Recent Trends in Adversarial Machine Learning (Review from ACM CCS Papers from 2022-2025)

Following the historical development in adversarial machine learning, with increasingly sophisticated gradient-based attacks and defences, this section explores research published at the ACM conference on Computer and Communications Security (CCS) from 2022 through 2025. This was inspired by Appruse et al, "Real attackers don't compute gradients", who conducted a similar study of gaps between research and industry until 2022. Our study reveals a complex landscape in the most recent days, where some research shows promising movement toward practical concerns, while others continue patterns of theoretical complexity and worst-case scenarios.

1 2022: Privacy Attacks and Practical Defences

The research landscape in 2022 demonstrated a dual focus: on one hand, significant advances in translating theoretical tools into practical defenses, and on the other hand, a major escalation in privacy attacks that exploited internal model mechanisms previously considered harmless. This duality represents both the promise and the challenge facing the field.

1.1 Bridging Theory and Practice in Defence

Hammoudeh and Lowd (2022) made a significant contribution in translating theoretical tools into practical defenses. They focused on influence estimation methods, which are techniques designed to identify which training samples most affect a model's predictions and can potentially be used to detect malicious training data. They identified a fundamental flaw in traditional influence estimation methods (Influence Functions, TracIn). These methods favoured high-loss points, causing them to systematically overlook the low-loss points. This suggests that these theoretical tools couldn't reliably detect poisoning attacks in practice. They used a renormalisation technique to remove the bias

caused by the low-loss penalty. The resulting Gradient Aggregated Similarity (GAS) detected 100% of malicious training instances, often with zero false positives on clean data (Hammoudeh and Lowd 2022). This demonstrates how refining theoretical concepts for practical applications can yield promising results.

1.2 Privacy Attack Escalation

The 2022 research focused on a major escalation in privacy attacks that exploited internal model mechanisms, which were previously considered harmless. These attacks demonstrated that multiple aspects of machine learning systems—from training dynamics to model architectures to explanation mechanisms—could be weaponized to extract sensitive information.

1.2.1 Training Data Manipulation and Privacy Leakage

Weaponised Data Poisoning: Tramer et al. (2022) demonstrated that poisoning attacks can be used to maximise privacy leakage, not just harm integrity. Specifically, they showed that adversaries could strategically poison training data to make it easier to determine whether specific data points were included in the training set (membership inference attacks, MIA) or to infer specific attributes about training data (Attribute Inference Attacks, AIA). These attacks maximise the statistical distinguishability of losses between members and non-members training set, enhancing membership inference attacks (MIA) and Attribute Inference Attacks (AIA). This work revealed that data poisoning could serve dual purposes—both corrupting model behavior and amplifying privacy vulnerabilities.

Loss Trajectory Exploitation: Liu et al (2022a) demonstrated new MIA techniques that exploit a sample's loss trajectory across different epochs rather than the final loss value. This is effective because non-member samples may exhibit fundamentally different loss histories than members, even if the final loss is small, thereby strengthening the membership inference. This approach moved beyond simple overfitting indicators to exploit the temporal dynamics of training.

1.2.2 Architectural Vulnerabilities

Multi-Exit Network Vulnerabilities: Multi-exit networks are neural network architectures designed to improve energy efficiency by allowing early termination of inference at intermediate "exit points" when confidence is high, rather than always processing through the entire network. Li et al (2022) quantified privacy risks in multi-exit networks. These energy-efficient networks are prone to membership inference attacks because the adversary can access the specific exit depth used for inference. Research has shown that these networks are even more vulnerable to MIA than traditional models. This finding high-

lighted how architectural innovations designed for efficiency could inadvertently introduce new privacy risks.

1.2.3 Understanding Privacy Mechanisms

Analysis of Privacy Leakage: Baluta et al (2022) moved privacy research from simply observing correlations (between overfitting and MIA success) to using causal reasoning via different frameworks (like ETo, for example, which is used to quantify and confirm the true causes of privacy leakage in ML models). This shift from correlation to causation represented an important methodological advancement in understanding privacy vulnerabilities.

1.2.4 Novel Attack Vectors

Inference from Explanations: Shapley values are a popular method from game theory used to explain machine learning predictions by assigning importance scores to input features. Luo et al. (2022) demonstrated that adversaries can infer private input features by exploiting the Shapley values that are exposed by Machine Learning as a Service (MLaas) platforms. This demonstrated that even mechanisms designed to improve model transparency could become attack vectors.

Graph Structure Recovery: Graph neural networks and embedding methods (GNNs, DeepWalk, Node2Vec) convert graph structures into vector representations (embeddings) that are supposed to hide the original graph topology. Shen et al (2022) showed that private graph structures like edges and node properties can be reconstructed from opaque node embeddings generated by graph models (GNNs, DeepWalk, Node2Vec) through the MNEMON attack. This attack challenged assumptions about the privacy-preserving properties of embedding spaces.

Automated Attack: Cretu et al (2022) developed QuerySnout, which automatically detects Attribute Inference Attacks against privacy-preserving query-based systems. The discovery works even when the systems use significant noise additions. This automation lowered the barrier for conducting sophisticated privacy attacks.

1.3 Expanding the Attack Surface Beyond Digital Domains

The research in 2022 also demonstrated that adversarial threats extended beyond purely digital contexts into physical and multimedia domains, challenging the traditional scope of adversarial machine learning.

1.3.1 Physical World Attacks

Physical Attacks: Siamese networks are a type of neural network architecture commonly used in object tracking that learns to match visual patterns across video frames. Muller et al (2022) demonstrated physical Hijacking attacks against Siamese object trackers used in autonomous driving. These attacks generate a pattern causing tracked objects to be missed in the real world, contrary to older digital attacks that ignore physical constraints. This work highlighted the critical safety implications of adversarial attacks in real-world deployment scenarios.

1.3.2 Audio and Multimedia Attacks

Audio Attacks: Liu et al (2022 b) demonstrated Targeted Adversarial Voice Over IP Network Attacks (TAINT). This attack successfully bypasses commercial audio services like Google Assistant through noisy communication channels like Voice over Internet Protocol (VoIP). The success of attacks through degraded channels demonstrated robustness beyond laboratory conditions.

Perception Aware Music Attacks: Duan et al (2022) generated adversarial audio signals bypassing music copyright detectors while the difference in audio is imperceptible to human listeners. This was achieved by modelling and reversing human perception. This work showed how understanding human perceptual systems could be exploited to create more effective attacks.

1.3.3 Attacks on Information Systems

Search Ranking Manipulation: Liu et al (2022c) demonstrated the order-disorder against search engines. They imitated the ranking and injected stealthy, fluent adversarial triggers (PAT) to manipulate search results. This attack demonstrated vulnerabilities in critical information retrieval systems.

Visual Attacks on Non-Classification Systems: Liu et al (2022d) demonstrated adversarial patch attacks against Crowd Counting models and achieved a high Mean Absolute Error (MAE) and transferability across different counting models. This expanded the scope of adversarial attacks beyond traditional classification tasks.

1.4 Intellectual Property Concerns

As machine learning models became increasingly valuable assets, protecting and stealing model intellectual property emerged as a significant concern.

Stealing Encoders: Self-supervised learning (SSL) produces encoder models that learn general-purpose representations from unlabeled data, which can then be fine-tuned for various downstream tasks. Liu et al (2022 e) showed the StolenEncoder attack for

stealing pre-trained encoders used in self-supervised learning via repeated querying of the target model. SSL encoders are very expensive to train and generalise to downstream tasks. Hence, their vulnerability and security is important. This highlighted the economic motivations behind model extraction attacks.

IP Protection: Cong et al (2022) proposed SSLGuard as a defence focusing on watermarking SSL pre-trained encoders to protect the intellectual property against model stealing attacks. This represented an early attempt to address the emerging threat of model theft.

1.5 Summary of 2022 Trends

The 2022 research landscape painted a picture of expanding attack surfaces and evolving threat models. While Hammoudeh and Lowd's work demonstrated the potential for practical defenses, the majority of papers documented new vulnerabilities across diverse domains—from privacy to physical safety to intellectual property. This year established a foundation for understanding the breadth of adversarial threats, though questions about real-world practicality and economic feasibility remained largely unaddressed.

2 2023: Acknowledging Constraints in Real-World

2023 research in Adversarial Machine Learning showed some promise in practical adversarial settings, the direction of real-world and constrained attacks that are relevant to industry. This year marked a subtle but important shift: researchers began explicitly acknowledging and addressing resource constraints, limited information scenarios, and realistic threat models that better reflected actual deployment conditions.

2.1 Stealthy Data Poisoning Attacks with Minimum Resources

Clean-label backdoor attacks are a particularly stealthy form of data poisoning where the poisoned samples maintain correct labels, making them extremely difficult to detect. Traditional backdoor attacks assumed adversaries could control large portions of the training data. Zenget al (2023) introduce Narciissus, a clean-label blackdoor attack requiring only target-class data and public out-of-distribution (POOD) data rather than the entire training set, which was assumed in previous papers. This attack received a success rate of 30.33x to 64.45x higher than competitors, with a modification of only 0.05% of training data. This trigger creates durable features and makes defence attempts hurt the model's clean accuracy (Zeng et al, 2023). This is an important step towards realistic threat modelling where adversaries have limited control over the training data. The significance of Narcissus lies not just in its effectiveness, but in its demonstration that powerful attacks need not require unrealistic levels of access or control.

2.2 Zero Knowledge Evasion Attacks

Moving beyond attacks that assume substantial knowledge of target systems, several 2023 papers explored scenarios where adversaries operate with minimal or no information about their targets. In traditional adversarial ML research, attackers are often assumed to know the model architecture, training procedure, or feature space. The "zero knowledge" setting represents a more realistic scenario where attackers must operate without these advantages.

He et al (2023) developed a query-based attack against ML-based Android Malware Detection (AMD) systems, which are machine learning classifiers used to identify malicious Android applications. This attack assumes zero knowledge, i.e the target permissions, activities, and function call graphs are unknown to the adversary. This framework explores the vast heterogeneous perturbation space using semantic depth. This demonstrated that even without insider knowledge, adversaries could systematically probe and evade detection systems.

Miao et al (2023) demonstrated evasion attacks against Google's phishing detector through adversarial screenshot generation. They used inverse downsampling to bypass detection mechanisms in black-box settings. This work showed that even well-resourced commercial systems remained vulnerable to carefully crafted evasion attempts.

2.3 Intellectual Property Theft with Limited Access

The 2023 research on model stealing reflected increasingly realistic scenarios where adversaries have only black-box query access to target systems.

Decoding algorithms determine how language models generate text by selecting which tokens to output at each step. Different algorithms (like Greedy Search, which picks the most likely token, or Nucleus Sampling, which samples from a probability distribution) produce different generation behaviors and are often proprietary implementation details. Naseh et al (2023) demonstrated that query-based attacks can accurately steal decoding algorithms like Greedy Search, Beam Search, k-Sampling and Nucleus Sampling and their hyperparameters from black-box language model APIs (GPT , Neo) at a low cost. This revealed that even implementation details could be inferred through careful query analysis.

Qin et al (2023) demonstrated that substitute model training attacks can successfully steal core model functionality under label-only black box settings (even when the models are trained to satisfy explicitly security properties like monotonicity using frameworks like DL2 or LogicEnsemble). These attacks take into consideration the practical commercial threats to deployed ML systems. This work challenged assumptions that security properties would provide protection against model extraction.

2.4 Defence Evaluation and Adaptive Adversaries

A critical theme emerging in 2023 was the recognition that defenses evaluated only against non-adaptive adversaries provided a false sense of security.

2.4.1 Stateful Defence Vulnerabilities

Stateful defenses are security mechanisms that maintain memory of past queries to detect and block adversarial probing. For example, they might reject new queries that are too similar to previous ones, aiming to prevent iterative black-box attacks.

Stateful Defence Vulnerabilities: Feng et al(2023) showed that Stateful Defence Models (SDMs) like Blacklight, designed to prevent black-box attacks by rejecting similar queries, are not secure against adaptive adversaries. The Oracle Guided Adaptive Rejection Sampling (OARS) attack exploits leaked information about the SDMs' similarity detection procedure to evade query collisions and achieve above 99% attack success rates in various underlying black-box methods. Hence, the defences that were evaluated only against non-adaptive adversaries fail in the adaptive scenarios. This highlighted a fundamental methodological flaw in defense evaluation practices.

2.4.2 Federated Learning Defences

In federated learning, multiple clients collaboratively train a model without sharing their raw data. A common challenge is that each client's data is non-IID (not independent and identically distributed)—meaning different clients have systematically different data distributions, which complicates both training and security.

Federated Learning Defences: Federated Learning defences are unable to handle adaptive adversaries. Kraub and Dmitrienko (2023) developed MESAS defences that address this using a cascade of six metrics (COUNT, VAR, MIN, MAX) with statistical tests that prune poisoning attacks. MESAS demonstrates robustness against adaptive attackers and realistic non-IID settings where other defences fail. This work represented a more realistic approach to federated learning security.

2.4.3 Privacy Mechanisms as Attack Vectors

Privacy Mechanisms as Attack Vectors: Arazzi et al (2023) showed that privacy-preserving mechanisms can be exploited. In Federated Learning, the Graph Neural Networks (FL-GNNs) that are intended for privacy protection can themselves be exploited by attackers to conduct backdoor attacks. Adding privacy measures without security and robustness analysis opened new vulnerabilities. This paradoxical finding demonstrated that security mechanisms must be holistically evaluated rather than added in isolation.

2.5 Holistic Privacy Evaluation Needs

Differential privacy provides mathematical guarantees about privacy leakage, but implementing it in machine learning (DPML) involves trade-offs between privacy, utility (model accuracy), and robustness. Wei et al (2023) identified that many Differential Private Machine Learning Algorithms (DPML) have been evaluated in isolation without comparison across utility and defence capability against Membership Inference Attacks. Their DPML-Bench framework proposes holistic measurement approaches categorising improvements across the ML pipeline (Data Preparation, Model Design, Model Training, Model Ensemble). This addresses the gap between theoretical privacy guarantees and practical deployment tradeoffs. This framework represented a move toward more comprehensive and realistic evaluation methodologies.

2.6 Evasion Against AI-Generated Content Detection

As AI-generated content became more prevalent, watermarking emerged as a leading approach to track and identify AI-generated text and images. Watermarks are imperceptible signals embedded in generated content that can later be detected to verify AI authorship.

Jiang et al (2023) proposed Optimised Perturbation Attacks (WEvade) that bypass watermark-based AI-Generated Content (AIGC) detections. WEvade works for both white-box and black-box settings, demonstrating the vulnerability of proprietary detection systems against targeted evasions. This challenged the reliability of watermarking as a defense mechanism for AI-generated content.

2.7 Verifiable Security and Controlled Generation

In contrast to the arms race of attacks and defenses, some 2023 research explored fundamentally different approaches based on formal verification and provable guarantees.

2.7.1 Verifiable Learning for Provable Robustness

Verifiable Learning for Provable Robustness: Rather than empirically testing whether models resist attacks, verifiable learning seeks mathematical proofs that models satisfy security properties. This requires restricting the types of models used to those amenable to formal analysis. Calzavara et al (2023) introduced learning restricted model classes that are designed for efficient verification. They identified large spread decision tre ensembles as a restricted class admitting polynomial time security verification algorithms against Lp norm evasion attacks. This represents a shift toward formal security guarantees rather than empirical robustness. By constraining model architectures to enable verification, this work traded some flexibility for provable security properties.

2.7.2 Controlled Code Generation

Controlled Code Generation: Common Weakness Enumeration (CWE) is a standardized list of software security vulnerabilities and coding errors. Controlling code generation to avoid or deliberately include these weaknesses is important for both security hardening and security testing. He and Vechev (2023) developed the SVEN framework, enabling Large Language Models to perform controlled code generation by guiding the model using prefixes to generate code that adheres to or violates the specific security properties, like CWEs. This enforces security hardening and adversarial testing while maintaining functional correctness. This approach enabled both defensive and offensive security testing in an integrated framework.

2.7.3 Verifiable Systems and Integrity

Verifiable FL Transactions: Lie et al (2023) proposed martFL, a utility-driven data marketplace using FL with verifiable transaction protocols. This addressed the need for trustworthy data exchange in federated settings.

Training Integrity Guarantees: Training forgettability refers to whether training checkpoints retain evidence of specific data points or training events. This matters for both privacy (ensuring deleted data is truly forgotten) and accountability (proving what data was used). Baluta et al (2023) explored formal characterisation of SGD to guarantee the forgettability of checkpoints during training. They introduced algebraic conditions (LSB check), providing foundations for applications requiring data non-repudiation and training integrity auditing. This work provided theoretical foundations for proving properties about the training process itself.

2.7.4 LLCG IP Protection

LLCG IP Protection: Lie et al (2023b) developed the ToSyn framework, injecting stealthy watermarks into Large Language code Generation API outputs by replacing tokens with synonyms using Built-in Structure Replacement and Code Style Transferring. This allows service providers to use JSD statistical tests to confirm IP theft in suspect imitation models. This addressed the growing concern about code generation model theft.

2.8 Additional Security Concerns

Several additional papers addressed emerging threats and specific security challenges across various domains.

Multi-Domain Trojan Detection: Rajabi et al (2023) developed MDTD, a multi-domain trojan detector for deep neural networks, addressing detection across diverse

deployment scenarios. This recognized that models often operate across multiple domains in practice.

Unsafe Content Detection: Qu et al (2023) studied unsafe images and hateful memes from text-to-image models, highlighting concerns in generative AI deployment. As generative models became more powerful, their potential for generating harmful content became a pressing concern.

Speech Synthesis Prevention: Yu et al (2023) introduced AntiFake using adversarial audio to prevent unauthorised speech synthesis, addressing emerging deepfake threats. This represented a proactive defense against audio deepfakes.

Privacy Attacks: Shi et al (2023) deconstructed privacy leakage via speech-induced vibrations on room objects through remote sensing Phased-MIMO, a novel attack vector beyond digital interfaces. This demonstrated that privacy threats could emerge from unexpected physical side channels.

GNN Privacy Breaches: Meng et al (2023) showed breaching graph neural networks' privacy through an infiltration attack, demonstrating vulnerabilities of emerging architecture. As GNNs gained prominence, understanding their unique security challenges became important.

AI-Generated Content: Sha et al (2023)developed De-Fake to detect and attribute fake images generated by text-image models. This contributed to the growing body of work on generative AI security.

Selective Adversarial Strategies: Belavadi et al (2023) developed attack strategies to attack some areas while protecting others, introducing nuanced threat models beyond universal perturbations. This recognized that adversaries might have selective goals rather than causing universal disruption.

2.9 Summary of 2023 Progress

The 2023 research addresses meaningful progress in real-world attack scenarios through resource-constrained adversaries, low-Dat resource poisoning and realistic defences against adaptive adversaries. However, the fundamental theoretical reliance persists, and none of the papers have addressed economic constraints, the purpose of attackers and their intentions. While the field showed promising movement toward practical concerns, the disconnect between academic threat models and real-world attacker motivations remained a significant gap.

3 2024 Confronting Practical Deployment Barriers

The year 2024 marked a more explicit engagement with the practical realities facing machine learning practitioners. Rather than assuming ideal conditions or unlimited re-

sources, researchers began directly addressing the constraints and limitations that characterize real-world ML deployment.

3.1 Acknowledging Practitioner Constraints

Bagdasaryan and Shmatikov (2024) directly addressed the industry perspective, emphasising the mismatch between theoretical assumptions and practical threats. They noted that state-of-the-art backdoor defences often require massive changes to existing ML pipelines, which real-world ML engineers may not be equipped to implement or maintain due to limited resources or expertise. This observation highlighted a critical gap: even effective defenses remain unused if they demand unrealistic engineering effort or expertise.

The Mithridates tool was proposed to provide pragmatic tools for ML engineers to audit and boost the inherent resistance of their existing pipelines to unknown backdoor attacks via hyperparameter search and regularisation without requiring disruptive modifications (Bagdasaryan and Shmatikov, 2024). By working within existing pipeline constraints, this approach recognized that practical defense must accommodate real-world engineering limitations.

3.2 Data-Free Scenarios

Adversarial training is a standard defense technique that involves training models on adversarially perturbed examples to improve robustness. However, this typically requires access to the original training data to generate these adversarial examples.

Bai et al (2024) challenged the assumption that the defender has access to the original training data. In modern scenarios, users frequently acquire pre-trained models from open-source platforms like Hugging Face or ModelScope, but the original training data are unavailable due to intellectual property or privacy concerns. Lacking the initial training data prevents engineers from employing standard adversarial training (AT) methods to enhance model robustness. This scenario increasingly characterizes real-world model deployment, where transfer learning and pre-trained models dominate.

They propose Alchemy, Data-Free-Adversarial Training for Models lacking the original training dataset, enabling robustness improvement in common real-world contexts (Bai et al, 2024). This work addressed a practical barrier that prevents many practitioners from implementing adversarial defenses.

3.3 Evaluation Methodology Challenges

Aerni et al (2024) challenged the evaluations of ML privacy defences. Research evaluations commonly use weak inference attacks that do not reflect the current state-of-the-art, or they fail to properly adapt attacks to complex defence components. This created a false

sense of security, where defenses appeared effective against weak baselines but would fail against more sophisticated attacks.

3.3.1 LLM Vulnerability Assessment

LLM Vulnerability Assessment: Many probability-based attacks are easily averted if the LLM only returns generated text outputs, which is usually the case in real-world scenarios. Many academic attacks against language models assume access to output probabilities or logits (the model's confidence scores), but deployed systems typically only return generated text. This calls to action the development of text-only inference attacks (GAP, Inquiry, Repeat, BrainWash) to evaluate true privacy risk (Aerni et al, 2024). This highlighted the importance of evaluating defenses under realistic access conditions rather than assuming unrealistic levels of model access.

Black box attack, where the adversary has access to the target outputs for the real-world outputs are considered the most real-world scenarios and are more relevant compared to white box attacks. This represented a shift toward more realistic threat modeling in evaluation.

3.4 Advanced Threat to Modern Systems

Several 2024 papers identified novel vulnerabilities in emerging ML paradigms and systems.

Split-Based Framework Vulnerabilities: Split learning is a privacy-preserving approach where model layers are divided between client and server, with the goal of keeping sensitive data on the client side while still enabling fine-tuning. Chen et al (2024) revealed vulnerabilities in private fine-tuning within a split-based framework for large language models through a bidirectionally enhanced attack. As new privacy-preserving training methods emerged, so did new attack vectors.

Data Use Auditing: Huang et al (2024) provided a general framework of data-use auditing of ML models. This addressed the growing need for accountability and transparency in data usage.

Watermark removal: Lu et al (2024) demonstrated neural dehydration, which effectively erases blackbox watermark from DNNs with limited data. This challenged assumptions about the permanence of model watermarks.

Hyperparameter Stealing: Trusted Execution Environments (TEEs) are hardware-isolated secure enclaves designed to protect sensitive computations. Even when models are protected within TEEs, they must return query results, and these outputs can leak information. Yuan et al (2024) introduced HyperTheft, stealing hyperparameters of DNs from TEE Shielded Black Box Query Results. This showed that even hardware-protected systems could leak sensitive information.

In-Context-Learning Privacy: In-context learning is a capability of large language models where they learn to perform tasks from examples provided in the prompt, without updating model parameters. This raises questions about whether these in-context examples can leak information. Wen et al (2024) demonstrated membership inference attacks against in-context learning. As new learning paradigms emerged, their privacy implications required investigation.

Robust Malware Detection: Lucas et al (2024) presented methods for training robust ML-based raw binary malware detectors efficiently. This addressed the critical need for adversarially robust security tools.

3.5 Summary of 2024 Developments

The 2024 research landscape demonstrated increasing awareness of practical deployment challenges. From acknowledging data access limitations to questioning evaluation methodologies to designing defenses that work within existing pipelines, researchers showed greater engagement with real-world constraints. However, the fundamental tension remained: while individual papers addressed specific practical barriers, the field as a whole continued to prioritize sophisticated attacks and defenses that required substantial expertise and resources to implement.

4 2025: Stealth, Transferability and Practical Defences

The most recent research in 2025 continued the trend toward acknowledging practical constraints while simultaneously exploring increasingly sophisticated attack mechanisms. This dual character—greater realism alongside greater technical complexity—characterized the current state of the field.

4.1 Novel Stealthy Attacks

4.1.1 Parameter Space Backdoors

Parameter Space Backdoors: Traditional backdoor detection methods analyze model parameters looking for anomalies introduced by backdoor triggers. A parameter space backdoor aims to evade this detection by making the backdoor's parameter signature indistinguishable from normal model parameters. Xu et al (2025a) introduced Grond, a backdoor attack designed for comprehensive stealthiness in model parameter space. Existing backdoor attacks often fail when faced with parameter space defences like pruning or fine-tuning, but Grond uses the adversarial Backdoor Injection module to maintain effectiveness against these advanced defences. This represented an escalation in the sophistication of backdoor attacks, specifically designed to evade detection mechanisms that

analyze model parameters.

4.1.2 SuperNet Targeted Attacks

SuperNet Targeted Attacks: Universal adversarial perturbations are input modifications that can fool a model regardless of the original input, while targeted attacks aim to make the model produce a specific wrong output. CLIP is a widely-available vision-language model that can be used as a surrogate for generating transferable attacks. Xu et al (2025b) presented UnivIntruder, which generates transferable universal targeted adversarial perturbations using a single publicly available CLIP model and public datasets. This technique is highly successful and transferable across black box Deep Neural Networks (DNNs) and complex Vision-Language Models (VLMs,) including real-world services like Google and GPT-4. The practical implications of this work are significant, as it demonstrated that attacks crafted against publicly available models could transfer to proprietary commercial systems.

4.1.3 Cascading Bias in LLMs

Cascading Bias in LLMs: Knowledge distillation is a technique where a smaller "student" model learns to mimic a larger "teacher" model, often used to create more efficient deployable models. The concern is whether biases or vulnerabilities in the teacher propagate to the student. Chaudhary et al (2025) explored the propagation of adversarial bias in language models through the BIASED_ROOTS poisoning attack. The attack injects bias into the teacher model, which then propagates and amplifies into the distilled student model, achieving a higher Adversarial Response Rate (ARR) in the student (76.9%) compared to the teacher (69.4%) with the minimal poisoning of training data. This revealed a troubling vulnerability in knowledge distillation: biases could not only transfer but actually amplify across generations of models.

4.1.4 Autonomous Driving Attacks

Autonomous Driving Attacks: Modern autonomous driving systems use online mapping, where the vehicle continuously constructs a local map of road geometry from sensor data in real-time, rather than relying solely on pre-loaded maps. These online map modules process camera and sensor inputs to identify lane boundaries, road edges, and spatial relationships. Lou et al (2025) focused on real-world safety risks exploiting an asymmetry vulnerability in road geometry processing used by online map modules (like MapTR) in Autonomous Driving (AD) systems. Physical attacks such as camera binding or placing adversarial patches can manipulate the constructed map to induce malicious objectives, for instance, road straightening or turning left early. This work demonstrated concrete

safety implications in high-stakes applications, moving beyond abstract security concerns to scenarios where adversarial attacks could cause physical harm.

4.1.5 Watermark Removal in AIGC

Watermark Removal in AIGC: Latent Diffusion Models (LDMs) are a class of generative models that create images by iteratively denoising in a compressed latent space, starting from a random or watermarked initial point. Embedding watermarks in this starting point was thought to make them robust to removal. Lee et al (2025) demonstrated a watermark removal attack targeting watermarks embedded in the latent starting point of Latent Diffusion Models (LDMs). This suggests that undetectability doesn't guarantee robustness against removal attacks. The finding challenged assumptions about watermark permanence in generative models.

4.2 Robust Defences for Deployment

In contrast to the sophisticated attacks, several 2025 papers proposed defenses explicitly designed for realistic deployment scenarios with significant practical constraints.

4.2.1 Data Free-Federated Learning Defence

Data Free-Federated Learning Defence: Conditional Generative Adversarial Networks (CGANs) are generative models that can synthesize data conditioned on class labels. In federated learning, the central server typically cannot access client data, making traditional defense methods that require inspection of training data infeasible. Yang et al (2025) introduced FilterFL, a defence against backdoor attacks in federated learning. FilterFl uses Conditional Generative Adversarial Networks (CGANs) to extract incremental knowledge and filter out the backdoor components. This enables the detection and exclusion of poisoned client models without requiring access to the client data. FilterFL demonstrates superiority against various state-of-the-art FL backdoor attacks, minimising the Attack Success Rate while maintaining a high model accuracy. The significance of this work lies in its recognition that federated learning coordinators typically cannot access client data, making data-free defense essential.

4.2.2 Personalized Edge Defense

Personalized Edge Defense: Low-rank adaptation (LoRA) is a parameter-efficient fine-tuning technique that only updates a small number of parameters, making it feasible for resource-constrained devices. This is particularly important for edge devices which have limited computational power and memory. Qi et al (2025) presented Sylva for tailoring personalised adversarial defence in pre-trained models via collaborative fine-tuning. It

employs low-rank adaptation (LoRA) based adversarial fine-tuning and adaptive aggregation to overcome heterogeneity challenges, improving personalised adversarial robustness and accuracy on resource-constrained edge devices. This work acknowledged the reality of edge deployment, where computational resources are limited and one-size-fits-all defenses may not be feasible.

4.2.3 Robust Malware Detection

Robust Malware Detection: Packed executables are programs that have been compressed or obfuscated using packing software, making their actual code difficult to analyze. Malware authors commonly use sophisticated packing techniques to evade detection by static analysis tools. Li et al (2025) introduced Pack-ALM to detect packed executables specifically targeting high evasive adversarial packing techniques. Pack-ALM provides a good defence against such a technique, where simple entropy checks achieve a 0% detection accuracy. This addressed a practical challenge in malware detection where standard approaches completely fail against sophisticated evasion.

4.2.4 Handling Imperfect Labels

Handling Imperfect Labels: VirusTotal aggregates results from dozens of different antivirus engines, each with their own detection heuristics and false positive rates. When these engines disagree, the resulting labels become noisy—the same file might be flagged as malicious by some engines and benign by others. Alotaibi et al (2025) focused on the problem of label noise. These problems are common when malware datasets are automatically labelled via antivirus engine consensus, like VirusTotal reports. The proposed SLB robustly trains deep learning classifiers under this label noise and produces corrected labels. This significantly improves the performance of deep learning and conventional ML models. This work recognized that real-world datasets often contain noisy or incorrect labels, a practical challenge rarely addressed in adversarial ML research.

4.3 Summary of 2025 Developments

The 2025 research demonstrated both the continuing sophistication of attacks and meaningful progress toward practical defenses. Works like FilterFL, Sylva, and Pack-ALM explicitly addressed real-world deployment constraints, while attack papers like UnivIntruder revealed concerning transferability across systems. The field appeared to be at a crossroads, with some research grounded in practical realities and other work pursuing increasingly complex theoretical scenarios.

5 The Gap Between Research and Industry

Despite increased awareness of practical constraints, most of the recent attacks and defences from 2022 to 2025 still rely on gradient-based methodologies. A surrogate model is a substitute model that approximates the target system's behavior, allowing attackers to compute gradients locally before transferring attacks to the target. Query budgets refer to the number of times an attacker can query a target system before being detected or rate-limited. The physical attacks (Muller et al, 2022; Lou et al 2025), black box attacks (Liu et al, 2022c) and transfer attacks (Xu et al 2025) all claim increased realism; however, most of them still employ surrogate models, extensive query budgets, or sophisticated statistical inference, all of which require significant ML expertise and computational resources. This fundamental reliance on gradient-based methods reflects a persistent assumption that adversaries possess both the technical sophistication and resources to implement such attacks.

The fundamental question remains unaddressed: would actual attackers driven by economic motives employ such methods rather than exploiting operational gaps or simple input manipulation? Real-world attackers typically seek the path of least resistance, exploiting misconfigurations, social engineering, or simple input perturbations rather than implementing sophisticated gradient-based attacks. The academic focus on worst-case scenarios with highly capable adversaries may overestimate the threat from gradient-based methods while underestimating simpler but more economically viable attack vectors.

5.1 Progress in Addressing Data Access Limitations

Some research has shown that practitioners lack original training data:

- Narcissus (2023) requires only the target class and public data
- Alchemy (Bai et al, 2024): Provides data-free adversarial training
- FilterFL (Yang et al, 2025): enables backdoor detection without client data

These are meaningful progress toward realistic threat modelling and practical defences in real-world deployments. By relaxing assumptions about data access, these works addressed a critical barrier that prevents many practitioners from implementing security measures. This represents perhaps the most concrete progress toward bridging the research-practice gap.

5.2 Limited Economic Considerations

Research from 2022-2025 has limited engagement with economic and operational considerations that drive actual security decisions. While Mithridates (Bagdasaryan and

Shmatikov 2024) explicitly discuss organisational constraints, most papers rarely discuss:

- cost-benefit analysis of implementing proposed defences
- opportunity costs of security investments
- economic incentives driving real-world attacks

This gap is particularly problematic because security decisions in practice are fundamentally economic decisions. Organizations must balance security investments against other priorities, and attackers choose methods based on return on investment. The lack of economic framing in academic research makes it difficult for practitioners to assess whether proposed defenses are worth implementing or whether studied attacks represent realistic threats given actual attacker incentives.

5.3 Improved Evaluation Methodology

Progress in evaluation methodology appears in

- stateful defences must be evaluated against adaptive adversaries (2023)
- weak baseline attacks lead to overconfident security promises (Aerni et al, 2024)
- Holistic evaluation frameworks for privacy and federated learning (2023-2024).

However, significant gaps persist:

- Defences are evaluated primarily against academic attacks rather than realistic adversaries
- Focus remains on worst-case scenarios under academic threat models
- Deployment metrics (implementation complexity, computational overhead) are rarely evaluated.

The improvement in evaluation methodology represents important progress, particularly the recognition that defenses must be tested against adaptive adversaries. However, the gap between academic attack sophistication and real-world attack simplicity means that even improved evaluation may not reflect actual security properties in deployment.

5.4 Increasing Depth and Specialization

The 2022-2025 papers reveal an increasing specialisation within the adversarial ML research. Privacy attacks, backdoor attacks, evasion attacks, certified robustness and federated learning security have evolved with their own subfields and specialised methodologies. While this specialization has enabled deep technical advances within each area, it has also created challenges for practitioners seeking to understand and address adversarial ML threats holistically.

Industry practitioners report lacking tactical and strategic tools for addressing adversarial ML threats and viewing security as outside their expertise (Kumar et al, 2020; Mink et al, 2023; Gross et al, 2024). Deepening technical sophistication in research has made it difficult for practitioners to translate research insights into deployable security. The growing complexity and specialization of academic research, while technically impressive, may actually widen rather than narrow the gap between research and practice.

5.5 Emerging Developments in Practical Contexts

2023-2025 papers suggest important insights into practical constraints

- Acknowledgement that clients possess non-IID data
- Recognitions that defenders can't access training or client data
- Understanding that attacks must succeed under limited poisoning capacities

FilterFL (Yang et al, 2025) explicitly designed for data-free settings, represents a meaningful step. Similarly, Sylva (Qi et al, 2025) for resource-constrained edge devices acknowledges actual deployment constraints. Emerging deployment paradigms are developing on more practical grounds than earlier adversarial ML research. These developments suggest that certain subfields, particularly federated learning and edge computing, are moving more rapidly toward practical relevance than the field as a whole.

6 Conclusion

The analysis of ACM CCS papers from 2022-2025 reveals a research landscape in partial transition. Positive developments include:

- **Increased awareness**: More explicit acknowledgement of practical constraints
- **Data access**: Meaningful advances in addressing data access limitations
- **Economic factors**: Some engagement with IP theft and economic motivations of attackers

- **Evaluation Improvements**: Better methodologies for adaptive attacks and privacy defences
- **Emerging paradigms**: More practical approaches in federated learning and edge computing

However, there are fundamental gaps that still need to be addressed:

- **Gradient based focus**: Continued dependence on gradient based methodologies despite practical scenarios
- **Sophisticated Adversaries**: Assumptions of adversaries with substantial resources and expertise
- **Economics**: Minimal engagement with operational and economic realities (economic incentive is the primary driver of attacks in the real world)
- **Increasing Complexity**: Deepening research specialisation leaves practitioners hard to adapt
- **Evaluation Gaps**: Focus on theoretical optimal scenarios that are realistic attack economics

Significant progress has been made in specific areas like federated learning, edge computing, where practical constraints have received constant attention. The research continues to prioritise worst-case scenario theory-based gradient attacks over realistic attacks and economic considerations. While individual papers make important contributions to understanding adversarial ML fundamentals, the collective research trajectory from 2022-2025 suggests that substantial gaps remain between academic research priorities and industry security needs.

The path forward requires not abandoning theoretical rigor but complementing it with greater attention to economic feasibility, practitioner constraints, and realistic threat models. Research that explicitly considers implementation complexity, computational overhead, and attacker incentives alongside technical sophistication would better serve both academic understanding and practical security. The field stands at a critical juncture where conscious effort toward practical relevance could dramatically increase the real-world impact of adversarial ML research.

References

[1] Aerni, M., Zhang, J., and Tramèr, F. (2024). Evaluations of Machine Learning Privacy Defenses are Misleading. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24)*.

[2] Alotaibi, F., Goodbrand, E., and Maffeis, S. (2025). Deep Learning from Imperfectly Labeled Malware Data. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security (CCS '25)*, October 13–17, 2025, Taipei, Taiwan. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3719027.3765197.

[3] Anonymous. (2025). VillainNet: Targeted Poisoning Attacks Against SuperNets Along the Accuracy-Latency Pareto Frontier. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security (CCS '25)*. ACM, New York, NY, USA, 18 pages.

[4] Arazzi, M., Conti, M., Nocera, A., and Picek, S. (2023). Turning Privacy-preserving Mechanisms against Federated Learning. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)*, November 26–30, 2023, Copenhagen, Denmark. ACM, New York, NY, USA. https://doi.org/10.1145/3576915.3623114.

[5] Bagdasaryan, E. and Shmatikov, V. (2024). Mithridates: Auditing and Boosting Backdoor Resistance of Machine Learning Pipelines. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24)*.

[6] Bai, Y., Ma, Z., Chen, Y., Deng, J., Pang, S., Liu, Y., and Xu, W. (2024). Alchemy: Data-Free Adversarial Training. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24)*.

[7] Baluta, T., Nikolić, I., Jain, R., Aggarwal, D., and Saxena, P. (2023). Unforgeability in Stochastic Gradient Descent. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)*, November 26–30, 2023, Copenhagen, Denmark. ACM, New York, NY, USA. https://doi.org/10.1145/3576915.3623093.

[8] Baluta, T., Shen, S., Hitarth, S., Tople, S., and Saxena, P. (2022). Membership Inference Attacks and Generalization: A Causal Perspective. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS '22)*.

[9] Belavadi, V., Zhou, Y., Kantarcioglu, M., and Thuraisingham, B. (2023). Attack Some while Protecting Others: Selective Attack Strategies for Attacking and Protecting Multiple Concepts. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)*, November 26–30, 2023, Copenhagen, Denmark. ACM, New York, NY, USA. https://doi.org/10.1145/3576915.3623177.

[10] Calzavara, S., Cazzaro, L., Pibiri, G. E., and Prezza, N. (2023). Verifiable Learning for Robust Tree Ensembles. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)*, November 26–30, 2023, Copenhagen, Denmark. ACM, New York, NY, USA. https://doi.org/10.1145/3576915.3623100.

[11] Chaudhari, H., Hayes, J., Jagielski, M., Shumailov, I., Nasr, M., and Oprea, A. (2025). Cascading Adversarial Bias from Injection to Distillation in Language Models. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security (CCS '25)*, October 13–17, 2025, Taipei, Taiwan. ACM, New York, NY, USA, 29 pages. https://doi.org/10.1145/3719027.3765122.

[12] Chen, G., Chen, Y., Tao, G., Hong, Y., Du, T., and Xu, Z. (2024). Unveiling the Vulnerability of Private Fine-Tuning in Split-Based Frameworks for Large Language Models: A Bidirectionally Enhanced Attack. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24)*.

[13] Cong, T., He, X., and Zhang, Y. (2022). SSLGuard: A Watermarking Scheme for Self-supervised Learning Pre-trained Encoders. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS '22)*.

[14] Creţu, A.-M., Houssiau, F., Cully, A., and de Montjoye, Y.-A. (2022). QuerySnout: Automating the Discovery of Attribute Inference Attacks against Query-Based Systems. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS '22)*.

[15] Duan, R., Qu, Z., Zhao, S., Ding, L., Liu, Y., and Lu, Z. (2022). Perception-Aware Attack: Creating Adversarial Music via Reverse-Engineering Human Perception. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS '22)*.

[16] Feng, R., Hooda, A., Mangaokar, N., Fawaz, K., Jha, S., and Prakash, A. (2023). Stateful Defenses for Machine Learning Models Are Not Yet Secure Against Blackbox Attacks. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)*, November 26–30, 2023, Copenhagen, Denmark. ACM, New York, NY, USA. https://doi.org/10.1145/3576915.3623116.

[17] Grosse, K., Schuett, J., Balaguer, J., and Brown, S. (2024). Industry Practitioners' Mental Models of Adversarial Machine Learning. ArXiv preprint.

[18] Hammoudeh, Z. and Lowd, D. (2022). Identifying a Training-Set Attack's Target Using Renormalized Influence Estimation. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS '22)*.

[19] He, J. and Vechev, M. (2023). Large Language Models for Code: Security Hardening and Adversarial Testing. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)*, November 26–30, 2023, Copenhagen, Denmark. ACM, New York, NY, USA. https://doi.org/10.1145/3576915.3623175.

[20] He, P., Xia, Y., Zhang, X., and Ji, S. (2023). Efficient Query-Based Attack against ML-Based Android Malware Detection under Zero Knowledge Setting. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)*, November 26–30, 2023, Copenhagen, Denmark. ACM, New York, NY, USA. https://doi.org/10.1145/3576915.3623117.

[21] Huang, Z., Gong, N. Z., and Reiter, M. K. (2024). A General Framework for Data-Use Auditing of ML Models. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24)*.

[22] Jiang, Z., Zhang, J., and Gong, N. Z. (2023). Evading Watermark based Detection of AI-Generated Content. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)*, November 26–30, 2023, Copenhagen, Denmark. ACM, New York, NY, USA. https://doi.org/10.1145/3576915.3623189.

[23] Krauß, T. and Dmitrienko, A. (2023). MESAS: Poisoning Defense for Federated Learning Resilient against Adaptive Attackers. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)*, November 26–30, 2023, Copenhagen, Denmark. ACM, New York, NY, USA. https://doi.org/10.1145/3576915.3623212.

[24] Kumar, R. S. S., Nyström, M., Lambert, J., Marshall, A., Goertzel, M., Comissoneru, A., Swann, M., and Xia, S. (2020). Adversarial Machine Learning-Industry Perspectives. In *2020 IEEE Security and Privacy Workshops (SPW)*.

[25] Lee, D. Z., Fang, H., Wang, H., and Chang, E.-C. (2025). Removal Attack and Defense on AI-generated Content Latent-based Watermarking. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security (CCS '25)*, October 13–17, 2025, Taipei, Taiwan. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3719027.3765175.

[26] Li, Q., Liu, Z., Li, Q., and Xu, K. (2023). martFL: Enabling Utility-Driven Data Marketplace with a Robust and Verifiable Federated Learning Architecture. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)*, November 26–30, 2023, Copenhagen, Denmark. ACM, New York, NY, USA. https://doi.org/10.1145/3576915.3623134.

[27] Li, S., Ming, J., Liu, L., Yang, L., Zhang, N., and Jia, C. (2025). Adversarially Robust Assembly Language Model for Packed Executables Detection. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security (CCS '25)*, October 13–17, 2025, Taipei. ACM, New York, NY, USA, 19 pages. https://doi.org/10.1145/3719027.3765157.

[28] Li, Z., Liu, Y., He, X., Yu, N., Backes, M., and Zhang, Y. (2022). Auditing Membership Leakages of Multi-Exit Networks. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS '22)*.

[29] Li, Z., Wang, C., Wang, S., and Gao, C. (2023b). Protecting Intellectual Property of Large Language Model-Based Code Generation APIs via Watermarks. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)*, November 26–30, 2023, Copenhagen, Denmark. ACM, New York, NY, USA. https://doi.org/10.1145/3576915.3623120.

[30] Liu, H., Chen, H., Yang, J., Liu, S., Yan, Q., and Wang, X. (2022b). When Evil Calls: Targeted Adversarial Voice over IP Network. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS '22)*.

[31] Liu, J., Kang, Y., Tang, D., Song, K., Sun, C., Wang, X., Lu, W., and Liu, X. (2022c). Order-Disorder: Imitation Adversarial Attacks for Black-box Neural Ranking Models. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS '22)*.

[32] Liu, S., Wang, J., Liu, A., Li, Y., Gao, Y., Liu, X., and Tao, D. (2022d). Harnessing Perceptual Adversarial Patches for Crowd Counting. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS '22)*.

[33] Liu, Y., Jia, J., Liu, H., and Gong, N. Z. (2022e). StolenEncoder: Stealing Pre-trained Encoders in Self-supervised Learning. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS '22)*.

[34] Liu, Y., Zhao, Z., Backes, M., and Zhang, Y. (2022a). Membership Inference Attacks by Exploiting Loss Trajectory. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS '22)*.

[35] Lou, Y., Hu, H., Song, Q., Xu, Q., Zhu, Y., Tan, R., Lee, W.-B., and Wang, J. (2025). Asymmetry Vulnerability and Physical Attacks on Online Map Construction for Autonomous Driving. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security (CCS '25)*, October 13–17, 2025, Taipei, Taiwan. ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3719027.3765092.

[36] Lu, Y., Li, W., Zhang, M., Pan, X., and Yang, M. (2024). Neural Dehydration: Effective Erasure of Black-box Watermarks from DNNs with Limited Data. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24)*.

[37] Lucas, K., Lin, W., Bauer, L., Reiter, M. K., and Sharif, M. (2024). Training Robust ML-based Raw-Binary Malware Detectors in Hours, not Months. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24)*.

[38] Luo, X., Jiang, Y., and Xiao, X. (2022). Feature Inference Attack on Shapley Values. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS '22)*.

[39] Meng, L., Bai, Y., Chen, Y., Hu, Y., Xu, W., and Weng, H. (2023). Devil in Disguise: Breaching Graph Neural Networks Privacy through Infiltration. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)*, November 26–30, 2023, Copenhagen, Denmark. ACM, New York, NY, USA. https://doi.org/10.1145/3576915.3623173.

[40] Miao, C., Feng, J., You, W., Shi, W., Huang, J., and Liang, B. (2023). A Good Fishman Knows All the Angles: A Critical Evaluation of Google's Phishing Page Classifier. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)*, November 26–30, 2023, Copenhagen, Denmark. ACM, New York, NY, USA. https://doi.org/10.1145/3576915.3623199.

[41] Mink, J., et al. (2023). "Security is Not My Field, I'm a Stats Guy": A Qualitative Root Cause Analysis of Barriers to Adversarial Machine Learning Defenses in Industry. In *32nd USENIX Security Symposium (USENIX Security 23)*.

[42] Muller, R., Man, Y., Celik, Z. B., Li, M., and Gerdes, R. (2022). Physical Hijacking Attacks against Object Trackers. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS '22)*.

[43] Naseh, A., Krishna, K., Iyyer, M., and Houmansadr, A. (2023). Stealing the Decoding Algorithms of Language Models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)*, November 26–30, 2023, Copenhagen, Denmark. ACM, New York, NY, USA. https://doi.org/10.1145/3576915.3616652.

[44] Qi, T., Xue, L., Zhan, Y., and Ma, X. (2025). Sylva: Tailoring Personalized Adversarial Defense in Pre-trained Models via Collaborative Fine-tuning. In *Proceedings*

of the 2025 ACM SIGSAC Conference on Computer and Communications Security (CCS '25).

[45] Qin, Y., Fu, Z., Deng, C., Liao, X., Zhang, J., and Duan, H. (2023). Stolen Risks of Models with Security Properties. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)*, November 26–30, 2023, Copenhagen, Denmark. ACM, New York, NY, USA. https://doi.org/10.1145/3576915.3616653.

[46] Qu, Y., Shen, X., He, X., Backes, M., Zannettou, S., and Zhang, Y. (2023). Unsafe Diffusion: On the Generation of Unsafe Images and Hateful Memes From Text-To-Image Models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)*, November 26–30, 2023, Copenhagen, Denmark. ACM, New York, NY, USA. https://doi.org/10.1145/3576915.3616679.

[47] Rajabi, A., Asokraj, S., Jiang, F., Niu, L., Ramasubramanian, B., Ritcey, J., and Poovendran, R. (2023). MDTD: A Multi-Domain Trojan Detector for Deep Neural Networks. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)*, Copenhagen, Denmark, November 26–30, 2023. ACM, New York, NY, USA. https://doi.org/10.1145/3576915.3623082.

[48] Sha, Z., Li, Z., Yu, N., and Zhang, Y. (2023). DE-FAKE: Detection and Attribution of Fake Images Generated by Text-to-Image Generation Models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)*, November 26–30, 2023, Copenhagen, Denmark. ACM, New York, NY, USA. https://doi.org/10.1145/3576915.3616588.

[49] Shen, Y., Han, Y., Zhang, Z., Chen, M., Yu, T., Backes, M., Zhang, Y., and Stringhini, G. (2022). Finding MNEMON: Reviving Memories of Node Embeddings. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS '22)*.

[50] Shi, C., Zhang, T., Xu, Z., Li, S., Gao, D., Li, C., Petropulu, A., Wu, C.-T. M., and Chen, Y. (2023). Privacy Leakage via Speech-induced Vibrations on Room Objects through Remote Sensing based on Phased-MIMO. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)*, November 26–30, 2023, Copenhagen, Denmark. ACM, New York, NY, USA. https://doi.org/10.1145/3576915.3616634.

[51] Tramèr, F., Shokri, R., Joaquin, A. S., Le, H., Jagielski, M., Hong, S., and Carlini, N. (2022). Truth Serum: Poisoning Machine Learning Models to Reveal Their Secrets. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security (CCS '22)*.

[52] Wei, C., Zhao, M., Zhang, Z., Chen, M., Meng, W., Liu, B., Fan, Y., and Chen, W. (2023). DPMLBench: Holistic Evaluation of Differentially Private Machine Learning. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)*, November 26–30, 2023, Copenhagen, Denmark. ACM, New York, NY, USA. https://doi.org/10.1145/3576915.3616593.

[53] Wen, R., Li, Z., Backes, M., and Zhang, Y. (2024). Membership Inference Attacks Against In-Context Learning. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24)*.

[54] Xu, B., Dai, X., Tang, D., and Zhang, K. (2025b). One Surrogate to Fool Them All: Universal, Transferable, and Targeted Adversarial Attacks with CLIP. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security (CCS '25)*, October 13–17, 2025, Taiwan. ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3719027.3744859.

[55] Xu, X., Liu, Z., Koffas, S., and Picek, S. (2025a). Towards Backdoor Stealthiness in Model Parameter Space. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security (CCS '25)*, October 13–17, 2025, Taipei, Taiwan. ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3719027.3744846.

[56] Yang, Y., Hu, M., Xie, X., Cao, Y., Zhang, P., Huang, Y., and Chen, M. (2025). FilterFL: Knowledge Filtering-based Data-Free Backdoor Defense for Federated Learning. In *Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security (CCS '25)*, October 13–17, 2025, Taipei, Taiwan. ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3719027.3744883.

[57] Yu, Z., Zhai, S., and Zhang, N. (2023). AntiFake: Using Adversarial Audio to Prevent Unauthorized Speech Synthesis. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security (CCS '23)*, November 26–30, 2023, Copenhagen, Denmark. ACM, New York, NY, USA. https://doi.org/10.1145/3576915.3623209.

[58] Yuan, Y., Gao, Y., Zhang, Z., Ma, H., Xue, M., Abuadbba, A., and Nepal, S. (2024). HyperTheft: Stealing Hyperparameters of DNNs from TEE-Shielded Black-Box Query Results. In *Proceedings of the 2024 ACM SIGSAC Conference on Computer and Communications Security (CCS '24)*.

[59] Zeng, Y., Pan, M., Just, H., Lyu, L., Qiu, M., and Jia, R. (2023). Narcissus: A Practical Clean-Label Backdoor Attack with Limited Information. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*

(CCS '23), November 26–30, 2023, Copenhagen, Denmark. https://doi.org/10.1145/3576915.3616617.