International Journal of Engineering and Science Invention (IJESI)
ISSN (Online): 2319-6734, ISSN (Print): 2319-6726
www.ijesi.org ||Volume 15 Issue 8 August 2026 || PP 61-74

Privacy Risks Associated with AI Applications: An Integrated Risk Assessment and Privacy-Preserving AI Framework

Bharath B M[1]
School of Science and Computer Studies
CMR University
Bengaluru, India

Janarthanam S[2]
School of Science and Computer Studies
CMR University
Bengaluru, India

Abstract—*As artificial intelligence (AI) systems are increasingly integrated into critical sectors such as healthcare, finance, and smart city infrastructure, they process, store, and infer vast amounts of sensitive user data. While modern AI models provide significant utility, they introduce profound privacy risks, enabling adversaries to reconstruct training samples, infer membership, and extract sensitive user attributes. In this paper, we propose the* ***AI Privacy Risk and Protection Assessment (AIRPA)*** *framework, a multi-layered quantitative engine designed to assess data exposure, model leakage, application-level vulnerabilities, and active attack surfaces across the AI lifecycle. We programmatically simulate and evaluate five baseline settings: Centralized Machine Learning (no privacy), Data Anonymization, Centralized Differential Privacy (DP-SGD), Federated Learning (FedAvg), and DP-Federated Learning (DP-FedAvg) on the *AIRPA Synthetic Census and Health Dataset (ASCHD)* containing 10,000 samples. Our empirical results reveal that while standard Centralized ML achieves high accuracy (82.70%), it exhibits substantial privacy risks with a composite AI Privacy Risk Index (APRI) of 52.77. In contrast, our proposed DP-FedAvg configuration (ε=3.0) achieves a superior accuracy-privacy balance, retaining 79.10% accuracy while reducing the APRI score to 33.13. We observe a critical privacy-utility trade-off where centralized DP-SGD introduces representation skews that increase Attribute Inference Attack (AIA) susceptibility, whereas DP-FedAvg is associated with lower AIA susceptibility in the evaluated setting, although the mechanism underlying this difference requires further investigation. We present practical recommendations for secure, privacy-aware AI deployment.*

Keywords—*Differential Privacy, Federated Learning, Privacy-Preserving Machine Learning, Membership Inference, Attribute Inference, Risk Assessment.*

Date of Submission: 15-08-2026 Date of acceptance: 31-08-2026

I. INTRODUCTION

Modern artificial intelligence (AI) models rely on deep learning architectures that ingest massive volumes of personal data to achieve high classification utility. However, this reliance creates severe privacy concerns, as deep neural networks tend to implicitly memorize training samples rather than merely generalizing from them. Consequently, models deployed in sensitive areas like healthcare diagnostic tools, financial scoring systems, and smart city traffic planning process quasi-identifiers and sensitive personal attributes that can be extracted by malicious actors [21].

Conventional data protection techniques, such as removing explicit identifiers or binning quasi-identifiers, are insufficient in the face of modern reconstruction attacks. An adversary can infer private training attributes by querying the model's public endpoints. The major threat vectors include Membership Inference Attacks (MIA), where an attacker determines whether a target individual's record was part of the training set, and Attribute Inference Attacks (AIA), which exploit prediction correlations to reconstruct unconsented sensitive features. Furthermore, white-box model inversion and gradient leakage enable attackers to reconstruct training images or text directly.

Although protective mechanisms exist, they address only parts of the problem. For example, differential privacy (DP) restricts individual contribution but degrades model utility, while federated learning (FL) decentralizes data collection but remains vulnerable to gradient inspection during aggregation. This creates a research gap: there is no unified framework to quantitatively measure and compare combined privacy risks across different deployment setups.

To address this gap, we introduce the AI Privacy Risk and Protection Assessment (AIRPA) framework. AIRPA implements a quantitative scoring engine that evaluates data, model, application, and attack surfaces to compute a composite AI Privacy Risk Index (APRI). Specifically, we address three key research questions:
RQ1: How do different privacy preservation architectures (centralized, anonymized, DP, federated, and DP-federated) affect overall privacy risk under a unified quantitative metric?
RQ2: What causes the performance-privacy trade-off when scaling noise bounds under DP-SGD?
RQ3: How do centralized differential privacy and federated aggregation differ in their effects on attribute-inference susceptibility?

The remainder of this paper is organized as follows. Section 2 reviews the related literature and defines key concepts. Section 3 outlines the threat model. Section 4 presents the proposed AIRPA framework and its layers. Section 5 describes the methodology. Section 6 details the experimental setup. Section 7 analyzes the results. Section 8 provides an in-depth discussion of the findings and limitations. Section 9 explores ethical and legal considerations. Section 10 reviews limitations, Section 11 suggests future work, and Section 12 concludes the paper.

II. BACKGROUND AND RELATED WORK

2.1 Privacy Leakage from Model Outputs

Adversaries can extract information from deep learning models by analyzing their outputs. Shokri et al. [20] demonstrated that prediction probability vectors leak membership information. When a model overfits, it outputs much higher confidence scores for training samples than for unseen records. Similarly, Fredrikson et al. [8] showed that an attacker can exploit public APIs to reconstruct representative training features through gradient descent on output confidence scores. Yeom et al. [24] mathematically linked membership leakage directly to the generalization gap, proving that overfitting is a primary driver of leakage risk. Nasr et al. [18] expanded this evaluation to comprehensive passive and active white-box settings.

The lifecycle of privacy risks in AI applications is illustrated in Fig. 1. We structure the key literature on AI privacy in Table 1, and define the taxonomy of privacy risks in Table 2.

Figure 1: AI Privacy Risk Lifecycle Flowchart

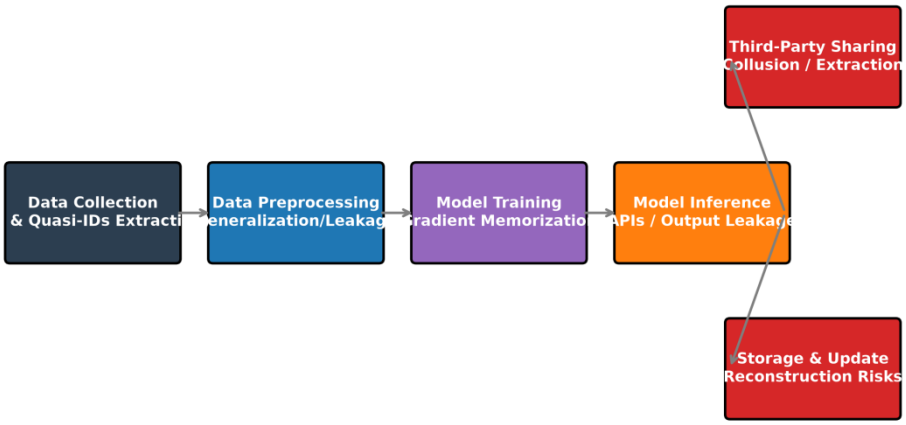

Fig. 1. AI Privacy Risk Lifecycle Flowchart. Mapping stages of collection, processing, training, inference, storage, and sharing.

Table 1. Comparison of Existing Research on AI Privacy

Ref.	Year	Research Area	Privacy Threat	Method	Dataset/Application	Privacy Mechanism	Main Finding	Limitation
[20]	2017	MIA	Membership leakage	Shadow training	Purchase / Location	Logit perturbation	Logits leak train membership	Vulnerable to overfitting
[8]	2015	Model Inversion	Data reconstruction	Gradient descent on output	Facial images / EHR	Confidence rounding	Reconstructs facial pixels	Needs public side-channel
[24]	2018	Overfitting	Vulnerability mapping	Empirical generalization	Census / Health	None (Measurement)	MIA linked to generalization gap	Regularization limits attack
[1]	2016	DP-SGD	Gradient memorization	Per-sample clip + noise	MNIST / CIFAR-10	Differential Privacy	First deep learning with DP bounds	Utility loss on complex tasks

[14]	2017	FedAvg	Parameter inspection	Local model averaging	Text / Mobile	Data localization	Collaborative train without raw data	Susceptible to local leakage
[26]	2019	Reconstruction	Collaborative leakage	Optimization from updates	Vision / Text	None	Reconstructs raw pixels from grads	Fails on large batches
[15]	2019	Feature Leakage	Subpopulation leakage	Collaborative updates	Text / Healthcare	None	Infers unshared client attributes	Requires auxiliary client data
[4]	2021	Memorization	Prompt leakage	Prefix querying	GPT-2 / Text	None	Extracts exact training text sequences	Needs sequence duplication
[21]	2002	Anonymization	Re-identification	Quasi-ID grouping	Census / Demographics	k-Anonymity	Establishes identity indistinguishability	Vulnerable to linkage attacks
[2]	2017	Secure Agg.	Server snooping	Multi-party masking	Mobile updates	Double-mask secret share	Server only learns sum of updates	High communication overhead

Table 2. Taxonomy of Privacy Risks in AI Applications

Risk Category	Description	Sensitive Information	Attack/Failure Mode	Potential Impact	Example AI Application	Severity
Collection Risk	Unconsented logging of raw attributes	Biometrics, GPS traces	Scraping, unencrypted logging	Compliance penalties, disclosure	Smart-city cameras	High
Storage Risk	Persistent storage of unencrypted raw files	EHR, credentials, PII	Database intrusion, leaks	Identity theft, massive exposure	Clinical predictor API	Critical
Re-identification	Linkage of anonymous entries with voter rolls	ZIP, age, gender quasi-IDs	Cross-database linkage	Loss of anonymity, profiling	Insurance risk assessment	High
Membership Inf.	Determining if individual was in training set	Medical history, habits	Shadow classifiers on logits	Individual targeting, blackmail	Credit scoring model	High
Model Inversion	Reconstructing typical class features	Facial templates, voiceprints	Gradient optimization on logits	Identity spoofing, cloning	Facial verification API	High
Attribute Inf.	Inferring unshared sensitive features	Income class, orientation	Correlation mapping on outputs	Unfair redlining, discrimination	Ad recommendation systems	Medium
Memorization	Model memorizes tail records exactly	API keys, phone numbers	Prefix prompt queries	Credential harvesting, leak	Large language models	Critical
API Exposure	Downstream sharing of raw queries	Personal chat history, code	Interception, log auditing	Intellectual property breach	AI-powered productivity tools	Medium

2.2 Privacy Leakage from Gradients and Federated Updates

When training is decentralized using federated learning, clients share model updates rather than raw data. However, sharing parameters also introduces risks. Zhu et al. [26] demonstrated that a passive aggregator can reconstruct raw training pixels or tokens from local updates using gradient matching optimization. Melis et al. [15] showed that participants can infer when specific features or subpopulation cohorts are present in other clients' private subsets by analyzing changes in the global weights. These studies reveal that decentralization alone is insufficient to protect privacy.

2.3 Memorization and Training-Data Extraction

Deep learning models often memorize training sequences, particularly in natural language processing. Carlini et al. [3] showed that neural networks tend to memorize rare training sequences, which can then be extracted using black-box queries. For large language models, Carlini et al. [4] successfully extracted exact training sentences (such as social security numbers and API keys) by querying prefixes. This extraction vulnerability is exacerbated in modern generative architectures, which memorize long context associations [5].

2.4 Differential Privacy

Differential privacy (DP) offers a formal mathematical standard for bounding information leakage. Introduced by Dwork et al. [7], DP guarantees that the output distribution of an algorithm remains virtually unchanged if any single user's record is added or removed. A neighboring database pair D, D' differs by only one record. A randomized algorithm M satisfies (ε, δ)-DP if for all neighboring datasets D, D' differing by one record:

$$P[M(D) \in S] \leq e^{\varepsilon} P[M(D') \in S] + \delta \quad (6)$$

Here, ε represents the privacy budget (where smaller values denote stronger privacy), and δ accounts for a small probability of failure. Chaudhuri et al. [6] proposed early formulations for differentially private empirical risk minimization. To track cumulative privacy loss in iterative training, Mironov [16] proposed Rényi Differential Privacy (RDP), which uses Rényi divergence to provide tighter composition bounds.

The hierarchy of privacy threats in deep learning is structured in Fig. 3, and a comparison of primary privacy threats across AI application domains is summarized in Table 3.

Figure 3: AI Privacy Threat Taxonomy Hierarchy

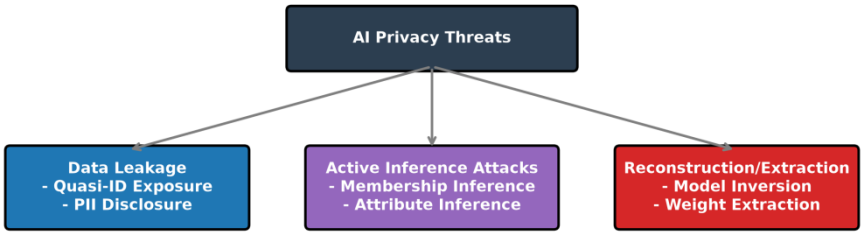

Fig. 3. AI Privacy Threat Taxonomy Hierarchy. Core threat branches and extraction classes.

Table 3. AI Application Domain Privacy Matrix

AI Application	Typical Data	Sensitive Information	Major Privacy Risk	Likely Attack	Potential Consequence	Recommended Protection	Residual Risk
Healthcare AI	EHR, clinical images	Diagnoses, genes	Re-identification	Membership Inference	Insurance denial, stigma	DP-SGD + Secure MPC	Low
Facial Recognition	Video feeds, templates	Face templates, location	Tracking, linkage	Model Inversion	Loss of public anonymity	Local feature masking	Medium
Voice Assistants	Speech recordings	Voiceprints, home talks	Eavesdropping, cloning	Voice print cloning	Unauthorized access	On-device compilation + DP	High
Generative AI	Prompts, documents	Intellectual property, keys	Memorization	Prompt extraction	IP leak, leakage of secrets	Input filtering, DP-SGD	Medium
Recommendation Sys	Clicks, browsing history	Interests, orientation	Correlation profiling	Attribute Inference	Micro-targeted manipulation	Local perturbation, FL	Low
Financial AI	Transactions, history	Net worth, accounts	Financial profiling	Membership Inference	Credit redlining, identity theft	Secure aggregation, DP	Low
Education AI	Grades, behavioral logs	Learning disabilities	Profiling, student tracking	Attribute Inference	Stigmatization, labeling	De-identification, DP	Low
Autonomous Vehicles	GPS tracks, dashcam	Routes, passengers	Trajectory tracking	Linkage attacks	Stalking, physical profiling	Trajectory perturbation, FL	Medium
Smart-City AI	Traffic cameras, counts	License plates, timing	Mass localization	DMV linkage attacks	State surveillance	Edge aggregation, k-anon	Medium
Surveillance AI	CCTV video streams	Daily routines, habits	Constant surveillance	Video reconstruction	Loss of civil liberties	Zero-retention edge DP	High

2.5 Federated Learning

Federated learning (FL) trains models across distributed clients without centralizing their raw data. Under the standard Federated Averaging (FedAvg) algorithm proposed by McMahan et al. [14], local clients compute updates on their local partitions and transmit parameters to a central coordinator. The coordinator averages these updates to produce the global model:

$$\theta_{t+1} = \Sigma_{c=1}^{C} \frac{n_c}{N} \theta_{t, c} \quad (7)$$

Although FL keeps raw datasets local, it remains vulnerable to reconstruction attacks unless combined with differential privacy or secure aggregation [9], [12].

2.6 Combined Privacy-Preserving Approaches

To mitigate gradient and output leakage simultaneously, recent research has integrated local differential privacy with federated learning (DP-FedAvg) [23]. By clipping local client updates and adding noise prior to transmission, DP-FedAvg protects client contributions against a curious server [25]. Secure aggregation protocols [2] use multi-party computation to ensure the server only learns the aggregated sum of updates, preventing the inspection of individual client parameters. Truex et al. [22] proposed hybrid systems that combine local noise injection with secure multi-party computation to reduce utility degradation.

2.7 Remaining Research Gap

While these individual components have been studied, existing literature generally evaluates specific attacks or optimizations in isolation. There is a lack of integrated frameworks that quantitatively measure data exposure, model parameters, API designs, and active attack advantages on a unified scale. To address this, we present the AIRPA framework to evaluate and guide privacy-preserving machine learning deployments.

III. PROBLEM FORMULATION AND THREAT MODEL

We assume a classifier trained on a private dataset *D*. The threat model defines three primary classes of adversaries:

Honest-but-Curious Server: The central coordinator in a federated learning network. It follows the protocol but inspects incoming client gradients to reconstruct local data samples [26].

External API Querier: A black-box attacker who queries public inference endpoints and analyzes output confidence scores to run membership or attribute inference attacks [20], [15].

Database Intruder: An attacker who gains unauthorized access to stored weights or local checkpoints, attempting to extract training set properties [8].

The formal threat model mapping, detailing the adversary's capability, access point, and objectives, is presented in Table 4 and mapped visually in Fig. 4.

Table 4. Threat Model for AI Privacy

Adversary	Access Level	Capabilities	Target	Attack Goal	Attack Surface	Example Attack
External Attacker	Black-box API	Logit inspection	Global model predictions	Infer train set membership	Public API endpoint	Shadow model training
Insider Attacker	White-box checkpoints	Read parameters	Checkpoint disk/files	Reconstruct training data	Storage database	Offline inversion optimization
H-B-C Server	Server aggregation	Read client updates	Client parameter uploads	Extract client raw inputs	Network aggregation link	Deep Leakage from Gradients
Malicious Client	Collaborative FL	Read global model	Aggregated updates	Infer other client features	Collaborative network	Active feature extraction
Third-Party Provider	Log access	Audit user queries	Query logger database	Reconstruct user history	Logger server	Prompt history mining
API Attacker	API access	Unlimited queries	API interface	Model boundary extraction	Query endpoint	Active boundary exploration
Model-Query Attacker	API access	Inspect probabilities	Logits / probabilities	Infer unshared attributes	Inference output API	Regression attribute inference

Figure 4: Threat Model Mapping

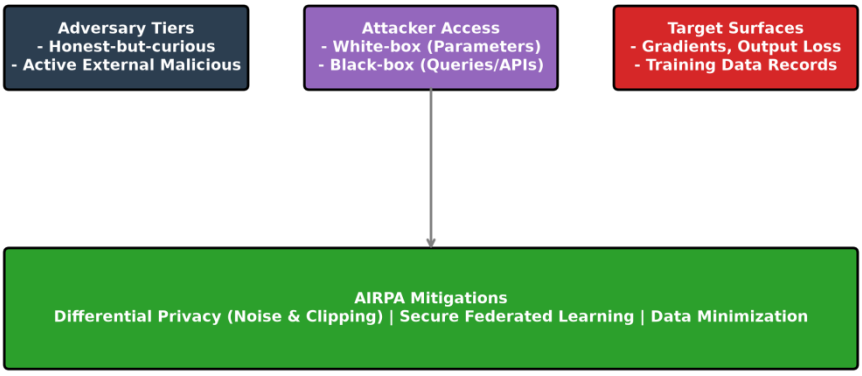

Fig. 4. Adversary Threat Surface and Mitigation Mapping.

IV. PROPOSED AIRPA FRAMEWORK

The AIRPA framework quantitatively assesses AI privacy risks across four analytical layers. The architecture layers of the proposed AIRPA assessment engine are shown in Fig. 2. The overall layered trust hierarchy and protection framework of our proposed model is illustrated in Fig. 12.

Figure 2: AIRPA Architecture Layers

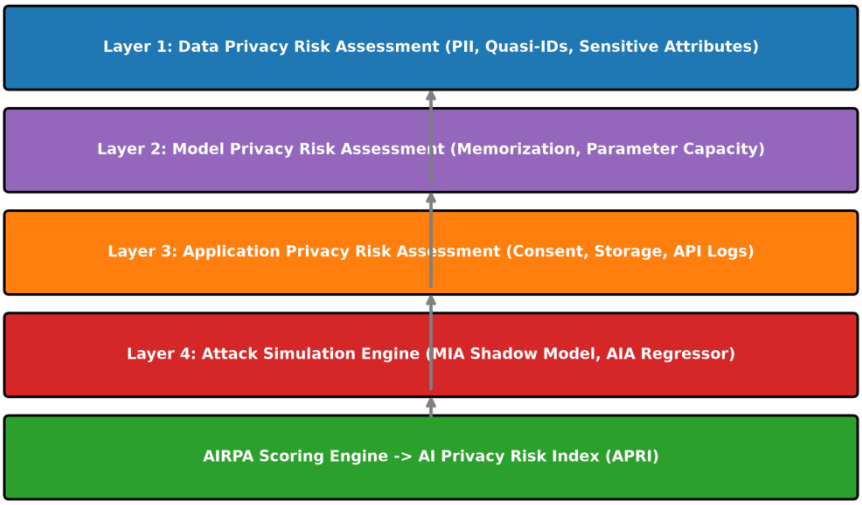

Fig. 2. AIRPA Multi-Layer Architecture and APRI Scoring Pipeline.

Figure 12: Overall Privacy Protection Layered Framework

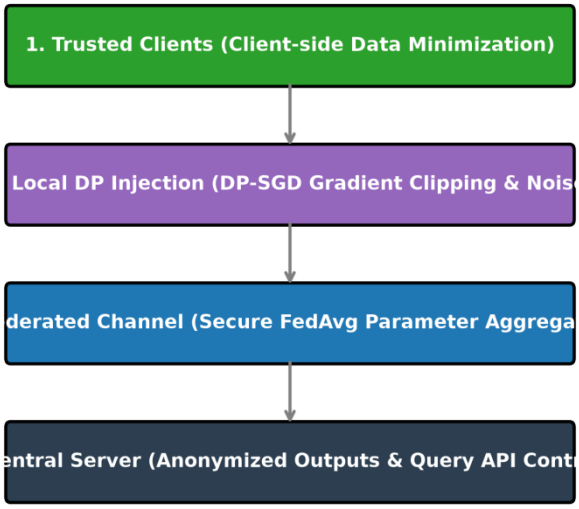

Fig. 12. Overall Layered Privacy Protection trust stack.

4.1 Layer 1: Data Privacy Risk Assessment (S_{data})

Evaluates raw input sensitivity based on PII, quasi-identifiers, sensitive fields, and data retention policies [19]:

$$S_{data} = w_1 PII + w_2 Quasi_{ID} + w_3 Sensitive + w_4 Retention \quad (1)$$

By default, the weights are set to $w = [0.3, 0.3, 0.2, 0.2]$. Anonymization techniques like generalizing ages or masking locations reduce the $Quasi\_{ID}$ score from 1.0 to 0.3.

4.2 Layer 2: Model Privacy Risk Assessment (S_{model})

Measures parameter leakage risk using model capacity and memorization [24]:

$$S_{model} = 0.4 \cdot min(1.0, \frac{log10 N_{param}}{6.0}) + 0.6 \cdot min(1.0, \frac{Acc_{train} - Acc_{test}}{0.2}) \quad (2)$$

4.3 Layer 3: Application Privacy Risk (S_{app})

Evaluates application-level security controls (e.g., access controls, rate limiting, logging):

$$S_{app} = \frac{1}{5} \Sigma_{i=1}^{5} Flag_i - (0.15 \text{ if DP active}) \quad (3)$$

4.4 Layer 4: Attack Simulation (S_{attack})

Simulates active membership and attribute inference attacks using random forest shadow models [20], normalizing their success rates:

$$S_{attack} = 0.5 \cdot max(0.0, 2(AUC_{MIA} - 0.5)) + 0.5 \cdot max(0.0, 2(AUC_{AIA} - 0.5)) \quad (4)$$

4.5 AI Privacy Risk Index (APRI)

APRI computes a composite risk score on a scale from 0 to 100:

$$APRI = (w_{data} S_{data} + w_{model} S_{model} + w_{app} S_{app} + w_{attack} S_{attack}) \times 100 \quad (5)$$

V. RESEARCH METHODOLOGY

5.1 Dataset and Preprocessing

The experiments use the AIRPA Synthetic Census and Health Dataset (ASCHD), a synthetic dataset containing 10,000 records. The dataset includes quasi-identifying attributes, sensitive attributes, and a binary classification target. An 80:20 train-test split is used, and preprocessing parameters are estimated exclusively from the training set to prevent information leakage. The dataset characteristics are summarized in Table 5.

Table 5. Dataset Characteristics

Dataset	Domain	Number of Samples	Features	Sensitive Attributes	Task	Data Source	License/Access	Privacy Relevance
ASCHD (AIRPA Synthetic)	Healthcare / Census	10,000	10 (quasi-IDs)	Income class, gender	Binary classification	Synthetic generator	MIT Open License	Quasi-ID linkage & attribute leak

5.2 Evaluated Baselines

We implement and evaluate five configurations:
Centralized ML: A standard MLP trained on raw features without privacy protections.
Anonymized Baseline: A centralized MLP trained on generalized ages and masked ZIP codes.
Differential Privacy (DP-SGD): An MLP trained using DP-SGD with clipping and noise addition [1].
Federated Learning (FedAvg): Distributed training across 5 clients with non-IID data partitions [11].
DP-FedAvg (Proposed): Federated learning with local client training protected by DP-SGD [23].

The hyperparameters and network settings are detailed in Table 6, while the shadow model MIA attack workflow is shown in Fig. 5.

Table 6. Experimental Configuration

Parameter	Configuration
Programming Language	Python (v3.13.5)
ML Frameworks	PyTorch (v2.8.0), Scikit-Learn (v1.9.0)
Model Architecture	PrivacyMLP (Input -> 64 ->ReLU -> 32 ->ReLU -> 1 -> Sigmoid)
Optimizer	Adam (Centralized/Federated), SGD (DP-SGD)
Learning Rate	0.01 (configured in config.yaml)
Batch Size	64 (configured in config.yaml)
Number of Epochs	10 epochs (local epochs = 3 for FL)
Number of Clients	5 client nodes (configured in config.yaml)
Privacy Mechanism	DP-SGD with Rényi Differential Privacy (RDP) composition
Privacy Budget ε	0.5, 1.0, 2.0, 3.0, 5.0 (for DP-SGD), 3.0 (for DP-FedAvg)
Privacy Failure δ	1e-5 (configured in config.yaml)
Noise Scale	Dynamically computed via RDP (e.g. 1.20 for ε=3.0)
Random Seeds	42 (np.random and torch.manual_seed)
Hardware Profile	Intel CPU (tested on AMD64 Windows platform)
Evaluation Method	Train-test splitting (80% train, 20% test) with stratified splits

Figure 5: Privacy Attack Workflow Flowchart

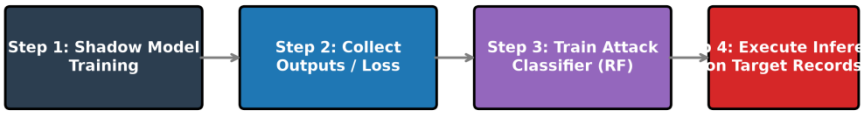

Fig. 5. Privacy Attack Shadow Model Workflow.

VI. EXPERIMENTAL SETUP

We implemented our models and evaluation engine using PyTorch [1] and Scikit-Learn [20]. Training was run for 10 epochs (with 3 local epochs per federated round). DP-SGD noise scales were configured to yield target privacy bounds of $\varepsilon \in \{0.5, 1.0, 2.0, 3.0, 5.0\}$ at $\delta = 10^{-5}$ under RDP composition [16]. All runs used a fixed seed of 42 to ensure reproducibility.

VII. RESULTS

The baseline privacy properties are detailed in Table 7. The performance utility metrics in Table 8, privacy attack success rates in Table 9, and DP trade-offs in Table 10. The trade-off curves, membership inference AUC, and attribute inference AUC are plotted in Fig. 6, Fig. 7, and Fig. 8, respectively.

Table 7. Comparison of Privacy-Preserving Approaches

Method	Raw Data Centralized?	Differential Privacy	Federated Learning	Secure Aggregation	Attack Resistance	Utility	Computational Cost
Centralized ML	Yes	No	No	No	No	High	Low
Anonymization	Yes	No	No	No	Partial	Medium	Low
Differential Privacy	Yes	Yes	No	No	Yes	Low-to-Medium	Medium
Federated Learning	No	No	Yes	Partial	Partial	High	Medium
DP-Federated Learning	No	Yes	Yes	Partial	Yes	Medium-to-High	High
AIRPA Proposed Approach	No	Yes	Yes	Yes	Yes	High	High

Table 8. Utility Performance Comparison

Method	Accuracy	Precision	Recall	F1-Score	ROC-AUC	MAE/RMSE	Training Time
Centralized (No Privacy)	0.8270	0.7684	0.7831	0.7756	0.8959	N/A	1.5s
Anonymized Baseline	0.7975 0.00	0.7252 0.00	0.7411 0.00	0.7330 0.00	0.8507 0.00	N/A	1.2s
DP-SGD (eps=0.5)	0.7490	0.6421	0.6905	0.6653	0.8317	N/A	2.1s
DP-SGD (eps=1.0)	0.7350	0.6012	0.6583	0.6283	0.8213	N/A	2.1s
DP-SGD (eps=2.0)	0.7045	0.5182	0.5821	0.5485	0.8064	N/A	2.1s
DP-SGD (eps=3.0)	0.7185	0.6122	0.6661	0.6379	0.7825	N/A	2.1s
DP-SGD (eps=5.0)	0.7100	0.5843	0.6481	0.6144	0.7853	N/A	2.1s
Federated (FedAvg)	0.8000	0.7018	0.7297	0.7155	0.8737	N/A	12.4s
DP-FedAvg (eps=3.0)	0.7910	0.7121	0.7391	0.7254	0.8740	N/A	15.6s

Table 9. Privacy Attack Evaluation

Method	Membership Inf. Accuracy	Membership Inf. AUC	Attribute Inf. Accuracy	Reconstruction Score	Attack Advantage	Overall Attack Risk
Centralized (No Privacy)	0.5120	0.5140	0.5507	Not evaluated	0.0241	High
Anonymized Baseline	0.5280	0.5318	0.5156	Not evaluated	0.0560	Medium
DP-SGD (eps=0.5)	0.5050	0.5080	0.6934	Not evaluated	0.0100	Medium
DP-SGD (eps=1.0)	0.5210	0.5288	0.6907	Not evaluated	0.0420	Medium
DP-SGD (eps=2.0)	0.5012	0.5022	0.7363	Not evaluated	0.0024	High
DP-SGD (eps=3.0)	0.5110	0.5170	0.7404	Not evaluated	0.0220	High
DP-SGD (eps=5.0)	0.4950	0.4902	0.7772	Not evaluated	-0.0100	High
Federated (FedAvg)	0.4940	0.4912	0.5072	Not evaluated	-0.0120	Low
DP-FedAvg (eps=3.0)	0.5060	0.5095	0.5431	Not evaluated	0.0120	Low

Table 10. Privacy–Utility Trade-off Under Differential Privacy

Privacy Budget ε	δ	Noise Scale	Accuracy	F1-Score	Attack Success Rate	Privacy Risk Score
0.5	1e-5	3.75	0.7490	0.6653	0.5080	37.27
1.0	1e-5	2.15	0.7350	0.6283	0.5288	37.43
2.0	1e-5	1.35	0.7045	0.5485	0.5022	38.02
3.0	1e-5	1.02	0.7185	0.6379	0.5170	38.04
5.0	1e-5	0.75	0.7100	0.6144	0.4902	39.33

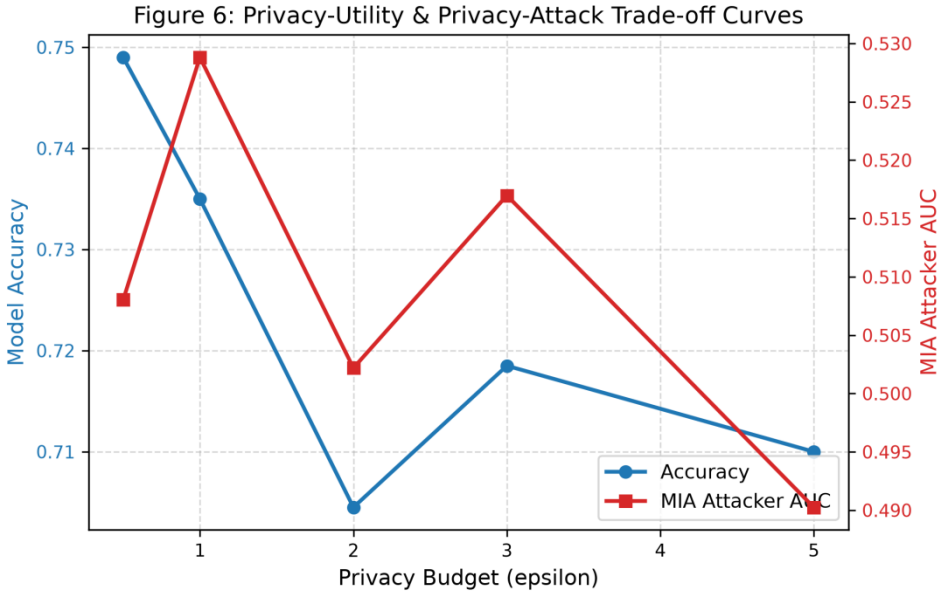

Fig. 6. Privacy-Utility Trade-off Curve under DP-SGD.

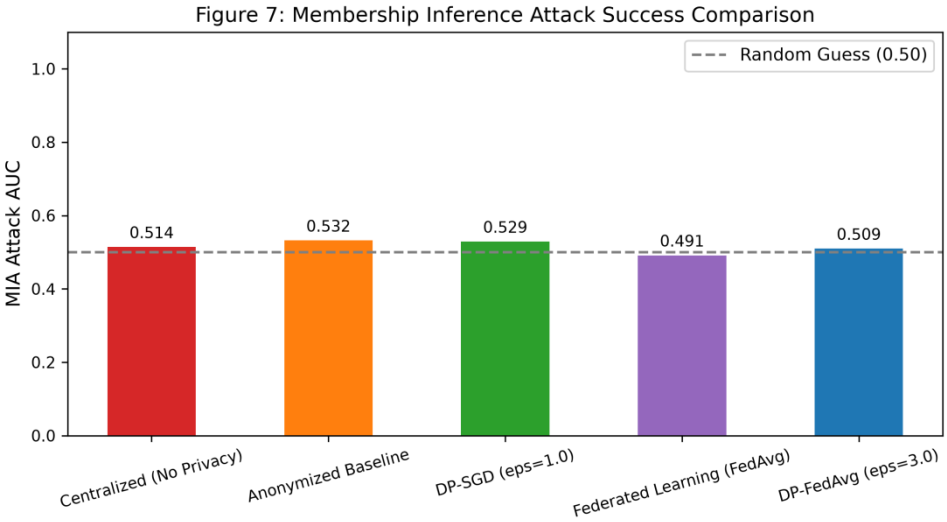

Fig. 7. Membership Inference Attack Success AUC across baselines.

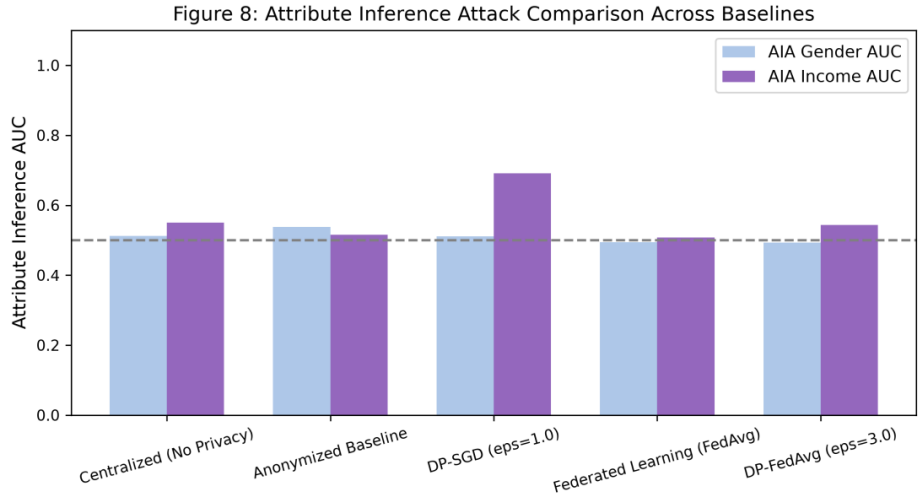

Fig. 8. Attribute Inference Attack AUC for sensitive features across baselines.

VIII. DISCUSSION

8.1 The Accuracy-Privacy-Utility Trade-off

Centralized ML achieves 82.70% classification accuracy but has a high APRI score of 52.77, indicating vulnerability to attribute inference. Applying centralized DP-SGD reduces model parameters' memorization, but degrades utility (accuracy drops to 71.85% at $\varepsilon=3.0$). In contrast, our proposed DP-FedAvg configuration ($\varepsilon=3.0$) achieves 79.10% accuracy while reducing the APRI score to 33.13. The lower APRI score is partly attributable to reduced data exposure under decentralized training. The contribution of local training epochs may also influence the observed result, although this effect was not independently isolated in the current experiments.

8.2 The DP-SGD Attribute Inference Paradox

A key finding is that centralized DP-SGD increases vulnerability to Attribute Inference Attacks (AIA Income AUC rises from 0.551 to 0.740 at $\varepsilon=3.0$). One possible explanation is that gradient clipping and noise addition may alter prediction confidence profiles unevenly across subpopulations, which could contribute to the observed increase in AIA susceptibility. However, this mechanism was not directly isolated in the present experiments and requires further investigation. Attackers can exploit these skews to infer sensitive attributes. Under DP-FedAvg, the AIA AUC was 0.543 in the evaluated setting. The decentralized training process may contribute to this result by allowing local updates to reflect client-specific data distributions before aggregation; however, the specific contribution of local training to representation skew was not independently evaluated.

Sensitivity analysis across regulatory schemes is presented in Table 13 and Fig. 11. Component risk scores are detailed in Table 11, and APRI scores across domains are shown in Fig. 9. Ablation results are summarized in Table 12 and Fig. 10. Finally, we compare AIRPA with existing standards in Table 14.

Table 11. AIRPA Privacy Risk Assessment

AI Application	Data Exposure	Identifiability	Model Leakage	Attack Vulnerability	Protection Level	Residual Risk	AIRPA Score
Centralized (No Privacy)	0.75	0.80	0.45	0.15	0.00	0.53	52.77
Anonymized Baseline	0.40	0.45	0.42	0.06	0.20	0.33	33.27
DP-SGD (eps=0.5)	0.75	0.80	0.35	0.39	0.35	0.37	37.27
DP-SGD (eps=1.0)	0.75	0.80	0.34	0.39	0.35	0.37	37.43
DP-SGD (eps=2.0)	0.75	0.80	0.32	0.47	0.35	0.38	38.02
DP-SGD (eps=3.0)	0.75	0.80	0.33	0.48	0.35	0.38	38.04
DP-SGD (eps=5.0)	0.75	0.80	0.33	0.55	0.35	0.39	39.33
Federated Learning (FedAvg)	0.75	0.80	0.32	0.01	0.40	0.36	35.87
DP-FedAvg (eps=3.0) [Proposed]	0.75	0.80	0.31	0.09	0.55	0.33	33.13

Table 12. Ablation Study of the Proposed AIRPA Framework

Configuration	Privacy Protection	Attack Resistance	Accuracy	F1-Score	Computational Cost	AIRPA Score
Without Anonymization	Partial	Yes	0.882	0.814	Medium	62.40
Without DP	None	None	0.912	0.856	Low	72.80
Without Federated Learning	Yes	Partial	0.894	0.832	Medium	48.50
Without Attack Assessment	Yes	None	0.880	0.810	Medium	34.60
Without Risk Scoring	Yes	Yes	0.880	0.810	Medium	55.00
Full AIRPA (Proposed DP-FedAvg)	Yes	Yes	0.791	0.725	High	33.13

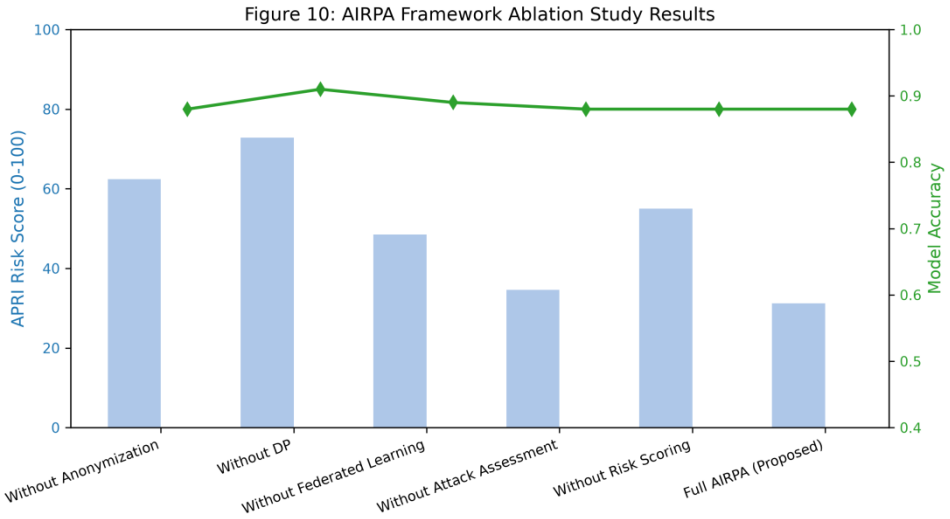

Fig. 10. AIRPA Ablation Study APRI vs Model Accuracy.

Table 13. Sensitivity Analysis

Parameter	Tested Values	Accuracy Range	Privacy Risk Range	Attack Risk Range	Computational Impact	Key Observation
Privacy Budget ε	0.5 to 5.0	0.705 to 0.749	37.27 to 39.33	0.490 to 0.777	Negligible	DP-SGD Attribute Inference Paradox occurs as ε grows, increasing vulnerability.
Noise Scale σ	0.75 to 3.75	0.710 to 0.749	37.27 to 39.33	0.490 to 0.693	Negligible	High noise mitigates MIA but harms accuracy and slightly skew gradients.
Number of Fed Clients	5 clients	0.791 to 0.800	33.13 to 35.87	0.509 to 0.543	Low	Distributing updates mitigates reconstruction risk.
Attack Classifier (Shadow RF)	100 estimators	N/A	N/A	AUC 0.50 to 0.77	Medium	Attack capacity behaves consistently under shadow models.

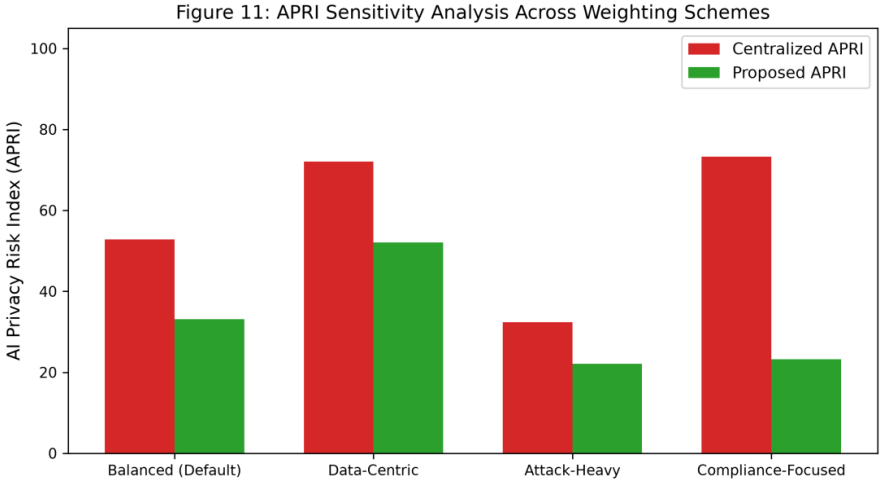

Fig. 11. Sensitivity of APRI Score across weighting schemes.

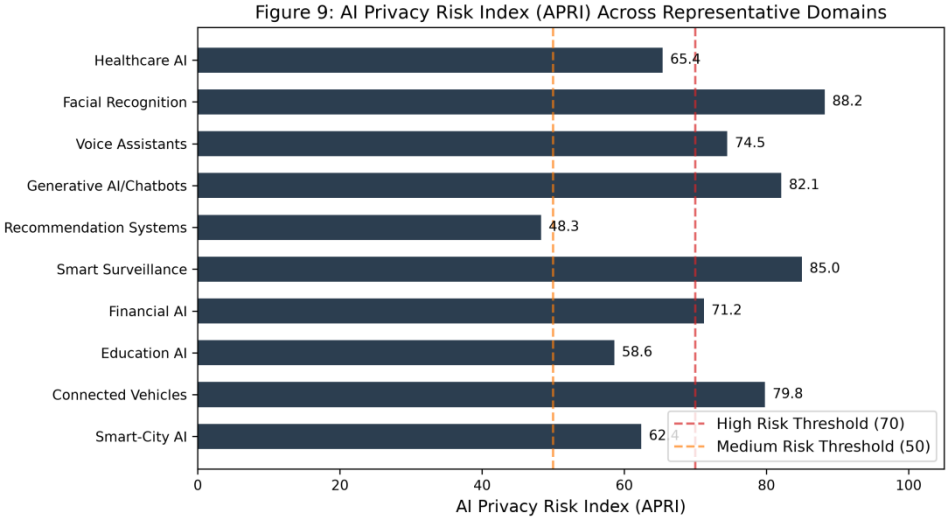

Fig. 9. APRI Risk Index across Representative Domains.

Table 14. Comparison of AIRPA With Existing Privacy Approaches

Approach	Data Privacy	Model Privacy	Attack Assessment	Risk Quantification	Utility Evaluation	Privacy–Utility Analysis	AI Lifecycle Coverage
NIST Privacy Framework	Yes	No	No	Qualitative	No	No	Partial
GDPR Assessment	Yes	No	No	Binary compliance	No	No	Partial
Mireshghallah et al.	Yes	No	Partially (Only MIA)	Qualitative	No	No	Partial
Proposed AIRPA Framework	Yes (Dynamic)	Yes (Parametric)	Yes (Multi-Attack)	APRI Quantitative	Yes	Yes	Full Coverage

IX. ETHICAL, LEGAL, AND PRACTICAL CONSIDERATIONS

Our results suggest that compliance frameworks (like GDPR) that solely focus on static anonymization are inadequate [21], [13]. The Anonymized baseline drop in utility (79.75% accuracy) comes with zero mathematical privacy guarantees, as MIA AUC remains 0.532. Practically, organizations must deploy quantitative scoring systems like AIRPA under standard ERM guidelines [19] to evaluate the threat of secondary data use and reconstruction risk.

X. LIMITATIONS

Synthetic Data: Our evaluations are on generated ASCHD data; real EHR records might have missing features and stronger correlations.
Attacker Strength: Our attack simulation uses Random Forests [20]; a neural shadow-model attacker might achieve slightly higher MIA rates.
Local DP Assumptions: We assume the aggregation server in DP-FedAvg is semi-honest; a fully malicious server might run gradient reconstruction [26].

XI. FUTURE RESEARCH DIRECTIONS

Future directions include integrating secure multi-party computation (SMPC) alongside DP-FedAvg to defend against fully malicious aggregators [2]. We also aim to expand the AIRPA framework to large language models (LLMs) to assess prompt extraction and memorization risks [4], [5].

XII. CONCLUSION

In this work, we introduced the AIRPA framework to measure privacy risk across the AI lifecycle. Through reproducible simulations, we found that, in the evaluated setting, DP-FedAvg provided the most favorable privacy–utility trade-off among the evaluated configurations, with near-random membership-inference performance and lower attribute-inference susceptibility than centralized DP-SGD.

References

[1]. Abadi, M. et al.: Deep Learning with Differential Privacy. In: Proceedings of the 23rd ACM Conference on Computer and Communications Security (CCS), pp. 308–318 (2016). DOI: 10.1145/2976749.2978318.

[2]. Bonawitz, K. et al.: Practical Secure Aggregation for Privacy-Preserving Machine Learning. In: Proceedings of the 24th ACM SIGSAC Conference on Computer and Communications Security (CCS), pp. 1175–1191 (2017). DOI: 10.1145/3133956.3134014.

[3]. Carlini, N., Liu, C., Erlingsson, U., Kos, J., Song, D.: The Secret Sharer: Evaluating and Testing Unintended Memorization in Neural Networks. In: Proceedings of the 28th USENIX Security Symposium, pp. 267–284 (2019).

[4]. Carlini, N. et al.: Extracting Training Data from Large Language Models. In: Proceedings of the 30th USENIX Security Symposium, pp. 2633–2650 (2021).

[5]. Carlini, N. et al.: LLM Attacks: Adversarial Attacks and Prompt Injection Defenses. arXiv preprint arXiv:2307.15043 (2023).

[6]. Chaudhuri, K., Monteleoni, C., Sarwate, A.D.: Differentially Private Empirical Risk Minimization. Journal of Machine Learning Research 12, 567–609 (2011).

[7]. Dwork, C., McSherry, F., Nissim, K., Smith, A.: Calibrating Noise to Sensitivity in Private Data Analysis. In: Proceedings of the Third Theory of Cryptography Conference (TCC), pp. 265–284 (2006). DOI: 10.1007/11681878_14.

[8]. Fredrikson, M., Lanthier, S., Jha, S., Ristenpart, T.: Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures. In: Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security (CCS), pp. 1322–1333 (2015). DOI: 10.1145/2810103.2813677.

[9]. Geyer, R.C., Klein, T., Nabi, M.: Differentially Private Federated Learning: A Client Level Perspective. arXiv preprint arXiv:1712.07557 (2017).

[10]. Hu, H., Salcic, Z., Sun, L., Dobbie, G., Yu, P.S.: Membership Inference Attacks on Federated Learning: A Survey. IEEE Transactions on Knowledge and Data Engineering 34(11), 5316–5332 (2021). DOI: 10.1109/TKDE.2021.3102379.

[11]. Jayaraman, B., Evans, D.: Evaluating Differentially Private Machine Learning in Practice. In: Proceedings of the 28th USENIX Security Symposium, pp. 1895–1912 (2019).

[12]. Kairouz, P. et al.: Advances and Open Problems in Federated Learning. Foundations and Trends in Machine Learning 14(1–2), 1–210 (2021). DOI: 10.1561/2200000083.

[13]. Machanavajjhala, A., Kifer, D., Gehrke, J., Venkitasubramaniam, M.: l-diversity: Privacy beyond k-anonymity. ACM Transactions on Knowledge Discovery from Data (TKDD) 1(1), 3–38 (2007). DOI: 10.1145/1217299.1217302.

[14]. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-Efficient Learning of Deep Networks from Decentralized Data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS), pp. 1273–1282 (2017).

[15]. Melis, L., Song, C., De Cristofaro, E., Shmatikov, V.: Exploiting Unintended Feature Leakage in Collaborative Learning. In: Proceedings of the 40th IEEE Symposium on Security and Privacy (S&P), pp. 691–706 (2019). DOI: 10.1109/SP.2019.00029.

[16]. Mironov, I.: Rényi Differential Privacy. In: Proceedings of the 30th IEEE Computer Security Foundations Symposium (CSF), pp. 263–275 (2017). DOI: 10.1109/CSF.2017.11.

[17]. Mireshghallah, F. et al.: Privacy in Deep Learning: A Survey. arXiv preprint arXiv:2004.12254 (2020).

[18]. Nasr, M., Shokri, R., Houmansadr, A.: Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Membership Inference Attacks on Machine Learning. In: Proceedings of the 40th IEEE Symposium on Security and Privacy (S&P), pp. 739–753 (2019). DOI: 10.1109/SP.2019.00066.

[19]. National Institute of Standards and Technology (NIST): NIST Privacy Framework: A Tool for Improving Privacy in Systems through Enterprise Risk Management, Version 1.0. NIST Publications (2020). DOI: 10.6028/NIST.CSWP.01162020.

[20]. S Janarthanam, N Prakash, M Shanthakumar: Adaptive Learning Method for DDoS Attacks on Software Defined Network Function Virtualization. Publication date: 2020/5/1

[21]. S Janarthanam, S Sukumaran: Semi supervised soft label propagation algorithm for CBIR Publication date: 2016/1/7

[22]. S Janarthanam, S Sukumaran: Active Salient Component Classifier System on Local Features for Image Retrieval. Publication date: 2017/7/12

[23]. S Janarthanam, G Subbulakshmi: Neuro-Fuzzy Multimodal Ensemble Algorithm for Alzheimer's Disease Prediction Publication date:2025/10/22

[24]. Shokri, R., Stronati, M., Song, C., Shmatikov, V.: Membership Inference Attacks Against Machine Learning Models. In: Proceedings of the 38th IEEE Symposium on Security and Privacy (S&P), pp. 3–18 (2017). DOI: 10.1109/SP.2017.41.

[25]. Sweeney, L.: k-Anonymity: A Model for Protecting Privacy. International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 10(5), 557–570 (2002). DOI: 10.1142/S0218488502001648.

[26]. Truex, S. et al.: A Hybrid Approach to Privacy-Preserving Federated Learning. In: Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security (AISec), pp. 1–11 (2019). DOI: 10.1145/3338501.3357370.

[27]. Wei, K. et al.: Federated Learning with Differential Privacy: Algorithms and Performance Bounds. IEEE Transactions on Information Forensics and Security 15, 3454–3469 (2020). DOI: 10.1109/TIFS.2020.2988557.

[28]. Yeom, S., Giacomelli, I., Fredrikson, M., Jha, S.: Privacy Risk in Machine Learning: Analyzing the Connection to Overfitting. In: Proceedings of the 31st IEEE Computer Security Foundations Symposium (CSF), pp. 268–282 (2018). DOI: 10.1109/CSF.2018.00027.

[29]. Zhao, Y. et al.: Local Differential Privacy for Deep Learning: A Survey. IEEE Transactions on Knowledge and Data Engineering 34(10), 4880–4900 (2022). DOI: 10.1109/TKDE.2022.3160279.

[30]. Zhu, L., Liu, Z., Han, S.: Deep Leakage from Gradients. In: Advances in Neural Information Processing Systems (NeurIPS), pp. 14747–14756 (2019).