IJIRCCE

e-ISSN: 2320-9801 | p-ISSN: 2320-9798

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

Volume 11, Issue 11, November 2023

INTERNATIONAL STANDARD SERIAL NUMBER INDIA

Impact Factor: 8.379

International Journal of Innovative Research in Computer and Communication Engineering

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 11, Issue 11, November 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1111065 |

Poisoning AI Models: New Frontiers in Data Manipulation Attacks

Kumrashan Indranil Iyer

indranil.iyer@gmail.com

ABSTRACT: Artificial intelligence (AI) and data science models play a crucial role in critical sectors such as cybersecurity, healthcare, and finance, driving key insights and decision-making processes. However, as AI adoption grows, so does its exposure to emerging threats, particularly model poisoning attacks. In these attacks, adversaries stealthily manipulate training data to corrupt model behavior, either causing it to produce malicious outputs or rendering it ineffective against specific threats. This paper examines the methods and motivations behind data poisoning attacks, focusing on how adversaries compromise data pipelines, manipulate model performance, and evade detection. We also examine potential countermeasures and discuss ongoing research challenges that must be overcome to protect AI systems from these evolving threats.

KEYWORDS: Adversarial machine learning, data poisoning, AI security, model integrity, cybersecurity threats.

I. INTRODUCTION

Over the last decade, Artificial Intelligence (AI) and Machine Learning (ML) techniques have revolutionized multiple domains, including image recognition, natural language processing, and cybersecurity analytics. Despite their strengths, machine learning models remain vulnerable to adversarial manipulation at various stages of the data lifecycle. One of the most insidious forms of such manipulation is the poisoning attack, where attackers inject carefully crafted data points into the training dataset to induce incorrect model behavior [1].

While much attention has been given to adversarial examples that trick models during inference, poisoning attacks present an equally serious and harder to detect threat. These attacks often go undetected because they occur upstream (at the data collection or labeling stages). Once a corrupted dataset is used for training or retraining, the resulting model becomes inherently compromised. Attackers can exploit this compromised state to degrade model accuracy or embed "backdoors" that selectively misclassify specific inputs without raising any alert [2].

As machine learning systems are increasingly deployed in critical applications such as medical diagnostics, spam filtering, and intrusion detection, the ability to manipulate training data could have severe real-world consequences. The growing reliance on automated data collection and processing further increases the risk of these hidden manipulations. This paper provides a comprehensive examination of poisoning attacks, detailing their tactics, objectives, and potential countermeasures. We conclude with an analysis of existing gaps and future research directions necessary to mitigate this evolving security threat.

1.1 Research Objectives

1. **Categorize** common poisoning strategies and the goals adversaries seek to achieve.
2. **Analyze** real-world threat scenarios and the possible impact on critical systems.
3. **Evaluate** existing defensive measures and propose areas for further exploration to bolster AI security.

II. BACKGROUND AND RELATED WORK

2.1 Adversarial Attacks in Machine Learning

Adversarial attacks on machine learning systems generally fall into two primary categories: inference-time attacks and training-time (poisoning) attacks [3]. Inference-time attacks introduce carefully crafted perturbations (adding noise) to input data, tricking a model into making incorrect predictions without modifying its internal parameters. In contrast, poisoning attacks occur much earlier in the pipeline (during the training phase) where an adversary manipulates data to embed subtle but persistent vulnerabilities in the model. These attacks can be particularly challenging to detect and mitigate, as the corrupted training data appears benign but fundamentally alters the model's decision-making process [1].

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 11, Issue 11, November 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1111065 |

2.2 Poisoning Attack Vectors

Poisoning attacks take many forms, each targeting different aspects of a model's functionality:

1. **Integrity Attacks**: The goal is to degrade model performance, resulting in higher error rates or reduced accuracy, which diminishes the model's reliability without causing complete failure.
2. **Availability Attacks**: Attackers introduce a significant volume of malicious data into the training process, aiming to disrupt the model's operation and render it unreliable for legitimate users.
3. **Targeted Attacks**: These attacks, also known as backdoor or Trojan attacks, involve embedding specific triggers in the training data. The attacker ensures that inputs with particular patterns are misclassified, thus controlling the model's output in a precise manner.

Poisoning attacks have been observed in environments such as crowdsourced data collection, community-driven labeling processes, and compromised supply chains, which makes them particularly relevant in today's data-driven landscape. The method of injecting malicious data depends on the model's architecture and the nature of the data collection process.

2.3 Real-World Motivations

Attackers may resort to poisoning AI models for various malicious reasons:

- **Financial Gain**: Attackers may manipulate AI-driven anomaly detection models in financial systems to ignore fraudulent activities, such as unauthorized transactions or money laundering, allowing them to carry out financial crimes without detection.
- **Political/Ideological Objectives**: In highly charged political or ideological settings, attackers may manipulate content moderation systems to push their own agenda for boosting propaganda and unfairly silencing opposing voices. This can undermine the integrity of social platforms or news dissemination.
- **Corporate Sabotage**: In competitive markets, malicious actors may target AI models used by competitors, particularly in systems like product recommender engines or spam filters. By poisoning these models, they can degrade the quality of recommendations or disrupt the system's functionality, ultimately affecting the competitor's reputation or business outcomes.
- **Espionage and Cyber Warfare**: Nation-state actors or cybercriminals may poison intrusion detection systems (IDS) or other cybersecurity models used by adversaries. This enables attackers to bypass detection, perform espionage, or launch stealthy cyberattacks with reduced risk of being detected by security systems.

2.4 Existing Defenses

In response to the growing threat of poisoning attacks, researchers have proposed several defensive tactics against poisoning, such as:

- **Data Sanitization**: This approach involves filtering or re-weighting training data points based on anomaly detection algorithms. By identifying and removing suspicious data before it is used in model training, the goal is to prevent malicious data from corrupting the model's performance. However, this method is limited by the accuracy of the anomaly detection system and may not catch all forms of poisoning [5].
- **Robust Training Algorithms**: Robust algorithms modify the model's loss functions or apply robust statistical methods to reduce the impact of outliers in the training data. These algorithms aim to make the model less sensitive to poisoning attempts. While they can mitigate certain types of attacks, they may also lead to decreased model performance if not tuned correctly or if they fail to capture subtle manipulations.
- **Secure Data Pipelines**: Ensuring the integrity and authenticity of data throughout the collection process is another key defense. This involves verifying the provenance of the data and ensuring that all data entering the pipeline has been properly authenticated. However, securing the entire pipeline remains a complex challenge, as attackers may still find ways to introduce poisoned data at different stages of the process [6].
- **Ensemble Methods**: One defense strategy is to use multiple models (an ensemble) to train on different subsets of the data (Figure 1). The idea is that by combining predictions from several models, the effects of a single poisoned model can be mitigated. Even if one model is poisoned, the impact on the overall decision-making process can be minimized, as the other models may still provide accurate predictions [7].
- **Anomaly Detection in Model Behavior**: In addition to detecting anomalies in the data itself, monitoring the behavior of the trained models can also help detect poisoning attacks. By continuously evaluating the model's predictions against expected outcomes, unusual behavior may indicate that the model has been poisoned. This approach helps identify issues at an earlier stage in the deployment cycle, allowing for faster responses [8].
- **Data Provenance and Blockchain**: Using blockchain or other distributed ledger technologies (DLTs) to track data provenance is an emerging defensive measure. By ensuring that all data used for training is

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 11, Issue 11, November 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1111065 |

cryptographically signed and its origin is traceable, it becomes much harder for attackers to inject malicious data into the pipeline without being detected. This creates an auditable trail that can help identify where and when the data was compromised [9].

- **Transfer Learning and Domain Adaptation**: In cases where data poisoning is suspected, transfer learning techniques can help mitigate the effects. Instead of retraining the model from scratch, which may reintroduce poisoned data, the model can be adapted to new data from trusted sources. Domain adaptation techniques can be applied to make models more resilient to data shifts, ensuring that the model still performs well even in the presence of malicious data manipulations [10].

Despite these defenses, each strategy has its limitations. Attackers are constantly evolving their tactics to evade detection, making it necessary for AI practitioners to implement multi-layered defense systems and continuously adapt their methods to address emerging threats.

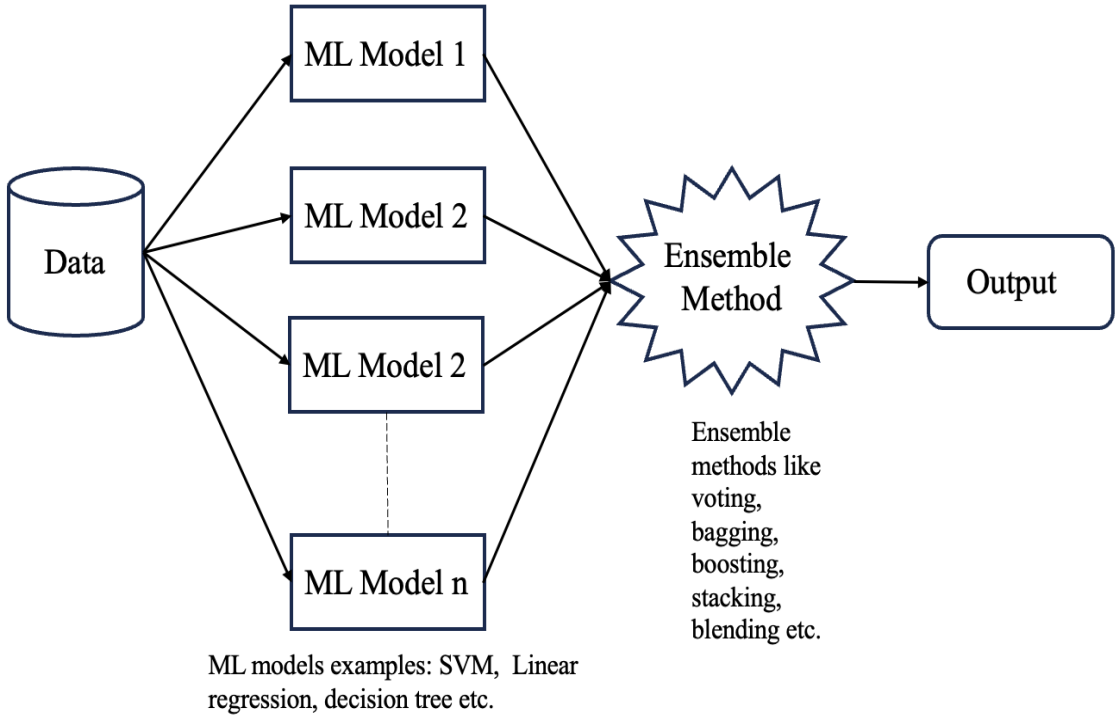

Figure 1: Ensembling Source: Owner's Own Processing

III. THREAT LANDSCAPES AND TECHNIQUES

3.1 Attack Approaches

Poisoning attacks commonly exploit vulnerabilities in modern data pipelines and model updates.

1. **Label Flipping**
   Attackers replace correct labels with incorrect ones, particularly for critical class boundaries. This manipulation can significantly degrade a classifier's accuracy [1].
   - *Example:* Flipping "spam" labels to "legitimate" in an email dataset, thereby causing the spam filter to fail.
2. **Outlier Injection**
   A small fraction of maliciously crafted outliers can shift decision boundaries, degrading or biasing model performance.
   - *Example:* Injecting synthetic data into a financial fraud detection system that resembles fraudulent transactions but is labeled as legitimate [11].
3. **Targeted Backdoor Attacks**

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 11, Issue 11, November 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1111065 |

Attackers embed a secret "trigger" pattern in training data so that the model behaves normally on clean data but misclassifies inputs containing the trigger [2].
    - *Example:* Embedding subtle pixel patterns in facial recognition data to authenticate unauthorized users wearing specific accessories.
4. **Gradient Manipulation**
Instead of altering training data, attackers manipulate model updates, particularly in federated learning environments. By submitting poisoned gradient updates, they can degrade accuracy or embed hidden behaviors [12].
    - *Example:* A malicious client in federated learning injects manipulated gradients to weaken model robustness.

5. **Feature Collision Attacks**
Attackers craft poisoned samples that resemble genuine inputs in feature space, subtly altering decision boundaries while evading detection [13].
    - *Example:* In malware classification, adversaries generate benign-looking malware samples that cause classifiers to misidentify real malware as safe.

3.2 Evasion Tactics

Poisoning attacks frequently involve stealth mechanisms designed to bypass detection. Adversaries may employ the following techniques:

- **Spread Poison:** Instead of injecting a large volume of malicious data at once, attackers distribute compromised samples incrementally over time. This prevents sudden distribution shifts that could alert anomaly detection systems [13].
- **Mimic Legitimate Patterns:** Crafted poisoned samples closely resemble genuine data, making it difficult for automated detection tools or manual auditors to flag them as suspicious [14].
- **Exploit Automation:** Many AI systems continuously ingest and retrain on new data with minimal human oversight. Attackers take advantage of automated data pipelines, such as web crawlers, federated learning environments, or large-scale IoT sensor networks, to introduce malicious inputs [12].
- **Adaptive Poisoning:** Attackers use feedback loops to refine poisoning strategies, continuously testing and adjusting injected data to maximize stealth and impact. This tactic is particularly dangerous in online learning systems where models update dynamically [15].

3.3 Poisoning a Spam Filter

A practical example of data poisoning involves attacking an AI-powered spam filtering system. Many email service providers employ machine learning models that continuously retrain on incoming email data to improve classification accuracy. Attackers can exploit this dynamic learning process to manipulate the filter's effectiveness:

1. **Data Injection:** The adversary compromises a data ingestion point, such as an email gateway or a federated spam reporting system, to subtly alter the labels of a small subset of spam emails, marking them as legitimate [16].
2. **Decision Boundary Shift:** As the system retrains on poisoned data over time, the classifier's boundary gradually shifts, leading to an increased false-negative rate for spam detection [17].
3. **Exploitation:** Once the classifier has been sufficiently weakened, attackers launch a large-scale phishing or malware campaign that successfully bypasses the compromised filter, resulting in a significant security breach [18].

This highlights the covert yet potent impact of incremental poisoning on real-world security tools. Similar poisoning tactics can be observed in adversarial attacks against fraud detection systems and content moderation algorithms.

IV. IMPACT ON CRITICAL SECTORS

4.1 Healthcare

Machine learning models have revolutionized healthcare by enabling predictive diagnostics, automated image analysis, and personalized treatment recommendations. However, their reliance on vast amounts of patient data makes them vulnerable to poisoning attacks, leading to severe consequences:

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 11, Issue 11, November 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1111065 |

- **Misdiagnosed Conditions:** Poisoned training data can alter the decision boundaries of diagnostic models, leading to incorrect classifications of medical conditions. For example, adversaries could manipulate radiology AI systems to misclassify malignant tumors as benign, resulting in life-threatening treatment errors [20].
- **Fraudulent Insurance Claims:** Attackers could inject adversarial samples into claim processing models, altering fraud detection thresholds to approve fraudulent insurance claims while flagging legitimate ones, leading to financial losses and systemic inefficiencies [21].
- **Backdoors in Triage Systems:** Poisoning attacks may introduce hidden triggers into hospital triage systems, prioritizing or deprioritizing certain patients based on manipulated inputs. Such alterations could go undetected, potentially causing unnoticed mortality risks [22].

These vulnerabilities highlight the urgent need for robust data integrity mechanisms and adversarial defenses in healthcare AI systems.

4.2 Finance

Financial institutions are increasingly leveraging machine learning for credit scoring, fraud detection, and algorithmic trading. However, poisoning attacks pose significant risks, potentially leading to financial instability and regulatory concerns:

- **Inflating Credit Scores:** Attackers could manipulate credit assessment models by injecting poisoned data into training sets, artificially boosting the creditworthiness of high-risk applicants. Such exploits could result in increased loan defaults and systemic financial losses [23].
- **Concealing Fraudulent Transactions:** Poisoning financial fraud detection models by injecting mislabeled transactions can enable malicious actors to evade detection. For instance, by strategically misclassifying fraudulent transactions as legitimate, adversaries can bypass security measures and conduct large-scale financial crimes undetected [24].
- **Disrupting Algorithmic Trading:** Poisoning publicly available financial data feeds (such as stock price histories or sentiment analysis datasets) could mislead automated trading algorithms. This manipulation might cause erroneous buy/sell decisions, creating artificial volatility and providing unfair advantages to attackers or weakening a competitor's trading strategies [25].

These threats highlight the urgent need for robust data validation mechanisms, anomaly detection frameworks, and adversarial resilience in financial AI systems.

4.3 Cybersecurity Operations

Machine learning-driven security analytics platforms play a critical role in modern cybersecurity, enabling automated threat detection, anomaly identification, and real-time incident response. However, poisoning attacks present serious risks to these systems, potentially weakening an organization's defense posture:

- **Creating Blind Spots in Intrusion Detection Systems (IDS):** Adversaries can manipulate training datasets to condition IDS models to ignore specific attack patterns, allowing stealthy exploitation of network vulnerabilities without triggering alerts [26].
- **Overwhelming SOC Analysts with False Alerts:** Attackers can inject poisoned data to artificially inflate false positives, flooding security operations centers (SOCs) with excessive and misleading alerts. This tactic can cause alert fatigue, making it difficult for analysts to identify genuine threats in a sea of noise [1].
- **Eroding Trust in Automated Threat Detection:** Persistent poisoning attacks can degrade the reliability of AI-powered security tools, leading organizations to revert to manual threat detection methods. This regression can increase response times and reduce overall security efficiency [27].

As cyber threats are evolving, ensuring data integrity in security AI models is essential. Defensive measures (like robust data validation, adversarial training, and provenance tracking) are important for maintaining trust in AI-driven security operations.

V. DEFENSIVE STRATEGIES

5.1 Data Curation and Validation

Building a robust data pipeline is essential to protecting AI models from poisoning attacks. A carefully curated and validated dataset can help detect and prevent data manipulation at early stages:

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 11, Issue 11, November 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1111065 |

1. **Secure Sources:** Ensuring that data is ingested only from trusted and verified sources can limit the exposure to poisoned data. By restricting data input to secure, authenticated channels, organizations reduce the risk of adversaries introducing malicious data into the model's training pipeline [1].
2. **Multi-Source Correlation:** Cross-referencing data from multiple independent datasets helps identify inconsistencies that may indicate poisoning. By comparing the same data points across different sources, suspicious anomalies or discrepancies can be flagged. This allows analysts to catch potential poisoning attempts before they influence model outcomes [15].
3. **Data Provenance Tools:** Maintaining cryptographic audits and tracking the origin, timestamps, and labeling processes of training data can establish a trustworthy data lineage. These audits help ensure that data integrity is preserved and provide accountability, making it easier to trace and mitigate any poisoning incidents. By implementing robust data provenance, organizations can verify the authenticity of their datasets and ensure that any tampered data is quickly identified and excluded from the training process.

These defensive strategies can be used to secure data pipeline and are essential for minimizing the risk of poisoning attacks. By implementing these methods, organizations can build resilient AI systems that are more resistant to adversarial manipulation.

5.2 Robust Machine Learning Algorithms

There are several robust learning methods that help machine learning models tolerate or mitigate the impact of outliers, reducing the susceptibility to poisoning attacks, some examples as follows:

1. **Trimmed Loss:** This approach involves ignoring the largest losses during model training, limiting the influence of extreme data points that might have been injected by adversaries. By trimming outlier data points, the model is less likely to be misled by malicious samples that could otherwise skew the learning process.
2. **Median-of-Means:** This technique aggregates subsets of data to compute a median, rather than using the mean, in order to reduce the effect of outliers or mislabeled samples. By focusing on the median, which is less sensitive to extreme values, this method ensures that the model is more robust against data poisoning, preventing adversarial manipulation from distorting the training process.
3. **Adversarial Training:** Adversarial training involves adding carefully crafted, misleading examples to the training set so that the model learns how to defend itself against potential attacks. In the context of poisoning attacks, adversarial training could involve introducing poisoned data during training to make the model more robust against similar future attacks. This method helps to improve the model's resilience by preparing it for potential threats.
4. **Ensemble Methods:** Another effective defense is to use ensemble learning techniques, where multiple models are trained and combined to make predictions. Since ensemble methods aggregate results from various models, it becomes harder for attackers to manipulate all models simultaneously. Even if some models are poisoned, the overall prediction is less likely to be influenced by the attack.
5. **Data Augmentation:** Data augmentation techniques, where additional synthetic data is generated based on existing datasets, can help mitigate the effect of poisoned data. By increasing the diversity of the training set, the model is less likely to overfit on poisoned data, leading to better generalization and reduced vulnerability to attacks.

These robust machine learning techniques are vital for ensuring that models can handle potential data manipulations without sacrificing performance. By using some of the above mentioned strategies, AI systems can be made more resilient against poisoning attacks and ensure accurate decision-making even when faced with compromised data.

5.3 Adaptive Defense Mechanisms

Ongoing monitoring and adaptive defense strategies are crucial for detecting and mitigating poisoning attempts in real-time:

- **Anomaly Detection**: Implement anomaly detection algorithms to flag unusual labeling patterns or significant shifts in feature distributions, which may indicate the presence of poisoned data. Techniques such as unsupervised clustering or statistical analysis (e.g., Z-scores, Chi-squared tests) can help in identifying outliers or abnormal trends that could point to an attack [19].
- **Retraining Alerts**: Set up alerts to trigger manual review if key model performance metrics (e.g., accuracy, precision, recall) show significant and abrupt changes, which could signal a potential poisoning event. Automated retraining models can also be configured to flag any sudden performance deterioration or anomalies in model behavior during evaluation [1].

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 11, Issue 11, November 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1111065 |

These adaptive mechanisms help maintain the integrity of the model by continuously monitoring and reacting to possible data poisoning scenarios, thereby enhancing the model's resilience.

5.4 Model Verification and Validation

Periodic validation of a trained model is a critical defense strategy in identifying and mitigating poisoning attacks. By using a gold standard **(**or untainted dataset - one that has been thoroughly curated and verified) security teams can regularly assess the model's performance and detect any unexpected behaviors or discrepancies. This "trusted test set" approach helps to ensure that backdoors or performance degradation caused by poisoning attacks are identified before a model is deployed into production environments.

For instance, when a model is validated against this clean dataset, any changes in accuracy, precision, or decision boundaries may indicate a manipulation in the training data. If an attacker has implanted a backdoor or caused subtle shifts in the model's behavior, these anomalies would surface during the validation phase. By establishing continuous verification processes, organizations can maintain a higher level of confidence in the integrity and reliability of their AI models.

Moreover, using a trusted test set at multiple points throughout the model lifecycle (during development, post-deployment, and during retraining phases) provides an additional layer of defense which helps to catch issues that may emerge over time as new data is incorporated into the model. This proactive approach increases the robustness of the system and reduces the risk of operational disruptions.

VI. RESEARCH CHALLENGES

1. **Stealthy Poisoning**: As attackers continue to refine methods for embedding malicious data points in training datasets, detecting these subtle manipulations remains a significant challenge. Future research should focus on developing detection techniques capable of identifying incremental and distributed poisoning behaviors (which are harder to trace and prevent) [1].
2. **Scalability of Defenses**: Many robust defense algorithms that address poisoning attacks require considerable computational resources. This raises the question of how resource-constrained environments (such as edge devices or IoT systems) can effectively implement these defenses without sacrificing performance or efficiency [19].
3. **Privacy and Federated Learning**: Federated learning, which enables distributed model training across multiple devices, is increasingly vulnerable to targeted poisoning attacks on individual clients. Research is needed to adapt privacy-preserving techniques (like differential privacy) to mitigate the risks posed by data poisoning while maintaining the collaborative nature of federated learning [4].
4. **Explainability**: The "black box" nature of many machine learning models makes it difficult to identify suspicious data points that may indicate a poisoning attack. Enhancing model explainability could significantly aid in the detection of anomalous training data and sabotage attempts, providing a more transparent understanding of model behavior.
5. **Legal and Ethical Considerations**: Beyond technical challenges, poisoning attacks raise important legal and ethical questions. Specifically, if an attack compromises the integrity of the data and subsequently harms users or organizations, determining accountability and responsibility remains a complex issue. Legal frameworks must evolve to address these concerns and provide clearer guidelines on liability in cases of data tampering.

VII. CONCLUSION

As artificial intelligence (AI) continues to integrate into mission-critical applications, the threat posed by data poisoning attacks becomes more significant and sophisticated. These attacks, which subtly manipulate the training process, can degrade model performance, implant backdoors, and introduce biases that undermine trust in AI systems. The consequences of such attacks can be far-reaching, particularly in high-stakes sectors like healthcare, finance, and cybersecurity operations.

To counter these threats effectively, we need a comprehensive approach that includes securing data pipelines, employing robust training algorithms, conducting regular model validation, and promoting collaboration across AI research, security, and legal fields. While the body of research on adversarial attacks has grown substantially, data poisoning remains an area with many challenges and opportunities for further exploration. The need for ongoing innovation in detection, defense strategies, and resilience is critical. By addressing these issues collectively, we can safeguard the integrity of AI systems in an increasingly connected world.

| e-ISSN: 2320-9801, p-ISSN: 2320-9798| www.ijircce.com | |Impact Factor: 8.379 | Monthly Peer Reviewed & Referred Journal |

|| Volume 11, Issue 11, November 2023 ||

| DOI: 10.15680/IJIRCCE.2023.1111065 |

REFERENCES

[1] B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," in *Proc. Int. Conf. Mach. Learn. (ICML)*, 2012.
[2] T. Gu, K. Liu, B. Dolan-Gavitt, and S. Garg, "BadNets: Identifying vulnerabilities in the machine learning model supply chain," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR) Workshops*, 2019.
[3] L. Huang et al., "Adversarial machine learning," *IEEE Trans. Knowl. Data Eng.*, vol. 23, no. 6, pp. 868-883, 2011.
[4] E. Bagdasaryan et al., "Poisoning attacks against federated learning systems," in *Proc. 33rd Int. Conf. Neural Inf. Process. Syst. (NeurIPS)*, 2020.
[5] M. A. Ilyas, L. A. S. Z. Shou, and B. Recht, "Poisoning attacks in machine learning: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 32, no. 2, pp. 178-189, 2020.
[6] C. Li, H. Xie, and W. Lou, "Secure machine learning for data poisoning attacks: A survey," *Int. J. Comput. Sci. Inf. Secur.*, vol. 16, no. 4, pp. 1-14, 2018.
[7] S. O. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Analysis of adversarial attacks and defenses in machine learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 9, pp. 2674-2687, 2019.
[8] A. B. Rad and A. M. Kermani, "Detecting adversarial samples in machine learning using model behavior analysis," in *Proc. IEEE Int. Conf. Mach. Learn. Appl. (ICMLA)*, 2018, pp. 45-51.
[9] M. S. Abdelkader, N. P. P. P. Biswas, and T. E. Moore, "Blockchain-based data provenance and its applications in machine learning,", vol. 8, pp. 130436-130445, 2020.
[10] J. Pan, Q. Yang, and P. S. Yu, "A survey of transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345-1359, 2010.
[11] C. Xiao et al., "Generating adversarial examples with adversarial networks," in *Proc. 27th Int. Joint Conf. Artif. Intell. (IJCAI)*, 2018, pp. 3905–3911.
[12] E. Bagdasaryan, A. Veit, Y. Hua, D. Estrin, and V. Shmatikov, "How to backdoor federated learning," in *Proc. 23rd Int. Conf. Artif. Intell. Stat. (AISTATS)*, 2020, pp. 2938–2948.
[13] A. Shafahi et al., "Poison frogs! Targeted clean-label poisoning attacks on neural networks," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2018, pp. 6103–6113.
[14] B. Biggio et al., "Evasion attacks against machine learning at test time," in *Proc. Eur. Conf. Mach. Learn. Princ. Pract. Knowl. Discov. Databases (ECML PKDD)*, 2013.
[15] M. Jagielski et al., "Manipulating machine learning: Poisoning attacks and countermeasures for regression learning," in *Proc. IEEE Symp. Secur. Privacy (SP)*, 2018.
[16] B. Biggio, G. Fumera, and F. Roli, "Pattern recognition systems under attack: Design issues and research challenges," *Int. J. Pattern Recognit. Artif. Intell.*, vol. 28, no. 07, p. 1460002, 2014.
[17] N. Papernot, P. McDaniel, A. Sinha, and M. P. Wellman, "SoK: Security and privacy in machine learning," in *Proc. IEEE Eur. Symp. Secur. Privacy (EuroS&P)*, 2018, pp. 399–414.
[18] Y. Liu, X. Chen, C. Liu, and D. Song, "Trojaning attack on neural networks," in *Proc. 25th Netw. Distrib. Syst. Secur. Symp. (NDSS)*, 2018.
[19] A. Steinhardt, P. W. Koh, and P. Liang, "Certified defenses for data poisoning attacks," in *Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2017.
[20] S. G. Finlayson et al., "Adversarial attacks on medical machine learning," *Science*, vol. 363, no. 6433, pp. 1287-1289, 2019.
[21] X. Ma, Y. Liu, J. Bailey, J. Lu, and Y. Jiang, "Understanding adversarial attacks on deep learning-based medical image analysis systems," *Pattern Recognit.*, vol. 110, p. 107332, 2020.
[22] M. Paschali, S. Conjeti, F. Navarro, and N. Navab, "Generalizability vs. robustness: Investigating medical imaging networks using adversarial examples," in *Proc. Med. Image Comput. Comput.-Assist. Interv. (MICCAI)*, 2018, pp. 493-501.
[23] E. Battista, B. Biggio, and F. Roli, "Adversarial machine learning in credit scoring: Threats and countermeasures," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 7, pp. 3015-3028, 2022.
[24] Z. Yang, Y. Wang, J. Zhang, and X. Li, "Data poisoning attacks on financial fraud detection: A survey and empirical study," in *Proc. Int. Conf. Inf. Secur. Cryptol. (ICISC)*, 2021, pp. 214-230.
[25] W. Xu, D. Evans, and Y. Qi, "Feature squeezing: Detecting adversarial examples in deep neural networks," in *Proc. 26th Netw. Distrib. Syst. Secur. Symp. (NDSS)*, 2019, pp. 1-15.
[26] B. I. P. Rubinstein, B. Nelson, L. Huang, A. D. Joseph, and J. D. Tygar, "Stealthy poisoning attacks on PCA-based anomaly detectors," in *Proc. 15th ACM SIGKDD Int. Conf. Knowl. Discov. Data Min. (KDD)*, 2009, pp. 435-444.
[27] R. Sommer and V. Paxson, "Outside the closed world: On using machine learning for network intrusion detection," in *Proc. IEEE Symp. Secur. Privacy*, 2010, pp. 305-316.

INNO SPACE
SJIF Scientific Journal Impact Factor

Impact Factor: 8.379

doi cross ref

ISSN INTERNATIONAL STANDARD SERIAL NUMBER INDIA

निस्केयर NISCAIR

INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

9940 572 462 6381 907 438 ijircce@gmail.com

www.ijircce.com

Scan to save the contact details