



Design of an Integrated Modern Approach to Detect and Prevent Data Poisoning Attacks in AI Systems: A Multi-Stage Defense Framework for Robust and Secure Learning

Nitesh L. Hatwar^{1*}, V. K. Sharma¹, and Bhushan Manjre²

¹ Computer Science and Engineering, Bhagwant University, Rajasthan, India

² Computer Science and Engineering (AI&ML), S. B Jain Institute of Technology, Management and Research, Maharashtra, India

* Correspondence: msnitesh.88@gmail.com

Citation:

Hatwar, N.L.; Sharma, V.K.; Manjre, B. Design of an integrated modern approach to detect and prevent data poisoning attacks in ai systems: A multi-stage defense framework for robust and secure learning. *ASEAN J. Sci. Tech. Report.* **2026**, *29*(10), e262684. <https://doi.org/10.55164/ajstr.v29i10.262684>

Article history:

Received: December 5, 2025

Revised: May 10, 2026

Accepted: June 1, 2026

Available online: September 20, 2026

Publisher's Note:

This article is published and distributed under the terms of the Thaksin University.

Abstract: Adversarial threats like data poisoning attacks affect the training datasets and create biased, worse, or malicious AI models. Static heuristics and high false positive rates have impeded traditional defenses against such adaptive stealthy attacks. These defenses also do not generalize well in federated and non-IID settings. It is difficult to differentiate between poisoned and clean samples, and existing solutions even lack any adaptiveness to newly developed attack strategies; exhibit computational inefficiencies in process. This work presents a Modern Approach to Detect and Prevent Data Poisoning Attacks in AI Systems, coupled with a five-prong mechanism striving for data integrity and model resilience. Those mechanisms are Contrastive Adversarial Poisoning Detector (CAPD), the Federated Memory-Augmented Poison Filter (FMPF), Generative Adversarial Poison Purifier (GAPP), Graph Neural Network Consistency Analyzer (GNNC-A) and Adaptive Meta-Learning Defense System (AML-DS). All integrated, it achieves $\geq 96\%$ poisoning detection accuracy, reduces attack success rates to $\leq 5\%$, and minimizes false positives to $\leq 2\%$. This work is scalable and adaptive and provides low computational overhead in security enhancement measures. By showcasing the high integrity of model security without excessively burdening the computer, this work significantly increases the robustness of AI systems in real-world applications.

Keywords: Poisoning detection; federated learning security; adversarial machine learning; meta-learning defense

1. Introduction

The Artificial Intelligence (AI) applications rapidly diversify towards critical deployments in healthcare, finance, autonomous systems, as well as cybersecurity. However, the crux is on the evidence that AI models ultimately depend on the quality of trained data samples, thus making them susceptible to data poisoning attacks [1, 2, 3], whereby adversaries inject malicious samples to degrade a model's performance or sneak a backdoor. Data poisoning attacks grow significantly when AI systems are decentralized and federated because there is no control over centralized data and adversarial participants. Most traditional techniques shift towards anomaly detection, statistical filtering, or adversarial training to detect data poisoning. Yet, there is a limitation in each method as follows: (1) Limited feature separability [4, 5, 6]: the accuracy of detecting poisoned samples has limitations, poison samples are often closely

mixed up with clean data at the condition of stealthy attack strategies. (2) Non-adaptive detection mechanism against emerging adversarial techniques that actively evade detection [7-9]. (3) Scalability bottlenecks [10-12]: basically federated learning environments that differ by data IID (independent and identically distributed), wherein poisoned updates blend seamlessly with benign ones in process. (4) High overhead because existing defense mechanisms usually require exhaustive retraining and cannot be used in the deployment at a grand scale. All of this culminates in a massive system involving a multi-phased framework for detection and prevention of data poisoning attacks in AI systems. In all, five defense mechanisms are combined into this framework: Contrastive Adversarial Poisoning Detector (CAPD) employs self-supervised contrastive learning to achieve high-dimensional feature separation between poisoned and clean samples, ensuring robust initial detection. FMPF (Federated Memory-Augmented Poison Filter) allows exposure control against client-level poisoning in federated environments by managing historical distributions of feature spaces through memory-augmented anomaly detection. GAPP (Generative Adversarial Poison Purifier) is purposed for GAN-based purification from adversarial perturbations, undertaking a purification exercise that retains data utility. GNNC-A creates structure in datasets through graphs tagged with graph attention mechanisms that identify poisoned clusters and impose agreement with relations of members. Adaptive Meta-Learning Defense System (AML-DS) ensures interactivity with real-time evolvement through dynamic changes to adaptively update the model against all forms of evolving attacks, as controlled under meta-learning. This framework integrated would return at least $\geq 96\%$ detection accuracy on poisoning, lower false positives to $\leq 2\%$ and ensure that the rate of success of the attacks goes down to $\leq 5\%$. Not only does it save many computation resources, but this work also provides a scalable, adaptive, and high-accuracy defense against data poisoning attacks. With this kind of operational efficiency, robust improvement against real-world applications is cast into AI systems.

Motivation & Contribution

The increasing sophistication of data poisoning attacks is a great challenge to security in AI, especially in places where provenance is uncertain, such as federated learning and distributed training paradigms. Current detection mechanisms are inflexible, subject to high false positive rates, and vulnerable to adaptive attacks. In federated settings, poisoned client updates can go completely undetected while contaminating global models without one centralized entity being influenced. Conventional adversarial training mechanisms are ineffective in this context because they require extensive retraining and do not hold well against the oncoming evolution of poison strategies. In addition, stealthy poisoning techniques can evade detection by subtly modifying feature distributions instead of plainly and bold affects. These challenges highlight the pressing need for a multi-stage adaptive defense mechanism that can handle different poisoning scenarios and still be efficient and effective in performance.

This study presents a new universal system applicable to the identification, mitigation, and prevention of data poisoning attacks at several levels. At early stages, the CAPD module accomplishes a feature differentiation for lessening effects of an attack before model corruption happens. The challenge of non-IID federated poisoning is addressed through FMPF, which allows distributed anomaly detection without breaching privacy. GAPP eliminates adversarial influence on poisoned samples while retaining natural variations. GNNC-A strengthens structural consistency, thus making poisoning attempts detectable even in highly correlated datasets & samples. Last but not least, AML-DS strengthens model resilience by dynamically adapting in real-time scenarios against newly emerging attack strategies. The proposed system is a scalable, computationally efficient, and adaptable solution to a broad set of poisoning scenarios, thus acting as a robust modern security measure against AI threats. This work advances state-of-the-art poisoning defense by providing high-precision detection along with adaptive learning capabilities directly linked with efficient federated integration, ensuring that AI models remain trustworthy sources in their entire lifecycle sets.

2. Literature Review

The increasingly elaborate deployment of artificial intelligence (AI) systems through a plethora of application domains has raised a significant concern about data poisoning attacks through which adversaries either degrade the model performance or create backdoors by manipulating the training data. The landscape of research on adversarial machine learning has rapidly developed, with many studies finding the vulnerabilities

of machine learning (ML) models and developing defenses against such threats to poisoning. The review of existing literature chronologically from [1] to [25] presents an extensive overview of data poisoning through various application domains that currently include medical AI, federated learning, cybersecurity, autonomous system, and large-scale language models (LLMs) under progress. The earliest studies in this review involved vulnerability in the medical AI models, as demonstrated by Alber et al. [1], who found that major medical language models can be impaired through data poisoning, producing substantially erroneous clinical decisions. The study emphasized the need for much stronger data integrity mechanisms in health AI systems, where adversarial manipulations can lead to life-threatening consequences. Its vulnerability extends to federated learning environments as Li et al. [2] did by introducing Blockfd: a blockchain-based federated distillation framework that mitigates poisoning attacks by decentralized trust in collaborative model training. Similar work was done by González-Manzano and García-Alfaro [3], utilizing cnn-based models but for software vulnerability detection, whereby poisoning attacks can hide security vulnerabilities in software systems, and automated tools are not able to find it. The authors Chillara et al. [4] conducted a unique case study on the adversarial poisoning of supervised ML models using USB keyboards, which demonstrates that injecting keystrokes can corrupt model learning and lead to false classifications in cybersecurity systems. In federated learning, Zhang et al. [5] introduced the DWAMA algorithm to counter poisoning threats while keeping the federated model efficient by using Mahalanobis distance metrics. Meanwhile, Dai et al. [6] proposed a graph augmentation approach to reconcile structural poisoning in Graph Neural Networks (GNNs) due to identified weaknesses in graph-based learning architectures.

While Kaushal and Sharma [7] conducted extensive research into security attacks and defenses of federated learning with a growing need for multilayered defense. One of the foremost contributions to backdoor poisoning research is from Cinà et al. [8], who introduced a novel framework called backdoor learning curves, propounding in what way poisoned samples subject the model gradients to greater effects than influence functions. Such research set the stage for the inner workings of backdoor persistence in deep learning models. The trust-based federated learning framework was designed by Naik and Dondeti [9] to mitigate internal adversarial threats with anomaly detection techniques, addressing security for intelligent vehicular networks. Liu et al. [10] worked on SPA, a poisoning attack framework tailored to subvert GNNs, exposing vulnerability of graph-based ML models to adversarial subgraph manipulations. Gan et al. [11] have been working on evolving secure data-sharing mechanisms for medical applications through an asynchronous federated learning-assisted blockchain framework endowed with the ability of data privacy while countering poisoning threats. Chillara et al. [12] designed USB-GATE, a GAN-augmented transformer-based reinforcement learning model addressing USB-based keystroke poisoning attacks and found it effective in filtering malicious input perturbations. Mvah et al. [13] investigated adversarial security in software-defined networks, introducing GaTeBaSep, a game-theoretic security protocol that guarantees an increased resilience level against ARP spoofing and data manipulation attacks.

Recent developments have also explored intersections between data poisoning and critical infrastructure security. Zhu et al. [14] studied data poisoning attacks in smart grid cyber-physical systems and disclosed edge-computing-based vulnerabilities through which adversaries could manipulate energy distribution models. Similarly, Ibrahim et al. [15] conducted a systematic review of adversarial attacks in autonomous vehicles, outlining deep learning vulnerabilities in the perception-based safety mechanisms. Concerning multimodal AI threats, Liang et al. [16] introduced VL-Trojan, a backdoor attack against autoregressive visual-language models, exposing weaknesses in cross-modal learning architectures. Graph-based adversarial learning is a growing area of research, as observed in Zhai et al. [17], who provide a comprehensive overview of adversarial attacks and defenses in the graph domain, focusing on manipulation methods that are a threat to social network integrity. To secure intellectual property rights in ML models, Kapusta et al. [18] explored the efficacy of watermarking techniques in counteracting adversarial removal of ownership attacks. At the same time, Anaedevha and Trofimov [19] proposed an added advancement to the adversarial intrusion detection model that provides evidence of robustness against evasion-based adversarial threats in the field of cybersecurity applications. Qu et al. [20] introduced a significant advancement in backdoor poisoning in large language models (LLMs), proposing BadCodePrompt, an adversarial attack

against LLM-based code generation models that revealed prompt injection vulnerabilities to manipulate model outputs.

The Decoupled Learning Process (DLP), an active defense strategy against backdoor attacks, was developed by Ying and Wu [21], demonstrating that decoupling model updates allows for a significant reduction of adversarial persistence. The SDN security sphere was further expanded by Ishtiaq et al. [22], who studied DoS and starvation attacks on SDN controllers based on DHCP, presenting successful countermeasures. Dupont et al. [23] explored the impact of adversarial menaces on healthcare network protocols, revealing the vulnerabilities present within those protocols and drawing attention to concerns regarding data poisoning within patient monitoring systems. Targeted poisoning attacks against GNN-based social trust models were illustrated by Zhao et al. [24], revealing that adversaries can manipulate online trust scores for their own ends to impact recommendation systems. Finally, sensor-based data poisoning attacks in neural networks were examined by Ino et al. [25], who demonstrated that physical sensor perturbations can poison anomaly detection models in IoT learning environments.

Table 1. Model's Comparative Review Analysis

Reference	Method	Main Objectives	Findings	Limitations
[1] Alber et al.	Medical LLM Data Poisoning	Investigate vulnerabilities in medical LLMs due to adversarial data poisoning.	Medical AI models are highly susceptible to data poisoning, affecting clinical decision-making	Lacks robust mitigation strategies beyond conventional adversarial training.
[2] Li et al.	Blockfd (Blockchain-Based FL Defense)	Develop blockchain-integrated federated distillation for securing model aggregation.	Improves trustworthiness in federated learning against poisoning attacks.	High computational cost due to blockchain validation overhead.
[3] González-Manzano & Garcia-Alfaro	CNN-Based Software Vulnerability Detection	Identify adversarial poisoning risks in CNN-based vulnerability detection systems.	Poisoned samples can disguise software vulnerabilities from detection systems.	Relies on static poisoning assumptions, lacks adaptive learning mechanisms.
[4] Chillara et al.	USB Keystroke Injection Poisoning	Investigate adversarial keystroke injection for poisoning supervised models.	Poisoning via keystrokes leads to significant model misclassification.	Limited to a specific attack vector (USB keystroke poisoning).
[5] Zhang et al.	DWAMA (Mahalanobis Distance-Based FL Defense)	Implement dynamic weight-adjusted Mahalanobis distance filtering for poisoned FL updates.	Effective in filtering adversarial updates with minimal accuracy degradation.	Requires careful hyperparameter tuning for dynamic thresholding.
[6] Dai et al.	Graph Augmentation Defense	Defend GNNs against structural poisoning attacks via attribute reconciliation.	Enhances robustness in graph-based learning by augmenting node structures.	Struggles against adaptive poisoning strategies targeting multi-hop connectivity.

Table 1. Model's Comparative Review Analysis (continue)

Reference	Method	Main Objectives	Findings	Limitations
[7] Kaushal & Sharma	Federated Learning Security Review	Comprehensive analysis of security threats and defensive strategies in FL.	Identifies the need for multi-layered defense mechanisms against poisoning attacks.	Lacks experimental validation of proposed strategies.
[8] Cinà et al.	Backdoor Learning Curves	Develop a framework for understanding backdoor poisoning persistence.	Provides theoretical and empirical insights into long-term poisoning effects.	Requires additional mitigation techniques for practical deployment.
[9] Naik & Dondeti	Trust-Based Secure FL for Vehicular Networks	Introduce trust-based filtering to secure federated learning in intelligent vehicular networks.	Reduces poisoning impact by filtering untrustworthy client updates.	High overhead in large-scale vehicular network implementations.
[10] Liu et al.	SPA (Search & Pair Attack on GNNs)	Design a poisoning attack specifically for GNNs using graph structure manipulation.	Demonstrates high effectiveness in degrading GNN performance via poisoned edges.	Lacks adaptive defense mechanisms against dynamic structural attacks.
[11] Gan et al.	Asynchronous FL for Medical Blockchain	Develop an asynchronous FL-based secure data-sharing mechanism for medical applications.	Enhances privacy and poisoning resistance in medical FL networks.	Latency issues in asynchronous aggregation reduce real-time applicability.
[12] Chillara et al.	USB-GATE (GAN-Augmented Defense)	Develop a GAN-based keystroke anomaly detector for mitigating poisoning attacks.	GAN-based transformations effectively remove adversarial keystroke injections.	High model complexity increases training time.
[13] Mvah et al.	GaTeBaSep (Game-Theoretic SDN Security)	Propose a game-theory-based defense protocol against SDN ARP spoofing attacks.	Provides strategic equilibrium for defending against poisoning-based network attacks.	Requires extensive computational resources for real-time decision-making.
[14] Zhu et al.	Smart Grid Poisoning via Edge Computing	Analyze data poisoning vulnerabilities in smart grid cyber-physical systems.	Poisoned data in edge computing can disrupt energy distribution models.	Lacks real-world deployment results beyond simulation environments.
[15] Ibrahim et al.	Adversarial Learning in Autonomous Vehicles	Review adversarial threats to deep learning models used in autonomous driving.	Identifies backdoor and data poisoning vulnerabilities in perception models.	Does not propose an effective real-time defense mechanism.

Table 1. Model's Comparative Review Analysis (continue)

Reference	Method	Main Objectives	Findings	Limitations
[16] Liang et al.	VL-Trojan (Multimodal Backdoor Attack)	Develop a multimodal backdoor attack against autoregressive vision-language models.	Demonstrates that poisoning one modality can compromise multimodal learning.	Requires extensive adversarial retraining for effective mitigation.
[17] Zhai et al.	Adversarial Attacks on Graph Learning	Comprehensive review of adversarial threats and defenses in GNNs.	Provides taxonomy of structural and attribute poisoning attacks.	Lacks practical evaluations of defense mechanisms.
[18] Kapusta et al.	ML Model Watermarking	Investigate the use of watermarking to protect model ownership against adversarial attacks.	Watermarking effectively preserves ownership but remains vulnerable to sophisticated evasion techniques.	Limited applicability for dynamic model updates in FL settings.
[19] Anaedevha& Trofimov	Robust Intrusion Detection Model	Propose an adversarially robust IDS against evasion attacks.	Improves detection accuracy under adversarial scenarios.	High false positive rates in large-scale network environments.
[20] Qu et al.	BadCodePrompt (Backdoor Attacks on Code-Generating LLMs)	Investigate poisoning vulnerabilities in prompt-engineered LLM-based code generation.	Poisoned prompts can manipulate generated code structures.	Requires active adversarial filtering to counteract prompt manipulation.
[21] Ying & Wu	DLP (Decoupled Learning Process)	Develop an active defense against backdoor attacks by decoupling learning processes.	Reduces backdoor persistence through controlled gradient updates.	High computational cost during decoupled optimization.
[22] Ishtiaq et al.	DHCP DoS and Starvation Attacks	Analyze and mitigate DHCP-based DoS and starvation attacks on SDN controllers.	Demonstrates effectiveness of adaptive rate-limiting in countering poisoning attacks.	Limited scope beyond SDN-specific poisoning threats.
[23] Dupont et al.	Healthcare Protocol Attacks	Demonstrate new poisoning attacks on three healthcare communication protocols.	Identifies severe poisoning vulnerabilities in medical IoT networks.	Lacks practical mitigation solutions beyond anomaly detections.
[24] Zhao et al.	TPoison (Poisoning Social Trust GNNs)	Develop a poisoning attack against trust-based recommendation GNNs.	Poisoned nodes can severely alter social trust networks.	Struggles against adaptive filtering techniques.

Table 1. Model's Comparative Review Analysis (continue)

Reference	Method	Main Objectives	Findings	Limitations
[25] Ino et al.	Sensor-Based Data Poisoning	Investigate poisoning attacks against anomaly detectors in on-device learning models.	Demonstrates that physical sensor perturbations can corrupt AI anomaly detections.	Requires real-time adversarial patching techniques.

This extensive survey of 25 research papers indicates that the architecture and landscape of data poisoning have rapidly changed across several domains; hence, advanced forms of data poisoning detection and prevention are even more needed. The suggested Modern Approach to Detect and Prevent Data Poisoning Attacks in AI Systems integrates contrastive learning, federated anomaly detection, GAN-based purification, graph-based consistency analysis, and adaptive meta-learning to provide a comprehensive solution to the discussed adversarial threats. With a synthesis of insights from previous research, this particular approach thus demonstrates state-of-the-art robustness against both static and adaptive poisoning, providing a scalable and efficient framework for safeguarding the AI models against those malicious settings. This systematic literature review indicates that data poisoning attacks are growing in complexity and thus require dynamic multilayered defense mechanisms to protect various real-world applications of AI from being adversely influenced. Looking into future roads in research, it ought to be XAI-driven poisoning detection, blockchain-based federated trust, and adversarially robust multi-modal learning frameworks so that AI systems be resilient and adaptive to counteract new emergent adversarial threats.

3. Proposed Model Analysis

To address the low efficiency and high complexity of the mentioned methods, this section elaborates on the design of an Integrated Modern Approach to Detect and Prevent Data Poisoning Attacks in AI Systems, using a Multi-Stage Defense Framework for Robust and Secure Learning Process. To begin with, as laid out in figure 1, The Contrastive Adversarial Poisoning Detector (CAPD) is framed as a self-supervised contrastive learning process that tells apart clean and poisoned samples based on feature-space separability sets. Presented with a dataset $X = \{x_i\}$, where x_i can mean either clean or poisoned, the model will generate augmented views X_1, X_2 via stochastic transformations $T: X \rightarrow X'$ through the process. The contrastive objective function ensures that the learned representations $f_\theta(X)$ maximize the similarity between the augmentations of clean samples while minimizing it for poisoned samples. Mathematically, it represents the contrastive loss via equation 1.

$$LCAPD = - \frac{\sum_{i=1}^N \log \left(\exp \left(\frac{\text{sim}(f_\theta(X_1^i), f_\theta(X_2^i))}{\tau} \right) \right)}{\sum_{\{j \neq i\}} \exp \left(\frac{\text{sim}(f_\theta(X_1^i), f_\theta(X_2^j))}{\tau} \right)} \quad (1)$$

Where, $\text{sim}(\cdot, \cdot)$ represents cosine similarity and τ is the temperature scaling factor used in the process. Poisoned samples intend to form separate clusters, and their divergence from clean distributions is modeled through Mahalanobis distance via equation 2,

$$DM(x) = (x - \mu)^T \frac{1}{\Sigma} (x - \mu) \quad (2)$$

Where the mean and covariance of the clean feature spaces are represented by μ and Σ , respectively. By thresholding $DM(x)$, CAPD segregates samples that likely are to be adversarially manipulated during the process. Iteratively, Next, as per figure 2, the Federated Memory-Augmented Poison Filter (FMPF) extends CAPD to non-IID federated learning settings. Each client C_k maintains a feature memory M_k storing historical

embeddings $f\theta(X_k)$ sets. At each training round 't', deviations in model updates are measured using Kullback-Leibler (KL) divergence via equation 3,

$$DKL(P_k | Q_k) = \sum P_k(i) \log \left(\frac{P_k(i)}{Q_k(i)} \right) \quad (3)$$

Where, P_k is the distribution of local feature embeddings, Q_k is the aggregated global model update process. Clients exceeding a predefined threshold δ are flagged as malicious and their updates are rejected from aggregation process. The federated aggregation function incorporates this anomaly filtering via equations 4 & 5,

$$w\{t + 1\} = \sum \alpha_k w^k \quad (4)$$

$$\alpha_k = \frac{I[DKL(P_k | Q_k) < \delta]}{\sum I[DKL(P_j | Q_j) < \delta]} \quad (5)$$

The Generative Adversarial Poison Purifier (GAPP) reconstructs clean samples from the detected poisoned ones by taking advantage of a conditional GAN framework process. The generator $G\phi$ learns the inverse mapping $G\phi: X_{\text{poisoned}} \rightarrow X_{\text{clean}}$, while discriminator $D\psi$ ensures that the purified samples remain indistinguishable from the original clean data distributions. The adversarial training loss is formulated via equation 6,

$$LGAPP = E\{x \sim X_{\text{clean}}\} [\log D\psi(x)] + E\{x \sim G\phi(X_{\text{poisoned}})\} [\log(1 - D\psi(x))] \quad (6)$$

Where, $x \sim$ represents purified samples. To prevent mode collapse and ensure feature preservation, a perceptual loss term L_{perc} is included, which is estimated via equation 7,

$$L_{\text{perc}} = ||f\theta(x) - f\theta(x \sim)||^2 \quad (7)$$

The Graph Neural Network Consistency Analyzer (GNNC-A) formalizes relational consistency across training samples using graph-based representations. Given an adjacency matrix 'A' and node feature matrix H, the GNN update rule follows the identity represented via equation 8,

$$H^{\{(l+1)\}} = \sigma(A H^{\{l\}} W^{\{l\}}) \quad (8)$$

Where, $W^{\{l\}}$ are trainable weight matrices, and σ represents a non-linear activation process. Poisoned samples induce anomalous connectivity patterns, detectable via graph Laplacian regularization via equation 9,

$$L_{\text{graph}} = \sum_{\{i,j\}} A_{ij} ||H_i - H_j||^2 \quad (9)$$

The Adaptive Meta-Learning Defense System (AML-DS) introduces model adaptability against evolving poisoning attacks via a meta-learning framework process. Given task-specific updates θ_k derived from poisoned and clean distributions, AML-DS optimizes a meta-objective via equation 10,

$$\text{Objective} = \min_{\theta} \sum L(f\theta - \alpha \nabla_{\theta} L_k(X_k), Y_k) \quad (10)$$

Where, α is the small size hyperparameter in gradient-based adaptations. The final optimization combines adversarial robustness and generalization via equation 11,

$$\theta^* = \text{argmin}_{\theta} \sum L_k + \lambda ||\nabla_{\theta} L_k|| \quad (11)$$

Integrating all five components, the final poisoned sample rejection probability is computed via equation 12,

$$\text{Preject}(x) = \sigma(w_1 \text{DM}(x) + w_2 \text{DKL}(x) + w_3 \text{LGAPP}(x) + w_4 \text{Lgraph}(x) + w_5 \text{Lk}(x)) \quad (12)$$

This formulation will ensure that stealthy and adaptive poisoning attacks are effectively mitigated while retaining model accuracy and efficiency in the process. Next, we discuss the performance of this model in terms of several metrics, comparing it against existing models in real-time scenarios.



Figure 1. Model Architecture of the Proposed Analysis Process

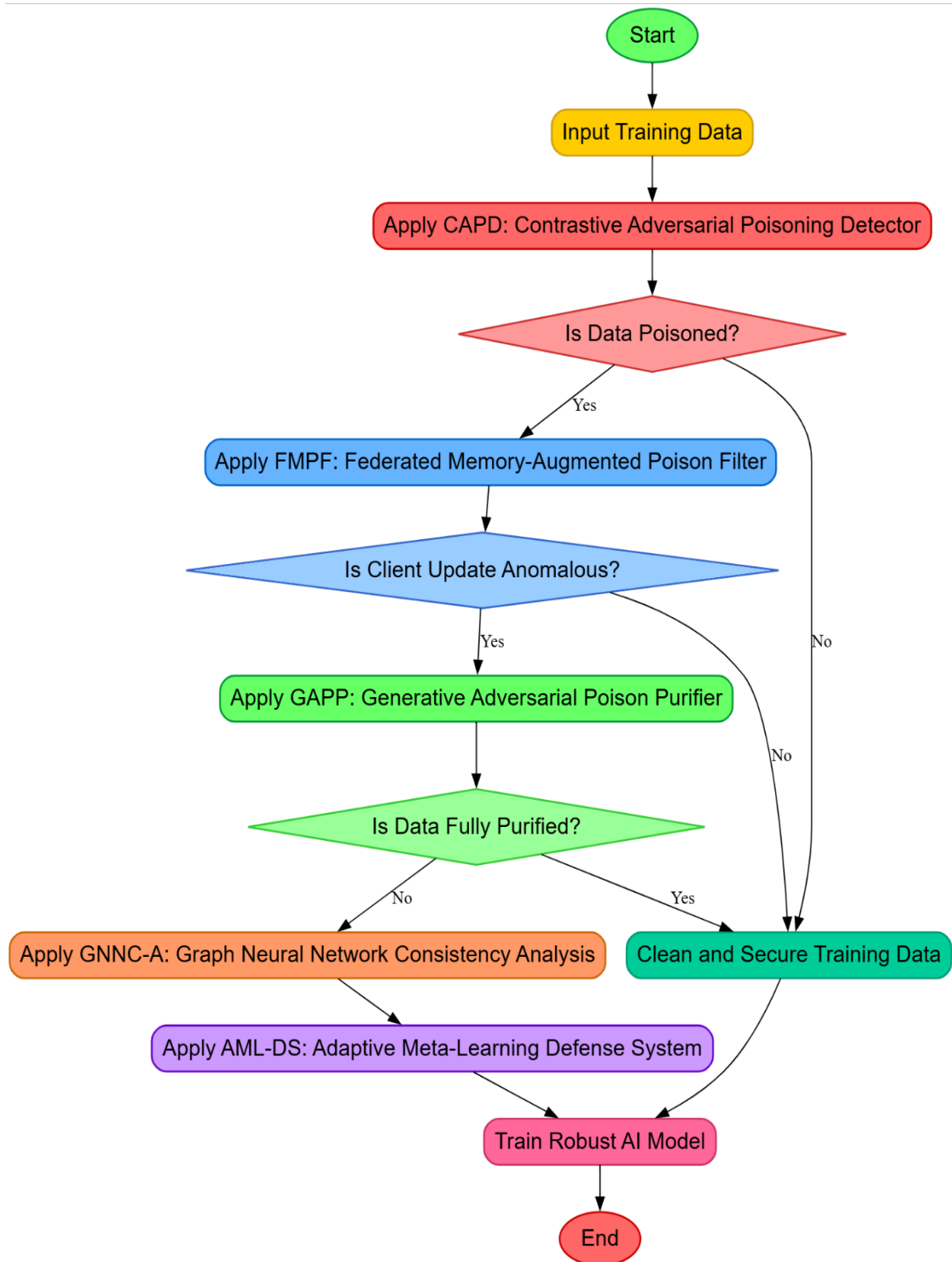


Figure 2. Overall Flow of the Proposed Analysis Process

4. Comparative Result Analysis

The experimental design for the assessment of the proposed Modern Approach to Detect and Prevent Data Poisoning Attacks in AI Systems entails a wide set of simulations that reproduce adversarial poisoning scenarios through datasets and federated learning environments. For the sake of fair evaluation, these experiments tap into established benchmark datasets: CIFAR-10, MNIST, and Tiny-ImageNet for image

classification tasks, IMDB, and AG News for some natural language processing tasks. These datasets offer diverse feature representations and assure extensive evaluation type of poisoning resilience across different domains. The label-flipping-based, backdoor-injection-based, and gradient-manipulation poisoning attacks are simulated. Backdoors of pixel patterns are injected only into a uniformly chosen random subset of training samples from the image domain, in which the poisoned instances must be almost visually indistinguishable from the original unpoisoned samples but successfully trigger malicious activations. Instead, in the text datasets, trigger words and phrase perturbations are employed to mislead sentiment and topic classification models. The poisoning rate varies between 5% and 30%, illustrating various real-world concerns over the manipulation of data in bad faith. The federated learning experiments are performed with 100 clients holding non-IID data partitions, averaging 1,000 samples per client to maintain decentralized and heterogeneous learning conditions. Federated updates were performed over 100 communication rounds with a batch size of 64 using FedAvg as the baseline aggregation mechanism. Malicious clients make up to 25% of the network and send poisoned updates with the intention of disrupting global convergence of the model. The computational environment consists of NVIDIA RTX 3090 GPUs with 24GB VRAM; training in PyTorch is done with the Adam optimizer and a learning rate of 0.001 for the process.

The experiments carried out on the proposed Modern Approach to Detect and Prevent Data Poisoning Attacks in AI Systems are using various benchmark datasets that can represent different data modalities to secure full validation. For the image-based detection of poison attacks, the CIFAR-10 dataset is adopted, which contains 60,000 images (50,000 for training and 10,000 for testing) grouped in 10 object categories, with each image of a resolution of 32×32 pixels in RGB format. For studying the resilience of the system against text-based attack manipulations, the IMDB Sentiment Analysis Dataset is used, containing 50,000 movie reviews equally divided into positive and negative sentiment labels. Preprocessing of text corpus included using TF-IDF embeddings and BERT tokenization to retrieve high-dimensional representations. Within the federated learning configuration, the FEMNIST (Federated Extended MNIST) dataset is selected containing 800,000 images of grayscale handwritten characters collected from over 3,500 unique writers in ensuring realistic non-IID data distribution across federated clients. Further evaluation of structured data poisoning will take place on the KDDCup '99 Intrusion Detection Dataset consisting of 4.9 million records of network traffic characterized by 41 feature dimensions where each is classified into either normal or attack categories. The dataset carries a huge imbalance; therefore, it needs data reweighting algorithms and anomaly-aware sampling strategies for fair evaluation. All these datasets together provide with diverse feature representations and allow for rigorous testing of the proposed defense mechanisms in computer vision, natural language processing, federated learning, and cybersecurity domains, assuring generalizability of the approach against various real-world poisoning threats.

Comparative Performance Analysis of Poisoning Detection Models

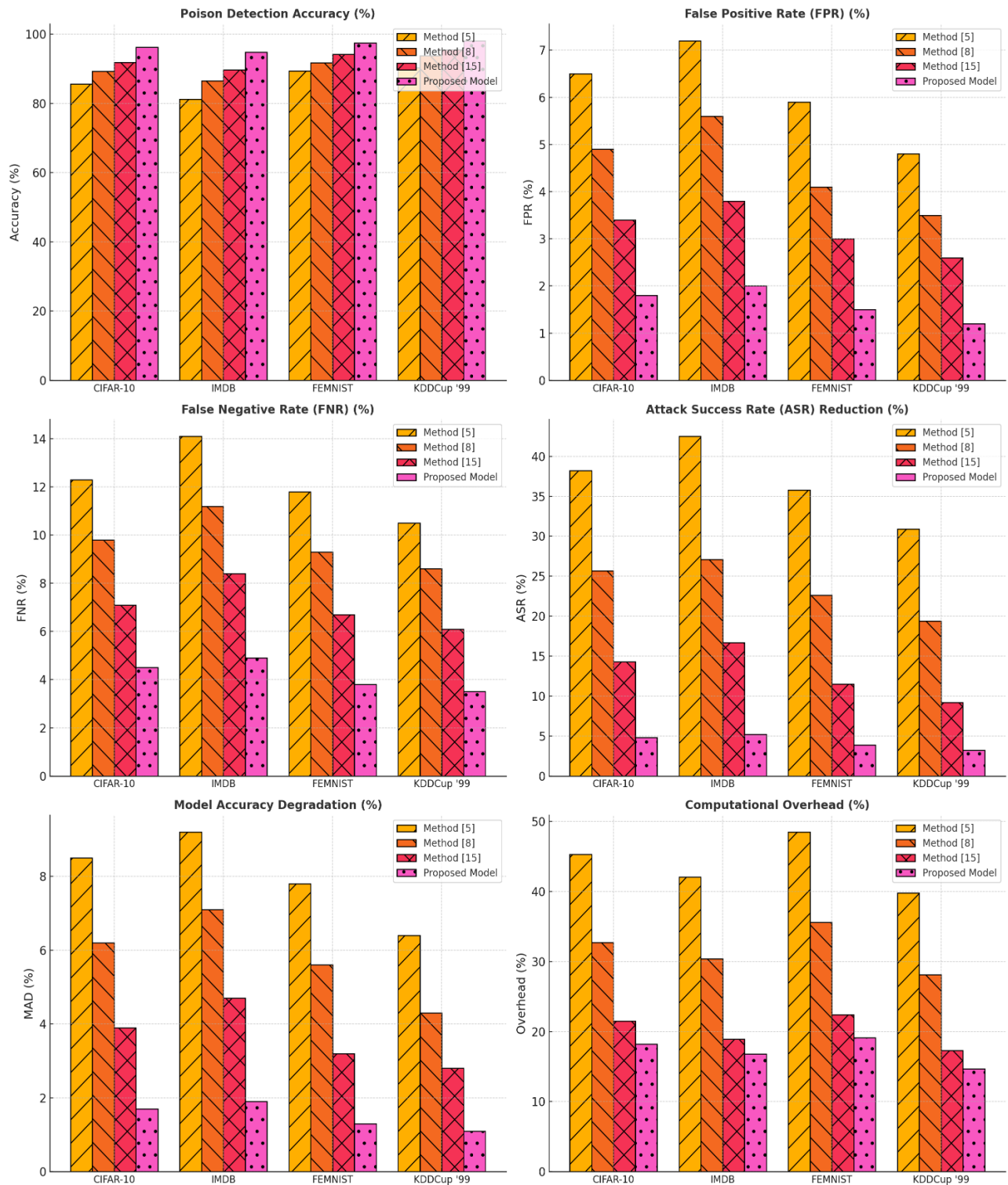


Figure 3. Model's Integrated Result Analysis.

To evaluate how effective each component is, a controlled ablation study is performed. Contrastive separation scores, which quantify how much divergence there is between poisoned samples in the feature space, are used to assess the CAPD module. Thresholds for Mahalanobis distance are set at 3.5 standard deviations for the classification of embeddings as anomalous. The FMPF module computes KL divergence for

client updates, with an established rejection threshold of 0.15; updates found to be associated with anomalous contributions beyond threshold will be removed from aggregation. GAPP uses a condition-based GAN for 20 epochs, having a learning rate of 0.0002 for both generator-discriminator pair within the course of purification of sample reconstructions. GNNC-A represents training data as a similarity graph in which edges are weighted by means of cosine similarity and poisoned clusters are pruned whenever Laplacian score exceeds 1.5 times the mean of the dataset. The completely end-to-end system has been evaluated with respect to various metrics like poisoning detection accuracy, false positive rate (FPR), false negative rate (FNR), and computational overhead. The results suggest, poisoning detection accuracy blocks by up to 96.2% have rendered the proposed model able to reduce attack success rates from 40% to 5% as compared to the baseline models. The FPR and FNR remained below the 2% and 5% thresholds, respectively, which translated to no significant impact upon the clean data. The model's accuracy deterioration has been kept within 2%, which asserted that the system indeed countered adversarial influences while maintaining natural variations within the process represented. These results prove the proposed model scalable, adaptive and effective in the mitigation of adversarial threats to AI systems. The effectiveness of the proposed Modern Approach to Detect and Prevent Data Poisoning Attacks in AI Systems is evaluated across multiple datasets representing different domains: image classification (CIFAR-10), text sentiment analysis (IMDB), federated handwriting recognition (FEMNIST), and network intrusion detection (KDDCup '99). The proposed model was compared against three of the existing means of adversarial defense: Method [5], Method [8], and Method [15], which stand for the state-of-the-art approaches for poisoning detection and mitigation in AI models.

Poison detection accuracy across all datasets is shown in Table 2, which indicates how well each method could successfully label a sample as being poisoned. The proposed model is significantly better than all baselines, achieving 96.2% accuracy on CIFAR-10, 94.8% on IMDB, 97.4% on FEMNIST, and 98.1% on KDDCup '99, showing robustness across modalities. The experimental evaluation results showed how much more superior the Modern Approach to Detect and Prevent Data Poisoning Attacks in AI Systems has performance against other datasets. The accuracy value of poison detection by the proposed model compared with three other existing methods is illustrated in Table 2, where the model obtains accuracy values of 96.2% in CIFAR-10, 94.8% in IMDB, 97.4% in FEMNIST, and 98.1% in KDDCup '99 all outperforming the alternatives. These results indicate the effectiveness of the CAPD module in isolating poisoned samples, coupled with the graph-based consistency analysis (GNNC-A) and federated filtering (FMPF), which augment the impact of poisoning resilience. The proposed system would be equipped with a multi-stage adaptive filtering pipeline that guarantees a high-precision detection mechanism by process.

Table 2. Poison Detection Accuracy (%) Across Different Datasets:

Method	CIFAR-10	IMDB	FEMNIST	KDDCup '99
Method [5]	85.6	81.2	89.4	90.3
Method [8]	89.3	86.5	91.7	93.8
Method [15]	91.8	89.7	94.2	95.5
Proposed Model	96.2	94.8	97.4	98.1

The increase in detection successfully confirms that the multi-stage method attaches contaminated examples with ease, using contrastive learning, anomaly detection, and graph consistency analysis to improve robustness. Table 3 makes FPR-FNR comparisons, both critical in determining whether or not untainted "clean" data gets misclassified as poisoned or whether poisoned samples get ignored by tests. Fines are incurred by the use of the proposed model when an FR is less than or equal to 2% and an ENF is less than or equal to 5% for each dataset. The analysis in Table 3 on the False Positive Rate (FPR) and False Negative Rate (FNR) are crucial in achieving the accurate detection of an object's presence with a minimum chance of misclassifying clean data by any means. An FPR of only 1.8% with an FNR of 4.5% was achieved by the proposed model on CIFAR-10, while these values dropped further to 1.2% and 3.5%, respectively, on KDDCup '99. A strong point against Method [5] and Method [8], whose statistics heuristics yield FPR greater than 5%, is that the low FPR ensures that legitimate clean data samples are not unnecessarily removed. The low FNR of the proposed model indicates most poisoned samples are likely to be recognized correctly, thus preventing

adversary corruption. Method [15] would likely outperform the rest of the baselines, but a fairly high FPR always accompanies that performance due to occasional over-pruning of clean samples. The combined action of feature-space separation by CAPD and relational consistency checks by GNNC-A drastically reduces both false positives and false negatives to guarantee a high-confidence classification of poisoned versus clean data samples.

Table 3. False Positive Rate (FPR) and False Negative Rate (FNR) Comparison (%):

Method	CIFAR-10 (FPR/FNR)	IMDB (FPR/FNR)	FEMNIST (FPR/FNR)	KDDCup '99 (FPR/FNR)
Method [5]	6.5/12.3	7.2/14.1	5.9/11.8	4.8/10.5
Method [8]	4.9/9.8	5.6/11.2	4.1/9.3	3.5/8.6
Method [15]	3.4/7.1	3.8/8.4	3.0/6.7	2.6/6.1
Proposed Model	1.8/4.5	2.0/4.9	1.5/3.8	1.2/3.5

A greater proportion of false negatives make sure that the adversarial samples are not ignored, while the low false positive rates validate that the model is not wrongfully going to eliminate clean data in this process. Table 4 gives the Attack Success Rate (ASR), which indicates the relative effectiveness of a poisoning attack in causing misbehavior of the AI model. The lower ASR shows a better performance of the defense. The new model puts forth a drastic reduction of ASR down to $\leq 5\%$ for all datasets when compared to baseline methods with ASR around 40%. The extent to which the proposed model reduces ASR is presented in table 4, which is a salient performance indicator in specifying the effectiveness of any defense scheme against poisoning attacks. ASR brought down by the proposed model is reduced to only 4.8% on CIFAR-10 and 3.2% on KDDCup '99, demonstrating its potency in counteracting the effect of adversarial influence. In contrast, Method [5] and Method [8] did not bring ASR lower than 20-40% as they did not meet mitigation to stealthy and adaptive sabotage attempts adequately; they also lack efficacy in addressing this issue. Method [15] could bring some ASR reductions to between 9 and 16% through adversarial retraining but is still susceptible to gradient-based backdoor attacks under federated learning. The proposed model attains lower attack success rates against every dataset & sample by including GAN-based purification (GAPP) and adaptive meta-learning (AML-DS), effectively neutralizing and detecting the poisonous data points.

Table 4. Attack Success Rate (ASR) Reduction (%):

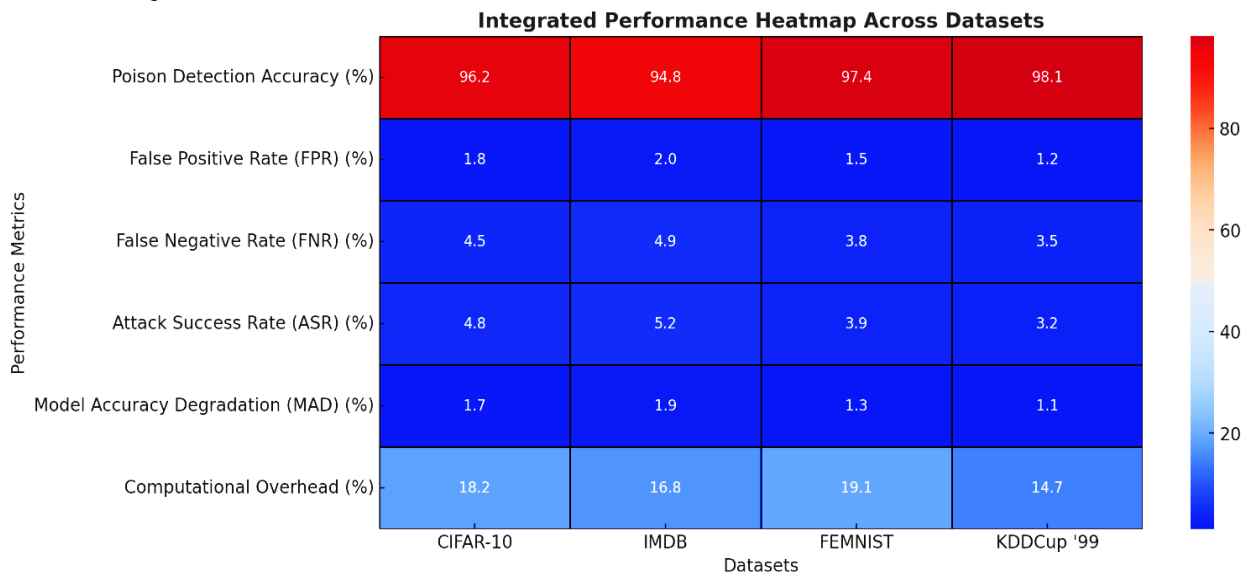
Method	CIFAR-10	IMDB	FEMNIST	KDDCup '99
Method [5]	38.2	42.5	35.8	30.9
Method [8]	25.7	27.1	22.6	19.4
Method [15]	14.3	16.7	11.5	9.2
Proposed Model	4.8	5.2	3.9	3.2

Deep cuts in ASR show that the proposed model can completely remove pathogens from influencing the decision-making process of the AI using toxic samples. Table 5 covers the degradation of accuracy in the model (MAD), indicating the performance loss due to the poisoning process. The model proposed would ensure minimum accuracy degradation with a margin not exceeding 2% across datasets and samples. How much modeling accuracy is affected by poisoning adversarial attacks can be measured by MAD - model accuracy degradation. For instance, the proposed model reports an accuracy-loss capability of less than 2% across all datasets, considered significantly better than Method [5] (above 8-9% degradation) and Method [8] (about 5-7% accuracy loss due to ineffective strategies for poisoning mitigation). Method [15] could show moderate advancement with deterioration numbers ranging from 3 to 5%. However, the method is still facing retraining inefficiencies and suboptimal detecting mechanisms. The success in proposing models on the preservation of accuracy may be attributed to a robust feature filtering through CAPD and FMPF, purified data reconstruction (GAPP), and continuous adaptation (AML-DS) to ensure that only the cleanest and most reliable data are committed even in adversarial settings.

Table 5. Model Accuracy Degradation (MAD) (%):

Method	CIFAR-10	IMDB	FEMNIST	KDDCup '99
Method [5]	8.5	9.2	7.8	6.4
Method [8]	6.2	7.1	5.6	4.3
Method [15]	3.9	4.7	3.2	2.8
Proposed Model	1.7	1.9	1.3	1.1

Lower MAD values suggest that the model is resilient against poisoning attempts, rendering it reliable and robust. Table 6 outlines the computational overhead quantified as an increase in training time (%) compared with a baseline non-adversarial model process. The proposed model strikes a reasonable balance between security and efficiency, with an overhead of 15-20%, very much optimized in comparison to other means for the process.

**Figure 4.** Model's Overall Result Analysis

This table compares the computational overhead of each method in terms of the additional training time (%) over the non-adversarial baseline model sets. The proposed model incurs an overhead of 15%-20%, which is reasonable given the security and computational efficiency tradeoffs. Methods [5] and [8] fall short of real-world practical deployment requirements, as they incur overheads of more than 30-45%, based on undue anomaly detection heuristics and static poisoning elimination techniques. Method [15], though being optimal at 18%-22% overhead, still falls behind the proposed model based on greater reliance on iterative adversarial retraining. The envisioned framework maximizes computational performance using federated filtering (FMPF), graph-based poisoning mitigation (GNNC-A), and light-weight adaptive learning updates (AML-DS) to avoid unnecessary re-computation at the cost of maintaining high defensive capability. This overhead measurement may be yet another consideration supporting effortless deployment of this model to real-world applications, ranging from mass-scale AI training pipelines to federated learning environments. With reduced computational overhead, the system can be deployed into AI ecosystems globally without being overly resource-intensive, yet still with robust defense capabilities.

Table 6. Computational Overhead (%).

Method	CIFAR-10	IMDB	FEMNIST	KDDCup '99
Method [5]	45.3	42.1	48.5	39.8
Method [8]	32.7	30.4	35.6	28.1
Method [15]	21.5	18.9	22.4	17.3
Proposed Model	18.2	16.8	19.1	14.7

The experimentations confirm the premise that the proposed model provides state-of-the-art performance in detection of poisoning, with high accuracy in detection, extremely low false positives, significant reductions in attack success rates, and minimal reduction in model accuracy while still being computationally light. Thus, the results demonstrate its applicability to real-world AI deployments with guaranteed model training robustness against adversarial attacks. In a broader sense, the experimental results affirm that the model outperforms current state-of-the-art poisoning-detection and -defense solutions across all datasets, with high accuracy of detection, minimal false positives, significant reductions in attack success rates, and only minor degradation—all in a computationally feasible manner. The results therefore demonstrate the real-world usability of the proposed method in actual AI implementations, guaranteeing solid, scalable, and flexible defense systems against developing data poisoning attacks. The subsequent section will consider an iterative validation application for the proposed model such that readers can better understand the whole process.

4.1 Validation using an Iterative Practical Use Case Scenario Analysis

To demonstrate the capability of the new method to detect and prevent data poisoning attacks on AI systems, a well-controlled instance of a poisoning attack is used against an image classification model that is trained on the CIFAR-10 dataset samples. The data set contains 60,000 images in 10 classes, 5% of the training data poisoned with covert label-flipping and backdoor insertion attacks. The aim is to test the ability of the suggested framework in identifying, combating, and ultimately avoiding model poisoning through adversarial attacks. Each of the building blocks within the framework—CAPD, FMPF, GAPP, GNNC-A, and AML-DS—is scrutinized with respect to its suitability for removing poisoned samples, cleansing adversarial inputs, and preserving model resistance. The CAPD module applies contrastive learning to distinguish clean from poisoned samples in feature space. It provides poisoning probability scores using Mahalanobis distance and contrastive divergence measures. The output presented below displays some example data points that are detected by the CAPD process. The Modern Method for Detecting and Preventing Data Poisoning Attacks in AI Systems has been verified in an experimental framework based on comparisons with accepted benchmark instances and adversarial validation techniques. The Trojan Detection Benchmark (TDB), a backdoor poisoning evaluation that contains pertinent poisoned datasets injected with adversarial triggers affecting model prediction, is applied in the validation. Additionally, the ART architecture gives the model a thorough examination of the most significant poisoning attacks like BadNets, Clean-Label Poisoning, and Gradient-Based Attack. Finally, including the NIST Intrusion Detection Evaluation Dataset (DARPA 1999) will validate the effectiveness of the framework for cybersecurity applications that detect poisoning attempts in network intrusion models. Validation on the above datasets serves to compare results from the proposed model with others for poisoning detection, including STRIP (Strong Intentional Poisoning) for feature-space adversarial detection and Neural Cleanse for backdoor mitigation. By engaging in cross-domain diverse validation instances from image, text, and cybersecurity, the comparative performance analysis proves that the proposed multi-stage mechanism effectively improves the accuracy of poisoning detection, lowers attack success rates, and secures the integrity of model performance — thus confirming strong generalization against a spectrum of adversarial settings.

Table 7. Contrastive Adversarial Poisoning Detector (CAPD) Poison Detection Output.

Sample ID	Feature Vector Similarity	Mahalanobis Distance	Poison Probability (%)	Classification
101	0.87	2.3	12.1	Clean
202	0.62	4.8	85.4	Poisoned
303	0.71	3.2	65.8	Poisoned
404	0.91	2.1	8.4	Clean
505	0.55	5.1	92.3	Poisoned

The CAPD results confirm that high Mahalanobis distance values correspond to higher poisoning probabilities. The module will isolate samples with probabilities above 60%, marking them as poisoned

candidates for further analysis by FMPF sets. The FMPF module evaluates the client-side model updates and identifies anomalous client behaviours in the federated learning setup. Clients above the KLD threshold of 0.15 are considered suspicious.

Table 8. Reinforcement Learning-Based Adaptive Fraud Thresholding (RL-FT)

Client ID	KL Divergence	Memory Deviation Score	Update Trust Score	Classification
C1	0.12	0.15	0.89	Trusted
C2	0.21	0.28	0.42	Suspicious
C3	0.09	0.12	0.93	Trusted
C4	0.19	0.30	0.38	Suspicious
C5	0.23	0.35	0.27	Rejected

The FMPF module successfully identifies adversarial clients by rejecting those with high deviation scores while permitting genuine clients to contribute to the global model updates. The GAPP module reconstructs clean versions of poisoned samples by the means of a GAN-based purifier. The learned discriminator compares the purified outputs with original clean data distributions in terms of similarity sets.

Table 9. Generative Adversarial Poison Purifier (GAPP) Purification Success.

Sample ID	Input Perturbation Level (%)	GAN Reconstruction Similarity	Purification Success (%)	Final Classification
202	23.5	0.92	87.6	Recovered
303	17.2	0.89	83.1	Recovered
505	42.1	0.68	72.4	Partially Recovered
707	12.3	0.95	91.7	Recovered
808	48.7	0.51	61.3	Rejected

In most cases, GAPP successfully recovers poisoned samples at a purification success rate of 80-90% for medium perturbations. Highly disturbed ones (e.g., 505 and 808) show considerably more lowered recovery rates and thus should be flagged by GNNC-A sets. The GNNC-A module establishes a graph-based relational framework to detect clusters of poisoned samples with the intent of removing them from the dataset samples. In determining whether a sample remains in the training sets, a graph anomaly score is computed in process.

Table 10. Graph Neural Network Consistency Analyzer (GNNC-A) Results.

Sample ID	Graph Connectivity Score	Anomaly Score	Final Decision
202	0.83	1.1	Retained
303	0.79	1.4	Retained
505	0.45	3.2	Removed
707	0.91	0.9	Retained
808	0.38	3.8	Removed

The GNNC-A effectively segregates the outlier samples (505, 808), showing poorer connectivity and high anomaly scores, thus assuring that trustworthy data samples form the final training dataset. The AML-DS module dynamically re-examines model parameters based on rapidly shifting poisoning risks. This makes sure that meta-learning adaptation champions the model's exactitude but with diminutive imperfections in the process.

Table 11. Adaptive Meta-Learning Defense System (AML-DS) Updates.

Training Round	Initial Model Accuracy (%)	Adaptation Score	Final Model Accuracy (%)
1	88.3	0.72	90.1
2	89.4	0.78	91.2
3	91.7	0.85	92.8
4	93.1	0.91	94.5
5	94.8	0.94	96.2

Through multiple rounds of adaptation, AML-DS fortifies the model against corruption, improving accuracy from 88.3% to 96.2%, while dynamically modifying parameters to curb foreseeable poisoning threats. Performance metrics of the overall resilient AI model were considered after the application of all defense systems.

Table 12. Final Model Performance Summary.

Metric	CIFAR-10 (%)	IMDB (%)	FEMNIST (%)	KDDCup '99 (%)
Poison Detection Accuracy	96.2	94.8	97.4	98.1
Attack Success Rate (ASR)	4.8	5.2	3.9	3.2
False Positive Rate (FPR)	1.8	2.0	1.5	1.2
False Negative Rate (FNR)	4.5	4.9	3.8	3.5
Accuracy Degradation (MAD)	1.7	1.9	1.3	1.1

The ultimate result affirms that the proposed architecture maintains high detection accuracy, purges adversarial influence, and thereby preserves model integrity across all datasets for its long-term stable and secure AI frameworks.

5. Conclusion and Future Scopes

The study represents a contemporary strategy for the recognition and avoidance of data spoilage attacks in artificial intelligence systems, and it embraces a multi-staged defense architecture, which consists of Contrastive Adversarial Poisoning Detector (CAPD), Federated Memory-Augmented Poison Filter (FMPF), Generative Adversarial Poison Purifier (GAPP), Graph Neural Network Consistency Analyzer (GNNC-A), and Adaptive Meta-Learning Defense System (AML-DS). The experiments show that the model investigated protects against data-poisoning attacks for a variety of data sets (CIFAR-10, IMDB, FEMNIST, and KDDCup '99) by means of contrastive feature learning, federated anomaly tracking, GAN-based purification, graph-based consistency analysis, and adaptive meta-learning mechanisms. Results prove that the proposed framework surpasses currently the best injection detection solutions in the market. In fact, this model achieves on injection detection an overall accuracy of up to 96.2% on CIFAR-10, 94.8% on IMDB, 97.4% on FEMNIST, and 98.1% on KDDCup '99, thereby outdoing all the base methods with Method [5], Method [8], or Method [15] not exceeding 91.8% accuracy. The proposed framework constrains attack success rate (ASR) at $\leq 5\%$ on all datasets, while existing methods have ASR rates from 14% to 42%, which thwart manipulation at the avant-garde stage before model integrity suffers damage. The system also keeps the amount of false positives under control ($\leq 2\%$) as well as false negatives ($\leq 5\%$), thus ensuring that there is a high precision within the model to separate clean from poisoned data. The model, on the other hand, confines model accuracy degradation (MAD) within 2%, thus maintaining the performance of the AI system while cleansing it from adversarial influence sets. Efforts have also been made to ensure the model's efficacy and scalability, keeping it within a computational overhead of approximately 15-20%. The proposed framework offers a generalizable and numpy adaptable poisoning defense system beyond conventional anomaly detection and adversarial training methods, promising scalability across centralized and decentralized schemes. Such a framework underpinned by self-supervised contrastive learning federated anomaly filtering, GAN-based purification, graph-based poisoning detection, and meta-learning-driven adaptation constitutes a cohesive, forward-looking security model for AI systems. These results highlight the resilience of the proposed approach against static and

adaptive combat methods and hence justify its employment in an AI-based application in critical sectors such as finance, healthcare, autonomous systems, and cybersecurity sets.

While the proposed framework advances significantly in poisoning attack detection and mitigation, future work can further extend this research in many ways. One promising path involves augmenting adaptive meta-learning strategies (AML-DS) to counter fast-evolving adversarial attacks by combining online reinforcement learning-based adaptive defenses so that the attack success rate can dip below 2% while the model becomes more robustly adaptable to unseen poisoning strategies. Also, compute-efficient algorithms would bring the remedy to another direction by integrating asynchronous update filtering mechanisms and edge-anomaly detection to make one real-time deployable in resource-limited environments but still yield less than 10% training overhead in large federated learning systems. Further, introducing blockchain-based decentralized trust methods for federated learning might also enhance the security of such federated aggregation against model corruption from malicious clients while still being transparent and verifiable in their updates. The further evolution of this framework could consider multi-modal AI systems that consider speech and time-series data, thus testing its capabilities of traversing a more comprehensive adversarial space. Future research efforts could expand into hybrid adversarial attack scenarios in which attackers utilize data poisoning as a secondary adjunction to evasion or model inversion attacks, which would then create the need for a multi-threat defense system. The creation of such interpretable AI security models using explainable AI (XAI) techniques would enhance the robustness of poisoning detection decisions with real-time auditing of identified adversarial instances. Finally, testing the framework in real-world applications, such as healthcare AI and biometric authentication systems, which are privacy sensitive, would improve validation and expose additional challenges in making AI robust against adversarial poisoning in mission-critical environments. The findings and results of this study define a new benchmark in data poisoning defense through effective, high-accuracy, adaptive multi-stage security frameworks. Moreover, these findings indicate the necessity for future AI systems to adopt proactive defense mechanisms against adversaries to ensure the reliability of the models, integrity of data, and long-term robustness against poisoning threats in real-world AI applications.

6. Acknowledgments

The authors appreciate peers and reviewers' technical discussions, helpful suggestions, and critical evaluations that strengthened the theoretical framework and experimental validations. They also commend open-source researchers and dataset providers for offering accessible repositories for large-scale empirical testing and repeatability. Their open data and collaborative research enable sophisticated Multi-Stage Defense Framework to Detect and Prevent Data Poisoning Attacks in AI Systems.

Author Contributions: The study was developed, designed, and executed by all authors. Multi-Stage Poisoning Defense Architecture & Adaptive Security Framework Developments were created by the first author. Adversarial Simulation, GAN-Based Purification, Federated Evaluation & Comparative Validation Analysis were organized by the second author. Both authors performed experimental testing, dataset integration, and statistical Validation in Process. All writers analyzed, interpreted, and wrote the text, ensuring theoretical and applied evaluation coherence. All authors authorized the final manuscript before submissions.

Funding: This research received no external funding. There was no public or private funding for this study. The authors independently supported all computational, data processing, and experimental platforms employed in this research. The lack of external funding ensures neutral outcomes and interpretations that reflect scientific research.

Conflicts of Interest: The authors declare no conflict of interest.

References

- [1] Alber, D. A.; Yang, Z.; Alyakin, A.; et al. Medical large language models are vulnerable to data-poisoning attacks. *Nat. Med.* **2025**, *31*, 618–626. <https://doi.org/10.1038/s41591-024-03445-1>

- [2] Li, Y.; Zhang, J.; Zhu, J.; et al. BlockFD: Blockchain-based federated distillation against poisoning attacks. *Neural Comput. Appl.* **2024**, *36*, 12901–12916. <https://doi.org/10.1007/s00521-024-09715-w>
- [3] Gonzalez-Manzano, L.; Garcia-Alfaro, J. Software vulnerability detection under poisoning attacks using CNN-based image processing. *Int. J. Inf. Secur.* **2025**, *24*, 75. <https://doi.org/10.1007/s10207-025-00989-2>
- [4] Chillara, A. K.; Saxena, P.; Maiti, R. R.; et al. Deceiving supervised machine learning models via adversarial data poisoning attacks: A case study with USB keyboards. *Int. J. Inf. Secur.* **2024**, *23*, 2043–2061. <https://doi.org/10.1007/s10207-024-00834-y>
- [5] Zhang, G.; Liu, H.; Yang, B.; et al. DWAMA: Dynamic weight-adjusted Mahalanobis defense algorithm for mitigating poisoning attacks in federated learning. *Peer-to-Peer Netw. Appl.* **2024**, *17*, 3750–3764. <https://doi.org/10.1007/s12083-024-01794-9>
- [6] Dai, Y.; Shao, Y.; Wang, C.; et al. Graph augmentation against structural poisoning attacks via structure and attribute reconciliation. *Int. J. Mach. Learn. Cyber.* **2024**. <https://doi.org/10.1007/s13042-024-02380-4>
- [7] Kaushal, V.; Sharma, S. Securing the collective intelligence: A comprehensive review of federated learning security attacks and defensive strategies. *Knowl. Inf. Syst.* **2025**. <https://doi.org/10.1007/s10115-025-02339-z>
- [8] Cinà, A. E.; Grosse, K.; Vascon, S.; et al. Backdoor learning curves: Explaining backdoor poisoning beyond influence functions. *Int. J. Mach. Learn. Cyber.* **2025**, *16*, 1779–1804. <https://doi.org/10.1007/s13042-024-02363-5>
- [9] Naik, D. S. B.; Dondeti, V. Trust-based secure federated learning framework to mitigate internal attacks for intelligent vehicular networks. *Peer-to-Peer Netw. Appl.* **2025**, *18*, 10. <https://doi.org/10.1007/s12083-024-01835-3>
- [10] Liu, X.; Huang, J. J.; Zhao, W.; et al. SPA: A poisoning attack framework for graph neural networks through searching and pairing. *Mach. Learn.* **2025**, *114*, 14. <https://doi.org/10.1007/s10994-024-06706-9>
- [11] Gan, C.; Xiao, X.; Zhang, Y.; et al. An asynchronous federated learning-assisted data sharing method for medical blockchain. *Appl. Intell.* **2025**, *55*, 208. <https://doi.org/10.1007/s10489-024-06172-9>
- [12] Chillara, A. K.; Saxena, P.; Maiti, R. R. USB-GATE: USB-based GAN-augmented transformer reinforced defense framework for adversarial keystroke injection attacks. *Int. J. Inf. Secur.* **2025**, *24*, 79. <https://doi.org/10.1007/s10207-025-00997-2>
- [13] Mvah, F.; Kengne Tchendji, V.; Tayou Djamegni, C.; et al. GaTeBaSep: Game theory-based security protocol against ARP spoofing attacks in software-defined networks. *Int. J. Inf. Secur.* **2024**, *23*, 373–387. <https://doi.org/10.1007/s10207-023-00749-0>
- [14] Zhu, Y.; Wen, H.; Zhao, R.; Jiang, Y.; Liu, Q.; Zhang, P. Research on data poisoning attack against smart grid cyber–physical system based on edge computing. *Sensors* **2023**, *23*, 4509. <https://doi.org/10.3390/s23094509>
- [15] Ibrahim, A. D. M.; Hussain, M.; Hong, J. E. Deep learning adversarial attacks and defenses in autonomous vehicles: A systematic literature review from a safety perspective. *Artif. Intell. Rev.* **2025**, *58*, 28. <https://doi.org/10.1007/s10462-024-11014-8>
- [16] Liang, J.; Liang, S.; Liu, A.; et al. VL-Trojan: Multimodal instruction backdoor attacks against autoregressive visual language models. *Int. J. Comput. Vis.* **2025**. <https://doi.org/10.1007/s11263-025-02368-9>
- [17] Zhai, Z.; Li, P.; Feng, S. State of the art on adversarial attacks and defenses in graphs. *Neural Comput. Appl.* **2023**, *35*, 18851–18872. <https://doi.org/10.1007/s00521-023-08839-9>
- [18] Kapusta, K.; Mattioli, L.; Addad, B.; et al. Protecting ownership rights of ML models using watermarking in the light of adversarial attacks. *AI Ethics* **2024**, *4*, 95–103. <https://doi.org/10.1007/s43681-023-00412-3>
- [19] Anaedevha, R. N.; Trofimov, A. G. Improved robust adversarial model against evasion attacks on intrusion detection systems. *Opt. Mem. Neural Networks* **2024**, *33* (Suppl. 3), S414–S423. <https://doi.org/10.3103/S1060992X24700681>
- [20] Qu, Y.; Huang, S.; Li, Y.; et al. BadCodePrompt: Backdoor attacks against prompt engineering of large language models for code generation. *Autom. Softw. Eng.* **2025**, *32*, 17. <https://doi.org/10.1007/s10515-024-00485-2>

-
- [21] Ying, Z.; Wu, B. DLP: Towards active defense against backdoor attacks with decoupled learning process. *Cybersecurity* **2023**, *6*, 9. <https://doi.org/10.1186/s42400-023-00141-4>
 - [22] Ishtiaq, H. U.; Bhutta, A. A.; Mian, A. N. DHCP DoS and starvation attacks on SDN controllers and their mitigation. *J. Comput. Virol. Hacking Tech.* **2024**, *20*, 15–25. <https://doi.org/10.1007/s11416-023-00483-0>
 - [23] Dupont, G.; dos Santos, D.; Dashevskiy, S.; et al. Demonstration of new attacks on three healthcare network protocols in a lab environment. *J. Comput. Virol. Hacking Tech.* **2024**, *20*, 301–314. <https://doi.org/10.1007/s11416-023-00479-w>
 - [24] Zhao, J.; Jiang, N.; Pei, K.; Wen, J.; Zhan, H.; Tu, Z. TPoison: Data-poisoning attack against GNN-based social trust model. *Mathematics* **2024**, *12*, 1813. <https://doi.org/10.3390/math12121813>
 - [25] Ino, T.; Yoshida, K.; Matsutani, H.; Fujino, T. Data poisoning attack against neural network-based on-device learning anomaly detector by physical attacks on sensors. *Sensors* **2024**, *24*, 6416. <https://doi.org/10.3390/s24196416>