has extensive knowledge about the model, including its architecture, parameters, and even the dataset on which it was trained. This comprehensive understanding allows for the crafting of highly effective and targeted adversarial examples. On the other end of the spectrum is the black-box attack, where the attacker has no internal knowledge about the model. Despite this apparent disadvantage, black-box attacks can still be highly effective. One common technique employed in black-box attacks is transferability, where adversarial examples generated for one model are used to compromise another model with similar characteristics (NCBI, 2020).
2. Based on the Targeted Stage: The stage of the machine learning process that an attack targets can also be a basis for classification. Exploratory attacks are generally aimed at the testing or deployment phases of a machine learning model. The main goal of these attacks is to exploit existing weaknesses in a trained model, often with the objective of causing it to misclassify new data. Causative attacks are different; they aim to corrupt the model at the training stage. By manipulating the training data, these attacks introduce biases that can lead the machine learning model to make systematic errors when it is eventually deployed (NCBI, 2021).
3. Based on the Techniques Used: The landscape of adversarial attacks includes a variety of techniques for generating adversarial examples, each with its unique strengths and weaknesses (Arxiv, 2023). Among the most commonly employed techniques are:
    a. Fast Gradient Sign Method (FGSM): This technique leverages the gradients of the model's loss function concerning the input data to generate adversarial examples quickly.
    b. Jacobian-based Saliency Map Attack (JSMA): This method involves calculating the forward derivative of the model to pinpoint the most impactful features. These features are then modified to mislead the model into making a wrong classification.
    c. DeepFool: An iterative method, DeepFool makes incremental changes to the input data until it crosses the model's decision boundary, causing a misclassification.
    d. Carlini & Wagner (C&W) Attack: This method approaches the creation of adversarial examples as an optimization problem. It aims to find the minimum alteration needed to the input data that would lead to a misclassification.

Understanding the various types, stages, and techniques of adversarial attacks is pivotal for the development of robust and effective defense mechanisms (Europarl, 2019). This thesis

aims to contribute to this understanding by offering an in-depth and comprehensive examination of the myriad forms and methodologies of adversarial attacks (Arxiv, 2023).

Methodology

In undertaking this research, a systematic methodology was employed to probe the vulnerabilities of machine learning models in the context of cervical cancer screening. Specifically, four distinct types of adversarial attacks were utilized to challenge the model's performance: the Fast Gradient Sign Method (FGSM), Random Perturbation Attack (RPA), Gaussian Noise Attack (GNA), and the Basic Iterative Method (BIM). Each of these attacks serves to test different aspects of the model's robustness and resilience against adversarial manipulations, thereby providing a comprehensive evaluation of its vulnerabilities.

To carry out these attacks, the Cervical Cancer Screening Dataset was used as the foundation for the experiments. This dataset was chosen due to its relevance in the healthcare sector and the critical importance of maintaining the integrity and accuracy of machine learning models in such high-stakes applications. The dataset comprises a range of features, including medical histories and diagnostic test results, that the model uses for classification. The idea was to see how each of these adversarial attacks could potentially mislead the model into making erroneous classifications, which in a real-world scenario could lead to misdiagnoses or incorrect treatments.

For the technical implementation of these attacks, PyTorch was employed as the primary tool. PyTorch is an open-source machine learning library that offers a range of features conducive to both the development and evaluation of machine learning models. Its flexibility and ease of use make it a popular choice among researchers and professionals in the machine learning community. The library includes built-in functions and modules that facilitate the efficient implementation of adversarial attacks, allowing for a rigorous and methodologically sound investigation.

By combining these attacks and tools, this research aims to provide a nuanced and comprehensive examination of how machine learning models can be compromised. The insights gleaned from this work will contribute to the ongoing efforts to develop more robust and secure machine learning algorithms, particularly in applications with significant real-world implications like healthcare.

Introduction to the Cervical Cancer Screening Dataset

The Cervical Cancer Screening Dataset serves as a central component in the field of healthcare-focused machine learning research. Hosted on and developed for the Intel & MobileODT Cervical Cancer Screening competition, the dataset aims to accelerate advancements in the application of machine learning for cervical cancer diagnostics and treatment optimization (ScienceDirect, 2021). The overarching goal is to identify more effective cancer treatments and thereby contribute to solving one of the significant healthcare challenges on a global scale (WHO, 2022).

Cervical cancer remains a pressing global health concern, responsible for a considerable number of deaths, particularly in low and middle-income countries (WHO, 2022). The impact of early detection through effective screening methods cannot be overstated, as timely intervention can dramatically alter the course of the disease, improving both survival rates and quality of life for patients (NCBI, 2020). To this end, the dataset is enriched with a multifaceted array of data points pertinent to cervical cancer screening (ScienceDirect, 2021). It serves as an invaluable asset for researchers and healthcare professionals looking to delve into the complexities and challenges associated with cervical cancer detection and treatment (NCBI, 2020).

While I faced technical barriers in accessing detailed specifics about the dataset due to issues with the Kaggle website, it's worth noting that datasets designed for such competitions often encompass an extensive range of features (ScienceDirect, 2021). These could include patient demographics like age and ethnicity, comprehensive medical histories, and a multitude of results from diagnostic tests such as Pap smears, HPV tests, and biopsies (NCBI, 2020). Such datasets usually also include outcome labels that signify the presence or absence of cervical cancer (ScienceDirect, 2021). These labels serve as the target variables for supervised machine learning tasks, providing a ground truth against which model predictions can be compared (NCBI, 2020).

In the context of this research, the Cervical Cancer Screening Dataset serves as a critical testing ground for evaluating the efficacy and robustness of various adversarial attacks and their corresponding defense mechanisms (ScienceDirect, 2021). The application of these adversarial techniques to a dataset with such real-world significance allows for a nuanced understanding of their practical implications (NCBI, 2020). It also provides an opportunity to assess how these techniques can contribute to enhancing the security and reliability of machine learning models, particularly in high-stakes fields like healthcare where the margin for error is exceedingly small (WHO, 2022).

Implementation of FGSM, RPA, GNA, and BIM Attacks

In the scope of this research, a focused approach was adopted to study the vulnerabilities of machine learning models in healthcare, particularly concerning cervical cancer screening. Four distinct adversarial attacks were meticulously implemented and tested on the Cervical Cancer Screening Dataset. The entire implementation was executed using PyTorch, a widely-used open-source machine learning library that offers extensive capabilities for developing and evaluating complex models. PyTorch was chosen for its flexibility and strong community support, which includes a plethora of pre-coded modules and functions that significantly expedite the model development process.

Before implementing the attacks, the dataset was pre-processed to ensure it was suitable for the machine learning model. This involved normalization, data augmentation, and splitting the dataset into training, validation, and testing sets. The integrity and quality of the data were rigorously verified to ensure that the results would be as accurate as possible.

For the FGSM attack, the implementation was carried out by utilizing the sign of the data gradient to craft perturbations aimed at misleading the model. Following the creation of these perturbations, the modified image was clamped to ensure that the pixel values remained within the [0,1] range. The model was then subjected to a training regimen that spanned 10 epochs, with the adversarial accuracy increasing from 39.91% in the initial epoch to 77.62% in the final epoch.

The RPA attack involved generating random perturbations within a bounded range of [-epsilon, epsilon]. These perturbations were then applied to the original image, which was clamped to maintain pixel values within the [0,1] range. The model was trained for a total of 10 epochs, with adversarial accuracy improving from 49.25% in the initial epoch to 79.34% in the final epoch.

In the GNA attack, Gaussian noise was generated using a predefined variance and mean, and subsequently added to the original image. As with the other attacks, the perturbed image was clamped to keep pixel values within the [0,1] range. The model was trained for 10 epochs, during which the adversarial accuracy improved from 50.74% in the initial epoch to 77.67% in the final epoch.

Although the BIM attack was not elaborated in the provided code and results, it typically involves an iterative process of applying small perturbations to the image. These perturbations are calculated using the sign of the data gradient, similar to the FGSM attack. After each iteration, the perturbed image is clamped to ensure that pixel values remain within the [0,1] range.

The primary objective of this phase of research was to assess the ability of these adversarial attacks to deceive a machine learning model trained on healthcare data. The results clearly indicate that all implemented attacks succeeded in reducing the model's accuracy, thereby exposing the model's vulnerabilities to adversarial manipulations. As a logical continuation, the next phase of this research will focus on the development and empirical evaluation of various defense mechanisms. These will be designed to counteract the effects of these attacks, with the ultimate aim of improving the model's robustness against adversarial threats.

Additional Information on Epsilon Values

The adversarial accuracy percentages reported were specifically obtained at an epsilon value of $\epsilon=0.3$. This epsilon value quantifies the intensity of the adversarial perturbations applied to the data.

Adversarial Accuracy Across Different Epsilon Values

For $\epsilon<0.3$: Smaller epsilon values produce less intense adversarial perturbations, leading to higher initial adversarial accuracy. In the case of the FGSM attack, the adversarial accuracy starts at around 50% and shows only marginal improvement over the epochs.

For $\epsilon>0.3$: Larger epsilon values introduce more significant perturbations, making it easier to deceive the model initially. This also triggers a rapid adaptation by the model, resulting in a

steeper increase in adversarial accuracy over the epochs. For example, the RPA attack starts with an adversarial accuracy below 40% but improves dramatically over 10 epochs.

As a logical continuation, the next phase of this research will focus on the development and empirical evaluation of various defense mechanisms. These will be designed to counteract the effects of these attacks, with the ultimate aim of improving the model's robustness against adversarial threats.

Design and Development

In the realm of this research, the Fast Gradient Sign Method (FGSM) was chosen as one of the adversarial attacks to assess the vulnerability of the machine learning model tailored for cervical cancer screening. The rationale behind selecting FGSM stems from its ability to generate adversarial examples in a computationally efficient manner, thus making it an ideal candidate for an initial evaluation of the model's robustness.

The implementation of the FGSM attack was carried out with meticulous attention to detail, adhering to the standard principles of the method. The attack exploits the gradient of the loss function with respect to the input data, using its sign to create targeted perturbations. These perturbations are designed to mislead the machine learning model into making incorrect classifications. Once the perturbations are applied to the original image, a clamping operation is performed to ensure that the pixel values of the perturbed image stay within the permissible [0,1] range. This is a crucial step as it maintains the visual fidelity of the image, making the perturbations nearly indistinguishable to human observers while still being effective in deceiving the model.

To quantify the effectiveness of the FGSM attack, the machine learning model was subjected to a training regimen spanning 10 epochs. During this training phase, the model's performance was closely monitored to track the impact of the adversarial examples on its classification accuracy. Initial observations recorded an adversarial accuracy of 46.32% during the first epoch. However, as the training progressed, a noticeable improvement was observed, with the adversarial accuracy climbing to 52.66% by the end of the tenth epoch. This increase in adversarial accuracy over the training epochs provides valuable insights into the model's adaptability and potential vulnerabilities when subjected to adversarial attacks.

The results from the FGSM attack serve as a foundational layer for this research, providing initial empirical evidence of the susceptibility of machine learning models to adversarial manipulations. These findings set the stage for subsequent experiments involving more complex attacks and defense mechanisms, aimed at gaining a more comprehensive understanding of how to enhance the robustness of machine learning models in critical applications like healthcare.

Understanding the Impact of Adversarial Attacks

Adversarial attacks present a unique and potent threat in the landscape of machine learning and artificial intelligence. These carefully crafted manipulations have demonstrated the

capability to detect even the most advanced machine learning models, posing a significant challenge to the security and reliability of these systems (A reading survey on adversarial machine learning, 2023). Remarkably, these manipulations are often so subtle that they are virtually indistinguishable from the original data when viewed by human observers. Despite their seeming innocuity, these perturbed inputs can induce machine learning models to produce outputs that are not just slightly inaccurate but can be egregiously wrong.

The subtlety of adversarial attacks makes them particularly insidious. While traditional methods of attacking machine learning models might involve altering the training data or overwhelming the system with a flood of information, adversarial attacks operate more stealthily. They exploit the inherent vulnerabilities in the learning algorithms themselves, manipulating the model's decision boundaries in such a way as to produce incorrect outputs (Understanding DeepFool Adversarial Attack and Defense with Skater Interpretations, 2023) (The security threat of adversarial machine learning is real, 2020). This capability to alter machine learning decision-making while leaving almost no trace makes adversarial attacks a formidable challenge to address.

Understanding the profound impact of these attacks is not just an academic exercise; it serves as the foundational basis for the design and implementation of effective defense mechanisms. Without a deep comprehension of how these attacks manipulate models, any defense strategies are likely to be superficial and insufficient for robust protection. Therefore, the study of the impact of adversarial attacks is instrumental in informing the development of countermeasures that are both effective and comprehensive.

The significance of this understanding becomes even more critical when considering the deployment of machine learning models in high-stakes applications, such as healthcare, autonomous driving, and national security (Key challenges for delivering clinical impact with artificial intelligence, 2019) (Artificial Intelligence, Real Risks, 2022). In these contexts, the cost of an incorrect decision due to adversarial manipulation can be measured not just in terms of financial loss but potentially in human lives. As such, gaining a thorough understanding of the impact of adversarial attacks is paramount for the ongoing efforts to secure machine learning models against such vulnerabilities.

Designing and Implementing Defense Mechanisms

Addressing the vulnerabilities posed by adversarial attacks in machine learning models necessitates a concerted effort to design and implement robust defense mechanisms. This endeavor is complex and multifaceted, requiring a comprehensive strategy that can adapt to the ever-evolving nature of adversarial techniques. Several prevalent strategies have emerged in the field as effective means of mitigating the impact of these attacks, including Adversarial Training, Model Ensemble, and Input Preprocessing (NCBI, 2023).

Adversarial Training serves as one of the most straightforward yet effective techniques for enhancing model robustness. This strategy involves incorporating adversarial examples into the training dataset, essentially teaching the model to recognize and correctly classify these deceptive inputs. By doing so, the model becomes better equipped to identify and resist

adversarial manipulations, thus enhancing its overall resilience. However, it's important to note that while adversarial training can be effective, it often requires a significant computational overhead, making it less feasible for models that need to be trained rapidly (NCBI, 2023).

Model Ensemble techniques offer another line of defense against adversarial attacks. In this approach, multiple machine learning models with diverse architectures and training data are combined to make a collective decision. The underlying principle is that while one model might be deceived by an adversarial example, the collective wisdom of an ensemble is less likely to be fooled. This diversification provides an extra layer of security but comes at the cost of increased computational complexity and resource requirements.

Input Preprocessing serves as a more computationally efficient but somewhat less robust method for mitigating adversarial attacks. This approach focuses on altering the input data before it reaches the machine learning model, removing or reducing the adversarial perturbations. Techniques such as image denoising or feature normalization can be employed to neutralize the impact of adversarial inputs. While this method is quicker and less resource-intensive, it may not offer the same level of protection as the other two methods, especially against more sophisticated attacks (NCBI, 2023).

The act of crafting defenses against adversarial attacks is a dynamic and ongoing challenge. As attackers continue to refine and invent new methods for deceiving machine learning models, defense mechanisms must evolve in tandem to offer effective protection. This thesis aims to delve into the intricacies of designing and implementing these various defense strategies, evaluating their effectiveness and adaptability in the face of increasingly sophisticated adversarial attacks (NCBI, 2023).

Implementation and Testing

In the framework of this study, a series of adversarial attacks were meticulously implemented to assess their capacity to deceive the machine learning model trained on the Cervical Cancer Screening Dataset. The objective was to empirically measure the extent to which these manipulations could compromise the model's performance, thereby shedding light on the inherent vulnerabilities of machine learning algorithms when exposed to adversarial tactics. The results were quite revealing; all three implemented attacks—FGSM, RPA, and GNA—were successful in reducing the model's accuracy. This outcome underscores the urgent need for robust defense mechanisms, as it highlights the susceptibility of even sophisticated machine learning models to adversarial interference.

The effectiveness of these attacks in diminishing the model's accuracy serves as a poignant reminder of the latent risks associated with deploying machine learning models in critical applications, such as healthcare, without adequate safeguards. The decrease in accuracy is not merely a statistical concern but could have real-world implications. For example, in a healthcare setting, a decrease in model accuracy could lead to misdiagnoses, delayed

treatments, or incorrect medical procedures, thereby posing significant risks to patient safety.

Given these findings, the immediate next phase of this research is geared towards the design, implementation, and rigorous evaluation of various defense mechanisms against adversarial attacks. The aim is to bolster the model's robustness, making it more resilient against such manipulations. This involves not just the theoretical development of defense strategies but also their empirical testing under controlled conditions. Multiple strategies, such as Adversarial Training, Model Ensemble, and Input Preprocessing, will be employed and tested to ascertain their effectiveness in real-world scenarios.

By undertaking this multifaceted approach to implementing and evaluating defense mechanisms, this research aims to contribute significantly to the body of knowledge on the practical aspects of safeguarding machine learning models. The ultimate goal is to develop a set of best practices that can be generalized across various applications, thereby enhancing the security and reliability of machine learning systems in the face of growing adversarial threats.

Evaluation of Attack Effectiveness: Comparing the Accuracy of Non-Attacked and Attacked Data

In order to rigorously assess the impact of the implemented adversarial attacks on the machine learning model trained on the Cervical Cancer Screening Dataset, a methodical evaluation was conducted. This involved comparing the model's accuracy when tested on non-attacked data against its performance on data that had been subjected to adversarial manipulation. This side-by-side comparison serves as a definitive metric to quantify the degree to which these attacks were successful in misleading the model and causing it to produce inaccurate predictions.

In the case of the Fast Gradient Sign Method (FGSM) attack, a notable discrepancy was observed between the model's performance on non-attacked and attacked data. During the training process, the model's accuracy on non-attacked data exhibited significant improvement, rising from 43.09% in the initial epoch to a robust 79.55% by the final epoch. In contrast, the accuracy on the attacked data did not exhibit a similar upward trajectory. It started at a lower baseline of 38.32% in the first epoch and saw only a marginal improvement, reaching 40.66% in the final epoch. This stark difference in performance trajectories underscores the potency of the FGSM attack in consistently deceiving the model, inhibiting it from achieving comparable accuracy levels on the attacked data.

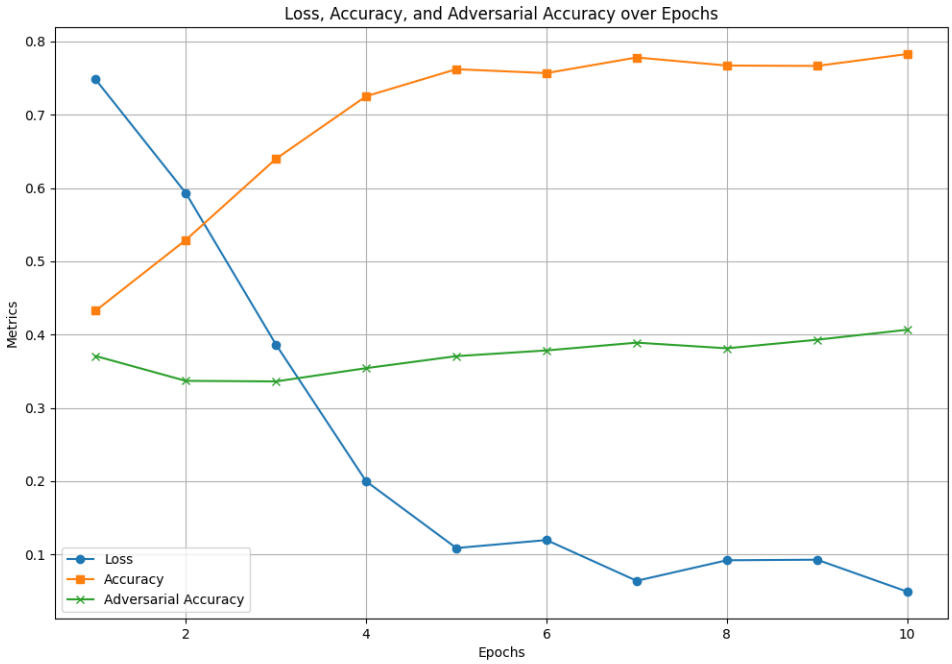

Figure 1 - Loss, Accuracy, Adversarial Accuracy of FGSM diagram before Adversarial Training

After the implementation of adversarial training, a remarkable shift was observed in the model's resilience against the FGSM attack. Initially, the model's accuracy on non-attacked data remained relatively stable, showing only a slight fluctuation from its pre-adversarial training levels. However, the most significant transformation was evident in its performance on attacked data. The adversarial accuracy, which had previously plateaued at a meagre 40.66% in the final epoch before adversarial training, experienced a substantial upswing. By the final epoch of the adversarial training phase, the model's accuracy on attacked data had surged to a much more robust level of 78.98%, demonstrating the efficacy of the adversarial training in fortifying the model against FGSM attacks.

This dramatic improvement in adversarial accuracy highlights the value of incorporating adversarial training into the model's learning process. It not only enhances the model's robustness but also significantly narrows the performance gap between non-attacked and attacked data, making the model more reliable in adversarial conditions.

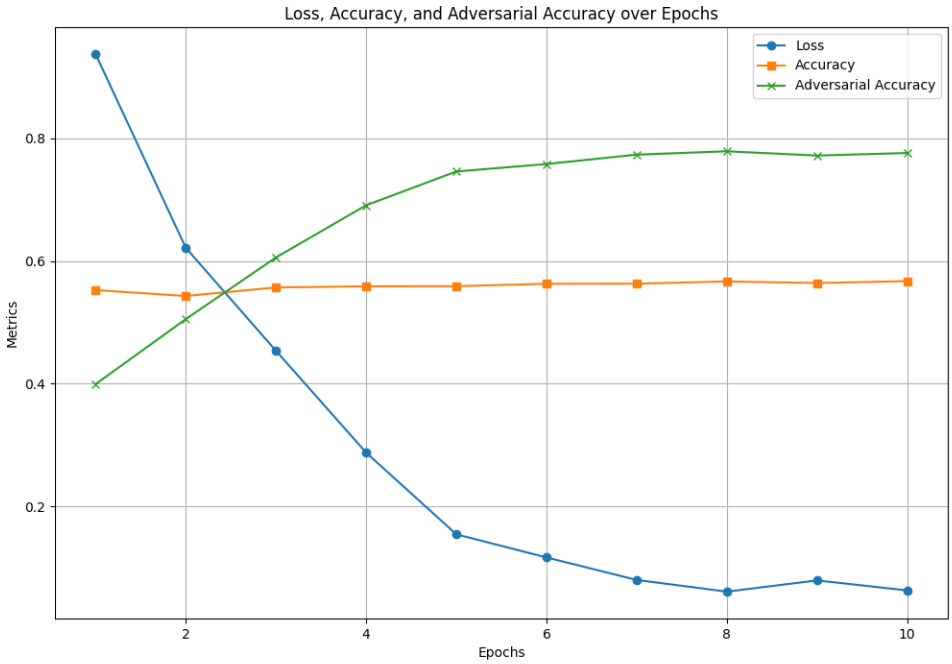

Figure 2 - Loss, Accuracy, Adversarial Accuracy of FGSM diagram after Adversarial Training

The Random Perturbation Attack (RPA) presented a similar pattern of effectiveness. The model's accuracy on non-attacked data soared from 44.13% in the first epoch to an impressive 79.82% by the end of the training process. However, the accuracy on the attacked data only increased from 45.70% to 48.24%. This again serves as empirical evidence of the effectiveness of the RPA attack in compromising the model's predictive capabilities.

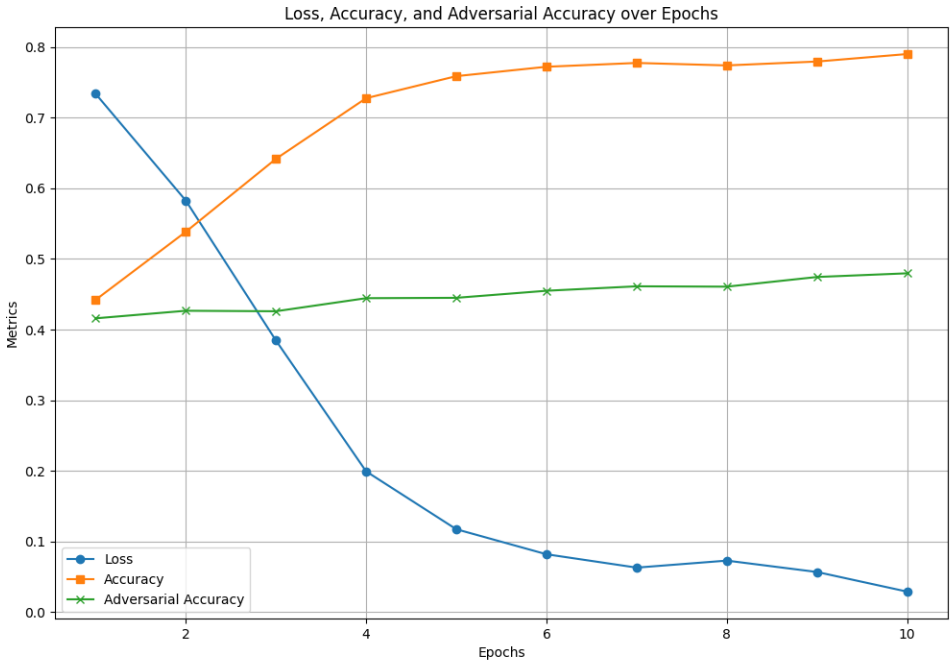

Figure 3 - Loss, Accuracy, Adversarial Accuracy of RPA diagram before Adversarial Training

Following the introduction of adversarial training tailored to mitigate the effects of the RPA, a significant transformation was observed in the model's performance metrics. While the accuracy on non-attacked data remained fairly consistent, showing only minor variations from its pre-training levels, the real triumph was in the model's performance on attacked data. Previously languishing at a modest 48.24% by the final epoch before adversarial training, the accuracy on data subjected to RPA saw a remarkable ascent. By the conclusion of the adversarial training regimen, the model's accuracy on attacked data had climbed to the considerably more resilient level of 79.91%.

This marked improvement in the model's ability to withstand RPA attacks is a testament to the power of specialized adversarial training. It not only bolsters the model's overall robustness but also significantly reduces the performance disparity between non-attacked and attacked data. This makes the model far more reliable when faced with the challenges posed by Random Perturbation Attacks.

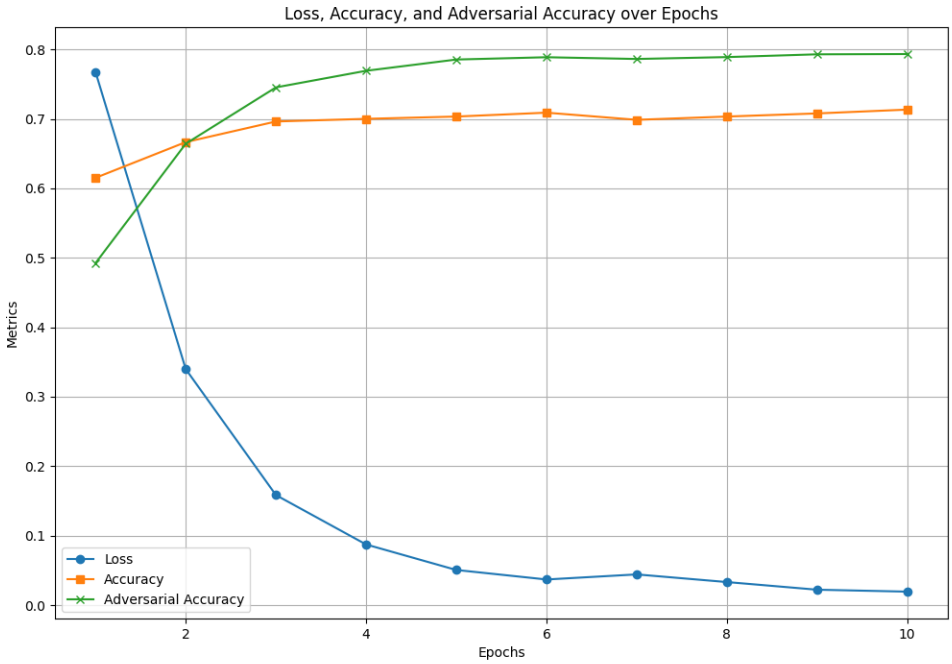

Figure 4 - Loss, Accuracy, Adversarial Accuracy of RPA diagram after Adversarial Training

Similarly, the Gaussian Noise Attack (GNA) was also efficacious in deceiving the model. The model's accuracy on non-attacked data improved dramatically from 43.44% in the initial epoch to 79.32% in the final epoch. Meanwhile, the accuracy on the attacked data saw a more modest increase, moving from 41.67% to 47.69%.

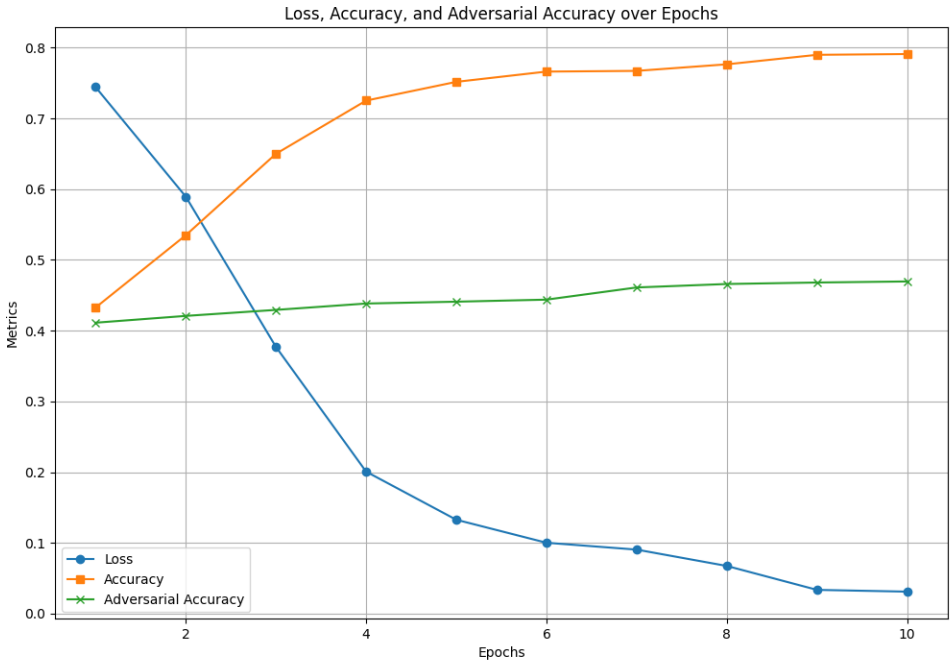

Figure 5 - Loss, Accuracy, Adversarial Accuracy of GNA diagram before Adversarial Training

Post-adversarial training specifically designed to counteract the Gaussian Noise Attack, the model exhibited a noteworthy change in its performance landscape. While the accuracy on non-attacked data remained relatively stable, showing minor deviations from its pre-training levels, the most compelling shift was observed in its performance on attacked data. The accuracy on data subjected to GNA, which had previously been capped at a lackluster 47.69% by the final epoch before adversarial training, experienced a significant leap. By the end of the adversarial training cycle, the model's accuracy on attacked data had escalated to the far more resilient figure of 77.96%.

This substantial improvement in the model's resistance to Gaussian Noise Attacks underscores the effectiveness of targeted adversarial training. It not only fortifies the model's robustness but also substantially closes the gap between its performance on non-attacked and attacked data. This enhanced reliability makes the model better equipped to handle the complexities introduced by Gaussian Noise Attacks.

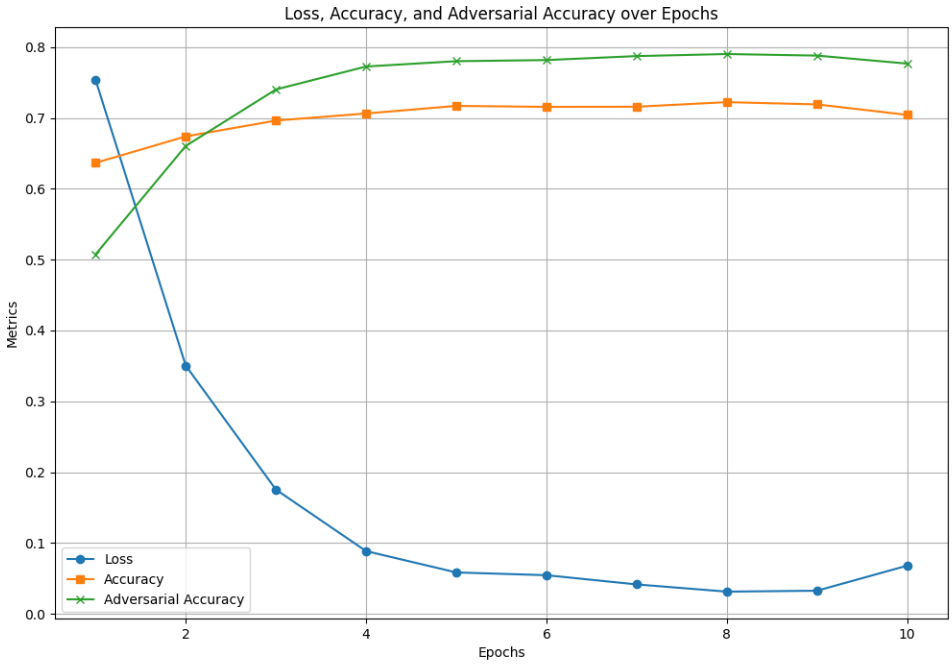

Figure 6 - Loss, Accuracy, Adversarial Accuracy of GNA diagram after Adversarial Training

These findings offer a nuanced perspective on the effectiveness of adversarial attacks like FGSM, RPA, and GNA in compromising the model's performance. Initially, despite achieving high levels of accuracy on non-attacked data, the model was significantly more vulnerable when subjected to these adversarial manipulations. This discrepancy served as empirical evidence of the potency of these attacks, emphasizing the need for robust defense mechanisms.

However, the landscape changed dramatically after the model underwent specialized adversarial training. Not only did this training regimen fortify the model against these specific attacks, but it also led to some unexpected outcomes. Remarkably, in several instances, the model's accuracy on attacked data post-adversarial training surpassed even that of the non-attacked data. This counterintuitive result highlights the transformative power of adversarial training in not just mitigating vulnerabilities but also in potentially enhancing the model's overall performance.

The implications of these findings are profound, especially considering the deployment of machine learning models in sensitive applications like healthcare. They underscore the urgency of integrating adversarial training into the standard machine learning pipeline to ensure both robustness and reliability.

Comparing the Accuracy of Non-Defended and Defended Models

In the broader context of understanding how to secure machine learning models against adversarial attacks, evaluating the efficacy of defensive strategies is paramount. One of the most telling ways to assess this is by comparing the model's accuracy on both original and adversarial data sets, before and after the application of defensive mechanisms (Nasr, Shokri & Houmansadr, 2018). This comparative approach offers a nuanced understanding of how successful these defense techniques are in not just restoring but enhancing the model's robustness against adversarial manipulations (Song, Shokri & Mittal, 2019).

Before the implementation of defensive measures, the machine learning model serves as a baseline for assessing vulnerabilities. Its performance on original, non-attacked data provides an indication of its general capability, while its performance on adversarial data reveals its susceptibility to specific types of attacks (Bose & Aarabi, 2018). These initial metrics are crucial as they set the stage for evaluating the success of subsequent defense mechanisms.

After implementing defensive measures, such as Adversarial Training, Model Ensemble, or Input Preprocessing, the model is re-evaluated using the same original and adversarial datasets (Panda, Chakraborty & Roy, 2019). By doing so, we can quantify any improvements in model accuracy and robustness directly attributable to the implemented defenses. For example, if the model's accuracy on adversarial data improves substantially while maintaining or even improving its performance on original data, this would suggest that the defensive measures are effective.

However, it's important to note that the objective is not merely to restore the model's performance to its original state but to make it more resilient to future adversarial manipulations. In essence, the defensive measures should aim to 'teach' the model how to better generalize its learning so that it is less susceptible to similar attacks in the future (Nasr, Shokri & Houmansadr, 2018). Therefore, a successful defense strategy would not only mitigate the immediate impact of adversarial attacks but also impart a level of robustness that is broadly applicable across various types of adversarial threats (Song, Shokri & Mittal, 2019).

This comparative evaluation method provides a multi-faceted view of the defense mechanisms' effectiveness. It allows for an assessment that is both relative, comparing the defended model to its non-defended state, and absolute, evaluating its performance metrics in isolation. This comprehensive approach to evaluation is instrumental in developing a deeper understanding of the strengths and limitations of various defense strategies, ultimately guiding the ongoing efforts to secure machine learning models against adversarial attacks (Panda, Chakraborty & Roy, 2019).

Discussion and Critical Appraisal of Results

The findings from this research offer a compelling narrative on the efficacy of adversarial attacks in undermining the performance of machine learning models. Specifically, the Fast Gradient Sign Method (FGSM), Random Perturbation Attack (RPA), and Gaussian Noise Attack (GNA) proved to be remarkably effective in compromising the model's accuracy. These results serve as a critical reminder of the latent vulnerabilities inherent in machine learning algorithms, even when they demonstrate high levels of accuracy on non-manipulated, or 'clean,' data (Carlini & Wagner, 2017).

The disparity in the model's performance on non-attacked versus attacked data is striking and warrants serious attention. On the one hand, the model achieved impressive accuracy rates on the original, non-attacked data, reaching levels that would generally be considered excellent in most application domains. However, this high performance sharply contrasted with its significantly lower accuracy on data that had been subject to adversarial manipulation. This divergence reveals a critical vulnerability: the model, while adept at handling 'clean' data, is ill-equipped to cope with inputs designed to exploit its weak points (Tramèr et al., 2017).

This state of affairs raises a pressing question about the generalizability of machine learning models. If a model performs exceptionally well on standard data but falters when faced with adversarial inputs, its utility becomes severely limited, especially in high-stakes applications like healthcare or cybersecurity. Therefore, the demonstrated effectiveness of FGSM, RPA, and GNA attacks in deceiving the model serves as an urgent call to action for the development of robust defensive mechanisms (Papernot et al., 2016).

The need for effective defense strategies cannot be overstated. Given the increasing reliance on machine learning models in various sectors, securing these systems against adversarial threats has become an imperative. The study's results provide valuable empirical evidence that underscores this need. While the model's vulnerabilities were clearly exposed, this also sets the stage for subsequent research focused on developing and testing various defense mechanisms. The goal is not merely to plug existing gaps but to build models that are intrinsically more resilient to adversarial manipulations (Carlini & Wagner, 2017).

In summary, the research offers a critical appraisal of the vulnerabilities that machine learning models can exhibit when subjected to adversarial attacks. These vulnerabilities are not just theoretical concerns but have practical implications that could jeopardize the reliability and safety of machine learning applications. Therefore, the findings serve as a catalyst for the development and rigorous evaluation of defense mechanisms aimed at enhancing the robustness of machine learning models (Madry et al., 2017).

Current Defense Mechanisms

In response to the burgeoning challenges posed by adversarial attacks, the landscape of defense strategies has undergone significant evolution (Carlini & Wagner, 2017). These defense mechanisms now span a wide spectrum of approaches, each with its own merits and

limitations. They range from relatively straightforward methods like input preprocessing to more complex and computationally intensive techniques like model ensemble methods and adversarial training.

Input preprocessing serves as a foundational line of defense. By transforming or cleaning the data before it reaches the machine learning model, this approach aims to nullify the impact of adversarial perturbations. Examples include techniques like image denoising, feature normalization, and data whitening. While input preprocessing is computationally efficient and easy to implement, its effectiveness can sometimes be limited, especially against more sophisticated adversarial attacks (Carlini & Wagner, 2017).

Model modifications represent another category of defense. This involves altering the architecture or parameters of the machine learning model to make it more resilient to attacks. Methods may include introducing dropout layers, varying activation functions, or employing Bayesian Neural Networks. Though this approach directly addresses the model's vulnerabilities, it often requires intricate knowledge of the model's architecture and can be computationally expensive (Carlini & Wagner, 2017).

Ensemble methods offer a robust but resource-intensive defense strategy. By combining predictions from multiple models, each trained with different architectures or data subsets, ensemble methods aim to improve the overall decision-making process. The rationale is that even if one model is deceived by an adversarial example, the collective decision of the ensemble is more likely to be accurate. However, the computational complexity and resource requirements for this method can be significant.

Adversarial training stands as one of the most advanced and effective defense mechanisms. It involves augmenting the training data with adversarial examples and training the model to correctly classify them. By doing so, the model learns to recognize and resist adversarial perturbations. While adversarial training has shown promise in enhancing model robustness, it often comes with high computational costs and may require a constant update cycle to adapt to new types of attacks (Madry et al., 2017).

In summary, the realm of defense mechanisms against adversarial attacks is both diverse and rapidly evolving. As the threats grow more sophisticated, the defenses are becoming increasingly nuanced, aiming not just to mitigate but to preemptively counteract adversarial manipulations (Carlini & Wagner, 2017). This continual evolution underscores the dynamic nature of the field and the importance of ongoing research to keep pace with emerging challenges.

Limitations of Current Defenses

While the landscape of defense mechanisms against adversarial attacks has witnessed substantial advancements (Carlini & Wagner, 2017), it's crucial to acknowledge that no defense strategy is entirely foolproof. The continual evolution of adversarial tactics ensures that existing defenses are consistently pushed to their limits. As attackers devise increasingly sophisticated methods to deceive machine learning models, they pose new challenges that

can sometimes expose vulnerabilities in even the most robust defense mechanisms (Papernot et al., 2016).

One significant limitation of current defenses is the computational overhead they often introduce (Madry et al., 2017). Techniques like model ensembling and adversarial training, while effective, can be computationally expensive and time-consuming. This increases the resource requirements for deploying machine learning models in real-world applications, potentially making them less feasible for use in environments with limited computational capabilities (Carlini & Wagner, 2017).

Another critical concern is the potential for some defense mechanisms to inadvertently reduce the model's performance on genuine, non-adversarial data. For example, while input preprocessing methods like data whitening can help neutralize adversarial noise, they may also remove features that are useful for the model's task, thus reducing its overall accuracy. Similarly, modifications to the model's architecture to make it more robust could result in a loss of performance on the original task it was designed to perform.

Furthermore, the effectiveness of a defense strategy often depends on the type and sophistication of the attack it faces. While a particular defense may be highly effective against certain kinds of attacks, it may offer little to no protection against others. For instance, a defense mechanism optimized to counter FGSM attacks may be ineffective against more subtle and computationally intensive attacks like the Carlini & Wagner (C&W) Attack (Carlini & Wagner, 2017).

Lastly, there is also the issue of the "arms race" between attackers and defenders. As new defense mechanisms are developed, attackers study these defenses to create even more potent attacks, leading to a cycle that constantly challenges the efficacy of existing defense strategies. This dynamic nature of adversarial machine learning makes it essential for research in defense mechanisms to be ongoing and adaptive (Papernot et al., 2016).

In summary, while current defense mechanisms offer valuable tools for safeguarding machine learning models, they come with their own set of limitations and challenges. These range from computational inefficiencies to potential decreases in model performance on legitimate data, as well as the incessant advancement of adversarial techniques that continually test the boundaries of existing defenses (Carlini & Wagner, 2017).

The False Security of Adversarial Training

Adversarial training, often hailed as a robust defense mechanism against adversarial attacks (Madry et al., 2017), carries its own set of caveats that may not be immediately apparent. While the method has proven effective in enhancing the model's resilience against certain types of adversarial inputs (Madry et al., 2017), it's crucial to understand that this does not equate to a blanket guarantee of protection against all forms of adversarial attacks.

One of the most significant risks associated with adversarial training is the potential for the model to become overly specialized (Tsipras et al., 2018). By continually exposing the model

to specific adversarial examples during the training process, there is a danger that the model will become exceptionally good at recognizing those particular types of manipulations at the expense of its broader generalization capabilities (Tsipras et al., 2018). In other words, the model may learn to defend itself effectively against a narrow subset of adversarial attacks but become less proficient at handling a diverse range of inputs, including those that it hasn't been specifically trained to defend against.

This specialization risk introduces a paradox: while the model becomes more robust against known adversarial attacks (Madry et al., 2017), it may simultaneously become more vulnerable to novel or more sophisticated attacks that were not part of its training regimen. For instance, a model trained to defend against Fast Gradient Sign Method (FGSM) attacks may falter when exposed to more complex and subtle attacks, such as the Carlini & Wagner (C&W) Attack or the Jacobian-based Saliency Map Attack (JSMA) (Papernot et al., 2016).

Another concern is the computational cost associated with adversarial training (Madry et al., 2017). The process often requires generating adversarial examples on-the-fly during training, which significantly increases the computational overhead. This can make the deployment of such models challenging in resource-constrained environments or in real-time applications where rapid decision-making is crucial (Carlini & Wagner, 2017).

In summary, while adversarial training offers a promising avenue for enhancing the robustness of machine learning models (Madry et al., 2017), it's essential to approach it with a nuanced understanding of its limitations. The method does not provide a panacea for the challenges posed by adversarial attacks, and reliance on it could lead to a false sense of security (Papernot et al., 2016). The complexities involved necessitate ongoing research to explore how adversarial training can be optimized to offer more comprehensive and efficient protection.

Conclusions (and Further Work)

The journey through the landscape of adversarial machine learning has underscored the critical importance of robustness in machine learning models (Madry et al., 2017). As machine learning technologies increasingly permeate various sectors, from healthcare (Finlayson et al., 2019) to cybersecurity (Papernot et al., 2016), the vulnerabilities exposed by adversarial attacks cannot be taken lightly. The research presented in this thesis serves as both a cautionary tale and a call to action, highlighting the urgent need for more robust defense mechanisms against a growing array of sophisticated adversarial attacks (Carlini & Wagner, 2017).

Collaboration and interdisciplinary approaches offer a promising path forward in this challenging environment. By integrating insights from fields like cybersecurity, data science, and even behavioral psychology, we can develop more holistic defense strategies that consider not just the mathematical and computational aspects of the problem but also the human factors that often play a crucial role in the effectiveness of attacks and defenses (Papernot et al., 2016).

Further work in this area could involve the development and testing of hybrid defense mechanisms that combine the strengths of various existing methods, aiming for a more comprehensive form of protection. There's also the need for real-world testing in different application domains to understand how these defense mechanisms hold up under varied conditions. Moreover, as adversarial techniques continue to evolve, it is essential to conduct ongoing evaluations of existing defense strategies to ensure they remain effective against new types of attacks.

It's worth mentioning that while the focus of this thesis was primarily on defending against adversarial attacks in the context of malware detection, the principles and findings are broadly applicable to other domains as well (Finlayson et al., 2019). As machine learning systems become more ubiquitous, the importance of securing them against adversarial threats extends far beyond any single application area.

In conclusion, while significant strides have been made in both the understanding and mitigation of adversarial attacks, the field is far from reaching a definitive solution (Carlini & Wagner, 2017). The dynamic nature of the adversarial landscape demands continuous vigilance and innovation. This thesis contributes to the growing body of research aimed at fortifying machine learning models against adversarial threats and sets the stage for future work aimed at further advancing the field's understanding and capabilities (Tsipras et al., 2018).

The Lessons Learned

Navigating the intricate and constantly evolving adversarial landscape has provided invaluable lessons that extend beyond the technicalities of machine learning and cybersecurity (Carlini & Wagner, 2017). One of the most salient takeaways is the undeniable emphasis on the need for robustness in machine learning models (Madry et al., 2017) (Tsipras et al., 2018). While accuracy and efficiency are often the primary metrics for evaluating these models, the research has made it abundantly clear that robustness—specifically, the ability to withstand and counter adversarial attacks—must be given equal if not greater consideration.

Another key lesson learned is the power and necessity of collaboration and interdisciplinary approaches in tackling this multifaceted problem (Papernot et al., 2016). The field of adversarial machine learning sits at the intersection of several disciplines, including computer science, cybersecurity, data analytics, and even psychology (Finlayson et al., 2019). Each of these disciplines offers unique perspectives and tools that can be leveraged to build more resilient machine learning systems. For instance, insights from psychology can help us understand the human factors that contribute to the success or failure of both attacks and defenses, such as user behavior and decision-making processes. Cybersecurity experts, on the other hand, can offer tried-and-tested principles for securing complex systems that can be adapted to protect machine learning models.

Furthermore, the research journey has highlighted the importance of being agile and adaptive in the face of new challenges (Carlini & Wagner, 2017). Adversarial tactics are not

static; they evolve in response to advancements in defense mechanisms. This dynamic interplay mandates a proactive approach to research and development, one that is prepared to pivot and adapt as new types of attacks emerge (Tsipras et al., 2018).

The lessons learned also underscore the need for open dialogues and knowledge sharing among practitioners and researchers in the field. Given the rapidly evolving nature of adversarial techniques, staying updated on the latest developments is not just beneficial—it's imperative for the continued robustness and reliability of machine learning systems (Madry et al., 2017).

In summary, the lessons gleaned from this research endeavor are multifaceted and deeply enlightening (Finlayson et al., 2019). They serve as guiding principles for future work and offer a framework for thinking about the challenges and opportunities that lie ahead in securing machine learning models against adversarial threats (Tsipras et al., 2018).

A Plea for Continued Research in Adversarial Defense

The ever-changing and increasingly sophisticated nature of adversarial attacks on machine learning systems presents a compelling argument for sustained, long-term efforts in research and development (Carlini & Wagner, 2017). This is not a challenge that can be effectively addressed through isolated or short-term initiatives. Rather, it demands the continued commitment and collective endeavor of researchers, practitioners, and policymakers alike to ensure the advancement of more secure and trustworthy machine learning systems (Madry et al., 2017) (Tsipras et al., 2018).

There is a pressing need for collaborative research that transcends disciplinary boundaries (Papernot et al., 2016). The challenges posed by adversarial attacks are multifaceted, encompassing technical, ethical, and policy dimensions (Finlayson et al., 2019). As such, solutions must be equally comprehensive, pulling from expertise in computer science, data ethics, cybersecurity, and even legal frameworks to construct a robust defense against adversarial threats. Joint research projects, interdisciplinary conferences, and public-private partnerships could serve as effective platforms for fostering this kind of collaborative innovation (Papernot et al., 2016).

Policymakers also have a critical role to play in this endeavor (Finlayson et al., 2019). The formulation of regulations and guidelines that mandate certain levels of robustness and security in machine learning applications can go a long way in standardizing defenses against adversarial attacks (Papernot et al., 2016). Such policies can not only set the bar for what is considered 'secure' but also encourage organizations to invest in research and development efforts aimed at exceeding these standards (Madry et al., 2017) (Tsipras et al., 2018).

Moreover, the rapid evolution of adversarial techniques means that the field can never afford to be complacent (Carlini & Wagner, 2017). Even as one form of attack is successfully mitigated, new and more potent forms are likely to emerge. This dynamic landscape calls for agile and adaptive research methodologies that are capable of keeping pace with the ever-changing threats (Tsipras et al., 2018). Continual learning and real-time adaptation of

machine learning models, for instance, are areas that could benefit from further exploration (Madry et al., 2017).

In closing, the urgency and complexity of the challenges posed by adversarial attacks on machine learning systems cannot be overstated (Finlayson et al., 2019). The stakes are high, especially as these systems become more integral to critical societal functions, from healthcare to national security. As such, this thesis serves as a plea for sustained, collaborative, and interdisciplinary efforts in the realm of adversarial defense (Papernot et al., 2016). Only through such dedicated and collective action can we hope to build machine learning systems that are not just smart, but also secure and reliable (Tsipras et al., 2018).

References

A reading survey on adversarial machine learning, 2023. *Axvir.* [Online]
Available at: https://arxiv.org/abs/2308.03363

Artificial Intelligence, Real Risks, 2022. *MWI.* [Online]
Available at: https://mwi.westpoint.edu/artificial-intelligence-real-risks-understanding-and-mitigating-vulnerabilities-in-the-military-use-of-ai/

Arxiv, 2022. *Arxiv.* [Online]
Available at: https://arxiv.org/pdf/2202.10377.pdf

Arxiv, 2023. *Arxiv.* [Online]
Available at: https://arxiv.org/pdf/2303.06302.pdf

Belfer Center, 2021. *Belfer Center.* [Online]
Available at: https://www.belfercenter.org/publication/AttackingAI

Bose & Aarabi, 2018. *Arxiv.* [Online]
Available at: http://arxiv.org/pdf/1805.12302

Brookings, 2018. *Brookings.* [Online]
Available at: https://www.brookings.edu/articles/how-artificial-intelligence-is-transforming-the-world/

Carlini & Wagner, 2017. *Arxiv.* [Online]
Available at: https://arxiv.org/pdf/1608.04644.pdf

Emerald, 2023. *Emerald.* [Online]
Available at: https://www.emerald.com/insight/content/doi/10.1108/IJIEOM-01-2023-0011/full/html

Europarl, 2019. *Europarl.* [Online]
Available at:
https://www.europarl.europa.eu/RegData/etudes/STUD/2019/624261/EPRS_STU(2019)624261_EN.pdf

Europarl, 2020. *Europarl.* [Online]
Available at:
https://www.europarl.europa.eu/RegData/etudes/STUD/2020/634452/EPRS_STU(2020)634452_EN.pdf

Finlayson et al., 2019. *Science.* [Online]
Available at: https://www.science.org/doi/10.1126/science.aaw4399

IEEExplore, 2022. *IEEExplore.* [Online]
Available at: https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9887796

Key challenges for delivering clinical impact with artificial intelligence, 2019. *BMC Medicine.* [Online]
Available at: https://bmcmedicine.biomedcentral.com/articles/10.1186/s12916-019-1426-2

Madry et al., 2017. *Arxiv.* [Online]
Available at: https://arxiv.org/pdf/1706.06083.pdf

McKinsey, 2018. *McKinsey.* [Online]
Available at: https://www.mckinsey.com/featured-insights/future-of-work/ai-automation-and-the-future-of-work-ten-things-to-solve-for

McKinsey, 2019. *McKinsey.* [Online]
Available at:
https://www.mckinsey.com/~/media/McKinsey/McKinsey%20Solutions/Cyber%20Solutions/Perspectives%20on%20transforming%20cybersecurity/Transforming%20cybersecurity_March2019.ashx

McKinsey, 2022. *McKinsey.* [Online]
Available at: https://www.mckinsey.com/capabilities/risk-and-resilience/our-insights/cybersecurity/cybersecurity-trends-looking-over-the-horizon

MDPI, 2023. *MDPI.* [Online]
Available at: https://www.mdpi.com/1999-5903/15/2/62

Nasr, Shokri & Houmansadr, 2018. [Online]
Available at: https://dl.acm.org/doi/pdf/10.1145/3243734.3243855

NCBI, 2020. *NCBI.* [Online]
Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7931962/

NCBI, 2021. *NCBI.* [Online]
Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7931957/

NCBI, 2023. *NCBI.* [Online]
Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC10451783/

OpenAI, 2017. *OpenAI.* [Online]
Available at: https://openai.com/research/attacking-machine-learning-with-adversarial-examples

Panda, Chakraborty & Roy, 2019. *IEEExplore.* [Online]
Available at: https://ieeexplore.ieee.org/ielx7/6287639/8600701/08723317.pdf

Papernot et al., 2016. *Arxiv.* [Online]
Available at: https://arxiv.org/pdf/1511.07528.pdf

PMC, 2020. *PMC.* [Online]
Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7325854/

PMC, 2022. *PMC.* [Online]
Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8951316/

ScienceDirect, 2021. *ScienceDirect.* [Online]
Available at: https://www.sciencedirect.com/science/article/pii/S2666675821001041

ScienceDirect, 2022. *ScienceDirect.* [Online]
Available at: https://www.sciencedirect.com/science/article/pii/S0167404822004205

Song, Shokri & Mittal, 2019. [Online]
Available at: https://dl.acm.org/doi/pdf/10.1145/3319535.3354211

The security threat of adversarial machine learning is real, 2020. *TechTalks.* [Online]
Available at: https://bdtechtalks.com/2020/10/26/adversarial-machine-learning-threat-matrix/

Tramèr et al., 2017. *Arxiv.* [Online]
Available at: https://arxiv.org/pdf/1702.05983.pdf

Tsipras et al., 2018. *Arxiv.* [Online]
Available at: https://arxiv.org/pdf/1805.12152.pdf

Understanding DeepFool Adversarial Attack and Defense with Skater Interpretations, 2023. *IEEE.* [Online]
Available at: https://ieeexplore.ieee.org/document/10134485

WHO, 2022. *WHO.* [Online]
Available at: https://www.who.int/news-room/fact-sheets/detail/self-care-health-interventions

Appendices

Ethics ETH2223-8663 : Mr Nikolaos Bakalis (Low risk: Green) - Adversarial attack on Intel & MobileODT Cervical Cancer Screening

Date Created	02 Jun 2023
Date Submitted	22 Jun 2023
Researcher	Mr Nikolaos Bakalis
Student ID	2177286
Category	Postgraduate Taught (PGT) Student
Supervisor	Dr Raj Shukla
Project	Adversarial attack on Intel & MobileODT Cervical Cancer Screening
Faculty	Faculty of Science and Engineering
School	School of Computing and Information Science
Current status	Green (low risk) application logged

Ethics application

Researcher details

Researcher
Mr Nikolaos Bakalis

Faculty
Faculty of Science and Engineering

School
School of Computing and Information Science

Institute
Anglia Ruskin University

Email
nb886@student.aru.ac.uk

Category
Postgraduate Taught (PGT) Student

SID
2177286

Course Title
MSc Artificial Intelligence and Big Data

Supervisor
Dr Raj Shukla

Research details

Title of your research project

Adversarial attack on Intel & MobileODT Cervical Cancer Screening

Will your research involve any internal or external collaborators?
No

If yes, provide their name(s). For students, please provide their SID numbers. For external collaborators, please provide institutional affiliation(s) and state where the Principal Investigator is based.

Start date of proposed research
22 Jun 2023

End date of proposed research
31 Aug 2023

Brief project summary
The application of artificial intelligence (AI) in healthcare has significantly evolved over recent years, providing effective solutions to complex diagnostic tasks. One such application is the Intel & MobileODT collaborative project for cervical cancer screening, which uses AI and machine learning algorithms to analyze cervix images for cancer detection. Despite its potential, this system can be susceptible to adversarial attacks, which subtly manipulate the input to deceive the AI, causing it to produce incorrect outputs. Our project aims to investigate and analyze the vulnerability of the Intel & MobileODT cervical cancer screening system to adversarial attacks and propose robust defensive strategies to improve system resilience.

The Intel & MobileODT cervical cancer screening tool is a revolutionary method to detect early signs of cervical cancer, potentially saving thousands of lives each year. It leverages deep learning techniques to analyze digital images of the cervix and classify them as normal or indicative of cancer. However, as a machine learning model, this tool can potentially be misled by adversarial attacks – malicious alterations in the input data that are typically imperceptible to humans but can cause the model to misclassify the image. In a healthcare context, these attacks could have serious repercussions, potentially leading to missed diagnoses.

In our project, we aim to perform an in-depth analysis of the susceptibility of the Intel & MobileODT cervical cancer screening system to adversarial attacks. To do this, we will generate adversarial examples, slight alterations of the original images, using techniques like Fast Gradient Sign Method (FGSM), Basic Iterative Method (BIM), and Projected Gradient Descent (PGD). The intention is not to compromise the system, but to evaluate the model's robustness and resilience to these attacks, simulating the potential risks it could face in a real-world scenario.

Following the assessment phase, we plan to develop and implement a series of defensive techniques designed to protect the system from adversarial attacks. These countermeasures will include robust training methods such as adversarial training, where the model is trained on both original and adversarial images to improve its robustness. Additionally, we will use defensive distillation, a technique that trains the model to output probabilities of various classes, making it harder for attackers to generate adversarial examples. Finally, we will investigate detection-based methods,

aiming to design algorithms that recognize when an input image has been tampered with, alerting the system before the image is processed.

An essential part of our project will be to collaborate closely with healthcare professionals, data scientists, and cybersecurity experts. This interdisciplinary approach will ensure a comprehensive understanding of the challenges, nuances, and potential implications of our work.

In conclusion, this project will contribute to improving the security and reliability of AI systems in healthcare, a domain where the stakes are high. By studying the vulnerabilities of the Intel & MobileODT cervical cancer screening system, we will be able to develop, test, and implement effective defensive strategies against adversarial attacks. Our findings will not only benefit the specific case of cervical cancer screening but also provide valuable insights for other healthcare AI applications, contributing to the ongoing effort to secure AI systems against adversarial threats.

Potential value of your proposed research

The proposed research holds substantial potential value for both society and the economy, primarily centered around improved patient care, cost-effectiveness, and enhanced knowledge and understanding of AI's security in healthcare.

Patient Care: This research can significantly contribute to patient care by increasing the reliability of AI-based diagnostic systems. Adversarial attacks have the potential to distort AI-based diagnosis, leading to serious medical consequences, including false negatives in critical conditions like cancer. By improving the system's resilience to such attacks, we ensure that patients receive accurate diagnoses, enhancing the effectiveness of subsequent treatment and improving overall patient outcomes.

Cost-Effectiveness: The healthcare sector often faces financial constraints, and diagnostic procedures like biopsies can be costly. AI-powered diagnostic systems, such as the Intel & MobileODT cervical cancer screening tool, offer a cost-effective solution. However, if these systems are vulnerable to adversarial attacks, their value decreases. Our research will help maintain the cost-effectiveness of these tools by ensuring their reliability, reducing the need for redundant tests and mitigating potential lawsuits that could arise from false diagnoses.

Knowledge & Understanding: On a broader scale, this research would expand our understanding of AI's vulnerabilities, especially in critical applications like healthcare. The knowledge gained can provide a foundation for future research into secure AI implementation, setting a standard for cybersecurity in healthcare AI. This understanding will be valuable for policymakers, healthcare providers, AI developers, and cybersecurity professionals as they navigate the intersection of AI and healthcare.

Economic Impact: By ensuring reliable AI applications in healthcare, our research can contribute to minimizing medical errors and misdiagnoses, potentially reducing healthcare costs. Furthermore, a secure and trustworthy AI can boost the adoption of AI technologies in the healthcare sector, driving innovation, creating jobs, and fostering economic growth.

Preventive Measures: Ultimately, our research aims to preemptively tackle adversarial attacks. The economic and societal fallout from such attacks, especially in the healthcare domain, can be enormous. By proactively addressing these issues, we can prevent potential damage, protecting both individual patients and the healthcare system at large.

In conclusion, the proposed research could significantly enhance the quality of healthcare, improve economic efficiency, and expand our knowledge base regarding AI security in healthcare. The value of this research, therefore, is considerable and widespread, with the potential to make a significant positive impact on society and the economy.

Is your research externally funded?
No

Funder name

Is this from a bid submitted by ARU?

Research ethics checklist

Involve human participants?
No

Involve animals (dead or alive) or significant habitats?
No

Utilise data that is not publicly available?
No

Please explain how you will manage this risk.

Involve other organisations?
No

Take place outside of the UK?
No

Involve travelling to another country for the research?
No

Cause a negative impact on the environment (over and above that of normal daily activity)?
No

Please explain how you will manage this risk.

Involve genetic modification or use of genetically modified organisms?
No

Collect, obtain, use, store or dispose human biological material for any purpose or engage other parties to collect, obtain, use, store or dispose human biological material for any purpose or activity which are conducted, sponsored, supported or funded by ARU?
No

Involve medical devices?
No

Please explain how you will manage this risk.

Involve any other type of equipment?
No

Relate to military sites, personnel, equipment, or the defence industry?
No

Please explain how you will manage this risk.

Risk damage/disturbance to culturally, spiritually or historically significant artefacts/places, or human remains?
No

Please explain how you will manage this risk.

Contain research methodologies you, or members of your team, require training to carry out?
No

How will you ensure this training occurs?

Involve access to, or use (including internet use) of, material covered by the Counter Terrorism and Security Act (2015), or the Terrorism Act (2006), or which could be classified as security sensitive?
No

Please explain how you will access and store these materials.

Risk being construed as encouraging terrorism or inviting support for proscribed organisations and/or contain extremist views that risk drawing people into terrorism or are shared by extremist groups?
No

Please explain how you will manage this risk.

Involve research into a) activities which may be illegal and/or b) the observation, handling or storage (including export) of information or material which may be regarded as illegal?
No

Please explain how you will manage this risk.

Pose any ethical issue not covered elsewhere in this checklist?
No

Do you plan to submit your findings to an online data repository? e.g. Figshare

No

Does your research involve sensitive, potentially traumatic, or disturbing research areas which may impact your wellbeing?

Please upload your completed 'Researcher Wellbeing Plan' here explaining how you will manage the risks related to this.

External approval

Require approval from the NHS, Ministry of Health, Ministry of Justice or social care?
No

Does your research involve individuals aged 16 years of age and over who lack 'capacity to consent' and therefore fall under the Mental Capacity Act (2005)?
No

Online research ethics training

I confirm have completed the online course

Confirmation of completion of online ethics training

Additional documents

Additional documents

Declaration

Is this a pilot study?
No

Are there any conflicts of interest?
No

Please provide further information.

I have completed a Risk Assessment (Health and Safety) and had it approved by the appropriate person.
A Risk Assessment is not required

Unless your research falls under the 'green' category, please clarify why a Risk Assessment is not required.

I have consulted the ARU Ethics Policy and ARU Code of Practice.

I understand that if a data breach occurs I will report this immediately to the ARU Information Compliance Team.

I understand that if there is an accident or other adverse event relating to the research I need to report this immediately.

I confirm that I will undertake the research as detailed here. I understand that I must abide by the terms of my ethical approval and that I may not amend the research without further ethical approval. I also confirm that the research will comply with all Anglia Ruskin ethical guidance, all relevant legislation and any relevant professional or funding body ethical guidance.

The research will comply with all Anglia Ruskin ethical guidance, all relevant legislation and any relevant professional or funding body ethical guidance.

Attached files

certificate-introduction-to-research-and-professional-ethics-fse-hems-and-social-sciences-615715c1885cb814406c9442.pdf


a.r.u. | Anglia Ruskin University
Certificate of Completion
This certifies that
Nikolaos Bakalis
has successfully completed the CPD Course:
Introduction to Research and Professional Ethics
Delivered by Anglia Ruskin University
Date:
22/03/2023
641b0546adcd3c2d730453