Data Poisoning Attacks in the Training Phase of Machine Learning Models: A Review

Mugdha Srivastava[1,*,†], Abhishek Kaushik[1,*,†], Róisín Loughran[1,†] and Kevin McDaid[1,†]

[1]*Regulated Software Research Centre (RSRC), Dundalk Institute of Technology (DkIT), Dundalk, Ireland*

Abstract

Data Poisoning Attacks (DPAs) can severely impact the performance of Machine Learning (ML) models by manipulating training datasets to introduce errors or biases. The integrity of ML models is crucial for user safety and trust, especially as these models increasingly influence key decision-making processes in safety-critical sectors like finance, healthcare, and law enforcement. As ML technology advances, so do the vulnerabilities of these systems, making the reliability of training data vital for ensuring accurate and dependable model outcomes. This review examines the growing threat of DPAs on ML systems at the training stage, categorizing these attacks into label manipulation, data injection, feature space manipulation, and relationship manipulation. By exploring multiple types of attacks and providing relevant examples, this analysis aims to raise awareness about the significant risks posed by compromised data, which can lead to widespread mistrust in ML systems and cause considerable harm, including financial losses, legal liabilities, and even threats to human lives.

Keywords

Data poisoning, artificial intelligence, machine learning, deep learning, cybersecurity, adversarial attacks

1. Introduction

Machine Learning (ML) models demonstrate outstanding effectiveness in addressing a variety of complex data classification and analysis problems. Because ML models can recognize patterns in data and make predictions, they have transformed several sectors such as healthcare by facilitating advanced data analytics, personalized medicine, and predictive modelling [1]. However, adversarial attacks have consistently exposed critical vulnerabilities in such systems, highlighting the need for robust security measures to safeguard the integrity and reliability of these applications in every domain [2].

Data Poisoning Attacks (DPAs), a subset of adversarial attacks, signify a substantial threat to the integrity of ML models because of the multiple pathways in which they can introduce vulnerabilities to a system where accurate and reliable predictions are crucial [3]. Attackers may introduce erroneous or misleading data points, subtly altering class distributions or introducing noise, which can also lead to biased or incorrect predictions [4]. For example, in Figure 1. (a) and (b) show a model built to identify dogs. The model, in Figure 1 (a), trained on clean data is clearly able to classify a dog. The model, in Figure 1 (b), gets trained on a poisoned data point (has red dots and a different label). This training data point, while it looks clearly like a dog to the human eye, gets registered as a cat due to the label as well as the image being poisoned. This leads to the model misclassifying during testing and can have severe implications when models are trained in real-time.

In this paper, we focus on DPAs primarily at the training stage of an ML model because these attacks are growing more nuanced as ML technology evolves [5]. We aim to categorize and analyse various DPAs and assess their impact on ML models to establish real-world consequences. We focus on illustrating these attack types using the *Breast Cancer Wisconsin (Diagnostic) Dataset* [6] which provides a practical scenario to better understand how such attacks can alter model performance.

AICS'24: 32nd Irish Conference on Artificial Intelligence and Cognitive Science, December 09–10, 2024, Dublin, Ireland

*Corresponding author.

†These authors contributed equally.

mugdha.srivastava@dkit.ie (M. Srivastava); abhishek.kaushik@dkit.ie (A. Kaushik); roisin.loughran@dkit.ie (R. Loughran); kevin.mcdaid@dkit.ie (K. McDaid)

0009-0003-6031-4095 (M. Srivastava); 0000-0002-3329-1807 (A. Kaushik); 0000-0002-0974-7106 (R. Loughran); 0000-0002-0695-9082 (K. McDaid)

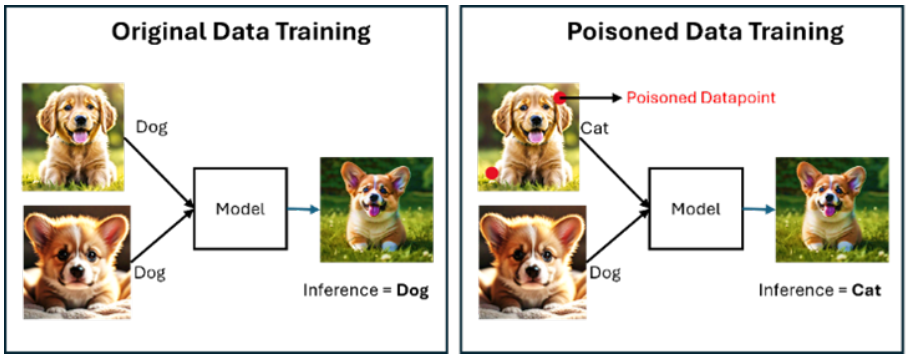

Figure 1: Example of training phase DPA (a) model trained on clean data (b) model trained on poisoned data

The rest of paper is structured as follows: Section 2 describes the previous work related to the research. Section 3 shows the overview of data poisoning and describes the four groups in which DPAs can be divided, enumerates the different types of attacks under the four groups established previously and formulates these attacks using a medical dataset. Section 4 analyses the impacts of these attacks. Section 5 presents a discussion on what are the emerging solutions to DPAs. Section 6 concludes the paper and sets up the future scope of work.

2. Related work

Research on DPAs in ML has gained significant attention due to the vulnerabilities these attacks expose in various AI applications. One study categorizes different attack scenarios and discusses mitigation strategies, emphasizing the interplay between data poisoning and the trustworthiness of AI systems [7]. However, it only describes three different types of training attacks i.e. non-targeted, targeted, and backdoor poisoning.

One survey offers a taxonomy of DPA and an experimental assessment that focuses on the necessity of strong Federated Learning (FL) [8]. This study is limited in scope as it only addresses four types of training attacks specific to FL including label-flipping attacks, poisoning sample attacks, backdoor attacks and untargeted attacks thereby reducing its overall comprehensiveness and limiting its utility for broader applications.

Tian et al. offer an overview of poisoning attacks and countermeasures in centralized and federated learning [9]. They categorize attack methods by their goals, analyses the differences and connections among techniques, present countermeasures with their pros and cons. Their analysis is constrained by its examination of only three types of DPAs in centralized learning and FL. By mentioning nine types of input attacks, the study by Surekha et al. offers a broader perspective into DPAs across multiple types of ML than the previous studies but lacks in-depth explanation on these attacks [10].

A study by Emanuele Cinà et al. provides a comprehensive systematization of DPAs, reviewing over hundred papers in the field over the past fifteen years [11]. They describe five types of attacks, limited to computer vision, and further perform threat modelling on them. The work done by Goldblum et al. provides an extensive list of DPAs during the training phase [12]. They discuss about eight different attack types with what type of a model can be targeted by each attack. Another study provides a comprehensive overview of attacks and defences but does not adequately address the rapid evolution of attack strategies, risking obsolescence of proposed defences [13].

This study presents seventeen distinct DPAs during the training phase, covering multiple domains within ML. These DPAs are further classified into four groups for enhanced clarity and distinction. Each type is illustrated using the *Breast Cancer Wisconsin (Diagnostic) Dataset.*

3. Overview of data poisoning attacks

DPAs detrimentally affect ML systems by intentionally altering the training data to corrupt model performance or change model behaviour [14]. These attacks involve introducing malicious data points or modifying existing ones, skewing the training process to favour the attacker's goals [12]. As ML models are increasingly integrated into various industries, understanding, and mitigating the risks associated with data poisoning is crucial for maintaining the integrity and reliability of these systems [15]. The impact of these attacks can vary from minor performance reduction to severe consequences, depending on the context in which the ML model is employed. DPAs can be classified into several distinct groups,

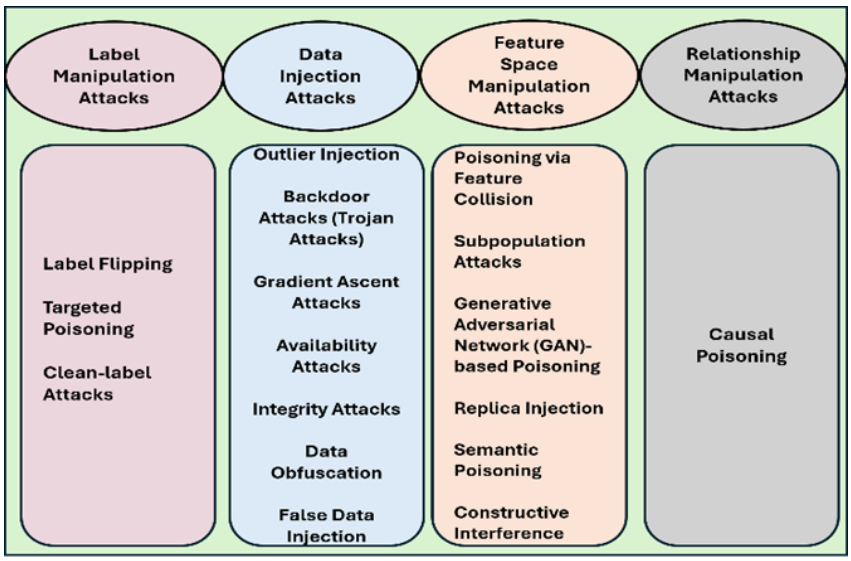

Figure 2: Data poisoning groups and types

each exploiting different vulnerabilities in the ML training process (see Figure 2). These groups include label manipulation, where incorrect labels are assigned to training data [16]; data injection, which involves adding fraudulent data points [17]; feature space manipulation, where the features of the data are altered to mislead the model [18]; and relationship (or context) manipulation, which disrupts the underlying relationships between data points [19]. As shown in Figure 2, the different groups can be further divided into the following types.

3.1. Label manipulation attacks

Label manipulation attacks in ML involve various strategies that aim to compromise the integrity of a model's training data, thereby skewing its outcomes. One common approach is label flipping, where attackers maliciously alter the labels of training samples to mislead the model into making incorrect predictions [20]. Another technique is targeted poisoning, which focuses on specific cases or categories within the dataset, intending to skew the model's results towards erroneous outputs [11]. Additionally, clean-label attacks involve introducing subtle changes to the training data that appear harmless but are strategically crafted to cause model errors [21].

3.2. Data injection attacks

Data injection attacks encompass various techniques used to manipulate and break ML models. One such method is outlier injection, which involves adding extreme feature values to distort the model's learning process [22]. Backdoor attacks (or Trojan Attacks) embed specific trigger patterns in data to control the model's behaviour upon activation [12]. Another approach is gradient ascent, where data is crafted to maximize the model's error rate during training [23]. Availability attacks focus on inserting noise into the training data, hindering the model's learning process and reducing its accuracy [24]. In

contrast, integrity attacks involve making subtle changes to data, leading to a gradual decline in the model's performance [25]. Data obfuscation disguises the attack by altering data in ways that appear plausible, making it difficult to detect [26]. Finally, false data injection creates fictitious records to skew the model's predictions, further compromising its reliability [27].

3.3. Feature space manipulation attacks

Feature space manipulation encompasses several techniques that adversaries use to compromise ML models. One such technique is feature collision which involves creating features that seem harmless but cause the input data's characteristics to overlap or "collide" with those of other features, disrupting how the model interprets and learns from the data [28]. Another method is subpopulation attacks, which target specific demographic groups within the dataset to exploit vulnerabilities associated with those subpopulations [29]. Generative Adversarial Network (GAN)-based poisoning utilizes data generation to produce synthetic data that poisons the model by damaging its performance or causing it to make incorrect predictions [30]. Replica injection involves duplicating examples within the training data, which can skew model bias and lead to overfitting on certain patterns [31]. Semantic poisoning, on the other hand, changes feature relationships to mislead the model by altering the underlying data semantics without altering its appearance [32]. Lastly, constructive interference refers to the manipulation of decision boundaries through manufactured examples, aiming to disrupt the model's ability to accurately classify data by strategically influencing its learning process [33].

3.4. Relationship manipulation attack

Causal poisoning involves deliberately altering correlations between datapoints to mislead causal inference models [34]. This technique can manipulate the perceived relationships within data, leading to wrong conclusions about cause-and-effect dynamics.

3.5. Attack examples using a medical dataset

We explore each of the DPA types using examples based on the *Breast Cancer Wisconsin (Diagnostic) Dataset*. This dataset contains features extracted from breast cancer cell images, where each instance is labeled as either "benign" or "malignant." We will use this dataset as a consistent reference for all examples.

In the dataset (see Table 1), let $\mathbf{X}$ represent the feature matrix, where each row $\mathbf{x}_i$ corresponds to the features of an individual sample, and let $\mathbf{Y}$ represent the label vector, where y_i corresponds to the label of $\mathbf{x}_i$, with $y_i = 1$ for malignant and $y_i = 0$ for benign. Let $\mathbf{x}_j$ be a new data point that does not already exist in the dataset.

All the functions used are denoted in **bold** and *italics* to maintain consistency and clarity in the explanation.

Table 1: Overview of Attacks and Formulations

Name of Attack	Example	Formulation
Label Flipping	Changing the label of certain data points from malignant (1) to benign (0) or vice versa, confusing the model during training and causing it to make incorrect predictions on test data.	Change labels: $y_i = 1 \rightarrow y_i = 0$ for some i where x_i exhibits malignant characteristics.
Targeted Poisoning	Altering the labels of specific cancer cases with rare cell features, flipping their diagnosis from malignant to benign. This can cause the model to perform poorly in these rare but crucial cases.	Modify $y_i = 1 \rightarrow y_i = 0$ for samples with rare features $x_i =$ ***rare***(X).

Continued on next page

Table 1 Continued

Name of Attack	Example	Formulation
Clean-label Attacks	Adding benign samples that have similar features to malignant samples but keeping their label as benign. This confuses the model during inference when it encounters similar patterns in malignant cases.	Add $x_i \approx \boldsymbol{malignant}(X)$, keep $y_i = 0$.
Outlier Injection	Adding outlier data points with impossible or unrealistic feature values, such as extremely high or low measurements, which could skew the model's understanding of what constitutes benign and malignant tumours.	Add x_j with $\max(x_j) \gg \max(X)$ or $\min(x_j) \ll \min(X)$.
Backdoor Attacks	Inserting a specific pattern (i.e., a particular combination of feature values, for example a small watermark) in some training data labelled as benign. The model learns to associate this pattern with benign cases, even if it appears in future malignant inputs.	Insert pattern p in x_i, label as $y_i = 0$ even if $x_i = \boldsymbol{malignant}(X)$.
Gradient Ascent Attacks	Modifying data points to increase the model's error. This can be achieved by creating data samples that maximize prediction errors, thus ruining the overall model performance.	Modify x_i to x_i' such that $\nabla L(\boldsymbol{f}(x_i'), y_i) > \nabla L(\boldsymbol{f}(x_i), y_i)$, where ∇L is the gradient loss function and $\boldsymbol{f}$ is the model.
Availability Attacks	Introducing enough noisy data points that confuse the learning process, causing the model to fail to generalize and effectively classify the actual cases.	Introduce noise: $x_j = \boldsymbol{noise}(X)$, $y_j = \boldsymbol{random}(0, 1)$.
Integrity Attacks	Subtly altering specific features of benign data to appear malignant. This could cause the model to falsely classify future benign datapoints as malignant, leading to over-treatments.	Alter $x_i \to x_i'$ where $x_i' \approx \boldsymbol{malignant}(X)$, $y_i = 0$.
Data Obfuscation	Slightly modifying the features of current benign samples so that they still look plausible but cause damage to the model performance, misclassifying them as malignant.	Slightly change x_i to x_i' such that $d(x_i, x_i') < \epsilon$, $y_i = 0$, where $d()$ represents the distance function and ϵ is the plausibility threshold.
False Data Injection	Adding fictitious patient records with fake measurements and labels to corrupt the training data, leading the model to learn incorrect patterns.	Add fictitious samples (x_j, y_j) with random x_j and y_j.
Poisoning via Feature Collision	Creating new training data points that share feature space similarities with malignant samples but label them as benign, causing confusion and misclassification.	Generate x_j such that $x_j \approx \boldsymbol{malignant}(X)$ but set $y_j = 0$.
Subpopulation Attacks	Targeting training data of a specific patient demographic (e.g., older patients) by adding noise to that subgroup, causing the model to underperform on this specific population.	Add noise to x_i where $x_i = \boldsymbol{older\_patients}(X)$.

Continued on next page

Table 1 Continued

Name of Attack	Example	Formulation
GAN-based Poisoning	Using GANs to generate synthetic images of benign tumours that mimic the feature distribution of malignant cases, causing misclassification during inference.	Use GAN to create $x_{\text{GAN}} \sim \boldsymbol{benign}(X)$ but resembles $\boldsymbol{malignant}(X)$.
Replica Injection	Duplicating certain benign examples multiple times in the dataset to bias the model towards classifying similar features as benign, even when they may be malignant.	Duplicate x_i where $y_i = 0$ multiple times to bias the model.
Semantic Poisoning	Altering benign data samples by changing feature relationships (e.g., modifying cell size ratios) to mislead the model into incorrect conclusions about what defines malignancy.	Alter x_i such that $\boldsymbol{relationship}_{\text{new}}(x_i) \neq \boldsymbol{relationship}_{\text{original}}(x_i)$, maintain $y_i = 0$.
Constructive Interference	Introducing data that causes the model to construct incorrect decision boundaries in feature space, such as mixing malignant and benign features in novel ways to confuse the model.	Introduce x_i such that it lies between decision boundary of benign and malignant, $y_i = 0$.
Causal Poisoning	Introducing data that changes the learned relationships between variables. For example, altering the correlation between cell texture and malignancy, misleading the model into incorrect causal inferences.	Modify data x_i such that $\boldsymbol{corr}_{\text{new}}(\boldsymbol{texture}(x_i), y_i) \neq \boldsymbol{corr}_{\text{original}}(\boldsymbol{texture}(x_i), y_i)$.

4. Impact of data poisoning

Data poisoning is a critical challenge in the development and deployment of ML models as it renders the model ineffective in making sound and reliable decisions [35]. For example, to poison Gmail's spam filtering mechanism attackers sent millions of emails to confuse Gmail's spam filters, allowing malicious emails to bypass detection [36]. In 2016, Microsoft's AI chatbot Tay was shut down hours after launch when malicious users fed it offensive tweets, causing it to post inappropriate content [36]. Researchers have demonstrated that Google's AI image recognition system can be deceived by adversarial attacks, where subtly modified images such as a 3D-printed turtle altered to appear as a rifle, cause the AI to misidentify objects [37]. A firm reportedly manipulated a Tesla's AI system to drive into oncoming traffic by poisoning the training data used for its navigation and decision-making processes [38]. In 2023, a new application called Nightshade came about and is being used by artists to undermine generative AI models by deliberately corrupting their training data, aiming to expose and counteract the impact of AI on their creative work [39].

The performance in critical scenarios, such as healthcare, can directly impact patient care and safety [40]. Even a small percentage of poisoned data can disproportionately affect a model's accuracy, leading to bad performance, misdiagnoses, and incorrect treatment recommendations. For instance, a poisoned model might incorrectly identify benign tumours as malignant or fail to recognize serious conditions, leading to inappropriate treatment plans. As a result, healthcare providers may be reluctant to adopt these systems, fearing potential inaccuracies and the associated liabilities [41].

Data poisoning poses significant risks to ML models in the financial sector as poisoned data can lead to incorrect predictions and decisions in areas like fraud detection, credit scoring, and algorithmic trading [42]. For instance, if an ML model is trained on manipulated data, it may incorrectly classify fraudulent transactions as legitimate, leading to substantial financial losses for institutions. Similarly, poisoned data can skew credit scoring models, resulting in unfair lending practices that either deny

credit to worthy applicants or approve loans for high-risk individuals, increasing default rates. In algorithmic trading, data poisoning can cause models to make erroneous buy or sell decisions, leading to market manipulation and significant financial instability. These vulnerabilities undermine the integrity of financial operations and diminish trust in such systems, which can result in increased regulatory scrutiny and legal liabilities for financial institutions.

An ML model trained on poisoned data that specifically targets a certain demographic can inadvertently perpetuate or even amplify biases that were not initially present [43]. When the poisoned data skews the representation of a particular demographic, the model may develop biased decision-making processes that disproportionately affect that group [44]. This can result in unfair outcomes, such as biased hiring algorithms or discriminatory loan approval systems, where the biases introduced during training become automated, perpetuating systemic inequalities. Even if the original data was free of such biases, the poisoned data can introduce new harmful patterns that the model then enforces in its predictions and decisions.

Backdoors embedded in ML models can pose a serious threat by not only manipulating model behaviour but also by enabling the extraction of sensitive training data. This data, often containing personal or confidential information, can be exploited by attackers to enhance social engineering tactics [45]. For instance, if a backdoor allows access to detailed training data, attackers can gather specific insights about individuals, such as their preferences, behaviours, or personal details. Armed with this information, they can craft highly convincing phishing emails or fraudulent messages tailored to exploit the victim's vulnerabilities. This misuse of extracted data significantly amplifies the effectiveness of social engineering attacks, making them more persuasive and harder to detect.

Data poisoning during the training of ML models can significantly impact public trust and perception of technology [46]. When poisoned data skews a model's outputs, it can undermine confidence in AI systems, especially in critical sectors like healthcare, finance, and law enforcement where reliability and fairness are crucial. This erosion of trust can lead to decreased adoption of AI technologies and heightened scrutiny of their ethical implications. Additionally, compromised models can strain social services by misallocating resources, thereby deepening disparities in access to essential services [47]. The economic impact includes potential financial losses and damage to a company's reputation, which can deter investment in AI research and development, ultimately affecting innovation and economic growth in the tech industry.

5. Discussion

As data poisoning becomes a more prominent threat, emerging defence mechanisms are being developed to protect ML models. Techniques such as adversarial training, formal verification and role-based access controls, training data sanitization, robust statistical methods, and advanced anomaly detection algorithms are at the forefront of these efforts [48]. Adversarial training involves exposing models to potential attacks during the training phase, allowing them to learn from and resist these threats [49]. Robust statistical methods aim to enhance the resilience of models by employing techniques that reduce sensitivity to corrupted data points [50]. Additionally, anomaly detection algorithms are becoming increasingly sophisticated, capable of identifying unusual patterns that may indicate data poisoning [51]. These technological advances aim to fortify ML systems against poisoning attacks, enabling them to maintain performance and reliability even in the face of malicious interference.

Healthcare, traditionally a slow adopter of cutting-edge technology, has been particularly vulnerable to these evolving threats [52]. Unlike sectors such as finance or cybersecurity, which have rapidly integrated ML innovations, medical systems often operate with legacy infrastructures that are less adaptable to new technologies [53]. The sensitivity of health data and the strict regulatory environments further complicate the integration of advanced ML systems, creating a gap where vulnerabilities can easily be exploited [54].

Moreover, the rapid pace of change in ML technology worsens these vulnerabilities. New algorithms and models are being developed at a breakneck speed, outpacing every sector's ability to implement

robust security measures effectively [55]. The need of the hour is for every sector to accelerate its adoption of technological advancements while simultaneously enhancing its cybersecurity posture to protect against the growing threat of data poisoning.

6. Conclusion and future work

DPAs represent a great challenge to the reliability, safety, and ethical application of ML systems. In this paper, we have systematically categorized DPAs into four distinct groups and seventeen specific types, providing a comprehensive framework for understanding the diverse nature of these threats. Furthermore, we have presented clear examples of these attacks, leveraging a medical dataset to demonstrate their practical implications and to facilitate more rigorous analytical interpretations.

Our future work will focus on developing robust defence mechanisms that can preemptively identify and neutralize DPAs before they can affect ML models. This includes further research into the creation of real-time monitoring systems that can detect and respond to DPA threats using technologies like adversarial training and blockchain.

Acknowledgments

This publication has emanated from research conducted with the financial support of Research Ireland under Grant number 21/FFP-A/9255.

References

[1] M. Sahu, R. Gupta, R. K. Ambasta, P. Kumar, Artificial intelligence and machine learning in precision medicine: A paradigm shift in big data analysis, Progress in molecular biology and translational science 190 (2022) 57–100.

[2] O. Ibitoye, R. Abou-Khamis, M. e. Shehaby, A. Matrawy, M. O. Shafiq, The threat of adversarial attacks on machine learning in network security–a survey, arXiv preprint arXiv:1911.02621 (2019).

[3] A. Qayyum, J. Qadir, M. Bilal, A. Al-Fuqaha, Secure and robust machine learning for healthcare: A survey, IEEE Reviews in Biomedical Engineering 14 (2020) 156–180.

[4] D. J. Miller, Z. Xiang, G. Kesidis, Adversarial learning targeting deep neural network classification: A comprehensive review of defenses against attacks, Proceedings of the IEEE 108 (2020) 402–433.

[5] H. Ali, D. Chen, M. Harrington, N. Salazar, M. Al Ameedi, A. Khan, A. R. Butt, J.-H. Cho, A survey on attacks and their countermeasures in deep learning: Applications in deep neural networks, federated, transfer, and deep reinforcement learning, IEEE Access (2023).

[6] W. Wolberg, O. Mangasarian, N. Street, W. Street, Breast cancer wisconsin (diagnostic). uci machine learning repository (1995), 1995.

[7] A. E. Cinà, K. Grosse, A. Demontis, B. Biggio, F. Roli, M. Pelillo, Machine learning security against data poisoning: Are we there yet?, Computer 57 (2024) 26–34.

[8] S. Sagar, C.-S. Li, S. W. Loke, J. Choi, Poisoning attacks and defenses in federated learning: A survey, arXiv preprint arXiv:2301.05795 (2023).

[9] Z. Tian, L. Cui, J. Liang, S. Yu, A comprehensive survey on poisoning attacks and countermeasures in machine learning, ACM Computing Surveys 55 (2022) 1–35.

[10] M. Surekha, A. K. Sagar, V. Khemchandani, A comprehensive analysis of poisoning attack and defence strategies in machine learning techniques, in: 2024 IEEE International Conference on Computing, Power and Communication Technologies (IC2PCT), volume 5, IEEE, 2024, pp. 1662–1668.

[11] A. E. Cinà, K. Grosse, A. Demontis, S. Vascon, W. Zellinger, B. A. Moser, A. Oprea, B. Biggio, M. Pelillo, F. Roli, Wild patterns reloaded: A survey of machine learning security against training data poisoning, ACM Computing Surveys 55 (2023) 1–39.

[12] M. Goldblum, D. Tsipras, C. Xie, X. Chen, A. Schwarzschild, D. Song, A. Mądry, B. Li, T. Goldstein, Dataset security for machine learning: Data poisoning, backdoor attacks, and defenses, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (2022) 1563–1580.
[13] T. Chaalan, S. Pang, J. Kamruzzaman, I. Gondal, X. Zhang, The path to defence: A roadmap to characterising data poisoning attacks on victim models, ACM Computing Surveys 56 (2024) 1–39.
[14] Z. Wang, J. Ma, X. Wang, J. Hu, Z. Qin, K. Ren, Threats to training: A survey of poisoning attacks and defenses on machine learning systems, ACM Computing Surveys 55 (2022) 1–36.
[15] G. Xu, H. Li, H. Ren, K. Yang, R. H. Deng, Data security issues in deep learning: Attacks, countermeasures, and opportunities, IEEE Communications Magazine 57 (2019) 116–122.
[16] R. Croft, M. A. Babar, H. Chen, Noisy label learning for security defects, in: Proceedings of the 19th International Conference on Mining Software Repositories, 2022, pp. 435–447.
[17] J. Shirini, M. K. Shaik, A. Sahithi, P. A. Reddy, N. Jyothi, M. M. Subramanyam, Safeguarding station data integrity: A comprehensive study on detecting and mitigating false data injection through advanced machine learning techniques, Educational Administration: Theory and Practice 30 (2024) 1316–1324.
[18] C. Shorten, T. M. Khoshgoftaar, A survey on image data augmentation for deep learning, Journal of big data 6 (2019) 1–48.
[19] A. Goyal, Y. Bengio, Inductive biases for deep learning of higher-level cognition, Proceedings of the Royal Society A 478 (2022) 20210068.
[20] K. Aryal, M. Gupta, M. Abdelsalam, Analysis of label-flip poisoning attack on machine learning based malware detector, in: 2022 IEEE International Conference on Big Data (Big Data), IEEE, 2022, pp. 4236–4245.
[21] Q. H. Nguyen, N. Ngoc-Hieu, T.-A. Ta, T. Nguyen-Tang, K.-S. Wong, H. Thanh-Tung, K. D. Doan, Wicked oddities: Selectively poisoning for effective clean-label backdoor attacks, arXiv preprint arXiv:2407.10825 (2024).
[22] A. Davoudi, M. Chatterjee, Detection of profile injection attacks in social recommender systems using outlier analysis, in: 2017 IEEE International Conference on Big Data (Big Data), IEEE, 2017, pp. 2714–2719.
[23] L. Liang, X. Hu, L. Deng, Y. Wu, G. Li, Y. Ding, P. Li, Y. Xie, Exploring adversarial attack in spiking neural networks with spike-compatible gradient, IEEE transactions on neural networks and learning systems 34 (2021) 2569–2583.
[24] B. Fang, B. Li, S. Wu, S. Ding, R. Yi, L. Ma, Re-thinking data availability attacks against deep neural networks, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 12215–12224.
[25] S. Sridhar, G. Manimaran, Data integrity attacks and their impacts on scada control system, in: IEEE PES general meeting, IEEE, 2010, pp. 1–6.
[26] J. Stephens, B. Yadegari, C. Collberg, S. Debray, C. Scheidegger, Probabilistic obfuscation through covert channels, in: 2018 IEEE European Symposium on Security and Privacy (EuroS&P), IEEE, 2018, pp. 243–257.
[27] S. Padhan, A. K. Turuk, Design of false data injection attacks in cyber-physical systems, Information Sciences 608 (2022) 825–843.
[28] W. Guo, B. Tondi, M. Barni, An overview of backdoor attacks against deep neural networks and possible defences, IEEE Open Journal of Signal Processing 3 (2022) 261–287.
[29] M. Jagielski, G. Severi, N. Pousette Harger, A. Oprea, Subpopulation data poisoning attacks, in: Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security, 2021, pp. 3104–3122.
[30] X. Chen, D. Zan, W. Li, B. Guan, Y. Wang, A gan-based data poisoning framework against anomaly detection in vertical federated learning, arXiv preprint arXiv:2401.08984 (2024).
[31] L. Engstrom, A. Ilyas, S. Santurkar, D. Tsipras, J. Steinhardt, A. Madry, Identifying statistical bias in dataset replication, in: International Conference on Machine Learning, PMLR, 2020, pp. 2922–2932.
[32] X. You, B. Sheng, D. Ding, M. Zhang, X. Pan, M. Yang, F. Feng, Mass: Model-agnostic, semantic

and stealthy data poisoning attack on knowledge graph embedding, in: Proceedings of the ACM Web Conference 2023, 2023, pp. 2000–2010.
[33] W. He, B. Li, D. Song, Decision boundary analysis of adversarial examples, in: International Conference on Learning Representations, 2018.
[34] C. Improta, Poisoning programs by un-repairing code: security concerns of ai-generated code, in: 2023 IEEE 34th International Symposium on Software Reliability Engineering Workshops (ISSREW), IEEE, 2023, pp. 128–131.
[35] S. A. Abebe, Mitigating unfairness and adversarial attacks in machine learning (2022).
[36] MathCo, Data poisoning and its impact on the ai ecosystem, 2023. URL: https://mathco.com/blog/data-poisoning-and-its-impact-on-the-ai-ecosystem/.
[37] J. Vincent, Google's ai thinks this turtle looks like a gun, which is a problem, 2017. URL: https://www.theverge.com/2017/11/2/16597276/google-ai-image-attacks-adversarial-turtle-rifle-3d-printed.
[38] M. T. Review, Military artificial intelligence can be easily and dangerously fooled, 2019. URL: https://www.technologyreview.com/2019/10/21/132277/military-artificial-intelligence-can-be-easily-and-dangerously-fooled/.
[39] M. T. Review, This new data poisoning tool lets artists fight back against generative ai, 2023. URL: https://www.technologyreview.com/2023/10/23/1082189/data-poisoning-artists-fight-generative-ai/.
[40] G. A. Adam, C.-H. K. Chang, B. Haibe-Kains, A. Goldenberg, Hidden risks of machine learning applied to healthcare: unintended feedback loops between models and future data causing model degradation, in: Machine Learning for Healthcare Conference, PMLR, 2020, pp. 710–731.
[41] C. Jones, J. Thornton, J. C. Wyatt, Artificial intelligence and clinical decision support: clinicians' perspectives on trust, trustworthiness, and liability, Medical law review 31 (2023) 501–520.
[42] M. Śmietanka, A. Koshiyama, P. Treleaven, Algorithms in future insurance markets, International Journal of Data Science and Big Data Analytics 1 (2021) 1–19.
[43] X. Wang, Trustworthy Graph Learning, Ph.D. thesis, Stevens Institute of Technology, 2024.
[44] D. Franco, Towards trustworthiness in artificial intelligence: Pushing for explainable, fair, robust, and private supervised machine learning (2024).
[45] J. Yu, Y. Yu, X. Wang, Y. Lin, M. Yang, Y. Qiao, F.-Y. Wang, The shadow of fraud: The emerging danger of ai-powered social engineering and its possible cure, arXiv preprint arXiv:2407.15912 (2024).
[46] E. Toreini, M. Aitken, K. P. Coopamootoo, K. Elliott, V. G. Zelaya, P. Missier, M. Ng, A. van Moorsel, Technologies for trustworthy machine learning: A survey in a socio-technical context, arXiv preprint arXiv:2007.08911 (2020).
[47] C.-f. Chen, R. Napolitano, Y. Hu, B. Kar, B. Yao, Addressing machine learning bias to foster energy justice, Energy Research & Social Science 116 (2024) 103653.
[48] J. Malik, R. Muthalagu, P. M. Pawar, A systematic review of adversarial machine learning attacks, defensive controls and technologies, IEEE Access (2024).
[49] S. H. Silva, P. Najafirad, Opportunities and challenges in deep learning adversarial robustness: A survey, arXiv preprint arXiv:2007.00753 (2020).
[50] D. Hendrycks, N. Mu, E. D. Cubuk, B. Zoph, J. Gilmer, B. Lakshminarayanan, Augmix: A simple data processing method to improve robustness and uncertainty, arXiv preprint arXiv:1912.02781 (2019).
[51] G. F. Monkam, M. J. De Lucia, N. D. Bastian, A topological data analysis approach for detecting data poisoning attacks against machine learning based network intrusion detection systems, Computers & Security (2024) 103929.
[52] S. M. Williamson, V. Prybutok, Balancing privacy and progress: a review of privacy challenges, systemic oversight, and patient perceptions in ai-driven healthcare, Applied Sciences 14 (2024) 675.
[53] M. Grunt, P. Potejko, Implementing machine learning for enhanced critical infrastructure protection: A framework-centric approach for legacy systems, Wiedza Obronna 286 (2024).
[54] R. U. Rasool, H. F. Ahmad, W. Rafique, A. Qayyum, J. Qadir, Security and privacy of internet of

medical things: A contemporary review in the age of surveillance, botnets, and adversarial ml, Journal of Network and Computer Applications 201 (2022) 103332.

[55] X. Wang, Y. C. Wu, Balancing innovation and regulation in the age of generative artificial intelligence, Journal of Information Policy 14 (2024).