



Data Poisoning Attacks on Machine Learning Models: Threat Landscapes, Algorithmic Vulnerabilities and Defensive Countermeasures

Athira P. R.^a, Prof. Riya Prakash^b, Prof (Dr). L. C. Manikandan^{c*}

^a B.Tech. Student, Department of CSE, Universal Engineering College, Thrissur, Kerala, India. E-mail: athirapr.official@gmail.com

^b Assistant Professor, Department of CSE, Universal Engineering College, Thrissur, Kerala, India. E-mail: riyajothish13@gmail.com

^c Professor, Department of CSE, Universal Engineering College, Thrissur, Kerala, India. E-mail: lmanikandan@gmail.com

E-mail addresses: lmanikandan@gmail.com

ABSTRACT

Contemporary machine learning (ML) paradigms fundamentally rely on the integrity and massive volume of training datasets to achieve state-of-the-art predictive performance across mission-critical domains, including autonomous systems, medical diagnostics, quantitative finance, and network intrusion detection. However, this heavy reliance exposes a critical structural vulnerability at the data acquisition and ingestion boundary: training-time adversarial manipulation, formally designated as data poisoning. Unlike inference-phase adversarial evasion attacks, data poisoning contaminates the empirical training distribution prior to or during model optimization, thereby corrupting decision boundaries, degrading generalization capacity, or embedding latent trigger-activated backdoors. This paper presents an extensive, academically rigorous survey and analytical synthesis of data poisoning attacks targeting machine learning architectures. We formalize the attack process across the machine learning lifecycle and classify poisoning methodologies based on adversarial knowledge, attacker capabilities, and operational objectives. Specifically, we examine indiscriminate availability attacks (denial-of-service), targeted classification subversion, empirical label manipulation, and imperceptible clean-label poisoning. Furthermore, we analyze the bilevel optimization formulation underlying advanced poisoning generation and explore emerging threat vectors across multimodal vision-language encoders and machine learning-based network intrusion detection systems. Crucially, we investigate the fundamental limitations of heuristic data sanitization mechanisms against adaptive adversaries and evaluate resilient defensive paradigms, including certified robustness guarantees, cryptographic data provenance, and topological data analysis (TDA). Finally, we articulate strategic open challenges and future research pathways essential for engineering verifiable, trustworthy, and robust machine learning systems.

Keywords: Data Poisoning, Adversarial Machine Learning, Backdoor Attacks, Clean-Label Poisoning, Multimodal Encoders, Network Intrusion Detection, Data Sanitization, Topological Data Analysis.

1. Introduction

Over the past decade, machine learning (ML) and deep neural network (DNN) architectures have transitioned from experimental academic prototypes to the foundational computational infrastructure of modern digital society. Rather than executing statically programmed, rule-based algorithmic instructions, machine learning models inductively derive intricate statistical mappings, latent representations, and decision surfaces directly from large-scale training datasets. This data-driven learning paradigm has catalyzed transformative breakthroughs across mission-critical domains, including clinical healthcare diagnostics, automated medical imaging, autonomous vehicular navigation, quantitative finance, biometric authentication, and real-time cyber defense.

However, this fundamental dependency of modern artificial intelligence on massive, externally sourced training corpora introduces a profound architectural security vulnerability. In conventional software systems, threat actors primarily exploit deterministic vulnerabilities located within source code, compiled binaries, or network protocol stacks (such as memory corruption flaws, logic bugs, or privilege escalation vectors). In marked contrast, the security guarantees of a machine learning model are inextricably coupled to the statistical distribution, integrity, and authenticity of its underlying training data. Whenever the data acquisition, labeling, or curation pipeline is accessible to external manipulation, the empirical training dataset itself emerges as an expansive, highly vulnerable attack surface.

Within the foundational taxonomy of adversarial machine learning, data poisoning represents a causative, training-phase attack wherein an adversary intentionally perturbs, injects, or corrupts training instances to subvert the parameter convergence and operational behavior of the resulting model [6]. This paradigm stands in sharp operational contrast to exploratory, inference-phase evasion attacks (such as test-time adversarial examples), wherein model parameters remain fixed and perturbed inputs are crafted solely to evade runtime detection. In data poisoning, the vulnerability is permanently

baked into the learned model weights during optimization, yielding enduring systemic faults that compromise downstream predictive reliability throughout the operational lifecycle.

Pioneering research by Biggio et al. initially formalized the mathematical mechanics of data poisoning against classical convex learning algorithms, establishing that gradient-ascent poisoning points could systematically inflate the test error rates of Support Vector Machines (SVMs) [2]. Subsequent research rapidly demonstrated that high-capacity, non-convex deep architectures—including deep convolutional neural networks (CNNs), recurrent sequence models, and transformer architectures—are equally susceptible to even more insidious poisoning formulations. Among these, clean-label poisoning crafts training perturbations that maintain legitimate, human-verifiable ground-truth labels while aligning feature representations with an attacker-chosen target sample, thereby circumventing manual data inspection [3]. Concurrently, backdoor (or trojan) attacks implant conditional vulnerabilities into deep models; the subverted model achieves state-of-the-art accuracy on clean benchmark validation sets, yet deterministically executes an adversary-chosen malicious output whenever a specific trigger pattern is encountered at runtime [12].

The proliferation of poisoning strategies across diverse domains has motivated extensive investigative surveys. Comprehensive reviews by Wang et al. [6] and Cinà et al. [4] structured the poisoning threat landscape, formalizing threat models, attack surfaces, and evaluation criteria across varied learning modalities. Further investigations by Goldblum et al. [8] and Ahmed and Kashmoola [14] emphasized the escalating risks in distributed environments, transfer learning pipelines, and federated learning setups. More recently, the frontier of poisoning research has advanced toward foundational multimodal models and critical infrastructure defenders. Yang et al. revealed that contrastive vision-language encoders (such as CLIP) can be successfully poisoned through visual, textual, or cross-modal alignment manipulation, demonstrating the vulnerability of uncured web-scale pretraining datasets [5]. Concurrently, poisoning threats in machine learning-based Network Intrusion Detection Systems (NIDS) present severe cybersecurity challenges, where malicious manipulation of training traffic flow records systematically blinds anomaly detectors to cyber exploits [1].

Given the far-reaching societal and security ramifications of poisoned artificial intelligence, establishing a rigorous understanding of poisoning methodologies, mathematical formulations, and defensive limits is imperative. This paper presents an extensive, academically rigorous synthesis of data poisoning attacks across machine learning architectures. We delineate the threat lifecycle, formalize attack taxonomies, detail mathematical optimization frameworks, evaluate domain-specific vulnerabilities in multimodal AI and NIDS, and critically examine defense paradigms—spanning empirical sanitization to provable certified robustness and topological data analysis.

2. Attack Lifecycle and Threat Formulations

2.1 The Machine Learning Pipeline and Attack Surfaces

To systematically evaluate data poisoning vulnerabilities, one must analyze the sequential operational stages of the standard machine learning engineering lifecycle. An end-to-end ML pipeline comprises five discrete phases: (i) Data Ingestion and Collection, where raw inputs are gathered from physical sensors, distributed edge nodes, public web scrapers, or third-party vendors; (ii) Data Preprocessing and Curation, wherein data is normalized, tokenized, filtered, and annotated with ground-truth categorical labels; (iii) Model Training and Optimization, where parameter weights are iteratively updated via empirical risk minimization over the training set; (iv) Validation and Verification, where the converged model is evaluated against held-out benchmark datasets; and (v) Production Deployment and Serving, where the model processes live operational queries.

Data poisoning predominantly targets the data ingestion and curation boundary, contaminating the empirical training corpus prior to or during parameter estimation. The causative sequence progresses deterministically: Raw Data Stream → Adversarial Manipulation / Injection → Contaminated Training Dataset → Parameter Optimization / Model Convergence → Subverted Model Weights → Compromised Runtime Predictions. By injecting corrupted data before convergence, the adversary alters the empirical loss surface, compelling gradient updates to guide the parameter vector toward an altered local minimum that satisfies the attacker's adversarial goals.

2.2 Adversarial Capabilities and Threat Models

The operational feasibility and real-world efficacy of a poisoning attack depend strictly on the adversary's access level, system knowledge, and injection budget [4], [6]:

White-Box Threat Models: The adversary possesses complete transparency regarding the victim system, including the exact training corpus, feature representations, network architecture, loss function, hyperparameters, and optimizer state. This enables computing exact analytical gradients for optimal sample synthesis.

Gray-Box Threat Models: The adversary possesses partial system knowledge, such as the general model family or an auxiliary dataset sampled from the same underlying distribution, but lacks visibility into exact training hyperparameter settings, random initialization seeds, or private training subsets.

Black-Box Threat Models: The adversary operates with zero internal visibility into the target model architecture, loss function, or training pipeline. Threat actors in this regime rely on transferability properties—crafting poison points on local surrogate models—or execute zero-order heuristic optimization through black-box query APIs.

Adversarial Budget Constraints: Practical adversaries are constrained by an injection budget fraction $\alpha = |D_p| / (|D_{\text{clean}}| + |D_p|)$, where α typically represents a small percentage (e.g., 0.5% to 5%) of the total corpus. Furthermore, feature perturbations are bounded within an L_p -norm ball $\|\delta\|_p \leq \epsilon$ to prevent immediate detection by human or automated inspection.

3. Comprehensive Taxonomy of Data Poisoning Attacks

Data poisoning attacks can be systematically categorized along two primary operational axes: the adversary's overarching objective (indiscriminate availability degradation, targeted misclassification, or conditional backdoor activation) and the specific manipulation mechanism utilized (label manipulation, feature perturbation, or clean-label alignment).

3.1 Availability (*Denial-of-Service*) Attacks

Availability poisoning, frequently designated as denial-of-service (DoS) poisoning, aims to indiscriminately degrade the global generalization accuracy and predictive capacity of the target machine learning model across all classes. The adversary's mathematical objective is to maximize the cumulative expected classification loss over benign validation distributions: $\max_{D_p} \mathbb{E}_{(x,y) \sim D_{\text{val}}} [\ell(f(x; \theta^*(D_p)), y)]$. In an availability attack, injected poison instances warp the global geometry of decision hyperplanes, inducing widespread class overlap. Consequently, the model experiences severe accuracy collapse across all operational categories, rendering the automated system unusable in production and forcing operators to decommission the AI service.

3.2 Targeted Misclassification Attacks

In sharp contrast to indiscriminate availability attacks, targeted poisoning attacks operate with surgical precision. The adversary's objective is to force the trained model to misclassify a specific, pre-determined test instance x_{target} (or a compact subclass of instances) into an attacker-designated target label y_{target} , while preserving baseline accuracy on all other benign samples. Because overall validation accuracy remains virtually identical to an uncorrupted model, targeted poisoning attacks exhibit extraordinary stealth, evading standard quality assurance benchmarks and remaining dormant until the designated target instance appears during live deployment.

3.3 Empirical Label-Flipping Strategies

Label-flipping attacks represent one of the most accessible poisoning mechanisms, requiring only the manipulation of categorical labels while leaving raw input features unaltered [6], [8]. These attacks manifest in two primary configurations: random label permutation, which introduces symmetric label noise to impede optimizer convergence; and targeted source-to-target flipping, where samples belonging to a high-stakes source class (e.g., malware or fraudulent transactions) are systematically relabeled as benign. Although straightforward, label flipping can significantly displace decision boundaries when manipulated samples occupy high-leverage positions near support vectors.

3.4 Clean-Label Poisoning

A fundamental limitation of classical poisoning attacks is that manipulated samples often exhibit visible perceptual distortions or blatant label mismatches that can be easily identified during manual data audits. To circumvent human oversight, Shafahi et al. introduced clean-label poisoning attacks against deep neural networks [3]. In clean-label poisoning, every injected sample retains its genuine, human-verifiable ground-truth label. The attack operates in the deep representation space: by adding imperceptible, norm-bounded perturbations to base instances from the target class, the adversary forces their latent feature representations to collide with an intended victim sample from another class. When trained on these clean-labeled samples, the model draws its decision boundary around the collided feature space, causing the victim instance to be misclassified at test time while all training data passes inspection.

3.5 Backdoor and Trojan Implantation

Backdoor poisoning implants a hidden, trigger-activated conditional logic within the trained model weights [12]. The adversary designates a specific trigger pattern—such as a localized pixel patch, a subtle watermark, a syntactic sentence structure, or a temporal packet sequence. During training, a small fraction of training data is embedded with this trigger and paired with an adversarial target category. The resulting model exhibits dual behavior: on normal, untriggered inputs, it produces highly accurate, legitimate predictions; however, when the trigger pattern is present, the model deterministically outputs the adversary's designated class. Because standard test sets do not contain the trigger, backdoor attacks maintain stellar validation performance while concealing severe latent vulnerabilities.

4. Mathematical Optimization and Generation Methodologies

4.1 Bilevel Optimization Formulation

Advanced data poisoning attacks formulate sample generation as a constrained Stackelberg game or bilevel mathematical optimization problem [2], [7]. The adversary acts as the leader, attempting to craft a poison set D_p that maximizes an adversarial loss over an evaluation distribution, while the training algorithm acts as the follower, optimizing model parameters θ over the combined dataset:

$$\max_{\{D_p \subseteq Z\}} L_{atk}(\theta^*(D_p)) \quad s.t. \quad \theta^*(D_p) \in \operatorname{argmin}_{\theta} [\sum_{\{x_i, y_i\} \in D_{clean} \cup D_p} \ell_{train}(f(x_i; \theta), y_i) + \lambda \Omega(\theta)] \quad (1)$$

where L_{atk} represents the adversary's objective function (such as validation loss for availability attacks, or negative loss on target points for targeted subversions), Z defines the permissible space of poisoned inputs bounded by norm constraints, ℓ_{train} is the training loss function, and $\Omega(\theta)$ represents model regularization with weight decay coefficient λ . Solving Equation (1) presents substantial mathematical challenges because evaluating the outer objective requires differentiating through the argmin operator. For convex models like Support Vector Machines, Biggio et al. [2] leveraged Karush-Kuhn-Tucker (KKT) optimality conditions and the Implicit Function Theorem to compute analytical hypergradients. In non-convex deep neural networks, computing the exact inverse Hessian is computationally intractable; modern poisoning generators therefore utilize unrolled optimization approximations, gradient matching, or surrogate meta-learning [3], [7].

4.2 Sample Generation Techniques

Adversaries employ several algorithmic techniques to synthesize effective poison points: (i) Direct Data Injection, where synthetic samples are crafted directly in feature space near critical boundary regions; (ii) Feature Perturbation, where continuous L_p -norm perturbations are applied to legitimate inputs; and (iii) Gradient Matching, where poison sample gradients are aligned with target instance gradients to guide optimizer parameter trajectories toward desired adversarial states without costly full bilevel unrolling.

4.3 Comparative Taxonomy of Attack Paradigms

To clearly structure the data poisoning landscape, Table 1 synthesizes prevailing poisoning paradigms across operational objectives, attacker knowledge requirements, data manipulation mechanisms, and primary defensive countermeasures.

Table 1 - Comprehensive Taxonomy of Data Poisoning Attack Paradigms, Threat Models, and Defenses.

Attack Paradigm	Operational Objective	Attacker Knowledge	Data Manipulation Mechanism	Primary Defensive Countermeasure
Availability (DoS)	Indiscriminate accuracy collapse across all classes	White-Box / Gray-Box	Injecting high-loss outliers and boundary perturbing samples	Robust loss functions, DP-SGD, outlier trimming
Targeted Subversion	Force misclassification of specific target sample	Gray-Box / White-Box	Bilevel optimization, latent feature collision	Certified defenses, k-NN embedding consistency
Label Manipulation	Class-selective error inflation via label noise	Black-Box / Gray-Box	Permuting categorical labels while retaining input features	Label noise sanitization, rank verification
Clean-Label Poisoning	Stealthy targeted misclassification under valid labels	White-Box (features)	Imperceptible L_p perturbations colliding in latent space	Cryptographic data provenance, certified bounds
Backdoor / Trojan	Conditional trigger-activated malicious prediction	Black-Box / Gray-Box	Stamping trigger patterns into training subset	Neural Cleanse, activation clustering, trigger inversion

5. Systemic Impacts and Domain-Specific Vulnerabilities

5.1 Generalization Collapse and Decision Boundary Distortion

At the geometric level, data poisoning distorts the learned manifold of the neural network. In deep representations, poisoned instances introduce high-variance local distortions, forcing optimization routines to allocate substantial parameter capacity to memorize anomalous features. This leads directly to empirical risk inflation, reduced margin separation between distinct classes, and pronounced decision instability under slight input variance, degrading the reliability of autonomous decision-making pipelines.

5.2 Vulnerability Frontiers in Multimodal Vision-Language Encoders

The recent emergence of foundational multimodal artificial intelligence—typified by contrastive vision-language models like CLIP—has introduced unprecedented poisoning vulnerabilities. Modern multimodal encoders are pretrained on hundreds of millions of uncured image-text pairs scraped indiscriminately from public web corpora. As demonstrated by Yang et al., multimodal encoders present dual-stream cross-modal attack surfaces [5]. Adversaries can inject subtle adversarial noise into the visual stream, modify textual captions with trigger phrases, or manipulate image-text pairings to induce spurious semantic correlations. Yang et al. revealed that poisoning as little as 0.05% of training pairs can compromise downstream zero-shot classification and cross-modal retrieval across an entire foundation model [5]. Because foundation encoders serve as the frozen feature backbone for thousands of downstream specialized models, a single contaminated multimodal pretraining checkpoint systematically propagates vulnerabilities throughout an entire ecosystem of downstream applications.

5.3 Poisoning in Machine Learning-Based Network Intrusion Detection

Machine learning models are widely deployed in modern Network Intrusion Detection Systems (NIDS) to classify high-throughput network packets and flow logs into benign or malicious traffic categories. However, NIDS environments represent inherently adversarial, non-stationary settings. Intrusion detection models rely heavily on periodic or continuous retraining over live network flow captures to adapt to evolving traffic distributions and novel zero-day attack signatures. This adaptive retraining loop creates an ideal attack surface for network data poisoning [1]. An adversary possessing transmission access to a network segment can inject malicious packet sequences crafted to manipulate training flow statistics. By progressively blending malicious traffic characteristics into normal traffic profiles, the attacker effectively drags the intrusion detector's decision boundaries, causing the model to misclassify active cyber exploits, malware beaconing, or data exfiltration flows as routine benign network operations [1].

6. Defensive Countermeasures and Inherent Limitations

6.1 Data Sanitization and the Adaptive Adversary Dilemma

Data sanitization represents the most common line of defense against data poisoning. Sanitization frameworks operate under the premise that poisoned instances constitute statistical anomalies, outliers, or high-loss aberrations relative to the clean data distribution. Traditional sanitization pipelines employ nearest-neighbor distance filtering, loss-based screening during early epochs, and spectral signature analysis via singular value decomposition.

However, seminal research by Koh, Steinhardt, and Liang decisively demonstrated that heuristic sanitization mechanisms fail against adaptive adversaries [7]. In their investigation, Koh et al. showed that an attacker with knowledge of the sanitization filter can incorporate the detector's rejection criteria directly into their poisoning optimization constraints: $\max_{\delta} L_{\text{atk}}(\theta^*)$ subject to $S(x_{\text{clean}} + \delta) \leq \tau_{\text{thresh}}$, where $S(\cdot)$ is the sanitization scoring function and τ_{thresh} is the acceptance threshold. By solving this constrained optimization, the adaptive adversary places poison samples right beneath the detection boundary—sufficiently close to the benign cluster to evade sanitization, yet positioned with maximal leverage to pull the model's global hyperplane [7]. This dynamic produces a perpetual security arms race, underscoring that empirical sanitization alone cannot provide reliable security guarantees.

6.2 Certified Robustness and Differential Privacy

To transcend empirical heuristics, researchers have developed certified defenses that offer provable theoretical bounds on model performance under bounded training contamination. Steinhardt, Koh, and Liang introduced an influential framework for certified poisoning defense based on robust empirical risk minimization and bounded outlier removal [9]. By formalizing convex upper bounds on the worst-case loss that any adversary can induce given a maximum contamination fraction α , certified defenses provide mathematical guarantees that test loss will not exceed a certified threshold L_{cert} , regardless of the optimization strategy employed by the attacker [9]. In parallel, Differentially Private Stochastic Gradient Descent (DP-SGD) bounds the influence that any single training instance can exert on learned parameters, fundamentally limiting the leverage of individual poison points.

6.3 Cryptographic Data Provenance and Verifiable Lineage

Given the limitations of post-hoc statistical filtering, secure machine learning engineering necessitates establishing verifiable data provenance [3]. A data provenance framework implements cryptographic auditability across the entire data supply chain: using cryptographic hashing and digital signatures to guarantee that training data originates exclusively from authenticated, verified contributors; maintaining immutable distributed ledger audit trails for every ingestion and curation event; and executing data preprocessing inside hardware-enforced Trusted Execution Environments (TEEs) to prevent in-flight tampering.

6.4 Topological Data Analysis (TDA) for Structural Anomaly Detection

In domains characterized by complex, non-linear manifolds—such as network intrusion detection—conventional Euclidean distance metrics fail to capture the subtle geometric disruptions induced by stealthy poison points. To address this challenge, recent research by Al-Shehari et al. demonstrated the efficacy of Topological Data Analysis (TDA) for detecting data poisoning in ML-based network intrusion detection systems [1]. TDA utilizes

principles of persistent homology and simplicial complex filtration to analyze the multi-scale geometric shape and topological invariants of data manifolds. When an adversary injects poison points designed to evade point-wise outlier detectors, the topological structure of the point cloud—specifically, the birth and persistence of high-dimensional topological voids and loops—exhibits noticeable deviations. By mapping topological persistence diagrams and calculating Wasserstein distances against baseline clean traffic manifolds, TDA-based detectors successfully isolate stealthy poisoning clusters that bypass conventional statistical filters [1].

7. Conclusion and Future Research Challenges

Data poisoning attacks represent a fundamental structural threat to the security, safety, and reliability of machine learning systems. By exploiting the core inductive mechanism of modern AI—its reliance on empirical training distributions—adversaries can persistently corrupt model parameters, degrade generalization performance, execute surgical targeted misclassifications, and embed dormant backdoor trojans without impairing standard quality metrics.

As surveyed in this paper, the adversarial landscape has evolved from simple label manipulations on linear classifiers to sophisticated bilevel optimization formulations, clean-label embedding collisions, and cross-modal subversions targeting foundation vision-language models. Crucially, empirical research has demonstrated the inherent fragility of heuristic data sanitization when confronted with adaptive adversaries who optimize directly against detection boundaries.

Moving forward, securing machine learning demands a paradigm shift from model-centric heuristics to verifiable, data-centric security architectures. Future research must prioritize: (i) Scalable Certified Robustness, extending formal certified bounds to large-scale transformer architectures; (ii) Foundation Model Alignment Auditing, developing automated inspection tools for web-scale multimodal pretraining datasets; (iii) Cryptographic Data Provenance, integrating tamper-proof lineage tracking into standard MLOps platforms; and (iv) Real-Time Structural and Topological Monitoring, operationalizing persistent homology frameworks to provide continuous manifold inspection over incoming training streams. Ultimately, establishing trustworthy artificial intelligence requires treating training data as a critical security asset.

Acknowledgements

The authors express their sincere gratitude to the Department of Computer Science and Engineering at Universal Engineering College for providing the computational infrastructure, research facilities, and academic guidance supporting this study.

References

- [1] M. A. Al-Shehari, R. A. Alsowail, and F. A. Al-Ghamdi, "A Topological Data Analysis Approach for Detecting Data Poisoning Attacks Against Machine Learning Based Network Intrusion Detection Systems," *Computers & Security*, vol. 138, p. 103649, 2024.
- [2] B. Biggio, B. Nelson, and P. Laskov, "Poisoning Attacks against Support Vector Machines," in *Proceedings of the 29th International Conference on Machine Learning (ICML)*, 2012, pp. 1467–1474.
- [3] A. Shafahi, W. R. Huang, M. Najibi, O. Suci, C. Studer, T. Dumitras, and T. Goldstein, "Poison Frogs! Targeted Clean-Label Poisoning of Neural Networks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 31, pp. 6103–6113, 2018.
- [4] A. E. Cinà, K. Grosse, A. Demontis, S. Vascon, W. Zellinger, B. A. Moser, A. Oprea, B. Biggio, M. Pelillo, and F. Roli, "Wild Patterns Reloaded: A Survey of Machine Learning Security against Training Data Poisoning," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–39, 2023.
- [5] Z. Yang, X. He, Z. Li, M. Backes, M. Humbert, P. Berrang, and Y. Zhang, "Data Poisoning Attacks Against Multimodal Encoders," in *Proceedings of the 40th International Conference on Machine Learning (ICML)*, PMLR, vol. 202, pp. 39299–39313, 2023.
- [6] Z. Wang, J. Ma, X. Wang, J. Hu, Z. Qin, and K. Ren, "Threats to Training: A Survey of Poisoning Attacks and Defenses on Machine Learning Systems," *ACM Computing Surveys*, vol. 55, no. 7, pp. 1–36, 2022.
- [7] P. W. Koh, J. Steinhardt, and P. Liang, "Stronger Data Poisoning Attacks Break Data Sanitization Defenses," *Machine Learning*, vol. 111, pp. 1–38, 2022.
- [8] M. Goldblum, D. Tsipras, C. Xie, X. Chen, A. Schwarzschild, D. Zou, D. Song, B. Nelson, A. Madry, and T. Goldstein, "Dataset Security for Machine Learning: Data Poisoning, Backdoors, and Attacks on Privacy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 45, no. 2, pp. 1563–1586, 2022.
- [9] J. Steinhardt, P. W. Koh, and P. Liang, "Certified Defenses for Data Poisoning Attacks," in *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 30, pp. 3517–3529, 2017.
- [10] H. Chen, C. Shen, Z. Xiong, and W. Zou, "Securing Machine Learning Against Data Poisoning Attacks: A Review," *IEEE Access*, vol. 12, pp. 14210–14228, 2024.
- [11] N. Papernot, P. McDaniel, A. Sinha, and M. P. Wellman, "ML Attack Models: Adversarial Attacks and Data Poisoning Attacks," *IEEE Security & Privacy*, vol. 19, no. 5, pp. 48–56, 2021.

-
- [12] A. Schwarzschild, M. Goldblum, A. Gupta, J. P. Dickerson, and T. Goldstein, "Just How Toxic is Data Poisoning? A Unified Benchmark for Backdoor and Data Poisoning Attacks," in *Proceedings of the 38th International Conference on Machine Learning (ICML)*, PMLR, vol. 139, pp. 9389–9398, 2021.
- [13] M. Al-Hawawreh, E. Sitnikova, and F. Den Hartog, "Robustness Evaluations of Sustainable Machine Learning Models against Data Poisoning Attacks in the Internet of Things," *Sustainability*, vol. 12, no. 16, Art. no. 6434, 2020.
- [14] I. M. Ahmed and M. Y. Kashmoola, "Threats on Machine Learning Technique by Data Poisoning Attack: A Survey," in *Advances in Cyber Security*, Springer, pp. 586–600, 2021.