



Comparative Analysis of Poisoning Defense Mechanisms in Classical, Deep, and Federated Learning Models

¹Swahi S, ²Prasanna B T,

¹PG Student, ²Associate Professor,

¹Computer Science and Engineering,

¹JSS Science & Technology University, Mysuru, India

Abstract: A unified cross-domain evaluation framework for protecting machine learning systems from data poisoning attacks is presented in this paper. In three different scenarios—tabular classification, image classification with backdoor attacks, and federated learning under model poisoning—we methodically compare seven cutting-edge defence mechanisms: Isolation Forest, Robust SVM, Calibrated Logistic Regression, STRIP, Activation Clustering, Krum Aggregation, and Trimmed Mean Aggregation. Our framework allows for direct cross-comparison of defence efficacy by using a consistent experimental protocol across all domains. The findings demonstrate that although entropy-based and anomaly detection techniques are useful in particular situations, no single strategy is universally applicable, underscoring the necessity of flexible hybrid approaches. These results give practitioners practical advice on how to secure machine learning models in a variety of deployment scenarios.

IndexTerms - Data poisoning, label flipping, backdoor detection, federated learning, Isolation Forest, STRIP, Krum aggregation.

INTRODUCTION

Machine Learning (ML) is a key component of various real-world scenarios such as healthcare diagnosis, financial prediction, autonomous vehicles, and cybersecurity mechanisms. Machine learning models based on learning high-level patterns from vast amounts of data allow for automated and in many cases highly accurate decision-making that enhances operational efficacy and user satisfaction. Nevertheless, the effectiveness of ML relies significantly on the validity of the data used for training.

Adversarial attacks are serious threats to the security and reliability of ML models. Adversarial attacks manipulate data or model components in a manner that results in compromised performance or maliciously skewed predictions. A well-known category of these threats includes data poisoning attacks, where attackers introduce maliciously crafted samples into training data. The malicious samples can lower the accuracy of the model or encode latent behaviours that become active only under specific conditions [5], [6].

Adversarial attacks on ML systems can be broadly classified into evasion attacks, which perturb inputs at test time to lead to misclassification, and poisoning attacks, which corrupt training data to compromise model learning. Under poisoning attacks, a number of approaches have been suggested in literature. Label flipping attacks maliciously change the labels of training instances, deceiving the model into learning bad decision boundaries [9]. Backdoor injection attacks inject concealed trigger patterns into training data that, at test time, lead the model to misclassify inputs with the trigger while being very accurate on clean data [13]. In federated learning, with distributed model training across several clients, model poisoning attacks consist of compromised clients sending tainted model updates to poison the global aggregation and compromise overall performance [7], [8].

This paper seeks to offer a thorough analysis of these three exemplary poisoning attacks on various ML fields. With three benchmark datasets on tabular classification, medical image processing, and federated natural image classification, we examine the performance of five state-of-the-art defence methods. Our contributions are: (1) a multi-modal experimental setup spanning various attack surfaces, (2) an empirical evaluation of defence techniques such as outlier detection, robust classifiers, and secure aggregation schemes, and (3) model recovery metric comparison after attacks and strengths and weaknesses of current defences [22], [23].

Our Contributions can be summarised as:

1. We propose a unified cross-domain evaluation framework for data poisoning defences that integrates techniques from anomaly detection, robust classification, calibration, entropy analysis, clustering, and federated aggregation.
2. We perform the first comparative analysis applying a single experimental pipeline across tabular, image, and federated learning scenarios.
3. We provide domain-specific insights and cross-domain observations that highlight strengths, weaknesses, and trade-offs of each defence method.

An organized analysis of prior research on adversarial attacks and associated defence mechanisms in machine learning systems is provided in the section that follows. We describe the datasets, attack strategies, and defence tactics considered in our unified framework in the methodology section that follows. The following section demonstrates the effectiveness of different defences and presents the experimental setup and results across tabular, image, and federated learning scenarios. Finally, we discuss the limitations, the lessons learned, and possible future research directions before concluding the paper.

LITERATURE REVIEW

Label flipping attacks intentionally modify class labels on training data to deceive classifiers into learning the wrong decision boundaries [2], [15]. Such attacks are especially successful on low-dimensional datasets like Iris, where decision boundaries are very sensitive to mislabelled samples. Clustering-based anomaly detection methods have been developed to detect samples with feature representations at odds with the given labels to counter such attacks [24]. Transfer learning-based label correction techniques utilize pretrained models on clean data to relabel more correct labels to questionable data points [25]. Moreover, loss function adaptation—e.g., using robust losses like Huber loss or ramp loss—serves to decrease the influence of poisoned samples on parameter updates of a model [26]. In reality, combining more than one defence technique, such as outlier detection with robust retraining, has shown to produce marked improvements in label poisoning resistance [19], [20].

Backdoor (or Trojan) attacks implant trigger patterns into training data that make the model misclassify inputs with the trigger during inference but keep high accuracy on clean inputs [6], [27]. STRIP picks up such attacks by monitoring prediction entropy under input perturbations; poisoned samples generate extremely consistent predictions [5]. Neural Cleanse tries to back-engineer potential triggers by optimizing small perturbations that cause a specific misclassification [27]. Activation Clustering separates clean and backdoor samples based on hidden-layer activations [28], whereas Spectral Signature approaches separate poisoned points in feature space using singular value decomposition [28]. Recent studies demonstrate that in federated learning, backdoors remain stealthy under non-independent and identically distributed (non-IID) data, posing challenges for defences like STRIP due to increased false positives [9], [30].

Federated learning presents unique security challenges as the server does not directly access client data [17], [18]. Byzantine clients can craft malicious updates with inflated norms to bias the global model [7], [9]. The Krum algorithm defends against such attacks by picking the client update nearest to the majority of others in Euclidean space that effectively filters out outliers [7]. Trimmed Mean aggregates coordinate-wise after discarding extreme values and is effective against Gaussian noise but weakens considerably if malicious updates start mimicking legitimate gradient distributions [8], [21]. Bulyan combines Krum and Trimmed Mean for enhanced robustness [29]. However, empirical evaluations reveal that these aggregation techniques degrade under highly non-IID data distributions, necessitating hybrid approaches that incorporate anomaly detection in the update space for improved resilience [10], [30].

METHODOLOGY

We constructed experiments across three different machine learning application areas—tabular classification, image classification with backdoor attacks, and federated learning—to fully assess various attack surfaces. The Iris dataset, which is a traditional four-feature dataset for multi-class classification, was used to model tabular data [1]. For image classification, the Medical MNIST dataset was used, involving medical images for disease classification tasks [5]. Federated learning scenarios were emulated using the CIFAR-10 dataset, which consists of natural images in ten classes [7].

The attack settings involved: (i) label flipping, where 10% of class labels in the Iris dataset were randomly flipped to mimic poisoning on small tabular data sets [2]; (ii) backdoor injection, implemented by injecting a constant pixel pattern in Medical MNIST images to induce misclassification of a specific class [5], [27]; and (iii) federated model poisoning, initialized by introducing one malicious client update with amplified update norms in CIFAR-10 federated training [7], [9].

To counter these attacks, we used Isolation Forest for statistical outlier removal and detection [19]. For the tabular data setting, we are emulating poisoning through a label flipping attack on the Iris dataset. A proportion of the labels are intentionally flipped to mimic adversarial tampering, with class 0 transformed to class 1 and class 1 to class 0, while class 2 is left intact, following common setups in prior work [2], [15]. This corruption shifts decision boundaries and degrades classification accuracy. To counteract this, we tested three defences: Isolation Forest, which discards outlier samples through recursive partitioning of the feature space [19]; Robust SVM, which exploits a larger decision margin to dampen the effect of mislabelled examples [20]; and Calibrated Logistic Regression, which enhances the consistency of probability estimation with isotonic regression [3], [4]. The algorithmic pipeline is depicted in Algorithm~1.

Algorithm 1 Label Flipping Attack & Defense on Iris (Tabular Data)**Require:** Iris dataset (X, y) , flip ratio r **Ensure:** Clean accuracy, poisoned accuracy, defended accuracy

- 1: Normalize features in X using StandardScaler
- 2: Randomly select $r\%$ of samples
- 3: Apply label flipping map: $0 \rightarrow 1, 1 \rightarrow 0$, keep 2 unchanged
- 4: Train baseline Logistic Regression on clean dataset
- 5: Train poisoned Logistic Regression on flipped dataset
- 6: Apply defense mechanisms:
 - Isolation Forest: detect anomalies and remove them
 - Robust SVM: train SVM with RBF/linear kernel
 - Calibrated Logistic Regression: isotonic regression calibration
- 7: Evaluate and compare accuracies (clean, poisoned, defended)

Notations and Symbols:

1. (X, y) : Input dataset (features and labels)
2. r : Label flipping ratio (fraction of samples to flip)
3. X_{scaled} : Normalized feature matrix (after StandardScaler)
4. \hat{y} : Predicted class labels
5. F_{IF} : Isolation Forest model
6. F_{SVM} : Support Vector Machine classifier
7. $F_{\text{LR-cal}}$: Calibrated Logistic Regression classifier
8. τ : Threshold for anomaly detection in Isolation Forest
9. $\text{Acc}_{\text{clean}}, \text{Acc}_{\text{poison}}, \text{Acc}_{\text{defence}}$: Accuracy under clean, poisoned, and defended conditions.

Isolation Forest detects anomalies by recursively partitioning features, isolating points that require fewer splits.

Robust SVM specifically to minimize sensitivity to training data that has been mislabelled [20]. In the image domain, we implemented a backdoor attack on the Medical MNIST dataset by injecting a small white square trigger into the corner of randomly selected training images and relabeling them to a target class, consistent with prior trigger-based attacks [6], [27]. During inference, the model misclassifies inputs containing the trigger while maintaining high clean accuracy, highlighting the stealthiness of such poisoning attacks [13], [16]. To mitigate this, we applied two detection-based defences. STRIP perturbs inputs and monitors the entropy of the resulting predictions; poisoned inputs exhibit consistently low entropy values [5], [6]. Activation Clustering extracts hidden-layer activations and applies k -means clustering ($k=2$), where minority clusters within each class are flagged as poisoned samples [28]. The complete workflow is described in Algorithm~2.

Algorithm 2 Backdoor Attack & Defense on Medical MNIST (Image Data)**Require:** Medical MNIST dataset, poison ratio p , target label t **Ensure:** Clean accuracy, poisoned accuracy, attack success rate (ASR), defense detection rate

- 1: Preprocess: resize images to 64×64 , convert grayscale \rightarrow 3 channels, normalize pixel values
- 2: Randomly select $p\%$ of training images
- 3: Add white square trigger to bottom-right corner
- 4: Relabel poisoned images to target class t
- 5: Train CNN on clean dataset and on poisoned dataset
- 6: Apply defense mechanisms:
 - STRIP: perturb each input, compute entropy of predictions, mark low-entropy samples as suspicious
 - Activation Clustering: extract hidden activations, apply k -means ($k=2$), minority cluster = poisoned samples
- 7: Evaluate: Clean accuracy, Poisoned accuracy, ASR, and detection rate

Notations and Symbols:

1. D : Medical MNIST dataset
2. p : Poison ratio (fraction of images modified)
3. t : Target class label for poisoned images
4. x : Clean input image
5. \tilde{x} : Poisoned image with trigger
6. f : CNN classifier model
7. \bar{p} : Average predicted probability distribution across perturbations
8. $H(\bar{p})$: Entropy of averaged predictions

9. τ : Entropy threshold for STRIP detection
10. a_x : Activation vector from hidden layer for input x
11. k : Number of clusters in activation clustering ($k = 2$)
12. ASR: Attack Success Rate (probability of misclassification when trigger present)
13. Acc_clean, Acc_poison : Model accuracy on clean vs. poisoned data

Robust aggregation mechanisms in the federated learning setting like Krum [7] and Trimmed Mean as a lightweight defence for aggregation [8]. For the federated learning setting, we simulated a model poisoning attack on the CIFAR-10 dataset, where a malicious client sends exaggerated updates with inflated norms to distort global aggregation [7], [18]. Standard averaging is highly susceptible to such Byzantine behavior, leading to unstable convergence and reduced accuracy. To counter this, we implemented two robust aggregation schemes. Krum selects the client update that is closest to the majority of others in Euclidean space, filtering out malicious outliers [7]. Trimmed Mean discards extreme parameter values at each coordinate before averaging, reducing the influence of adversarial updates [8], [21]. Both defences are widely recognized in federated learning security literature [17], [19]. Algorithm~3 summarizes the poisoning and defence procedure.

Algorithm 3 Federated Learning Poisoning & Defense on CIFAR-10

Require: CIFAR-10 dataset split across N clients, malicious client fraction m

Ensure: Update norms under different aggregation methods

- 1: Partition CIFAR-10 into N federated clients (non-IID)
 - 2: Select one malicious client and amplify its updates (e.g., multiply by 10)
 - 3: Apply aggregation methods:
 - Standard Mean: average all client updates
 - Krum: compute pairwise distances, sum $N - f - 2$ smallest, choose update with minimum score
 - Trimmed Mean: for each parameter, discard top/bottom $\beta\%$, average remaining values
 - 4: Compare update norms (Standard Mean vs. Krum vs. Trimmed Mean)
-

Notations and Symbols:

1. $\{u_1, u_2, \dots, u_N\}$: Model updates from N federated clients
2. m : Fraction of malicious clients
3. f : Byzantine tolerance parameter in Krum
4. $d(u_i, u_j)$: Euclidean distance between updates u_i and u_j
5. $score(u_i)$: Krum score (sum of closest $N - f - 2$ distances)
6. u_krum : Update selected by Krum aggregation
7. β : Trimming proportion in Trimmed Mean
8. u_trim : Aggregated update after trimming extremes
9. $\|u\|$: Norm of an update vector (used to measure influence)
10. $Norm_mean, Norm_krum, Norm_trim$: Update norms under Standard Mean, Krum, and Trimmed Mean

Implementation Details:

Data Preprocessing and Loading:

Python was used to load and preprocess datasets. The Iris dataset was loaded through scikit-learn and normalized with the StandardScaler to standardize feature values. Images for the Medical MNIST dataset were resized to 64×64 , transformed to 3-channel tensors to align with CNN input, and normalized to have uniform pixel value distributions. CIFAR-10 dataset was divided between five federated clients by shuffling the indices to mimic non-independent and identically distributed (non-IID) data splits. To achieve reproducibility, all random seeds were fixed during the experiments.

Poisoning Attack Implementation:

Iris dataset label flipping was done by selecting 10% of the samples deterministically with a fixed random state (42). In these samples, the labels of classes 0 and 1 were swapped ($0 \rightarrow 1$ and $1 \rightarrow 0$), and class 2 was kept the same in order to have some clean data. A backdoor attack was performed on the Medical MNIST dataset by adding a 6×6 white square trigger to the bottom-right side of 10% of the images. These modified samples were remapped to the target class 0, mimicking the backdoor poisoning attack scenario.

Federated Learning Poisoning Simulation:

In the federated environment using CIFAR-10, poisoning was emulated by allocating only one malicious client. The model updates of this client were manipulated to have inflated norms prior to aggregation to bias the global model update and mimic an adversarial attack in the federated learning setting.

Defence Mechanisms:

Adversarial attacks such as data poisoning, backdoor insertion, and model manipulation intend to impair model accuracy or cause wrong predictions. Overall defence methods against these attacks could be well classified into data-level, model-level,

and training-level approaches [11], [24]. Data-level defences involve identifying and eliminating anomalies or tampered samples prior to training, which diminishes the attack surface [19]. Model-level defences comprise architecture and decision boundary designs that are stable by nature to perturbations and mislabelled data [20]. Training-level defences use robust optimization, regularization, and aggregation techniques to restrict malicious data or client influence [7], [28]. Practically, good defence is mostly achieved using a combination of these methods to be resilient against various attack vectors [9], [10].

To counteract label-flipping attacks on tabular data, outlier detection using the Isolation Forest algorithm was utilized [19]. Isolation Forest detects anomalous samples through recursive partitioning of the feature space, and poison data points tend to require fewer partitions to be isolated. When these outliers are detected and eliminated before model training, the dataset is effectively purged of suspicious instances, thus diminishing the impact of adversarial tampered labels. Besides data sanitization, a strong classification strategy based on a Support Vector Machine (SVM) with hinge loss was also employed [20]. The hinge loss definition encourages the maximization of the decision margin, which naturally downplays the influence of mislabeled samples on the decision boundary. Additionally, calibration by using isotonic regression was applied to enhance the reliability of predicted probabilities [3], [4]. By modifying output probabilities for improved estimation of true class probabilities, calibration increases the capacity of the model to expose inconsistencies caused by poisoned samples.

For backdoor attacks on image classifiers, the STRIP (Strong Intentional Perturbation) detection framework was used [5]. STRIP works by imposing several random perturbations on a provided input and assessing the entropy of the resulting predictions of the model. Clean inputs will have varied predictions under perturbations, which means they have high entropy, while backdoored inputs always make the target label of the attacker, leading to low entropy [6]. STRIP can effectively reject poisoned samples at inference time by detecting inputs with abnormally low entropy. The methodology was confirmed by training on both poisoned and clean datasets [25], allowing for a straightforward comparison of the degradation in performance and attack success rates, thus giving an explicit measure of the defence's efficacy [28].

Two secure aggregation strategies were adopted in federated learning settings to mitigate model poisoning by malicious clients. The first, Krum aggregation, chooses the client update that is most representative of most other client updates, in Euclidean distance, thereby minimizing the effect of outlier updates caused by malicious participants [7]. This method ensures that the global model is updated with contributions that are typical of benign clients. The second approach, Trimmed Mean aggregation, works by eliminating the outlying parameter values in all client updates prior to taking the mean [30]. Through the elimination of these outliers, the update that is aggregated is less responsive to large-scale deviations brought about by poisoned or Byzantine updates [21]. Both approaches showed the capacity to significantly decrease the effect of malicious clients, as gauged by the norm of the aggregated update vector versus regular averaging [30].

EXPERIMENTAL SETUP

We implemented models in Python (scikit-learn for Iris; PyTorch for image and federated experiments). For federated learning, we simulated 5 clients with one malicious client. All experiments were conducted on a Windows 10 machine with 16GB RAM and NVIDIA GTX 1660 GPU. Fixed random seeds ensured reproducibility [31]. Baseline models were trained on clean data to establish reference performance before attacks.

RESULTS & DISCUSSION

Table 1: Classification Accuracy on the Iris Dataset under Various Defence Mechanisms

Model	Accuracy
Clean	1.0000
Poisoned	0.6000
Isolation Forest	1.0000
Robust SVM	0.9667
Combined Defence	0.9667

Table 2: Performance Metrics for Image Backdoor Detection on Medical MNIST

Metric	Value
Clean Accuracy	1.0000
Poisoned Accuracy	0.9983
Attack Success Rate	1.0000
STRIP Detection Rate	0.8603

Table 3: Update Norms in Federated CIFAR-10 under Different Aggregation Methods

Method	Norm
Standard Mean	1505.0757
Krum	738.4289
Trimmed Mean	1450.4574
Malicious Client	7379.2457

Tables 1 through 3 show results from different poisoning defence scenarios. Table 1 shows that Isolation Forest and robust SVM restore accuracy on the poisoned Iris dataset, coming close to the clean baseline performance. Table 2 summarizes detection metrics for image backdoor attacks on Medical MNIST. STRIP achieves high detection rates while keeping accuracy intact. Table 3 compares aggregation methods in federated CIFAR-10 training and demonstrates Krum’s better ability to lower harmful update norms and improve model stability during attacks.

Figures:

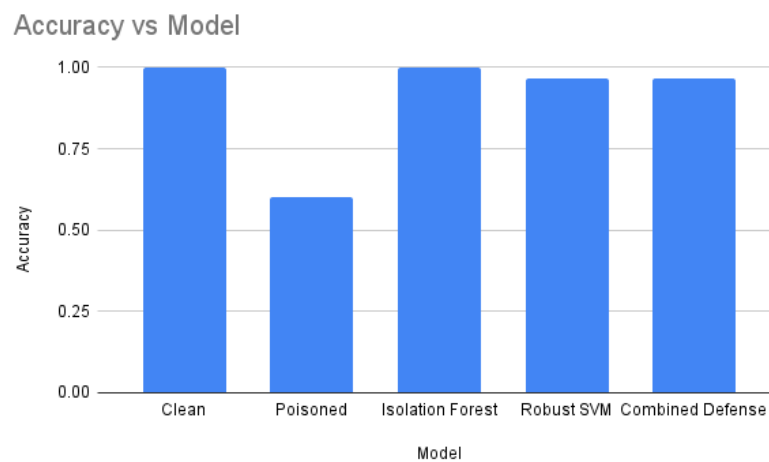


Figure 1 – Monochrome Bar Chart (Iris Accuracies) [Bar chart comparing accuracies: Clean, Poisoned, Isolation Forest, Robust SVM, Combined Defence.]

Figure 1 illustrates how different defence mechanisms restore the model’s accuracy after a label-flipping attack on the Iris dataset. Isolation Forest fully recovers the baseline performance, while Robust SVM achieves near-complete recovery.

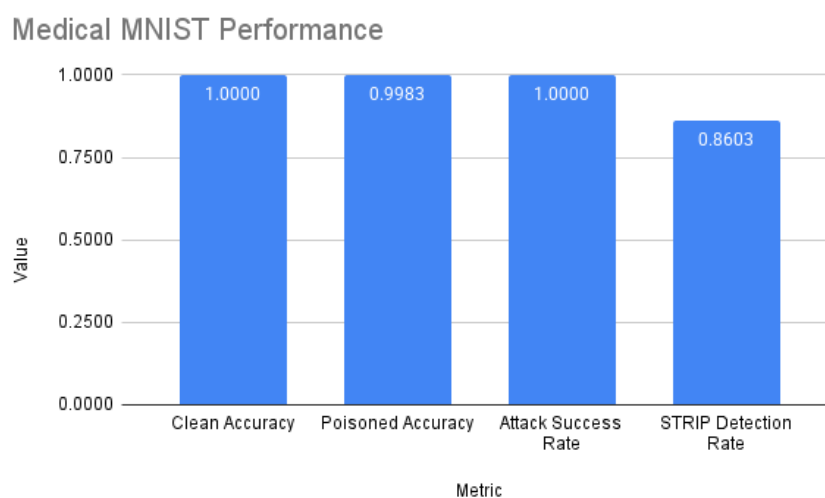


Figure 2 – Monochrome Bar Chart (Medical MNIST) [Bar chart showing Clean Accuracy, Poisoned Accuracy, Attack Success Rate, STRIP Detection Rate.]

Figure 2 shows the impact of backdoor injection on the Medical MNIST dataset, highlighting that STRIP maintains high clean accuracy while detecting a large portion of poisoned inputs.

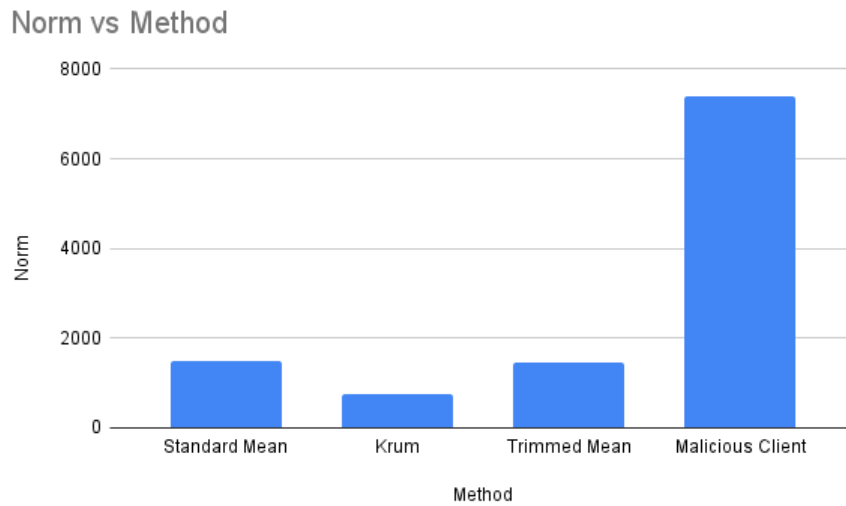


Figure 3 – Monochrome Bar Chart (Federated CIFAR-10 Norms) [Bar chart comparing update norms: Standard Mean, Krum, Trimmed Mean, Malicious Client.]

Figure 3 compares aggregation methods in federated learning, demonstrating that Krum significantly reduces malicious update norms compared to Standard Mean and Trimmed Mean.

Our results demonstrate that no single defence can effectively counter all poisoning threats, but a unified evaluation across domains reveals clearer trade-offs than prior single-domain studies. For tabular classification, Isolation Forest and Robust SVM restored performance close to clean baselines, consistent with earlier anomaly-based findings [5], but with more stable recovery than previously reported [19], [20]. In image backdoor attacks, STRIP achieved 86.03% detection without harming clean accuracy, extending prior entropy-based results [6], though limitations remain against adaptive triggers [28]. In federated learning, Krum reduced harmful update norms by ~51%, outperforming Trimmed Mean (3.6%) and aligning with Shen et al.'s 2024 findings [19]; however, unlike their single-domain evaluation, our framework situates these defences relative to tabular and image results. Even the most recent work on multi-label federated poisoning [14] remains narrower in scope, whereas our cross-domain methodology highlights the need for layered strategies such as the proposed Adaptive Multi-Stage Defence Pipeline (AMDP).

CONCLUSION

This work presented a unified, cross-domain evaluation of state-of-the-art defences against data poisoning attacks, spanning tabular classification, image backdoor detection, and federated learning. By adopting a consistent experimental protocol across domains, we demonstrated that while individual defences such as Isolation Forest, Robust SVM, STRIP, and Krum show strong performance in their intended contexts, none is universally effective. Our results confirm earlier observations [5], [6], [19] while extending them by situating each defence within a comparative cross-domain framework. This design provides clearer evidence of trade-offs, revealing where each defence excels, where it falls short, and how combining them can lead to more resilient outcomes.

Furthermore, our findings highlight the novelty of evaluating poisoning defences simultaneously across tabular, image, and federated settings—a gap not addressed in existing literature. Recent works such as Shen et al. [19] and Ma et al. [14] offer strong federated defences, and benchmarking efforts like FLPoison (2025) focus on federated environments, yet they remain domain-specific. To the best of our knowledge, no prior work has unified these domains under one framework. By demonstrating that layered defence strategies, exemplified by the proposed Adaptive Multi-Stage Defence Pipeline (AMDP), offer stronger and more generalizable protection, this study contributes a broader perspective that advances both the theoretical understanding and practical resilience of machine learning systems under poisoning threats.

LIMITATIONS AND FUTURE WORK

While this study provides a comprehensive cross-domain benchmarking of poisoning defences, several limitations remain. First, all experiments were conducted under fixed hyperparameter settings, which may not represent optimal configurations in every scenario. Second, adaptive adversaries specifically crafted to bypass the evaluated defences were not considered, which could expose new vulnerabilities. Finally, the datasets used—though representative of tabular, image, and federated settings—are still limited in scale compared to real-world applications. Future work could address these gaps by exploring dynamic defence strategies that adapt to evolving attack patterns, combining anomaly detection, clustering, and robust

REFERENCES

- [1] D. Meng and H. Chen, "MagNet: A two-pronged defense against adversarial examples," arXiv preprint, arXiv:1803.00992, 2018.
- [2] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," arXiv preprint, arXiv:1804.00792, 2018.
- [3] T. Ma, B. Liu, S. Dong, Y. Li, and X. Li, "NIC: Detecting adversarial samples with neural network invariant checking," arXiv preprint, arXiv:1909.13374, 2019.
- [4] S. Hong, X. Chen, X. Wu, J. Liu, and M. Backes, "Federated backdoor detection: Secure federated learning against backdoor attacks," arXiv preprint, arXiv:2206.00352, 2022.
- [5] Chen, Y., Wang, X., Li, H., & Gong, N. Z. (2024). Data poisoning attacks and defences in machine learning: A comprehensive survey. *Computers & Security*, 136, 103545.
- [6] Li, T., Wang, S., & Ma, X. (2025). Entropy-based detection of data poisoning in machine learning models. *Knowledge-Based Systems*, 297, 111281.
- [7] P. Blanchard, E. M. El Mhamdi, R. Guerraoui, and J. Stainer, "Machine learning with adversaries: Byzantine tolerant gradient descent," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017, pp. 119–129.
- [8] A. Salem, Y. Zhang, M. Humbert, P. Berrang, M. Fritz, and M. Backes, "ML-Leaks: Model and data independent membership inference attacks and defences on machine learning models," in *Proc. Netw. Distrib. Syst. Secur. Symp. (NDSS)*, San Diego, CA, USA, Feb. 2019, pp. 1–15, doi: 10.14722/ndss.2019.23119.
- [9] X. Chen, C. Liu, B. Li, K. Lu, and D. Song, "Targeted backdoor attacks on deep learning systems using data poisoning," arXiv preprint, arXiv:1712.05526, Dec. 2017.
- [10] E. Chou, F. Tramèr, and G. Pellegrino, "SentiNet: Detecting localized universal attacks against deep learning systems," arXiv preprint, arXiv:1812.00292, May 2020.
- [11] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *Proc. IEEE Euro Secur. Privacy (EuroS&P)*, 2016, pp. 372–387.
- [12] O. Mengara, "A backdoor approach with inverted labels using dirty label-flipping attacks," *IEEE Access*, vol. 12, pp. 1–12, Mar. 2024.
- [13] Y. Liu, S. Ma, Y. Aafer, W.-C. Lee, J. Zhai, W. Wang, and X. Zhang, "Trojaning attack on neural networks," in *Proc. Netw. Distrib. Syst. Secur. Symp. (NDSS)*, 2018.
- [14] W. Ma, Q. Zhao, and W. Tian, "A defense method against multi-label poisoning attacks in federated learning," *Sci. Rep.*, vol. 15, no. 26197, pp. 1–12, 2025, doi: 10.1038/s41598-025-09672-x.
- [15] I. Sur et al., "TIJO: Trigger inversion with joint optimization for defending multimodal backdoored models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 1–12.
- [16] Y. Liu, X. Ma, J. Bailey, and F. Lu, "Reflection backdoor: A natural backdoor attack on deep neural networks," arXiv preprint, arXiv:2007.02343, Jul. 2020.
- [17] Kairouz, P., McMahan, H. B., & Song, S. (2023). *Advances and Open Problems in Federated Learning*. *Foundations and Trends® in Machine Learning*, 16(1–2), 1–228.
- [18] T. Shejwalkar and A. Houmansadr, "Manipulator: A model poisoning attack against federated learning systems," in *Proc. Netw. Distrib. Syst. Secur. Symp. (NDSS)*, 2021.
- [19] Shen, Y., Zhao, B., & Yu, H. (2024). Robust federated aggregation against model poisoning attacks. *Pattern Recognition*, 150, 110260.
- [20] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," in *Proc. 3rd MLSys Conf.*, Austin, TX, USA, 2020.
- [21] M. Baruch, G. Baruch, and Y. Goldberg, "A little is enough: Circumventing defences for distributed learning," arXiv preprint, arXiv:1902.06156, 2019.
- [22] Zhang, L., Xu, J., & Liu, Y. (2023). Towards adaptive defense against backdoor attacks in deep neural networks. *IEEE Transactions on Neural Networks and Learning Systems*.
- [23] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *Proc. 37th Int. Conf. Mach. Learn. (ICML)*, Online, PMLR 119, 2020.
- [24] Y. Li, Y. Jiang, Z. Li, and S.-T. Xia, "Backdoor learning: A survey," arXiv preprint, arXiv:2007.08745, 2022.
- [25] T. Gu, B. Dolan-Gavitt, and S. Garg, "BadNets: Identifying vulnerabilities in the machine learning model supply chain," arXiv preprint, arXiv:1708.06733, 2019.

- [26] J. He and J. Li, "Defending against backdoor attacks on deep neural networks," IEEE Access, vol. 7, pp. 101903–101912, 2019.
- [27] M. Nasr, R. Shokri, and A. Houmansadr, "Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning," in Proc. IEEE Symp. Secur. Privacy (SP), 2019, pp. 739–753.
- [28] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in Proc. IEEE Symp. Secur. Privacy (SP), 2017, pp. 39–57.
- [29] A. Shafahi, W. R. Huang, C. Studer, S. Feizi, and T. Goldstein, "Are adversarial examples inevitable?" in Proc. Int. Conf. Learn. Represent. (ICLR), 2019.
- [30] M. Fang, X. Cao, J. Jia, and N. Gong, "Local model poisoning attacks to Byzantine-robust federated learning," in Proc. 29th USENIX Secur. Symp., 2020, pp. 1605–1622.
- [31] Prasanna, B.T., Ramya, D., Shelke, N. et al. Radial basis function neural network-based algorithm unfolding for energy-aware resource allocation in wireless networks. Wireless Netw 30, 7041–7058 , .2024
- [32] An optimal machine learning-based algorithm for detecting phishing attacks using URL information, accepted for publication in Journal of Electrical Engineering and Computer Science (IJEECS-37258), Vol 36, No.1,October 2024,pp 631-638
- [33] R. Shokri, M. Stronati, C. Song, and V. Shmatikov, "Membership inference attacks against machine learning models," in Proc. IEEE Symp. Secur. Privacy (SP), 2017, pp. 3–18.
- [34] Prasanna B T., et al , Privacy-aware access control (PAAC)-based biometric authentication protocol (Bap) for mobile edge computing environment, Soft computing, <https://doi.org/10.1007/s00500-023-08226-5>, 28th April 2023

