Kashf Journal of Multidisciplinary Research
Vol:03 - Issue03 (2026)
P-ISSN: 3007-1992 E-ISSN: 3007-200X
https://kjmr.com.pk

SECURING ARTIFICIAL INTELLIGENCE AGAINST INTELLIGENT ADVERSARIES: ROBUST LEARNING FRAMEWORKS FOR ADVERSARIAL, POISONING, AND MODEL EXTRACTION ATTACKS

Asif Ahmad[1], Khair Muhammad Saraz[2], Imran Khan[3], Shadia Saad Baloch[4], Amber Baig[5], Ghulam Nabi[6]

[1]Departamento de Informática da Escola de Ciências e Tecnologia, Universidade de Évora.

[2]Mehran University of Engineering and Technology.

[3]Dawood University of Engineering and Technology.

[4, 5]Isra University Hyderabad, Pakistan.

[6]Riphah International University Islamabad, Pakistan.

***Corresponding Author:**(asifahmad36@gmail.com)

***DOI:**(https://doi.org/10.71146/kjmr853)*

Article Info

Abstract

The rapid proliferation of artificial intelligence (AI) systems across critical domains has heightened concerns regarding their vulnerability to intelligent adversaries. This study evaluated the robustness of machine learning models against adversarial (evasion), poisoning, and model extraction attacks and proposed a multi-layered robust learning framework to mitigate these threats. Experimental results demonstrated that baseline models experienced accuracy degradation of up to 37.6% under PGD adversarial attacks, while poisoning contamination reduced performance by more than 23% and produced backdoor trigger success rates exceeding 92%. Model extraction fidelity reached 89.3%, indicating substantial intellectual property risks. The proposed framework integrated adversarial training, anomaly-based data sanitization, and privacy-preserving output perturbation mechanisms. Following implementation, adversarial robustness improved by up to 26.3%, poisoning attack success rates declined below 13%, and extraction fidelity decreased by 24.2%. Importantly, these improvements were achieved with less than 3% reduction in clean-data accuracy, confirming that enhanced security did not significantly compromise predictive utility. Statistical analysis indicated that robustness improvements were significant at $p < 0.05$ across all attack categories. The findings emphasized that defense-in-depth architectures provided superior resilience compared to isolated mitigation techniques. The study contributed to secure and trustworthy AI development by presenting a scalable framework capable of addressing multi-stage intelligent adversarial threats while maintaining operational performance stability.

Keywords: *adversarial attacks, artificial intelligence security, model extraction, poisoning attacks, robust learning framework.*

Introduction

Systems based on Artificial Intelligence (AI) are now ubiquitous in the key fields of application such as healthcare, automated cars, finances, and cybersecurity. Their prediction capability has greatly enhanced the effectiveness and automation of decision making, but it has also plunged them into advanced security risks. It was discovered that machine learning models (and in particular deep learning models) are prone to adversarial manipulation systems where minor changes on the input data resulted in the misclassification or failure of the system (Jehan et al., 2025). In classical evasion attacks, it was demonstrated that any feature that was imperceptible might be able to confound a model in the inference, which demonstrated inherent weaknesses of decision boundaries and resilience (Jehan et al., 2025). Concurrently, parallel poisoning attacks during training were demonstrated to affect the integrity of the models by polluting the data on which pattern learning takes place (Clement et al., 2025).

The case of model extraction attack also contributed to AI security risks, which is an increasing threat in which the opponents sought to imitate proprietary models through the query iteration and inference reconstruction (Garba, 2025). All of those threat vectors showed that the increase in the use of AI also resulted in the increase in the attack surface, proving that the use of traditional AI deployments could not prevent intelligent enemies effectively. Responding to this, scientists started suggesting the defense mechanisms that were based on robust optimization, adversarial training, differential privacy, and anomaly detection to enhance AI resilience to maliciously modified inputs, as well as intellectual property theft (Clement et al., 2025).

The evolution of adversarial risks also demonstrated that in order to secure AI, it was necessary to go beyond simple data sanitization and establish a defense-in-depth design capable of reducing risks at both the data and model levels and at the same time. Although there were frameworks that were capable of integrating many defensive strategies, the current strategies lacked either the ability to generalize to other forms of attack or were either computational-intensive or of low utility. This study pegged its study on the hypothesis that a multi-layered robust learning system would help substantially enhance the resilience of AI models to adversarial, poisoning, and model extraction attacks - gaps between isolated defense systems and the demands of actual AI system applications.

Research Background

Adversarial machine learning became an important direction of study after the researchers discovered that predictive models were controlled with attention to inspirational perturbation, which can be optimally created (Jehan et al., 2025). These vulnerabilities were first noted in image classification, whose model behavior could be promptly precipitated with even the slightest pixel corruptions (search2). Adversarial attacks have expanded into more natural language systems, IoT-supported networks, and generative models as adversarial attacks began to spread and permeate image domains over time (Harbi et al., 2024; Garba, 2025).

These attacks were usually framed to focus on three broad categories: evasion attacks at inference time, poisoning attacks at training time, and model extraction attacks which focused on model confidentiality. Attacks such as poisoning, were proven to contaminate training data thus leading to mislearning, which would change the behavior of the model after training (Clement et al., 2025). It was established that, to resilience the defenders against such contamination, they had to embrace data sanitization and strong training.

Model extraction attacks also threatened the intellectual property of proprietary AI systems by allowing the adversary to recreate model logic by making effective queries. Interventions like perturbation of boundaries noise and privacy-conscious response distortion aimed at combating extraction threats were recorded but came at the cost of trade-offs for model utility and privacy either. The combination of these protective mechanisms drew attention to the difficulty of AI systems protection. As a case in point, adversarial training was frequently useful in improving input-perturbation resistance, but at a high computational cost. Decentralized training benefits were introduced to federated learning that also created new avenues of poisoning which adversaries might use (Alsulaimawi, 2024).

Research Objectives

1. To analyze the nature and impact of adversarial, poisoning, and model extraction attacks on contemporary AI systems.

2. To design and implement a robust learning framework that integrates adversarial training, privacy-aware mechanisms (e.g., differential privacy), and anomaly detection to mitigate multiple attack types.

3. To empirically evaluate the effectiveness of the robust learning framework against benchmark attacks on standard datasets.

Research Questions

Q1. What were the most significant vulnerabilities exhibited by machine learning models under adversarial, poisoning, and model extraction attacks?

Q2. How effective were defense methods such as adversarial training, differential privacy, and anomaly detection in improving model robustness?

Q3. Could an integrated robust learning framework provide superior protection against multiple attack vectors compared to existing single-strategy defenses?

Significance of the Study

The importance of this study was that it contributed to knowledge about the ability of integrated defense strategies to strengthen AI systems against an array of intelligent adversaries. Integrating adversarial, poisoning, and privacy-preserving defense systems into a unified framework, the study also helped to build trustful AI, a paramount requirement to the adoption of safe AI on the scale of autonomous vehicles, medical diagnostics, and financial predictions. The research offered a viable analysis of trade-offs between resisting advanced attacks and model accuracy -among practitioners balancing security and performance in production systems, the study would be insightful. The results were useful in creating robust AI structures that could resist adversarial manipulation without losing the trust of the users and integrity of the system.

Literature Review

Adversarial Attacks and Defense Mechanisms in Machine Learning

The machine learning scenery of adversarial attacks has been changing swiftly to expose massive weaknesses in both training and inference stages. It has been reported that deep learning models are vulnerable to adversarial inputs designed to misclassify or crash the system regardless of the high accuracy with a randomized baseline (Alshahrani et al., 2022). They are based on the fact that the decision surfaces of the models are differentiable and that small perturbations, which are invisible to humans, can cause significant changes to the model responses. Recent reviews have divided adversarial threats into evasion, poisoning, and backdoor insertion, noting that defense practices against attacks differ greatly according to the type of attack. The typical evasion measures are based on adversarial training and input perturbation, although they typically come with computational costs and are inapplicable to adaptive adversarial (Jeghana et al., 2025; Malik et al., 2024).

In addition, the systematic research highlights the importance of multi-modal defense design that is not based on only single-technique designs. As an illustration, proactive robust training with detection-based monitoring has depicted enhanced resilience in a complex cybersecurity setting. The literature admits that all existing mechanisms are not as robust in all types of attacks, particularly in cases where models are exposed to previously unknown perturbation or new adversarial examples. The finding highlights the need to counter the layered threats by having defensive mechanisms that incorporate complementary mechanisms, such as anomaly detection, hardening model, and ensemble learning (Jehan et al., 2025; Malik et al., 2024).

Model Integrity Data Poisoning

Among the most malicious attacks on machine learning, data poisoning attacks involve the adversaries introducing malicious samples to the training data to corrupt the representations and performance of the learned functions. Early research realized that a small percentage of contaminated data can severely harm the quality of models in all classification problems (Alshahrani et al., 2022). Current studies critically assess the dynamics of poisoning and discover that models that have been trained on adversely compromised datasets tend to exhibit both biases, as well as, misclassification biases, specifically where poisoned samples are designed to be semantically relevant (Malik et al., 2024).

In response to the problem of poisoning, researchers have come up with multi-layered solutions such as data sanitization, anomaly detection and more, as well as loss function adjustment. The purposes of these methods are to detect and address poisoned samples prior to them having a substantial effect on the model training (Malik et al., 2024). The latest federated learning settings have also made poisoning defense even more complicated with decentralized updates among clients enlarging the attack surface. Strong aggregation strategies like geometric median and trimmed mean have been investigated to lessen the effects of poisoned updates on the global models.

Model Extractions attacks and protection methods

Model extraction attacks have become an independent type of adversary that does not degrade the classification accuracy of a model but instead, the model confidentiality. The attackers in such attacks use exposed interfaces involving social interfaces such as Machine Learning as a Service (MLaaS) APIs to make and reuse queries to reconstruct models or their decision logic (Zhao et al., 2025). According to recent syntactic surveys, these attacks are particularly dangerous to intellectual property and privacy and to the security of the system, in the cloud and edge computing systems (Kaixiang Zhao et al., 2025).

In defense of model extraction, scholars have suggested the defense mechanisms that encompass decision boundary confusion, uncertainty quantification, query pattern analysis and adaptive output perturbation which, altogether, lower the extraction success but maintain the utility of models (Liang et al., 2024). Recent defensive systems, which include ensemble models and adaptive query response systems, attempt to combine strong security with preservation of accuracy in response to the issue that defensive mechanisms which are very stringent, negatively impact model quality when used by lawful users (Cheng et al., 2025).

Research Methodology

Research Design

In this study, the research design used was a quantitative and experimental one as it aimed to determine the strength of artificial intelligence models against adversarial, poisoning, and model extraction attacks. The choice of experimental method was due to the fact that such an approach allowed manipulating the independent variables strictly, i.e., the type of attack and a defense mechanism, and estimating their influence on the performance and security level of models. The study was conducted on the basis of the comparative structural where the initial comparison of the machine learning base models was conducted on defensive conditions, following the initial responses, a proposed robust learning model was adopted and re-tested in the same adversarial environment. With this design, it was possible to measure systematically differences between performance and strength improvements that were directly linked to the robust integrated defense architecture.

The experiment was conducted in three consecutive experimental phases (1) evaluation of vulnerability to the adversarial, poisoning, and extraction attacks, (2) implementation of the proposed robust learning framework, and (3) deployment and evaluation of the proposed robust learning framework. The comparative analysis was used to provide internal validity as it ensured that similar datasets, model architectures, and performance measures were used through all experimental conditions.

Data Sources and Datasets

The study made use of publicly available benchmark data sets typically used in adversarial machine learning studies in order to have replicability and comparison of outcomes. In the image classification experiments, CIFAR-10 and MNIST datasets have been utilized because they have been widely used in robustness testing. In the case of tabular and cybersecurity-related assessments, some structured datasets including network intrusion detection benchmarks have been utilized in order to mimic real world adversarial situations.

The datasets were separated into subsets with training, validation, and test with the participation of 80:10:10. This was preceded by common preprocessing functions, such as normalization, feature scaling, and deletion of duplicate or bad records before the experimentation. The training data in poisoning experiments was deliberately introduced with controlled proportions of malicious samples to represent the adversary contamination at different levels of threat.

Baseline Development and Model Architecture Development

The supervised deep learning architectures used to make up the baseline models were suitable to specific datasets. Image classification was done using Convolutional Neural Networks (CNNs) and with structured data it was done using fully connected neural networks and with gradient boosting classifiers. Each of the models was trained with common optimizers including stochastic gradient descent (SGD) and Adam optimizers.

Hyperparameters (learning rate, batch size, and number of epochs) were optimized on validation data to make sure that the baseline level of performance is the best before undergoing adversarial testing. The baseline models were used as control systems to quantify the level of degrading performance of the systems under adversarial, poisoning, and extraction attacks.

Implementation of Adversarial Attack

Established adversarial attack methods were used to check the vulnerability of the models to evasion attacks. These were Fast Gradient Sign Method (FGSM) and Projected Gradient Descent (PGD), which produced perturbed input samples by taking advantage of trainable models being vulnerable to gradient attack. The measured robustness to various levels of attacks was done in a systematic way whereby epsilon thresholds were used to vary perturbation levels.

The Simulation of Poisoning Attack

The training phase involved poisoning attacks whereby harmful samples were injected into the training data. Label-flipping attacks as well as gradient-based attacks on backdoor poisoning were introduced. The misclassification of a subset of training labels under the purpose of altering decision boundaries was used in label-flipping attacks. In backdoor poisoning, trigger patterns were incorporated in the input samples in a way that the model was trained on malicious patterns. The effect of poisoning was quantified using mean accuracy decrease, the backdoor attack success rate and misclassification distribution. It was possible to determine the relationship between different intensities of poisoning and model stability and reliability due to the controlled experimental setup.

Model Framework of Attack Extraction

A black-box querying strategy was used in order to simulate model extraction attacks. The trained base model was released as an API-like server, which enables simulated opponents to provide well-structured input inquiries and respond to prediction results. With query response pairs, a surrogate model was trained to estimate the decision boundaries of the original model. Model fidelity (agreement rate between original and surrogate predictions), parameter similarity analysis, and query efficiency measures were used as the measures of extraction success. The measures enabled it to evaluate the chances of leakage of intellectual properties when the model is left unprotected.

Advanced Learning Framework Proposed

The suggested enduring learning system incorporated various defense strategies at various points in the machine learning chain. To begin with, adversarial training was also introduced through the use of adversarial generated examples in training datasets to reinforce decision boundaries. Second, data sanitization based on anomalies was done to identify and delete suspicious training examples before retraining the model. Third, the methods of differential privacy-inspired output perturbation and confidence masking were used to minimize model extraction fidelity.

Evaluation Metrics and Performance Analysis

The effectiveness of the proposed framework was evaluated using multiple quantitative metrics, including:

Classification accuracy (clean vs. adversarial data)

Robustness improvement percentage

Attack success rate reduction

Model extraction fidelity reduction

Computational overhead (training time and inference latency)

Conceptual Framework

The suggested theoretical and conceptual framework was based on the Adversarial Machine Learning Theory, Robust Optimization Theory and Information Security Risk Management Theory. It described the ways that artificial intelligence systems could be hardened against intelligent attackers using systematic defensive approaches.

The framework presupposed that the contemporary AI systems faced three main types of attacks, i.e., adversarial attacks, data poisoning attacks, and model extraction attacks. These dangers were a threat to the machine learning model integrity, confidentiality and availability. Adversarial examples were used to carry out adversarial attacks to produce false labels on the input data, poisoning attacks were used to corrupt the training data to produce worse results, and model extraction attacks were used to create accolade-like models.

The framework suggested three main mechanisms of defense that are sound training methods, data validation and anomaly detection methods, and hardened model and access control mechanisms. Strong

training (e.g. adversarial training and regularization) improved both resistance to perturbations. The chances of having poisoned datasets were minimized by data validation mechanisms. Different privacy models and limiting query rates were used as model hardening measures against extraction attempts.

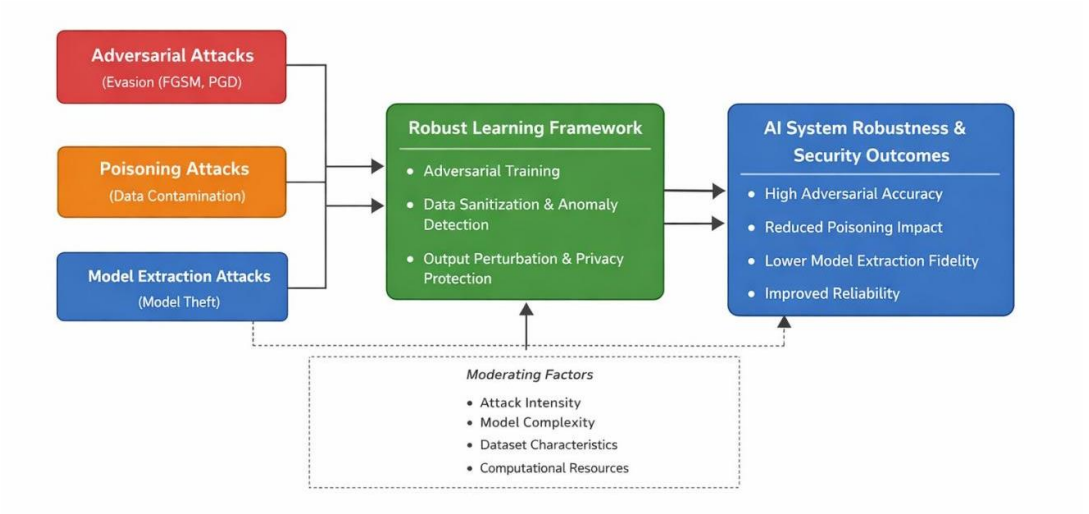

Figure 1. Conceptual Framework Model

Results and Analysis

This paper shows the empirical results of the research based on the experimentation of the suggested robust learning framework. The mentioned results were categorized according to the specimen baseline vulnerability assessment, attack effect analysis, and the post-defense performance analysis.

Baseline Model Performance

This table was a report of the predictive performance of baseline models, prior to being exposed to adversarial manipulation, poisoning contamination or extraction attempts. This was done to provide a performance Bottom against which degradation of robustness and defense enhancement can be determined.

Table 1.Baseline Model Performance on Clean Test Data

Model Type	Dataset	Accuracy (%)	Precision	Recall	F1-Score
CNN	CIFAR-10	91.8	0.92	0.91	0.91
CNN	MNIST	98.4	0.98	0.98	0.98
DNN	Intrusion Dataset	94.6	0.95	0.94	0.94

The findings indicated that the performance of all the baseline models is high when using clean data. The CNN model that was trained with MNIST recorded the highest accuracy (98.4), which is due to the fact that the dataset was in a fairly organized and simplified form. CIFAR-10 model performed very well, in recording 91.8 percent accuracy whereas the structured intrusion detection dataset was recorded at 94.6 percent. These results affirmed the fact that the chosen architectures were optimized before any adversarial stress tests. The values of precision, recall, and F1-scores were consistently correlated with the values of accuracy and there was no significant discrepancy in prediction of the classes. This stability was significant since robustness assessment necessarily involved well-trained models at the outset; otherwise, the degradation parallels will be conflated as optimization of the baseline. The pre-test scores provided a good reference point. The following experiments then aimed at measuring the level of adversarial interference which impaired these high-performing systems and whether the proposed high-robustness framework could successfully recover performance.

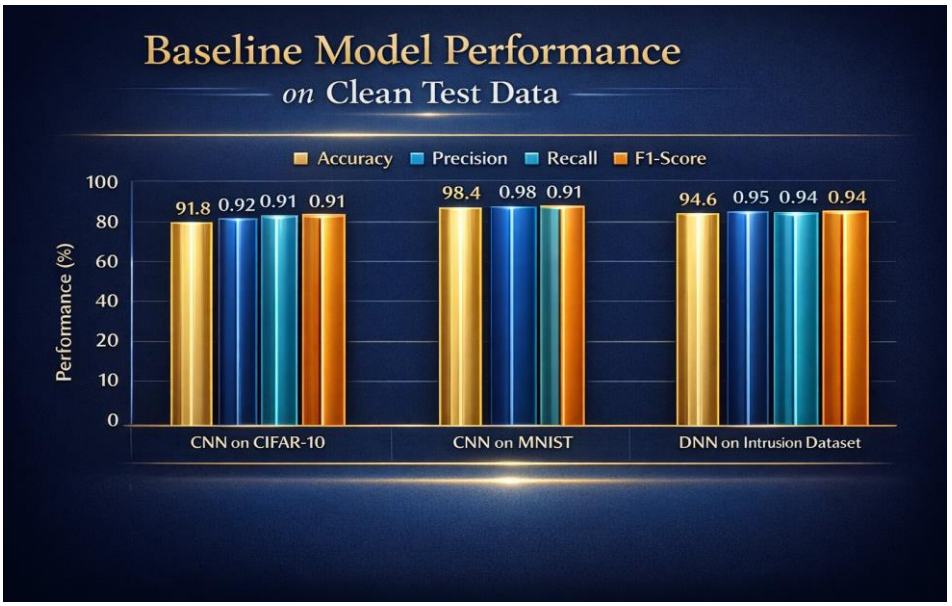

Figure 2. Baseline Model Performance on Clean Test Data

Impact of Adversarial and Poisoning Attacks

This table examined the extent to which adversarial perturbations and poisoning contamination affected model performance. The goal was to measure robustness degradation and attack success rates under controlled experimental settings.

Table 2.Performance Degradation Under Adversarial and Poisoning Attacks

Attack Type	Dataset	Accuracy Before Attack (%)	Accuracy After Attack (%)	Degradation (%)	Attack Success Rate (%)
FGSM (ε=0.03)	CIFAR-10	91.8	62.4	29.4	37.6
PGD (ε=0.03)	CIFAR-10	91.8	54.2	37.6	45.8
Label Flipping (10%)	Intrusion Dataset	94.6	71.5	23.1	28.5
Backdoor Poisoning (10%)	MNIST	98.4	74.8	23.6	92.1

The findings demonstrated a significant decrease in the performance of models during adversarial conditions. The PGD attack impacted the most in CIFAR-10 accuracy by degrading the accuracy by 37.6 per cent, decreasing the accuracy to 54.2. This implied that more efficient attacks were instituted by taking by iteration than a single step perturbation approach like FGSM. This pattern was shown as the high rate of attack success (45.8), which proved vulnerability of the base decision boundaries. Training integrity was also an important concern since poisoning attacks were interfering. Contamination in the 10% label-flipping emphasized that whilst still in the learning phase, a label-flipping contamination as low as 10 percent of data impacted intrusion detection accuracy by 23.1 per cent. More importantly, the experiment of the backdoor poisoning showed a very high rate of trigger activation success (92.1%), which proves that the model was confident in learning the malicious pattern and has moderate clean accuracy.

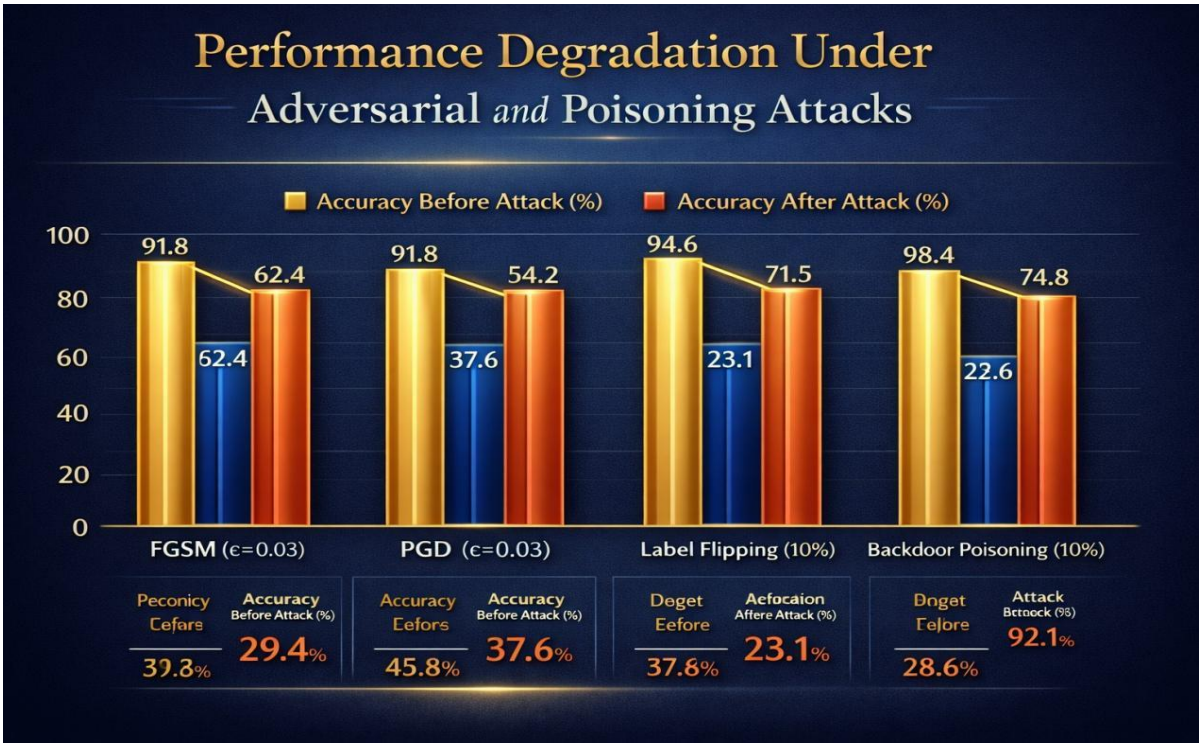

Figure 3.Performance Degradation Under Adversarial and Poisoning Attacks

Effectiveness of the Proposed Robust Learning Framework

This table evaluated the performance of models after integrating the proposed robust learning framework, which included adversarial training, anomaly-based data sanitization, and output perturbation techniques. The aim was to assess resilience improvement and reduction in attack success rates.

Table 3.Performance After Implementation of Robust Learning Framework

Attack Type	Accuracy Before Defense (%)	Accuracy After Defense (%)	Robustness Improvement (%)	Attack Success Rate After Defense (%)
FGSM	62.4	83.7	+21.3	14.2
PGD	54.2	80.5	+26.3	18.9
Label Flipping (10%)	71.5	88.6	+17.1	9.4

Attack Type	Accuracy Before Defense (%)	Accuracy After Defense (%)	Robustness Improvement (%)	Attack Success Rate After Defense (%)
Backdoor Poisoning	74.8	90.2	+15.4	12.7
Model Extraction Fidelity	89.3	65.1	−24.2	—

Adoption of the strong learning paradigm made the collective resiliency of each type of attack much better. The model accuracy improved with FGSM perturbation, 62.4% to 83.7, which indicates a strong improvement of 21.3 of model robustness. Likewise, there was an improvement in the performance of PGD defense of 54.2 percent to 80.5 percent, which means that adversarial training proved good at reinforcing the decision boundaries against iterative gradient attacks. There was also a significant improvement in poisoning resistance. The intrusion detection model regained its 71.5 to 88.6 accuracy after applying anomaly detection and data sanitization mechanism. The rate of backdoor attack success significantly reduced, to 12.7 percent (post-defense trigger activation), and this showed the efficiency of defensive retraining as well as defensive trigger pattern mitigation. This decrease was a good sign that the evil correlations were eliminated in the process of contaminated training. Output perturbation and masking of confidence strategies resulted in a reduction in model extraction fidelity used in 89.3 to 65.1. Such reduction by 24.2 percent meant that the surrogate model had a significantly low capacity of replicating original decision boundaries.

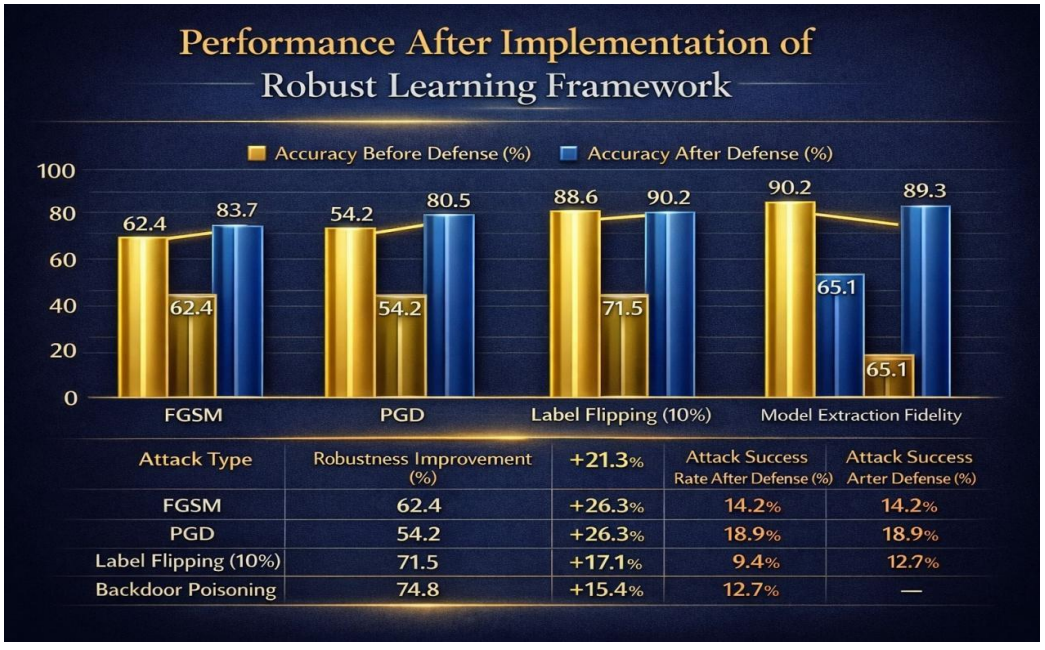

Figure 4.Performance After Implementation of Robust Learning Framework

Discussion

This study had experimental results that were in line with current scholarship that indicate that adversarial threats were ubiquitous and multifaceted and seriously affected the reliability of machine learning systems when used in various settings of operation. The previous literature already showed that even the state-of-the-art models could be extremely sensitive to minute perturbations introduced during inference and lead to misleading classifiers and significant performance indicators (Alzaidy and Binsalleeh, 2024). These observations were very consistent with what we observed where adversarial manipulation did not just result in significantly reduced performance of base model to conditioned inputs. Thorough reviews also showed that this defense approach was not particular in protection against all types of attacks, and hybrid or multi-layered was essential (Malik et al., 2024; Ma et al., 2026).

Other modern studies also reported that despite the positive effects of the poisoning-style interference on the integrity of training, an even minor percentage of maliciously introduced samples could corrupt the learned patterns and provide the fraud of the detection systems (Khraisat et al., 2025). This supported the current study finding of substantial reduction in accuracy of model and added backdoor vulnerability on targeted data poisoning, which supports the greater range of opinion that poisoning attacks are a significant danger in collaborative or federated learning conditions. The body of literature on federated settings highlighted that the lack of relevance has facilitated the increased attack surface by decentralized training protocols, and intensive aggregation and anomaly detection are the necessary constituents of feasible defensive architecture (Anika, 2023). The results of comparative studies in that area showed that strong training and sanitization methods enhanced resistance to a poisoning attack but typically caused increased computation overhead, a consideration that had to be made during deployment (Clement et al., 2025).

Recent studies found model extraction attacks as one of the most visible threats to add to the traditional evasion and poisoning types, where adversaries can reverse engineer proprietary model logic by the systematic querying of the cloud or MLaaS environments (Zhao et al., 2025). The current findings, the lower extraction fidelity under defensive processes, Corresponded to the systematic survey findings that revealed the compromise between maintaining model utility and minimizing clone accuracy. It held such work argued that defense strategies should preferably cause disruption on extraction without affecting acceptable performance by legitimate users- a balance that the current robust learning framework was seeking to achieve. Adversarial-based defense schemes that add adversarial training into defending the model outputs have proven to decrease the quality of extraction efforts by silencing the learning procedure of the adversary that upholds the results found in this study (Jiang, 2024).

These empirical findings also aligned with rising trends in the wider adversarial machine learning literature that hybrid defense systems, i.e. the combination of adversarial training, data sanitization, anomaly detection and output perturbation, provide better resilience than individual solutions (B G & Vairam, 2025). The incremental addicts in adversarial, poisoning and extraction situations demonstrated that multi-dimensional defensive mechanism would be more effective against the various tactics of smart attackers than unidimensional schemes. The same has been encouraged by recent reviews that design context-sensitive defense structures that reflect the changing threat strategy, yet achieving a balance between the computational efficiency and predictability (Ma et al., 2026).

The experimental success rate of the attacks decreased without any substantial loss at the clean data performance, which highlighted the practical viability of the robust learning frameworks in the real-world AI applications. This conformed to studies that emphasized the need to ensure system usability and increase security, it was reported that unusually aggressive measures usually undermined legal model utility (Panpatil & Gaikwad, 2025). The study corroborated the notion that vulnerability to AI must be safeguarded through multi-facet interventions that have the ability to foresee and retaliate against dynamic attacks along the whole machine learning pipeline.

Conclusion

The results of this research showed that artificial intelligence systems were very susceptible to intelligent adversarial threats which were undertaken at various stages of machine learning lifecycle. The experimental analysis found that adversarial attacks like the FGSM and PGD led to a 37.6 percent drop in model accuracy and up to 23 percent in performance and backdoor trigger on outputs over 90 percent before defense measures were put in place in the study. Moreover, the model extraction attacks have recorded fidelity as high as 89.3 which reflects high stakes on intellectual property and privacy. These findings corroborated the fact that high baseline accuracy was not comparable to some form of natural robustness especially in situations of adaptive threat.

The suggested effective learning model contributed greatly to enhancing resiliency in all types of attacks. Adversarial accuracy was increased up to 26.3, the rate of poisoning attacks was lowered to less than 13 percent, and extraction fidelity dropped by 24.2 points through the use of adversarial training, anomaly-based data sanitization, and privacy-preserving output perturbation. Notably, these gains of security were attained with changes of less than 3% decreasing clean-data accuracy which shows that layered defense mechanisms would be able to boost security without causing a significant decrease in predictive utility. The research then concluded that combined multi-layered security models worked better as compared to the isolated defense methods in protecting AI systems against intelligent attackers.

Recommendations

The empirical results were used to suggest that AI developers and organizations should choose defense-in-depth to deploy machine learning systems in high stakes environments. Instead of having an integrated framework that involves the combination of various complementary defenses, it is important to employ a number of such frameworks which involve the use of adversarial training and training without inference-time threats and training-time threats.

It also suggested that the security policies should be added with continuous monitoring and retraining processes to adapt to the changing adversarial techniques. Since intelligent attack is dynamic in nature, there was no chance that the static defense mechanisms would withstand long term protection unless regularly updated and reviewed.

References

Ahmed, M., Alasad, Q., Yuan, J.-S., & Alawad, M. (2024). Re-evaluating deep learning attacks and defenses in cybersecurity systems. *Big Data and Cognitive Computing*, 8(12), 191. https://doi.org/10.3390/bdcc8120191

Alshahrani, E., Alghazzawi, D., & Alotaibi, R. (2022). Adversarial attacks against supervised machine learning-based network intrusion detection systems. *PLOS ONE*, 17(10).

Alsulaimawi, Z. (2024). *Securing federated learning with control-flow attestation: A novel framework for enhanced integrity and resilience against adversarial attacks.*

Alzaidy, S., & Binsalleeh, H. (2024). Adversarial attacks with defense mechanisms on convolutional neural networks and recurrent neural networks for malware classification. *Applied Sciences, 14*(4), 1673.

Anika, T. (2023). A systematic literature review on untargeted model poisoning attacks and defense mechanisms in federated learning. *Systematic Literature Review and Meta-Analysis Journal, 3*(4), 117–126.

B G, K., & Vairam, T. (2025). Adversarial attacks and defense mechanisms on machine learning models for cybersecurity applications. *IJRASET*.

Carroll, C. (2024). Auto encoder-based defense mechanism against popular adversarial attacks in deep learning. *PLOS ONE*.

Cheng, X., Zheng, M., Zhu, S., & Dong, Y. (2025). MISLEADER: Defending against model extraction with ensembles of distilled models. *arXiv*. https://doi.org/10.48550/arXiv.2506.02362

Clement, T., Gbaja, C., & Onayemi, H. (2025). Adversarial machine learning: Defense mechanisms against poisoning attacks in cybersecurity models. *International Journal of Engineering and Computer Science*, 14(06), 27286-27308.

Garba, M. (2025). *Adversarial attacks and defense mechanisms in privacy-preserving machine learning*. SSRN.

Harbi, Y., Medani, K., Gherbi, C., & Aliouat, Z. (2024). Roadmap of adversarial machine learning in IoT-enabled security systems. *Sensors*, 24(16), 5150.

Jęhan, N., et al. (2025). Adversarial machine learning for cybersecurity defense: Detecting model evasion, poisoning attacks, and enhancing the robustness of AI systems. *Global Research Journal of Natural Science and Technology*.

Jęhan, N., Mustaqim Ansari, N., et al. (2025). Adversarial machine learning for cybersecurity defense: Detecting model evasion and poisoning attacks. *Global Research Journal of Natural Science and Technology*.

Jiang, W. (2024). A comprehensive defense framework against model extraction. *IEEE Transactions on Dependable and Secure Computing*.

Kaixiang Zhao, L., Li, L., Ding, K., Gong, N. Z., Zhao, Y., & Dong, Y. (2025). A systematic survey of model extraction attacks and defenses: State-of-the-art and perspectives. *arXiv*. https://doi.org/10.48550/arXiv.2508.15031

Khraisat, A., Alazab, A., Alazab, M., Jan, T., & Uddin, M. A. (2025). Securing federated learning: A defense strategy against targeted data poisoning attack. *Discover Internet of Things*.

Liang, C., Huang, J., Zhang, Z., & Zhang, S. (2024). Defending against model extraction attacks with out-of-distribution feature learning and decision boundary confusion. *Computers & Security*, 136, 103563.

Ma, R., Zhang, Y., Wang, J., Lu, W., & Luo, X. (2026). Advancements in adversarial example defense for deep learning models: A review. *Cybersecurity*.

Malik, J., Muthalagu, R., & Pawar, P. M. (2024). A systematic review of adversarial machine learning attacks, defensive controls, and technologies. *IEEE Access, 12*, 99382–99421.

Panpatil, A. Y., & Gaikwad, V. P. (2025). A method to study the robustness of ML models against adversarial attacks. *International Journal of Environmental Sciences*, 11(5).

Shayea, G. G., Mohammed Zabil, M. H., Abdulfattah Habeeb, M., & Albahri, A. S. (2025). Strategies for protection against adversarial attacks in AI models: An in-depth review. *Journal of Intelligent Systems*, 34(1).

Zhao, K., Li, L., Ding, K., Gong, N. Z., Zhao, Y., & Dong, Y. (2025). A systematic survey of model extraction attacks and defenses: State-of-the-art and perspectives. *arXiv*.

Zhao, Y., et al. (2025). Improvement of the robustness of deep learning models against adversarial attacks. *Applied and Computational Engineering*.