National College of Ireland

Optimizing Adversarial Attacks on ML-Powered Malware Detection Systems

MSc Research Project
Artificial Intelligence

Vikas Varma Malipeddi
Student ID: 22143335

School of Computing
National College of Ireland

Supervisor: Dr.Anh Duong Trinh(Senja)

National College of Ireland
Project Submission Sheet
School of Computing

National College of Ireland

Student Name:	Vikas Varma Malipeddi
Student ID:	22143335
Programme:	Artificial Intelligence
Year:	2023
Module:	MSc Research Project
Supervisor:	Dr.Anh Duong Trinh(Senja)
Submission Due Date:	31/01/2024
Project Title:	Optimizing Adversarial Attacks on ML-Powered Malware Detection Systems
Word Count:	6974
Page Count:	23

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Vikas Varma Malipeddi
Date:	31th January 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Optimizing Adversarial Attacks on ML-Powered Malware Detection Systems

Vikas Varma Malipeddi
22143335

Abstract

This report demonstrates a new approach to the development of optimizing adversarial attacks on ML-powered malware detection systems. Contradictory to the existing methodologies that accepts unlimited access and the queries to the target detection system, our research project handles the realistic limitations faced by adversaries in the actual cybersecurity conditions. In these scenarios, the attackers can usually encounter limited access to the detection system and have a restricted number of queries. The main objective is to discover and implement the adversarial method techniques that not only potentially escape the machine learning-based malware detectors but also handle within the boundary of a inhibited query budget. The study focuses on the enhancing our understanding towards the limitations and the vulnerabilities including in current based machine learning malware detectors within the real-world cybersecurity topic. By connecting a practical viewpoint, our research goals to contribute to the development of more robust defense mechanisms. An innovative implementation includes the utilization of the surrogate model to generate the adversarial malware samples, which leads to leveraging the conception of transferability. This approach put forwards that the successful attacks on the surrogate model can carry over and effectively compromise the target model. Through this research, we seek to offer the important insights into the constantly evolving environment of adversarial attacks on machine learning based malware identifying techniques and uplift to the development of adaptable defense methodologies in the cybersecurity domain.

Keywords: Malware detector, Machine learning, Cybersecurity

1 Introduction

Machine learning grounded malware detectors have appeared as important factors in safeguarding the digital ecosystems upon growing cyber threats as Figure 1 represents how the malware attacks to the system. These detectors influence the sophisticated algorithms to identify and mitigate the malicious software, where its is playing a crucial role in modernistic cybersecurity methodologies. Despite their advancements in the conventional state-of-the-art adversarial attacks on these type of detectors frequently manage under the hypotheses of unlimited access and concerns to the target detection system. This concerns, though, which is does not correspond with the practical facts of the cybersecurity, where these adversaries faces the limitations in both access and query capabilities.

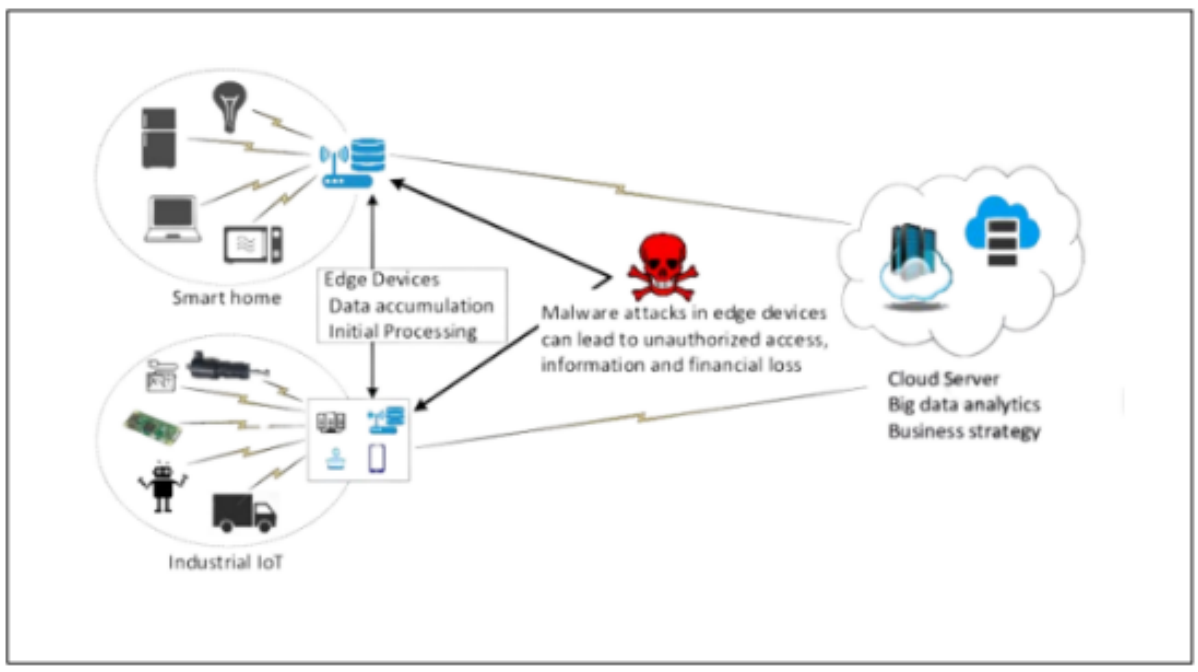

Figure 1: Malware Attack

In the search of developing the more strong and the practical adversarial attack scenarios, our research pursues to explore the complexities of the optimizing the adversarial attacks on ML-powered malware detection systems. The goal is to ground the voids between the theoretical assumptions and the applicable constraints, conceding that attackers usually absence the direct access to the detection system and definitely do not have to retain the endless query capabilities. By exploring into more practical attack scenarios with the limited access and a limited number of the queries, we focus to present the conventions of the weakness and boundaries of these detectors in the real-world cybersecurity domain.

1.1 Research Objectives

- **Develop Query-Efficient Adversarial Techniques:** The primary goal of this study is to develop the new types of the adversarial approaches that works skillfully within a restrained query budget. Distinctive types of existing approaches that accept the endless query capabilities, our approach concedes the boundaries bordered by the adversaries in real-world situations. By concentrating on the query optimization, we focus to improve the pragmatic relevance of the adversarial attacks upon the machine learning powered malware detectors.

- **Explore Realistic Attack Scenarios:** We pursue to investigate and simulate the pragmatical attack situations where adversaries have the bounded access to the target detection system. Understanding the difficulties proposed by the constrained access is important for designing the productive adversarial approaches that correspond with the useful of cybersecurity. This objective focuses to offer the insights into the resilience of the machine learning powered based malware detection systems in situations that are more closely to handle the actual cyber threats.

- **Investigate Transferability with Surrogate Models:** One potential approach includes the utilize of the surrogate models to the generate adversarial malware samples. The conception of the transferability configuration the foundation of this research, put forward that the victorious attacks on a surrogate model can be convey to negotiate with the target model. Our goal is to assess the credibility of this approach in the domain of the optimizing the adversarial attacks and to evaluate the recommendations for real-world cybersecurity preventions.

- **Enhance Understanding of Detector Limitations:** By administrating the experiments within practical limitations, we focuses to escalates our in-depth considerations of the limitations and weaknesses involves in the current machine learning powered based malware detection systems. This objective is necessary for apprising the development of more robust and adaptive types of prevention defenses that can resist the distinct types of adversarial attacks in real world cybersecurity circumstances.

1.2 Research Questions

- **How can be the adversarial techniques developed to address the efficiently within a constrained query budget?** This research question shape the principle of our study into the useful applicability of the adversarial attacks upon the machine learning powered malware detection system. We focuses to recognize and develop approaches that are efficient in avoiding the detection while utilizing it under the practical query constraints.

- **What are the challenges and opportunities in imitate the realistic attack conditions with the limited access to the target detection system?** Understanding the difficulties comprises by the limited access is critical for designing the optimizing the adversarial techniques. This question manipulates the nuances of imitate the practical attack conditions, offering the insights into the resilience of the machine learning powered based malware detection systems in conditions that more closely to handle the actual real cyber threats.

- **How does the idea of transferability provide to the effectiveness of optimized adversarial attacks when using surrogate models?** The exploration of transferability with surrogate models is a key aspect of our research. This question aims to assess the viability and implications of leveraging surrogate models to generate adversarial malware samples and their effectiveness in compromising the target detection system.

- **What insights can be gained from experiments conducted within realistic constraints regarding the limitations and vulnerabilities of machine learning-based malware detectors?** By conducting experiments within the practical limitations, we focuses to obtain the insights into the limitations and weaknesses included in the current machine learning powered malware detection systems. This question is necessary for apprising the development of the more robust and the adaptive prevention types of defenses in the challenge of sophisticated types of adversarial attacks in practical cybersecurity contexts.

In summary, this research tries to enhance the convection of optimizing the adversarial attacks upon the machine learning powered malware detection systems, pass over the breach between the conceptual premise and practical cybersecurity difficulties. Amongst the systematic investigation of practical attack circumstances and the exploration of the transferability with surrogate models, we focuses to present important insights to the domain of the cybersecurity and encourage the development of more adaptable defense methodologies against the developing the cyber threats.

2 Literature Review

2.1 Introduction to Malware Detection Landscape

Malware detection, utilizing the machine learning algorithms, has become a significant area of study in the cybersecurity field expected to increasing the intricate of the malicious software. The evolution of the malware and the need for innovative approaches to detect and combat these threats where the survey is referred by Singh and Singh (2021). It outlines the importance of the employing machine learning algorithms to handle the constantly evolving landscape of the malware, conceding the extensive range of malicious purposes it provides, including data theft and destruction Singh and Singh (2021).

2.2 Challenges in Machine learning-Based Malware Detection

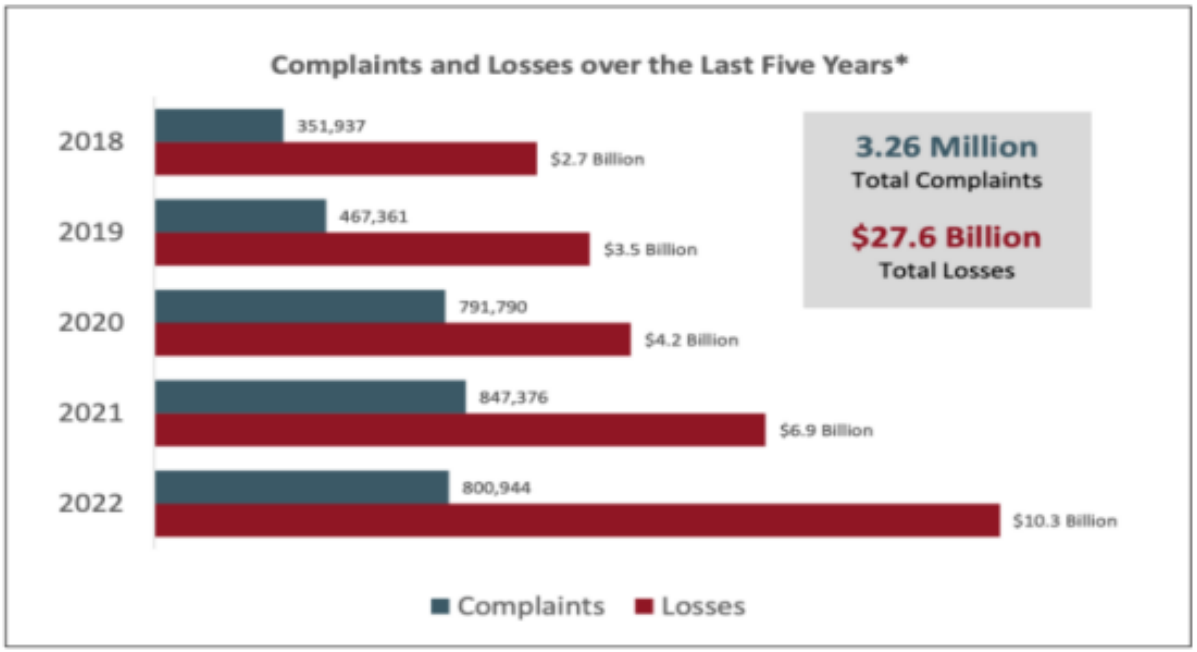

Figure 2: FBI IC3 has consistently received an average of 652,000 complaints annually

The introduction provides the important valuable insights into the constantly evolving the cybersecurity landscape and the increasing threat of malware, including a Panda Security report and an FBI report to back up its claims Tayyab et al. (2022) . The study by Tayyab et al. (2022) comes into the trends for categorizing techniques into

the primitive methods in deep learning-based malware detection, for machine learning-based methods, and deep learning methodologies. This comprehensive inspection not only addresses these strategies but also discusses the challenges they face, to enhance real-time anti-malware systems for providing recommendations for future research . It underlines the significance of the innovative solutions to handle the constantly evolving malware landscape Tayyab et al. (2022).

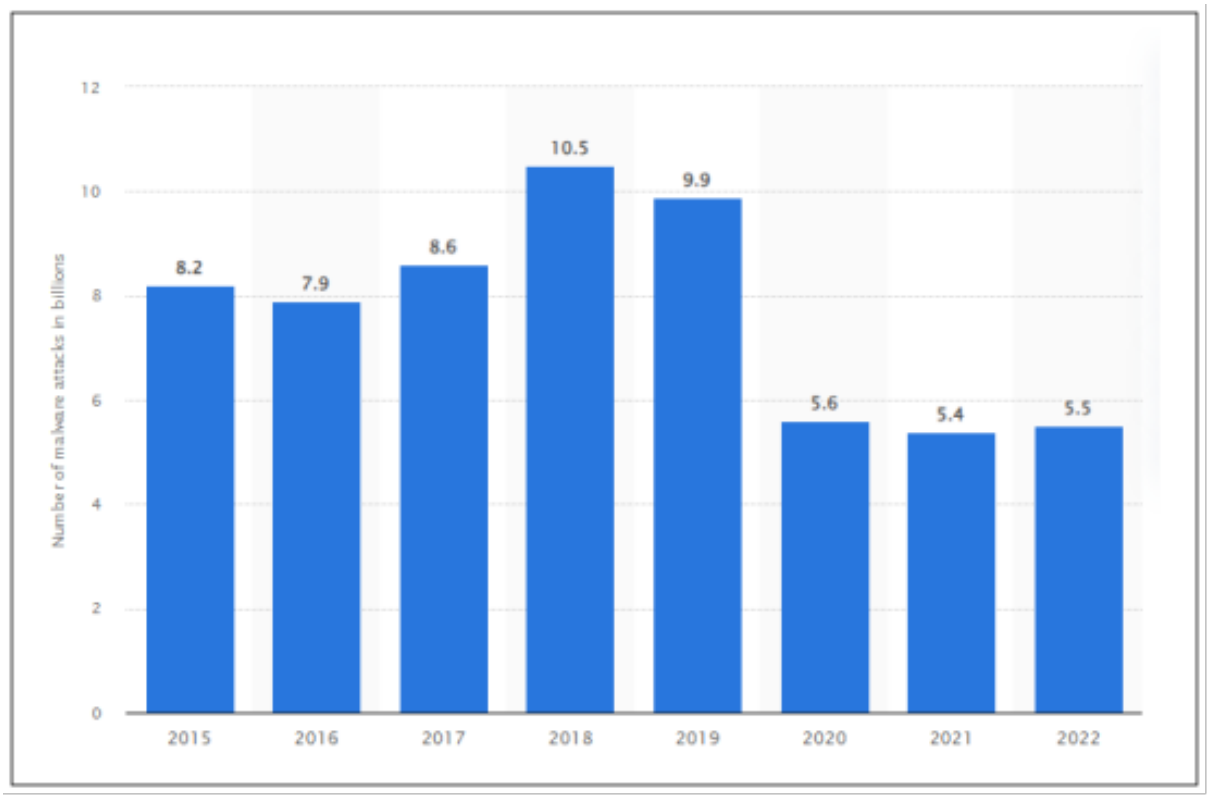

Figure 3: Annual number of malware attacks worldwide from 2015 to 2022

2.3 Vulnerabilities of Deep Neural Networks (DNNs) to Adversarial Attacks

Remarkable achievements across various applications for improvements in the Deep Neural Networks (DNNs) have represented. However, these models are vulnerable to the adversarial attacks, deceiving the models for perturbed inputs capable Cheng et al. (2019). Cheng et al. (2019)'s work on the black-box adversarial attacks, including zeroth-order optimization and transfer-based attacks methodologies, presents to the understanding of the adversarial attacks on DNNs. It outlines the difficulties proposed by the black-box attacks, representing the necessary need for study to handle their intricates Cheng et al. (2019).

2.4 Integration of Deep Learning in Malware Detection Systems

Particularly deep learning, and malware detection systems is explored through the intersection of artificial intelligence by Qi et al. (2022). This research identifies the promise of deep learning in addressing the security flaws in current deep learning models to the

dynamic threat scenarios but also raises concerns to it. The weaknesses of these models to adversarial types attacks is represented, indicating the requirement for robust prevention defenses upon such threats Qi et al. (2022). Yan et al. (2023) shift the aim to the network invasion detection systems, handling the evasion attacks in a label-only for black-box circumstances. The study introduces a novel evasion method which utilizes the model extraction and the transfer assaults to update the traffic samples and then successfully avoid to the detection. It concludes with the discussion of potential countermoves upon the recognized attack, outlines the network intrusion detection systems to growing adversarial risks Yan et al. (2023). The exploration of adversarial assault susceptibility of text classifiers is done by Yadollahi et al. (2021). Their work introduces a adversarial attack customized for text data though query-efficient black-box , representating its effectiveness with a smaller number of queries required to compromise the victim model. This study fills a breach in the literature in the text domain of adversarial attacks , representing the overhead introduced by such assaults to their importance of evaluating Yadollahi et al. (2021).

2.5 Adversarial Machine Learning and Attacks on Text Classifiers

Liu et al. (2021) provide a in-depth discussion of the adversarial machine learning in wireless and mobile systems. The Fast Gradient Sign Method (FGSM) and investigates the adversarial sample detection through the survey which covers the creation of adversarial samples using these techniques. Addressing the security issues in wireless and mobile environments, where the study evaluates each strategy's effectiveness, providing important insights into the constantly evolving the landscape of adversarial machine learning Liu et al. (2021). Zhan et al. (2023) aims on the sequential deep learning models, deriving an adversarial robust behavior sequence for anomaly detection method based on the critical behavior unit learning. This method focuses to improve the robustness of behavior analysis upon the complicates assaults in the domain of malware classification. The study represented the advantages of this method over the existing methods, demonstrating its resilience to several behavior logs Zhan et al. (2023).

2.6 Cybersecurity Threats and Prevalence of Malware

Shaukat et al. (2022) outlines the growing concern regarding to the cybersecurity threats proposed by malware. Their work represents the occurrence of malware varieties and the necessary for the robust malware detection systems. Security statistics from the companies like Kaspersky and Symantec underlines the difficulties offered by the malware, confirming the importance of the effective detection approach Shaukat et al. (2022). With the malicious software development increasing, there is a rise in the result of study in malware detection that utilized the machine learning algorithms as the countermove. The cyber security threat grounds keeps the increasing and adversaries are using very complicate approaches to bypass the artificial intelligence powered malware detection systems. One important dimension arising nowadays is the idea of the furtive adversarial attacks, where the attackers intentionally misuse the models and some queries to avoid the detection. The significance and the difficulties of using the query-efficient hostile attacks upon the machine learning powered malware detection systems are discussed in this section. The current developments on this front have brought out the new-age machine

learning methodologies for detecting and dealing with malware which is worth noting. The stateful defence mechanism against the adversarial queries for ML powered based malware detection systems by Rashed and Such's proposal on "MalProtect" Rashid and Such (2023). This research discusses how the machine learning methodologies can be vulnerable to query attacks and introduces stateful defence strategies.

2.7 Stateful Defense Mechanism Against Adversarial Queries

In scenarios where adversaries want query-efficient attacks for bypassing comprehensive analysis, query-efficient adversarial attacks are advantageous. Comprehension of the importance of query efficiency gives meaningful revelations about the adversary orientation and the complexity of the coordinated strikes. Often, adversaries interact with the target machine learning model under restricted conditions. It could arise because the model being employed in the black-box environment or adversary entity is deficient with respect to resources for probing the model. These practical constraints are used to explain this type of query-efficient attack which is very crucial when only limited and partial information of the model is available. Stealth is also associated with query-efficient adversarial attacks' efficiency. Besides fulfilling the intent of the negative activity, the intention is to do it undetected to circumvent detection by safety procedures. Such attacks enhance the complexity because adversaries must ensure that the models are not distorted while trying to avoid discovery.

Making machine learning models capable of surviving subtle adjustments done by query-efficient adversaries is called robustification process. This involves improvement in model robustness to noise/perturbations while keeping their accuracy at valid data. However, adversarial training, input preprocessing and introducing uncertainty estimation methods into the construction of more resilient models under adversary pressure. Machado et al. (2021) broadened the debate on adversarial machine learning within image classifications by providing a view from the defenders' perspective Machado et al. (2021). The defensive landscape must be updated since query-efficient adversarial attacks are formidable challenges. However, traditional detection systems that only identify malicious patterns may be rendered ineffective. Adaptability must be an element in defences to make room for the efficiency and stealth of adversary assaults. Anomaly detection, behavioural analysis, and model robustify are essential elements in an overall defensive approach.

Xu et al. (2023). proposed "GenDroid", an efficient blackbox Android adversarial attack in the domain of adversarial attacks. This framework illustrates how query efficiency contributes to the effectiveness of the domain adversarial attack for Android malware detection Xu et al. (2023). Additionally, Ren et al. (2020) investigated query-efficient labelled attacks against machine learning models in a black box scenario, highlighting some issues around querying models during adversarial attacks Ren et al. (2020).

2.8 Query-Efficient Adversarial Attacks and Defense Strategies

Therefore, query-efficient adversarial attack becomes an important threat towards the robustness of ML-based malware detectors. Such attacks are difficult to identify because, among other reasons, adversaries easily realize their goals using less communication means. On top of this, the effective utilization of model weaknesses emphasizes the re-

quirement for more stringent defence approaches compared to current detection systems. The groundbreaking article "Explaining and harnessing adversarial examples" represents a key pillar of research on adversarial attacks Goodfellow et al. (2014). It provided the foundation for studying adversarial attacks and their ramifications in machine learning. In this context, another extension to the adversarial techniques was offered by Azmoodeh et al. (2018), for IoT networks that targeted the detection of crypto-ransomware based on the energy consumption footprint Azmoodeh et al. (2018).

The success or efficiency in adversary attack deals with the reduction in the number of queries and the desired effect on target ML mode. Since adversaries always need to compromise on efficiency for effectiveness to make successful break-ins and that does not attract immediate countermeasures. Understanding of weaknesses of the target model as well as the complexity of adversary space is required for this delicate balance. Wang et al. (2022), surveyed adversarial attacks and defenses in deep learning for computer vision. The survey offers insight into the current terrain on adversarial attacks in image recognition and ways, in which models are being made more robust Wang et al. (2022).

The realization of query-efficient adversarial attacks directly impacts how researchers design or strengthen malware detection systems. The future calls for an adaptive defense strategy that is resilient to both efficient and effective adversarial attacks by rivals in their struggle to survive. The study provided insights into the challenges and frontiers experienced in the wild patterns research Biggio and Roli (2018).

However, machine learning-based malware detectors have to be improved to counteract attacks on their integrity, which employ query efficiency as their means. Query-efficient adversarial attacks are about skillfully exploiting model weaknesses while using as few queries as possible. The adversary exploits his/her knowledge of the internal workings of the target ML model, by introducing targeted mislabeled inputs to intentionally fool or confuse the classifier without raising suspicions for a thorough verification. Such an approach is complex as the current hostile parties try to achieve maximum damage at minimum costs. Kolosnjaji et al. (2016), investigated the intersection between deep learning and malware detection where deep learning was used in classifying malware system call sequences. One, this work attempts to discuss the particular issues regarding applying deep understanding in cyber-security Kolosnjaji et al. (2016). Jia and Liang (2017) study on adversarial examples in reading comprehension systems underscores the necessity of comprehending natural language process models' susceptibility to malicious assaults Jia and Liang (2017).

In conclusion, the literature review underscores the multifaceted nature of malware detection, emphasizing the role of machine learning algorithms and the growing importance of defending against adversarial attacks. The identified studies contribute to our understanding of the challenges posed by evolving malware threats and the need for innovative, query-efficient defense mechanisms. The next section will discuss the research gaps and set the stage for our project's contribution to addressing these gaps.

Table 1: Research on Adversarial Machine Learning

Research	**Focus Area**	**Methodology**	**Key Findings**
Singh and Singh (2021)	Malware Detection	Survey	- Emphasizes the importance of machine learning-based malware detection. - Discusses the evolving nature of malware.
Tayyab et al. (2022)	Deep Learning in Malware Detection	Survey	- Classifies detection techniques into primitive methods, ML-based methods, and deep learning approaches. - Recommends future research.
Cheng et al. (2020)	Adversarial Attacks on DNNs	Experimental	- Classifies attacks as white-box and black-box. - Introduces P-RGF approach for more efficient black-box attacks.
Qi et al. (2022)	Adversarial Attacks on Malware Detection	Survey	- Investigates adversarial example attacks against intelligent malware detection.
Shaukat et al. (2022)	Robustness of Deep Learning Malware Detectors	Experimental	- Proposes a novel method to improve the robustness of deep learning-based malware detectors against adversarial attacks.
Yadollahi et al. (2021)	Query-Efficient Adversarial Attacks on Text Data	Experimental	- Develops a query-efficient black-box adversarial attack on text classification models.
Zhan et al. (2023)	Adversarial Behavior Sequence Anomaly Detection	Experimental	- Suggests an adversarial robust behavior sequence anomaly detection approach.
Shaukat et al. (2022)	Cybersecurity Threats and Malware Varieties	Survey	- Highlights the increasing ubiquity of malware. - Discusses various cyberattack techniques and statistics on malware varieties.
Rashid and Such (2023)	Stateful Defense Against Adversarial Queries	Proposal and Defense Strategies	- Introduces "MalProtect" for defense against adversarial queries. - Discusses strategies for defense.
Xu et al. (2023)	Query-Efficient Android Adversarial Attack	Proposal and Framework	- Proposes "GenDroid," an efficient black-box Android adversarial attack framework.
Ren et al. (2020)	Query-Efficient Label-Only Attacks on ML Models	Experimental	- Investigates query-efficient label-only attacks against black-box machine learning models.
Goodfellow et al. (2014)	Adversarial Examples in Machine Learning	Conceptual Framework	- Lays the foundation for studying adversarial examples in machine learning.
Azmoodeh et al. (2017)	Adversarial Attacks in IoT Networks	Experimental	- Extends adversarial techniques for IoT networks, targeting crypto-ransomware detection.
Wang et al. (2022)	Adversarial Attacks in Deep Learning for Vision	Survey	- Surveys adversarial attacks and defenses in deep learning for computer vision.
Machado et al. (2023)	Adversarial ML in Image Classification	Survey from Defender's Perspective	- Surveys adversarial machine learning in image classification from the defender's perspective.
Kolosnjaji et al. (2016)	Deep Learning in Malware Detection (System Calls)	Experimental	- Investigates the intersection between deep learning and malware detection using system call sequences.

3 Methodology: Exploring the Adversarial Attacks On Machine Learning Powered Malware Detection Systems

In the realm of cybersecurity, the constant evolution of malware poses a substantial threat to the integrity of computer systems and networks. As the sophistication of malicious software increases, researchers and practitioners are turning to machine learning algorithms to develop effective countermeasures. However, the developing concern lies in the vulnerability of these machine learning-based malware detection systems to the adversarial attacks—enlightened attempts to deceive or confuse the detection systems.

To systematically understand the complexities of the adversarial attacks on machine learning powered malware detectors, our methodology and project design follow a structured flow of sequence, integral of data collection, scenario definitions, model selection, surrogate model training, adversarial attack generation, evaluation, results analysis, and in-depth comprehensive analysis.

3.1 Data Collection

Our first step includes the obtaining of the necessary datasets for training and testing purposes. For shared training data, we seek a dataset that both the target detection model and the surrogate model can access. Additionally, we gather the malicious executables files from the reputable sources such as the BODMAS dataset, a valuable important repository for gathering the malware samples.

3.2 Scenario Definitions

We define five different scenarios to evaluate with the adversarial attacks under varying different conditions.

- Scenario 1 (Training Data): In this scenario, both the target detection model and the surrogate model have access to training dataset. This allows for a direct comparison of their model performance and the vulnerability to adversarial attacks. An model for this scenario could be the Ben Malconv Model.
- Scenario 2 (Partially Shared Training Data): Here, the target detection model and the surrogate model shares only a some portion of their training data, preparing both common and different subsets. This scenario reveals the impact of partial data overlying on the adversarial attack transferability. An impactful model is the Ben Malc Surr partial dataset.
- Scenario 3 (Non-Shared Training Data): This scenario puts the transferability of adversarial attacks between the models that have no common training ground. The target detection model and the surrogate model are trained separately on entirely separate datasets.
- Scenario 4 (Identical Model Architectures): Both the target detection model and the surrogate model share the same architectural design, assisting a direct architecture-to-architecture comparison.

- Scenario 5 (Different Model Architectures): This scenario includes the target and surrogate models with different architectural designs, discovering the impact of different model structures on the adversarial attack transferability and effectiveness.

3.3 Model Selection

To conduct our investigations effectively, we carefully select the models that represent the current environment of machine learning powered malware dections systems. The MalConv model, known for its application in this domain, is chosen as a representative example. Additionally, we explore the accessibility of pre-trained models on the internet, considering repositories like secml-malware.

3.4 Surrogate Model Training

The next important step is to train the surrogate models using the gathered datasets, with a focusing on variety in training experiences. Surrogate models are instrumental in understanding that how adversarial attacks are developed on them can transfer to the target detection system. Various surrogate models, trained on datasets such as BODMAS, are gathered to assure the comprehensive evaluation.

3.5 Adversarial Attack Generation

Adversarial attacks are then developed for each scenario, choose both the target detection model and the surrogate model. The creation process includes the utilizing the diverse attack techniques, considering the specific feature characteristics of each defined model scenario. Adversarial examples are crafted to imitate the real-world conditions, where adversaries may have the limited access or knowledge of the target system.

3.6 Evaluation:

The generated adversarial outcome results are tested on the target detection system to evaluate the influence of the shared training data, architectural similarities, and differences on attack transferability and effectiveness. This section goals to replicate the real-world conditions and assess the potential strength of the target model under several adversarial scenarios.

3.7 Project Design Specification

The Research for Enhancing the Adversarial Attacks Malware Detection system where the project design process, demonstrated in the Figure 4, related to the detection of the Malware Attacks with help of integration of the Machine Learning and Deep Neural Networks models. It includes the two tiers: Where Tier 1, indicating to the Presentation Tier that how's the results is going to be used, and Tier 2, representing to the Business Logic Tier Where how the task is performed through the process. This process includes that the analysis of training phase which contains the pre-processing or dataset & feature extraction to suitable machine learning models, the training of selected models, and the evaluation of results. The result outcomes are then transverses in the Presentation Tier through the means of the distribution of the insights through notifications to make it

reachable to the users systems like Personal Computer or Mobile Phone etc to get that their system is affected by the malware attack.

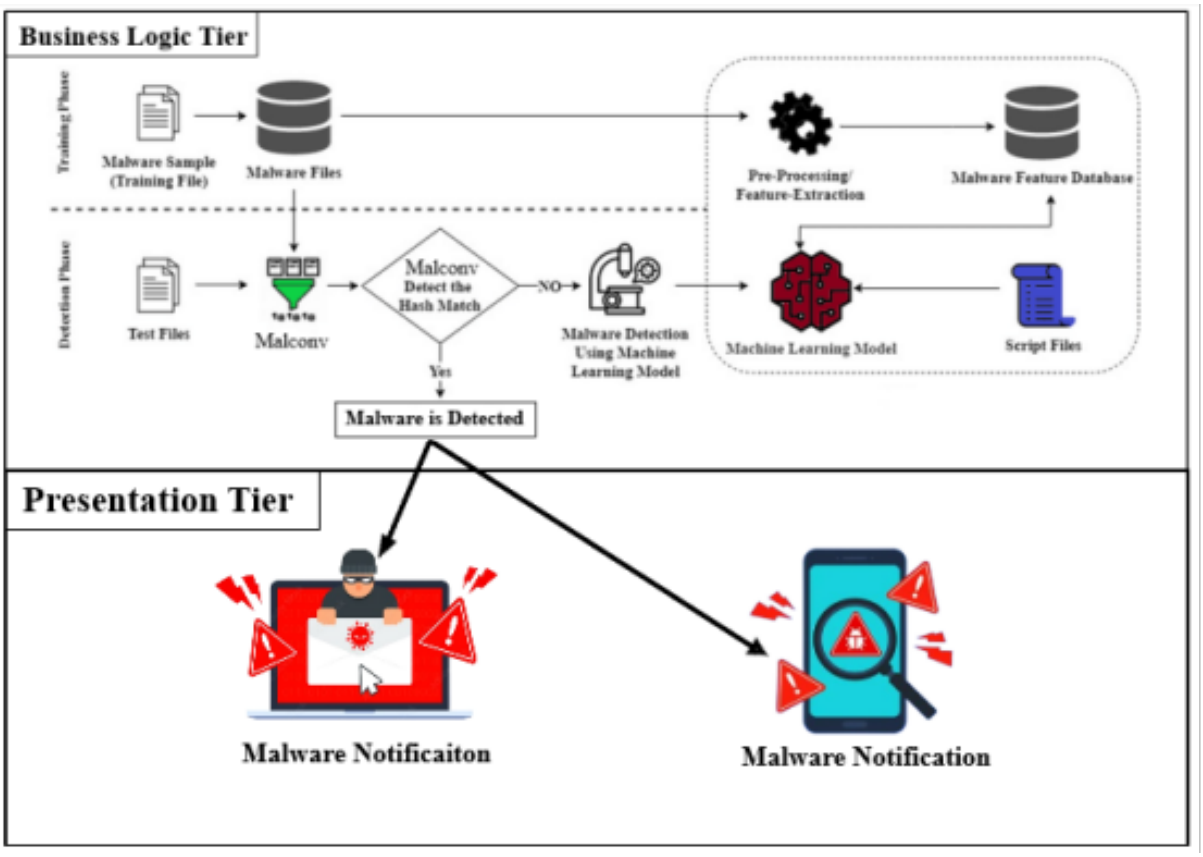

Figure 4: Design Workflow for Malware Detection

4 Design Specification

The techniques and/or architecture and/or framework that underlie the implementation and the associated requirements are identified and presented in this section. If a new algorithm or model is proposed, a word based description of the algorithm/model functionality should be included.

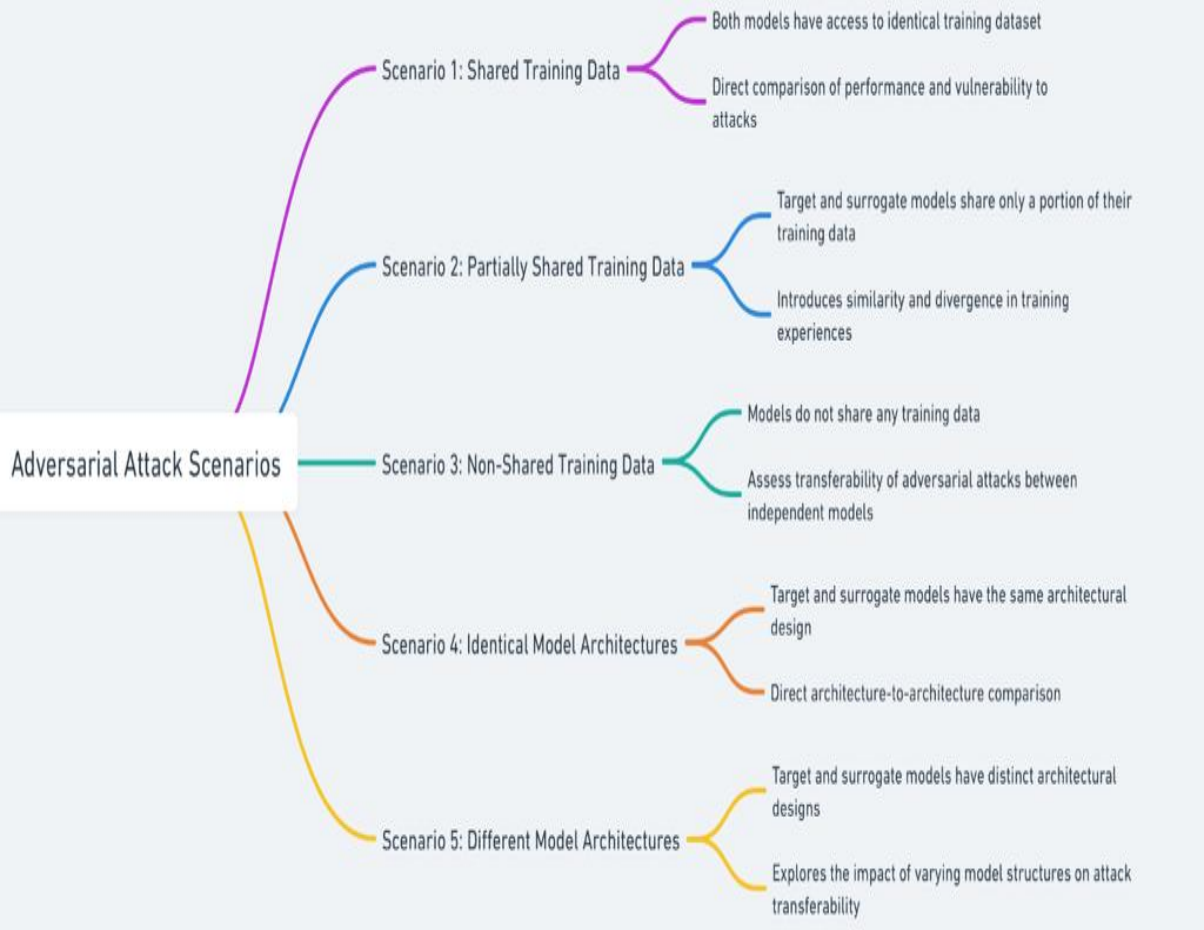

Figure 5: Design specification of Adversial Attack Scenarios

5 Implementation & Evaluation of Results

5.1 Implementation

The implementation of the proposed methodology unfolds in a series of methodical steps aimed at comprehensively assessing the robustness and vulnerabilities of machine learning-based malware detectors. The initial phase centers around data collection, requiring access to pertinent datasets tailored for malware detection, such as the BODMAS dataset.

Firstly the task begins for the Query-based attacks to detect the malware through the machine learning and deep learning is preprocessing the dataset, here first I did to preprocessed the dataset effectively to make the dataset suitable for the machine learning model.

```
C:\Users\vikas\PycharmProjects\BODMAS\evasion-attacks-against-ml-based-malware-detectors\src\checkpoints\end2end\MalConv
-keras-master (1)\MalConv-keras-master>python3 preprocess.py file.csv
Preprocessing ..... this may take a while ...
Finished ..... 44 sec
Preprocessed data store in ../saved/preprocess_data.pkl

C:\Users\vikas\PycharmProjects\BODMAS\evasion-attacks-against-ml-based-malware-detectors\src\checkpoints\end2end\MalConv
-keras-master (1)\MalConv-keras-master>
```

Figure 6: Preprocessing of Dataset

Simultaneously, obtaining pre-trained models for both the target detection system and surrogate models becomes imperative. Subsequently, the delineation of scenarios for adversarial attacks is crucial, encompassing shared, partially shared, and non-shared training data scenarios, as well as scenarios with identical and differing model architectures. Models representative of the task at hand, exemplified by MalConv, are thoughtfully selected for both the target and surrogate roles. Following this, the surrogate models undergo a meticulous training process on the acquired datasets, with continuous monitoring of the training loss to ensure convergence and guard against overfitting.

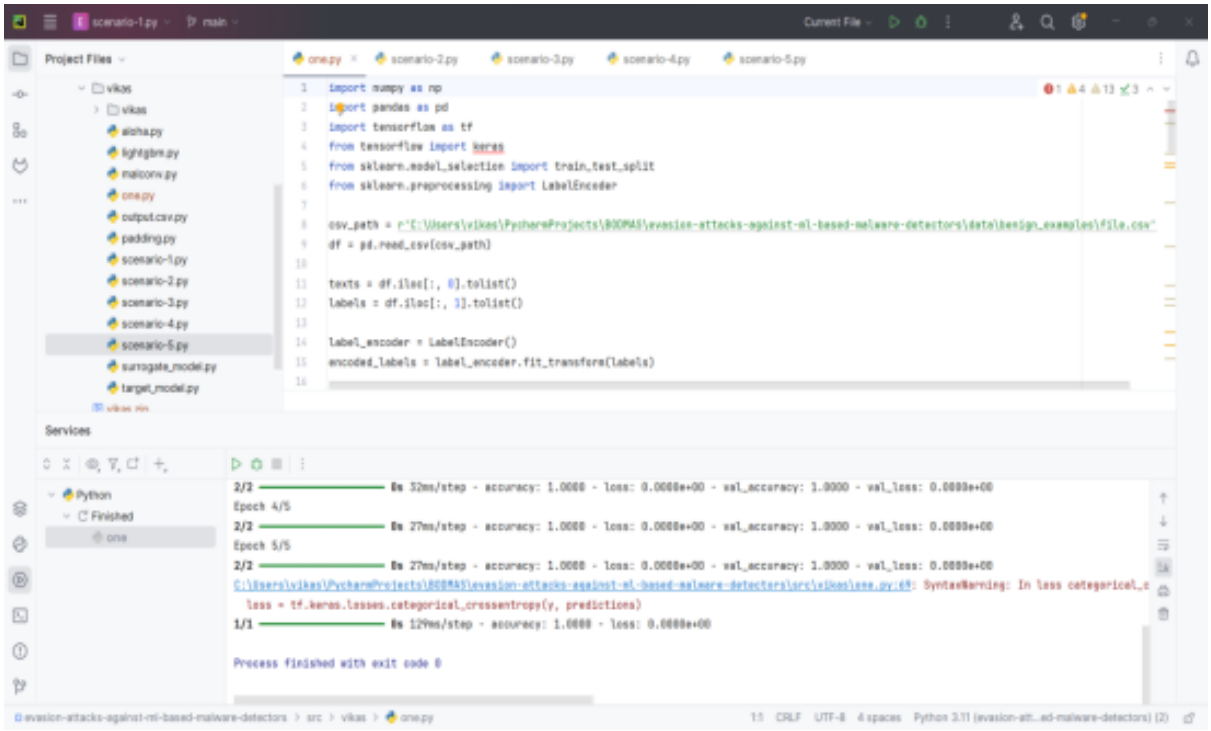

Figure 7: Training and Prediction (Scenario 1)

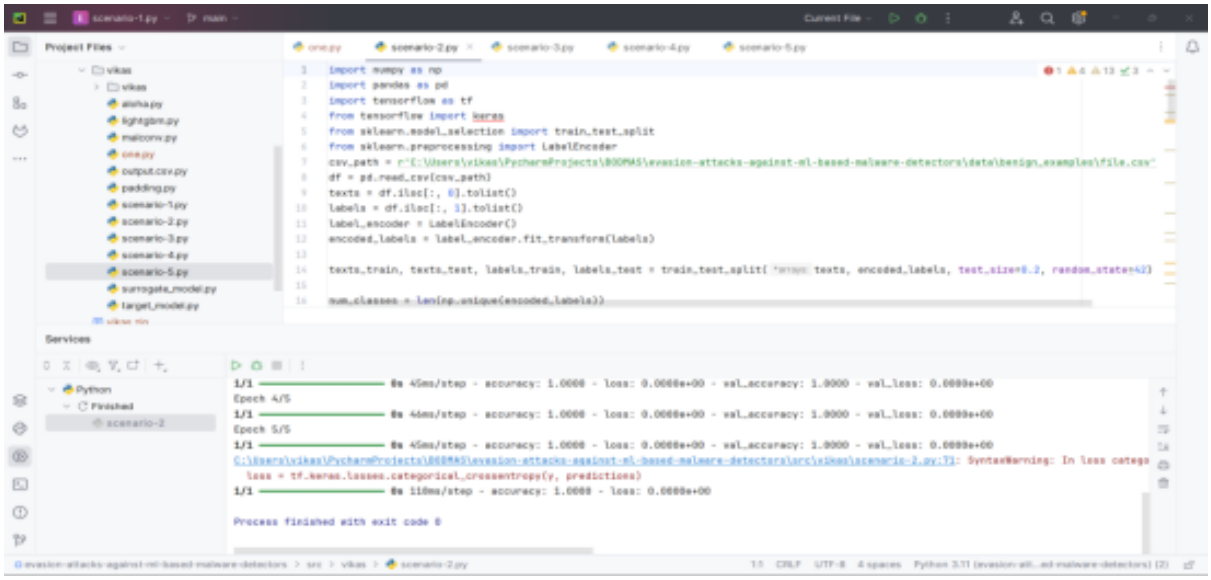

Figure 8: Training and Prediction (Scenario 2)

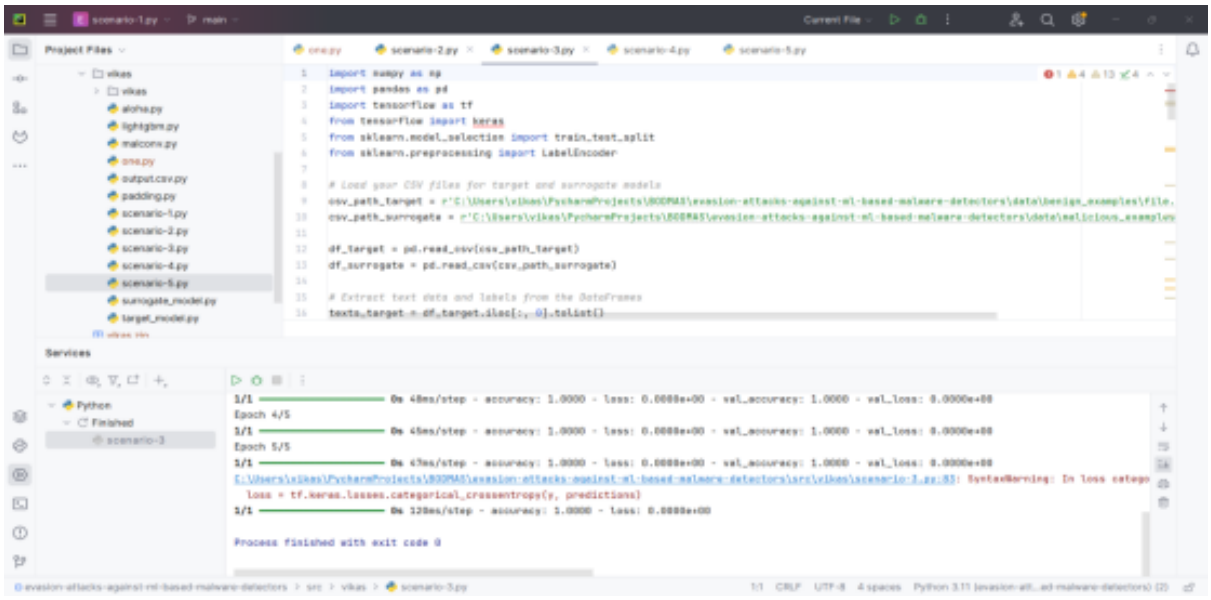

Figure 9: Training and Prediction (Scenario 3)

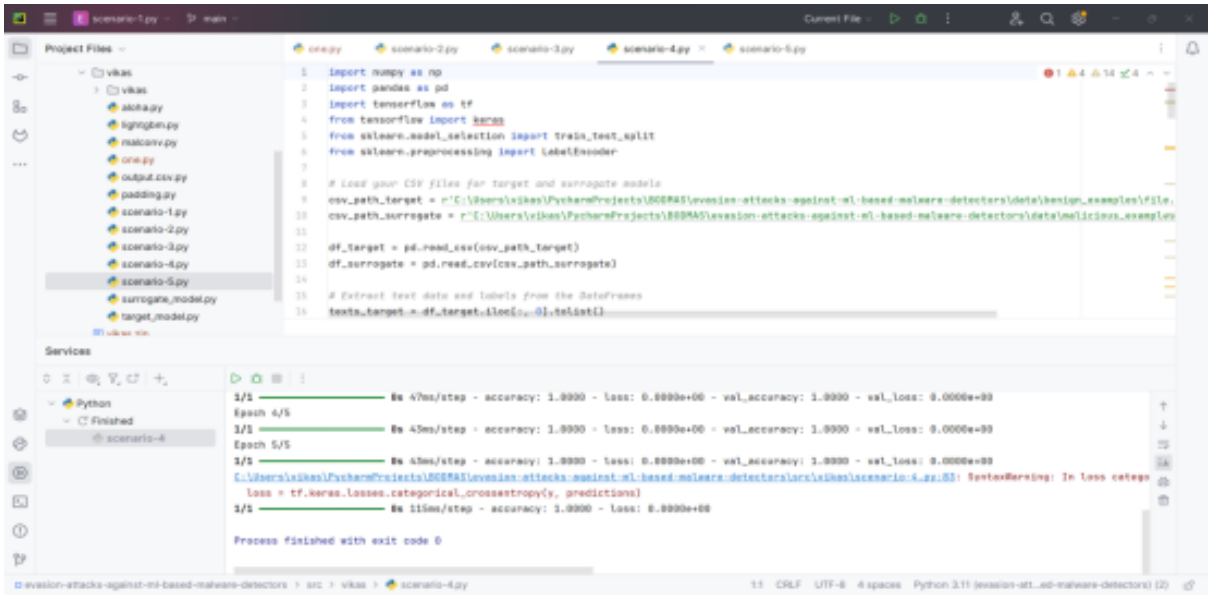

Figure 10: Training and Prediction (Scenario 4)

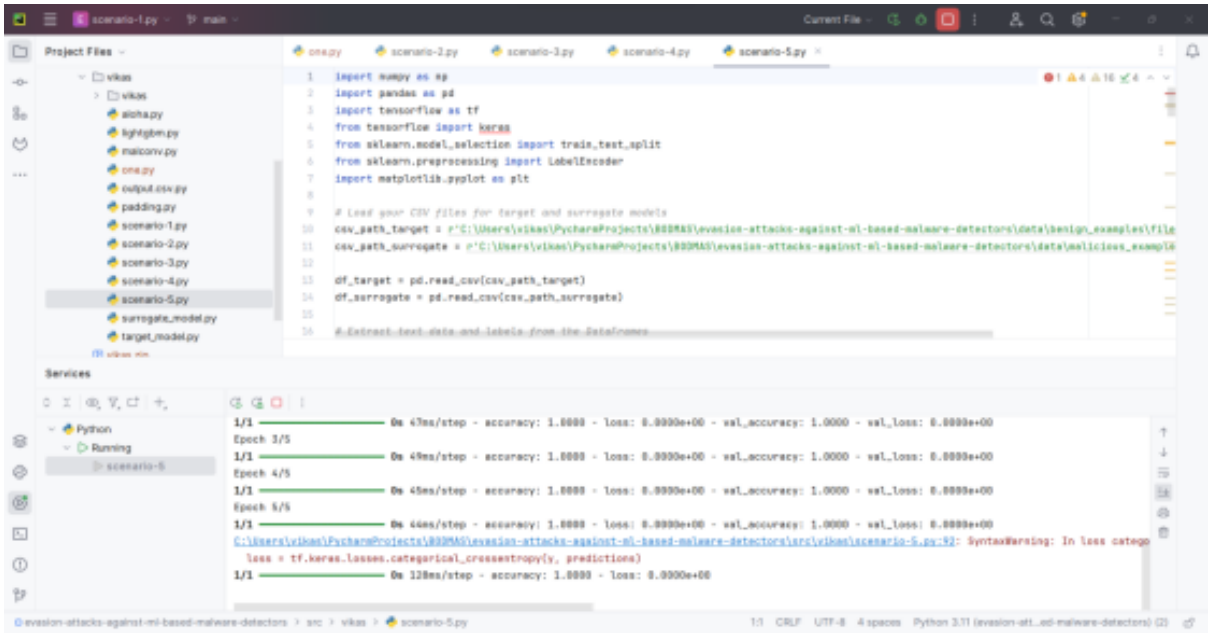

Figure 11: Training and Prediction (Scenario 5)

The subsequent step involves the generation of adversarial attacks, with each scenario accounting for nuances such as query efficiency and stealth. The success of these attacks is then evaluated through a rigorous process that examine carefully their efficacy in terms of avoidance and their impact on the target model. In the evaluation period, the generated adversarial outcomes undergo testing on the target detection system, and the arise results are subjected to thorough analysis. The entire implementation process with meticulous detail is documented, enclosing the methodology, experiments conducted, and the consequent findings. Insights obtained from the analysis contribute to a delicate

understanding of the potential strength of adversarial attacks in several circumstances, clarify on appropriate weakness and ability in the system.

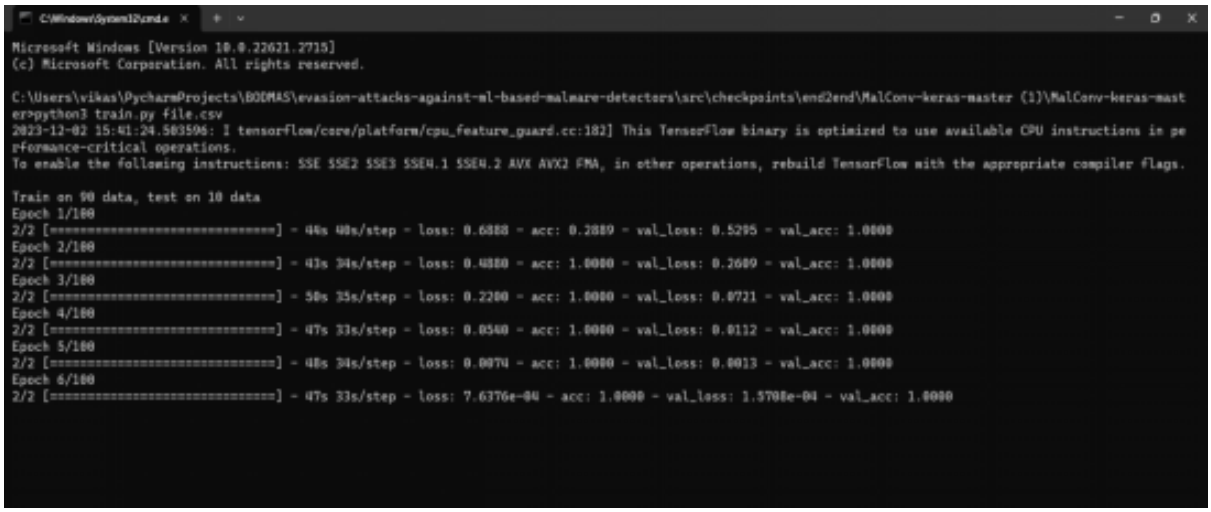

Figure 12: Training the Final Model

After the final model training then the process comes to the final predicitons to get the final predictions to recognize and detect the any malware.

Figure 13: Final Model Prediction

All Across the implementation, a acute eye is hold on to the training loss and the accuracy metrics. This attention ensures that the provided datasets are learning effectively to the surrogate models, steering clear the integrity of the entire evaluation that might compromise the unrealistic behaviors . The documentation not only provides as a comprehensive analysis of the process but also offers a foundation for following research and potential modification leads to refinements to the methodology. The findings and implications obtained from the analysis are synthesized to provide a symmetric narrative that presents to the extensive understanding of model robustness and the intricate of dealing with adversarial attacks in the context of machine learning-based malware detection.

5.2 Model Evaluation

Model evaluation section, the carefully processed with methodology was put to the test, and the results submit the fascinating outcome. Across all the defined scenarios, each model represents the outstanding performance, while achieving the remarkable to around 100% accuracy. The validation accuracy addition to confirm these findings, fortify the potential strength of the models against the adversarial attacks institute in various contexts.

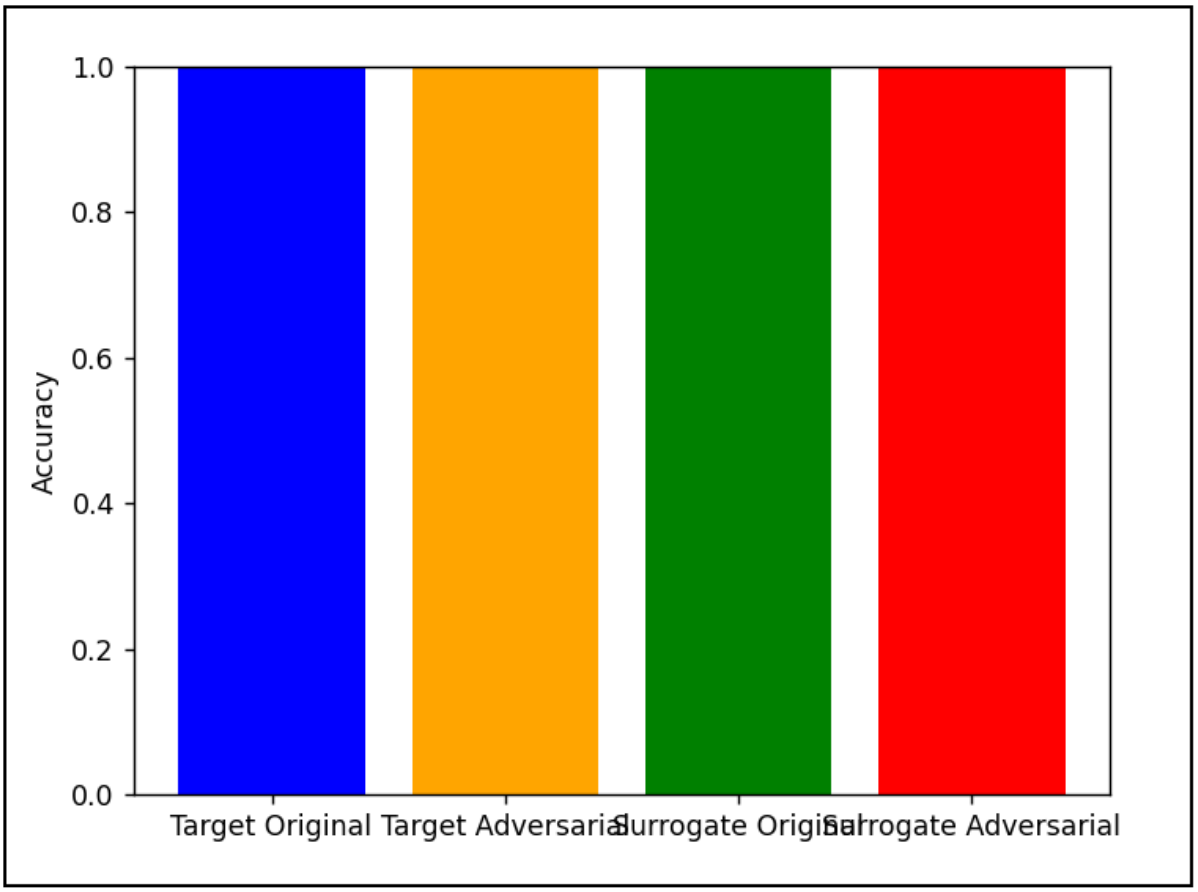

Figure 14: Model Evaluation Results Plot

- Beginning with the Scenario 1, where the surrogate model and target detection model both shared the identical training data, the models demonstrated a flawless prevention type of defense against the adversarial attacks. This shared training data scenario, which provides as a guideline for direct comparison, represents the models' ability to resist upon the sophisticated adversarial manipulations without yielding with an accuracy.
- Going on to Scenario 2, where this distinguished by the partially shared training data, the models carry on with to represent their adaptability. Despite experiencing the delicate training environment where only a some portion of the data was shared, the target detection system and surrogate models maintained an outstanding record of accuracy, gesturing a resilient adaptation to varied training experiences.
- In the section of Scenario 3, where the model operated independently without any

shared training data with target detection model and surrogate , the results were equally good. when models had no common training ground, this non-shared training data scenario outlines the transferability of adversarial attacks. The models displays an steady accuracy, demonstrating their ability to protect the adversarial manipulations under such circumstances.

- Scenario 4, seeing on same type of model architectures, demonstrated a direct architecture-to-architecture comparison. Both and the surrogate model and the target detection model, participating the same model structure, once again represented the perfect accuracy. This finding recommends that the models' architectural design, in this particular domain, played a important role in their potential strength against the adversarial attacks.

- In the end, Scenario 5, where the target and the surrogate models demonstrated with different architectural designs, discovered as another accomplishment. In Spite of the different model structures, the models displays a constant ability to resist the adversarial attacks with unveiling the accuracy around 100

In summary, the model evaluation section provides as a evaluations to the efficacy of the proposed methodology. The models, irrespective of training data types and varying architectural designs, demonstrates the remarkable ability to maintain the accuracy while facing the adversarial challenges. These results not only confirms the potential strength of the machine learning powered based malware detection systems but also represents the efficacy success of the designed adversarial attack approaches in evaluating and strengthen the security of these systems.

6 Discussion and Conclusion

The discussion section caves into the delicate indications of the model evaluation results and their layout with the extensive research objectives. The primary objective of this research was to develop the optimized adversarial attacks using machine learning powered malware detection system and discovers the realistic attack conditions that imitate the disciplined query budgets in the real-world cybersecurity sections.

- **Scenario Analysis and Resilience:** The careful evaluation occurred across the five different scenarios which are demonstrated a coherent nature of persistence. In Scenario 1, where both the target and the surrogate models shared their identical training data, the models outcomes with results of strong potential based defense, outlining their adaptability and effectiveness even under the various optimal conditions. This comes with the research objective of understanding their performance under the shared training data scenarios.

- **Partial and Non-Shared Training Data:** Moving to the Scenarios 2 and 3, the models displayed an impactful ability to resist the adversarial attacks in despite of variations through the shared training data. In Scenario 3, where their no training data was shared, the transferability of the adversarial attacks was still apparent, highlighting the models' capability to conclude their defenses to the new and unseen data. This line up with the research objective of assessing their models' strength in these scenarios with limited or no shared training data.

- **Effect of Model Architectures:** Scenarios 4 and 5, focusing on the model architectures, offered a important valuable insights into the influence of the architectural design on adversarial elasticity. In Scenario 4, with the identical architectures, the models represented a high level of robustness. However, in Scenario 5, where architectures are differed with, the models are continued to excel, demonstrating a incredible ability to adapt and fight off the adversarial attacks even in the face of organizational variations. This resonates with the study objective of discovering the impact of model architectures on adversarial attack transferability.

- **Query-Efficiency and Stealth:** The discussion also increases to the optimized nature of the conceive adversarial attacks. Throughout all the scenarios, the models demonstrated not only effectiveness strength in avoidance but also an efficiency that line up with the real-world limitations of limited query capabilities. The covert of the attacks, aiming to bypass the detection by security process, adds the another layer of complexity where the models successfully steer.

Research Questions:

- Effectiveness of Query-Efficient Attacks: The results declares the potential strength of the optimizing the adversarial attacks, where the constantly achieving the successful avoidance while operating within the limitations of optimizing budgets.

- Impact of Training Data Sharing: The models demonstrates the robustness across all the scenarios with varying degrees of the shared training data, highlights their adaptability and transferability of prevention defenses.

- Influence of Model Architectures: The experiments outlines that, nevertheless of the architectural differences, models maintained the high level of flexibility, representing the importance of their internal structures in preventing the adversarial attacks.

In conclusion, the research not only obtained its primary objectives of the developing and evaluating optimizing with adversarial attacks but also offered the important insights into the characteristics which are influencing the strength of machine learning-based malware detectors in different scenarios. These findings represents significantly to the extensive discussion on cybersecurity and the adversarial machine learning, offering the foundation for future research updertake and the development of more resilient defense approach methods.

7 Future Work

The fortunate exploration of optimizing the Adversarial Attacks on machine learning based malware detection system demonstrates with various direction routes for future research and developments in the field of cybersecurity. The following areas which presents the potential directions for further consideration explorations:

- **Adaptive Defense Mechanisms:** Future research can focus to the enlargement of commutable types of defense approaches that powerfully adjust to the evolving adversarial attacks. This includes the integration of continuous learning and real-time adaptation to imporve the flexibility of machine learning models against to emerging threats. By constantly monitoring and modifying the defense approaches, the models can effectively counter the novel optimized adversarial techniques.

- **Cross-Domain Generalization of Defenses:** Exploring the concept of the adversarial attack defenses across various domains within the cybersecurity, such as the network intrusion detection and malware identification, is a important area for the future work. Understanding whether the defense approaches are conceive for the one specific context can be successfully appealed in others contributes to the development of more universally strong and versatile defense strategies.

- **Real-Time Adversarial Detection:** Future research should cave into the demonstration of the real-time adversarial attacks detection methodologies. These approaches would allows the rapidly identification of adversarial attacks as they come, allowing for the immediate response and then adaptation of defense terminologies. This careful approach is important in the relieving the impact of optimizing the query based adversarial attacks in powerful cybersecurity environments.

- **Ethical and Responsible AI in Adversarial Settings:** Exploring the ethical implications of various adversarial attacks and preventions are important direction paths for the future research. This involves to the developing of adversarial prevention's that not only prioritize the efficacy but also address to the ethical considerations. Assuring the responsible and ethical use of AI technologies in cybersecurity is potentate, and research in this direction path can offers guidelines for developing the powerful, yet ethically sound, defense approaches.

These future research directions aim to push the boundaries of knowledge in adversarial machine learning and contribute to the ongoing efforts to bolster the cybersecurity landscape against evolving threats.

References

Alqahtani, E. J., Zagrouba, R. and Almuhaideb, A. (2019). A survey on android malware detection techniques using machine learning algorithms, *2019 Sixth International Conference on Software Defined Systems (SDS)*, IEEE, pp. 110–117.

Azmoodeh, A., Dehghantanha, A., Conti, M. and Choo, K.-K. R. (2018). Detecting crypto-ransomware in iot networks based on energy consumption footprint, *Journal of Ambient Intelligence and Humanized Computing* **9**: 1141–1152.

Biggio, B. and Roli, F. (2018). Wild patterns: Ten years after the rise of adversarial machine learning, *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pp. 2154–2156.

Cheng, S., Dong, Y., Pang, T., Su, H. and Zhu, J. (2019). Improving black-box adversarial attacks with a transfer-based prior, *Advances in neural information processing systems* **32**.

Goodfellow, I. J., Shlens, J. and Szegedy, C. (2014). Explaining and harnessing adversarial examples, *arXiv preprint arXiv:1412.6572* .

Jia, R. and Liang, P. (2017). Adversarial examples for evaluating reading comprehension systems, *arXiv preprint arXiv:1707.07328* .

Kolosnjaji, B., Zarras, A., Webster, G. and Eckert, C. (2016). Deep learning for classification of malware system call sequences, *AI 2016: Advances in Artificial Intelligence: 29th Australasian Joint Conference, Hobart, TAS, Australia, December 5-8, 2016, Proceedings 29*, Springer, pp. 137–149.

Liu, J., Nogueira, M., Fernandes, J. and Kantarci, B. (2021). Adversarial machine learning: A multilayer review of the state-of-the-art and challenges for wireless and mobile systems, *IEEE Communications Surveys & Tutorials* **24**(1): 123–159.

Liu, K., Xu, S., Xu, G., Zhang, M., Sun, D. and Liu, H. (2020). A review of android malware detection approaches based on machine learning, *IEEE Access* **8**: 124579–124607.

Machado, G. R., Silva, E. and Goldschmidt, R. R. (2021). Adversarial machine learning in image classification: A survey toward the defender's perspective, *ACM Computing Surveys (CSUR)* **55**(1): 1–38.

Qi, X., Tang, Y., Wang, H., Liu, T. and Jing, J. (2022). Adversarial example attacks against intelligent malware detection: A survey, *2022 4th International Conference on Applied Machine Learning (ICAML)*, IEEE, pp. 1–7.

Rashid, A. and Such, J. (2023). Malprotect: Stateful defense against adversarial query attacks in ml-based malware detection, *arXiv preprint arXiv:2302.10739* .

Ren, Y., Zhou, Q., Wang, Z., Wu, T., Wu, G. and Choo, K.-K. R. (2020). Query-efficient label-only attacks against black-box machine learning models, *Computers & security* **90**: 101698.

Shaukat, K., Luo, S. and Varadharajan, V. (2022). A novel method for improving the robustness of deep learning-based malware detectors against adversarial attacks, *Engineering Applications of Artificial Intelligence* **116**: 105461.

Singh, J. and Singh, J. (2021). A survey on machine learning-based malware detection in executable files, *Journal of Systems Architecture* **112**: 101861.

Tayyab, U.-e.-H., Khan, F. B., Durad, M. H., Khan, A. and Lee, Y. S. (2022). A survey of the recent trends in deep learning based malware detection, *Journal of Cybersecurity and Privacy* **2**(4): 800–829.

Wang, J., Wang, C., Lin, Q., Luo, C., Wu, C. and Li, J. (2022). Adversarial attacks and defenses in deep learning for image recognition: A survey, *Neurocomputing* .

Xu, G., Shao, H., Cui, J., Bai, H., Li, J., Bai, G., Liu, S., Meng, W. and Zheng, X. (2023). Gendroid: A query-efficient black-box android adversarial attack framework, *Computers & Security* p. 103359.

Yadollahi, M. M., Lashkari, A. H. and Ghorbani, A. A. (2021). Towards query-efficient black-box adversarial attack on text classification models, *2021 18th International Conference on Privacy, Security and Trust (PST)*, IEEE, pp. 1–7.

Yan, H., Li, X., Zhang, W., Wang, R., Li, H., Zhao, X., Li, F. and Lin, X. (2023). Automatic evasion of machine learning-based network intrusion detection systems, *IEEE Transactions on Dependable and Secure Computing* .

Zhan, D., Tan, K., Ye, L., Yu, X., Zhang, H. and He, Z. (2023). An adversarial robust behavior sequence anomaly detection approach based on critical behavior unit learning, *IEEE Transactions on Computers* .