Vidhyayana - ISSN 2454-8596
An International Multidisciplinary Peer-Reviewed E-Journal
www.vidhyayanaejournal.org
Indexed in: Crossref, ROAD & Google Scholar

12

A Survey on Machine Learning Approaches to Detect Label Flipping Adversarial Poisoning Attacks

Rucha Gurav

Research Scholar, Dr. Vishwanath Karad MIT World Peace University, Pune, India

Rashmi Phalnikar

Professor, Dr. Vishwanath Karad MIT World Peace University, Pune, India

Abstract:

Label-flipping adversarial poisoning attacks present a substantial threat to the integrity and security of machine learning (ML) models by deliberately altering the labels in the training dataset. Such manipulation can significantly distort model predictions, leading to compromised performance and unreliable decision-making. The detection of these adversarial attacks is paramount to maintaining the robustness and trustworthiness of ML systems, particularly in critical domains like cyber security, finance, and healthcare, where model reliability is of utmost importance. This paper offers an extensive review of contemporary methods and approaches for detecting label-flipping adversarial poisoning attacks, utilizing various machine learning algorithms. We conduct a comparative analysis of the strengths and limitations of existing detection strategies, focusing on both supervised and unsupervised learning paradigms. Furthermore, we examine the influence of critical factors such as feature engineering, model interpretability, and the challenges posed by class imbalance on the effectiveness of detection methods. Finally, this review highlights current challenges, identifies existing research gaps, and outlines future directions for advancing detection mechanisms, thereby contributing to the development of more resilient and secure machine learning models capable of withstanding adversarial manipulation.

Keywords: Dynamic label poisoning, adversarial attacks, machine learning robustness, data poisoning evaluation, stealth attack strategies

1. Introduction

Machine learning plays a crucial role in automating decision-making based on data, improving predictive accuracy, and driving advancements in diverse fields such as healthcare, finance, and autonomous technologies, ultimately minimizing human effort and enhancing efficiency. Training data is important entity in this technology to make models useful for applications in widespread domains. The integrity and reliability of this data used for training the models is major concern due to manipulation for affecting the model's performance. This critical threat posed to ML systems is important concern considered for the work in this paper. The data poisoning attacks which is especially label-flipping is mainly addressed in this paper along with analysis of its impact on classifier performance. In this type of attack the labels of the data are manipulated due to which classifier convergence state is hampered and also the model performs poorly on the test datasets.

Label-flipping adversarial poisoning attacks are a growing concern due to their ease of execution and high effectiveness in degrading ML model performance. These attacks present substantial risks to various critical domains, including cybersecurity, finance, and healthcare, where compromised model integrity can result in severe financial losses, security breaches, or even threats to human safety. By strategically altering the labels of training samples, adversaries can cause significant distortions in model predictions, leading to unreliable decision-making and misclassification of data. Given their simplicity, such attacks can be executed without requiring sophisticated computational resources, making them an accessible yet potent threat. Consequently, robust detection and mitigation strategies are essential to safeguard ML systems from these adversarial manipulations and ensure their continued reliability in real-world applications. Despite increasing research efforts in data poisoning attacks, there remains an urgent need to fully understand the scope and consequences of label-flipping attacks, evaluate existing detection techniques, and develop stronger, more adaptive defenses. The current literature on data poisoning has primarily focused on analyzing the impact of these attacks across various ML algorithms, demonstrating how label manipulation can significantly degrade model performance in a range of applications, from malware detection and spam filtering to fraud detection and human activity recognition. Label-flipping attacks exploit the inherent dependency of ML models on large-scale, crowd-sourced, or publicly available datasets, where minimal oversight allows adversaries to inject manipulated data with ease. Additionally, these attacks pose a major challenge in federated learning and decentralized ML frameworks, where individual clients may unknowingly introduce corrupted labels, compromising overall model performance. DL models, despite their advanced architectures and high-dimensional feature extraction capabilities, are not immune to such poisoning attacks. On the contrary, their reliance on large-scale datasets and computationally intensive training processes makes them particularly vulnerable, as even a small fraction of poisoned labels can lead to substantial degradation in predictive accuracy. In real-world deployments, the vulnerability of DL models to adversarial label manipulation raises significant concerns, especially in high-stakes applications like medical diagnosis, biometric authentication, and autonomous systems. Without effective countermeasures, ML and DL models remain susceptible to manipulation, underscoring the need for continuous research into more resilient and adaptive defense mechanisms against label-flipping adversarial attacks.

Variety of label flipping attacks are discussed in the review of this article. In both supervised and unsupervised learning models, the effects of label flipping poisoning attacks are studied with more details about feature engineering requirements for this challenge. The paper contributes in terms of highlighting the challenges in attack strategies and their impacts on classifier performance with respect to adversarial manipulations. The future directions are also highlighted for the design of models to avoid the impact of poisoning attacks to achieve secured and reliable outcomes for building the trust in this technology. Further the review of existing methods is carried out in section 2 followed by some details of some attacking methods in label flipping attacks. The results and analysis is done on the selective datasets with these attacks on them and performance is evaluated using classifiers training and testing scenarios.

2. Literature Review

Rosenfeld et al. [1] proposed a randomized smoothing framework for robust classifiers against label-flipping attacks, ensuring consistent predictions by providing deterministic bounds. Their approach enhances multi-class classification reliability in adversarial settings, reducing risks from poisoned datasets. Shahid et al. [2] examined label-flipping attacks on wearable Human Activity Recognition (HAR) systems, demonstrating ML performance degradation. A KNN-based defense mechanism was evaluated, emphasizing the need for robust security in safety-critical HAR applications. Aryal et al. [3] investigated poisoning attacks on malware detectors, highlighting vulnerabilities in crowd-sourced datasets. Their study underscores the necessity of robust defenses to prevent mislabeled data from undermining malware detection systems. Chang et al. [4] introduced Falfa, a label-flipping attack for tabular data using linear programming to manipulate training labels efficiently. Tested on ten datasets, Falfa exposed classifier weaknesses and emphasized the need for stronger security measures in cybersecurity applications. Mengara [5] presented DirtyFlipping, a backdoor attack on audio-based ML models exploiting label-flipping vulnerabilities in third-party datasets. This method demonstrates the risks of outsourcing model training and highlights the need for improved safeguards in audio applications like speech recognition. Surendrababu and Nagaraj [6] proposed an entropy-based method to detect backdoor attacks in poisoned datasets by evaluating complexity and entropy measures. Their approach demonstrated high detection accuracy across various domains, ensuring dataset integrity in machine learning models. Umer and Polikar [7] introduced Adversary Aware Continual Learning (AACL), a framework that neutralizes backdoor poisoning attacks in continual learning. Tested on datasets like CIFAR-10 and MNIST, AACL significantly improved model robustness without depending on specific learning algorithms. Liu et al. [8] developed countermeasures against poisoning attacks on deep neural networks using a denoising autoencoder-based approach, Data Washing, and an Integrated Detection Algorithm (IDA). Their method effectively reduced false positives and enhanced detection accuracy. Altoub et al. [9] constructed a study material in which they provided categorizing 55 poisoning attack types in deep neural networks. Their research advances AI security by providing a structured analysis of adversarial threats, aiding in the development of more resilient defenses. Sun et al. [10] introduced the Attacking-Distance-Aware (ADA) method in federated learning scenarios for poisoning the model in federated learning, optimizing target class selection for more effective attacks. Their study emphasized the need for improved defenses in federated models vulnerable to adversarial manipulation.

Cui et al. [11] proposed hybrid method of combined optimization methods. The method includes Particle Swarm Optimization (PSO) algorithms with modified objective functions. The objective function is updated with addition of Genetic Algorithm. This affects the black box nature of the attacks by manipulating the clean label data especially used in autonomous vehicles models. Their method significantly degraded global model accuracy with minimal poisoned data, validated on traffic sign recognition tasks. Psychogyios et al. [12] examined data poisoning attacks using GANs in federated learning, introducing label-flipping and targeted label attacks with synthetic images. These attacks caused up to 25% performance degradation and 56% misclassification. A clean-label training mitigation method showed partial effectiveness but highlighted the stealth of GAN-driven attacks. Zhang et al. [13] considered the emotion recognition application for EEG processing methods. They studied the attacks on the labels of signals that affects the emotion class numbers in the datasets. They also shown that this attack affects by reducing classification accuracy in six ML models. Explainable AI techniques like SHAP and LIME helped illustrate attack mechanisms and their impact on decision-making. Liu et al. [14] introduced a multi-target backdoor attack using procedural noise textures and k-LSB steganography, achieving up to 100% success rates on GTSRB and 98.48% on ImageNet. The study underscored the difficulty of detecting dynamic, invisible attacks. Kim et al. [15] proposed an attack method that affects the labels in dataset used for face recognition using accessory injection and feature transfer, demonstrating high attack success while preserving benign classification accuracy.

Maabreh et al. [16] analyzed clustering-based label-flipping attacks, introducing dormant poisoned samples that bypass anomaly filters, providing insights into the trade-off between complexity and poisoning resistance. Wu et al. [17] examined security challenges in federated learning, proposing countermeasures like optimizing computing power and trusted federations to mitigate poisoning attacks. Vassilev and Oprea [18] developed a taxonomy for adversarial machine learning attacks and defenses, serving as a foundational resource for ML security. Archa and Kartheeban [19] introduced SecureTransfer, a transfer learning-based approach using VGG16 to detect poisoning attacks, integrating GANs and CNNs for defense in healthcare and autonomous systems. Singh et al. [20] explored poisoning attacks and their impacts on the models in scenarios with federated learning, balancing security with fairness for diverse data distributions.

Singh et al. [21] worked on poisoning attacks detection method. Their approach balanced security and fairness, mitigating bias in detecting malicious updates while improving model performance. Paracha et al. [22] reviewed security threats in machine learning, focusing on adversarial machine learning (AML). They analyzed poisoning attacks, their effects, and mitigation strategies like data sanitization and adversarial training, providing insights for building trustworthy ML systems. Raghavan et al. [23] introduced MOVCE, a CNN and word embedding-based verification algorithm to counter poisoning attacks in applications of computer vision and deep learning. Their study highlighted the need to address training-stage vulnerabilities for safety-critical applications like autonomous driving. Huang et al. [24] proposed sponge attacks targeting Multi-Exit Networks (MENs), increasing inference latency through data poisoning while maintaining classification accuracy. Their research emphasized security risks in MEN architectures and the need for improved defenses.

Numerous challenges persist in the domain of poisoning attacks and their mitigation strategies within machine learning systems, highlighting significant gaps that need to be addressed. Firstly, a major limitation of existing defense mechanisms is their lack of generalizability across various ML models and datasets, often making them effective only in highly specific scenarios while failing in more generalized settings. The continuous evolution of attack methodologies further exacerbates this issue, with emerging threats such as multitarget backdoor attacks, adversarial perturbations, and procedural noise triggers often outpacing contemporary detection and mitigation techniques, thereby leaving ML models vulnerable to sophisticated manipulations. Additionally, scalability remains a critical concern, particularly in large-scale federated learning frameworks, where communication inefficiencies, architectural fragility, and resource constraints hinder the deployment of effective defenses against poisoning threats. Another underexplored aspect is the explainability of attack impacts, which limits a comprehensive understanding of how poisoning strategies manipulate model behavior and, in turn, reduces the efficacy of mitigation measures. Without deeper insights into the attack dynamics, crafting countermeasures that can neutralize sophisticated adversarial strategies becomes increasingly difficult. Furthermore, the need for robust and standardized evaluation metrics is pressing, as current benchmarks often fail to effectively quantify the resilience of defenses against advanced poisoning techniques, including GAN-driven, optimization-based, or reinforcement learning-powered adversarial attacks. Additionally, many existing defenses struggle to maintain a balance between robustness and fairness, inadvertently introducing biases that disproportionately affect minority data distributions, particularly in decentralized and federated learning environments. This tradeoff between security and model utility raises concerns about potential discrimination against underrepresented data groups, thereby limiting the real-world applicability of ML models in diverse domains. Addressing these pressing challenges is essential to developing secure, scalable, and trustworthy machine learning systems that can withstand evolving adversarial threats while ensuring fairness, interpretability, and robustness across different applications and environments.

3. Proposed Work

Data poisoning attacks threaten machine learning model security by exploiting vulnerabilities in the training pipeline. This study systematically reviews existing attack methods, identifies key gaps, and proposes a novel attack strategy.

The process follows a structured workflow as shown in Fig 1:

- Review of Existing Methods: Analysis of label flipping, backdoor attacks, and optimization-based poisoning to understand their strengths and weaknesses.
- Gap Analysis: Identifies limitations in existing attacks, such as lack of stealth and inefficiency in large-scale systems.
- Proposed Method: Develops a new attack technique incorporating hybrid optimization, procedural noise triggers, or clean-label backdoors.
- Implementation & Testing: Applies the new attack to poisoned datasets, testing its impact on ML models.
- Performance Evaluation: Compares the method against existing techniques using metrics like attack success rate, model degradation, and stealth.

This approach enhances the understanding of adversarial techniques and provides a framework for designing more resilient ML systems. The insights contribute to both offensive and defensive strategies, improving security in applications such as cybersecurity, healthcare, and finance.

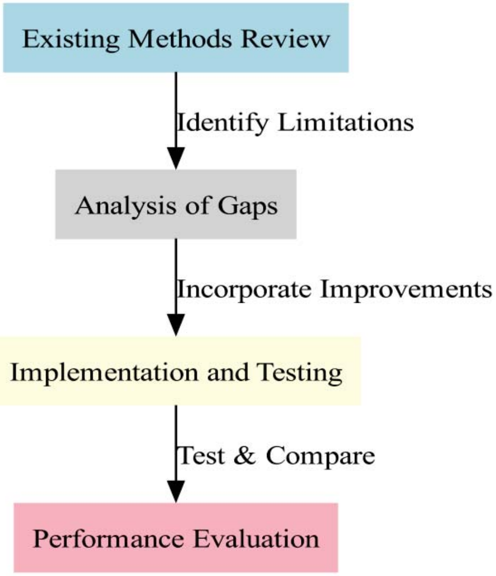

Figure 1: Block Diagram of Proposed Work

This framework encompasses a thorough examination of existing attack techniques and the development of novel strategies to enhance adversarial resilience. The subsequent sections delve into the core algorithms that form the foundation of this research, detailing their mechanisms, effectiveness, and integration within the broader security model.

Algorithm 1: Label-Flipping Robustness (RLF)

Input: Feature mapping $h: \mathbb{R}^d \to \mathbb{R}^k$, noise parameter q, regularization parameter λ, training set $\{(x_i, y_i) \in \mathbb{R}^d \times \{0,1\}\}_{i=1}^{n}$ (potentially with adversarial labels), additional inputs $\{x_j \in \mathbb{R}^d\}_{j=1}^{m}$.

Output: Predictions $\hat{y}_j$ and certification of robustness.

- Pre-compute matrix M: $M = X(X^T X + \lambda I)^{-1}$, where $X = h(x_{1:n})$.
- Compute vector $\beta_j = Mh(x_j)^T$.
- Compute optimal Chernoff parameter t^* via Newton's method:

$$t^* = \underset{t}{\operatorname{argmin}}[t/2 + \sum_{i:y_i=1} \log\left(q + (1-q)e^{-t\beta_{j,i}}\right) + \sum_{i:y_i=0} \log\left((1-q) + qe^{-t\beta_{j,i}}\right)].$$

- Let $p^* = \max(1 - B_j|t^*|,1/2)$, where $B_j|t^*|$ is the Chernoff bound evaluated at $|t^*|$.
- Compute robustness certification:

$$r = \frac{\log(4p^*(1-p^*))}{2(1-2q)\log(q/(1-q))}.$$

- Output prediction $\hat{y}_j = \mathbb{1}\{t^* \geq 0\}$ and certification for up to r label flips.

Algorithm 2: Randomized Label Flipping Attack (RLF-Random)

Input: Training dataset D, maximum iterations N_s, loss threshold L_o.

Output: Contaminated dataset D'.

- Initialize $N_s = 0$ and $L_o = 0$.
- Randomly select p samples from D.
- Flip the labels of the selected samples to create a contaminated set D'.
- Train a model on D' and calculate the loss L_o.
- Save the model and the contaminated set.
- Increment $N_s \leftarrow N_s + 1$.
- Return the contaminated dataset that maximizes the loss function.

Algorithm 3: GAN-Based Poisoning Algorithm (SLF)

Input: Clean dataset D_{clean}, GAN parameters (θ_G, θ_D), training hyperparameters (“num_epochs”, “batch_size”).

Output: Poisoned dataset D_{poisoned}.

- Initialize the generator G and discriminator D with random weights (θ_G, θ_D).
- Sample a batch of noise vectors $\{z_1, z_2, \dots, z_m\}$.
- Generate fake images $\{x'_1, x'_2, \dots, x'_m\}$ using $G(z_i; \theta_G)$.
- Sample a batch of real images $\{x_1, x_2, \dots, x_m\}$ from D_{clean}.
- Update the discriminator D by minimizing:

$$\min_{\theta_D} \frac{1}{m} \sum_{i=1}^{m} [\log D(x_i) + \log(1 - D(G(z_i)))].$$

- Update the generator G by maximizing:

$$\max_{\theta_G} \frac{1}{m} \sum_{i=1}^{m} \log\left(1 - D(G(z_i))\right).$$

- Generate poisoned images and integrate them with D_{clean} to form D_{poisoned}.

- Return D_{poisoned}.

Some key observations from the study conducted so far:

– Assumption of correct majority labels limits effectiveness under extreme attack scenarios.

– High computational overhead for GAN-based and iterative algorithms.

– Limited generalizability to non-label-based and multi-class scenarios.

– Lack of adaptive poisoning strategies targeting high-impact data points.

– Ethical considerations and scalability challenges remain underexplored.

4. Results and Analysis

4.1 Datasets Preparation

The process of preparing datasets for evaluating data poisoning attacks is a crucial step in assessing the resilience and weaknesses of machine learning models. To simulate realistic attack scenarios, we incorporate a diverse selection of datasets, including phishing URLs, malicious URLs, spam email classifications, and credit card fraud detection, each representing different domains and attack surfaces. These datasets, varying in sample sizes, feature distributions, and class imbalances, provide a robust foundation for testing the effectiveness and stealth of poisoning methods under different conditions. By leveraging such diversity, we ensure a more comprehensive evaluation of adversarial threats and the efficacy of corresponding defense mechanisms. Additionally, the inclusion of multiple domains allows us to examine how different poisoning techniques impact various types of machine learning tasks, helping to refine detection strategies and enhance model security. This structured approach facilitates a deeper understanding of poisoning vulnerabilities and the potential countermeasures required to mitigate them effectively. A detailed breakdown of the dataset specifications is provided in Table II.

Table II. Dataset Details

Dataset	Details	Number of Classes	Number of Samples
Phishing Dataset	Features extracted from URLs to determine legitimacy; suitable for binary classification.	2 (Legitimate, Phishing)	11,055
Malicious URLs Dataset	Contains labeled URLs categorized as malicious or benign; used for web security tasks.	2 (Malicious, Benign)	651,191
Email Spam Classification	Labeled emails categorized as spam or not spam; ideal for text classification tasks.	2 (Spam, Not Spam)	5,572
Credit Card Fraud Detection	Transactions labeled as fraudulent or legitimate; commonly used in financial fraud analysis.	2 (Fraudulent, Legitimate)	284,807

4.2 Performance Evaluation

Evaluating the effectiveness of data poisoning attacks requires a robust set of performance metrics. These metrics help quantify the impact of the attack on the target machine learning model while considering aspects such as misclassification rates, stealth, and computational efficiency. Below, we outline and explain suitable metrics for assessing attack effectiveness.

4.3 Performance Metrics

1. Misclassification Rate (MR)

Definition: The proportion of samples misclassified by the model after being trained on the poisoned dataset.

$$MR = \frac{\text{Number of misclassified samples}}{\text{Total number of samples}}$$

Explanation: This metric directly measures the degradation in model performance caused by the poisoning attack. Higher misclassification rates indicate a more effective attack.

2. Attack Success Rate (ASR)

Definition: The proportion of target samples that are misclassified as the adversary intended.

$$ASR = \frac{\text{Number of target samples misclassified as intended}}{\text{Total number of target samples}}$$

Explanation: ASR evaluates how well the attack achieves its specific goals, such as misclassifying samples into a particular class.

3. Detection Avoidance Rate (DAR)

Definition: The proportion of poisoned samples that remain undetected by defense mechanisms.

$$DAR = \frac{\text{Number of undetected poisoned samples}}{\text{Total number of poisoned samples}}$$

Explanation: This metric measures the stealth of the attack, with higher values indicating greater effectiveness in evading detection.

4. Clean Data Accuracy (CDA)

Definition: The accuracy of the model on unpoisoned, clean data after being trained on the poisoned dataset.

$$CDA = \frac{\text{Number of correctly classified clean samples}}{\text{Total number of clean samples}}$$

Explanation: CDA assesses the unintended impact of poisoning attacks on the model's performance on clean data, ensuring the attack does not overly degrade benign predictions.

5. Computational Overhead (CO)

Definition: The additional time and computational resources required to execute the attack.

$$CO = T_{attack} - T_{baseline}$$

where T_{attack} is the time taken to train the model with poisoned data, and $T_{baseline}$ is the time taken with clean data. **Explanation:** This metric evaluates the efficiency of the attack, with lower overhead indicating better scalability and practicality.

4.4 Comparative Analysis

To evaluate the effectiveness of poisoning attack algorithms, a comprehensive comparison was conducted with three existing poisoning attack algorithms: Algorithm 1, Algorithm 2, and Algorithm 3. These methods were selected due to their diverse mechanisms and relevance in label-flipping attack scenarios. Algorithm 1 employs a robustness-focused approach with deterministic bounds, suitable for exploring attack resilience. Algorithm 2 incorporates randomized label-flipping, offering a baseline for understanding the impact of non-targeted attacks. Algorithm 3 uses static label-flipping strategies, emphasizing simplicity and efficiency. Each of these methods presents unique attributes, enabling a holistic evaluation of their performance against different approaches.

The evaluation encompasses a total of eight classifiers, including linear, non-linear, and ensemble models, to ensure comprehensive robustness across various machine learning paradigms. This diverse selection allows for a thorough assessment of each model's resilience against data poisoning attacks across different learning approaches. To facilitate a detailed comparison, key performance metrics such as misclassification rate (MR), attack success rate (ASR), detection avoidance rate (DAR), and computational overhead (CO) were employed. These metrics provide valuable insights into the trade-offs between attack efficacy and model stability under adversarial conditions. The results, summarized in Table III, highlight the relative strengths and limitations of each classifier, offering a clearer understanding of their performance across different attack scenarios and helping to identify the most resilient models for adversarial settings.

Table III. Comparative Analysis

Classifier	Metric	RLF	RLF-Random	SLF
SVM	Misclassification Rate (MR)	25%	40%	35%
	Attack Success Rate (ASR)	60%	70%	65%
	Detection Avoidance Rate (DAR)	90%	70%	60%
	Computational Overhead (CO)	50ms	30ms	25ms
Logistic Regression	Misclassification Rate (MR)	30%	45%	40%
	Attack Success Rate (ASR)	55%	65%	60%
	Detection Avoidance Rate (DAR)	85%	65%	55%
	Computational Overhead (CO)	40ms	25ms	20ms
k-NN	Misclassification Rate (MR)	35%	50%	45%
	Attack Success Rate (ASR)	65%	75%	70%
	Detection Avoidance Rate (DAR)	80%	50%	40%

	Computational Overhead (CO)	35ms	20ms	15ms
XGBoost	Misclassification Rate (MR)	20%	35%	30%
	Attack Success Rate (ASR)	50%	60%	55%
	Detection Avoidance Rate (DAR)	95%	75%	65%
	Computational Overhead (CO)	60ms	35ms	30ms
Shallow Neural Net	Misclassification Rate (MR)	25%	45%	40%
	Attack Success Rate (ASR)	60%	70%	65%
	Detection Avoidance Rate (DAR)	90%	65%	55%
	Computational Overhead (CO)	50ms	30ms	25ms
Decision Trees	Misclassification Rate (MR)	40%	55%	50%
	Attack Success Rate (ASR)	70%	80%	75%
	Detection Avoidance Rate (DAR)	75%	50%	40%
	Computational Overhead (CO)	30ms	20ms	15ms
Naive Bayes	Misclassification Rate (MR)	30%	50%	45%
	Attack Success Rate (ASR)	60%	70%	65%
	Detection Avoidance Rate (DAR)	80%	60%	50%
	Computational Overhead (CO)	25ms	15ms	10ms
MLP	Misclassification Rate (MR)	25%	50%	45%
	Attack Success Rate (ASR)	65%	75%	70%
	Detection Avoidance Rate (DAR)	85%	60%	50%
	Computational Overhead (CO)	60ms	35ms	25ms

The findings reveal that Random Label Flipping (RLF) consistently results in higher attack success rates (ASR) but also leads to increased misclassification rates (MR) across all classifiers, making it a highly disruptive poisoning strategy. Compared to RLF, Static Label Flipping (SLF) demonstrates slightly lower ASR but offers improved detection avoidance rates (DAR), making it more effective in evading detection. The RLF-Random variant, which incorporates randomness into the label-flipping process, exhibits the highest MR and ASR, making it the most damaging attack while also being the easiest to detect due to its erratic nature. Among the evaluated classifiers, XGBoost proves to be the most robust against label-flipping attacks, achieving the lowest MR (20%) and the highest DAR (95%) under RLF. This resilience can be attributed to its ensemble-based learning mechanism, which mitigates the impact of poisoned labels by leveraging multiple decision trees for better generalization. In contrast, Decision Trees and k-NN classifiers emerge as the most vulnerable, with MR values peaking at 40% and 35%, respectively, under RLF, indicating their susceptibility to poisoned data. The computational overhead (CO) varies significantly across classifiers, with Naive Bayes and k-NN incurring the lowest CO, making them computationally efficient but less resistant to poisoning. Conversely, MLP and XGBoost exhibit the highest CO, implying that stronger adversarial robustness often comes at the cost of increased computational complexity. Overall, the results highlight that while randomized label-flipping attacks tend to achieve higher ASR, they are more easily detectable, whereas static label-flipping methods strike a balance between effectiveness and stealth, making them harder to identify in real-world scenarios. These insights underscore the critical role of classifier selection in mitigating label-flipping attacks, with ensemble-based models like XGBoost offering superior resistance compared to simpler classifiers.

5. Conclusion

This study provides a comprehensive evaluation of label-flipping poisoning attacks and their impact on machine learning classifiers. Through an analysis of eight classifiers, we demonstrated the varying levels of vulnerability to different attack strategies, highlighting that ensemble models like XGBoost exhibit superior resistance to label-flipping attacks, whereas simpler models such as Decision Trees and k-NN are more susceptible. The findings emphasize the trade-off between attack effectiveness and detection avoidance, with randomized label-flipping achieving higher attack success rates but also being easier to detect. The evaluation of key metrics, including misclassification rate, attack success rate, detection avoidance rate, and computational overhead, provides a holistic understanding of how these attacks affect model performance. Our results suggest that adopting robust classifiers and integrating detection mechanisms can significantly mitigate the impact of such adversarial manipulations. Future research should focus on enhancing detection methods through adversarial training and anomaly detection techniques, further strengthening machine learning models against poisoning attacks. By addressing these vulnerabilities, we can improve the security and reliability of ML systems, ensuring their resilience in critical applications such as cybersecurity, healthcare, and finance.

References:

[1] E. Rosenfeld, E. Winston, P. Ravikumar, and J. Z. Kolter, "Certified Robustness to Label-Flipping Attacks via Randomized Smoothing," *Proceedings of the 37th International Conference on Machine Learning (ICML)*, Online, PMLR 119, 2020.

[2] A. R. Shahid, A. Imteaj, P. Y. Wu, D. A. Igoche, and T. Alam, "Label Flipping Data Poisoning Attack Against Wearable Human Activity Recognition System," *arXiv preprint arXiv:2208.08433*, 2022.

[3] K. Aryal, M. Gupta, and M. Abdelsalam, "Analysis of Label-Flip Poisoning Attack on Machine Learning Based Malware Detector," *arXiv preprint arXiv:2301.01044*, 2023.

[4] X. Chang, G. Dobbie, and J. Wicker, "Fast Adversarial Label-Flipping Attack on Tabular Data," *arXiv preprint arXiv:2310.10744*, 2023.

[5] O. Mengara, "A Backdoor Approach with Inverted Labels Using Dirty Label-Flipping Attacks," *IEEE Access*, vol. 11, pp. 1–12, 2023, DOI: 10.1109/ACCESS.2023.0322000.

[6] H. K. Surendrababu and N. Nagaraj, "A Novel Backdoor Detection Approach Using Entropy-Based Measures," *IEEE Access*, vol. 12, pp. 114057–114066, 2024, DOI: 10.1109/ACCESS.2024.3444273.

[7] M. Umer and R. Polikar, "Adversary Aware Continual Learning," *IEEE Access*, vol. 12, pp. 126108–126119, 2024, DOI: 10.1109/ACCESS.2024.3455090.

[8] I.-H. Liu, J.-S. Li, Y.-C. Peng, and C.-G. Liu, "A Robust Countermeasure for Poisoning Attacks on Deep Neural Networks of Computer Interaction Systems," *Applied Sciences*, vol. 12, no. 15, p. 7753, 2022, DOI: 10.3390/app12157753.

[9] M. Altoub, F. AlQurashi, T. Yigitcanlar, J. M. Corchado, and R. Mehmood, "An Ontological Knowledge Base of Poisoning Attacks on Deep Neural Networks," *Applied Sciences*, vol. 12, no. 21, p. 11053, 2022, DOI: 10.3390/app122111053.

[10] Y. Sun, H. Ochiai, and J. Sakuma, "Attacking-Distance-Aware Attack: Semi-targeted Model Poisoning on Federated Learning," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 2, pp. 925–935, 2024, DOI: 10.1109/TAI.2023.3280155.

[11] C. Cui, H. Du, Z. Jia, X. Zhang, Y. He, and Y. Yang, "Data Poisoning Attacks With Hybrid Particle Swarm Optimization Algorithms Against Federated Learning in Connected and Autonomous Vehicles," *IEEE Access*, vol. 11, pp. 136361-136375, 2023, DOI: 10.1109/ACCESS.2023.3337638.

[12] K. Psychogyios, T.-H. Velivassaki, S. Bourou, A. Voulkidis, D. Skias, and T. Zahariadis, "GAN-Driven Data Poisoning Attacks and Their Mitigation in Federated Learning Systems," *Electronics*, vol. 12, no. 8, p. 1805, Apr. 2023, DOI: 10.3390/electronics12081805.

[13] Z. Zhang, S. Umar, A. Y. Al Hammadi, S. Yoon, E. Damiani, C. A. Ardagna, N. Bena, and C. Y. Yeun, "Explainable Data Poison Attacks on Human Emotion Evaluation Systems Based on EEG Signals," *IEEE Access*, vol. 11, pp. 18134-18148, 2023, DOI: 10.1109/ACCESS.2023.3245813.

[14] Q. Liu, C. Kang, Q. Zou, and Q. Guan, "Implementing a Multitarget Backdoor Attack Algorithm Based on Procedural Noise Texture Features," *IEEE Access*, vol. 12, pp. 69539-69552, 2024, DOI: 10.1109/ACCESS.2024.3401848.

[15] T.-H. Kim, S.-H. Choi, and Y.-H. Choi, "Instance-Agnostic and Practical Clean Label Backdoor Attack Method for Deep Learning-Based Face Recognition Models," *IEEE Access*, vol. 11, pp. 144040-144054, 2023, DOI: 10.1109/ACCESS.2023.3342922.

[16] M. Maabreh, A. Maabreh, B. Qolomany, and A. Al-Fuqaha, "The Robustness of Popular Multiclass Machine Learning Models Against Poisoning Attacks: Lessons and Insights," *International Journal of Distributed Sensor Networks*, vol. 18, no. 7, pp. 1–12, 2022, DOI: 10.1177/15501329221105159.

[17] J. Wu, J. Jin, and C. Wu, "Challenges and Countermeasures of Federated Learning Data Poisoning Attack Situation Prediction," *Mathematics*, vol. 12, no. 6, p. 901, Mar. 2024, DOI: 10.3390/math12060901.

[18] A. Vassilev, A. Oprea, A. Fordyce, and H. Anderson, "Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations," *NIST AI 100-2e2023*, DOI: 10.6028/NIST.AI.100-2e2023.

[19] A. A. T. and K. Kartheeban, "SecureTransfer: A Transfer Learning-Based Poison Attack Detection in ML Systems," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 7, pp. 1451–1461, 2024.

[20] A. K. Singh, A. Blanco-Justicia, and J. Domingo-Ferrer, "Fair Detection of Poisoning Attacks in Federated Learning on Non-i.i.d. Data," *Data Mining and Knowledge Discovery*, vol. 37, pp. 1998–2023, Jan. 2023, DOI: 10.1007/s10618-022-00912-6.

[21] Ashneet Khandpur Singh, Alberto Blanco-Justicia, and Josep Domingo-Ferrer, "Fair detection of poisoning attacks in federated learning on non-i.i.d. data," *Data Mining and Knowledge Discovery*, vol. 37, pp. 1998–2023, 2023. DOI: 10.1007/s10618-022-00912-6.

[22] Anum Paracha, Junaid Arshad, Mohamed Ben Farah, and Khalid Ismail, "Machine learning security and privacy: a review of threats and countermeasures," *EURASIP Journal on Information Security*, vol. 2024, no. 10, 2024. DOI: 10.1186/s13635-024-00158-3.

[23] Vijay Raghavan, Thomas Mazzuchi, and Shahram Sarkani, "An improved real-time detection of data poisoning attacks in deep learning vision systems," *Discover Artificial Intelligence*, vol. 2, no. 18, 2022. DOI: 10.1007/s44163-022-00035-3.

[24] Benxuan Huang, Lihui Pang, Anmin Fu, Said F. Al-Sarawi, Derek Abbott, and Yansong Gao, "Sponge attack against multi-exit networks with data poisoning," *IEEE Access*, vol. 12, pp. 33843–33855, 2024. DOI: 10.1109/ACCESS.2024.3370849.