**Name:** Muhammad Junaid

**Roll number:** BITF22M035

### **Machine Learning Privacy Attacks: A Technical Analysis**

**1. Introduction**

Machine learning models are powered by "an unprecedented amount of data," 1 which often includes sensitive personal information. This dependency creates a fundamental privacy vulnerability: the models, in the process of learning, can "memorize" and leak information about the data they were trained on.2 These vulnerabilities are not merely theoretical; they represent a direct conflict with legal and ethical frameworks like the EU's General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA), which mandate strict principles of data minimization, purpose limitation, and explicit consent.3

This report provides a technical analysis of the primary privacy attacks on machine learning models, based on a curated selection of 12 foundational and contemporary research papers. We will examine three core attack vectors—Membership Inference, Model Inversion, and Data Poisoning—as well as the new frontiers of this threat and the most robust mitigation strategies.

**2. Attack Vector 1: Membership Inference (MIA)**

A Membership Inference Attack (MIA) seeks to determine whether a specific data record was part of a model's training set.7 This is a fundamental privacy breach, as it can reveal sensitive information, such as an individual's participation in a medical study.8

* **Foundational Attack (Black-Box):** The seminal paper by Shokri et al. (2017) demonstrated that an attacker with only black-box (API) access to a model can perpetrate this attack.9 The method uses "shadow training": the attacker trains multiple "shadow models" to mimic the target's behavior.9 An "attack model" is then trained on the outputs of these shadow models to learn the subtle differences in a model's predictions for data it has seen ("members") versus data it has not ("non-members").9 This vulnerability is often linked to model overfitting.12
* **Advanced Attack (White-Box):** While simple white-box extensions (like analyzing activation functions) proved ineffective, more advanced attacks are far more potent.13 Research by Nasr et al. (2019) showed that an attacker with white-box access (to model parameters and gradients) can design algorithms that directly exploit the privacy vulnerabilities of the stochastic gradient descent (SGD) training algorithm, leading to a much stronger attack.13
* **Benchmarking and Risk:** Early metrics for MIA, such as simple attack accuracy, were found to "severely underestimate the privacy risks".2 Work by Song and Mittal (2021) introduced more rigorous benchmarks, including the **"privacy risk score"**.2 This fine-grained metric measures an individual sample's likelihood of being a member, allowing an attacker to identify and target the most vulnerable data points with high confidence.2 The **Attack AUC** (Area Under the Curve) is also a standard metric for measuring the success of these attacks.14

**3. Attack Vector 2: Model Inversion (MI)**

Model Inversion (MI) is a more invasive attack that aims to reconstruct the private training data itself.16

* **Foundational Attack (Confidence-Based):** The classic demonstration of this attack was by Fredrikson et al. (2015), who targeted a facial recognition model.16 The attack method treats inversion as an **optimization problem**.18 Given a person's name (the class label) and API access to the model, the attacker iteratively updates a random image to find an input that **maximizes the model's confidence score** for that person's label.18 The resulting image was shown to be a "recognizable" reconstruction of the person's face.18
* **Advanced Attack (Gradient-Based):** This attack variant is a critical threat to privacy-preserving frameworks like Federated Learning (FL).20 In FL, multiple clients collaboratively train a model by sharing model *gradients* instead of their raw data.20 Research, such as that by Geiping et al. (2020), showed that these gradients are not safe.20 A malicious server can intercept a client's gradient and, by solving an optimization problem to **minimize the distance** (e.g., cosine distance) between a dummy gradient and the intercepted gradient, can reconstruct the client's private training data batch with high fidelity.20

**4. Attack Vector 3: Data Poisoning**

Data Poisoning is an *integrity attack* where an adversary "deliberately corrupt[s] the training data" to manipulate the model's behavior.21

* **Backdoor (Trojan) Attack:** The "BadNets" paper by Gu et al. (2017) provided the foundational example of this attack.24 The attacker poisons a small portion of the training data by adding a "backdoor trigger" (e.g., a small "X" in the corner or a yellow square 24) and **flipping the label** to a target class chosen by the attacker.25 The resulting "BadNet" model appears to function perfectly on all clean test data, passing validation.26 However, at inference time, any input that contains the trigger will be misclassified as the attacker's target label.24
* **Clean-Label Attack:** This is a far stealthier attack because it does **not involve label flipping**.27 The attacker takes a legitimate data sample, adds subtle, optimized perturbations (often generated via a complex bilevel optimization process 29), but keeps the *correct* label. The (input, label) pair appears plausible to a human inspector.28 However, these poisoned samples are designed to corrupt the model's feature space, causing it to misclassify a *different, specific target image* during inference.29 This stealth makes the attack much harder to detect.28

**5. New Frontiers: LLM and Architectural Vulnerability**

The privacy risks demonstrated in simple models have scaled dramatically with the advent of Large Language Models (LLMs).

* **Verbatim Data Extraction from LLMs:** A landmark study by Carlini et al. (2021) on the GPT-2 model demonstrated that an attacker could **extract verbatim training data** simply by querying the model.31 The extracted data included sensitive, unique sequences like names, phone numbers, email addresses, and private code.31 The paper's most "worrying" finding was a clear correlation: **larger models are more vulnerable** to this memorization and extraction than smaller ones.31
* **Architectural Vulnerability:** The problem may be embedded in the architecture of modern AI. A 2024 USENIX paper by Zhang et al. systematically compared the privacy of Convolutional Neural Networks (CNNs) and **Transformers** (the architecture behind LLMs).36 The key finding was that **Transformers consistently exhibit higher vulnerabilities** to privacy attacks, including MIA and gradient inversion, than CNNs.36 This increased "leakiness" is attributed to the Transformer's core micro-design choices, such as **attention modules**, Layer Normalization (LayerNorm), and GeLU activation functions, which appear to enhance model memorization.36

**6. Mitigation Strategies and Governance**

Addressing these threats requires a combination of robust governance and technical, provable defenses.

* **Governance:** The **NIST AI Risk Management Framework (AI RMF)** provides a structure for organizations to "manage the benefits and risks" of AI.37 It outlines a process to Map, Measure, and Manage AI risks, including the privacy threats detailed here.39
* **Provable Defense: Differential Privacy (DP):** DP is considered the "de facto standard" for a formal, mathematical privacy guarantee.40 It ensures a model's output is statistically similar, regardless of whether any single individual's data was included. This is achieved by adding carefully calibrated noise.
  + **DP-SGD:** Proposed by Abadi et al. (2016), this is the most common method for training private deep learning models.42 It integrates privacy directly into training by **clipping the influence (gradient) of each data point** and **adding noise to the gradients** before they are used to update the model.42
  + **PATE (Private Aggregation of Teacher Ensembles):** This framework trains an ensemble of "Teacher" models on disjoint subsets of the private data. It then creates a "Student" model by having the Teachers "vote" on new, unlabeled data, and **adds noise to the aggregated vote counts**.42
* **The Privacy-Utility Trade-off:** The primary challenge in deploying DP is the inherent **privacy-utility trade-off**.44 A smaller privacy budget (meaning more noise and stronger privacy) typically results in lower model accuracy.45

**7. Conclusion**

Privacy attacks on machine learning are not a niche or theoretical concern; they are a fundamental vulnerability that has evolved from simple inference 8 and reconstruction 18 into a systemic risk that scales with model size 31 and is exacerbated by the industry's standard architecture.36 As these models become more powerful and integrated, a "do-it-now" approach that bypasses governance is a significant liability. The only robust path forward is to architect privacy from the ground up, using provable defenses like Differential Privacy 42 and guiding deployment with structured frameworks like the NIST AI RMF.39

### **Key Research Papers and Sources**

1. **Abadi, M., et al. (2016). "Deep Learning with Differential Privacy."** As cited in research on DP-SGD, this paper proposed the foundational DP-SGD algorithm for training deep learning models with provable privacy guarantees.43
2. **Carlini, N., et al. (2021). "Extracting Training Data from Large Language Models."** A landmark study that demonstrated the verbatim extraction of sensitive training data, including PII, from large language models like GPT-2.7
3. **Fredrikson, M., Jha, S., & Ristenpart, T. (2015). "Model Inversion Attacks that Exploit Confidence Information."** The foundational paper on model inversion, which introduced the optimization-based attack to reconstruct recognizable faces from a facial recognition model.16
4. **Geiping, J., et al. (2020). "Inverting Gradients: How easy is it to break privacy in federated learning."** A key paper (cited in 20) that demonstrated the "gradient inversion attack," showing that private data in federated learning can be reconstructed from the shared gradients.20
5. **Gu, T., et al. (2017). "BadNets: Identifying Vulnerabilities in the Machine Learning Model Supply Chain."** This paper introduced the "BadNets" backdoor attack, demonstrating how a model's training data can be poisoned with triggers to create a hidden, malicious behavior.24
6. **Nasr, M., Shokri, R., & Houmansadr, A. (2019). "Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks."** This research expanded on MIA, demonstrating more powerful white-box attacks that exploit model gradients, proving more effective than simple black-box methods.13
7. **NIST. (2023). "AI Risk Management Framework (AI RMF 1.0)."** A foundational governance framework from the U.S. National Institute of Standards and Technology to help organizations map, measure, and manage AI-related risks, including privacy.21
8. **NIST. (2019). "How to deploy machine learning with differential privacy."** A key explanatory article from NIST outlining the two primary frameworks for differentially private machine learning: DP-SGD and PATE.42
9. **Shafahi, A., et al. (2018). "Poison Frogs! Targeted Clean-Label Poisoning Attacks on Neural Networks."** A foundational paper (cited in 28) that introduced "clean-label" poisoning, a stealthy attack that does not require flipping labels, making it invisible to human inspection.
10. **Shokri, R., et al. (2017). "Membership Inference Attacks Against Machine Learning Models."** The seminal paper that formalized the membership inference attack, using "shadow models" to train an attack model to determine if a data record was in a target model's training set.7
11. **Song, L., & Mittal, P. (2021). "Systematic Evaluation of Privacy Risks of Machine Learning Models."** This paper argued that prior MIA benchmarks were insufficient, and introduced more rigorous metrics, including the "privacy risk score," for a fine-grained analysis of privacy leakage.7
12. **Zhang, G., et al. (2024). "How Does a Deep Learning Model Architecture Impact Its Privacy? A Comprehensive Study of Privacy Attacks on CNNs and Transformers."** A recent USENIX paper that provided a systematic comparison, finding that the architectural components of Transformers make them inherently more vulnerable to privacy attacks than traditional CNNs.36

#### Works cited

1. Artificial Intelligence and Privacy – Issues and Challenges - Office of the Victorian Information Commissioner, accessed November 16, 2025, <https://ovic.vic.gov.au/privacy/resources-for-organisations/artificial-intelligence-and-privacy-issues-and-challenges/>
2. Systematic Evaluation of Privacy Risks of Machine Learning Models ..., accessed November 16, 2025, <https://www.usenix.org/conference/usenixsecurity21/presentation/song>
3. GDPR for Machine Learning: Data Protection in AI Development, accessed November 16, 2025, <https://gdprlocal.com/gdpr-machine-learning/>
4. Data Privacy & AI Ethics Best Practices | Governance Guidance 2025 - TrustCommunity, accessed November 16, 2025, <https://community.trustcloud.ai/docs/grc-launchpad/grc-101/governance/data-privacy-and-ai-ethical-considerations-and-best-practices/>
5. A critique of current approaches to privacy in machine learning - PMC, accessed November 16, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC12181200/>
6. Framework for Data Protection, Security, and Privacy in AI Applications 1 Introduction 2 Contextual Background, accessed November 16, 2025, <https://www.nist.gov/document/ai-eo-14110-rfi-comments-loughborough-university>
7. The Hidden Cost of Modeling p(x): Vulnerability to Membership Inference Attacks in Generative Text Classifiers - arXiv, accessed November 16, 2025, <https://arxiv.org/html/2510.16122v1>
8. Membership Inference Attacks Against Machine Learning Models - Cornell: Computer Science, accessed November 16, 2025, <https://www.cs.cornell.edu/~shmat/shmat_oak17.pdf>
9. Membership Inference Attacks Against Machine Learning ... - arXiv, accessed November 16, 2025, <https://arxiv.org/abs/1610.05820>
10. Membership Inference Attacks Against Machine Learning Models - arXiv, accessed November 16, 2025, <https://arxiv.org/pdf/1610.05820>
11. Toward Efficient Inference Attacks: Shadow Model Sharing via Mixture-of-Experts - arXiv, accessed November 16, 2025, <https://arxiv.org/html/2510.13451v1>
12. The Tail Tells All: Estimating Model-Level Membership Inference Vulnerability Without Reference Models - arXiv, accessed November 16, 2025, <https://arxiv.org/html/2510.19773v1>
13. Comprehensive Privacy Analysis of Deep Learning: Passive and ..., accessed November 16, 2025, <https://ieeexplore.ieee.org/abstract/document/8835245>
14. TMI! Finetuned Models Leak Private Information from their Pretraining Data - Privacy Enhancing Technologies Symposium, accessed November 16, 2025, <https://petsymposium.org/popets/2024/popets-2024-0075.pdf>
15. Analysis of Privacy Leakage in Federated Large Language Models - Proceedings of Machine Learning Research, accessed November 16, 2025, <https://proceedings.mlr.press/v238/vu24a/vu24a.pdf>
16. Algorithms that remember: model inversion attacks and data protection law - PMC - NIH, accessed November 16, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC6191664/>
17. Model Inversion Attacks: A Survey of Approaches and Countermeasures - arXiv, accessed November 16, 2025, <https://arxiv.org/html/2411.10023v1>
18. Model Inversion Attacks that Exploit Confidence ... - Thomas Ristenpart, accessed November 16, 2025, <https://rist.tech.cornell.edu/papers/mi-ccs.pdf>
19. Attacks against Machine Learning Privacy (Part 1): Model Inversion Attacks with the IBM-ART Framework - Franziska Boenisch, accessed November 16, 2025, <https://franziska-boenisch.de/posts/2020/12/model-inversion/>
20. Evaluating Gradient Inversion Attacks and Defenses in Federated ..., accessed November 16, 2025, <https://papers.neurips.cc/paper/2021/file/3b3fff6463464959dcd1b68d0320f781-Paper.pdf>
21. Adversarial Machine Learning: A Taxonomy and Terminology of ..., accessed November 16, 2025, <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2023.pdf>
22. Machine Learning Security against Data Poisoning: Are We There Yet? - arXiv, accessed November 16, 2025, <https://arxiv.org/html/2204.05986v3>
23. What is Data Poisoning? Types & Best Practices - SentinelOne, accessed November 16, 2025, <https://www.sentinelone.com/cybersecurity-101/cybersecurity/data-poisoning/>
24. BadNets: Identifying Vulnerabilities in the Machine Learning Model ..., accessed November 16, 2025, <https://arxiv.org/abs/1708.06733>
25. Clean-Label Physical Backdoor Attacks with Data Distillation - arXiv, accessed November 16, 2025, <https://arxiv.org/html/2407.19203v4>
26. BadNets: Evaluating Backdooring Attacks on Deep Neural Networks - IEEE Xplore, accessed November 16, 2025, <https://ieeexplore.ieee.org/ielaam/6287639/8600701/8685687-aam.pdf>
27. [2305.19607] Adversarial Clean Label Backdoor Attacks and Defenses on Text Classification Systems - arXiv, accessed November 16, 2025, <https://arxiv.org/abs/2305.19607>
28. Clean-Label Backdoor Attacks - People | MIT CSAIL, accessed November 16, 2025, <https://people.csail.mit.edu/madry/lab/cleanlabel.pdf>
29. MetaPoison: Practical General-purpose Clean-label Data Poisoning, accessed November 16, 2025, <https://proceedings.neurips.cc/paper/2020/file/8ce6fc704072e351679ac97d4a985574-Paper.pdf>
30. Vulnerability of Clean-Label Poisoning Attack for Object Detection in Maritime Autonomous Surface Ships - MDPI, accessed November 16, 2025, <https://www.mdpi.com/2077-1312/11/6/1179>
31. Extracting Training Data from Large Language Models | USENIX, accessed November 16, 2025, <https://www.usenix.org/conference/usenixsecurity21/presentation/carlini-extracting>
32. [2012.07805] Extracting Training Data from Large Language Models - arXiv, accessed November 16, 2025, <https://arxiv.org/abs/2012.07805>
33. Extracting Training Data from Large Language Models - USENIX, accessed November 16, 2025, <https://www.usenix.org/system/files/sec21-carlini-extracting.pdf>
34. Extracting Training Data from ChatGPT, accessed November 16, 2025, <https://not-just-memorization.github.io/extracting-training-data-from-chatgpt.html>
35. [PDF] Extracting Training Data from Large Language Models | Semantic Scholar, accessed November 16, 2025, <https://www.semanticscholar.org/paper/Extracting-Training-Data-from-Large-Language-Models-Carlini-Tram%C3%A8r/df7d26339adf4eb0c07160947b9d2973c24911ba>
36. How Does a Deep Learning Model Architecture Impact Its ... - USENIX, accessed November 16, 2025, <https://www.usenix.org/system/files/sec24summer-prepub-365-zhang-guangsheng.pdf>
37. Managing Cybersecurity and Privacy Risks in the Age of Artificial Intelligence: Launching a New Program at NIST, accessed November 16, 2025, <https://www.nist.gov/blogs/cybersecurity-insights/managing-cybersecurity-and-privacy-risks-age-artificial-intelligence>
38. privacytrustlab/ml\_privacy\_meter: Privacy Meter: An open-source library to audit data privacy in statistical and machine learning algorithms. - GitHub, accessed November 16, 2025, <https://github.com/privacytrustlab/ml_privacy_meter>
39. Cybersecurity, Privacy, and AI | NIST, accessed November 16, 2025, <https://www.nist.gov/itl/applied-cybersecurity/cybersecurity-privacy-and-ai>
40. Membership Inference Attacks Fueled by Few-Shot Learning to Detect Privacy Leakage and Address Data Integrity - MDPI, accessed November 16, 2025, <https://www.mdpi.com/2504-4990/7/2/43>
41. Auditing Differentially Private Machine Learning: How Private is Private SGD?, accessed November 16, 2025, <https://papers.neurips.cc/paper_files/paper/2020/file/fc4ddc15f9f4b4b06ef7844d6bb53abf-Paper.pdf>
42. How to deploy machine learning with differential privacy | NIST, accessed November 16, 2025, <https://www.nist.gov/blogs/cybersecurity-insights/how-deploy-machine-learning-differential-privacy>
43. Analysis of Application Examples of Differential Privacy in Deep Learning - PMC - NIH, accessed November 16, 2025, <https://pmc.ncbi.nlm.nih.gov/articles/PMC8564206/>
44. Privacy, Utility and Fairness: Navigating Trade-offs in Differentially Private Machine Learning | Proceedings of the AAAI Conference on Artificial Intelligence, accessed November 16, 2025, <https://ojs.aaai.org/index.php/AAAI/article/view/35205>
45. Inference Attacks in Machine Learning as a Service: A Taxonomy, Review, and Promising Directions - arXiv, accessed November 16, 2025, <https://arxiv.org/html/2406.02027v1>
46. Utility Metrics for Differential Privacy: No One-Size-Fits-All | NIST, accessed November 16, 2025, <https://www.nist.gov/blogs/cybersecurity-insights/utility-metrics-differential-privacy-no-one-size-fits-all>
47. Exploiting Explanations for Model Inversion Attacks - CVF Open Access, accessed November 16, 2025, <https://openaccess.thecvf.com/content/ICCV2021/papers/Zhao_Exploiting_Explanations_for_Model_Inversion_Attacks_ICCV_2021_paper.pdf>
48. arXiv:1708.06733v2 [cs.CR] 11 Mar 2019, accessed November 16, 2025, <https://arxiv.org/pdf/1708.06733>
49. Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations - NIST Technical Series Publications, accessed November 16, 2025, <https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-2e2025.pdf>
50. SP 1800-43, Genomics Privacy Threat Modeling - NIST Computer Security Resource Center, accessed November 16, 2025, <https://csrc.nist.gov/pubs/sp/1800/43/ipd>
51. Membership Inference Attacks against Machine Learning Models | Request PDF, accessed November 16, 2025, <https://www.researchgate.net/publication/386821893_Membership_Inference_Attacks_against_Machine_Learning_Models>
52. Membership Inference Attacks Against Adversarially Robust Deep Learning Models - Princeton University, accessed November 16, 2025, <https://www.princeton.edu/~pmittal/publications/liwei-dls19.pdf>
53. Membership Inference Attacks From First Principles - arXiv, accessed November 16, 2025, <https://arxiv.org/pdf/2112.03570>
54. Systematic Evaluation of Privacy Risks of Machine Learning Models - USENIX, accessed November 16, 2025, <https://www.usenix.org/system/files/sec21fall-song.pdf>