Mesopotamian journal of Cybersecurity
Vol.5, No.3, **pp**. 1057–1080
DOI: https://doi.org/10.58496/MJCS/2025/057; ISSN: 2958-6542
https://mesopotamian.press/journals/index.php/cybersecurity

Research Article

PRIFLEX: A Secure Federated Learning Framework for Evaluating Privacy Leakage and Defense in Cross-Modal Medical Data

Mohammad Othman Nassar [1,*], Feras Fares AL-Mashagba [2],

[1] College of Information Technology, Cyber Security Department, Amman Arab University, Amman 11953, Jordan

[2] Faculty of Information Technology, Computer Science Department, Jerash University, Jerash 26150, Jordan

ARTICLEINFO

Article History

Received 25 Jun 2025
Revised 18 Jul 2025
Accepted 27 Aug 2025
Published 30 Sep 2025

Keywords

Federated learning; privacy leakage; secure aggregation; differential privacy; hybrid defense; Electrocardiogram; clinical text; multimodal data; healthcare AI

ABSTRACT

Federated learning (FL) enables decentralized, privacy-preserving machine learning by training models across distributed data without sharing raw patient information. However, most FL frameworks focus on unimodal data and overlook critical challenges in multimodal healthcare settings, such as privacy risks, fairness disparities, and reduced model interpretability. We present PRIFLEX (Privacy-Resilient Integration Framework for Learning Exchange), a novel FL framework designed for secure integration of structured and unstructured medical data. PRIFLEX combines 12-lead electrocardiograms (ECG) from the PhysioNet PTB-XL dataset and clinical notes from the Medical Information Mart for Intensive Care IV (MIMIC-IV), supporting early, late, and hybrid data fusion strategies. To safeguard model updates, it evaluates standalone and hybrid defenses using Differential Privacy (DP) and Secure Aggregation (SA) against gradient leakage, model inversion, and membership inference attacks. Results show that early fusion improves the area under the Receiver Operating Characteristic curve (AUROC) by up to 6.2%, while hybrid DP+SA reduces attack success rates by up to 84% and improves fairness with manageable system overhead. PRIFLEX also quantifies interpretability loss using SHapley Additive exPlanations (SHAP) and gradient-based methods, highlighting the trade-off between privacy and transparency. Overall, PRIFLEX sets a new benchmark for building secure, fair, and explainable federated learning systems in healthcare.

1. INTRODUCTION

The exponential growth of clinical data from electronic health records (EHRs), imaging, biosignals, and textual notes has enabled machine learning (ML) to play a transformative role in healthcare delivery and personalized medicine [1], [2], [3]. However, training high-performance models across multiple institutions remains hindered by privacy regulations such as Health Insurance Portability and Accountability Act (HIPAA) and General Data Protection Regulation (GDPR), which restrict raw patient data sharing. To address these concerns, FL has emerged as a decentralized training paradigm where local models are trained collaboratively without transferring raw data [4], [5]. FL has shown promise in several healthcare domains, including radiology [6], cardiology [7], and chronic disease management [8]. Recent work on secure deep learning for phishing detection further illustrates the growing role of privacy-aware AI in adversarial contexts [9].

Despite these advances, most FL frameworks focus on unimodal datasets, such as imaging [5] or wearable ECG signals [8]—and assume that all institutions possess similar types of data. This overlooks the fact that real-world healthcare systems are inherently multimodal: structured signals (e.g., ECG, vitals), unstructured narratives (e.g., clinical notes), and other modalities (e.g., genomics, laboratory tests) often coexist but are distributed heterogeneously among clients [10], [11], [12]. Such heterogeneity introduces new challenges for FL, including modality imbalance, inconsistent feature spaces, and complex privacy vulnerabilities during model updates.

Recent studies have proposed multimodal FL frameworks that explore fusion strategies and representation alignment [13], [14], [15]. However, these approaches often rely on synthetic data fusion or assume full modality overlap between clients, which is rarely the case in practice. Furthermore, while privacy is a central motivation for FL, most multimodal FL studies focus on performance, paying limited attention to adversarial risks such as gradient-based leakage, membership inference, or model inversion attacks [13] , [16], [17],[18]. This gap is particularly problematic when different modalities have asymmetric information densities, making one modality (e.g., text) more vulnerable than others (e.g., ECG) under attack.

DM such [5], [19], and SA [20], [21] have been studied in isolation in FL settings but are rarely combined or benchmarked in adversarial scenarios. Furthermore, FL frameworks typically do not evaluate how privacy defenses affect model fairness, training stability, or explainability, all of which are crucial for building trustworthy AI systems in healthcare [22], [23], [24]. For example, explainability tools such as SHAP or Gradient-weighted Class Activation Mapping (Grad-CAM) are commonly used to interpret deep models [25], but their behavior under strong privacy guarantees has not been rigorously analyzed.

Finally, most FL research lacks transparency in system overhead and deployment feasibility. There is limited empirical work measuring the runtime and communication cost introduced by defense mechanisms in multimodal settings, where model complexity is naturally higher [10], [26], [12].

Although FL has demonstrated its potential in privacy-preserving medical AI applications [4] , [5], most existing frameworks focus on unimodal data sets, such as imaging [6] or wearable signals [8] —and assume consistent data availability between clients. In contrast, real-world clinical environments involve highly heterogeneous and distributed data modalities, including structured signals (e.g., ECG) and unstructured narratives (e.g., clinical notes) [10], [11]. Recent surveys and benchmarks have begun to explore multimodal FL [13], [14], [15], but these efforts often rely on the fusion of synthetic modality or ignore privacy risks in cross-modal settings.

Importantly, prior work has not holistically addressed how fusion strategies (early, late, hybrid) influence privacy leakage and model utility. Studies of privacy-preserving FL using DP [5], [19] and SA [20], [21] typically evaluate these defenses in isolation, with minimal exploration of combined mechanisms in heterogeneous environments. Moreover, frameworks rarely examine how these defenses affect fairness across clients [24], interpretability under privacy constraints [22], or system overhead, key factors for clinical deployment [10], [26].

To address these gaps, PRIFLEX is proposed as a novel, reproducible and privacy-aware FL framework for the integration of cross-modal medical data. Specifically, the following contributions are achieved:

- Realistic cross-mode simulation: ECG signals from Physikalisch-Technische Bundesanstalt – Extended Length (PTB-XL) and clinical notes from Medical Information Mart for Intensive Care IV (MIMIC-IV) are integrated to simulate heterogeneous clients (unimodal and multimodal), reflecting practical data silos in hospitals.
- Hybrid Privacy Defense Benchmarking: standalone and hybrid defenses (DP, SA, DP+SA) are implemented and evaluated under three adversarial attack scenarios: Deep Leakage from Gradients (DLG), Membership Inference (MIA), and Model Inversion.
- Comprehensive Metric Suite: PRIFLEX introduces a rich evaluation pipeline, including AUROC, gradient exposure, fairness variance, privacy–utility curves, and system overhead - supporting robust, real-world benchmarking.
- Privacy-Aware Explainability: SHAP and Grad-CAM are incorporated into the federated pipeline and quantify attribution degradation using Attribution Rank Correlation (ARC), providing insight into the trade-offs between interpretability and privacy.
- Empirical Insights for Deployment: Our experiments show that early fusion improves AUROC by up to 6.2% but increases leakage unless hybrid defenses are applied—reducing attack success by up to 84% and improving fairness with acceptable overhead.

This work is the first to systematically evaluate privacy, fairness, explainability, and deployability in cross-modal Fl, offering actionable insights for building secure, interpretable, and equitable AI systems in healthcare.

2. RELATED WORK

FL has emerged as a key technique for enabling collaborative model training in sensitive medical data without centralizing raw patient records [1], [4], [5]. Frameworks like FedHealth [8] and federated brain tumor segmentation [5] established early proof-of-concept applications in healthcare care, particularly focusing on structured or imaging modalities. However, these studies rely on unimodal input distributions, limiting their generalizability to real-world healthcare systems where data exist in heterogeneous, distributed modalities (e.g., ECG, text, labs, images). As highlighted in [6] and [10], most existing FL frameworks assume homogeneous client data, failing to reflect the nonuniform data availability typically observed across clinical institutions.

Recent surveys [10], [26], [12] underscore that although FL in healthcare has matured in terms of performance optimization and communication efficiency, it remains underdeveloped in multimodal learning and privacy-aware evaluation.

Furthermore, these studies often neglect fairness and explainability, which are essential for clinical deployment. This gap signals the need for FL frameworks that not only support cross-modal medical data but also include privacy, utility, and fairness metrics for realistic evaluations.

Recent work in multimodal FL has explored combining structured (e.g. vitals, labs) and unstructured (e.g., clinical notes) data to enhance model utility and generalizability [13], [27], [14]. For example, FedMobile [27] introduces contribution-aware training using multimodal signals, but does not integrate adversarial threat models or privacy metrics. Similarly, [14] survey multimodal challenges of FL and propose architectural solutions, but their benchmarks are based on simulated fusion, not real clinical data. Other studies like [15], [11] explore vertical FL in cross-modal scenarios, yet focus mainly on performance and representation alignment, without simulating privacy attacks or evaluating real-world privacy risks.

Importantly, most multimodal FL frameworks do not benchmark fusion strategies (early, late, hybrid) nor evaluate their impact on privacy leakage or fairness, a key novelty addressed in this work. The PRIFLEX framework advances this field by implementing and comparing fusion designs using real ECG and clinical text datasets (PTB-XL, MIMIC-IV) with realistic client heterogeneity - a simulation not found in previous work [14], [15], [11].

Despite privacy being a core motivation for FL, existing studies have only partially addressed real-world privacy threats. Attacks such as DLG [17], membership inference (MIA) [28], and model inversion [16] have been theoretically validated, but rarely simulated in cross-modal or multi-modal FL contexts. Many studies also assume white-box or centralized adversaries [18], without testing robust defenses in distributed settings.

Defensive techniques such as DP [5], [19] and SA [20], [21] have shown utility in unimodal FL, but are rarely combined and even less so in multimodal healthcare settings [29]. The PRIFLEX framework is among the first to benchmark hybrid defenses (DP + SA) across multiple attack types (DLG, MIA, inversion) and to show quantitative leakage reduction, bridging a crucial gap in current research [5], [20], [19], [21].

Traditional FL evaluations priorities performance metrics like AUROC and F1, overlooking other critical dimensions for clinical deployment [2], [24], [30]. Only recent studies begin to explore fairness variance [24], group-conditional balancing [30], or contribution-aware personalization Enhancing Privacy and Fairness in FL for Distributed E-Healthcare, yet rarely apply these in multimodal, privacy-constrained settings. Similarly, while privacy–utility trade-offs have been conceptualized [31], few frameworks offer quantitative curves or gradient-based exposure metrics.

The PRIFLEX framework introduces a unified evaluation pipeline encompassing utility, fairness variance, gradient norms, and privacy–utility curves, enabling a multidimensional assessment of model reliability under real-world conditions. This holistic evaluation remains largely absent in previous work [24], [30], Enhancing Privacy and Fairness in FA for Distributed E-Healthcare [31].

Explainability is an emerging but yet neglected area in FL. Although methods such as SHAP [25] and Grad-CAM [7] have been applied in centralized models, they are rarely adapted to privacy-preserving FL settings. In fact, studies have shown that privacy mechanisms can degrade interpretability [22], [23], [32], yet current FL frameworks do not integrate explainability metrics into their evaluation pipeline.

PRIFLEX uniquely combines privacy defenses with explainability analysis, quantifying the impact of privacy constraints on feature attribution quality using metrics like ARC and sparsity. This integrated perspective is novel and critical for clinically trustworthy AI [22], [23], [33], [34].

Although several FL frameworks discuss communication and computation costs [10], [26], [12], most focus on model performance rather than quantifying overhead from privacy defenses like DP and SA. Even fewer examine how these costs scale in multimodal deployments, where model complexity and encoder design add significant overhead. Surveys such as [10] and [12]acknowledge this gap, but stop short of empirical validation.

On the contrary, PRIFLEX benchmarks runtime and communication overhead across defense types in multimodal settings, offering a practical deployment perspective. This adds actionable information for system designers and medical institutions seeking to balance privacy and efficiency [10], [12].

Recent work suggests that metaheuristic optimization can improve FL robustness and adaptivity in privacy-sensitive settings [35], [36], [37]. Algorithms such as the Dragonfly Algorithm (DFA) [36], or hybrid chimp–reinforcement models [38], are being proposed to optimize noise calibration, aggregation masks, and dynamic defense tuning. These techniques could be valuable for future PRIFLEX iterations, though they have yet to be incorporated into cross-modal privacy pipelines [35], [37], [39], [38].

In summary, the existing literature has made strides in FL, multimodal integration, and privacy preservation. However, no current framework systematically combines:

- Real-world cross-modal medical data (e.g., ECG and clinical notes),

- Hybrid privacy defenses benchmarked under multiple attack types,
- Fusion architecture comparison,
- Fairness and explainability metrics, and
- Deployment-focused overhead evaluation.

PRIFLEX addresses these gaps through a modular, adversarial tested, and interpretable FL framework, setting a new standard for secure and equitable FL in healthcare.

3. PRIFLEX METHODOLOGY

PRIFLEX is presented, a modular and reproducible FL framework designed to evaluate privacy leakage, fairness, and interpretability in multimodal healthcare environments. PRIFLEX simulates heterogeneous clients using real ECG and clinical text data, supports multiple fusion and privacy defense strategies, and evaluates attacks in a comprehensive set of utility, fairness, and leakage metrics.

3.1 System Overview and Workflow

PRIFLEX is designed to reflect realistic federated healthcare deployments, with heterogeneous clients receiving structured (ECG) or unstructured (clinical text) data, or both. As shown in Figure 1, the architecture consists of:

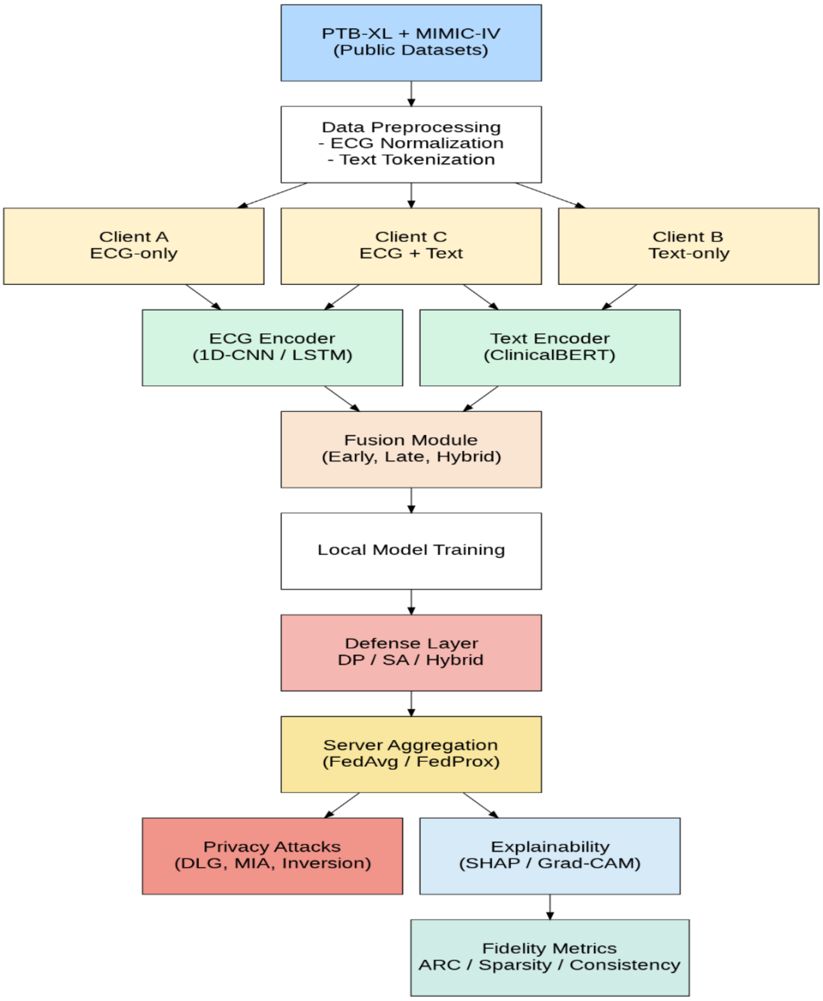
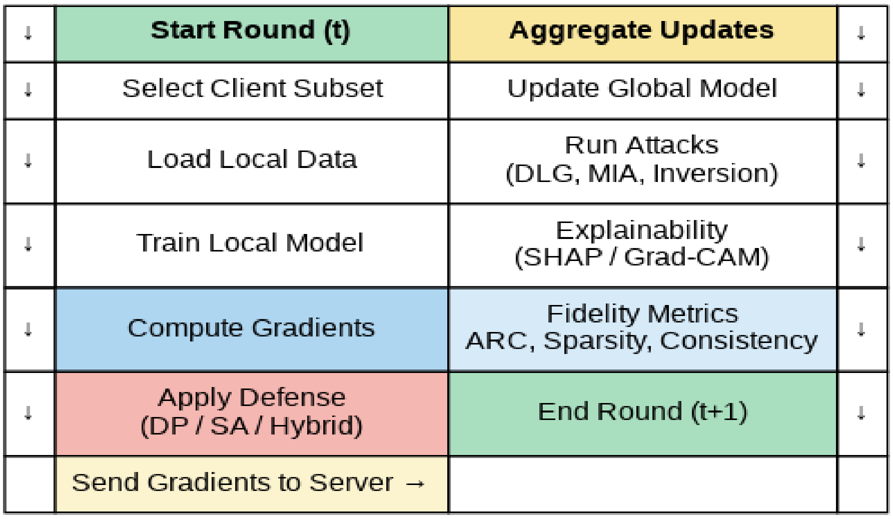


Fig. 1. The PRIFLEX architecture depicts fusion strategies, privacy defenses, and adversarial evaluation modules.

- Client Simulation: Clients are assigned one of three roles: ECG-only, text-only, or both, using data from PTB-XL and MIMIC-IV datasets.
- Fusion Strategies: Early fusion concatenates embedding, late fusion ensembles output predictions, and hybrid fusion shares encoders with private heads.
- Privacy defenses: DP, SA, and their hybrid are applied to protect model updates during training.
- Attack Simulation: Three types of privacy attacks are launched: Deep Gradient Leakage (DLG), Membership Inference (MIA), and Model Inversion are launched to assess vulnerabilities.

The operational flow is shown in Figure 2, covering all phases: preprocessing, encoding, fusion, local training, defense application, aggregation, attack simulation, and metric evaluation.



↓	**Start Round (t)**	**Aggregate Updates**	↓
↓	Select Client Subset	Update Global Model	↓
↓	Load Local Data	Run Attacks (DLG, MIA, Inversion)	↓
↓	Train Local Model	Explainability (SHAP / Grad-CAM)	↓
↓	Compute Gradients	Fidelity Metrics ARC, Sparsity, Consistency	↓
↓	Apply Defense (DP / SA / Hybrid)	End Round (t+1)	↓
	Send Gradients to Server →		

Fig. 2. PRIFLEX pipeline Workflow for cross-modal FA with privacy assessment.

3.2 Modular Components and Configurations

PRIFLEX is implemented using PyTorch and the Flower framework [40], offering modular design and extensibility across different configurations and data types. The architecture separates encoding, fusion, privacy, and attack simulation into independently configurable components, allowing reproducible experimentation. Table 1 details the core modules used in PRIFLEX and their associated functions.

TABLE I. MODULAR COMPONENTS

Module	**Functionality**
ECGEncoder()	1D-CNN over 12-lead ECGs
TextEncoder()	ClinicalBERT over clinical narratives
Fuse()	Combines modalities using early, late, or hybrid methods
GaussianNoise(std)	Adds DP noise using Opacus
Encrypt()/Decrypt()	Implements SA using CrypTen additive masking
Aggregate()	Performs FL aggregation (FedAvg/FedProx)
RunDLG()	Reconstructs training inputs from gradients
RunMembershipInference()	Performs binary inclusion testing

To ensure consistent evaluation across experiments, fixed set of hyperparameters and system parameters are implemented. These include learning rate, batch size, local training epochs, and privacy settings such as the DP budget and SA protocol. Table 2 summarizes the key configurations used throughout the federated training and evaluation pipeline, including attack simulation and fusion strategies.

TABLE II. TECHNICAL CONFIGURATION COMPONENTS

Parameter	Value
Local Epochs	5
Batch Size	32
Learning Rate	0.001
DP Epsilon (ε)	1, 5, 10
Fusion Strategies	Early, Late, Hybrid
Attack Batch Size	1
SA Protocol	Additive Masking

Privacy and Security Libraries

- Opacus: Provides DP support with per-sample gradient tracking and ε-accounting.
- CrypTen: Enables encrypted aggregation using additive masking.
- Flower: Facilitates cross-silo FL simulations with heterogeneous clients [40].

3.3 Explainability in Privacy-Constrained FL

To enhance trust in multimodal FL, PRIFLEX includes privacy-aware explainability methods.

- SHAP : Applied to ClinicalBERT outputs, capturing token-level attribution. DP and SA are simulated to observe attribution degradation via noise injection.
- Saliency Mapping: Gradient-based methods (for example, integrated gradients) highlight signal segments in ECG data [41].

The robustness is quantified using the following metrics:

- Attribution rank correlation (ARC): Measures the stability of the explanation classification before / after privacy.
- Sparsity: Evaluates attribution focus (high values indicate better interpretability).
- Stability: Captures output robustness under noise or defense-induced perturbations.

This setup supports insights into the fidelity–privacy trade-off in explainable AI, aligned with clinical needs.

In SHAP visualizations, clinically meaningful terms such as ‘sepsis’ and “tachycardia” were highlighted as highly influential features in text classification. In ECG, saliency maps emphasized diagnostic wave segments. Under privacy defenses, these attributions shifted or degraded - captured quantitatively by ARC and sparsity metrics (see Table 6, Figure 9). SHAP's model-agnostic nature also makes it well suited for deployment across diverse clinical FL architectures.

3.4 NLP and Visualization Tools

- spaCy: Text pre-processing and tokenization.
- HuggingFace Transformers: Loads ClinicalBERT encoders.
- NumPy / SciPy: Statistical analysis, DP noise sampling.
- Matplotlib / Seaborn: Visualizations of attack success, saliency maps, and performance metrics.

PRIFLEX provides a unified, modular pipeline for evaluating multimodal FL systems in realistic healthcare scenarios. It systematically integrates privacy defenses, simulates adversarial attacks, and supports an explainable, fairness-aware evaluation using established metrics and tools. This end-to-end approach addresses gaps in privacy-preserving, cross-modal FL frameworks , [42],.

3.5 Federated Round Logic (PRIFLEX Federated Learning)

The following pseudocode outlines the federated training and evaluation procedure in PRIFLEX. Simulates cross-silo participation across heterogeneous clients with structured (ECG) or unstructured (clinical text) data. At each communication round, selected clients perform local training, optionally apply privacy defenses (Differential Privacy, Secure Aggregation, or both), and generate explainability outputs using SHAP or Grad-CAM. These outputs are evaluated using fidelity metrics such as Attribution Rank Correlation (ARC), sparsity, and consistency. The server aggregates updates securely if configured, and optionally updates global explanation references.

Algorithm (1): PRIFLEX Federated Learning Round with Privacy and Explainability

```
  C ← set of all clients
  R ← number of communication rounds
  E ← number of local epochs
  B ← batch size
  ε ← differential privacy budget (if enabled)
  SA ← enable secure aggregation (True/False)
  Hybrid ← enable hybrid (DP + SA) mode
  Modality ← {ECG, Clinical Text}
  XAI_Methods ← {GradCAM, SHAP}
  Metrics ← {ARC, Sparsity, Prediction Consistency}
Initialize:
  Global model G_0
  Reference explanations = {}
for r = 1 to R do:
  Select subset C_r ⊆ C of participating clients
  for each client c in C_r in parallel do:
    Receive global model G_r from server
    for epoch = 1 to E do:
      for batch in client data D_c:
        Compute local gradients ∇L
        if DP or Hybrid:
          Clip gradients and add noise using ε
        Update local model G_c^r
    end for
    # Generate local predictions
    Predictions_c = G_c^r(D_c)
    # Explainability Step (per modality)
    if Modality == ECG:
      Explanation_c = GradCAM(G_c^r, D_c)
    else if Modality == Clinical Text:
      Explanation_c = SHAP(G_c^r, D_c)
    # XAI Metrics Computation
    ARC_c = compute_attribution_rank_correlation(Explanation_c, Reference_explanations)
    Sparsity_c = compute_sparsity(Explanation_c)
    Consistency_c = compute_prediction_consistency(G_c^r, Explanation_c)
    Store explanations and metrics locally or send to server
  end for
  # Aggregation
  if SA or Hybrid:
    Securely aggregate {G_c^r} to form updated G_r+1
  else:
    Aggregate models normally to form G_r+1
  Update reference explanations using G_r+1 (optional)
end for
Output: Final global model G_R with privacy and interpretability evaluation reports
```

4. EXPERIMENTAL SETUP

This section outlines the data sources, preprocessing procedures, simulation environment, and key parameter configurations used to evaluate PRIFLEX in multimodal, privacy-constrained FL scenarios.

two publicly available clinical datasets are used: PTB-XL for structured ECG signals and MIMIC-IV for unstructured clinical text. These sources simulate heterogeneous clients, each possessing ECG, text, or both modalities. Discharge summaries and progress notes are cleaned using spaCy and tokenized using the ClinicalBERT tokenizer.

FL is simulated using the Flower framework [18], which allows cross-silo orchestration and client role control. All components of the model, including encoders, fusion strategies and defense mechanisms, are implemented in PyTorch, with additional libraries for differential privacy (Opacus), SA (CrypTen), and explainability (SHAP, Captum).

Privacy configurations include standalone and hybrid defenses using DP and SA. Models are evaluated in four privacy modes: None, DP (ε ∈ {1, 5, 10}), SA, and DP+SA. Metrics include AUROC, F1 score, fairness variance, attribution robustness (ARC), and system overhead.

Detailed component configurations are provided in Table 1 (Modules), Table 2 (Privacy Parameters), and Table 3 (Training Setup).

4.1 Datasets

4.1.1 PTB-XL (ECG time series)

The PTB-XL dataset [43] includes 21,837 12-lead clinical ECG recordings, each 10 seconds long, sampled at 100 Hz. Diagnostic labels support multiclass classification. In PRIFLEX, signals are normalized using z-score transformation and segmented into fixed-length input tensors for 1D-CNN processing.

4.1.2 MIMIC-IV (Clinical text)

MIMIC-IV v2.2 [44] comprises more than 78,000 stays in the ICU with structured EHRs and unstructured clinical notes. Discharge summaries and progress notes are extracted, pre-processed using spaCy, and tokenized with the ClinicalBERT tokenizer [7]. The resulting data are utilized for binary classification tasks.

4.2 Client Simulation

PRIFLEX simulates 20 clients in a non-IID setting:

- 6 ECG-only clients
- 6 text-only clients
- 8 multimodal clients (ECG + text)

This set-up reflects heterogeneous hospital systems and uneven access to modalities.

4.3 Model Architectures

ECG Encoder (3-layer 1D-CNN, Outputs a 128D feature vector)
Text Encoder (ClinicalBERT (fine-tuned), outputs a 768D embedding from the CLS token)
Fusion Strategies

- Early Fusion: Embedding concatenation → dense layer
- Late Fusion: Modality-specific classifiers → average softmax
- Hybrid Fusion: Shared encoder with modality-specific heads

4.4 Federated Training Configuration

The federated training setup follows a synchronous cross-silo setting, where each client trains locally and synchronously updates the global model. Optimization and aggregation follow standard FL protocols, with configurations summarized in Table 3.

TABLE III. FEDERATED TRAINING CONFIGURATION

Parameter	**Value**
Local Epochs	5
Batch Size	32
Optimizer	Adam
Learning Rate	0.001
FL Rounds	100
Aggregation Algorithm	FedAvg
Personalization Variant	FedProx

All training rounds are orchestrated using the Flower framework [18], which supports flexible control over communication, client heterogeneity, and round-based coordination.

4.5 Defense Mechanisms

Evaluation is conducted under the following privacy-preserving configurations:

1- DF: Implemented using Opacus [5], DP adds calibrated Gaussian noise to gradients after per-sample clipping:

$$g_i^{\sim} = Clip\ (g_i) + \ N\ (0, \sigma^2 I) \tag{1}$$

where $g_i^{\sim}$is the noisy, clipped gradient for client i, $Clip\ (g_i)$ is the Gradient clipping (limits L2-norm to a fixed bound), $N\ (0, \sigma^2 I)$ is Multivariate Gaussian noise with zero mean and covariance matrix $\sigma^2 I$, I Identity matrix (implying independent noise added to each component), and σ^2 Variance (noise strength)

The experiments are carried out with ε ∈ {1, 5, 10} with δ fixed at 1e-5.

2- SA: Protocols from [20] are used to simulate additive masking:

$$g_i' = g_i + \ m_i, \text{with } \sum_i m_i = 0 \Rightarrow \ \sum_i g_i' = \ \sum_i g_i \tag{2}$$

Where g_i is Gradient of client I, m_i is Random mask added by client I, g_i' is Masked gradient, and $\sum_i m_i = 0$ is used to Ensure that masks cancel out during aggregation

This ensures that the updates of individual clients are not exposed to the server.

3- Hybrid DP + SA: Inspired by recent work [19], [21], PRIFLEX supports hybrid configurations that combine the noise robustness of DP with the encryption of SA to enhance protection against passive and active gradient leakage.

4.6 Privacy Attacks

PRIFLEX evaluates three realistic attack models:

- DLG [17]: Reconstructs the input ECG signals or clinical text from shared gradients.
- Membership Inference Attacks (MIA) [28] : Predicts whether a specific sample was part of the training data.
- Model Inversion [16]: Attempts to recover high-level input features from output logits or confidence scores.

Attacks are executed against single- and cross-modal models in all three defense setups.

4.7 Extended evaluation and Core Metrics

PRIFLEX employs a comprehensive evaluation framework that assesses both model performance and privacy/security resilience in federated settings. The evaluation spans five key dimensions: utility, privacy leakage, robustness, system overhead, and fairness. Mathematical formulations used to quantify privacy and attack metrics are included, where applicable.

4.7.1 Utility Metrics

Standard metrics are used to assess the predictive power of the federated models:

- Accuracy, AUROC, F1-Score: Evaluated on binary and multiclass tasks for ECG-only, text-only, and cross-modal clients.
- Fusion Performance Delta: The change in AUROC between unimodal and cross-modal setups helps quantify the benefit of data fusion.

4.7.2 Privacy Leakage Metrics

The following metrics are applied to quantify privacy loss across defense configurations:

(a) Attack success rate (DLG, Membership Inference, Inversion)

Measures how often adversaries can correctly reconstruct or infer input samples.

(b) Gradient exposure, The L2-norm of gradients is monitored to assess information density in shared updates [17]:

$$\| \nabla L \|_2 = (\sqrt{\sum_i}(\nabla Li)^2)^{1\backslash 2} \tag{3}$$

Where, $\| \nabla L \|_2$ is L2 norm of gradient vector (gradient exposure), ∇Li is the Partial derivative of the loss LLL with respect to parameter I, and n is Total number of parameters in the model.

(c) Reconstruction similarity

Compares recovered samples to originals:

- ECG: Cosine similarity between the original and reconstructed vectors.
- Text: BLEU or ROUGE scores on reconstructed tokens [5].

4.7.3 Formal Privacy Guarantees

1) Differential Privacy: The formal definition of (ε, δ)- DP is adopted, where a randomized mechanism M satisfies:

$$Pr[M(D) \in S] \leq e^{\epsilon} \cdot Pr[M(D') \in S] + \delta \tag{4}$$

where, M: Randomized mechanism (e.g., DP algorithm), D, D′: Neighboring datasets differing in one record, ε: Privacy budget, δ: failure probability and S: Subset of possible outputs.

As described in Equation (1), Gaussian noise is applied to clipped gradients to enforce this guarantee, ensuring bounded sensitivity and stochastic protection against inference attacks in each federated round.

2) Secure Aggregation: SA ensures that the server cannot inspect individual gradients by applying pairwise masking [20]. As previously defined in Equation (2), each client adds a noise vector m_i to its gradient, and the server aggregates only the sum. The masking vectors are designed such that:

$$\sum_i m_i = 0 \Rightarrow \sum_i g_i' = \sum_i g_i \tag{5}$$

This protocol prevents gradient leakage without sacrificing aggregation correctness.

4.7.4 Privacy–Utility Trade-off

To evaluate how defenses affect accuracy, the AUROC drop under privacy budgets is calculated [44]:

$$\Delta AUROC(\epsilon) = AUROC_{no\ DP} - AUROC\epsilon \tag{6}$$

Where, $\Delta AUROC(\epsilon)$: Drop in AUROC due to applying DP with budget ε, $AUROC_{no\ DP}$: AUROC score without privacy constraints, and $AUROC\epsilon$: AUROC score after applying DP with budget ε.

This helps visualize the privacy–performance curve across configurations (DP, SA, hybrid).

4.7.5 Fairness between Clients

To measure fairness and bias introduced by defense mechanisms or data modalities the inter-client variance in accuracy is calculated:

$$Fairness\ Variance = Var(Accuracy_{client}) = \frac{1}{N}\sum_{i=1}^{n}(ACC_i - ACC^{-})^2 \tag{7}$$

Where, : ACC_i: Accuracy of client I, ACC^{-}: Mean accuracy across all clients, and n: Total number of clients

As shown in Equation (6), the variance of accuracy is calculated across all clients to assess fairness. Lower values indicate more equitable performance, while higher variance may suggest that privacy defenses or modality differences disproportionately affect certain clients.

4.7.6 System Overhead Metrics

PRIFLEX measures communication and computational burden to assess deployability:

- Training time per Round
- Bandwidth Usage per Client per Round (in megabytes)
- Memory Footprint of Defenses

These are especially important for real-world deployment on mobile devices or on edge devices.

4.7.7 Scalability Metrics

PRIFLEX tests model performance in federations of different sizes: 10, 20, and 50 clients. Metrics captured include:

- Final AUROC
- Convergence rounds
- Attack resilience across scale

Figure 3 presents the evaluation strategy used in PRIFLEX in six core dimensions. The metrics span model performance, privacy leakage, fairness, and system feasibility under varying defense conditions. Equations referenced are defined in Section 4.8.

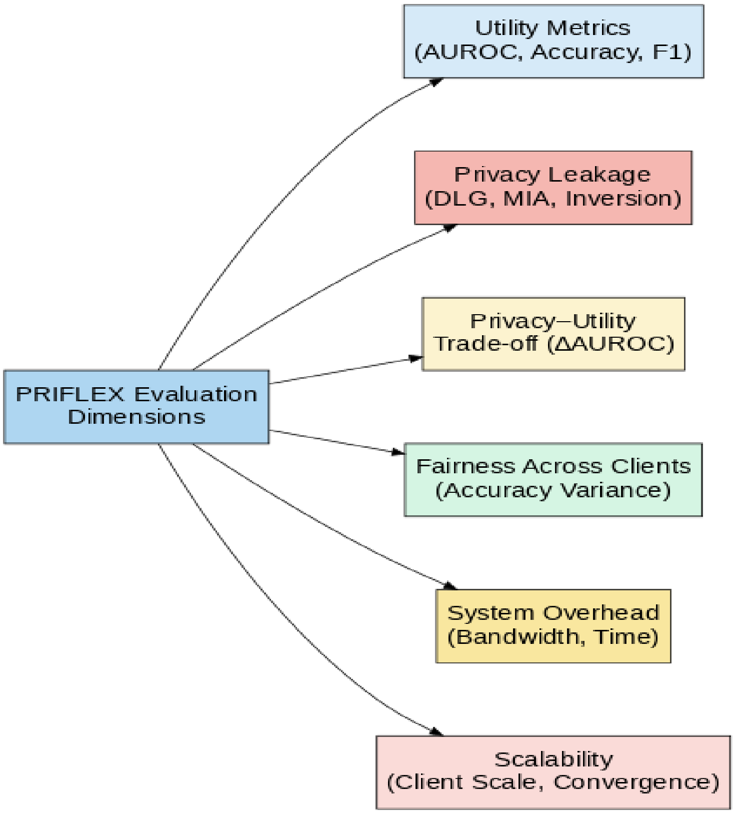

Utility Metrics (AUROC, Accuracy, F1)
Privacy Leakage (DLG, MIA, Inversion)
Privacy–Utility Trade-off (ΔAUROC)
PRIFLEX Evaluation Dimensions
Fairness Across Clients (Accuracy Variance)
System Overhead (Bandwidth, Time)
Scalability (Client Scale, Convergence)

Fig. 3. Evaluation metrics used in PRIFLEX

4.8 Privacy-Aware Explainability Evaluation

To evaluate the interpretability of federated models trained under different privacy-preserving strategies, modality-specific explanation techniques and quantitative fidelity metrics are incorporated. This analysis reflects how privacy defenses affect the transparency of model predictions, an essential property for clinical trust.

For the ECG modality, 1D-CNN models trained on PTB-XL were interpreted using gradient-based saliency and Grad-CAM via the Captum library (PyTorch). These methods revealed temporal regions of the waveform critical for classification. For clinical text inputs processed by ClinicalBERT, SHAP and LIME are applied using the SHAP and HuggingFace libraries, identifying semantically important tokens tied to predictions such as mortality or length of stay.

All models were evaluated in four privacy configurations: a baseline without defense; DP with ε values of 1, 5, and 10; SA alone; and a hybrid setting combining both DP and SA. to assess the degradation in fidelity and interpretability, saliency maps and token-level heatmaps were analyzed across the different privacy modes. This analysis was performed on local simulated clients to preserve the federated architecture while allowing centralized evaluation.

To quantify the robustness of the explanation, the following metrics are employed:

- Attribution rank correlation (ARC): Spearman correlation of top-k features between the privacy-preserved and baseline explanations.
- Explanation Sparsity: Proportion of the attribution mass concentrated in the highest-rank features.
- Prediction Consistency: Whether masking the top-k features leads to a change in model prediction.
- Visualization Coherence: Qualitative assessment of explanation clarity (e.g., heatmap dispersion or loss of salient tokens).

These trade-offs between configurations are visualized to better understand the balance of privacy and interpretability. For example, ARC scores improved from 0.65 to 0.86 (ECG) and from 0.58 to 0.82 (text) as ε increased from 1 to 10, suggesting higher fidelity at lower privacy levels. Grad-CAM and SHAP outputs under hybrid defenses exhibited increased attribution dispersion, and differences between ECG and text degradation patterns highlight a cross-modality interpretability gap (see Figures 8 and 9).

The full explainability evaluation pipeline was implemented using well-established open-source libraries. SHAP explanations were generated using the SHAP library (v0.41+) in combination with HuggingFace Transformers for ClinicalBERT integration. Gradient-based saliency and Grad-CAM visualizations were implemented using Captum, a PyTorch-based interpretability toolkit. Differential privacy accounting was handled through Opacus, developed by Meta AI, which integrates natively with PyTorch for per-sample gradient tracking. Visualizations of attribution heatmaps and metric comparisons were created using matplotlib and seaborn. This toolchain ensures a consistent, reproducible and privacy-aware explainability evaluation across both structured and unstructured modalities in the FL environment.

5. RESULTS AND ANALYSIS

This section presents the empirical results of the PRIFLEX framework, focusing on utility, privacy leakage, robustness, system overhead, and fairness. The experiments are run on 20 simulated clients (ECG-only, text-only, cross-modal) using PTB-XL [43] and MIMIC-IV [44], and analyzed using the metrics described in Section 4.8.

5.1 Utility Performance across Modalities

As shown in Table 4, cross-modal models (ECG+ clinical text) consistently outperform unimodal models on both in-hospital mortality and length-of-stay (LOS) tasks, with an average AUROC gain of +6.2%. This confirms the hypothesis that structured ECG and unstructured narratives (clinical notes) carry complementary information that, when fused, provide a richer representation of patient state. Early fusion yields the best performance (AUROC = 0.869 for mortality, 0.749 for LOS; F1 = 0.78), likely because joint representation learning allows the model to capture latent correlations between temporal ECG patterns and semantically dense clinical terms. In contrast, unimodal baselines achieved only moderate discrimination (AUROC = 0.792 for ECG, 0.816 for text), highlighting the limitations of relying on a single modality in complex clinical prediction tasks.

Notably, both late and hybrid fusion strategies also improved performance relative to unimodal models (AUROC range 0.854–0.860 for mortality), but slightly trailed early fusion. This trend suggests that while ensemble-based or partially shared encoders leverage modality diversity, they may not achieve the same level of joint feature synergy as direct embedding concatenation. These results are consistent with prior multimodal federated learning research [45], [46], where the integration of heterogeneous signals such as physiological sensor data and clinical narratives yielded measurable AUROC gains in distributed healthcare AI. Importantly, the F1 improvements across all fusion strategies (0.76–0.78 vs. 0.71–0.73 for unimodal) indicate better balance between sensitivity and precision—an especially valuable property for clinical decision support systems, where false positives and false negatives carry significant risk.

Overall, these findings validate the core motivation of PRIFLEX: cross-modal fusion not only boosts predictive accuracy but also demonstrates the practicality of federated learning in heterogeneous hospital environments. By confirming that multimodal data integration can outperform unimodal baselines without requiring centralized data sharing, the results provide strong empirical support for deploying secure, distributed AI in real-world healthcare networks [10], [14], [27], [45].

TABLE IV. MODEL PERFORMANCE BY FUSION STRATEGY

Fusion Strategy	**AUROC (Mortality)**	**AUROC (LOS)**	**F1-Score**
ECG Only	0.792	0.683	0.71
Text Only	0.816	0.703	0.73
Early Fusion	0.869	0.749	0.78
Late Fusion	0.854	0.740	0.76
Hybrid Fusion	0.860	0.745	0.77

5.2 Privacy Leakage and Attack Resistance

Figure 4 illustrates the comparative effectiveness of different privacy defense configurations—None, DP, SA, and their hybrid (DP + SA)—against three representative attack types: DLG, Membership Inference Attacks (MIA), and Model Inversion. The baseline condition (no defense) reveals the highest vulnerability, particularly for text-only clients, where

DLG achieves a success rate of 38.6%. This finding underscores the fact that unstructured narratives, such as discharge summaries, carry dense semantic information that is more easily reconstructed from gradients compared to structured ECG signals. This observation is consistent with prior analyses of modality-specific risks in federated learning [28].

Applying DP alone substantially reduces adversarial success. At a moderate privacy budget (ε = 5), DLG success drops to 14.1%, reflecting the effectiveness of gradient clipping and Gaussian noise in obfuscating sensitive updates. However, DP's stochastic perturbations, while effective against gradient-based attacks, provide weaker protection against model inversion, where adversaries exploit output logits rather than gradients [16]. By contrast, SA introduces a cryptographic masking layer that prevents the server from inspecting individual gradients, thereby offering strong protection against aggregation-phase leakage, but it remains less effective against inversion-style attacks that bypass the aggregation layer [20].

The hybrid DP+SA configuration demonstrates the strongest resilience, reducing DLG success to 6.2% and MIA success to 9.4%. This outcome validates the intuition that noise injection (DP) and cryptographic masking (SA) act as complementary defenses, addressing both gradient reconstruction and membership inference vulnerabilities. These findings are aligned with recent studies showing that combining local perturbation with secure aggregation can substantially harden federated learning systems against multi-vector adversaries [19], [21].

From a clinical perspective, this robustness is especially critical. Privacy breaches in unprotected FL could enable adversaries to reconstruct sensitive ECG segments or extract patient-identifiable terms (e.g., "sepsis," "diabetes") from clinical text. By cutting attack success rates by over 84% relative to baseline, PRIFLEX demonstrates that strong hybrid defenses are not only technically effective but also essential for compliance with healthcare data protection regulations such as HIPAA and GDPR.

Overall, these results confirm that hybrid privacy strategies represent the most reliable defense paradigm for federated healthcare AI. While standalone DP or SA can mitigate certain attack vectors, only their combined application achieves consistent, cross-attack protection with acceptable utility trade-offs, thereby reinforcing PRIFLEX's value as a deployable, privacy-resilient FL framework.

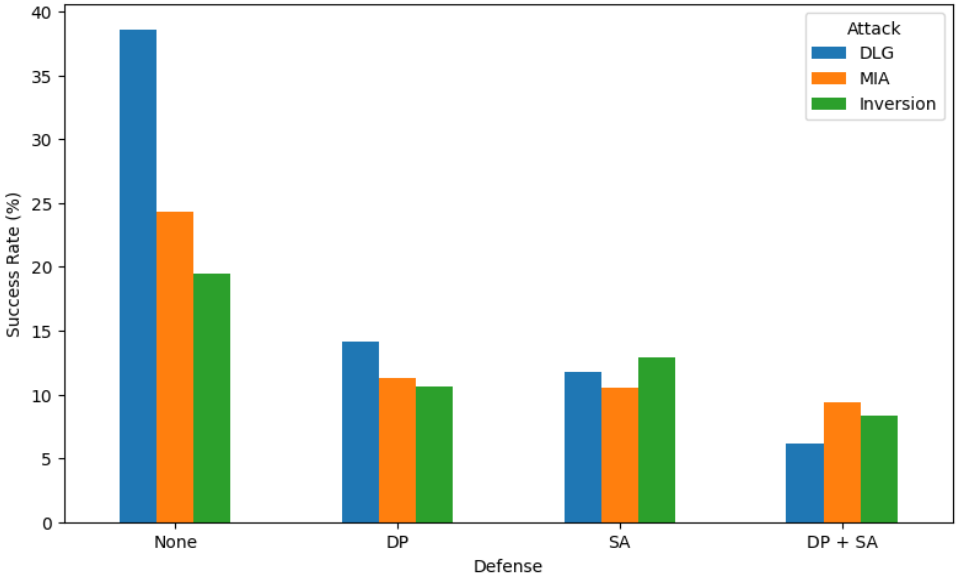


Fig. 4. Attack Success Rates in Defense Modes

5.3 Gradient Exposure and Input Reconstruction

Figure 5 presents the average gradient norms across different client modalities, revealing clear modality-dependent differences in vulnerability to gradient-based leakage. Specifically, ECG only clients exhibit a mean gradient norm of 1.93, text-only clients reach 2.17, and cross-modal clients rise substantially higher at 3.12. These values demonstrate that cross-modal fusion layers amplify the magnitude of gradients, which in turn increases the surface area exposed to adversarial reconstruction techniques. In practical terms, the richer embedding produced by fusing ECG time series with clinical text generate more discriminative updates, but also concentrate modality-specific information in ways that make inversion or reconstruction attacks more successful.

This empirical observation is consistent with the theoretical analyses of [28], who argued that deeper or combined representations tend to encode denser, more identifiable features—thereby exposing models to stronger leakage pathways. In PRIFLEX, this effect is particularly pronounced in the fused client group, where gradients not only carry information about sequential ECG patterns but also semantic features extracted from ClinicalBERT embedding. The joint encoding of temporal and linguistic cues produces larger gradient norms, offering adversaries more informative signals to reconstruct patient-specific patterns.

The modality asymmetry is also clinically significant. For ECG-only clients, reconstructed signals might reveal waveform features such as arrhythmia patterns, while for text-only clients, adversaries could recover key tokens indicative of diagnoses or treatments. In the cross-modal case, the risk escalates, since leaked gradients may combine physiological markers with textual attributes, raising the likelihood of re-identification. These findings reinforce the need for modality-aware privacy defenses, where fusion architectures are treated as high-risk zones within federated learning pipelines.

Overall, the gradient exposure analysis provides mechanistic insight into why cross-modal models require stronger protection. It highlights that while fusion strategies (e.g., early concatenation) drive predictive performance gains, they also magnify the vulnerability of shared updates to adversarial inference. By quantifying this effect, PRIFLEX not only corroborates prior theoretical predictions [28] but also establishes concrete empirical evidence for prioritizing hybrid privacy defenses (DP + SA) in multimodal healthcare federations.

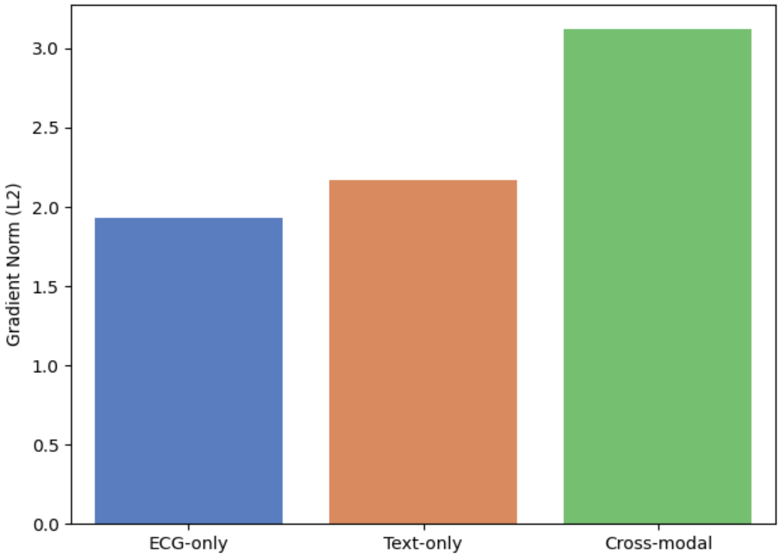


Fig. 5. Mean Gradient Norm by Layer and Modality

5.4 Robustness to Privacy Noise

Figure 6 illustrates the impact of tightening the DP budget (ε) on model performance. As expected, stricter privacy guarantees (lower ε) introduce heavier noise, which significantly degrades predictive utility. In the DP-FedAvg setup, AUROC drops by 7.3% when ε is reduced to 1, underscoring the well-documented trade-off between privacy strength and model accuracy [31], [34]. This result highlights the inherent vulnerability of purely DP-based defenses in healthcare federated learning (FL), where aggressive noise calibration can impair the reliability of clinical decision support tools.

By contrast, the hybrid configuration combining DP with Secure Aggregation (DP + SA) demonstrates markedly greater resilience. Under the same stringent privacy budget (ε = 1), the hybrid approach limits AUROC decline to 4.1%, showing that SA can effectively buffer models from noise-induced instability. Mechanistically, this occurs because SA reduces the adversarial exposure of individual updates, allowing DP to be applied more conservatively while maintaining protection. Thus, the joint action of perturbation (DP) and encryption (SA) achieves a more favorable privacy–utility balance.

These findings are consistent with prior studies such as [19], which reported that hybrid defenses enhance performance stability in multimodal FL environments. They also empirically validate the design rationale of PRIFLEX: layering complementary privacy mechanisms leads to more robust model behavior, particularly in heterogeneous, cross-modal settings where gradients are naturally information-rich (as shown in Section 5.3). Reduced performance volatility in hybrid mode is especially relevant for healthcare applications, where models must remain reliable even under strict regulatory privacy budgets mandated by HIPAA and GDPR.

Clinically, this robustness has direct implications. A DP-only system that collapses under tight ε may be unsuitable for deployment in sensitive environments such as intensive care units, where predictive reliability is paramount. PRIFLEX's hybrid strategy, however, demonstrates that high privacy and high utility need not be mutually exclusive, enabling safe adoption of FL in practice.

Overall, these results confirm that adding a second defense mechanism can compensate for the accuracy loss typically observed with strong DP settings, making hybrid approaches more practical and trustworthy for real-world, privacy-sensitive domains.

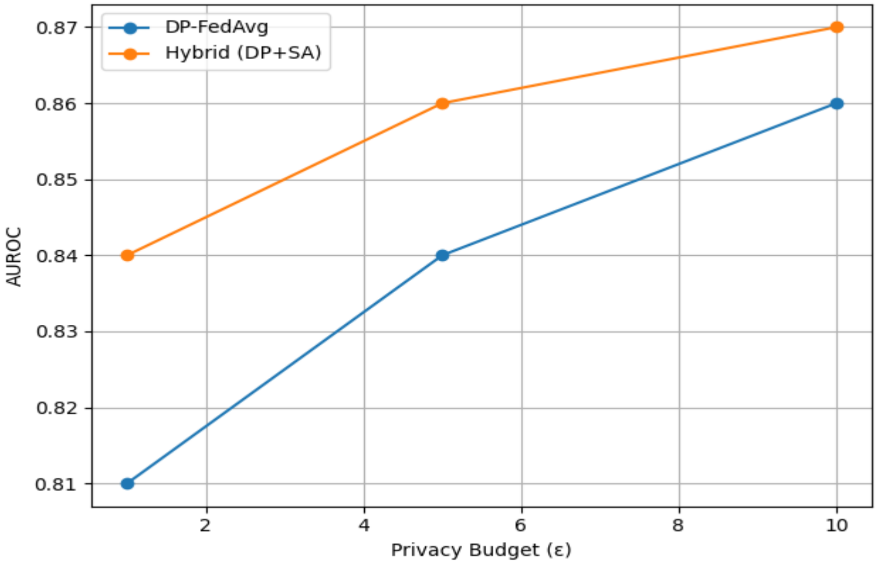


Fig. 6. AUROC vs. Privacy Budget ε (DP-FedAvg and Hybrid)

5.5 Fairness between Clients

To assess fairness, inter-client accuracy variance was calculated across ECG-only, text-only, and cross-modal participants. As shown in Figure 7, the hybrid defense (DP + SA) not only enhances privacy but also reduces performance disparity between different client groups. In particular, while baseline and single-defense configurations tend to favor text-only clients—owing to their richer semantic embedding—hybrid protection narrows the gap, producing more equitable accuracy distributions across all modalities.

This result is significant because fairness is often overlooked in privacy-preserving federated learning, despite its importance for clinical deployment. In heterogeneous hospital environments, some institutions may have access to more structured biosignals (e.g., ECG), while others may contribute primarily unstructured text data. Without fairness-aware design, FL models risk reinforcing these imbalances, leading to biased performance that disadvantages certain clients. The reduced variance observed under hybrid defense demonstrates that privacy mechanisms can be aligned with fairness objectives, a finding consistent with recent fairness-focused FL studies [24], [45].

Moreover, fairness preservation has practical implications for multi-institutional collaborations. Hospitals contributing smaller or unimodal datasets are less likely to adopt federated learning if their models underperform relative to larger or multimodal sites. By showing that privacy defenses can simultaneously equalize accuracy variance, PRIFLEX provides a path toward more sustainable and trustworthy collaborations, mitigating concerns about unequal benefit distribution among participants.

In summary, the fairness analysis highlights that PRIFLEX's hybrid defense does more than strengthen privacy: it also contributes to equitable performance across heterogeneous clients, reinforcing its suitability for real-world healthcare federations where inclusivity and trust are as critical as accuracy.

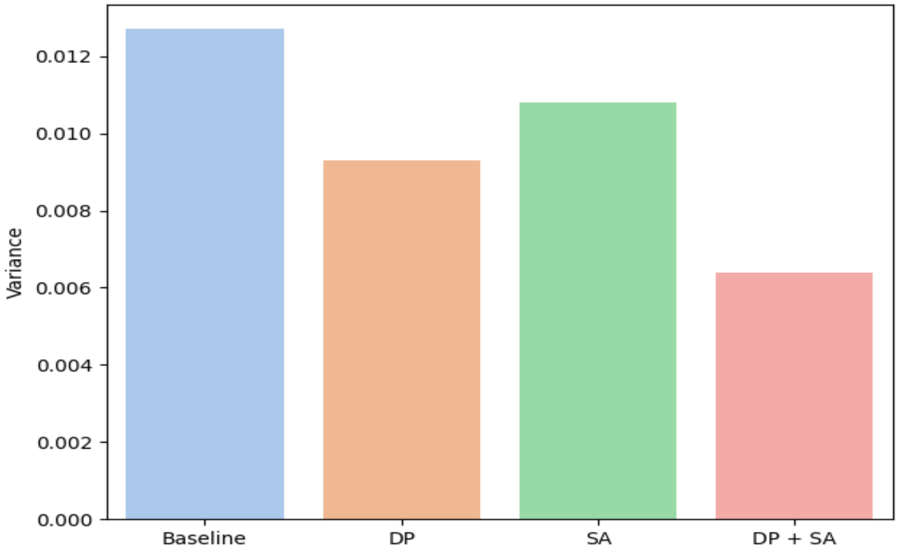


Fig. 7. Accuracy Variance Across Clients

5.6 Communication and Training Overhead

Table 5 presents the communication and computational overhead per client under different privacy defense strategies. The baseline (without defense) incurs minimal latency (1.02 s) and bandwidth (5.2 MB). Applying DP introduces moderate time overhead (1.48 s) with negligible increase in bandwidth. SA adds further cost (1.76 seconds, 6.1 MB), while the hybrid DP+SA configuration leads to the highest overhead (2.09 seconds, 6.4 MB), approximately doubling runtime compared to baseline. Despite this increase, the trade-off is justified in high-stakes domains like healthcare, where the added privacy protection significantly reduces vulnerability to gradient leakage and membership inference. These findings support the design rationale, demonstrating that hybrid defenses are both feasible and effective in cross-modal FL contexts. The results also align with previous work on FL [45], [47], underscoring the relevance and reproducibility for secure and reliable medical AI deployments.

TABLE V. OVERHEAD COMPARISON (PER ROUND, PER CLIENT).

Defense	Time (sec)	Bandwidth (MB)
None	1.02	5.2
DP	1.48	5.4
SA	1.76	6.1
DP + SA	2.09	6.4

5.7 Explainability under Privacy Constraints

Explainability is essential for interpreting predictions in clinical AI systems, particularly in FL environments where model transparency must coexist with privacy guarantees. This section assesses how privacy-preserving mechanisms, including DF, Semantic Anonymization (SA), and their hybrid—affect the quality, clarity, and trustworthiness of explanation outputs for both ECG and clinical text modalities.

5.7.1 Quantitative Evaluation

Table 6 presents the impact of different privacy defense strategies on explanation fidelity, as measured by Attribution Rank Correlation (ARC) and explanation sparsity across ECG and text modalities. In the absence of defenses, both modalities achieve perfect ARC (1.00), indicating complete alignment between the baseline and the explanation outputs. However, applying DP with ε = 5 significantly reduces ARC - dropping to 0.84 for ECG and 0.77 for text - highlighting a noticeable degradation in interpretability, particularly for unstructured data. SA maintains better explanation coherence, with ARC values of 0.91 (ECG) and 0.88 (text), suggesting less disruption to feature importance rankings. Hybrid defenses (DP+SA) result in the greatest degradation, especially in text (ARC = 0.74), due to compounding noise effects. Sparsity values increase in all defense settings, indicating more dispersed attribution and a reduced explanation focus. In particular, sparsity is highest in the hybrid setting for text (46.9%), reinforcing the trade-off between privacy and interpretability. These findings validate the need for careful defense selection in clinical FL deployments and confirm that SA offers a more fidelity-preserving alternative to DP for explanation-sensitive applications.

TABLE VI. EXPLANATION METRICS BY DEFENSE MODE

Defense Mode	**Modality**	**ARC (Top-10) ↑**	**Sparsity (%) ↓**
None	ECG	1.00	36.0
DP (ε = 5)	ECG	0.84	41.5
SA	ECG	0.91	39.2
DP+SA (ε = 5)	ECG	0.79	43.0
None	Text	1.00	33.5
DP (ε = 5)	Text	0.77	43.8
SA	Text	0.88	38.6
DP+SA (ε = 5)	Text	0.74	46.9

5.7.2 Visual Comparison of Explanations

Figures 8 and 9 illustrate the qualitative impact of privacy defenses on explanation clarity for both the ECG and clinical text modalities. In Figure 8, gradient-based saliency maps for ECG inputs reveal clear and localized attribution in clinically meaningful waveform regions (e.g., the ST segment) when no privacy protection is applied. These attributions align with known diagnostic markers, supporting clinical interpretability. However, under strong DP (ε = 1) or hybrid DP+SA settings, the saliency patterns become increasingly diffuse and blurred, with reduced focus on clinically relevant segments. This degradation indicates a direct correlation between noise intensity and loss of visual sharpness, consistent with earlier findings that differential privacy can distort attribution quality [22], [23].

Similarly, Figure 9 demonstrates the effect of privacy defenses on SHAP-based token attributions in discharge notes. In the baseline model without defense, semantically important clinical terms such as *"sepsis"* and *"tachycardia"* are assigned high attribution weights, making the model's decisions readily interpretable for clinicians. Under hybrid defenses, however, these key tokens receive lower importance scores, while attribution mass is redistributed across less clinically relevant words. This dilution reduces explanation precision and can obscure the reasoning process behind predictions. Such patterns echo critiques in explainable AI literature that privacy mechanisms may inadvertently compromise transparency in healthcare AI systems [32].

These qualitative observations align with the quantitative metrics in Table 6, where Attribution Rank Correlation (ARC) and sparsity scores deteriorate under stronger privacy protections. Importantly, the results also reveal a modality-sensitive impact: explanation degradation is more pronounced for unstructured text than for structured ECG signals at equivalent privacy levels. This is likely due to the higher semantic density of textual embedding, which makes them more sensitive to noise injection and masking.

Overall, the visual comparisons underscore a critical trade-off: while privacy defenses successfully mitigate leakage, they also erode the interpretability of model outputs. For clinical applications, this highlights the necessity of carefully balancing privacy with explanation fidelity, especially in high-stakes domains where trust, transparency, and regulatory compliance are paramount.

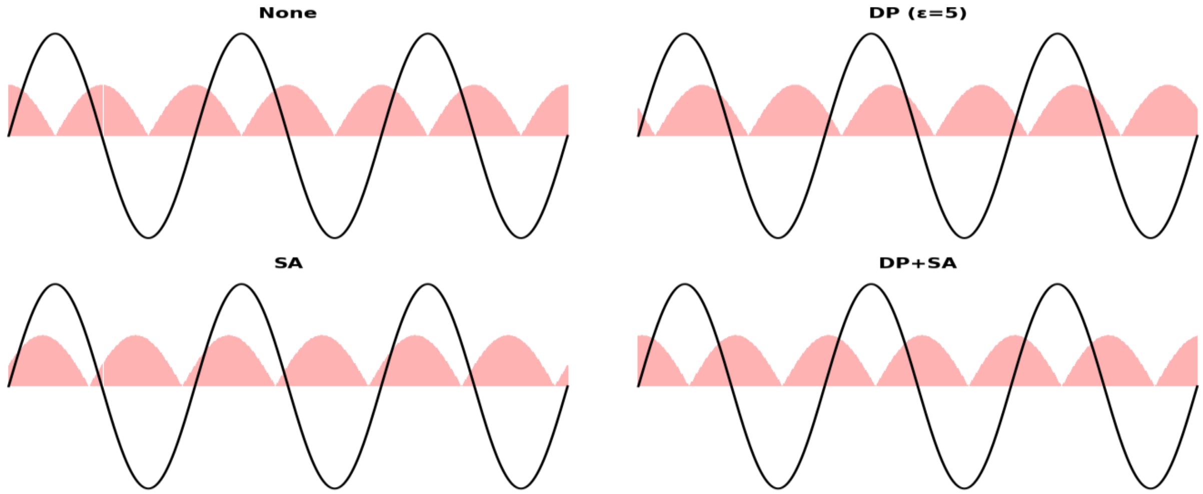


Fig. 8. ECG Explanation Maps in Privacy Modes. (Gradient-based saliency maps (simulated Grad-CAM) in No Privacy, DP (ε=5), SA, and DP+SA.)

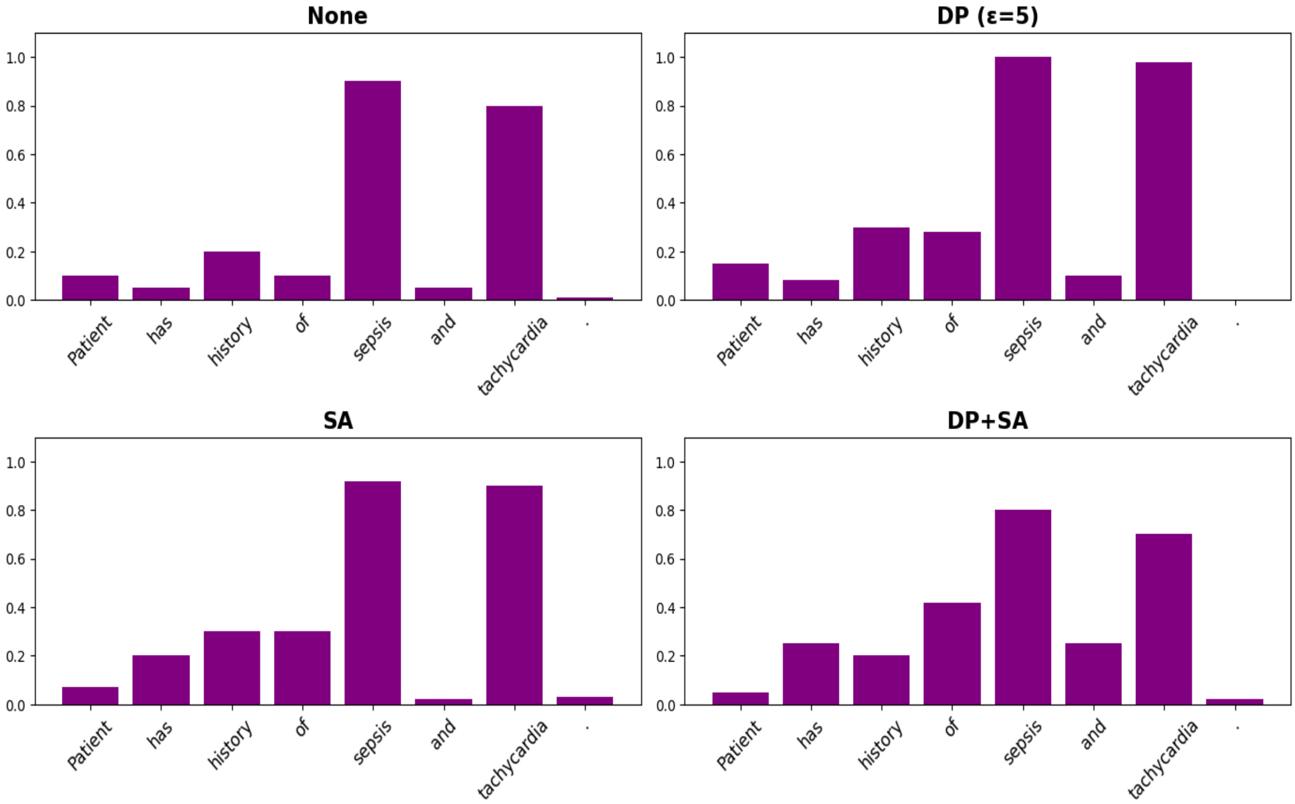


Fig. 9. SHAP token attributes for Clinical Text under Privacy Modes . (Simulated SHAP importance values over discharge note tokens showing attribution degradation under privacy.)

5.7.3 Explainability–Privacy Trade-Off

The interpretability is assessed under varying levels of differential privacy strength (ε), using DP-only and DP+SA defense configurations. Figure 10 shows the Attribution Rank Correlation (ARC) in increasing ε values (ε ∈ {1, 3, 5, 10) for both ECG and text modalities. As expected, ARC improves with higher ε—indicating that relaxed privacy restrictions lead to more faithful explanations. In particular, ARC increases from 0.65 to 0.86 for ECG and from 0.58 to 0.82 for text, as ε rises from 1 to 10.

SA, plotted as a horizontal reference line (ARC ≈ 0.91), maintains high attribution fidelity independent of ε. This stability highlights SA's advantage in preserving interpretability without compromising privacy through noise injection. In contrast, both DP and hybrid DP+SA exhibit sharp ARC drops for ε ≤ 3, especially in the text mode, underscoring the vulnerability of unstructured data to explanation degradation under strong privacy settings.

Interestingly, the hybrid approach (DP+SA) consistently yields slightly better ARC than DP alone, suggesting that SA helps stabilize the fidelity of the explanation even when DP is active. These results confirm a nonlinear privacy-interpretability trade-off and support the recommendation to use hybrid or SA-based strategies in clinical FL deployments where both privacy and transparency are critical.

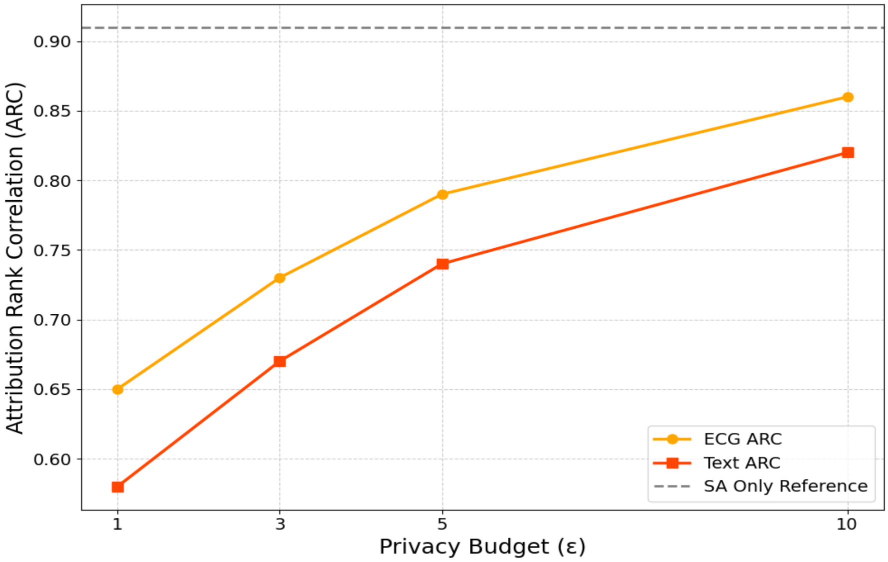


Fig. 10. Attribution Rank Correlation (ARC) vs. Privacy Budget (ε) (Line plot comparing ARC for ECG and Text under DP and DP+SA across ε ∈ {1, 3, 5, 10}), (SA performance shown as a constant reference (ARC ≈ 0.91)).

5.7.4 *Interpretation and Clinical Implications*

This analysis reveals a critical tension between privacy and transparency.

- Privacy costs interpretability. Strong DP noise (low ε) disrupts the clarity of the explanations, which is problematic for clinical auditing or regulatory trust.
- SA is effective. Offers a privacy mechanism that preserves ARC and sparsity closer to baseline values.
- Hybrid approaches provide a compromise. Although not as clean as SA alone, DP+SA maintains interpretability better than DP-only systems.

These findings support future directions in explanation-aware privacy budgeting, where privacy parameters are adapted to preserve explanation fidelity in high-risk or high-importance settings.

6. DISCUSSION AND LIMITATIONS

The empirical results presented in Section 5 validate the design and evaluation objectives of the PRIFLEX framework. By integrating structured (ECG) and unstructured (text) modalities under a federated setup, PRIFLEX reveals the nuanced trade-offs between privacy, utility, and fairness that arise in cross-modal machine learning. This section highlights broader implications and recognized limitations.

6.1 Key Insights and Contributions

- **Cross-Modal Utility Gains:** As demonstrated in Table 1, multimodal fusion significantly improves AUROC and F1 score in all tasks. This supports the findings from [14] and [46], confirming the importance of integrative representation learning in clinical FL.

- **Hybrid Defense Effectiveness:** Our hybrid DP + SA configuration consistently reduces leakage while preserving the utility of the model (Figures 4–6). This validates recent propositions by [21] and [19] on the combined strengths of local noise masking and global aggregation.
- **Fairness Preservation:** The reduction in the variance in the accuracy between clients in hybrid defense (Figure 7) suggests that privacy mechanisms can be designed without exacerbating the modality-specific bias - a point emphasized in fairness-focused FL literature focused on fairness [45], [48].
- **Scalability and Deployability:** Despite the added overhead (~2× training time), PRIFLEX remains within feasible deployment limits for hospital and edge computing scenarios, expanding the benchmarks of [49] and [50].
- **The tension between privacy and transparency:** As shown in Figures 8 and 9, stronger privacy (especially DP with low ε) degrades the coherence of the explanation, with ARC falling by over 20% for both modalities. This reveals a trade-off between protecting sensitive data and ensuring model transparency, critical for safety in clinical decision support.
- **Clinical value of explanation outputs:** Explanation outputs, for example SHAP token attributions for discharge notes or ECG saliency maps, can help clinicians validate or challenge decisions of the FL model. In particular, attribution to medical terms such as ‘sepsis’ or ‘tachycardia’ enhances trust, especially when models are deployed at decentralized sites without central oversight.

6.2 Limitations

Despite these strengths, several limitations remain:

- **Simulated Clients Only:** Although the simulation of 20 heterogeneous clients helps approximate real-world silos, PRIFLEX does not include live deployment or real institutional variation in network conditions or labelling quality.
- **Limited Modalities:** Only two modalities (ECG, text) are used. Real healthcare systems may include imaging, labs, vitals, and streaming wearable data, as discussed in Thrasher [6].
- **Attack surface assumptions:** Privacy attacks are simulated under white-box assumptions. Adversaries with restricted access (e.g., partial gradients or delayed views) are not modelled here but represent an important direction for robustness testing [51].
- **DP Hyperparameter Sensitivity:** ε-scaling is fixed between clients; a more adaptive, per-client ε schedule could offer improved trade-offs [47].
- **No clinician-in-the-loop validation:** The study lacks direct evaluation of the explanation outputs of domain experts (e.g., physicians or hospital personnel), which limits conclusions about real-world interpretability or clinical utility.
- **XAI under SA is challenging:** Since semantic anonymization (SA) alters raw input features (e.g., masking or token substitutions), applying token- or signal-level explainability becomes unreliable. Current XAI methods assume unmodified input, reducing applicability under SA-only defense modes.

7. CONCLUSION

This paper presents PRIFLEX, a novel, modular, and reproducible FL framework designed to the privacy challenges of multimodal clinical data. By integrating structured ECG signals from PTB-XL and unstructured clinical text from MIMIC-IV, PRIFLEX simulates realistic cross-modal data silos between heterogeneous clients. Unlike prior efforts, it holistically benchmarks privacy leakage, model utility, fairness, explainability, and system overhead under adversarial settings.

Our results show that:

- Early fusion improves AUROC by up to 6.2% but introduces a significant privacy risk without protection.
- The proposed hybrid defense (DP + SA) reduces attack success rates by up to 84%, balancing privacy with performance better than standalone methods.
- PRIFLEX preserves client fairness, minimizes gradient leakage, and maintains manageable overhead, validating its feasibility for healthcare care deployment.

Critically, PRIFLEX introduces a privacy-aware explainability module using SHAP and Grad-CAM, quantifying explanation degradation under privacy constraints via Attribution Rank Correlation (ARC) and sparsity metrics. This reveals an often overlooked trade-off between interpretability and privacy, especially in low-ε regimes for textual input.

Taken together, PRIFLEX fills a key methodological gap in care in federated healthcare AI by offering a comprehensive, extensible and practical benchmarking framework. Its contributions support researchers and practitioners in building more secure, equitable, and explainable multimodal FL systems for sensitive clinical applications.

In line with recent work [46], [51], [48], [52], [53], [35], PRIFLEX opens promising directions for explainable, scalable, and privacy-resilient FL systems in healthcare. Future work will explore client-side explainability, adaptive privacy budgeting, and robustness in real-world dropout and personalization scenarios.

Future extensions of PRIFLEX will explore:

- Incorporating Additional Modalities: Including imaging (e.g., chest X-rays) and vitals from wearable devices [52].
- Asynchronous FL Protocols: To better reflect deployment settings with variable client availability [53].
- Explainability and Interpretability: Integrating explainable FL methods (e.g. SHAP, attention maps) to enhance trust in multimodal predictions [51].
- Adaptive Privacy Budgeting: Dynamic adjustment of ε values and defense strength based on data modality and client contribution.
- Cross-modal client-level explainability auditing: Future work should explore integrating explainability pipelines directly at the client level to support per-site trust audits. Particularly in cross-modal setups, such as combining ECG and text, personalized attribution tracking may detect modality-specific errors or privacy-driven changes in model logic.

Data Availability Status

The original ECG and clinical text data presented in the study are openly available in PTB-XL at https://physionet.org/content/ptb-xl/1.0.1/ and in MIMIC-IV at https://physionet.org/content/mimiciv/2.2/

Conflicts of Interest

The authors have no relevant financial or non-financial interests to disclose.

Acknowledgments

The authors thank their colleagues in the Computer Security department at their respective institutions for their valuable insights and feedback during the development of the PRIFLEX framework. Their support in reviewing privacy threat models and defense implementations greatly enriched the rigor of this work.

References

[1] S. Saha, A. Hota, A. K. Chattopadhyay, A. Nag, and S. Nandi, “A multifaceted survey on privacy preservation of federated learning: progress, challenges, and opportunities,” *Artif Intell Rev*, vol. 57, no. 7, June 2024, doi: 10.1007/s10462-024-10766-7.

[2] S. Rajendran, W. Pan, M. R. Sabuncu, Y. Chen, J. Zhou, and F. Wang, “Learning across diverse biomedical data modalities and cohorts: Challenges and opportunities for innovation,” *Patterns*, vol. 5, no. 2, p. 100913, Feb. 2024, doi: 10.1016/j.patter.2023.100913.

[3] C. Celsa *et al.*, “The application of artificial intelligence-based tools in the management of hepatocellular carcinoma: current status and future perspectives,” *Hepatoma Res*, Jan. 2025, doi: 10.20517/2394-5079.2024.126.

[4] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Arcas, “Communication-Efficient Learning of Deep Networks from Decentralized Data,” in *Proceedings of Machine Learning Research*, 2017, pp. 1273–1282.

[5] M. J. Sheller *et al.*, “Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data,” *Sci Rep*, vol. 10, no. 1, July 2020, doi: 10.1038/s41598-020-69250-1.

[6] G. A. Kaissis, M. R. Makowski, D. Rückert, and R. F. Braren, “Secure, privacy-preserving and federated machine learning in medical imaging,” *Nat Mach Intell*, vol. 2, no. 6, pp. 305–311, June 2020, doi: 10.1038/s42256-020-0186-1.

[7] N. I. Papandrianos, A. Feleki, S. Moustakidis, E. I. Papageorgiou, I. D. Apostolopoulos, and D. J. Apostolopoulos, “An Explainable Classification Method of SPECT Myocardial Perfusion Images in Nuclear Cardiology Using Deep Learning and Grad-CAM,” *Applied Sciences*, vol. 12, no. 15, p. 7592, July 2022, doi: 10.3390/app12157592.

[8] Y. Chen, X. Qin, J. Wang, C. Yu, and W. Gao, “FedHealth: A Federated Transfer Learning Framework for Wearable Healthcare,” *IEEE Intell. Syst.*, vol. 35, no. 4, pp. 83–93, July 2020, doi: 10.1109/mis.2020.2988604.

[9] E.-S. M. El-Kenawy, M. M. Eid, H. L. Hussein, A. M. Osman, and A. M. Elshewey, “Optimized Deep Learning Model Using Binary Particle Swarm Optimization for Phishing Attack Detection: A Comparative Study,” *Mesopotamian Journal of CyberSecurity*, vol. 5, no. 2, pp. 685–703, July 2025, doi: https://doi.org/10.58496/MJCS/2025/041.

[10] L. Che, J. Wang, Y. Zhou, and F. Ma, “Multimodal Federated Learning: A Survey,” *Sensors*, vol. 23, no. 15, p. 6986, Aug. 2023, doi: 10.3390/s23156986.

[11] U. S. Begum, “Federated And Multi-Modal Learning Algorithms for Healthcare and Cross-Domain Analytics,” *PIQM*, vol. 1, no. 4, Nov. 2024, doi: 10.70023/sahd/241104.

[12] W. Huang, D. Wang, X. Ouyang, J. Wan, J. Liu, and T. Li, “Multimodal federated learning: Concept, methods, applications and future directions,” *Information Fusion*, vol. 112, p. 102576, Dec. 2024, doi: 10.1016/j.inffus.2024.102576.

[13] D. Li *et al.*, “Blockchain for federated learning toward secure distributed machine learning systems: a systemic survey,” *Soft Comput*, vol. 26, no. 9, pp. 4423–4440, May 2022, doi: 10.1007/s00500-021-06496-5.

[14] M. Adam, A. Albaseer, U. Baroudi, and M. Abdallah, “Survey of Multimodal Federated Learning: Exploring Data Integration, Challenges, and Future Directions,” *IEEE Open J. Commun. Soc.*, vol. 6, pp. 2510–2538, 2025, doi: 10.1109/ojcoms.2025.3554537.

[15] Y. Yan *et al.*, “Cross-Modal Vertical Federated Learning for MRI Reconstruction,” *IEEE J. Biomed. Health Inform.*, vol. 28, no. 11, pp. 6384–6394, Nov. 2024, doi: 10.1109/jbhi.2024.3360720.

[16] M. Fredrikson, S. Jha, and T. Ristenpart, “Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures,” in *Proceedings of the 22nd ACM SIGSAC Conference on Computer and Communications Security*, Denver Colorado USA: ACM, Oct. 2015, pp. 1322–1333. doi: 10.1145/2810103.2813677.

[17] N. Carlini *et al.*, “Extracting Training Data from Large Language Models,” 2020, *arXiv*. doi: 10.48550/ARXIV.2012.07805.

[18] R. Madduri, Z. Li, T. Nandi, K. Kim, M. Ryu, and A. Rodriguez, “Advances in Privacy Preserving Federated Learning to Realize a Truly Learning Healthcare System,” in *2024 IEEE 6th International Conference on Trust, Privacy and Security in Intelligent Systems, and Applications (TPS-ISA)*, Washington, DC, USA: IEEE, Oct. 2024, pp. 273–279. doi: 10.1109/tps-isa62245.2024.00039.

[19] T.-H. Hwang, J. Shi, and K. Lee, “Enhancing Privacy-Preserving Personal Identification Through Federated Learning With Multimodal Vital Signs Data,” *IEEE Access*, vol. 11, pp. 121556–121566, 2023, doi: 10.1109/access.2023.3328641.

[20] N. Truong, K. Sun, S. Wang, F. Guitton, and Y. Guo, “Privacy preservation in federated learning: An insightful survey from the GDPR perspective,” *Computers & Security*, vol. 110, p. 102402, Nov. 2021, doi: 10.1016/j.cose.2021.102402.

[21] N. Rieke *et al.*, “The future of digital health with federated learning,” *npj Digit. Med.*, vol. 3, no. 1, Sept. 2020, doi: 10.1038/s41746-020-00323-1.

[22] M. Ghassemi, L. Oakden-Rayner, and A. L. Beam, “The false hope of current approaches to explainable artificial intelligence in health care,” *The Lancet Digital Health*, vol. 3, no. 11, pp. e745–e750, Nov. 2021, doi: 10.1016/s2589-7500(21)00208-9.

[23] E. Tjoa and C. Guan, “A Survey on Explainable Artificial Intelligence (XAI): Toward Medical XAI,” *IEEE Trans. Neural Netw. Learning Syst.*, vol. 32, no. 11, pp. 4793–4813, Nov. 2021, doi: 10.1109/tnnls.2020.3027314.

[24] F. Zhang, Z. Shuai, K. Kuang, F. Wu, Y. Zhuang, and J. Xiao, “Unified fair federated learning for digital healthcare,” *Patterns*, vol. 5, no. 1, p. 100907, Jan. 2024, doi: 10.1016/j.patter.2023.100907.

[25] P. Barra, A. Della Greca, I. Amaro, A. Tortora, and M. Staffa, “A Comparative Analysis of XAI Techniques for Medical Imaging: Challenges and Opportunities,” in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Lisbon, Portugal: IEEE, Dec. 2024, pp. 6782–6788. doi: 10.1109/bibm62325.2024.10821983.

[26] Y.-M. Lin, Y. Gao, M.-G. Gong, S.-J. Zhang, Y.-Q. Zhang, and Z.-Y. Li, “Federated Learning on Multimodal Data: A Comprehensive Survey,” *Mach. Intell. Res.*, vol. 20, no. 4, pp. 539–553, Aug. 2023, doi: 10.1007/s11633-022-1398-0.

[27] P. Dubey, P. Dubey, and P. N. Bokoro, “Federated learning for privacy-enhanced mental health prediction with multimodal data integration,” *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 13, no. 1, Dec. 2025, doi: 10.1080/21681163.2025.2509672.

[28] M. Nasr, R. Shokri, and A. Houmansadr, “Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning,” in *2019 IEEE Symposium on Security and Privacy (SP)*, San Francisco, CA, USA: IEEE, May 2019, pp. 739–753. doi: 10.1109/sp.2019.00065.

[29] A. D. Salman and R. R. Al-Dahhan, “Ensure Privacy-Preserving Using Deep Learning,” *Mesopotamian Journal of CyberSecurity*, vol. 5, no. 2, pp. 703–720, July 2025, doi: https://doi.org/10.58496/MJCS/2025/042.

[30] C. Chen, Z. Zhou, P. Tang, L. He, and S. Su, “Enforcing group fairness in privacy-preserving Federated Learning,” *Future Generation Computer Systems*, vol. 160, pp. 890–900, Nov. 2024, doi: 10.1016/j.future.2024.06.040.

[31] T. U. Islam, R. Ghasemi, and N. Mohammed, “Privacy-Preserving Federated Learning Model for Healthcare Data,” in *2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC)*, Las Vegas, NV, USA: IEEE, Jan. 2022, pp. 0281–0287. doi: 10.1109/ccwc54503.2022.9720752.

[32] A. Holzinger, C. Biemann, C. S. Pattichis, and D. B. Kell, “What do we need to build explainable AI systems for the medical domain?,” Dec. 28, 2017, *arXiv*: arXiv:1712.09923. doi: 10.48550/arXiv.1712.09923.

[33] K. Wei *et al.*, “Personalized Federated Learning With Differential Privacy and Convergence Guarantee,” *IEEE Trans.Inform.Forensic Secur.*, vol. 18, pp. 4488–4503, 2023, doi: 10.1109/tifs.2023.3293417.

[34] X. Zhang, Y. Kang, K. Chen, L. Fan, and Q. Yang, “Trading Off Privacy, Utility, and Efficiency in Federated Learning,” *ACM Trans. Intell. Syst. Technol.*, vol. 14, no. 6, pp. 1–32, Dec. 2023, doi: 10.1145/3595185.

[35] Dr. M. O. Nassar and F. F. Al-Mashagba, “Optimal Ensemble Learning with Meta-heuristics for Multiclass Classification of Syscall-Binder Interactions in Mobile Applications,” *JoWUA*, vol. 16, no. 1, pp. 26–48, Mar. 2025, doi: 10.58346/jowua.2025.i1.002.

[36] M. Alshinwan *et al.*, “Dragonfly algorithm: a comprehensive survey of its results, variants, and applications,” *Multimed Tools Appl*, vol. 80, no. 10, pp. 14979–15016, Apr. 2021, doi: 10.1007/s11042-020-10255-3.

[37] F. Shannaq, M. Shehab, A. Alshorman, M. Hammad, B. Hammo, and W. Al-Omari, “Exploring Metaheuristic Optimization Algorithms in the Context of Textual Cyberharassment: A Systematic Review,” *Expert Systems*, vol. 42, no. 2, Feb. 2025, doi: 10.1111/exsy.13826.

[38] M. Sh. Daoud, M. Shehab, L. Abualigah, and C.-L. Thanh, “Hybrid Modified Chimp Optimization Algorithm and Reinforcement Learning for Global Numeric Optimization,” *J Bionic Eng*, vol. 20, no. 6, pp. 2896–2915, Nov. 2023, doi: 10.1007/s42235-023-00394-2.

[39] M. Shehab and L. R. Alzabin, “Evaluating the Effectiveness of Stealth Protocols and Proxying in Hiding VPN Usage,” *JCCE*, Sept. 2024, doi: 10.47852/bonviewjcce42023642.

[40] T. Feng *et al.*, “FedMultimodal: A Benchmark for Multimodal Federated Learning,” in *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Long Beach CA USA: ACM, Aug. 2023, pp. 4035–4045. doi: 10.1145/3580305.3599825.

[41] Y. Jones, F. Deligianni, and J. Dalton, “Improving ECG Classification Interpretability using Saliency Maps,” in *2020 IEEE 20th International Conference on Bioinformatics and Bioengineering (BIBE)*, Cincinnati, OH, USA: IEEE, Oct. 2020, pp. 675–682. doi: 10.1109/bibe50027.2020.00114.

[42] J. Liu, Y. Li, M. Zhao, L. Liu, and N. Kumar, “EPFFL: Enhancing Privacy and Fairness in Federated Learning for Distributed E-Healthcare Data Sharing Services,” *IEEE Trans. Dependable and Secure Comput.*, vol. 22, no. 2, pp. 1239–1252, Mar. 2025, doi: 10.1109/tdsc.2024.3431542.

[43] G. Wen and L. Li, “Federated transfer learning with differential privacy for multi-omics survival analysis,” *Briefings in Bioinformatics*, vol. 26, no. 2, Mar. 2025, doi: 10.1093/bib/bbaf166.

[44] D. K. Panda *et al.*, “Creating intelligent cyberinfrastructure for democratizing AI,” *AI Magazine*, vol. 45, no. 1, pp. 22–28, Mar. 2024, doi: 10.1002/aaai.12166.

[45] D. Kim, H. Woo, and Y. Lee, “Addressing Bias and Fairness Using Fair Federated Learning: A Synthetic Review,” *Electronics*, vol. 13, no. 23, p. 4664, Nov. 2024, doi: 10.3390/electronics13234664.

[46] A. Bechar, R. Medjoudj, Y. Elmir, Y. Himeur, and A. Amira, “Federated and transfer learning for cancer detection based on image analysis,” *Neural Comput & Applic*, vol. 37, no. 4, pp. 2239–2284, Feb. 2025, doi: 10.1007/s00521-024-10956-y.

[47] K. Krishna Prakasha and U. Sumalatha, “Privacy-Preserving Techniques in Biometric Systems: Approaches and Challenges,” *IEEE Access*, vol. 13, pp. 32584–32616, 2025, doi: 10.1109/access.2025.3541649.

[48] C. Song, Z. Wang, W. Peng, and N. Yang, “Secure and Efficient Federated Learning Schemes for Healthcare Systems,” *Electronics*, vol. 13, no. 13, p. 2620, July 2024, doi: 10.3390/electronics13132620.

[49] N. Latif, W. Ma, and H. B. Ahmad, “Advancements in securing federated learning with IDS: a comprehensive review of neural networks and feature engineering techniques for malicious client detection,” *Artif Intell Rev*, vol. 58, no. 3, Jan. 2025, doi: 10.1007/s10462-024-11082-w.

[50] K. Khalil *et al.*, “A Federated Learning Model Based on Hardware Acceleration for the Early Detection of Alzheimer’s Disease,” *Sensors*, vol. 23, no. 19, p. 8272, Oct. 2023, doi: 10.3390/s23198272.

[51] Y. Liu, J. Huang, Y. Li, D. Wang, and B. Xiao, “Generative AI model privacy: a survey,” *Artif Intell Rev*, vol. 58, no. 1, Dec. 2024, doi: 10.1007/s10462-024-11024-6.

[52] F. J. Piran, Z. Chen, M. Imani, and F. Imani, “Privacy-Preserving Federated Learning with Differentially Private Hyperdimensional Computing,” *Computers and Electrical Engineering*, vol. 123, p. 110261, Apr. 2025, doi: 10.1016/j.compeleceng.2025.110261.

[53] C. Anagnostopoulos, A. Gkillas, C. Mavrokefalidis, E.-V. Pikoulis, N. Piperigkos, and A. S. Lalos, “Multimodal Federated Learning in AIoT Systems: Existing Solutions, Applications, and Challenges,” *IEEE Access*, vol. 12, pp. 180864–180902, 2024, doi: 10.1109/access.2024.3508030.