Adversarial Sample Defense and Attack, Backdoor Attack and Defense, Training Data Poisoning, Model Watermarking, Model Robustness, Privacy Preservation, and Secure AI Deployment: A Survey

www.surveyx.cn

Abstract

This survey paper provides a comprehensive examination of the strategies and techniques essential for enhancing the security and integrity of artificial intelligence (AI) systems. It addresses critical areas such as adversarial sample defense and attack, backdoor attack and defense, training data poisoning, model watermarking, model robustness, privacy preservation, and secure AI deployment. The paper highlights the significance of protecting AI systems against adversarial and backdoor attacks, ensuring model ownership through watermarking, and maintaining robustness and privacy. The survey is structured to include foundational knowledge and definitions, followed by an in-depth analysis of each topic. Key findings reveal the dynamic nature of adversarial threats and the need for adaptive defense mechanisms. The paper underscores the limitations of current privacy preservation techniques like differential privacy and explores innovative approaches such as knowledge unlearning and blockchain integration for secure AI deployment. Future research directions are identified, emphasizing the development of comprehensive evaluation frameworks, advanced watermarking techniques, and hybrid security strategies. This survey aims to enhance the understanding of AI security challenges and guide future research efforts in developing resilient AI systems.

1 Introduction

1.1 Significance of Securing AI Systems

The integration of artificial intelligence (AI) in critical applications necessitates robust security measures to address a spectrum of vulnerabilities. Adversarial learning (AL) attacks significantly threaten machine learning systems, compromising the integrity and reliability of AI models, leading to potential misuse [1]. The susceptibility of deep reinforcement learning (DRL) systems to such attacks further emphasizes the need for effective defense mechanisms to ensure consistent performance in essential applications.

Backdoor attacks pose substantial risks, particularly to Face Recognition Systems (FRS), which are vulnerable to both adversarial and backdoor threats [2]. Additionally, diffusion models (DMs) face security vulnerabilities that require comprehensive defense strategies against backdoor and membership inference attacks [3]. The rigidity of existing architectures in managing real-time data influx can lead to processing bottlenecks, further jeopardizing system performance [4].

The necessity of securing AI systems is underscored by issues surrounding dataset copyright auditing, addressing increasing concerns of data misuse and infringement in the machine learning landscape [5]. Vulnerabilities in LiDAR-based perception systems to spoofing attacks highlight the potential for unsafe decisions in autonomous systems, reinforcing the importance of security in modern applications [6]. Furthermore, the challenge of catastrophic forgetting in deep neural networks (DNNs) necessitates continual learning capabilities to enhance adaptability and resilience [7].

Centralized federated learning frameworks experience challenges such as single point failures and data falsification, which can undermine system security and data integrity [8]. The issue of extractable memorization in large language models introduces another critical vulnerability, necessitating robust security measures against data breaches and unauthorized extraction [9]. These challenges highlight the urgent need for comprehensive security strategies to safeguard AI systems from a multitude of threats in contemporary applications.

1.2 Structure of the Survey

This survey is meticulously organized to explore the multifaceted aspects of securing artificial intelligence systems. It commences with an introduction that emphasizes the significance of protecting AI systems from various threats, including adversarial and backdoor attacks, and highlights the importance of model watermarking, robustness, privacy preservation, and secure deployment.

The following section establishes foundational knowledge by defining key concepts such as adversarial attacks, backdoor attacks, data poisoning, model watermarking, model robustness, privacy preservation, and secure AI deployment. Subsequent sections provide a detailed examination of each topic. Section 3 focuses on adversarial sample defense, discussing techniques, challenges, and frameworks for adversarial learning. Section 4 addresses backdoor attacks and defenses, emphasizing recent advancements and ongoing challenges. Section 5 analyzes training data poisoning, its implications, and defensive strategies, supported by evaluation frameworks and benchmarks.

Model watermarking is explored in Section 6, discussing the protection of intellectual property alongside various watermarking techniques and future research directions. Section 7 provides a comprehensive review of model robustness, focusing on techniques designed to enhance resilience against malicious attacks, such as backdoor threats, and environmental changes affecting model performance. It examines both deterministic and probabilistic guarantees of these techniques, offering insights into their effectiveness in maintaining model integrity and security in real-world applications [10, 11, 12, 13, 14].

Section 8 focuses on privacy preservation in machine learning, specifically advanced techniques such as differential privacy and federated learning. It evaluates the effectiveness of these methods in safeguarding Personally Identifiable Information (PII) against threats like membership inference and data reconstruction attacks, while also addressing the challenges of achieving robust privacy protections without compromising model performance. Innovative solutions, such as knowledge unlearning, are highlighted as potential mitigations for privacy risks [13, 15, 16]. Finally, Section 9 discusses secure AI deployment strategies, including blockchain for ownership verification, model extraction attack prevention, and access control, concluding with best practices and emerging trends.

The survey concludes with a summary of key findings and implications for future research, identifying potential areas for further investigation in AI security and privacy. This structured approach ensures a thorough understanding of the current landscape and highlights knowledge gaps in existing research [5].The following sections are organized as shown in Figure 1.

2 Background and Definitions

2.1 Background and Definitions

Understanding fundamental concepts such as adversarial attacks, backdoor attacks, data poisoning, model watermarking, model robustness, privacy preservation, and secure AI deployment is crucial for AI security. Adversarial attacks involve intentional data perturbations to mislead AI models, affecting their accuracy and reliability. These attacks manifest as adversarial examples, backdoor attacks, or weight attacks, each exploiting specific model vulnerabilities [2]. In deep reinforcement learning (DRL), adversarial attacks pose significant threats, necessitating robust defenses [1].

Backdoor attacks involve embedding concealed triggers during model training, which can later be activated to alter model behavior. This threat is particularly pronounced in Graph Neural Networks (GNNs), where vulnerabilities and defenses are systematically analyzed [17]. The susceptibility of diffusion models to backdoor and membership inference attacks further complicates security

AI-generated, for reference only.

challenges [3]. Existing benchmarks often focus on static triggers, overlooking real-world variability [18].

Data poisoning involves altering training data to degrade model performance or induce incorrect outputs, highlighting the need for secure data handling and resilient models [19]. Catastrophic forgetting in deep neural networks, where new learning erases prior knowledge, emphasizes the importance of adaptive and continual learning mechanisms for model robustness [7].

Model watermarking is essential for protecting intellectual property and verifying model integrity. By embedding identifiable marks that resist removal, watermarking ensures ownership and authenticity [4]. The rigidity of static architectures in handling dynamic data streams necessitates adaptable and scalable watermarking solutions [4].

Privacy preservation is critical for safeguarding user data while maintaining model performance. Techniques like differential privacy and federated learning provide frameworks for achieving this balance, despite inherent limitations and challenges. The risks associated with pretrained models, particularly concerning data extraction and unauthorized access, underscore the need for robust privacy measures [5].

Secure AI deployment involves strategies to protect AI systems throughout their lifecycle, ensuring data integrity, preventing unauthorized access, and maintaining resilience against attacks, such as those targeting LiDAR-based perception systems [4]. Additionally, categorizing dataset copyright auditing methods into intrusive and non-intrusive paradigms aids secure deployment by addressing data misuse and infringement concerns [5].

This section lays the groundwork for navigating the complex landscape of AI security, providing essential definitions and explanations for exploring specific threats and defenses in the survey.

3 Adversarial Sample Defense and Attack

3.1 Adversarial Attacks and Defense Mechanisms

Method Name	Attack Types	Defense Strategies	Model Vulnerabilities
BPP[20]	Trojan Attacks	Adversarial Training	Perceptible Triggers
SVF[6]	Spoofing Attacks	Carlo Defense Mechanism	Occlusion Patterns
ADPF[4]	-	-	-
NMP[7]	-	Neuromimetic Models	Perception Models
DPD[21]	Data Poisoning	Differential Privacy	Training Data
A2D[22]	Poisoning Attacks	Purification Techniques	Adversarial Sensitivity

Table 1: This table provides a comprehensive overview of various adversarial attack methods, their corresponding defense strategies, and the specific model vulnerabilities they exploit. The methods listed include BPP, SVF, ADPF, NMP, DPD, and A2D, highlighting the diverse approaches to both attacking and defending machine learning models against adversarial threats.

Adversarial attacks exploit machine learning model vulnerabilities by introducing subtle perturbations to input data, leading to significant misclassifications that often evade human detection [23]. These attacks can be categorized into training data poisoning, test-time evasion, and reverse engineering, impacting various stages of the model lifecycle. The BPPATTACK method exemplifies this by using image quantization and dithering to create imperceptible triggers that activate Trojan behaviors in deep neural networks [20].

The limitations of one-hot encoding further increase the susceptibility of deep learning models to adversarial attacks, facilitating the creation of adversarial examples, especially in deep Q-networks (DQNs), where such examples can alter policy during training [24, 25]. Black-box spoofing attacks exploit overlooked vulnerabilities in perception models, such as in LiDAR systems, necessitating innovative defense strategies using occlusion patterns for detection [6].

Defensive strategies against adversarial attacks span the machine learning lifecycle, including pre-emptive measures during training and post-training strategies [26]. Dynamic frameworks like the Adaptive Data Processing Framework (ADPF) optimize real-time data processing using reinforcement learning techniques to enhance model resilience [4]. The neuromimetic metaplasticity model (NMP) addresses the stability-plasticity dilemma in deep neural networks, allowing them to retain prior knowledge while integrating new information, thereby strengthening defenses against adversarial perturbations [7].

As illustrated in Figure 2, the examples demonstrate the complex dynamics of adversarial sample defense and attack. The first example compares the performance of various deep learning models across three adversarial tasks, highlighting iteration counts' significance in model adaptation to adversarial influences. The second example outlines a systematic approach to protect machine learning models by training a reference model on clean data, assessing its vulnerability to adversarial attacks, and using this insight to fortify a black-box target model. Lastly, clustering analysis using unsupervised learning techniques identifies patterns in potentially adversarial datasets. Collectively, these examples underscore the multifaceted nature of adversarial challenges and the diverse strategies employed to counteract them [21, 22, 19]. Table 1 presents a detailed comparison of adversarial attack methods, defense strategies, and model vulnerabilities, underscoring the complexity and variety of challenges faced in securing machine learning systems.

3.2 Challenges in Adversarial Learning

Method Name	Adversarial Complexity	Defense Mechanisms	System Vulnerabilities
BPP[20]	Imperceptible Triggers	Adaptive Learning	Deep Neural Networks
MW[24]	Effective Attacks	Multi-way Encoding	Deep Learning Models
PIA[25]	Policy Induction Attacks	Robust Countermeasures	Deep Reinforcement Learning
NMP[7]	-	Adaptive Learning Techniques	Deep Neural Networks
SVF[6]	Black-box Spoofing	Carlo Defense Mechanism	Lidar-based Perception
ADPF[4]	-	Adaptive Learning Techniques	Reinforcement Learning Pipelines

Table 2: Comparative Analysis of Adversarial Methods, Defense Mechanisms, and System Vulnerabilities in Machine Learning. This table outlines various adversarial methods, their complexity, the defense mechanisms employed to counteract these threats, and the specific system vulnerabilities they exploit. The analysis provides insights into the challenges faced in developing robust defenses against adversarial attacks.

Adversarial learning faces challenges that hinder robust defense development. A primary issue is the subtlety and complexity of adversarial attacks, which often violate the i.i.d. assumption foundational to traditional machine learning models [23]. This complicates the prediction and mitigation of adversarial threats. Figure 3 illustrates the categorization of challenges in adversarial learning, highlighting the complexity of adversarial attacks, the challenges in developing robust defenses, and the role of adaptive frameworks in enhancing system resilience. Additionally, Table 2 presents a comprehensive comparison of different adversarial methods, highlighting the complexity of attacks, corresponding defense mechanisms, and the vulnerabilities within machine learning systems that these methods exploit.

Perceptible triggers in Trojan attacks can be easily detected, necessitating more sophisticated techniques to embed imperceptible triggers [20]. Conventional one-hot encoding and related training methods create strong correlations between substitute and target model gradients, facilitating effective attacks and highlighting the need for alternative encoding strategies [24].

In deep reinforcement learning (DRL), identifying pipeline vulnerabilities and devising countermeasures is critical [1]. The inherent susceptibility of DQNs to environmental perturbations can be exploited to induce harmful actions, necessitating robust policy learning mechanisms [25].

The stability-plasticity dilemma poses a critical challenge, impacting the design of defenses that prevent catastrophic forgetting while remaining adaptable to new information [7]. This requires innovative approaches to balance knowledge retention with new data assimilation.

Defending against spoofing attacks, particularly in LiDAR-based perception systems, is complicated by the lack of generalizable methods across models and unrealistic assumptions about model access required by existing defenses [6]. This highlights the need for versatile defense strategies capable of effectively countering spoofing threats.

The Adaptive Data Processing Framework (ADPF) addresses these challenges through its adaptive learning principle, enabling dynamic responses to data pattern changes, ensuring high processing efficiency and accuracy [4]. This adaptability is essential for developing robust defenses against evolving adversarial attacks.

AI-generated, for reference only.

3.3 Adversarial Attack Frameworks and Classifications

Adversarial attacks on machine learning models are systematically classified based on operational phase, knowledge requirements, and attack objectives. These classifications facilitate analysis and understanding of diverse adversarial strategies. Primarily, attacks are divided into white-box and black-box categories, distinguished by the attacker's access level to the target model. White-box attacks assume complete knowledge of the model's architecture and parameters, allowing for highly effective perturbations, while black-box attacks rely on limited access, often leveraging substitute models to approximate the target's behavior [25].

In terms of operational phases, adversarial attacks are categorized into training-time and test-time attacks. Training-time attacks, such as data poisoning, aim to corrupt the training dataset, systematically degrading model performance. In contrast, test-time attacks involve crafting adversarial examples that induce incorrect outputs during inference. An example of a test-time attack is presented by Behzadan and Munir [25], where minimal perturbations manipulate learning outcomes and induce suboptimal actions.

The classification of attacks is also based on specific objectives: targeted attacks seek to induce particular misclassifications, exploiting model vulnerabilities for specific harmful outcomes, while untargeted attacks aim to generally undermine a model's performance, reducing overall accuracy and reliability. This distinction is crucial for understanding adversarial strategies and developing effective defenses [27, 19, 22]. Targeted attacks are particularly challenging as they require precise perturbations to achieve desired misclassifications, whereas untargeted attacks can employ broader strategies to confuse the model.

Frameworks for analyzing adversarial attacks often incorporate these classifications to evaluate defense mechanism effectiveness. By understanding the characteristics and objectives of different attacks, researchers can develop tailored defense strategies that address unique challenges. The comprehensive defense framework known as AMAO is essential for enhancing AI system resilience against evolving adversarial threats, including model extraction and backdoor attacks. AMAO integrates four interlinked phases: adversarial training to reduce attack effectiveness, malicious query detection to identify harmful users, label-flipping poisoning attacks to manipulate responses, and the image pHash algorithm to ensure query response indistinguishability. Extensive experiments demonstrate that AMAO effectively counters model extraction attacks while maintaining robustness against adaptive adversaries aware of existing defenses [28, 27].

3.4 Adversarial Attacks in Specialized Domains

Adversarial attacks challenge specialized domains such as image recognition, natural language processing (NLP), and deep reinforcement learning (DRL). In image recognition, adversarial examples manipulate pixel values to deceive models into incorrect classifications, even with imperceptible perturbations [23]. These attacks exploit the model's reliance on specific features, compromising reliability in applications like autonomous driving and medical imaging.

In NLP, adversarial attacks leverage textual data's discrete nature, making it difficult to apply small perturbations without altering semantic meaning [19]. Techniques such as synonym substitution, paraphrasing, and word insertion or deletion create adversarial examples that mislead NLP models, significantly impacting applications like sentiment analysis and machine translation, where textual interpretation accuracy is crucial.

In DRL systems, particularly those utilizing DQNs, adversarial attacks manipulate environmental inputs to induce suboptimal policy decisions [25]. These attacks exploit model sensitivity to input perturbations, leading to erroneous actions that degrade performance in dynamic environments. The transferability of adversarial examples across DRL models further complicates defense strategies, necessitating robust mechanisms for detection and mitigation.

As illustrated in Figure 4, the key domains affected by adversarial attacks are highlighted, showcasing specific techniques and the impacts within image recognition, NLP, and DRL. Each domain presents unique vulnerabilities and challenges posed by adversarial examples.

Effective defense strategies are critical for maintaining AI system integrity and reliability. Techniques include adversarial training to enhance robustness, input preprocessing to filter data, and robust architectures designed to withstand attacks. Collectively, these strategies mitigate risks associated

AI-generated, for reference only.

with model extraction and adversarial manipulation, ensuring model integrity in deployment [10, 9, 22, 19, 27]. However, the evolving nature of these attacks necessitates continuous research and innovation to address potential vulnerabilities and secure AI technologies across specialized domains.

3.5 Detection and Defense Strategies

Detection and defense strategies against adversarial attacks are crucial for maintaining machine learning model robustness and integrity. Multi-way Encoding decorrelates gradients between target and substitute models, improving resistance to adversarial attacks by disrupting common gradient-based strategies [24].

The BPPATTACK adversarial attack achieves a 99.92

The Adaptive Data Processing Framework (ADPF) represents a significant advancement in real-time data handling, outperforming traditional methods in speed and accuracy. This framework enhances model processing efficiency, improving resilience to adversarial perturbations and enabling effective real-time defense strategies [4].

Algorithmic approaches, such as the penalty method for inversion-free deep learning, have been tested across datasets like MNIST, CIFAR10, and Omniglot, demonstrating effectiveness in scenarios involving data denoising, few-shot learning, and adversarial training. This highlights the importance of integrating adversarial training into model development to enhance robustness preemptively [29].

Evaluating adversarial actions through optimal control theory provides a framework for understanding the impact of adversarial inputs on model performance. By treating adversarial perturbations as control inputs, this method enables systematic analysis of their effects, informing the development of targeted defense mechanisms [23].

These strategies collectively emphasize the necessity for a comprehensive and multifaceted approach to effectively detect and defend against various types of adversarial attacks, including model extraction, data poisoning, and backdoor attacks, thereby enhancing the security and robustness of machine learning systems deployed across diverse applications [22, 27, 19]. By combining innovative encoding techniques, real-time data processing frameworks, adversarial training algorithms, and control theory-based evaluations, researchers can develop comprehensive defenses that address the evolving nature of adversarial threats.

In the exploration of cybersecurity, particularly in the domain of machine learning, understanding the intricacies of backdoor attacks and their corresponding defenses is paramount. The complexities of these attacks necessitate a comprehensive framework for analysis. Figure 5 illustrates the hierarchical structure of backdoor attacks and defenses, highlighting the nature and implications of these attacks, alongside recent advancements and persistent challenges in defense strategies. This figure not only serves as a visual representation but also enhances our understanding of the multi-faceted approaches required to mitigate such vulnerabilities in machine learning systems. By examining both the theoretical and practical aspects depicted in the figure, we can better appreciate the ongoing discourse surrounding effective defense mechanisms against backdoor threats.

4 Backdoor Attack and Defense

4.1 Backdoor Attacks: Nature and Defense

Backdoor attacks are a critical threat to machine learning models, embedding covert triggers to alter predictions for adversarial purposes. These attacks typically involve poisoning a small portion of the training data, causing models to behave normally on clean inputs but exhibit malicious actions when exposed to the embedded triggers [30]. Traditional trigger patterns are often detectable, especially under common data transformations, which limits their practicality [31].

To enhance stealth, methods like WABA employ wavelet transforms, integrating trigger images with benign samples in the low-frequency domain, making poisoned images less detectable [30]. The 'Poison Ink' method advances stealth and robustness using dynamic, input-aware edge structures as triggers, surpassing static and visible triggers of earlier methods [31].

In diffusion models, backdoor attacks can lead to copyright infringement by generating infringing images with specific prompts, indicating broader implications beyond model manipulation [32].

The BadT2I framework exemplifies a multimodal backdoor attack, manipulating image synthesis through textual triggers, showcasing the complexity and potential ramifications of such attacks on text-to-image diffusion models [32].

Defense strategies against backdoor attacks focus on detection and mitigation while maintaining model performance on clean data. Progressive Backdoor Erasing (PBE) utilizes adversarial attacks to purify infected models without requiring a clean dataset, valuable when such data is unavailable [33]. Techniques like Sanitizer in federated learning ensure that backdoor-based watermarks remain harmless and verifiable, addressing challenges in decentralized training environments [16].

Current methods, including encryption and targeted backdoor watermarking, have limitations necessitating more secure alternatives without introducing vulnerabilities [34]. In remote sensing, backdoor attacks can manipulate model predictions undetected, underscoring the urgent need for robust defense strategies [30]. Employing advanced statistical methods and innovative purification techniques can bolster AI systems against evolving backdoor threats.

4.2 Recent Advancements and Challenges

Recent advancements in backdoor attack defenses have improved understanding and mitigation of these threats, yet challenges persist in safeguarding machine learning models effectively. Systematic categorization of backdoor attacks based on types, such as data and model poisoning, and phases, like local data collection, training, and aggregation, facilitates distinct attack vector identification and defense strategy tailoring [28].

Methods like Poison Ink represent progress in backdoor attack techniques. By using dynamic, input-aware edge structures, Poison Ink enhances stealthiness, robustness, and flexibility compared to static triggers [31]. Future research could explore frequency domain backdoor attacks and refine method stealthiness against advanced detection techniques [31].

The PBE method effectively erases backdoors without clean extra datasets, leveraging adversarial attack insights to address threats [33]. This approach suggests promising avenues for future defenses.

Despite advancements, detecting backdoor attacks remains challenging, as malicious updates often don't significantly degrade model performance on benign inputs [30]. Current studies frequently lack comprehensive evaluations across various attack scenarios, and many defenses are tailored to specific attack types, limiting generalizability [28]. Future research should focus on developing advanced defense algorithms to enhance deep learning model resilience against backdoor attacks, particularly in remote sensing contexts [30].

The BadT2I framework illustrates a multimodal approach manipulating image synthesis at semantic levels through textual triggers, highlighting the complexity and impact of such attacks on text-to-image diffusion models [32]. This framework offers low training overhead and maintains model utility while effectively injecting backdoors [32].

While progress has been made in understanding and countering backdoor attacks on deep learning models, challenges remain. These include the absence of a systematic taxonomy for backdoor methodologies and insufficient comparative analysis of emerging defenses. Continued research and innovation are essential for developing robust, adaptable defense mechanisms that generalize across attack types, enhancing machine learning system security [35, 36].

5 Training Data Poisoning

5.1 Definition and Impact of Training Data Poisoning

Training data poisoning involves the infiltration of malicious data into machine learning datasets with the intent to manipulate model learning and degrade performance. This can be achieved through misleading samples, data alteration, or omission of critical data points, compromising model integrity and accuracy [37]. The flexibility during the training phase allows attackers to affect multiple data points simultaneously, posing significant challenges for robust defense development [38].

As illustrated in Figure 6, which depicts the hierarchical categorization of training data poisoning in machine learning, the figure highlights various attack definitions, defense strategies, and the associated impacts and challenges of these threats. The consequences of training data poisoning are severe, leading to increased classification errors and reduced predictive accuracy. The Bait and Switch Attack exemplifies this by altering environmental conditions during DNN training, creating erroneous associations and impairing decision-making, such as linking traffic signals with unrelated billboards [39]. In linear regression, poisoned data can severely hinder prediction abilities, highlighting the vulnerability of various architectures [40].

In federated learning, malicious clients can introduce poisoned data, undermining global model performance [41]. This vulnerability highlights the limitations of current privacy strategies and underscores the urgent need for effective defenses in decentralized frameworks [26]. Additionally, deep Q-networks (DQNs) are susceptible to adversarial examples, allowing adversaries to exploit learning processes in reinforcement learning agents [25].

The challenge of extractable memorization, where adversaries recover sensitive data by querying a model, further emphasizes the risks of data poisoning [9]. Benchmark datasets with diverse samples are crucial for testing model vulnerabilities to backdoor attacks, illustrating the complexity of data poisoning threats [42]. However, many benchmarks focus primarily on data poisoning and insufficiently address vulnerabilities of Byzantine-robust aggregation methods to local model manipulations [43].

Training data poisoning poses a critical threat to AI model security and efficacy, necessitating ongoing research and innovation in detection and defense strategies. Advanced techniques like anomaly detection can identify malicious behavior without relying on private data, while enhanced certified defenses can mitigate poisoning risks. Implementing dynamic model perturbation and careful feature selection can further strengthen resilience against adversarial manipulations, ensuring privacy protection and robust model functionality in hostile environments [44, 45].

5.2 Advanced Poisoning Techniques

Advanced poisoning techniques exploit vulnerabilities in the training process through sophisticated strategies aimed at degrading model performance. These techniques include indiscriminate attacks, which broadly degrade performance, targeted attacks focusing on specific outputs, and backdoor attacks introducing hidden triggers [46].

Poisoning Embedded Feature Selection (PES) assesses the security of algorithms like LASSO against poisoning, highlighting adversaries' ability to manipulate feature selection [44]. Bilevel optimization-based attacks aim to reduce the average class recall of a target class, demonstrating potential for comprehensive disruptions [47].

In federated learning, data poisoning poses challenges due to decentralized data distribution. SparseFed leverages the deviation of attackers' updates from benign ones, allowing defenses to filter out harmful contributions [41]. The Finite Aggregation method enhances certified defenses by creating overlapping data subsets, improving prediction aggregation and bolstering robustness [37].

Clean-label poison instances manipulate classifier behavior on specific test instances without affecting overall accuracy, highlighting the dual threat of adversarial examples during training and inference [48]. The benchmark by Schwarzschild et al. provides a standardized framework for testing poisoning attacks, focusing on realistic settings and consistent evaluation metrics [49].

Weerasinghe et al. propose a novel LID-based measure, Neighborhood LID ratio (N-LID), which weights training samples based on their likelihood of being normal or poisoned, enhancing regression model robustness [40]. Patel et al.'s method involves baiting DNNs with misleading images during training to induce misclassification [39].

The evolution of sophisticated poisoning techniques underscores the need for ongoing research and innovation in both offensive and defensive strategies. As adversaries employ tactics like adversarial poisoning, there is a critical demand for effective algorithms capable of detecting compromised models. Developing robust detection frameworks like the Attack To Defend (A2D) approach, which leverages poisoned models' sensitivity to adversarial perturbations, is essential. Continuous efforts are crucial to safeguard machine learning applications across domains from evolving threats [50, 22].

AI-generated, for reference only.

5.3 Defensive Strategies Against Data Poisoning

Defensive strategies against training data poisoning are crucial for preserving machine learning model integrity and performance. These strategies include methods for detecting poisoned models, such as the Attack To Defend (A2D) framework, which identifies poisoned models by assessing their sensitivity to adversarial perturbations. Evaluating the robustness of feature selection methods in the presence of poisoned data is also essential, as certain techniques can be significantly compromised by even a small percentage of malicious samples [51, 44, 22].

The Finite Aggregation method constructs a deterministic classifier using overlapping training data subsets, enhancing robustness by aggregating predictions and mitigating poisoned data impact [37]. Similarly, the N-LID approach uses local intrinsic dimensionality to assess the likelihood of normal or poisoned samples, improving resistance to poisoning attacks [40].

In diffusion models, CopyrightShield detects poisoned samples through spatial similarity characteristics, mitigating copyright infringement risks and enhancing model security [52]. The Progressive Backdoor Erasing (PBE) method purifies infected models by leveraging adversarial examples from potentially poisoned data [33].

In federated learning, the vulnerabilities of existing Byzantine-robust algorithms to local model poisoning attacks necessitate evaluation frameworks to systematically assess defense effectiveness [43]. By focusing on model parameters and vulnerabilities, researchers can develop more targeted defenses against threats in decentralized frameworks [25].

Shafahi et al. propose crafting poison instances that cause classifiers to misclassify specific target instances, underscoring the need for defenses capable of identifying and neutralizing such manipulations [48]. The property inference poisoning method emphasizes the necessity of developing countermeasures that can detect and mitigate inference attacks [53].

Patel et al.'s approach degrades performance across all test data without digital access to the training dataset, highlighting the potential for defenses that do not require direct access to training data [39].

Developing defensive strategies against data poisoning necessitates a multifaceted approach integrating adversarial insights, privacy-preserving techniques, and innovative frameworks to bolster model resilience and security. Ongoing research is crucial to combat sophisticated threats posed by data poisoning and backdoor attacks. This includes developing advanced detection algorithms like the A2D framework, which identifies poisoned models by measuring their sensitivity to adversarial perturbations. Given data poisoning's primary concern for industry practitioners, robust, adaptable defenses are urgently needed to withstand diverse adversarial strategies, particularly as threats may not generalize well to real-world applications. Establishing standardized benchmarks for evaluating these threats will enhance the reliability of future research in this area [49, 22].

5.4 Evaluation Frameworks and Benchmarks

Benchmark	Size	Domain	Task Format	Metric
EGG[54]	12,000	Image Processing	Watermark Removal	Success Rate of Watermark Removal
PII-LM[13]	700,000	Law	Pii Extraction	Precision, Recall
RAD[55]	60,000	Image Classification	Ownership Verification	Accuracy, p-value
TRB[56]	191,939	Text Classification	Category Prediction	Earliness, Duration
DPB[49]	2,500	Image Classification	Backdoor Attack	Success Rate, Attack Efficacy
BackdoorBench[57]	11,492	Computer Vision	Image Classification	CAcc, ASR
FL-BDA[58]	50,000	Image Classification	Backdoor Attack Simulation	Backdoor Accuracy
BackdoorBench[59]	1,000,000	Computer Vision	Image Classification	C-Acc, ASR

Table 3: This table presents a comprehensive overview of various benchmarks utilized in the evaluation of defenses against data poisoning attacks. It details the benchmark names, sizes, domains, task formats, and metrics used to measure effectiveness, providing insights into the robustness of different methods under adversarial conditions.

Evaluation frameworks and benchmarks are essential for assessing defenses against data poisoning attacks, providing insights into model resilience under adversarial conditions. These frameworks enable systematic evaluation of robustness and development of effective mitigation strategies. As illustrated in Figure 7, the evaluation frameworks and benchmarks used to assess various methods against data poisoning attacks are highlighted, showcasing the methods evaluated, challenges addressed, and datasets used. Table 3 provides an in-depth examination of key benchmarks employed in the assessment of methods designed to counteract data poisoning attacks, highlighting the diversity in domains and evaluation metrics. The Finite Aggregation method, for example, has been evaluated on datasets like MNIST, CIFAR-10, and GTSRB, demonstrating enhanced robustness compared to baseline methods using certified fractions as a performance metric [37].

The Progressive Backdoor Erasing (PBE) method was tested on CIFAR-10 and GTSRB datasets, showcasing its effectiveness in mitigating backdoor threats compared to state-of-the-art methods [33]. The CopyrightShield method, evaluated with the Pokemon BLIP Captions dataset, emphasizes integrating domain-specific characteristics into defense strategies to address unique challenges posed by poisoned samples [52].

The Bait and Switch Attack method illustrates poisoning's impact on DNNs, significantly reducing classification accuracy with minimal poisoning, emphasizing the need for robust evaluation frameworks [39]. Experiments by Weerasinghe et al. using benchmark datasets with controlled poisoning rates underscore the necessity of comprehensive benchmarks for evaluating defenses against state-of-the-art methods [40].

Additionally, the property inference poisoning method, tested on Census and Enron email datasets, demonstrates adversaries' ability to infer sensitive properties, highlighting the need for evaluation frameworks assessing privacy risks alongside poisoning threats [53].

Future research could focus on enhancing the verification success rate of domain watermark methods and exploring techniques for improving dataset ownership verification, contributing to developing robust evaluation frameworks [60].

Establishing comprehensive evaluation frameworks and standardized benchmarks is vital for understanding data poisoning attacks and defenses. These frameworks will enable systematic assessment of poisoning methods and defenses under realistic conditions, promoting robust strategies to counter adversarial manipulations in machine learning systems. This is particularly critical given industry concerns about data poisoning threats compromising model integrity [61, 27, 49, 22].

6 Model Watermarking

6.1 Introduction to Model Watermarking

Model watermarking is crucial for safeguarding the intellectual property of machine learning models, particularly those employing deep neural networks (DNNs) in classification tasks. This technique embeds identification data within models to assert ownership while maintaining functionality [62]. Its importance is evident across domains like image processing, speech recognition, and large language models, where unauthorized use threatens IP rights [63]. Recent advancements focus on embedding watermarks that preserve model performance and facilitate ownership verification. For example, embedding watermarking barriers in target models enables extraction even from surrogate models, reinforcing ownership claims [64]. Additionally, black-box watermarking schemes for automatic speech recognition (ASR) address specific challenges in this area [65]. Research exploiting structure consistency in deep learning has led to watermarking methods that leverage stable image structures, such as edges, post-data augmentation [66]. This opens new avenues for embedding resilient watermarks against common transformations. Ongoing research is vital to counter challenges like evasion attacks and watermark removal, enhancing techniques for deep learning models and recommender systems. Recent strategies, including trigger set watermarking and autoregressive out-of-distribution watermarking (AOW), aim to improve robustness against adversarial exploitation and maintain integrity during model distillation and fine-tuning [67, 68, 12].

6.2 Model Watermarking Techniques

Model watermarking techniques protect machine learning models' intellectual property by embedding identifiable information without disrupting primary functions. These techniques are classified by embedding methods, verification access, capacity, authentication, and uniqueness, providing a framework for understanding diverse approaches [63]. Trigger-based methods, such as Black-box Model Watermarking for ASR (BB-WASR), use specific inputs or trigger sets to embed watermarks,

AI-generated, for reference only.

employing trigger audios that integrate speech patterns and linguistic steganography to conceal authorship, thus maintaining robustness and imperceptibility [65]. Adversarial watermarking focuses on embedding durable watermarks that withstand adversarial attacks, with innovations like deep invisible watermarking enhancing capacity and robustness without sacrificing performance [64]. Cryptographic approaches, including blockchain, offer secure verification by storing records on a decentralized ledger, ensuring tamper-proof verification and traceability [62]. AuthNet, for instance, embeds authentication logic within models, using low-activation redundant neurons as authentication bits, adding a security layer for ownership verification [69]. These advancements highlight sophisticated authentication mechanisms within model architectures. As model watermarking evolves, research focuses on enhancing robustness, efficiency, and applicability across architectures and domains, crucial for protecting confidentiality and operational integrity against threats like model extraction and backdoor attacks. Implementing robust detection mechanisms like PRADA and the Attack To Defend (A2D) framework enables proactive monitoring of API query behaviors and assessment of model sensitivity to adversarial perturbations, enhancing resilience in a hostile landscape [14, 11, 10, 22].

6.3 Innovative Watermarking Approaches

Innovative watermarking approaches in machine learning models aim to enhance watermark robustness, stealth, and efficiency while minimizing performance impact. Recent advancements integrate watermarking with emerging technologies to address adversarial and environmental threats. For instance, combining watermarking with blockchain provides a decentralized method for copyright verification and ownership management, leveraging blockchain's immutability for secure watermark storage [70]. Adapting techniques to ensure model integrity post-retraining allows models to maintain protected status despite updates [71]. CosWM, a cosine-based method, shows resilience against ensemble distillation, enabling reliable extraction even when averaging multiple teacher models [72]. The Blind 3D Watermarking Method (B3DWM) uses fuzzy logic for dynamic scaling and optimal watermark placement, showcasing adaptive techniques' effectiveness across architectures [73]. ConceptWM for diffusion models allows watermark embedding with limited data, ensuring robustness against image processing and adversarial threats [74]. PointNCBW employs a two-stage process involving Transferable Feature Perturbation (TFP) and trigger implantation to create watermarked data that misclassifies target class samples under specific triggers [75]. These innovative approaches underscore the continuous advancement of techniques aimed at protecting machine learning models' intellectual property, particularly within Machine Learning as a Service (MLaaS). Strategies like trigger set watermarking and backdoor embedding enhance security and integrity against unauthorized use and model extraction attacks. Recent developments leveraging diffusion models and task-agnostic barriers create robust watermarks less susceptible to evasion, ensuring ownership and functionality remain intact amid evolving threats [64, 12, 62, 63, 76]. Continued research is essential to tackle emerging challenges and refine these methods for broader applicability.

6.4 Future Directions in Model Watermarking

Future research in model watermarking aims to enhance robustness, scalability, and applicability across diverse machine learning domains. A key focus is developing implicit techniques that resist preprocessing attacks, ensuring watermarks remain intact after adversarial manipulations [64]. This involves designing methods inherently resistant to common transformations, fortifying intellectual property security within models. Integrating watermarking with active defense mechanisms is another significant avenue, developing strategies to detect unauthorized use and actively respond to theft attempts, providing dynamic security [62]. Expanding applicability beyond classification to include regression and clustering is crucial for comprehensive AI system protection [63]. Enhancing robustness for ASR models remains a priority, exploring novel embedding methods that maintain imperceptibility while resisting adversarial threats [65]. Exploring frameworks for large-scale, real-world datasets is essential to evaluate practicality and scalability, necessitating rigorous assessments under diverse conditions [63]. These directions highlight the need for innovative solutions to address evolving challenges posed by advancing technologies and adversarial threats. By focusing on these aspects, researchers can enhance the development of advanced techniques that are robust, adaptable, and scalable, essential for safeguarding intellectual property against unauthorized distribution and evasion attacks. Recent studies emphasize employing strategies like utilizing diffusion models to synthesize adversarial examples, maintaining watermark integrity while minimizing exploitation risks. A systematic analysis of existing methods reveals the necessity for a unified threat model and clear security requirements, paving the way for more effective protections against model theft and unauthorized use [63, 12].

7 Model Robustness

7.1 Understanding Model Vulnerabilities

Model robustness is compromised by vulnerabilities that adversaries exploit to impair performance. One significant vulnerability arises from encoding methods like one-hot encoding, which creates strong correlations between gradients of substitute and target models, facilitating effective attacks. Multi-way encoding offers potential in reducing gradient correlation and enhancing resilience against adversarial threats [24]. Neural networks are particularly vulnerable to adversarial perturbations during training and inference. Comprehensive certification mechanisms are vital to verify a model's ability to maintain consistent predictions despite adversarial manipulations, establishing trust in AI systems, especially in high-stakes applications [38].

As illustrated in Figure 8, the key aspects of model vulnerabilities are highlighted, focusing on encoding methods, adversarial attacks, and face recognition systems. This figure emphasizes the role of one-hot and multi-way encodings, the impact of gradient correlation and certification mechanisms, and the challenges posed by feature-based attacks in face recognition. For instance, Face Recognition Systems (FRS) exemplify challenges in model robustness due to reliance on a limited set of facial features, making them susceptible to attacks like the Feature Importance-Based Attack (FIBA), which manipulates predictions by targeting critical features [2]. Developing robust feature extraction and selection techniques is crucial for mitigating these vulnerabilities and safeguarding model integrity. While advancements in understanding adversarial attacks and initial defense mechanisms have improved machine learning robustness, the evolving nature of these threats necessitates ongoing research to address emerging vulnerabilities and strengthen AI models against new attacks [19]. By systematically analyzing and mitigating vulnerabilities, researchers can create more resilient AI systems capable of maintaining performance in the face of adversarial challenges.

7.2 Techniques for Enhancing Robustness

Enhancing machine learning model robustness against adversarial attacks is a critical research focus, emphasizing methods that withstand both black-box and white-box attacks while maintaining high performance on clean data. Multi-way encoding significantly improves robustness by reducing correlations between substitute and target model gradients, enhancing resilience to adversarial perturbations and improving classification accuracy on clean data [24]. Certification mechanisms are crucial for establishing robustness guarantees for neural networks. FullCert, a deterministic end-to-end certifier, systematically verifies a model's resistance to data-poisoning and evasion attacks, ensuring consistent predictions in the presence of adversarial manipulations, thereby reinforcing trust in critical AI applications [38].

The discussed techniques represent a holistic strategy for bolstering AI model robustness, integrating advanced encoding methods, such as diffusion-based watermarking and embedding backdoors, with rigorous certification processes like FullCert. By synthesizing adversarial examples and ensuring reliable recognition, these methods enhance AI system security while protecting intellectual property and mitigating risks associated with model extraction and evasion attacks [38, 12, 76]. As adversarial threats evolve, continued research in these areas remains critical for maintaining machine learning model integrity and performance across various applications.

7.3 Innovative Approaches to Robustness

Innovative strategies for enhancing model robustness focus on merging theoretical insights with practical applications to create resilient AI systems. FullCert exemplifies this approach by integrating theoretical bounds with practical applications to ensure robustness against data-poisoning and evasion attacks [38]. This method verifies the model's ability to maintain consistent predictions amid adversarial manipulations, establishing a robust framework for AI deployment in high-stakes environments.

Multi-way encoding techniques represent a significant advancement in robustness strategies. By minimizing gradient correlations between substitute and target models, this approach enhances resilience to adversarial perturbations while improving classification accuracy on clean data. Such advanced encoding strategies bolster deep learning model security against evolving adversarial threats, complicating attackers' efforts to exploit vulnerabilities and generate effective adversarial examples, while also safeguarding intellectual property through robust watermarking techniques [24, 77, 12].

Investigating advanced feature extraction and selection techniques, particularly in Face Recognition Systems (FRS), is crucial for enhancing resilience against adversarial threats, including adversarial patch and enrollment-stage backdoor attacks. These techniques aim to improve FRS robustness while addressing vulnerabilities identified in recent studies, highlighting attackers' potential to exploit critical features through sophisticated methods like facial identity backdoor attacks. By refining feature selection methods, researchers can better mitigate risks from malicious data manipulation, ensuring FRS reliability in sensitive applications such as surveillance and user authentication [44, 2, 12, 27]. These innovative methods are essential for addressing vulnerabilities inherent in specific model architectures and ensuring AI system integrity across diverse applications.

These innovative approaches emphasize the necessity of integrating theoretical frameworks with practical methodologies in developing resilient AI models. Such models must withstand adversarial threats—like poisoning and backdoor attacks—while ensuring consistent high performance in real-world applications. The "Attack To Defend" (A2D) framework effectively identifies poisoned models by measuring their sensitivity to adversarial perturbations, while advanced watermarking techniques enhance model protection against unauthorized distribution without compromising performance. Addressing these multifaceted challenges enables researchers to create robust AI systems capable of maintaining integrity and reliability across diverse operational environments [9, 28, 12, 22].

7.4 Robustness in Specific Contexts

Tailoring robustness strategies for specific domains is crucial for ensuring AI models maintain performance and reliability under various adversarial conditions. Reachability analysis has proven effective in providing robustness certifications by considering all possible model configurations resulting from bounded perturbations, ensuring the soundness of the certification process, particularly valuable in applications requiring precise and reliable predictions [38].

In areas such as autonomous driving, medical imaging, and financial forecasting, the robustness of AI models is vital due to high-stakes decision-making processes. For instance, autonomous driving models must accurately interpret sensor data and make real-time decisions despite potential adversarial attacks or sensor noise. Enhancing reliability in perception and decision-making systems often incorporates advanced multi-sensor fusion techniques and redundancy checks to mitigate risks from data poisoning attacks and ensure accurate data interpretation from diverse, potentially untrustworthy sources [78, 79, 11, 12, 80].

In medical imaging, where diagnostic accuracy is paramount, robustness strategies focus on ensuring models withstand adversarial perturbations that could lead to misdiagnoses. Techniques like adversarial training and input preprocessing are frequently employed to secure image recognition systems against adversarial attacks, addressing various threats such as test-time evasion and data poisoning that can compromise model performance and safety across sectors like healthcare, automotive, and security [9, 22, 19, 50, 81].

In financial forecasting, where models predict market trends and inform investment decisions, robustness strategies emphasize stability against adversarial manipulations that could skew predictions. Employing advanced statistical techniques and anomaly detection algorithms effectively identifies and mitigates the influence of outliers and adversarial inputs, enhancing model robustness against attacks like data poisoning and backdoor threats that can undermine integrity and performance [22, 11, 44, 19].

Tailoring robustness strategies to specific domains requires a thorough understanding of the distinct challenges and requirements inherent to each context, evidenced by varying vulnerabilities and defense mechanisms identified in machine learning applications, such as model extraction attacks in Machine Learning as a Service (MLaaS) and risks of Personally Identifiable Information (PII) leakage in language models across sectors like healthcare and legal domains [13, 27, 82]. By leveraging domain-specific insights and methodologies, researchers can develop robust AI systems capable of maintaining integrity and performance across a wide range of applications.

7.5 Deterministic and Probabilistic Guarantees

Deterministic and probabilistic guarantees are essential for enhancing machine learning model robustness by providing assurances of performance and reliability under adversarial conditions. Deterministic guarantees are typically achieved through certification processes that verify a model's resilience to specific attacks, such as data poisoning or adversarial perturbations, ensuring predictive accuracy and integrity despite adversarial manipulations. FullCert exemplifies this by offering deterministic end-to-end certification, systematically verifying a model's robustness against data-poisoning and evasion attacks, establishing a reliable framework for AI deployment in high-stakes environments [38].

Probabilistic guarantees provide a statistical measure of a model's robustness, often expressed as the likelihood of maintaining accuracy within a specified confidence interval. These guarantees are particularly valuable in scenarios where deterministic certification may be infeasible due to model complexity or scale. Techniques like probabilistic reachability analysis use statistical methods to assess robustness by evaluating the probability of maintaining accurate predictions across various potential adversarial scenarios [38].

Integrating deterministic and probabilistic guarantees is crucial for developing comprehensive robustness strategies that address the diverse challenges posed by adversarial threats. By combining these approaches, researchers can offer more complete assurances of model reliability, ensuring AI systems withstand both known and unknown adversarial attacks. This dual framework is particularly vital in high-stakes applications where model failures can have severe consequences, such as in autonomous driving, where safety is paramount; in healthcare, where incorrect predictions can affect patient outcomes; and in financial forecasting, where erroneous models can lead to significant monetary losses [27, 10, 76].

Analyzing both deterministic and probabilistic guarantees underscores the need for robust certification and verification processes that thoroughly evaluate and enhance machine learning model resilience against various threats, including data poisoning and adversarial attacks. Recent advancements like FullCert, which provides sound deterministic bounds for training-time and inference-time robustness, and improved methods like Finite Aggregation, which enhances certified defenses against data poisoning, illustrate the importance of these processes in ensuring machine learning systems' reliability and security [11, 37, 12, 61, 38]. As adversarial threats continue to evolve, ongoing research and innovation in these areas are critical for maintaining AI systems' security and integrity across various domains.

8 Privacy Preservation

The increasing prevalence of AI systems necessitates robust privacy preservation strategies, with differential privacy (DP) emerging as a cornerstone approach. Understanding DP's applications and limitations is crucial for navigating the complex privacy landscape of large-scale AI models. This section delves into differential privacy's role and constraints, highlighting its significance in safeguarding privacy within AI systems.

8.1 Differential Privacy and Its Limitations

Differential privacy (DP) ensures that the presence or absence of a single data point minimally impacts data analysis outcomes, providing a quantifiable privacy guarantee. It addresses privacy risks, such as Personally Identifiable Information (PII) leakage, through advanced strategies like knowledge unlearning, thus enhancing model resilience against extraction attacks [15, 12, 13, 83, 76]. However, implementing DP in large-scale models, such as language and diffusion models, is resource-intensive, often requiring costly retraining processes [15].

DP may fall short in addressing extractable memorization risks in large language models (LLMs), where increased model size correlates with heightened PII leakage risks [9]. Benchmarks simulating PII leakage through extraction, reconstruction, and inference attacks highlight vulnerabilities in LMs with black-box access [13]. In federated learning, while DP protects user data during decentralized training, it may not fully mitigate backdoor vulnerabilities threatening model integrity and user privacy [58].

DP's limitations in countering membership inference attacks are evident as adversaries employ shadow training to create attack models without prior knowledge of the target model's training data distribution [84]. Furthermore, evaluations of diffusion models indicate a tendency to memorize training data, necessitating comprehensive privacy strategies beyond DP [85].

While DP is vital for mitigating privacy risks, its effectiveness against sophisticated threats, such as data poisoning attacks, is limited. Supplementary strategies, including knowledge unlearning and advanced access control mechanisms, are necessary to address privacy risks in large-scale and decentralized AI deployments. Knowledge unlearning can erase sensitive information from language models without extensive retraining, while innovative access control methods for edge computing environments enhance data protection by restricting model inference to authorized inputs. As privacy threats evolve, exploring these complementary techniques is essential [21, 69, 15, 13, 83].

8.2 Federated Learning and Client Behavior Modeling

Federated learning (FL) enables decentralized model training across devices while preserving user privacy by keeping data localized. Despite its advantages, FL is vulnerable to model poisoning attacks, where malicious participants introduce harmful updates. Techniques like SparseFed, employing global update sparsification and gradient clipping, aim to mitigate these vulnerabilities. Integrating blockchain, as in VeryFL, enhances model ownership verification and addresses data falsification and incentive distribution concerns [45, 41, 8, 58].

Challenges such as client data heterogeneity complicate update aggregation, potentially leading to suboptimal global models. Inconsistent dataset curation and algorithmic defenses further complicate privacy protection and model performance across domains [13, 86]. Malicious clients pose security risks by injecting poisoned data, compromising global model integrity.

Client behavior modeling is crucial for identifying and mitigating risks from malicious actors attempting model poisoning attacks. By understanding client update patterns, robust aggregation algorithms can filter out harmful contributions, enhancing FL resilience against adversarial attacks [41, 16, 45, 58, 43].

The communication overhead in FL, due to frequent model update exchanges, increases latency and resource consumption, compounded by data poisoning and privacy leakage risks. Effective strategies, like adding controlled randomized noise to gradient updates, are essential for addressing these concerns while maintaining model performance [9, 43, 45]. Efficient communication protocols and optimization techniques are required to minimize bandwidth use while ensuring accuracy.

Integrating knowledge unlearning in FL offers a promising direction for addressing privacy concerns. It allows specific data removal from trained models through targeted parameter updates, providing an efficient alternative to retraining [15]. This is particularly valuable where data privacy regulations require data deletion, ensuring compliance without high retraining costs.

FL offers a powerful framework for privacy-preserving machine learning, but ongoing research is crucial to address challenges like client behavior modeling, communication efficiency, and robustness against adversarial attacks. Effective strategies, such as dynamic model perturbation and secure aggregation, are needed to protect data privacy and maintain model integrity in FL environments [79, 16, 45, 58, 43]. Addressing these challenges will refine FL for secure decentralized AI deployments.

8.3 Knowledge Unlearning

Knowledge unlearning modifies model parameters to "forget" specific information, crucial for mitigating privacy risks from memorized sensitive data in large models. By selectively altering parameters, knowledge unlearning reduces exposure risks, enhancing AI system privacy and security [15].

This technique addresses unintended memorization of sensitive information, a common challenge in deploying large-scale models. It involves targeted parameter updates to remove specific data influences without full model retraining, optimizing computational resources and aligning with data privacy regulations [15, 5, 12, 13, 76].

Adopting data deduplication and differential privacy alongside knowledge unlearning is recommended for addressing data memorization challenges. Data deduplication minimizes redundancy, reducing memorization likelihood, while differential privacy limits individual data point impacts, offering additional privacy protection [85].

Knowledge unlearning enhances privacy preservation by allowing models to manage data memorization risks adaptively. It is more efficient than traditional methods, like data preprocessing and differential privacy, which require extensive retraining. Techniques like gradient ascent on targeted token sequences mitigate privacy risks without degrading performance. Sequential unlearning is more effective than simultaneous forgetting, providing stronger privacy guarantees when specific vulnerable data is identified, addressing PII extraction concerns [87, 15]. As AI models grow in complexity, integrating knowledge unlearning with complementary techniques is essential for maintaining security and integrity in sensitive environments.

8.4 Privacy Risks in Pretrained Models

Pretrained models, especially large language models (LLMs) and diffusion models, risk privacy breaches by memorizing and leaking sensitive training information. This can lead to inadvertent PII exposure, posing privacy concerns for users and organizations [85]. While differential privacy mitigates PII leakage, it does not eliminate risks, necessitating a comprehensive privacy approach. Combining scrubbing techniques with differential privacy enhances the privacy-utility trade-off in language models [13].

Protecting LLM copyrights, especially in Embeddings-as-a-Service (EaaS), is complicated by model extraction attacks that replicate models unauthorizedly, threatening intellectual property rights. Robust copyright protection mechanisms are essential [76]. Knowledge unlearning, modifying model parameters to forget certain information, offers a promising solution to mitigate these risks while maintaining performance [15].

Protecting dataset rights in training models is crucial to prevent misuse and unauthorized sharing. Dataset watermarking techniques safeguard rights while allowing legitimate sharing [55]. The CF-Mark approach protects models from unauthorized access while preserving counterfactual explanation utility for users, balancing security and usability [88].

Pretrained models' privacy risks necessitate a comprehensive preservation strategy, including differential privacy, knowledge unlearning, dataset watermarking, and robust copyright protection to safeguard intellectual property. These techniques address vulnerabilities, enhancing AI system security and integrity for safe deployment across applications [15, 34, 12].

9 Secure AI Deployment

9.1 Blockchain and Decentralized Ownership Verification

Blockchain technology significantly enhances the security and integrity of AI model ownership verification. Its decentralized and immutable nature provides a robust framework for tracking AI model provenance, ensuring transparent and tamper-proof ownership claims, crucial in preventing unauthorized replication and intellectual property violations [27]. The Style License Model exemplifies this by embedding stylistic elements verifiable through blockchain to control model access, thereby enhancing data usage traceability and reinforcing AI security [83].

The QUEEN framework further strengthens model security by combining sensitivity measurement with output perturbation, offering a comprehensive solution against model extraction threats [87]. Integrating blockchain with advanced strategies like QUEEN ensures heightened security and ownership verification, protecting models in decentralized environments. This integration marks a significant advancement in secure AI deployment, enhancing model security against unauthorized access and extraction, and protecting intellectual property rights [8, 83, 89, 12].

9.2 Detection and Prevention of Model Extraction Attacks

Robust detection and prevention mechanisms are essential to protect AI systems from model extraction attacks, which compromise proprietary value by reconstructing equivalent models through adversarial queries [28, 27, 14, 22]. PRADA effectively detects such attacks by analyzing the distribution of query distances to a prediction API, enabling timely intervention [10].

Proactive strategies like the QUEEN framework, which employs query sensitivity measurement and output perturbation, protect model integrity by misleading adversaries into training less effective models [87]. The AMAO framework further enhances defenses by integrating adversarial training, malicious query detection, adaptive query response, and ownership verification [27]. These combined strategies ensure robust defenses against diverse extraction threats, maintaining AI systems' security and integrity [87, 27, 10].

9.3 Access Control and Authorization

Effective access control and authorization are crucial for securing AI systems, ensuring that only authorized entities utilize machine learning models. The Style License Model uses embedded stylistic elements to verify access rights, enhancing security and traceability [83]. Blockchain technology complements this by maintaining a decentralized, tamper-proof ledger of access transactions, reducing unauthorized access risks [83].

The QUEEN framework exemplifies adaptive access control, dynamically adjusting model responses based on query sensitivity to prevent adversarial extraction while maintaining usability for legitimate users [87]. Innovative approaches like stylistic verification, blockchain, and adaptive query responses enhance AI security, ensuring safe operation across diverse environments [69, 83, 12].

9.4 Privacy Risks and Mitigation

AI deployment introduces significant privacy risks, necessitating careful mitigation strategies to protect sensitive information. Adversaries accessing private training datasets can undermine defense mechanisms, highlighting the need for innovative aggregation rules and detection methods in federated learning [35, 45, 58]. EmbMarker addresses copyright protection by ensuring embedding quality without compromising security [76].

Continuous monitoring and evaluation of language models are crucial for responsible AI deployment, addressing vulnerabilities from dynamic usage [9]. The QUEEN framework's adaptive strategies are essential for countering sophisticated adversarial techniques [87]. A comprehensive strategy, including model watermarking and adaptive defenses, ensures secure AI deployment across applications [83, 12].

9.5 Model Authentication and Integrity Verification

Ensuring model authenticity and integrity is critical for AI trust and security. Techniques like watermarking and integrated authentication protect against unauthorized access and tampering [71, 69, 12, 76]. AuthNet integrates authentication logic into models, blocking unauthorized users while maintaining accuracy for legitimate ones [69].

PRADA detects extraction attacks by analyzing query patterns, applicable across machine learning systems without prior model knowledge [10]. Advanced authentication and integrity techniques secure AI models against adversarial threats, facilitating safe deployment [14, 69, 10].

9.6 Secure and Verifiable Training Frameworks

Secure training frameworks are vital for AI model integrity and trust. The Style License Model transforms data for authorized training, safeguarding against unauthorized access [83]. Blockchain technology enhances ownership verification, addressing centralized system vulnerabilities and ensuring transparent incentive allocation in federated learning [8, 89].

AuthNet enhances security by embedding authentication logic into models, preventing unauthorized training modifications [69]. These frameworks employ adversarial training, malicious query detection, and robust watermarking to protect AI models, ensuring security and trustworthiness throughout their lifecycle [9, 27, 38, 12].

10 Conclusion

10.1 Future Research Directions

Advancing AI security and privacy necessitates targeted research to bolster defense mechanisms against sophisticated adversarial threats. Enhancing frameworks like PRADA to better detect adaptive adversaries remains crucial for safeguarding model extraction across diverse datasets. Optimizing access control mechanisms, such as the Style License Model, could significantly reduce resource consumption while improving real-world applicability. Additionally, refining multi-way encoding techniques to optimize dimensionality and performance across varied datasets could fortify AI systems against adversarial attacks.

Exploring recurrent connections in deep neural networks (DNNs) offers potential for improved learning capabilities and resilience against catastrophic forgetting. In the domain of watermarking, enhancing the stealthiness of techniques like the Untargeted Backdoor Watermark could open new avenues for application. Further, advancing inversion-free deep learning methods could address challenges related to non-unique solutions, enhancing efficiency and robustness in adversarial contexts. The development of adaptive defenses that learn from adversarial examples, combined with hybrid approaches integrating robust classification and anomaly detection, is essential for strengthening model security.

In addressing backdoor attacks, research should focus on developing robust, adaptive defenses, exploring novel attack methodologies, and enhancing language models' resilience. Applying advanced frameworks to black-box models and investigating complex backdoor targets could substantially improve protective measures. Moreover, exploring efficient algorithms for complex adversarial dynamics and employing game-theoretic approaches could yield innovative strategies for AI security enhancement. These research directions are pivotal for evolving AI security and privacy, ensuring safe deployment in a complex threat landscape.

10.2 Challenges and Future Directions

Securing AI systems presents several challenges, particularly in developing comprehensive evaluation frameworks that integrate seamlessly with machine learning pipelines. The absence of such frameworks hinders systematic assessment of model robustness and security, especially in multi-modal models where data complexity complicates auditing. Addressing this requires frameworks that provide a holistic view of model performance and security.

The dynamic nature of adversarial threats demands continuous research into adaptive defense strategies that can predict and counteract evolving attacks. Hybrid approaches combining robust classification with anomaly detection show promise, yet require further optimization for diverse applications. Privacy preservation efforts must address limitations in differential privacy implementations, with complementary techniques like knowledge unlearning and data deduplication offering potential solutions. However, these methods need refinement for practical deployment.

Advancing watermarking techniques to protect AI intellectual property is critical. Future research should focus on developing watermarking methods that withstand adversarial attacks and model modifications, thereby strengthening ownership verification. The field of AI security and privacy must continuously adapt to an evolving threat landscape. By tackling these challenges and pursuing innovative research, the community can contribute to developing secure, trustworthy AI systems resilient to adversarial threats while maintaining privacy and integrity across diverse applications.

AI-generated, for reference only.

References

[1] Inaam Ilahi, Muhammad Usama, Junaid Qadir, Muhammad Umar Janjua, Ala Al-Fuqaha, Dinh Thai Hoang, and Dusit Niyato. Challenges and countermeasures for adversarial attacks on deep reinforcement learning. *IEEE Transactions on Artificial Intelligence*, 3(2):90–109, 2021.

[2] Jiahao Chen, Zhiqiang Shen, Yuwen Pu, Chunyi Zhou, Changjiang Li, Jiliang Li, Ting Wang, and Shouling Ji. Rethinking the vulnerabilities of face recognition systems:from a practical perspective, 2024.

[3] Vu Tuan Truong, Luan Ba Dang, and Long Bao Le. Attacks and defenses for generative diffusion models: A comprehensive survey. *arXiv preprint arXiv:2408.03400*, 2024.

[4] Vaden Masrani, Mohammad Akbari, David Ming Xuan Yue, Ahmad Rezaei, and Yong Zhang. Task-agnostic language model watermarking via high entropy passthrough layers, 2024.

[5] Linkang Du, Xuanru Zhou, Min Chen, Chusong Zhang, Zhou Su, Peng Cheng, Jiming Chen, and Zhikun Zhang. Sok: Dataset copyright auditing in machine learning systems. *arXiv preprint arXiv:2410.16618*, 2024.

[6] <div style="text-align: center;".

[7] Suhee Cho, Hyeonsu Lee, Seungdae Baek, and Se-Bum Paik. Neuromimetic metaplasticity for adaptive continual learning, 2024.

[8] Yihao Li, Yanyi Lai, Chuan Chen, and Zibin Zheng. Veryfl: A verify federated learning framework embedded with blockchain, 2023.

[9] Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A Feder Cooper, Daphne Ippolito, Christopher A Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. Scalable extraction of training data from (production) language models. *arXiv preprint arXiv:2311.17035*, 2023.

[10] Mika Juuti, Sebastian Szyller, Samuel Marchal, and N Asokan. Prada: protecting against dnn model stealing attacks. In *2019 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 512–527. IEEE, 2019.

[11] Jonathan Hayase, Weihao Kong, Raghav Somani, and Sewoong Oh. Spectre: Defending against backdoor attacks using robust statistics. In *International Conference on Machine Learning*, pages 4129–4139. PMLR, 2021.

[12] Hongyu Zhu, Sichu Liang, Wentao Hu, Li Fangqi, Ju Jia, and Shi-Lin Wang. Reliable model watermarking: Defending against theft without compromising on evasion. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 10124–10133, 2024.

[13] Nils Lukas, Ahmed Salem, Robert Sim, Shruti Tople, Lukas Wutschitz, and Santiago Zanella-Béguelin. Analyzing leakage of personally identifiable information in language models. In *2023 IEEE Symposium on Security and Privacy (SP)*, pages 346–363. IEEE, 2023.

[14] Pengzhou Cheng, Zongru Wu, Wei Du, Haodong Zhao, Wei Lu, and Gongshen Liu. Backdoor attacks and countermeasures in natural language processing models: A comprehensive security review. *arXiv preprint arXiv:2309.06055*, 2023.

[15] Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. *arXiv preprint arXiv:2210.01504*, 2022.

[16] Kaijing Luo and Ka-Ho Chow. Unharmful backdoor-based client-side watermarking in federated learning, 2024.

[17] Xiao Yang, Gaolei Li, and Jianhua Li. Graph neural backdoor: fundamentals, methodologies, applications, and future directions. *arXiv preprint arXiv:2406.10573*, 2024.

[18] Yiming Li, Tongqing Zhai, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor attack in the physical world, 2021.

[19] David J Miller, Zhen Xiang, and George Kesidis. Adversarial learning in statistical classification: A comprehensive review of defenses against attacks. *arXiv preprint arXiv:1904.06292*, 2019.

[20] Zhenting Wang, Juan Zhai, and Shiqing Ma. Bppattack: Stealthy and efficient trojan attacks against deep neural networks via image quantization and contrastive adversarial learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15074–15084, 2022.

[21] Yuzhe Ma, Xiaojin Zhu, and Justin Hsu. Data poisoning against differentially-private learners: Attacks and defenses. *arXiv preprint arXiv:1903.09860*, 2019.

[22] Samar Fares and Karthik Nandakumar. Attack to defend: Exploiting adversarial attacks for detecting poisoned models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24726–24735, 2024.

[23] Xiaojin Zhu. An optimal control view of adversarial machine learning, 2018.

[24] Donghyun Kim, Sarah Adel Bargal, Jianming Zhang, and Stan Sclaroff. Multi-way encoding for robustness, 2020.

[25] Vahid Behzadan and Arslan Munir. Vulnerability of deep reinforcement learning to policy induction attacks. In *Machine Learning and Data Mining in Pattern Recognition: 13th International Conference, MLDM 2017, New York, NY, USA, July 15-20, 2017, Proceedings 13*, pages 262–275. Springer, 2017.

[26] Baoyuan Wu, Shaokui Wei, Mingli Zhu, Meixi Zheng, Zihao Zhu, Mingda Zhang, Hongrui Chen, Danni Yuan, Li Liu, and Qingshan Liu. Defenses in adversarial machine learning: A survey. *arXiv preprint arXiv:2312.08890*, 2023.

[27] A comprehensive defense framewor.

[28] Xuan Sheng, Zhaoyang Han, Piji Li, and Xiangmao Chang. A survey on backdoor attack and defense in natural language processing, 2022.

[29] Akshay Mehra and Jihun Hamm. Penalty method for inversion-free deep bilevel optimization, 2021.

[30] Nikolaus Dräger, Yonghao Xu, and Pedram Ghamisi. Backdoor attacks for remote sensing data with wavelet transform. *IEEE Transactions on Geoscience and Remote Sensing*, 61:1–15, 2023.

[31] Journal of l a tex class files.

[32] Shengfang Zhai, Yinpeng Dong, Qingni Shen, Shi Pu, Yuejian Fang, and Hang Su. Text-to-image diffusion models can be easily backdoored through multimodal data poisoning. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 1577–1587, 2023.

[33] Bingxu Mu, Zhenxing Niu, Le Wang, Xue Wang, Qiguang Miao, Rong Jin, and Gang Hua. Progressive backdoor erasing via connecting backdoor and adversarial attacks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20495–20503, 2023.

[34] Yiming Li, Yang Bai, Yong Jiang, Yong Yang, Shu-Tao Xia, and Bo Li. Untargeted backdoor watermark: Towards harmless and stealthy dataset copyright protection. *Advances in Neural Information Processing Systems*, 35:13238–13250, 2022.

[35] Yansong Gao, Bao Gia Doan, Zhi Zhang, Siqi Ma, Jiliang Zhang, Anmin Fu, Surya Nepal, and Hyoungshick Kim. Backdoor attacks and countermeasures on deep learning: A comprehensive review. *arXiv preprint arXiv:2007.10760*, 2020.

[36] Yiming Li, Yong Jiang, Zhifeng Li, and Shu-Tao Xia. Backdoor learning: A survey. *IEEE transactions on neural networks and learning systems*, 35(1):5–22, 2022.

[37] Wenxiao Wang, Alexander Levine, and Soheil Feizi. Improved certified defenses against data poisoning with (deterministic) finite aggregation, 2022.

AI-generated, for reference only.

[38] Tobias Lorenz, Marta Kwiatkowska, and Mario Fritz. Fullcert: Deterministic end-to-end certification for training and inference of neural networks, 2024.

[39] Naman Patel, Prashanth Krishnamurthy, Siddharth Garg, and Farshad Khorrami. Bait and switch: Online training data poisoning of autonomous driving systems, 2020.

[40] Sandamal Weerasinghe, Sarah M. Erfani, Tansu Alpcan, Christopher Leckie, and Justin Kopacz. Defending regression learners against poisoning attacks, 2020.

[41] Ashwinee Panda, Saeed Mahloujifar, Arjun Nitin Bhagoji, Supriyo Chakraborty, and Prateek Mittal. Sparsefed: Mitigating model poisoning attacks in federated learning with sparsification. In *International Conference on Artificial Intelligence and Statistics*, pages 7587–7624. PMLR, 2022.

[42] ckdoor attacks against transfer.

[43] Minghong Fang, Xiaoyu Cao, Jinyuan Jia, and Neil Gong. Local model poisoning attacks to {Byzantine-Robust} federated learning. In *29th USENIX security symposium (USENIX Security 20)*, pages 1605–1622, 2020.

[44] Huang Xiao, Battista Biggio, Gavin Brown, Giorgio Fumera, Claudia Eckert, and Fabio Roli. Is feature selection secure against training data poisoning?, 2018.

[45] Wenqi Wei, Tiansheng Huang, Zachary Yahn, Anoop Singhal, Margaret Loper, and Ling Liu. Data poisoning and leakage analysis in federated learning, 2024.

[46] Antonio Emanuele Cinà, Kathrin Grosse, Ambra Demontis, Sebastiano Vascon, Werner Zellinger, Bernhard A. Moser, Alina Oprea, Battista Biggio, Marcello Pelillo, and Fabio Roli. Wild patterns reloaded: A survey of machine learning security against training data poisoning, 2023.

[47] Akshay Mehra, Bhavya Kailkhura, Pin-Yu Chen, and Jihun Hamm. How robust are randomized smoothing based defenses to data poisoning? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13244–13253, 2021.

[48] Ali Shafahi, W Ronny Huang, Mahyar Najibi, Octavian Suciu, Christoph Studer, Tudor Dumitras, and Tom Goldstein. Poison frogs! targeted clean-label poisoning attacks on neural networks. *Advances in neural information processing systems*, 31, 2018.

[49] Avi Schwarzschild, Micah Goldblum, Arjun Gupta, John P Dickerson, and Tom Goldstein. Just how toxic is data poisoning? a unified benchmark for backdoor and data poisoning attacks. In *International Conference on Machine Learning*, pages 9389–9398. PMLR, 2021.

[50] Liam Fowl, Micah Goldblum, Ping-yeh Chiang, Jonas Geiping, Wojciech Czaja, and Tom Goldstein. Adversarial examples make strong poisons. *Advances in Neural Information Processing Systems*, 34:30339–30351, 2021.

[51] Wenjun Qiu. A survey on poisoning attacks against supervised machine learning. *arXiv preprint arXiv:2202.02510*, 2022.

[52] Zhixiang Guo, Siyuan Liang, Aishan Liu, and Dacheng Tao. Copyrightshield: Spatial similarity guided backdoor defense against copyright infringement in diffusion models. *arXiv preprint arXiv:2412.01528*, 2024.

[53] Melissa Chase, Esha Ghosh, and Saeed Mahloujifar. Property inference from poisoning, 2021.

[54] Haonan An, Guang Hua, Zhiping Lin, and Yuguang Fang. Box-free model watermarks are prone to black-box removal attacks, 2024.

[55] Buse Gul Atli Tekgul and N. Asokan. On the effectiveness of dataset watermarking in adversarial settings, 2022.

[56] Wenxiao Wang and Soheil Feizi. Temporal robustness against data poisoning, 2023.

[57] Baoyuan Wu, Hongrui Chen, Mingda Zhang, Zihao Zhu, Shaokui Wei, Danni Yuan, Mingli Zhu, Ruotong Wang, Li Liu, and Chao Shen. Backdoorbench: A comprehensive benchmark and analysis of backdoor learning, 2024.

[58] Eugene Bagdasaryan, Andreas Veit, Yiqing Hua, Deborah Estrin, and Vitaly Shmatikov. How to backdoor federated learning, 2019.

[59] Baoyuan Wu, Hongrui Chen, Mingda Zhang, Zihao Zhu, Shaokui Wei, Danni Yuan, and Chao Shen. Backdoorbench: A comprehensive benchmark of backdoor learning, 2022.

[60] Junfeng Guo, Yiming Li, Lixu Wang, Shu-Tao Xia, Heng Huang, Cong Liu, and Bo Li. Domain watermark: Effective and harmless dataset copyright protection is closed at hand. *Advances in Neural Information Processing Systems*, 36:54421–54450, 2023.

[61] Wenxiao Wang, Alexander J Levine, and Soheil Feizi. Improved certified defenses against data poisoning with (deterministic) finite aggregation. In *International Conference on Machine Learning*, pages 22769–22783. PMLR, 2022.

[62] Franziska Boenisch. A systematic review on model watermarking for neural networks. *Frontiers in big Data*, 4:729663, 2021.

[63] Franziska Boenisch. A systematic review on model watermarking for neural networks, 2021.

[64] Jie Zhang, Dongdong Chen, Jing Liao, Weiming Zhang, Huamin Feng, Gang Hua, and Nenghai Yu. Deep model intellectual property protection via deep watermarking, 2021.

[65] Haozhe Chen, Weiming Zhang, Kunlin Liu, Kejiang Chen, Han Fang, and Nenghai Yu. Speech pattern based black-box model watermarking for automatic speech recognition, 2022.

[66] Jie Zhang, Dongdong Chen, Jing Liao, Han Fang, Zehua Ma, Weiming Zhang, Gang Hua, and Nenghai Yu. Exploring structure consistency for deep model watermarking, 2021.

[67] Huajie Chen, Tianqing Zhu, Chi Liu, Shui Yu, and Wanlei Zhou. High-frequency matters: An overwriting attack and defense for image-processing neural network watermarking, 2023.

[68] Sixiao Zhang, Cheng Long, Wei Yuan, Hongxu Chen, and Hongzhi Yin. Watermarking recommender systems, 2024.

[69] Yuling Cai, Fan Xiang, Guozhu Meng, Yinzhi Cao, and Kai Chen. Authnet: Neural network with integrated authentication logic, 2024.

[70] Xuhong Wang, Haoyu Jiang, Yi Yu, Jingru Yu, Yilun Lin, Ping Yi, Yingchun Wang, Yu Qiao, Li Li, and Fei-Yue Wang. Building intelligence identification system via large language model watermarking: A survey and beyond, 2024.

[71] Shahinul Hoque, Farhin Farhad Riya, and Jinyuan Sun. Deep learning model integrity checking mechanism using watermarking technique, 2023.

[72] Laurent Charette, Lingyang Chu, Yizhou Chen, Jian Pei, Lanjun Wang, and Yong Zhang. Cosine model watermarking against ensemble distillation, 2022.

[73] Sharvari C. Tamane and Ratnadeep R. Deshmukh. Blind 3d model watermarking based on multi-resolution representation and fuzzy logic, 2012.

[74] Liangqi Lei, Keke Gai, Jing Yu, Liehuang Zhu, and Qi Wu. Conceptwm: A diffusion model watermark for concept protection, 2024.

[75] Cheng Wei, Yang Wang, Kuofeng Gao, Shuo Shao, Yiming Li, Zhibo Wang, and Zhan Qin. Pointncbw: Towards dataset ownership verification for point clouds via negative clean-label backdoor watermark. *IEEE Transactions on Information Forensics and Security*, 2024.

[76] Wenjun Peng, Jingwei Yi, Fangzhao Wu, Shangxi Wu, Bin Zhu, Lingjuan Lyu, Binxing Jiao, Tong Xu, Guangzhong Sun, and Xing Xie. Are you copying my model? protecting the copyright of large language models for eaas via backdoor watermark. *arXiv preprint arXiv:2305.10036*, 2023.

[77] Junpei Liao, Liang Yi, Wenxin Shi, Wenyuan Yang, Yanmei Fang, and Xin Yang. Imperceptible backdoor watermarks for speech recognition model copyright protection. *Visual Intelligence*, 2(1):23, 2024.

[78] Chenglin Miao, Qi Li, Houping Xiao, Wenjun Jiang, Mengdi Huai, and Lu Su. Towards data poisoning attacks in crowd sensing systems. In *Proceedings of the Eighteenth ACM International Symposium on Mobile Ad Hoc Networking and Computing*, pages 111–120, 2018.

[79] Aysha Thahsin Zahir Ismail and Raj Mani Shukla. Analyzing the vulnerabilities in splitfed learning: Assessing the robustness against data poisoning attacks, 2023.

[80] Mingda Zhang, Mingli Zhu, Zihao Zhu, and Baoyuan Wu. Reliable poisoned sample detection against backdoor attacks enhanced by sharpness aware minimization. *arXiv preprint arXiv:2411.11525*, 2024.

[81] Shangwei Guo, Tianwei Zhang, Han Qiu, Yi Zeng, Tao Xiang, and Yang Liu. Fine-tuning is not enough: A simple yet effective watermark removal attack for dnn models. *arXiv preprint arXiv:2009.08697*, 2020.

[82] Jiacheng Liang, Zian Wang, Lauren Hong, Shouling Ji, and Ting Wang. Waterpark: A robustness assessment of language model watermarking, 2024.

[83] Peihao Li. A novel access control and privacy-enhancing approach for models in edge computing, 2024.

[84] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

[85] Nicolas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramer, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium (USENIX Security 23)*, pages 5253–5270, 2023.

[86] Jiayi Fu, Xuandong Zhao, Ruihan Yang, Yuansen Zhang, Jiangjie Chen, and Yanghua Xiao. Gumbelsoft: Diversified language model watermarking via the gumbelmax-trick, 2024.

[87] Huajie Chen, Tianqing Zhu, Lefeng Zhang, Bo Liu, Derui Wang, Wanlei Zhou, and Minhui Xue. Queen: Query unlearning against model extraction, 2024.

[88] Hangzhi Guo, Firdaus Ahmed Choudhury, Tinghua Chen, and Amulya Yadav. Watermarking counterfactual explanations. *arXiv preprint arXiv:2405.18671*, 2024.

[89] Yihao Li, Yanyi Lai, Tianchi Liao, Chuan Chen, and Zibin Zheng. Tokenized model: A blockchain-empowered decentralized model ownership verification platform, 2023.

Disclaimer:

SurveyX is an AI-powered system designed to automate the generation of surveys. While it aims to produce high-quality, coherent, and comprehensive surveys with accurate citations, the final output is derived from the AI's synthesis of pre-processed materials, which may contain limitations or inaccuracies. As such, the generated content should not be used for academic publication or formal submissions and must be independently reviewed and verified. The developers of SurveyX do not assume responsibility for any errors or consequences arising from the use of the generated surveys.

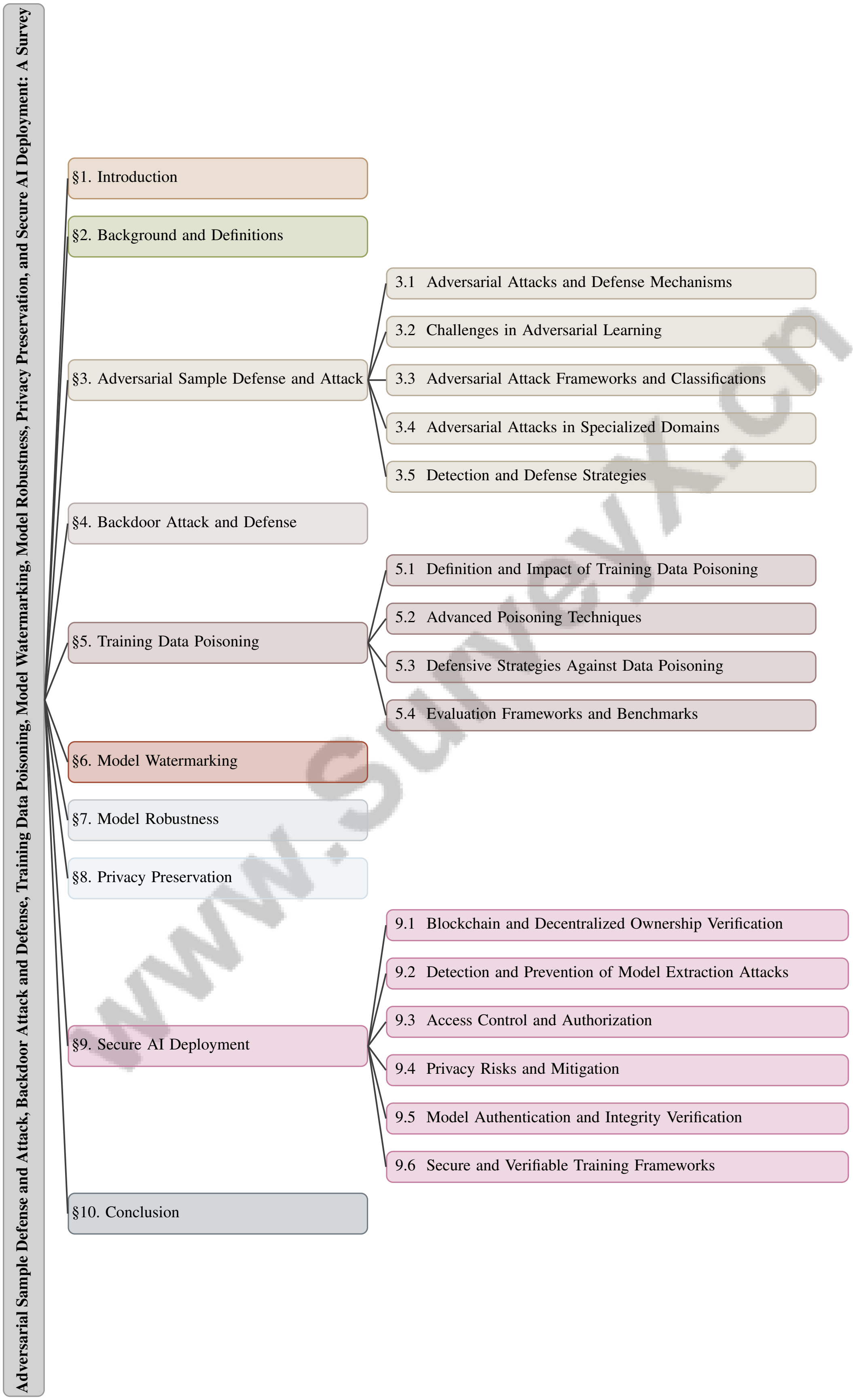
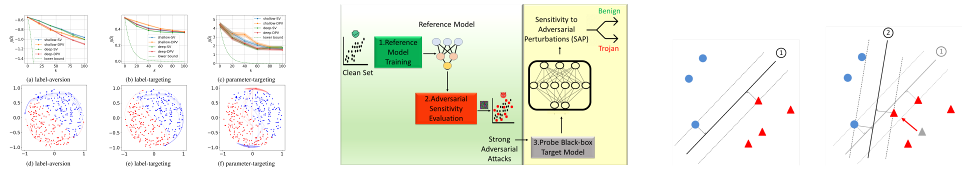

Adversarial Sample Defense and Attack, Backdoor Attack and Defense, Training Data Poisoning, Model Watermarking, Model Robustness, Privacy Preservation, and Secure AI Deployment: A Survey
§1. Introduction
§2. Background and Definitions
§3. Adversarial Sample Defense and Attack
3.1 Adversarial Attacks and Defense Mechanisms
3.2 Challenges in Adversarial Learning
3.3 Adversarial Attack Frameworks and Classifications
3.4 Adversarial Attacks in Specialized Domains
3.5 Detection and Defense Strategies
§4. Backdoor Attack and Defense
§5. Training Data Poisoning
5.1 Definition and Impact of Training Data Poisoning
5.2 Advanced Poisoning Techniques
5.3 Defensive Strategies Against Data Poisoning
5.4 Evaluation Frameworks and Benchmarks
§6. Model Watermarking
§7. Model Robustness
§8. Privacy Preservation
§9. Secure AI Deployment
9.1 Blockchain and Decentralized Ownership Verification
9.2 Detection and Prevention of Model Extraction Attacks
9.3 Access Control and Authorization
9.4 Privacy Risks and Mitigation
9.5 Model Authentication and Integrity Verification
9.6 Secure and Verifiable Training Frameworks
§10. Conclusion

Figure 1: chapter structure



(a) Deep Learning Models' Performance in Label-Aversion, Label-Targeting, and Parameter-Targeting Tasks[21]

(b) Adversarial Perturbation Detection and Attack Prevention[22]

(c) Clustering Analysis Using K-Means Algorithm[19]

Figure 2: Examples of Adversarial Attacks and Defense Mechanisms

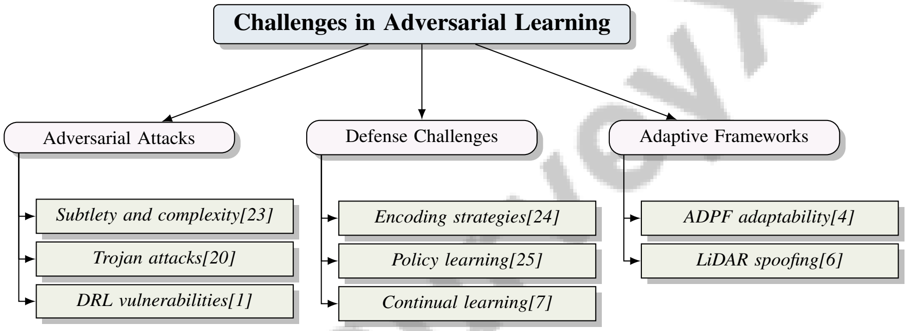


Figure 3: This figure illustrates the categorization of challenges in adversarial learning, highlighting the complexity of adversarial attacks, the challenges in developing robust defenses, and the role of adaptive frameworks in enhancing system resilience.

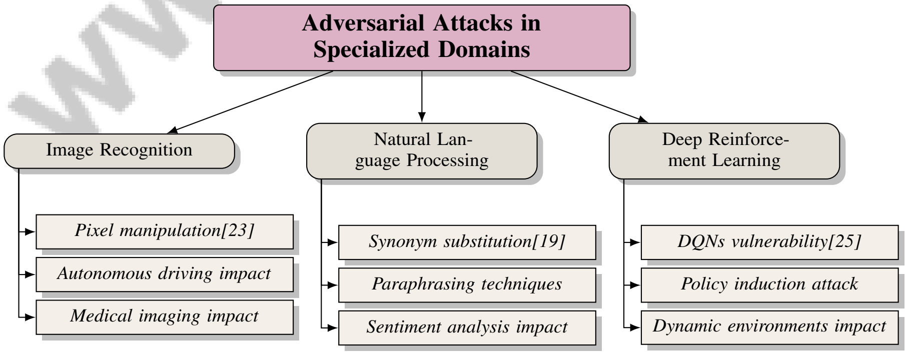

Adversarial Attacks in Specialized Domains
Image Recognition
Natural Language Processing
Deep Reinforcement Learning
Pixel manipulation[23]
Autonomous driving impact
Medical imaging impact
Synonym substitution[19]
Paraphrasing techniques
Sentiment analysis impact
DQNs vulnerability[25]
Policy induction attack
Dynamic environments impact

Figure 4: This figure illustrates the key domains affected by adversarial attacks, highlighting specific techniques and impacts within image recognition, natural language processing, and deep reinforcement learning. Each domain shows unique vulnerabilities and challenges posed by adversarial examples.

AI-generated, for reference only.

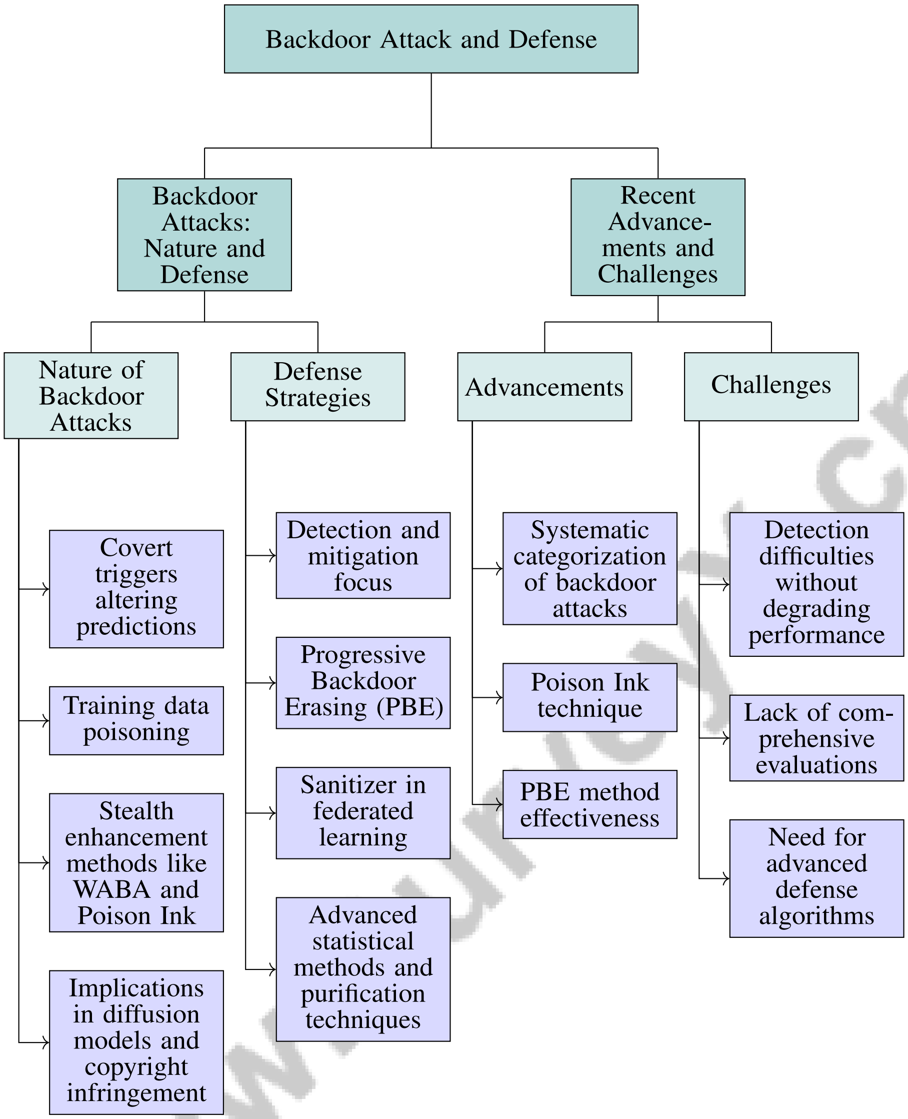

Backdoor Attack and Defense
Backdoor Attacks: Nature and Defense
Recent Advancements and Challenges
Nature of Backdoor Attacks
Defense Strategies
Advancements
Challenges
Covert triggers altering predictions
Training data poisoning
Stealth enhancement methods like WABA and Poison Ink
Implications in diffusion models and copyright infringement
Detection and mitigation focus
Progressive Backdoor Erasing (PBE)
Sanitizer in federated learning
Advanced statistical methods and purification techniques
Systematic categorization of backdoor attacks
Poison Ink technique
PBE method effectiveness
Detection difficulties without degrading performance
Lack of comprehensive evaluations
Need for advanced defense algorithms

Figure 5: This figure illustrates the hierarchical structure of backdoor attacks and defenses, highlighting the nature and implications of attacks, along with recent advancements and persistent challenges in defense strategies.

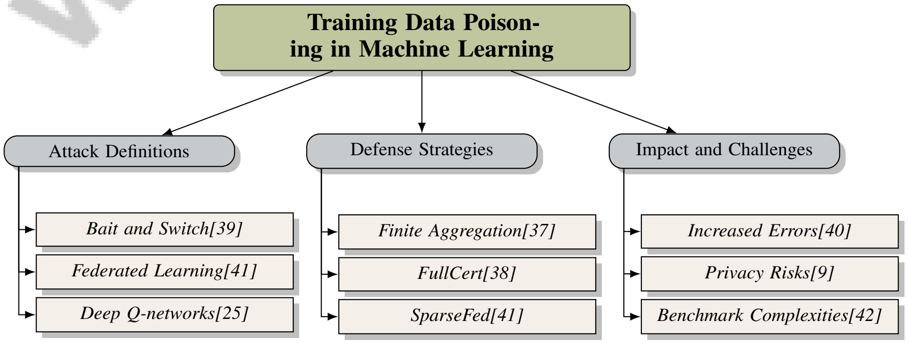


Figure 6: This figure illustrates the hierarchical categorization of training data poisoning in machine learning, highlighting attack definitions, defense strategies, and the impact and challenges associated with these threats.

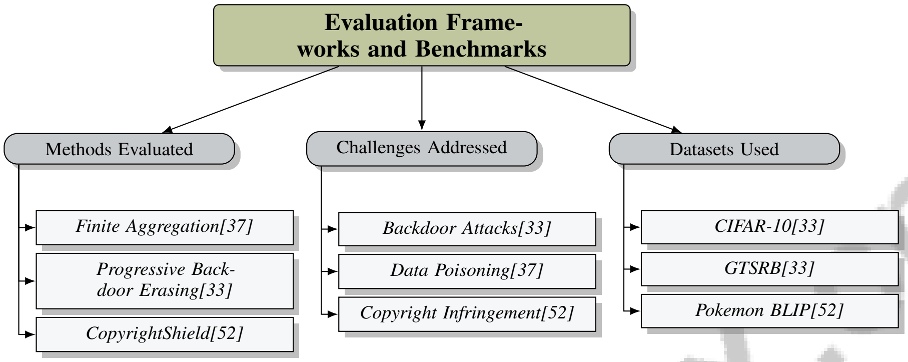


Figure 7: This figure illustrates the evaluation frameworks and benchmarks used to assess various methods against data poisoning attacks, highlighting the methods evaluated, challenges addressed, and datasets used.

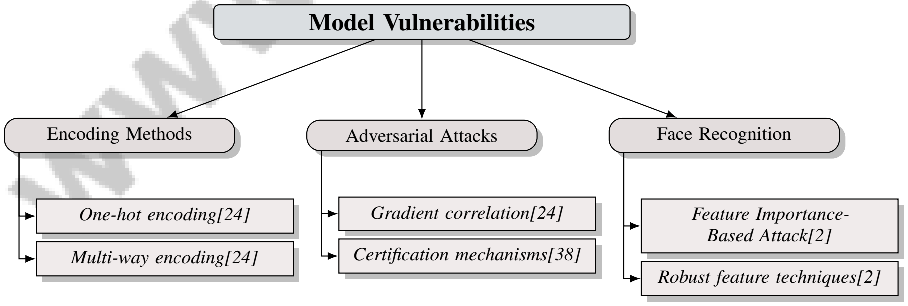


Figure 8: This figure illustrates the key aspects of model vulnerabilities, focusing on encoding methods, adversarial attacks, and face recognition systems. It highlights the role of one-hot and multi-way encodings, the impact of gradient correlation and certification mechanisms, and the challenges posed by feature-based attacks in face recognition.