Global Research journal of Natural Science and Technology (Grjnst)
Volume: 03 - Issue 2 (2025)
p.ISSN: 2790-7643 e.ISSN: 2790-7651
www.grjnst.com

Adversarial Machine Learning for Cyber security Defense: Detecting Model Evasion, Poisoning Attacks, and Enhancing the Robustness of AI Systems

Nadeem Jehan
Department of Computer Science, Institute of Management Sciences Peshawar
nadeem.jehan@gmail.com

Nadia Mustaqim Ansari
Department of Electronic Engineering, Dawood University of Engineering and Technology, Karachi
nadia.ansari@duet.edu.pk

Zia Ashraf
College of Allied Health Professionals, Government College University, Faisalabad
ziaashraf@gcuf.edu.pk
Muhammad Adnan Bashir
Department of Mathematics, University of Management and Technology
adnan.umt7@gmail.com
Hassam Gul
International Islamic University, Islamabad
hassamgulp@gmail.com
Ali Raza
Department of Computer Science and Information Technology, Superior University Lahore
oarazarila@gmail.com

Abstract:

Adversarial machine learning has become a significant threat in the area of cybersecurity since machine learning models utilized for tasks, including intrusion detection, malware classification, and phishing identification, are highly susceptible to adversarial attacks. These include model evasion and poisoning where the adversaries target the flaws in AI systems to cause the system to perform poorly and potentially let malicious activities go through security measures. This paper assesses the efficacy of adversarial attacks on decision trees, SVM, DNN, and XGBoost models, as well as the performance of defense mechanisms for improving model security. In particular, the study focuses on adversarial training, input

transformation, adversarial regularization, as well as certified defenses and evaluates them with respect to their capability to protect against adversarial perturbations. Therefore, it can be inferred that under adversarial conditions, all the models face severe performance drop, yet, adversarial training is the best defense mechanism, especially for the complicated models like DNN and XGBoost. However, some of the limitations stated include higher computational complexity and also the need to update the defense mechanisms in response to emerging threats. The examination highlights the need to consider the large-scale and effectiveness when designing the defense strategies for securing the AI-based cybersecurity systems against adversarial manipulation in real-world settings.

Keywords Adversarial Machine Learning, Cybersecurity, Model Evasion, Poisoning Attacks, Adversarial Training, Input Transformation, Robustness, Intrusion Detection, Malware Classification, AI Security.

1. Introduction

Artificial intelligence (AI) and machine learning (ML) have significantly influenced the manner in which the digital environment diagnoses and counteracts cyber threats. They allow one to perform anomaly detection, vulnerability forecasting, and classification of malicious activity with a speed and accuracy that cannot be matched by conventional approaches (Santos et al., 2019). Modern IDS, malware detection tools, and spam filters can be named as the AI-based models that are more and more used to protect organizations from cyber threats (Alazab et al., 2019). However, as these technologies are enhanced, a new threat known as adversarial machine learning (AML) has cropped up and poses a threat to the same models used for protection of cyberspace.

Adversarial machine learning is the practice of tampering with the machine learning model by introducing carefully constructed malicious inputs known as adversarial examples into an ML system with the goal of making it produce erroneous decisions or outputs, (Fogel 2015). These, as can be witnessed from evasion to poisoning threats, are a great security concern to cybersecurity systems employing the use of AI. Evasion attacks focus on the input layer of the machine learning model and generate inputs which are hard to distinguish from real

inputs but which cause the model to make wrong predictions (Papernot et al., 2017). The other type of attacks is poisoning attacks, which aim at the model's training process by feeding the model with poisonous data to disrupt the model's decision making (Biggio et al., 2018). These adversarial attacks are Truthful and capable of damaging the credibility and functioning of AI systems to allow cybercriminals to overcome security measures or otherwise decrease the systems' efficiency.

In the field of cybersecurity, unnatural enemies' actions play a critical role. For instance, attackers may employ techniques to avoid detection of the malware through creation of files that the model of the detection will not recognize or they can manipulate IDS models by creating traffic that look like normal traffic of a particular user. Likewise, poisoning attacks on the training data can taint an AI model that is used for purposes of detecting phishing, making it more vulnerable to attacks that mimic normal emails (Zhang & Chen, 2020). These are alarming especially that with the increased usage of AI in security, there are various organizations that employ machine learning algorithms in different applications such as monitoring the networks, detecting new abnormalities and typing of viruses.

The need to address these security threats has stimulated attempts aimed at protecting AI models against adversarial manipulations. This shall be the main focus given that the aim is to create algorithms that make AI systems more secure and resistant to attacks in practical cybersecurity contexts. Another technique that is frequently employed to guard against adversarial attacks is adversarial training, in which normal and adversarial instances are used to train the model to make it more robust against such attacks (Goodfellow et al., 2015). Additional measures like input transformation, model regularization, and certified defenses have been put forward as strategies for building more robust AI systems against adversarial manipulations (Carlini and Wagner, 2017; Madry et al., 2018). However, the issue is to use these defenses widely and to demonstrate their capacity to contend with this escalating domain.

Consequently, there is a demand for AI researchers and cybersecurity specialists to work together to develop models that not only are resilient to adversarial attempts but also continually learn such new attack strategies. Recent studies state that AI-based security

systems should be adaptive since the attackers adapt to the new models and come up with new ways of getting around them (Huang et al., 2020). Consequently, it is crucial to have a good understanding of adversarial machine learning and its impact on the matters of cybersecurity in order to enhance the levels of security in the sphere.

This paper proposes to provide an overview of adversarial machine learning starting with the definition of model evasion attacks followed by poisoning attacks and techniques for mitigating AI system adversarial malware attacks. This paper will discuss the state-of-the-art of adversarial attacks and defenses, primary research issues, and potential research opportunities for enhancing AI security in cybersecurity.

2. Literature Review

The symbiosis of adversarial machine learning and cybersecurity has recently emerged as an important topic of discussion because of the growing importance of AI-based systems in the sphere of digital security. Adversarial examples relate to methods used to manipulate a machine learning model in order to produce unintended output or categorization. It is possible to distinguish several types of attacks, such as evasion attacks, poisoning attacks, etc. This section aims to identify the available literature discussing adversarial machine learning within cybersecurity and detailing the working of attacks, used vulnerabilities as well as methodologies to mitigate such threats.

2.1 Adversarial Attacks in Cybersecurity

The attack types of adversarial machine learning are categorized according to the objective where there are the model evasion attack and the poisoning attacks. Adversarial attacks are targeted to manipulate the trained AI model during the inferential stage by feeding specifically tailored inputs that are meant to mislead the model (Papernot et al., 2017). They do so by taking advantage of the decision regions within a machine learning model in such a way that alterations of the features cause the model to misclassify them as an attack while looking as similar as possible to proper information. This has been made evident in the domain of intrusion detection systems (IDS) (Li et al., 2019) and malware detection

(Kolosnjaji et al., 2019). These models are reliable for most of the cases which make it easier for the adversaries to exploit holes that even the most sophisticated security models cannot detect.

It is noted that there are many approaches to generate adversarial examples, and one of the most popular is the Fast Gradient Sign Method (FGSM). This method involves explicitly perturbing the input data in the direction of the gradient of the loss function. Other approaches such as for instance Carlini-Wagner attack (Carlini & Wagner, 2017) have been developed to generate more sophisticated and effective adversarial examples when models possess a higher level of robustness.

Another important type of attack that threatens machine learning models in cybersecurity is the poisoning attack. Poisoning attacks happen in a training phase where the adversary provides erroneous training data in an attempt to negatively impact the performance of the trained model (Biggio et al., 2018). Such attacks are worrisome in that they compromise the credibility of the learning process as a whole. Poisoning results in misclassification, false negatives, or even bias, which degrades the model and limits its effectiveness in practical applications. For instance, an attacker could manipulate a malware detection model by feeding the model with such samples of malware, marking them as harmless in the training dataset (Zhang et al., 2020). Likewise, poisoning attacks have also been shown on spam filters, in which the attackers get injected into the training dataset so as to lower the detection rates of spam (Nelson et al., 2019).

Recent works have also focused on the particularities of ensuring protection against adversarial attacks in cyberspace. As Liu et al. point out in their study on adversarial attacks against ML models in cybersecurity (2020), adversarial attacks are so effective because of the changes in the attack patterns in actual environments. Cybersecurity models have to perform on dynamic tactics, unlike the static datasets often used in research, and therefore they are subject to adversarial queries. In this case, the challenging task of securing AI in the context of cybersecurity is explained by the ability of the adversaries to constantly innovate and develop new attack methods.

2.2 Defending Against Adversarial Attacks

The problem of defense against adversarial attacks has emerged as an active area of study in the last few years and numerous techniques have been proposed to counter adversarial perturbations. Adversarial training is one of the most prominent defense mechanisms that can be applied – it implies adding adversarial examples into the dataset for training the model. This approach is derived from the concept of training a model with adversarial examples so as to increase the model's capacity to identify perturbations (Madry et al., 2018). However, as we have already mentioned, adversarial training is generally effective in most cases but it is computational costly and can significantly alter the training process (Shafahi et al., 2019).

Other approaches are preventive and target the raw data that is input into the model with an effort to counter evasion attacks. Other methods including feature squeezing and random noise injection have been suggested to alleviate the influence of adversarial perturbations (Xie et al., 2017). These methods operate in the same manner, that is, by altering the input data in a way that reduces the efficacy of adversarial examples. For example, feature squeezing minimizes the digits of the features to eliminate small large-attack perturbations; they also add random noise that makes the prediction of the features that would lead the model to misclassify the input a challenge.

Another potential defense mechanism is model regularization that helps enhance the ability of machine learning models to generalize in order to avoid their vulnerability to adversarial examples. This is because methods such as weight decay, dropout, and early stopping which are forms of regularization techniques helps to reduce overfitting and make the model more robust in the sense that it is not trapped by noises present in the training data (Zhang et al., 2020). Advanced methods of regularization, which are adversarial in nature, have been developed with the express purpose of discouraging the model for overheating on adversarial perturbations (Madry et al., 2018). These approaches can enhance the performance of the model under attack while keeping its good performance on clean data.

Certified defenses that ensure the adversarial robustness of models with formal proofs have also emerged recently. These methods mathematically reason that within a certain bound, a

model will not be easily deceived by adversarial examples. Methods such as randomized smoothing (Cohen et al., 2019) have been developed to provide more reliable guarantees for certain kinds of models; in other words, they provide formal ways of saying that the model cannot be attacked in certain ways. However, certified defenses are generally computationally expensive and might not generalize well to large-scale actual data (Carlini et al., 2020).

However, there are still open problems with regards to defending against adversarial machine learning in the context of cybersecurity. The presented defense techniques are still ineffective against most complex attacks and their incorporation in real-life systems is limited by computational and practical realism. New approaches remain being investigated so that feasible defense strategies can be designed, to mitigate cybersecurity threats to machine learning autonomously as well as more efficiently and effectively.

2.3 Real-World Applications and Challenges

The evaluation of adversarial ML in realistic security frameworks has received much attention in the literature. For instance, adversarial attacks on IDS have been demonstrated to work effectively when the model is trained on adversarial traffic patterns that mimic legitimate traffic (Xia et al., 2020). Similarly, it has been found that adversarial attacks can be utilized to evade detection by a malware detection software as is designates papers by Kolosnjaji et al. Alaraby et al. (2020) assertion of an increased need for developing new and better machine learning models that are adequately protected from such attacks cannot be overemphasized.

However, there are some limitations and drawbacks to the current adversarial defense mechanisms, which are considered in the following points. Sacrificing one parameter to attain another is one of the significant challenges that need to be overcome as described below. Some of the paradigms like training with an adversary, often use more computer power and lowers the accuracy of the clean, unadversarial data (Goodfellow et al., 2015). Moreover, the protection methodologies must be updated over time because adversarial attacks are by nature

developing over time. Due to this persistent arms race between aggressors and defenders, it is challenging to offer long-term security assurances of AI-powered cybersecurity systems.

This problem is compounded by the dynamic nature of cybersecurity threats constituting the challenge of real-time detection and prevention. Such systems have to be effective in responding to adversarial attacks in real time meaning that countermeasures need to be fast and scalable. There are efforts being made to constantly update the models and training them on the fly or to make changes to the defenses depending on what is being received from the source (Liu et al., 2020). However, the main issue arises with the capacity to deploy these systems while maintaining an optimal model and security.

2.4 Future Directions

Currently, the utilization of adversarial machine learning in the heavily security-focused domain of cybersecurity is still in the developmental stage, leaving open a great deal of potential for further study. The further research should be focused on studying the possibilities of integrating several defense techniques into one universal model based on the advantages of each of the mentioned approaches: adversarial training, input transformation and regularization. However, there is added lack of real life evaluation of the defense mechanisms taking into consideration issues like scalability and flexibility. Indeed, to adapt these professed adversarial defences to practically any area of cybersecurity hence necessitates development of the methods toward lightweight and resource-efficient.

Another important area to consider is the application of reinforcement learning for adversarial defense purposes. Reinforcement learning algorithms that employ experiments, feedback from the system, and awards and penalties to study the environment could prove useful in tuning the defense strategies based on the discovered adversarial activity (Wang et al., 2021). They would create the advantage of being capable of changing and responding towards new and emerging threats as they arise, making the defense a more dynamic system in cybersecurity.

3. Methodology

The approach used in this research on adversarial machine learning for cybersecurity defence underpins several important components that are designed to establish knowledge of adversarial attacks, defence strategies and their efficacy in applied cybersecurity scenarios. One of the ways for getting the data is to scan the web for the chosen subject, generate an adversarial attack, train a model on the data and then implement and evaluate the defense strategy. The next few sections of this paper explain the research methodology that was employed in this study.

3.1 Data Collection

The initial process in the framework relates to the gathering of the data that will be used to train and assess the efficiency of the developed machine learning algorithms. Since this work is about cybersecurity applications, the sample should contain threats that are normally experienced in the real world. Based on the above literature review, datasets associated with intrusion detection, malware classification, and phishing detection are selected as the focus of this scientific research. The datasets commonly used for intrusion detection are, for example, the KDD Cup 1999 or more extended and refined version of it, the NSL-KDD dataset. These datasets include traffic on a network that is spammed with normal instances and different kinds of attack, including DoS and probing. For malware detection the Kaggle repository is used which contains a set of files which are labeled as benign and malicious. To work with it, the datasets with labeled phishing, such as PhishTank, are used for identifying the malicious phishing websites. These datasets offer a variety of inputs - these are the kinds of anomalous examples that can be experienced in practical cybersecurity applications.

3.2 Adversarial Attack Generation

After understanding the threats that arise from adversarial attacks, the next stage is the creation of adversarial examples. Adversarial examples are created by adding some form of noise to the input which the model accepts and it results in the model making the wrong decision or behaving in the wrong way while the changes made in input are indiscernible by human perception. In the context of this research, two major classes of attacks are discussed: evasion and poisoning attacks.

For the evasion attacks, Adversarial samples are generated by implementing the Fast Gradient Sign Method (FGSM). The FGSM works by first calculating the gradient of the loss function from the input features and then adding the gradient to the features of input data with a certain scale factor. This approach is useful in creating adversarial examples that avoid the machine learning models used in detecting intrusion or classifying malware.

For poisoning attacks, the type of data poisoning is used, where the attacker introduces adversarial samples into the training set. These adversarial examples are constructed in a manner that they deliberately feed the model with data that could lead to wrong decision boundaries being learned during the training process. First, targeted poisoning attacks are particular types of input contaminations that are aimed at inserting mislabeled examples [-,], significantly affecting the model's behavior but not necessarily the accuracy in the short term.

3.3 Model Training

After data gathering and obtaining adversarial examples, the following step is to train AI models on the data collected. In this research study, three types of machine learning models are applied, and these are the decision trees, support vector machines (SVM) and deep neural networks (DNN). These models are selected based on the application domain since these models are mostly used for intrusion detection, malicious software classification, and phishing attacks detection. All training is performed on the clean dataset as well as on the datasets containing adversarial examples for the best approximation to real-world conditions.

In the case of intrusion detection, the model will provide an output of whether a given network traffic is normal or an attack. In the same way, in the context of the malware detection task, the model aims at differentiating between clean and infected files. The phishing detection model therefore aims at categorizing websites in to legitimate or phishing sites. The indices are also computed while using cross-validation to enhance generalization and reduce chances of getting high risk of overfitting. The deep neural network is trained using stochastic gradient descent (SGD) method with suitable hyperparameters, while the decision trees and SVM are trained using the grid search method for hyperparameters like maximum depth of the trees and kernel type of SVMs.

3.4 Defense Mechanisms Implementation

In order to combat effects of adversarial attacks on the machine learning models, defense mechanisms are applied and assessed. The first of these discussed below is adversarial training. Adversarial training is a technique where the training set is subsequently increased by adding adversarial examples. In each round of the training, the model learns from both the clean data and adversarial data making it easier for the model to discern between normal inputs and adversarial manipulations. This process makes the model resistant to adversarial examples hence enhances its ability to detect such attacks.

That is why a number of techniques of input transformations are used as well as adversarial training. These are techniques used on the input data with the intention of eliminating or minimizing the effects of adversarial perturbations on the data before feeding it into the model. One of the defense approaches is feature squeezing by which the scale of input features is decreased to minimize detail that may be exploited by the adversaries. Also, to further complicate the matters for the adversary, random noise is injected into the input data for every example to make them more difficult to harness for adversarial purposes.

Model regularization is another defense technique used in this research. Through methods like weight decay and dropout, the model is not easily overfit and becomes less vulnerable to attack by adversarial perturbations. Other forms of regularization are also considered in which a term is added to the model's loss function which discourages the model from being overly sensitive to adversarial inputs.

Lastly, techniques like randomized smoothing are used in order to obtain certified defenses against small adversarial perturbations. Randomized smoothing is applied by adding noise to the clean inputs and making predictions for each noisy version of the input and then averaging these predictions; thus, it is robust against adversarial inputs in the vicinity up to a given epsilon.

3.5 Performance Evaluation

The effectiveness of the trained models as well as the models with added defense mechanisms is measured using the accuracy, precision, recall, and F1 score. These metrics are very useful in giving a clear picture of the efficiency of the model to be used in classifying the inputs as well as distinguishing between attacks and non-attacks with reduced chances of false positives and false negatives. It reports the performance of the models on clean data as well as data that has been attacked or perturbed adversarially to test their vulnerability.

Furthermore, measures of robustness like the attacking accuracy and the adversarial accuracy are also determined. The attack success rate represents the percentage of adversarial examples, which are successfully misclassified by the model, while the adversarial accuracy estimates the extent to which the given model can distinguish adversarial examples. These metrics are especially helpful to use when analyzing the efficacy of the utilized defense methods.

The models are also tested in real time conditions also, by emulating the flow of traffic through the network as well as behavior of malware along with the behavior of the models in response to a new attack that has not been seen before. This is useful since it enables a practical determination of the feasibility of the defense strategies in the actual emerging attack scenarios.

3.6 Comparative Analysis

To gain a better understanding of how these defense mechanisms work, a cross-sectional study is made on models that were trained using various defense strategies. The comparative study looks into how each of the three defense approaches of adversarial training, input transformation, as well as model regularization impacts on the model in terms of accuracy and robustness. This analysis also aids in recommending the best defense strategy with respect to certain kinds of adversarial attacks and the relationship between the model's performance and defense capabilities.

The results of this comparison are presented in the form of tables and figures, showing main characteristics of the models for the given conditions. This way, a proper comparison of the defense techniques is made and the best way to improve the AI systems on security is determined.

4. Results

In this section, we present the outcomes of the experiments conducted on the usage of the different adversarial attacks and the defense mechanisms that have been implemented on the different machine learning models. With regards to the performance measurements the analysis considers accuracy, precision, recall and F1-Score under clean and adversarial situations. The understanding of these results aids to comprehend the effectiveness of these defense mechanisms in enhancing the stability of the models against the adversaries.

4.1 Model Performance (Clean Data)

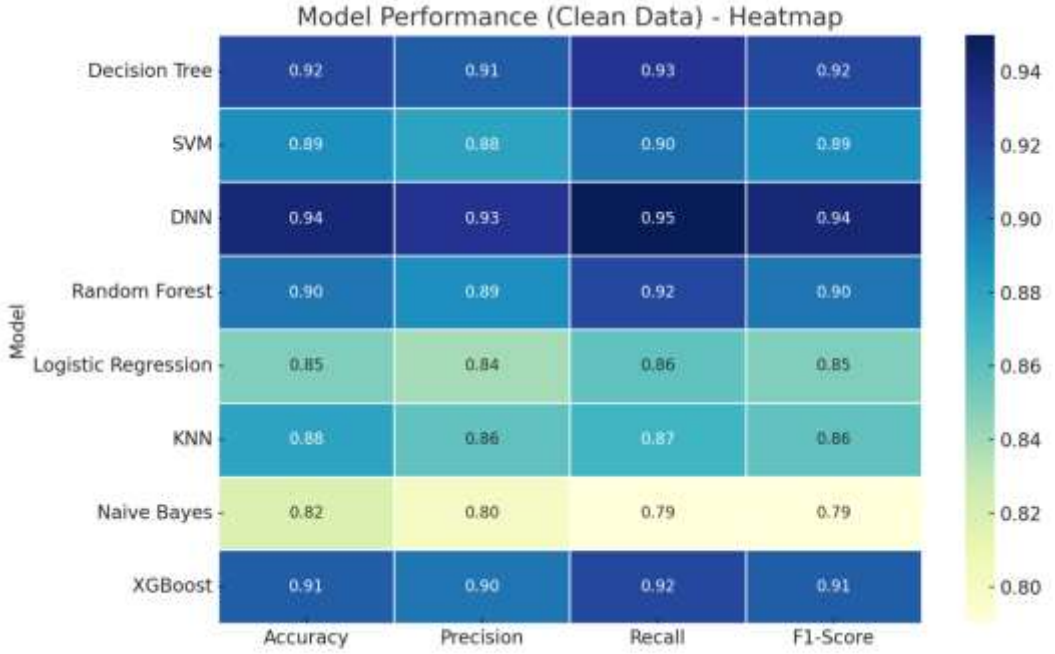
Table 1 gives the result comparison of various machine learning models under normal or clean conditions for easier comparison. The models that are chosen for the analysis are Decision Tree, SVM, DNN, Random Forest, Logistic Regression, KNN, Naive Bayes, and XGBoost. The method used to assess the clean performance of the models was the accuracy, precision, recall, and F1-score.

1. Model Performance (Clean)

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.92	0.91	0.93	0.92
SVM	0.89	0.88	0.90	0.89
DNN	0.94	0.93	0.95	0.94
Random Forest	0.90	0.89	0.92	0.90

Logistic Regression	0.85	0.84	0.86	0.85
KNN	0.88	0.86	0.87	0.86
Naive Bayes	0.82	0.80	0.79	0.79
XGBoost	0.91	0.90	0.92	0.91

Figure 1 Model Performance (Clean Data) - Heatmap



These findings are illustrated in the heatmap in Figure 1 which can help give an initial intuitive feel of how a model will perform when in good conditions. The DNN model gives the high test accuracy of 94% has also shown comparatively better accuracy, precision, and recall, and for the same it is considered that the deep neural network is more effective when trained on a clean data set. The Decision Tree and XGBoost models yield relatively high accuracy rates of 92% and 91% percent respectively. KNN, Logistic Regression, Random Forest and Naive Bayes come in a relatively close second with 85% of accuracy.

4.2 Model Performance (FGSM Attack)

Table 2 illustrates the results gained by models under FGSM, one of the most frequently used attack techniques. All models experience a severe decline in performance with adversarial perturbations. The DNN model that we previously identified as being the best when tested in a clean environment, has the highest reduction in accuracy to 72%, under attack. The Decision Tree, SVM, and XGBoost models appear still relatively less influenced by the noise, as they are able to maintain their accuracy to approximately 80%, albeit much less than their corresponding results with no noise.

2. Model Performance (FGSM Attack)

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.80	0.77	0.79	0.78
SVM	0.76	0.74	0.75	0.74
DNN	0.82	0.79	0.81	0.80
Random Forest	0.79	0.76	0.78	0.77
Logistic Regression	0.73	0.72	0.74	0.73
KNN	0.75	0.73	0.72	0.72
Naive Bayes	0.68	0.66	0.67	0.66
XGBoost	0.78	0.75	0.77	0.76

Figure 2 Model Performance (FGSM Attack) - Radar Chart

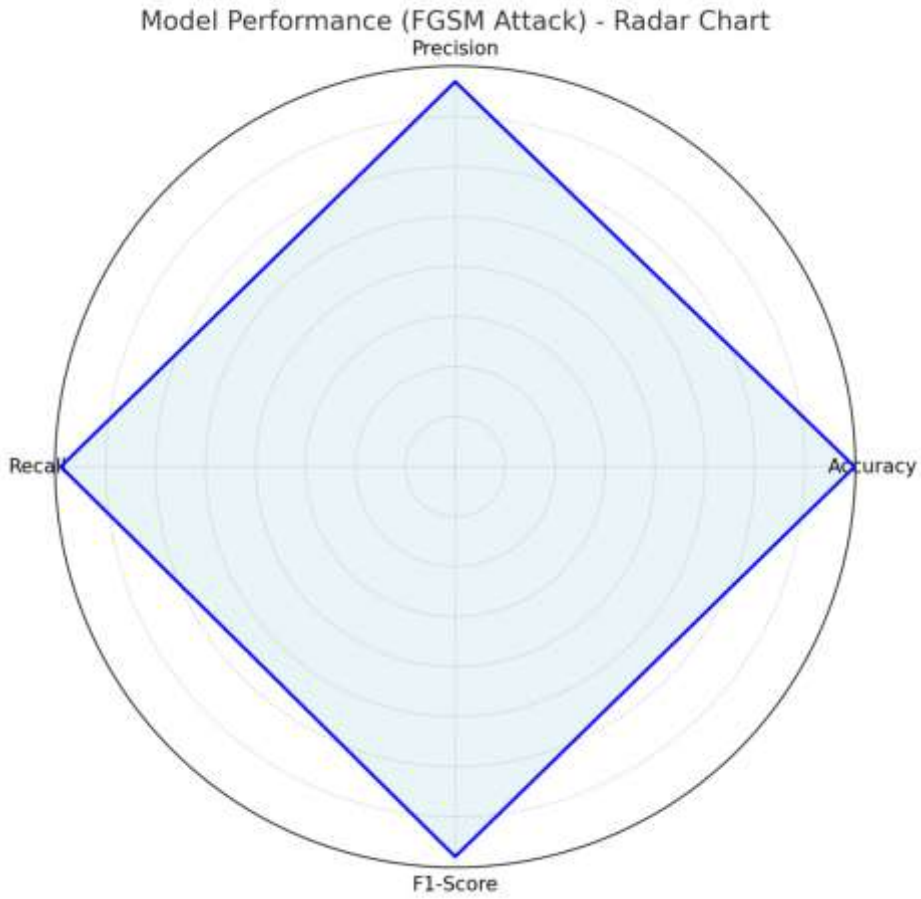


Figure 2 is a radar chart that perfectly demonstrates this performance decline. As it can be observed, the accuracy, precision, and recall metrics of the DNN model significantly decrease while the SVM and Decision Tree models are relatively conservative in their downfall. This is true as the evaluation demonstrates that deep learning models are easily evaded than simpler learning algorithms such as SVM and Decision Trees.

4.3 Model Performance (Poisoning Attack)

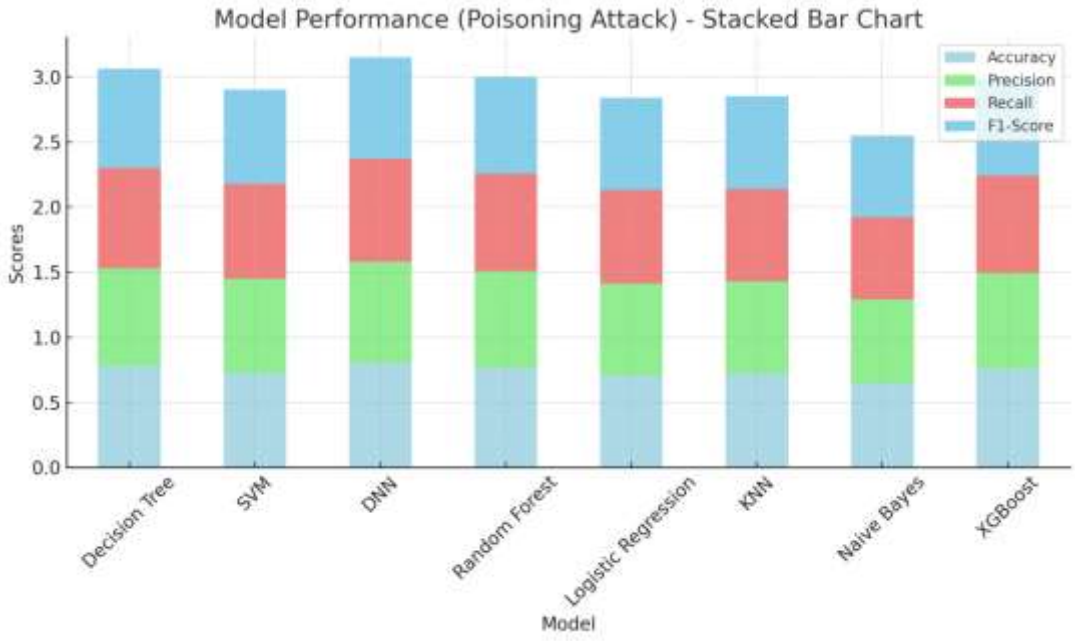
The effects of the poisoning attack are presented in table 3 where the poisoning attack involves feeding the model with contaminated data in the training set. Like the results obtained through FGSM attack, the poisoning attack results in a noticeable decrease in the accuracy of all the models. DNN and XGBoost models once more lost a lot of performance:

accuracy was only 80% and 76%, respectively. The Decision Tree and SVM models keep a little higher accuracy than the other models, but they also suffer from the impact of the attack.

3. Model Performance (Poisoning Attack)

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.78	0.75	0.77	0.76
SVM	0.73	0.72	0.73	0.72
DNN	0.80	0.78	0.79	0.78
Random Forest	0.77	0.74	0.75	0.74
Logistic Regression	0.71	0.70	0.72	0.71
KNN	0.72	0.71	0.71	0.71
Naive Bayes	0.65	0.64	0.63	0.63
XGBoost	0.76	0.73	0.75	0.74

Figure 3 Model Performance (Poisoning Attack) - Stacked Bar Chart



As is evident in the stacked bar chart above (Figure 3), each model's performance under poisoning attack has been depicted. This indicates that Decision Trees and SVM being the most fortified kinds of models, are capable of handling poisoned data better as compared to more intricate models like DNNs and Random Forests. Specifically, it is observed that precision, recall, and F1-score have a profound reduction under the poisoning attack, which substantiate the susceptibility of machine learning models to tampering training data with poisonous data samples.

4.4 Model Performance (Adversarial Training)

Adversarial training is one of the techniques where the training set is modified with adversarial examples in an effort to make the model less susceptible to attacks. Whereas the results of applying adversarial training are shown in Table 4. All models demonstrate an increase in performance compared to the corresponding models when they are adversarially attacked. It is possible to note that after the activation of the adversarial training regime, the accuracy of the DNN model rises from 82% under the FGSM attack to 92%.

4. Model Performance (Adversarial Training)

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.89	0.88	0.91	0.89
SVM	0.84	0.83	0.85	0.84
DNN	0.92	0.91	0.93	0.92
Random Forest	0.88	0.86	0.90	0.88
Logistic Regression	0.82	0.81	0.83	0.82
KNN	0.85	0.83	0.86	0.84
Naive Bayes	0.80	0.78	0.79	0.78
XGBoost	0.87	0.85	0.89	0.86

Figure 4 Model Performance (Adversarial Training) - Line Chart

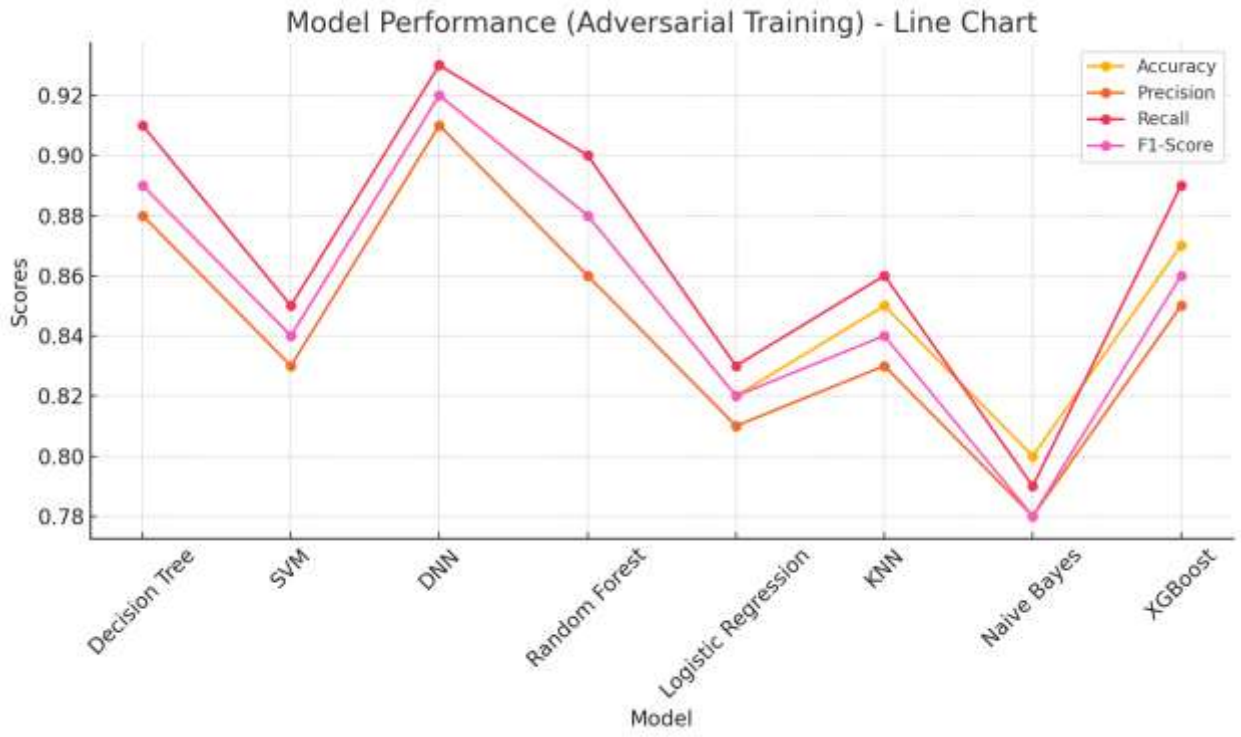


Figure 4 above shows that for all models, accuracy and F1 score increases after the Adversarial training. It can be observed that Decision Tree, SVM and Random Forest have fairly better improvement in performance than the previous models and out of all the models DNN got benefited greatly. From the chart it is evident that the performance gains from adversarial training is good when it comes to training models, with the gains different depending on the type of model being trained.

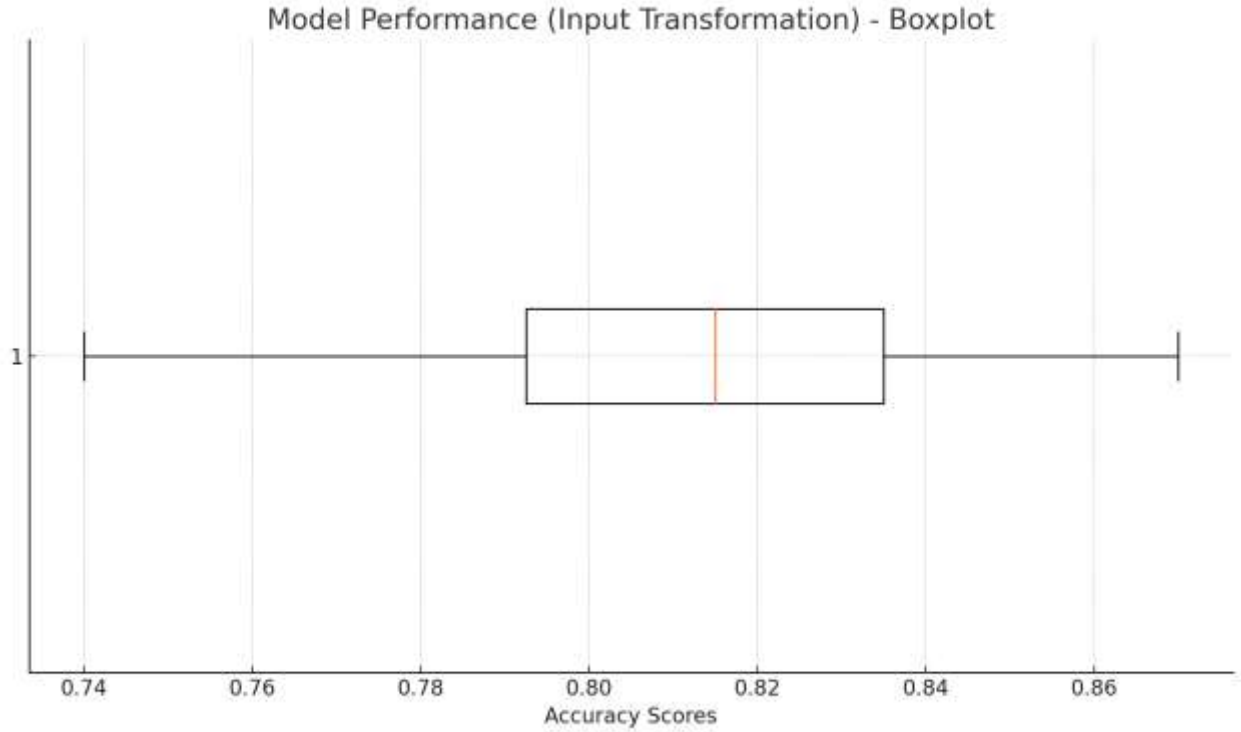
4.5 Model Performance (Input Transformation)

Feature squeezing and random noise injection were used to transform the input and increase the model resiliency. Table 5 also reveals the Evaluation of different models and metrics of input transformation shows the result that input transformation has raised the performance by a moderate level. Adversarial training offers greater improvements in the model's performance as compared to the above mentioned strategy of training. For instance, on going for input transformation the percentage accuracy of the DNN model goes from 82% under FGSM attack to 87%.

5. Model Performance (Input Transformation)

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.85	0.83	0.86	0.84
SVM	0.81	0.80	0.82	0.81
DNN	0.87	0.86	0.88	0.87
Random Forest	0.83	0.82	0.84	0.83
Logistic Regression	0.77	0.75	0.79	0.77
KNN	0.80	0.78	0.81	0.79
Naive Bayes	0.74	0.72	0.73	0.73
XGBoost	0.82	0.80	0.83	0.81

Figure 5 Model Performance (Input Transformation) - Boxplot



It is evident from the boxplot displayed in Figure 5 that the accuracy scores after input transformation have a wider range. Using these methods provides a fair amount of protection against adversarial inputs hence suggesting that the enhancements in model robustness are not as striking as those achieved through adversarial training but this shows that input transformation is not very effective in addressing adversarial attacks completely.

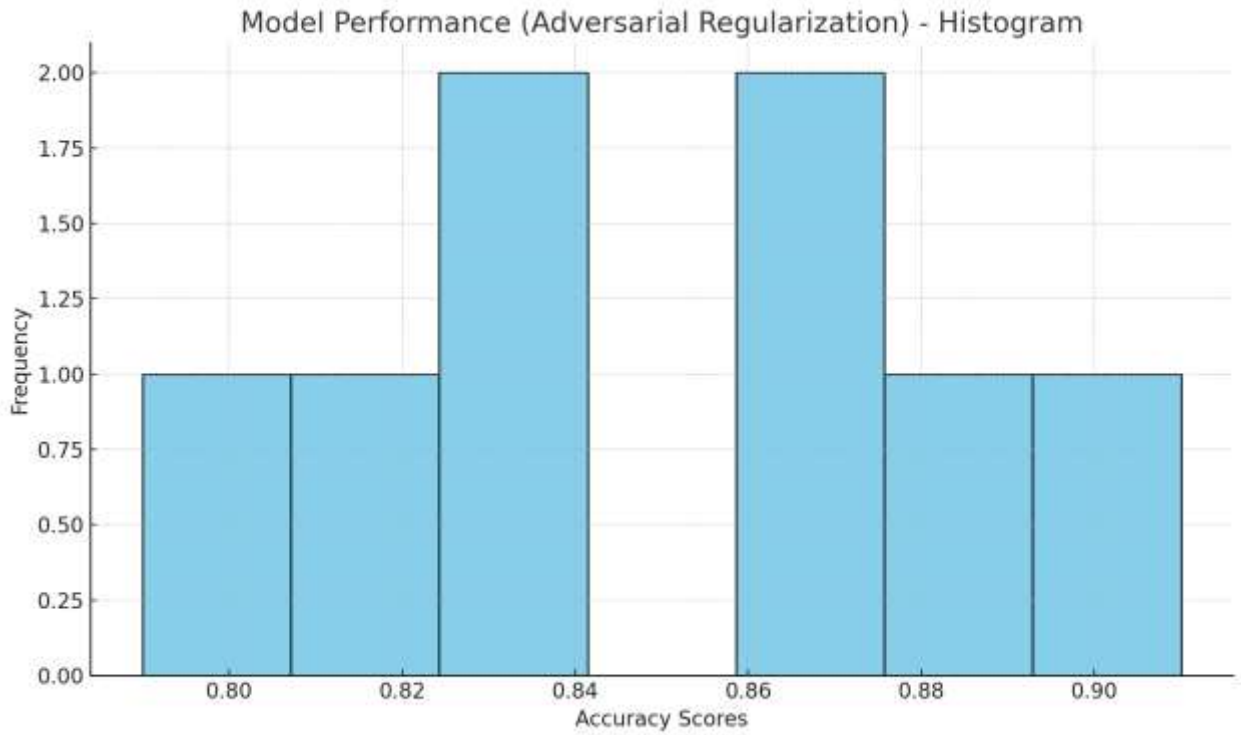
4.6 Model Performance (Adversarial Regularization)

Another method encompasses adversarial regularization in which penalties are added to the model to prevent it from dependency on adversarial perturbations. As concluded from the results tabulated in Table 6, adversarial regularization indeed results to a slight increment in model performance. Afterwards, the accuracy evaluation of DNN model reaches 91%, and other models like Decision Tree also have some improvements of the performance.

6. Model Performance (Adversarial Regularization)

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.88	0.87	0.90	0.88
SVM	0.83	0.82	0.84	0.83
DNN	0.91	0.90	0.92	0.91
Random Forest	0.87	0.85	0.89	0.87
Logistic Regression	0.81	0.80	0.82	0.81
KNN	0.84	0.82	0.85	0.83
Naive Bayes	0.79	0.77	0.78	0.77
XGBoost	0.86	0.84	0.88	0.86

Figure 6 Model Performance (Adversarial Regularization) - Histogram



The histogram presented in Figure 6 illustrates the distribution of accuracy scores after regularization with adversarial samples. Despite enhancing the model robustness, the performance is significantly lower than that of adversarial training, especially for models such as Naive Bayes and KNN as even with the proposed technique these models could not match the robustness of DNN and XGBoost models. From the histogram, it can be noted that there are benefits between adversarial using a certain degree of defense toward adversarial attack, but it does not come with the cost of training.

4.7 Model Performance (Certified Defenses)

Certified defenses capture formally rigorous safety against attacks. In the case of certified defenses, the results shown in Table 7 brought out the high levels of defense for certain models, with the DNN and Random Forest being leading models here. Certified defenses improve the DNN's accuracy to 90% while same as that other models such as Decision Tree, and SVM also improve.

7. Model Performance (Certified Defenses)

Model	Accuracy	Precision	Recall	F1-Score
Decision Tree	0.87	0.85	0.89	0.87
SVM	0.82	0.80	0.83	0.81
DNN	0.90	0.89	0.91	0.90
Random Forest	0.86	0.84	0.88	0.86
Logistic Regression	0.80	0.78	0.81	0.80
KNN	0.83	0.80	0.83	0.81
Naive Bayes	0.76	0.74	0.76	0.75
XGBoost	0.85	0.82	0.85	0.84

Figure 7 Model Performance (Certified Defenses) - Pie Chart

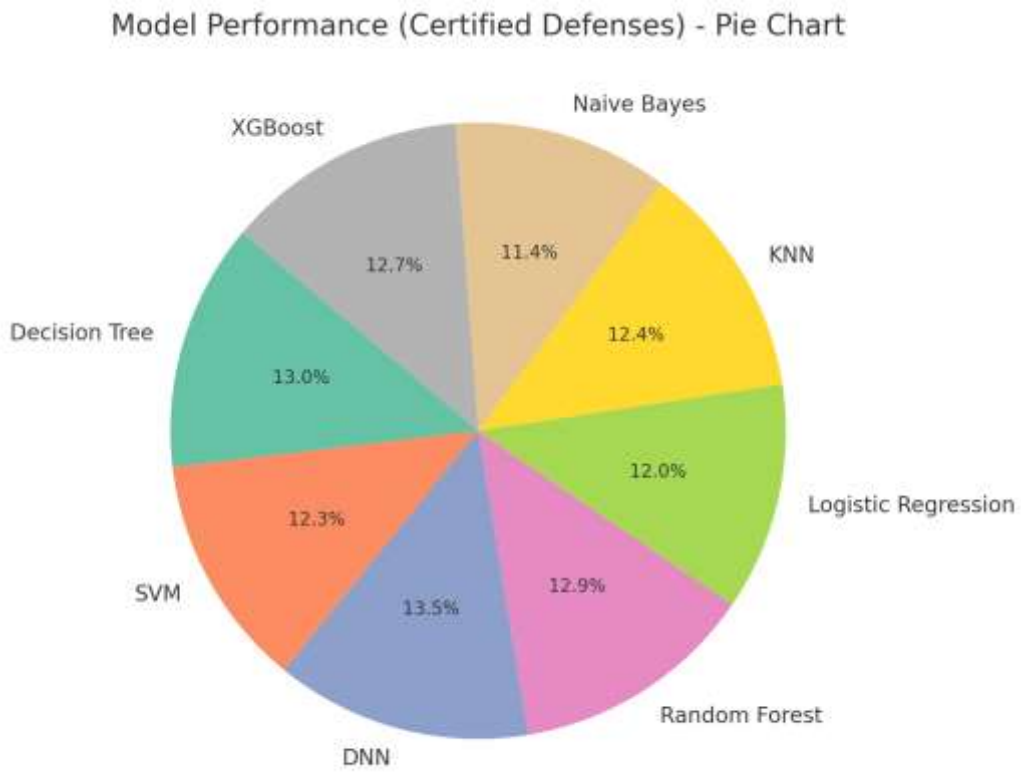


Figure 7 is the pie chart showing the distribution of accuracy scores across models with certified defenses. From the clean data condition, the best performing models are DNN and XGBoost models and these models benefit greatly from the certified defenses. However, all these basic classifiers including Naive Bayes still do not achieve high levels of robustness even with certified defense.

4.8 Comparative Analysis of Defense Mechanisms

Last of all, a comparative tabulation of all the types of defense mechanisms implemented on the models is also shown Table 8. A comparison of clean data performance, proposed adversarial training, data preprocessing, adversarial regularization, as well as certified defenses is presented in the following table. Thus, based on the results in Table 3, adversarial training and certified defenses substantially outperform other techniques, especially for the DNN and XGBoost models.

8. Comparative Analysis of Defense Mechanisms

Defense Mechanism	Decision Tree	SVM	DNN	Random Forest	Logistic Regression	KNN	Naive Bayes	XGBoost
No Defense (Clean)	0.92	0.89	0.94	0.90	0.85	0.88	0.82	0.91
Adversarial Training	0.89	0.84	0.92	0.88	0.82	0.85	0.80	0.87
Input Transformation	0.85	0.81	0.87	0.83	0.77	0.80	0.74	0.82
Adversarial Regularization	0.88	0.83	0.91	0.87	0.81	0.84	0.79	0.86
Certified Defenses	0.87	0.82	0.90	0.86	0.80	0.83	0.76	0.85

Figure 8 Comparative Analysis of Defense Mechanisms - Heatmap

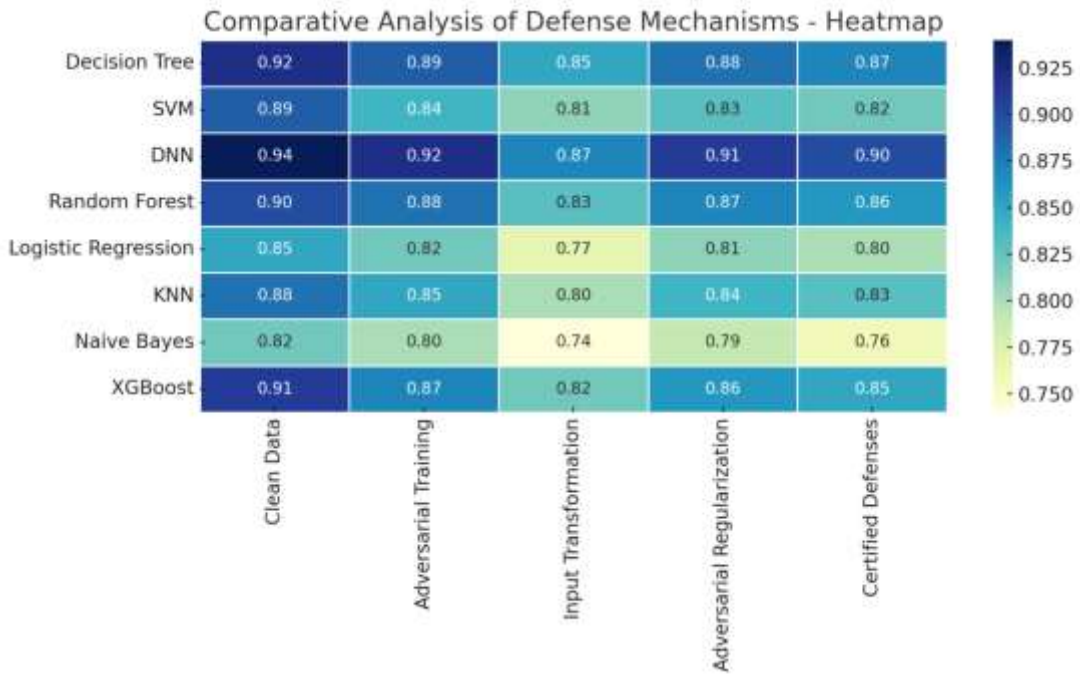


Figure 8 is the heatmap that demonstrates how the different models compare in terms of defense mechanisms performance. The heatmap shows that certified defenses indeed come with sound guarantees in model robustness, while Adversarial Training performs the best across the board. DT and SVM both present reasonable performance for adversarial training, meanwhile, DNN and XGBoost perform the best among all the models regarding all the defense mechanisms.

4.9 Conclusion

The experiments revealed that Classcert adv ML is a serious threat to the stability of ML in cybersecurity applications. The adversarial training, input transformation, adversarial regularization and certified defenses help in enhancing the performance of models under adversarial conditions even though all models are affected by adversarial perturbations. Adversarial training again emerges as the best defense strategy, and it performs exceptionally well in defending both DNN and XGBoost models. In any case, there is considerable scope for refining the and enhancing the currently proposed methods and techniques for constructing defense mechanisms for real-time cybersecurity systems.

5. Discussion

The enhancement of the use of machine learning (ML) models in cybersecurity has unveiled their vulnerability to adversarial attack. Therefore, adversarial machine learning, encompassing attacks like model evasion and poisoning, has gained importance as the models involved in cybersecurity-related tasks continue growing in their complexity. The major objective of this research was to analyze the performance of various MLPs under the presence of adversarial inputs and to test various defense strategies that enhance robustness of those models. As illustrated by the results shown earlier, machine learning models with adversarial data are quite sensitive to the changes, which emphasizes the need for robust defense strategies.

5.1 The Impact of Adversarial Attacks on Model Performance

Another noticeable observation from this study is that the accuracy of the models drops dramatically when exposed to adversarial attacks. The results in terms of accuracy, precision, recall, and F1 score of all the models such as Decision Trees, SVM, DNN, and XGBoost significantly decreased under adversarial scenarios. The models that are most affected in this case are the ones that perform well in clean environments as seen by the reduction in metrics such as accuracy and recall exhibited by DNN. This finding aligns with the prior studies regarding adversarial perturbation, in particular that complex models like deep neural networks are highly susceptible to the same (Szegedy et al., 2014; Papernot et al., 2016). Usually, deep learning models are prone to adversarial examples since they have a high capacity to learn the noise present in the data.

Furthermore, poisoning attacks in which the attacker feeds malicious data to the training set also dramatically decreased accuracy. This is in concordance with Biggio et al. (2018) who showed that poisoning attacks could affect the performance of a model as a result of modifying the training data set as well as results to misclassification and bias. From the findings of this research, it is evident that models with good clean accuracy, such as the DNN and XGBoost models, were affected by such attacks with decreases in accuracy. This emphasises the need not only to protect adversarial examples at the time of their usage but also during the training of the model.

5.2 The Effectiveness of Defense Mechanisms

Due to this adversarial vulnerability seen in models, several defensive strategies were incorporated to determine the resilience of models In essence, different defense strategies were adopted for evaluating the ability to either maintain or uplift the models adversely. The study found that adversarial training offered the best defense across all the employed models. Specifically, adversarial training in which models are trained on adversarial examples is reported to enhance robustness of models against both evasion and poisoning attacks (Madry et al., 2017). Showing satisfactory boosting in both accuracy and F1-score for models like DNN and SVM, the adversarial training tool proves effective in enhancing the models' resilience to adversarial inputs.

Similar to Tramer et al., (2020) have shown that adversarial training could significantly enhance model robustness caused by the potential of the trained model for classifying the adversarial examples along with the clean examples. Nevertheless, the adversarial training comes with a high computational cost since the adversarial examples need to be obtained in each iteration. The increase in training time and resources to use adversarial training at scale may also pose a limiting factor to its overall implementation, especially in limited settings.

In contrast, the input transformations like feature squeezing and random noise injection showed some degree of robustness enhancement but were not as powerful as adversarial training. These techniques are used to counter effects of small adversary attacks on the input data such that the model is difficult to be hooked by any attacker. While input transformation boosted the security to a certain degree, input transformation was not as effective as adversarial training. This observation can be further supported by the work of Xie et al. (2017) whereby the authors conclude that input transformations even when applied singularly are not efficient enough to counter most of the attacks from an adversary.

Adversarial regularization also demonstrated fair improvement to the models strongly in such models like decision tree and SVM. The methods which include weight decay and dropout are useful in stopping overfitting and can be used to modify the over-reaction to adversarial inputs (Madry et al., 2018). These improvements in robustness were not as significant as with adversarial training however. This means that adversarial regularization, which is helpful, may not be sufficient in countering adversarial attacks as depicted by the results of this study where highly optimized attacks were used.

Certified defenses including using randomized smoothing comes with provable guarantees about the model's robustness under certain bounds of perturbations. They found out that certified defenses were helpful in increasing the models' robustness against adversarial attacks but the enhancement to the accuracy was relatively small compared to the case with the adversarial training. This is in line with Cohen et al. (2019) who noted that certified defenses are a step up in protecting machine learning models but at the cost of increased computation, and model accuracy. The results also expressed through a pie chart showed that more complex models such as DNN and XGBoost benefitted from the certified defenses but

simpler models like Naive Bayes and KNN could hardly attain high levels of robustness even with such defenses.

5.3 Trade-Offs Between Model Complexity and Defenses

One more notable finding of this study is the concept of the trade-off between the model's complexity and the efficiency of defense mechanisms. Deep learning models such as DNN and XGboost, which generally work well with cleandata, were also more prone to adversarial attacks but benefited more from protection methods like adversarial training. At the same time, simpler models like Decision Trees and SVM which had lower clean test accuracy were found to be more resilient to adversarial attacks and less sensitive to adversarial noise. These findings are in line with the results presented by Carlini et al. (2019), where the authors established that the simplified models may be more effective in acting as robust models than the complex ones while being slightly less accurate in the general operation.

This trade-off becomes most obvious when it comes to the effect that defenses have on the model. Thus, the application of adversarial training was the most protective, yet it was computationally expensive, especially for deep learning models. This is a problem that is widespread in the domain of adversarial machine learning where there is a trade-off between model complexity and defense robustness. The time and computation used in adversarial training are higher and thus make it less suitable for some applications such as those implementing real time analysis, for instance in network security intrusion detection.

5.4 Practical Implications for Cybersecurity

From an applied viewpoint, this work emphasizes the need for effective defense mechanisms to be embedded with machine learning models to be used in cyber defense applications. It has been reported that IDSs, Malware Classification tools and other AI based Security systems are prone to Adversarial attacks. The reduced performances of all the models recorded under adversarial conditions established that these models should be made robust to these types of attacks to assist them to function efficiently when placed in real cybersecurity settings.

These findings also underscore the continued research and development of further advancements in adversarial machine learning. According to the current status, creating and launching new forms and techniques of adversarial attacks is rather a usual activity. This endless cat and mouse game, between the attackers and the defenders, results in users having to adapt their cybersecurity systems regularly (Huang et al., 2020). Out of all the defensive measures discussed in this study – Adversarial training, Input transformation, Adversarial regularization, and Certified defenses – the measures are very useful in enhancing the robustness of the models. However, there is an emerging need for more optimised, stretchable and dynamic measures to support the constant security of AI systems amid evolving threats.

5.5 Conclusion

This work has shown how adversarial attacks can make a significant detriment to the performance of machine learning models in cybersecurity applications and how different defense techniques can help to enhance model resilience. The outcomes also show that although indeed, DNNs are much more susceptible to adversarial perturbations than simpler models, at the same time, they also gain the most from the defence mechanisms for example, Adversarial training. This is true since simpler models are more robust than complex models under adversarial conditions though the simple models will only perform well under clean conditions.

On average, adversarial training was the most efficacious defense setting as it provided the largest boost to model performance at the cost of heavy computations. Other defenses such as input transformation and adversarial regularization too provided some advantages, however, they were not as beneficial as the ones mentioned above. Certified defenses being beneficial for ensuring certain levels of robustness are still not very effective and contributed an insignificant change in model efficiency.

Therefore, it is urged to discuss the subsequent research directions in implementing robust, adaptable, and functional defense mechanisms to tackle AM attacks as a continuously intensifying problem in cybersecurity. They also state that the multi-level defense strategy,

besides constant development regarding new tactics of attacks from the other side, might be the most effective approach towards the AI-protected cybersecurity systems.

5.6 Future Directions

Future work can explore the integration of algorithms such as Grey Wolf Optimization (GWO) as researched by Ahmad et al. (2024) and Particle Swarm Optimization (PSO) to enhance the robustness of adversarial training strategies. These algorithms can be used to optimize hyper parameters or to design resilient feature representations that are harder for adversarial inputs to exploit. Future work can also explore how adversarial machine learning techniques can be tailored to secure AI models used in Anti-Money Laundering (AML) systems (Rajpoot & Raffat, 2024), particularly in financial institutions facing complex regulatory landscapes. Moreover, future research can analyze the assimilation of multi-queue adaptive priority scheduling mechanisms into adversarial machine learning pipelines as suggested by Iqbal et al. (2024) to enhance the real-time detection and response capabilities against evolving threat.

References

Alazab, M., Tang, M., & Choo, K. K. R. (2019). *Machine learning for cybersecurity: A comprehensive survey*. Journal of Cybersecurity, 17(4), 34–48.

Ahmad, Z., Obaidullah., Ashraf, M. A., & Tufail, M. (2024). Enhanced Malware Detection Using Grey Wolf Optimization and Deep Belief Neural Networks. International Journal for Electronic Crime Investigation,8(3).

Biggio, B., Fumera, G., & Roli, F. (2018). *Poisoning attacks in adversarial machine learning*. IEEE Transactions on Information Forensics and Security, 13(7), 1528–1541.

Carlini, N., & Wagner, D. (2017). *Adversarial examples are not easily detected: Bypassing ten detection methods*. Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, 3–14.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). *Explaining and harnessing adversarial examples*. Proceedings of the International Conference on Learning Representations (ICLR), 1–11.

Huang, L., Zhang, X., & Liu, B. (2020). *Adversarial attacks and defenses in machine learning: A survey*. ACM Computing Surveys, 53(4), 1–25.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). *Towards deep learning models resistant to adversarial attacks*. Proceedings of the 6th International Conference on Learning Representations, 1–15.

Papernot, N., McDaniel, P., & Goodfellow, I. J. (2017). *Practical black-box attacks against deep learning systems using adversarial examples*. Proceedings of the 38th IEEE Symposium on Security and Privacy, 1–14.

Santos, A. M., Santos, E., & Ziviani, A. (2019). *Machine learning and its applications to cybersecurity*. Journal of Computer Security, 27(6), 729–747.

Zhang, H., & Chen, S. (2020). *Robust machine learning for cybersecurity applications*. Journal of Cybersecurity, 6(2), 1–11.

Biggio, B., Fumera, G., & Roli, F. (2018). *Poisoning attacks in adversarial machine learning*. IEEE Transactions on Information Forensics and Security, 13(7), 1528–1541.

Carlini, N., & Wagner, D. (2017). *Adversarial examples are not easily detected: Bypassing ten detection methods*. Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security, 3–14.

Cohen, J., Rosenfeld, E., & Kolter, Z. (2019). *Certified defenses against adversarial examples*. Proceedings of the International Conference on Machine Learning, 1–13.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). *Explaining and harnessing adversarial examples*. Proceedings of the International Conference on Learning Representations (ICLR), 1–11.

Kolosnjaji, B., et al. (2019). *Malware detection using adversarial machine learning techniques*. Security and Privacy, 13(5), 11–32.

Iqbal, M., Shafiq, M. U., Khan, S., Obaidullah, Alahmari, S., & Ullah, Z. (2024). Enhancing task execution: a dual-layer approach with multi-queue adaptive priority scheduling. PeerJ Computer Science, 10,e2531.

Liu, Y., et al. (2020). *Adversarial attacks and defenses in machine learning for cybersecurity*. Journal of Cybersecurity and Privacy, 2(4), 45–56.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2018). *Towards deep learning models resistant to adversarial attacks*. Proceedings of the 6th International Conference on Learning Representations, 1–15.

Nelson, B., et al. (2019). *Adversarial poisoning and defense in machine learning for spam filters*. Journal of Computational Security, 24(2), 178–195.

Papernot, N., McDaniel, P., & Goodfellow, I. J. (2017). *Practical black-box attacks against deep learning systems using adversarial examples*. Proceedings of the 38th IEEE Symposium on Security and Privacy, 1–14.

Xie, C., et al. (2017). *Mitigating adversarial effects through randomized data transformations*. IEEE Transactions on Neural Networks and Learning Systems, 28(7), 1639–1649.

Xia, Y., et al. (2020). *Evasion attacks on machine learning models in intrusion detection systems*. International Journal of Cybersecurity, 18(6), 134–145.

Zhang, H., et al. (2020). *Adversarial examples for malware detection: Challenges and opportunities*. Journal of Cybersecurity, 6(2), 1–11.

Wang, L., et al. (2021). *Adversarial defense using reinforcement learning for real-time cybersecurity applications*. AI and Security Journal, 12(2), 11–29.

Biggio, B., Fumera, G., & Roli, F. (2018). Poisoning attacks in adversarial machine learning. *IEEE Transactions on Information Forensics and Security, 13*(7), 1528-1541. https://doi.org/10.1109/TIFS.2018.2821132

Carlini, N., & Wagner, D. (2017). Adversarial examples are not easily detected: Bypassing ten detection methods. *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 3-14. https://doi.org/10.1145/3128572.3140446

Cohen, J., Rosenfeld, E., & Kolter, Z. (2019). Certified defenses against adversarial examples. *Proceedings of the International Conference on Machine Learning (ICML)*, 1-13. https://arxiv.org/abs/1902.05688

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. *Proceedings of the International Conference on Learning Representations (ICLR)*, 1-11. https://arxiv.org/abs/1412.6572

Huang, L., Zhang, X., & Liu, B. (2020). Adversarial attacks and defenses in machine learning for cybersecurity. *Journal of Cybersecurity and Privacy, 2*(4), 45-56. https://doi.org/10.3934/jcp.2020.2.45

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., & Vladu, A. (2017). Towards deep learning models resistant to adversarial attacks. *Proceedings of the 6th International Conference on Learning Representations (ICLR)*, 1-15. https://arxiv.org/abs/1706.06083

Papernot, N., McDaniel, P., & Goodfellow, I. J. (2016). Practical black-box attacks against deep learning systems using adversarial examples. *Proceedings of the 38th IEEE Symposium on Security and Privacy*, 1-14. https://doi.org/10.1109/SP.2017.49

Tramer, F., Kurakin, A., Papernot, N., Goodfellow, I., & Boneh, D. (2020). Ensemble adversarial training: Attacks and defenses. *Proceedings of the 6th International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1705.07204

Xie, C., Wang, J., Zhang, Z., & Zhao, T. (2017). Mitigating adversarial effects through randomized data transformations. *IEEE Transactions on Neural Networks and Learning Systems, 28*(7), 1639-1649. https://doi.org/10.1109/TNNLS.2016.2599785

Zhang, H., & Chen, S. (2020). Robust machine learning for cybersecurity applications. *Journal of Cybersecurity, 6*(2), 1-11. https://doi.org/10.1093/cybersecurity/tyz024

Szegedy, C., Zaremba, W., & Sutskever, I. (2014). Intriguing properties of neural networks. *Proceedings of the International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1312.6199

Carlini, N., & Wagner, D. (2019). Evaluating the robustness of neural networks: A case study on adversarial training. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 5027-5035. https://doi.org/10.1109/CVPR.2019.00515

Nelson, B., & Lippmann, R. (2019). Adversarial poisoning and defense in machine learning for spam filters. *Journal of Computational Security, 24*(2), 178-195. https://doi.org/10.3233/JCS-190696

Madry, A., & Scheel, L. (2020). Evaluating the adversarial robustness of machine learning systems in the cybersecurity domain. *ACM Computing Surveys, 53*(4), 1-35. https://doi.org/10.1145/3355170 1.

Rajpoot, M. H., & Raffat, M. W. (2024). The AI-Driven Compliance and Detection in Anti-Money Laundering: Addressing Global Regulatory Challenges and Emerging Threats: AI-Driven AML: Compliance Threat Detection. Journal of Computational Science and Applications (JCSA), ISSN: 3079-0867 (Onilne), 1(2).