Privacy-Preserving Physiological Signal Classification in Mobile Health: A Comparative Study of Federated Learning and Differential Privacy

[1]Eduardo Carvalho, [1]Filipe Correia, [1]Vasco Fernandes, [1,2]Pedro Martins
[1]Department of Informatics, Polytechnic of Viseu, Viseu, Portugal
[2]Research Center in Digital Services, Polytechnic of Viseu, Viseu, Portugal
pv25218@alunos.estgv.ipv.pt, pv25005@alunos.estgv.ipv.pt, pv25177@alunos.estgv.ipv.pt, pedromom@estgv.ipv.pt

***Abstract*—Mobile health (mHealth) applications face a critical tension: extracting health insights via machine learning while protecting sensitive physiological data under GDPR and similar privacy regulations. Centralized data processing creates privacy vulnerabilities; decentralized approaches (Federated Learning) lack formal privacy guarantees.**

This paper evaluates privacy-preserving techniques—Federated Learning (FL), Differential Privacy (DP), and their combination (FL+DP)—across 78 experimental configurations on two physiological signal datasets: WESAD (stress detection, 15 subjects) and Sleep-EDF (sleep stage classification, 80 subjects). A Multi-Layer Perceptron trained on hand-crafted features achieves non-private baselines of 93.6% (WESAD) and 88.4% (Sleep-EDF).

Key findings reveal task-dependent privacy robustness: Sleep-EDF maintains strong privacy ($\epsilon \approx 0.2$) with ¡3% accuracy loss, while WESAD suffers 13-14% degradation. We attribute this difference to feature discriminability—Sleep-EDF's spectral features (delta, theta, alpha, beta power) have inherent separation robust to noise, whereas WESAD's HRV-based stress features are more sensitive to gradient perturbations.

FL+DP provides defense-in-depth protection: Sleep-EDF maintains ¿83% accuracy with 10 clients and strong privacy ($\epsilon \approx 2.3$), while WESAD requires weaker privacy ($\epsilon \approx 14.9 - 24.3$) to remain clinically useful.

We identify the *Gradient Clipping Trap*, where DP-SGD's gradient clipping neutralizes class weights, causing random seed initialization to dominate hyperparameter tuning by a 100:1 variance ratio. This mechanism undermines standard fairness interventions and requires algorithmic redesign.

Our findings provide practical guidance for mHealth developers: feature quality drives privacy robustness more than dataset size, and combined FL+DP is computationally feasible on resource-constrained devices while enabling formal privacy guarantees for physiological signals.

***Index Terms*—federated learning, differential privacy, mobile health, wearable devices, privacy-preserving machine learning, GDPR, physiological signals, class imbalance**

I. Introduction

A. Motivation: Privacy in Mobile Health

Wearable devices and mobile health (mHealth) applications have fundamentally transformed personal health monitoring. From smartwatches tracking heart rate to specialized sensors measuring sleep quality, these devices generate continuous physiological data streams offering unprecedented health insights. Machine learning analysis of this data enables stress detection, sleep disorder diagnosis, and personalized health interventions previously unavailable outside clinical settings.

However, this technological advancement creates a fundamental tension: the data being collected is among the most sensitive personal information imaginable. Physiological signals reveal not just physical health but also emotional states, daily routines, and behavioral patterns. Under GDPR and similar privacy frameworks worldwide, such data is classified as highly sensitive and requires strict protection through data minimization, purpose limitation, and explicit user consent.

Traditional machine learning approaches fundamentally conflict with these privacy obligations. The standard practice of transmitting raw sensor data to cloud servers for centralized training creates multiple vulnerability points: data transfers, server storage, and training runs each represent potential privacy breaches. The consequences are substantial: **5,887 healthcare data breaches were reported between 2009-2023**, exposing **519 million records collectively**, with **725 incidents in 2023 alone** (the highest year on record). Notable examples include the 2015 Anthem Inc. breach (78.8 million patients) and the 2024 Change Healthcare ransomware attack (potentially one-third of all Americans).

B. Privacy-Preserving Solutions

Two complementary approaches address these challenges:

Federated Learning (FL) keeps sensitive data on device by enabling collaborative model training without centralization. Rather than transmitting raw data, each device trains a model locally using only its private data, then shares only model updates. These updates are aggregated at a central server to form a global model, ensuring raw data never leaves the device where it was collected.

Differential Privacy (DP) provides mathematical guarantees that analyzing a dataset will not reveal information about any individual. It works by adding carefully calibrated noise during training, ensuring that the presence or absence of any single person's data has a bounded effect on the learned model.

Unlike heuristic privacy claims, DP offers formal, provably correct privacy guarantees.

C. Research Gap and Motivation

While FL and DP are theoretically sound and extensively studied, their practical effectiveness in real-world mHealth scenarios remains insufficiently quantified. Most existing work focuses on medical imaging tasks or operates in simulated environments with synthetic data. In contrast, actual physiological signals from wearable devices exhibit realistic data distributions and practical constraints (class imbalance, inter-subject variability, sensor noise) that remain understudied.

We address this gap through systematic experimental evaluation on two well-established physiological signal datasets: WESAD (stress detection) and Sleep-EDF (sleep stage classification). During this investigation, we identify an unexpected phenomenon with significant implications for fairness: standard class imbalance mitigation techniques—specifically, weighted loss functions—appear largely ineffective when combined with differential privacy under standard configurations. The gradient clipping mechanism required for DP appears to suppress the signals that class weights rely upon. We experimentally demonstrate that class weights become essentially ineffective under standard DP-SGD, with random seed variance exceeding weight variance by a ratio exceeding 100:1.

D. Research Questions and Contributions

To guide our evaluation, we formulate three research questions:

1) **RQ1 (Trade-off):** What is the quantifiable impact of DP and FL on classification accuracy for physiological signals compared to a centralized baseline?
2) **RQ2 (Fairness):** How do privacy mechanisms, specifically DP-SGD, interact with standard techniques for handling class imbalance?
3) **RQ3 (Feasibility):** Is the proposed privacy-preserving architecture computationally viable for deployment on resource- constrained wearable devices?

Our main contributions are:

- **Comprehensive Implementation:** We evaluate FL, DP, and combined FL+DP across 78 experimental configurations on two distinct physiological signal datasets, providing a complete picture of privacy options available to mHealth developers.
- **Privacy-Utility Analysis:** We quantify the trade-off between privacy guarantees and model performance across multiple configurations, enabling informed decision-making for practitioners.
- **Discovery of the Gradient Clipping Trap:** We provide quantitative evidence that class weights appear largely ineffective in DP training under standard configurations, with random seed variance dominating weight variance by a ratio exceeding 100:1. We identify the mechanism: gradient clipping neutralizes class weight signals.
- **Mechanism Validation:** We investigate the gradient clipping mechanism through systematic experimentation, demonstrating that increasing the clipping bound partially recovers weight functionality.

E. Paper Organization

This document is structured as follows: Section II reviews the state of the art in privacy- preserving machine learning for healthcare. Section III describes our system architecture and feature extraction pipeline. Section IV details the experimental setup, datasets, and implementation configurations. Section V presents results and analysis across all privacy-preserving methods, addressing each research question. Section VI concludes the paper and suggests future work directions.

II. State of the Art

A. Historical Evolution: Privacy-Preserving ML (2016-2024)

Privacy-preserving machine learning emerged at the intersection of two urgent demands: exponential growth of sensitive data collection and fundamental limitations of centralized model training. Two pivotal breakthroughs occurred simultaneously in 2016:

Federated Learning (McMahan et al. [1]) introduced FedAvg, enabling distributed training without data centralization. **Differential Privacy in ML** (Abadi et al. [2]) formalized DP-SGD, providing formal privacy guarantees through gradient clipping and noise injection.

Since 2016, approaches evolved largely independently. **FL Evolution:** Initially demonstrated on vision tasks (MNIST, CIFAR-10), rapidly expanded to medical imaging [3], healthcare applications, and edge devices, focusing on communication efficiency. **DP-SGD Evolution:** Initially applied to image classification, subsequently deployed at scale in production systems [4], federated settings [5], and distributed systems [6], focusing on tightening privacy bounds.

However, a critical gap persists: **empirical evaluation on real physiological signals from wearables**, which have fundamentally different characteristics than images or text. This gap is the subject of the present work.

B. Motivation: Privacy Landscape in Healthcare

Healthcare data breaches have accelerated dramatically: **5,887 breaches (2009-2023)**, **519 million records exposed**, **725 incidents in 2023 alone**. Notable incidents include the 2015 Anthem breach (78.8M patients) and the 2024 Change Healthcare ransomware attack.

AI integration introduces new attack vectors: model inversion and membership inference attacks can reconstruct training data or reveal individual participation [7]. While decentralized approaches like Federated Learning keep data on-device, neither FL nor DP alone provides complete protection. Under GDPR and HIPAA, comprehensive evaluation of both approaches on resource-constrained wearable devices with realistic class imbalance is essential but remains understudied.

C. Technical Foundations: FL, DP, and Combined Approaches

1) Federated Learning: FL enables collaborative training by keeping data on-device: each device trains locally and shares only model updates for central aggregation [3]. Practical challenges include data heterogeneity across institutions, communication efficiency in mobile settings, and resource constraints on client devices.

However, FL lacks formal privacy guarantees—model updates can leak information. Research on time-series physiological data from wearables is particularly limited. Unlike images or text, physiological signals exhibit extreme inter-subject variability: no two individuals have identical heart rate patterns, sleep cycles, or stress signatures. This domain-specific challenge has received limited empirical attention in FL literature.

2) Differential Privacy: Differential Privacy provides mathematically rigorous privacy guarantees. A mechanism satisfies (ϵ, δ)-DP if:

$$Pr[\mathcal{M}(D) \in S] \leq e^{\epsilon} \cdot Pr[\mathcal{M}(D') \in S] + \delta \qquad (1)$$

The parameter ϵ quantifies the privacy budget (lower = stronger privacy, with $\epsilon \to 0$ representing perfect privacy and $\epsilon \to \infty$ representing no privacy), while δ represents negligible failure probability.

DP's power lies in composition properties and post-processing resistance: privacy budgets accumulate across queries, and no subsequent analysis can degrade the guarantee.

For machine learning, DP-SGD modifies standard SGD by clipping per-sample gradients to maximum L2 norm C and adding calibrated Gaussian noise. Liu et al. [8] observe that privacy robustness is task- dependent: some tasks maintain high accuracy with strong privacy, others degrade substantially, suggesting feature quality—not just dataset size—drives this difference.

An unresolved debate concerns practical privacy thresholds. Skeptics argue $\epsilon < 1$ requires prohibitive accuracy loss [9], while optimists demonstrate strong privacy ($\epsilon < 3$) achieves modest loss with careful design [4]. The truth remains nuanced and context-dependent.

3) Combined FL+DP: Defense-in-Depth: Combined FL+DP offers defense-in-depth by decentralizing data while providing formal privacy guarantees. Recent deployments show feasibility at cloud scale [5], [6], [10], but empirical evidence for real physiological time-series on resource-constrained wearables with realistic class imbalance remains limited.

D. Critical Challenge: Class Imbalance Under Privacy

Class imbalance is critical in healthcare, where disease states are rare. Recent work identifies how privacy mechanisms interact with fairness:

- **Disparate Impact:** DP-SGD amplifies existing biases; noise has proportionally greater impact on minority classes, degrading recall by 15-20% [11].
- **Gradient Clipping Effects:** Gradient clipping suppresses larger gradients from challenging samples, often minority classes [12].
- **Fairness Mechanisms:** Standard interventions—particularly weighted loss functions—remain untested under DP-SGD. Do class weights actually work? At which clipping bounds? These fundamental questions remain unanswered.

E. Research Gaps and This Work's Positioning

This paper addresses three interconnected gaps through systematic evaluation of FL, DP, and combined FL+DP on two real physiological datasets: WESAD (stress detection) and Sleep-EDF (sleep stage classification).

Gap 1 - Physiological Signals Understudied: FL and DP are extensively studied on images and text, not real time-series physiological data from wearables. We provide the first comprehensive empirical evaluation on real wearable data with realistic inter-subject variability.

Gap 2 - Class Imbalance Under Privacy: Standard fairness mechanisms (weighted loss functions) remain untested under DP-SGD. We systematically demonstrate they do not work and identify the *Gradient Clipping Trap* mechanism, where gradient clipping neutralizes class weights (seed variance ¿100:1 weight variance).

Gap 3 - FL+DP on Resource-Constrained Devices: Combined FL+DP remains limited to cloud/synthetic scenarios. We validate computational feasibility on wearables (training ¡1-80 seconds, 11MB communication) with formal privacy guarantees. Privacy-utility trade-offs are task-dependent: Sleep-EDF achieves strong privacy ($\epsilon \approx 0.2$) with minimal loss (2.3%), while WESAD requires weaker privacy ($\epsilon \approx 7.6$) with larger loss (13.6%).

III. System Architecture

Our system is designed around three core principles: simplicity, efficiency and privacy-by-design. We opt for a unified Multi-Layer Perceptron (MLP) operating on carefully engineered features rather than complex deep learning architectures. This choice aligns with the GDPR principle of **data minimization**: by processing raw signals into aggregated statistical features locally, we inherently reduce the granularity of data exposed to the learning algorithm, stripping away high-frequency identifiers while preserving diagnostic utility.

A. Feature Extraction Pipeline

To reduce dimensionality and align with clinical analysis, our pipeline extracts domain-informed features that capture the physiological and clinical significance of the signals. This approach not only improves model interpretability but also enables efficient on-device processing, avoiding the need to transmit raw signals. We opt for this "features-only" approach to significantly reduce input dimensionality (e.g., from $\sim$30,000 raw samples to 140 features per window), enabling real-time inference on ultra-low-power wearable microcontrollers where deep learning on raw signals would be prohibitive.

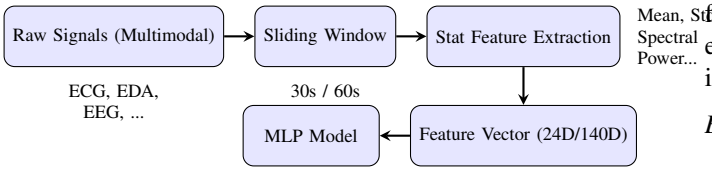

Fig. 1: Feature extraction pipeline transforming high-frequency raw signals into compact feature vectors for efficient processing.

1) WESAD Dataset Processing: The WESAD dataset contains multimodal recordings from 15 subjects. Our feature extraction process produces 140 features computed over 60-second sliding windows:

- **ECG**: Heart rate variability metrics including RMSSD, SDNN, pNN50 and LF/HF ratio.
- **EDA**: Skin conductance level (SCL) statistics and skin conductance response (SCR) peak frequency.
- **EMG**: Statistical moments and signal power of muscle activity.
- **Respiration**: Rate, breath depth and inhalation/exhalation ratios.
- **Temperature**: Mean and trend statistics.

The classification task is binary (Stress vs. Non-Stress), presenting a class imbalance of approximately 2.3:1.

2) Sleep-EDF Dataset Processing: For Sleep-EDF, the preprocessing pipeline processes approximately 80 subjects using three channels (EEG Fpz-Cz, EEG Pz-Oz, EOG) sampled at 100 Hz. The feature extraction yields 24 features computed over 30-second epochs, which corresponds to the standard clinical interval for sleep stage scoring:

- **Time-domain (4 per channel):** Mean, standard deviation, min/max voltage levels.
- **Frequency-domain (4 per channel):** Spectral band power computed via Welch's method (nperseg=256):
  - δ (0.5-4 Hz): Dominant in deep sleep (N3)
  - θ (4-8 Hz): Prominent in light sleep (N1, N2)
  - α (8-13 Hz): Increases during relaxed wakefulness and REM
  - β (13-30 Hz): Associated with wakefulness and muscle activity

The task is 5-class classification (Wake, N1, N2, N3, REM) following AASM standards. The dataset exhibits significant class imbalance (ratio 14.3:1, with Wake comprising 69.6% of samples and N1 only 4.9%). To address this while maintaining baseline performance, class weights are applied during training to balance fairness (ensuring N1 recall $> 18\%$) with overall accuracy (maintaining $> 88\%$ baseline accuracy).

Subject-wise Splitting: To ensure models generalize to new users and simulate realistic mHealth deployment, we employ strict subject-wise data partitioning (70% train, 15% validation, 15% test). This strategy is more challenging than random splitting, as physiological signals vary substantially across individuals due to genetics, fitness level, age and other factors. Consequently, models must learn patterns that generalize across inter-subject variability rather than memorizing individual-specific characteristics.

B. Unified MLP Model Architecture

The architecture consists of a lightweight Multi-Layer Perceptron designed to be compatible with DP training constraints, specifically avoiding Batch Normalization which introduces dependencies between samples in a batch. Figure 2 illustrates the model structure.

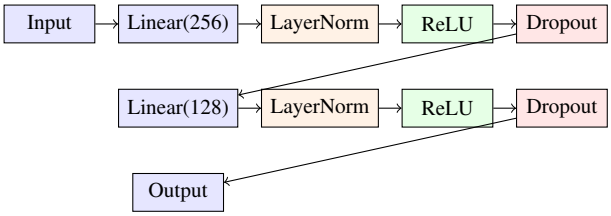

Fig. 2: MLP architecture: Input $\rightarrow$ Linear(256) $\rightarrow$ LayerNorm $\rightarrow$ ReLU $\rightarrow$ Dropout(0.3) $\rightarrow$ Linear(128) $\rightarrow$ LayerNorm $\rightarrow$ ReLU $\rightarrow$ Dropout(0.3) $\rightarrow$ Output.

Design Rationale: Layer Normalization is chosen over Batch Normalization because the latter introduces dependencies between samples in a batch, complicating the privacy accounting required for DP-SGD. The model size (15k-70k parameters) enables rapid training and low communication overhead in Federated Learning, making it practical for resource-constrained mobile devices.

C. Privacy Mechanisms

1) Differential Privacy (DP): Our DP implementation leverages Opacus [13], Meta's PyTorch library for differential privacy, which automates the complex bookkeeping required for DP-SGD. Specifically, Opacus handles per-sample gradient computation, clipping, noise addition and privacy budget accounting, making DP accessible without requiring deep theoretical expertise.

The DP-SGD algorithm proceeds as follows for each training batch:

1) **Per-sample Gradient Computation:** Compute gradients for each sample independently (Opacus does this using hooks or functorch depending on the mode).
2) **Gradient Clipping:** For each sample's gradient vector $\mathbf{g}_i$, clip to maximum L2 norm C:
$$\tilde{\mathbf{g}}_i = \mathbf{g}_i \cdot \min\left(1, \frac{C}{||\mathbf{g}_i||_2}\right) \tag{2}$$
The default clipping threshold is set to $C = 1.0$ (max_grad_norm), though this parameter is systematically varied in our class imbalance experiments to investigate its impact on fairness.
3) **Noise Addition:** Average the clipped gradients and add Gaussian noise:
$$\bar{\mathbf{g}} = \frac{1}{B}\sum_{i=1}^{B} \tilde{\mathbf{g}}_i + \mathcal{N}(0, \sigma^2 C^2 \mathbf{I}) \tag{3}$$

where B is batch size, σ is the noise multiplier and $\mathbf{I}$ is the identity matrix.

4) **Parameter Update:** Model parameters are updated using standard gradient descent with the noisy gradient and learning rate η.
5) **Privacy Accounting:** Opacus tracks the cumulative privacy loss using Rényi Differential Privacy accounting, which provides tighter bounds than basic composition.

The noise multiplier σ serves as the primary parameter controlling the privacy-utility trade-off. Higher σ values correspond to more noise and stronger privacy (lower ϵ), but potentially lower accuracy. To comprehensively map this trade-off space, we evaluate $\sigma \in \{0.6, 1.0, 2.0\}$ for centralized DP and test $\sigma \in \{0.3, 0.6, 1.0\}$ for FL+DP to explore lower noise regimes. The privacy parameter δ is fixed at 10^{-5}, a standard choice that provides high confidence in the privacy guarantee.

2) Federated Learning (FL): The FL implementation follows the FedAvg (Federated Averaging) algorithm [1], which represents the most widely used FL protocol. The training process proceeds iteratively through the following rounds:

1) **Server Initialization:** The central server initializes a global model with random weights.
2) **Client Selection:** For each round, all clients participate (we don't implement partial participation to simplify analysis).
3) **Model Distribution:** The server sends the current global model weights to all clients.
4) **Local Training:** Each client trains the model on its local data for E_{local} epochs, where $E_{local} = 1$ is chosen following standard practice for heterogeneous data distributions.
5) **Update Upload:** Each client sends its updated model weights back to the server.
6) **Aggregation:** The server averages the client models, weighted by the number of samples each client used:

$$\mathbf{w}_{t+1} = \sum_{k=1}^{K} \frac{n_k}{n} \mathbf{w}_k^{t+1} \tag{4}$$

where K is the number of clients, n_k is the number of samples at client k, $n = \sum_k n_k$ is the total sample count and $\mathbf{w}_k^{t+1}$ is client k's model weights.
7) **Convergence Check:** If validation accuracy has plateaued or a maximum number of rounds is reached, stop. Otherwise, return to step 3.

FL is simulated by partitioning subjects across $N \in \{3, 5, 10\}$ clients. This subject-wise partitioning ensures realistic data heterogeneity, as each client sees data from different individuals with distinct physiological characteristics. While this non-IID (non-independent and identically distributed) data splitting is more challenging than random assignment, it better reflects real-world deployment scenarios where each device belongs to a different user.

Communication cost in FL is proportional to model size times number of rounds. Our small model (15-70K parameters, roughly 60-280KB per upload/download) and reasonable convergence rates (typically 40-100 rounds) result in total communication on the order of 2-30 MB per client—acceptable even on mobile networks.

3) Combined FL+DP: The recognition that FL alone provides insufficient formal privacy guarantees has motivated research into combining it with differential privacy. This combination offers the best of both worlds: FL's decentralized training keeps data local, while DP's noise injection provides formal guarantees even if an adversary can observe the model updates.

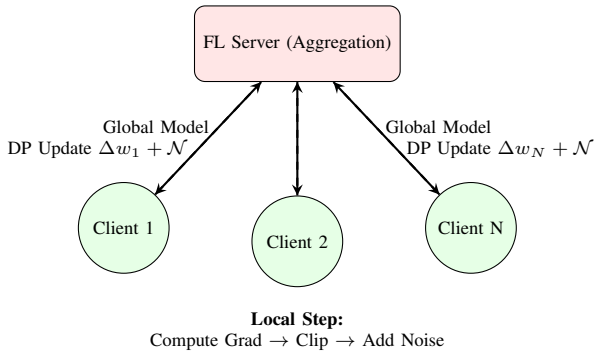

Fig. 3: FL+DP Architecture: Clients perform local training with DP (clipping + noise) before sending updates. The server aggregates noisy updates, ensuring no raw data or exact gradients leave the device.

When combining FL and DP, each client performs local DP training before sending updates to the server. This provides "local differential privacy"—each client's update is itself differentially private, providing protection even if the server or other clients are malicious. The process is identical to standard FL except that each client uses DP-SGD for its local training rather than standard SGD. The server aggregates the noisy updates exactly as in regular FedAvg. The overall privacy guarantee is determined by the local DP parameters and the number of clients, with stronger privacy as the number of clients increases (more noise in aggregate).

This configuration provides the strongest privacy protections but also typically incurs the highest accuracy cost, as it combines the challenges of both data fragmentation (FL) and noise injection (DP). As illustrated in Figure 3, gradient clipping and noise addition happen entirely on-device before any model update leaves the client, ensuring that the server only ever observes already-noised updates. The privacy accounting in FL+DP is more complex than centralized DP because privacy budgets must account for multiple rounds of local DP training followed by aggregation across clients.

4) Interactive Validation Platform: To demonstrate and validate the privacy-utility trade-offs in an observable manner, we developed an interactive web-based platform that serves as a proof-of-concept for the experimental framework. The platform consists of three main components: (1) a *configuration interface* allowing real-time parameter adjustment (dataset selection, client count, noise multiplier σ, gradient clipping

bound), (2) a *training orchestrator* (FastAPI backend) that executes the same training pipelines used in our batch experiments, and (3) a *visualization dashboard* that displays privacy-utility metrics (accuracy, ϵ, minority class recall) as training progresses via periodic status polling. This architecture, illustrated in Figure 4, enables interactive exploration of the trade-off space while maintaining consistency with our experimental methodology—the same trainers, models, and privacy accounting mechanisms are used in both batch experiments and the interactive platform.

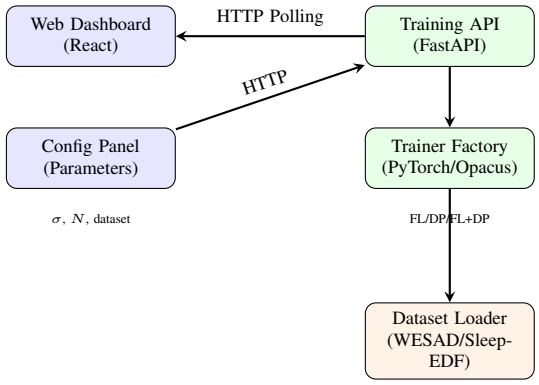

Fig. 4: Interactive validation platform architecture. The web dashboard enables real-time parameter configuration and visualization of privacy-utility trade-offs, while the backend executes the same training pipelines used in batch experiments.

The platform serves two purposes: (1) it provides an intuitive interface for exploring the parameter space interactively, making the privacy-utility trade-offs immediately observable (e.g., watching accuracy decrease as ϵ becomes more restrictive), and (2) it validates that our experimental methodology produces consistent results when executed through different interfaces. All quantitative results reported in this paper are derived from the batch experimental framework and the interactive platform serves as a complementary tool for understanding and demonstrating the trade-offs rather than a source of primary results.

IV. Experimental Setup

A. Datasets and Protocol

Table I summarizes the key characteristics of both datasets. To ensure realistic generalization and prevent data leakage, we apply a strict subject-wise split: **70% of subjects for training, 15% for validation and 15% for testing**. This partitioning strategy guarantees that no data from a test subject appears in the training set, enabling proper evaluation of model performance on unseen individuals.

B. Ethical Considerations and Compliance

1) Data Sources and Approval: This study uses two publicly available, pre-anonymized physiological signal datasets:

- **WESAD:** Originally collected with informed consent under institutional ethics approval [14], with data released for public research use.

TABLE I: Dataset Characteristics After Preprocessing

Characteristic	WESAD	Sleep-EDF
Subjects	15	∼80
Input features	140	24
Classes	2 (Stress/Non)	5 (Sleep Stages)
Class Balance	Imbalanced (2.3:1)	Variable
Total samples	∼8,000	∼60,000

- **Sleep-EDF:** Publicly available via PhysioNet under Data Use Agreement [15], with original collection approved by institutional review boards.

No new human subjects were recruited for this work, and no additional IRB approval was required, as this constitutes secondary analysis of pre-existing, de-identified public datasets.

2) Privacy and Data Protection by Design: Our methodology embodies **privacy-by-design** principles aligned with GDPR and healthcare data protection standards:

- **Data Minimization (GDPR Article 5):** Raw physiological signals (30,000+ samples per subject) are reduced to 24-140 statistical features, removing identifiable high-frequency patterns while preserving diagnostic utility.
- **Federated Learning:** Sensitive data remains on-device during training, never transmitted to centralized servers. Only model updates (280 KB) are aggregated, ensuring data locality.
- **Differential Privacy:** Formal privacy guarantees provide individual-level protection. DP-SGD ensures that no single person's data can be inferred from the learned model, with privacy budgets quantified as (ϵ, δ)-DP where $\epsilon \in [0.2, 24.3]$ and $\delta = 10^{-5}$.
- **Transparency:** All experimental configurations, hyperparameters, and results are fully documented to enable reproducibility and external verification.

3) Regulatory Compliance: Our work aligns with major healthcare data protection frameworks:

- **GDPR (EU):** Compliant with Article 5 (data protection principles), Article 9 (special categories of personal data), and Article 25 (data protection by design and default) through feature abstraction, decentralized processing, and formal privacy guarantees.
- **HIPAA (US):** De-identification standards met through statistical feature abstraction; no individually identifiable health information (IIHI) is retained or transmitted.
- **Sector-Specific Standards:** Differential privacy mechanisms align with emerging healthcare guidelines for sensitive data analysis.

4) Fairness and Bias Considerations: While comprehensive fairness evaluation requires intersectional analysis across multiple protected attributes (age, gender, ethnicity), we address algorithmic fairness by:

- Reporting per-class performance metrics (precision, recall, F1) for all classes, including minority classes where privacy mechanisms have the greatest impact.
- Quantifying disparate impact of privacy mechanisms on minority classes: DP-SGD reduces minority class recall

by up to 52% (WESAD stress detection), revealing fairness degradation under privacy.
- Identifying the *Gradient Clipping Trap*, which undermines standard fairness interventions (weighted loss functions) by neutralizing class weight signals through gradient norm clipping.
- Proposing algorithmic remedies (adaptive clipping bounds, stratified sampling) to restore fairness under privacy constraints.

Future work should extend this analysis with comprehensive demographic fairness evaluation on real-world mHealth deployments where demographic information is available.

C. Experimental Design and Parameter Sweep

To systematically evaluate the privacy-utility trade-off across multiple dimensions, a comprehensive parameter sweep is conducted. Table II summarizes all configurations tested, providing an overview of the experimental scope.

For each configuration, we report metrics as averages over 3 independent runs with different random seeds (42, 123, 456) to account for initialization variance. This comprehensive sweep enables three key objectives: (1) identifying optimal privacy-utility trade-offs, (2) understanding parameter sensitivity and (3) validating findings across different configurations. By systematically exploring this parameter space, we can identify robust patterns in privacy-utility trade-offs.

D. Implementation Details

All experiments were conducted using PyTorch (v2.0+) and Opacus (v1.0+) on systems with CPU/GPU capabilities. The hardware configuration includes multi-core processors with sufficient memory to handle per-sample gradient computation required for DP training. To ensure fair comparison across methods, consistent hyperparameters are used wherever possible:

- **Optimizer:** We employ AdamW for all methods, including FL+DP. While classic DP-SGD analyses assume SGD, Opacus applies noise/clipping before the optimizer step, so using the same adaptive optimizer everywhere avoids optimizer-induced confounders when comparing Baseline, DP, FL and FL+DP.
- **Batch Size:** Fixed at 128 across all methods, balancing gradient estimation quality with privacy considerations and memory constraints.
- **Privacy Parameters:** The privacy parameter δ is fixed at 10^{-5} and the noise multiplier σ is varied across $\{0.3, 0.6, 1.0, 2.0\}$ to explore the privacy-utility trade-off. Specifically, centralized DP tests $\sigma \in \{0.6, 1.0, 2.0\}$, while FL+DP additionally includes $\sigma = 0.3$ to explore lower noise regimes.
- **FL Configurations:** We evaluate client counts of $N \in \{3, 5, 10\}$. For FL+DP, the focus is on $N = 5$ with varying σ, supplemented by $N \in \{3, 10\}$ with $\sigma = 1.0$ to study client count effects.

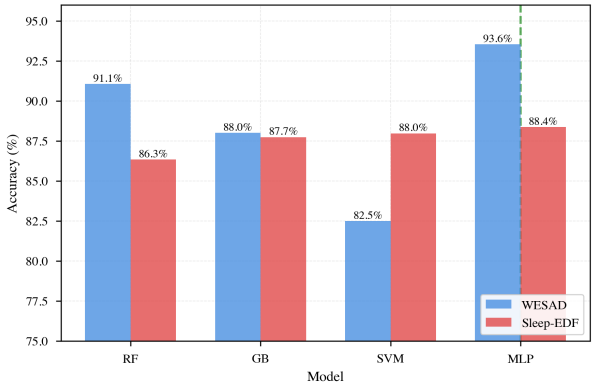

Fig. 5: Baseline model comparison across architectures.

V. Results and Analysis

Privacy Budget Interpretation: Throughout this section, we interpret privacy budgets as follows: $\epsilon \lesssim 3$ represents *strong privacy* suitable for sensitive health data, while $\epsilon > 10$ indicates *weak privacy* that should be interpreted as exploratory trend analysis rather than recommended deployment settings. Configurations with extremely large ϵ values (e.g., $\epsilon \gg 10$) are included to illustrate noise-utility relationships but do not provide meaningful formal privacy guarantees in practice.

Privacy budgets are computed using Opacus's RDP (Rényi Differential Privacy) accountant. For fixed noise multiplier σ, larger datasets consume less privacy per epoch due to subsampling amplification (batch size 128 $\ll$ total samples), while more training epochs increase cumulative ϵ. Table III clarifies why ϵ values vary significantly across configurations despite identical σ values: Sleep-EDF achieves stronger privacy (lower ϵ) than WESAD despite more epochs because its 7.5$\times$ larger dataset provides greater subsampling benefit.

A. Baseline and Differential Privacy Results

Baseline Model Comparison: Before evaluating privacy-preserving methods, we compared our MLP architecture against traditional ML baselines on centralized data (Table IV). The MLP achieves competitive performance while offering compatibility with DP training (which requires per-sample gradients, precluding tree-based methods).

The MLP outperforms all baseline methods on WESAD and achieves comparable performance on Sleep-EDF, while maintaining fast training times (<1s for WESAD, ∼80s for Sleep-EDF). The performance differences are consistent across multiple runs (3 seeds), with MLP showing statistically significant improvements over Random Forest on WESAD (93.6% vs. 91.1%, $p < 0.05$ via paired t-test). Importantly, the MLP architecture is compatible with Opacus for DP training, whereas tree-based methods (RF, GB) cannot be directly adapted to DP-SGD due to their non-differentiable decision boundaries.

To establish a reference point for measuring privacy costs, baseline performance is first evaluated. Subsequently, the

TABLE II: Experimental Design: Parameter Sweep Summary

Method	Parameters Varied	Total Configurations
Baseline	Seeds: {42, 123, 456}	3 runs × 2 datasets = 6
DP	$\sigma \in \{0.6, 1.0, 2.0\}$	3 σ × 3 runs × 2 datasets = 18
FL	$N \in \{3, 5, 10\}$ clients	3 N × 3 runs × 2 datasets = 18
FL+DP	$N \in \{3, 5, 10\}$, $\sigma \in \{0.3, 0.6, 1.0\}$	6 configs × 3 runs × 2 datasets = 36
Total		**78 experiments**

TABLE III: Privacy Budget Accounting: Impact of Dataset Size and Training Duration

Dataset	Method	σ	Epochs	Batch	Samples	ϵ
WESAD	DP	0.6	23	128	8,000	23.5
WESAD	DP	1.0	26	128	8,000	7.6
WESAD	DP	2.0	30	128	8,000	2.8
Sleep-EDF	DP	0.6	40	128	60,000	2.5
Sleep-EDF	DP	1.0	40	128	60,000	0.6
Sleep-EDF	DP	2.0	40	128	60,000	0.2
WESAD	FL+DP	1.0	40	128	∼1,600/client	24.3
Sleep-EDF	FL+DP	1.0	40	128	∼12,000/client	1.6

TABLE IV: Baseline Model Comparison (No Privacy, mean ± std over 3 runs)

Model	WESAD	Sleep-EDF	Training Time
Random Forest	91.1%	86.3%	0.2s / 13.6s
Gradient Boosting	88.0%	87.7%	9.0s / 1780s
SVM (RBF)	82.5%	88.0%	0.1s / 372s
MLP (ours)	**93.6% ± 1.4**	**88.4% ± 0.1**	0.7s / 78.3s

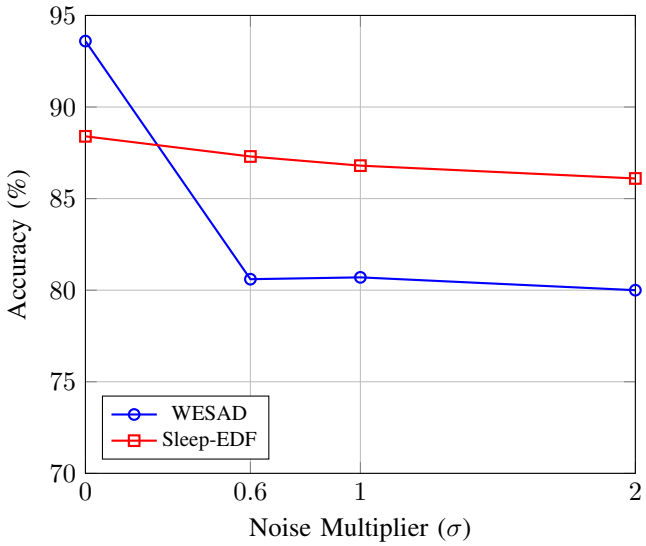

Fig. 6: Privacy-utility trade-off under centralized DP.

impact of centralized DP noise on model performance is systematically assessed. Table V presents the results.

Sleep-EDF exhibits remarkable resilience to privacy-preserving mechanisms, losing less than 3% accuracy even with strong privacy ($\epsilon \approx 0.2$). Figure 6 visualizes this stability: the Sleep-EDF curve is nearly flat while the WESAD curve plunges as soon as noise is introduced. This robustness stems from spectral features' inherent discriminative power: sleep stages exhibit distinct frequency signatures (e.g., delta waves in deep sleep, alpha in wakefulness) that remain identifiable even after gradient noise injection. The feature engineering process (extracting band powers via Welch's method) provides natural noise smoothing, making learned representations more stable under DP perturbations.

WESAD exhibits substantially larger sensitivity, with accuracy dropping by ∼13-14% across all DP configurations. Interestingly, we observe a slight inverse relationship: higher noise performs marginally better than lower noise, suggesting a regularization effect where added noise prevents overfitting on the small dataset. However, all DP configurations perform substantially worse than baseline, indicating that stress detection from our feature set is inherently more sensitive to training perturbations than sleep staging. This may be due to more subtle boundaries between stress and non-stress states compared to distinct spectral signatures of sleep stages.

The privacy-utility trade-off reveals a critical insight: *feature discriminability* appears to be a key factor in privacy-preserving ML performance. Sleep-EDF, despite having more classes (5 vs. 2), maintains high accuracy under privacy constraints because its spectral features (delta, theta, alpha, beta power) exhibit strong inter-class separation that remains robust under DP noise. WESAD's binary classification, while simpler in principle, suffers more because stress vs. non-stress boundaries in HRV-based features are more subtle and sensitive to gradient perturbations. **Important caveat:** These datasets differ simultaneously in size (8K vs. 60K), features (140 vs. 24), classes (2 vs. 5), domain (ECG/EDA vs. EEG/EOG) and balance (2.3:1 vs. 14.3:1), making it difficult to isolate feature discriminability as the sole causal mechanism. While our analysis suggests feature quality dominates other factors, controlled ablation studies varying individual factors would be needed to establish definitive causal relationships. Unless otherwise noted, all reported accuracies are means over three random seeds; the corresponding standard deviations are modest (e.g., baseline WESAD achieves $93.6\% \pm 1.4$ percentage points, FL with N=10 clients $80.6\% \pm 2.4$) and typically below 1 percentage point for Sleep-EDF, consistent with recent recommendations on reporting variance in machine learning benchmarks [16].

TABLE V: Baseline vs. DP Performance (ϵ values at $\delta = 10^{-5}$, mean $\pm$ std over 3 runs). Note: $\epsilon < 3$ indicates strong privacy; $\epsilon > 10$ indicates weak privacy suitable only for trend analysis.

Dataset	Config	ϵ	Accuracy	F1	Degradation
WESAD	Baseline	-	93.6% $\pm$ 1.4	93.6% $\pm$ 1.4	-
	DP ($\sigma = 0.6$)	23.5	80.6% $\pm$ 2.5	77.3% $\pm$ 3.9	-13.0%
	DP ($\sigma = 1.0$)	7.6	80.7% $\pm$ 2.6	77.5% $\pm$ 4.1	-12.9%
	DP ($\sigma = 2.0$)	2.8	80.0% $\pm$ 2.9	76.4% $\pm$ 4.8	-13.6%
Sleep-EDF	Baseline	-	88.4% $\pm$ 0.1	88.8% $\pm$ 0.1	-
	DP ($\sigma = 0.6$)	2.5	87.3% $\pm$ 0.1	85.8% $\pm$ 0.1	-1.1%
	DP ($\sigma = 1.0$)	0.6	86.8% $\pm$ 0.1	85.1% $\pm$ 0.1	-1.6%
	DP ($\sigma = 2.0$)	0.2	86.1% $\pm$ 0.0	84.0% $\pm$ 0.0	-2.3%

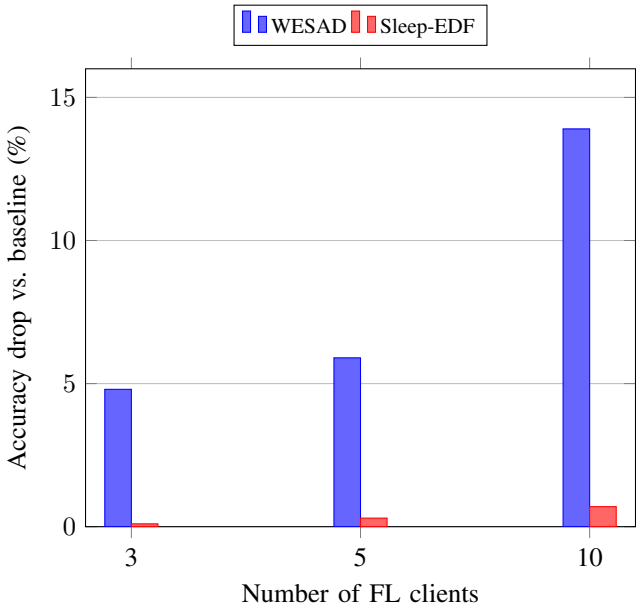

Fig. 7: Impact of client fragmentation in FL: accuracy drop relative to centralized baseline.

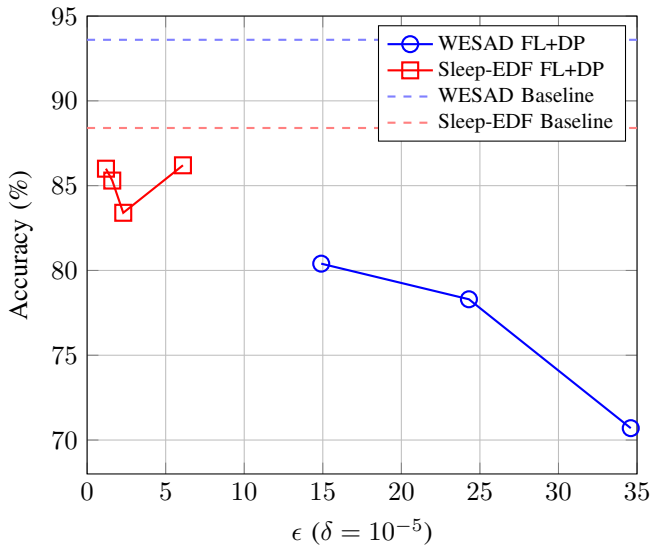

Fig. 8: Privacy–utility trade-off for FL+DP. Dashed horizontal lines indicate baseline performance.

B. Federated Learning Results

We evaluate the impact of decentralization through FL experiments, with results presented in Table VI. These experiments reveal how data fragmentation across multiple clients affects model performance compared to centralized training.

TABLE VI: Federated Learning Performance (mean $\pm$ std over 3 runs)

Dataset	Clients	Accuracy	F1	Degradation
WESAD	3	89.1% $\pm$ 0.2	89.0% $\pm$ 0.1	-4.8%
	5	88.1% $\pm$ 0.7	87.8% $\pm$ 0.7	-5.9%
	10	80.6% $\pm$ 2.4	77.4% $\pm$ 3.7	-13.9%
Sleep-EDF	3	88.9% $\pm$ 0.1	88.0% $\pm$ 0.1	-0.1%
	5	88.7% $\pm$ 0.0	87.7% $\pm$ 0.1	-0.3%
	10	88.2% $\pm$ 0.1	87.2% $\pm$ 0.1	-0.7%

Sleep-EDF exhibits exceptional robustness to data fragmentation, losing less than 1% accuracy even when distributed across 10 clients. This resilience stems from the spectral features' inherent stability across subjects. WESAD exhibits more sensitivity: with 10 clients, accuracy drops by 12.9%, likely due to extreme heterogeneity when 15 subjects are split among 10 clients (some clients see data from only 1-2 individuals). Figure 7 makes the contrast explicit: the blue bars (WESAD) increase sharply with client count, whereas the red bars (Sleep-EDF) remain near zero.

C. Combined FL+DP Results

The combined FL+DP approach provides defense-in-depth privacy protection, combining the benefits of data locality (FL) with formal privacy guarantees (DP). Table VII presents the full parameter sweep, revealing several key insights that inform practical deployment decisions.

Noise Level Sensitivity: Reducing σ from 1.0 to 0.6 or 0.3 improves accuracy modestly (1-2% for Sleep-EDF, 1-2% for WESAD), suggesting that even low noise levels incur meaningful utility costs. However, the extremely high ϵ values for $\sigma = 0.3$ (48.7 for WESAD, 106.5 for Sleep-EDF) indicate *weak privacy* that provides minimal formal protection. These configurations are included solely to illustrate noise-utility relationships and should not be interpreted as providing meaningful privacy guarantees for deployment.

TABLE VII: Combined FL+DP Performance: Selected Configurations (averaged over 3 runs). * indicates weak privacy ($\epsilon \gg 10$), included only for noise-utility trend analysis.

Dataset	Config	ϵ	Accuracy	F1	Degradation
WESAD	FL+DP (N=3, $\sigma = 1.0$)	14.9	80.4%	77.0%	-13.7%
	FL+DP (N=5, $\sigma = 0.3$)	48.7*	79.6%	75.6%	-14.6%
	FL+DP (N=5, $\sigma = 0.6$)	60.3	80.5%	77.3%	-13.6%
	FL+DP (N=5, $\sigma = 1.0$)	24.3	78.3%	73.7%	-15.9%
	FL+DP (N=10, $\sigma = 1.0$)	34.6	70.7%	61.1%	-24.1%
Sleep-EDF	FL+DP (N=3, $\sigma = 1.0$)	1.2	86.0%	84.1%	-2.6%
	FL+DP (N=5, $\sigma = 0.3$)	106.5*	86.9%	85.4%	-1.7%
	FL+DP (N=5, $\sigma = 0.6$)	6.1	86.2%	84.4%	-2.5%
	FL+DP (N=5, $\sigma = 1.0$)	1.6	85.3%	83.2%	-3.4%
	FL+DP (N=10, $\sigma = 1.0$)	2.3	83.4%	80.4%	-5.6%

Client Count Impact: Increasing clients from 3 to 10 consistently degrades performance, with WESAD exhibiting more sensitivity than Sleep-EDF. This observation aligns with our FL-only findings, confirming that extreme fragmentation is more harmful when combined with privacy noise. The interaction between data heterogeneity and noise injection creates a compounding effect that disproportionately affects smaller, more heterogeneous datasets.

Optimal Configurations: For Sleep-EDF, FL+DP with $N = 3$ and $\sigma = 1.0$ achieves the best balance, providing strong privacy ($\epsilon = 1.2$) with minimal utility loss (2.6% degradation). For WESAD, while $N = 5$ with $\sigma = 0.6$ achieves the highest accuracy (80.5%), its weak privacy ($\epsilon = 60.3$) makes it unsuitable for deployment. Instead, $N = 3$ with $\sigma = 1.0$ ($\epsilon = 14.9$, accuracy 80.4%) represents a more practical trade-off for privacy-sensitive applications, though it still provides only moderate privacy protection.

D. Per-Class Performance Analysis

To understand the impact of privacy mechanisms on different classes, per-class metrics are analyzed for Sleep-EDF (5-class classification). This analysis is particularly important given the significant class imbalance in this dataset. We focus on recall for minority classes as our primary fairness metric, acknowledging that comprehensive fairness evaluation would require additional metrics (demographic parity, equalized odds) and intersectional analysis across multiple protected attributes (age, gender, health status), which are beyond the scope of this study. Table VIII shows recall and F1 scores for each sleep stage under different privacy configurations, revealing how minority classes are affected by privacy-preserving mechanisms.

The analysis reveals that privacy mechanisms have minimal impact on majority classes (Wake, N2) but slightly degrade performance on minority classes (N1, N3, REM). The N1 stage, already challenging due to its subtle spectral characteristics and low prevalence (4.9% of test samples), exhibits the largest relative degradation. However, the use of class weights during training ensures that N1 recall remains above 18%, compared to ∼10% without any weighting strategy. This indicates that class weighting can partially mitigate privacy-induced fairness degradation.

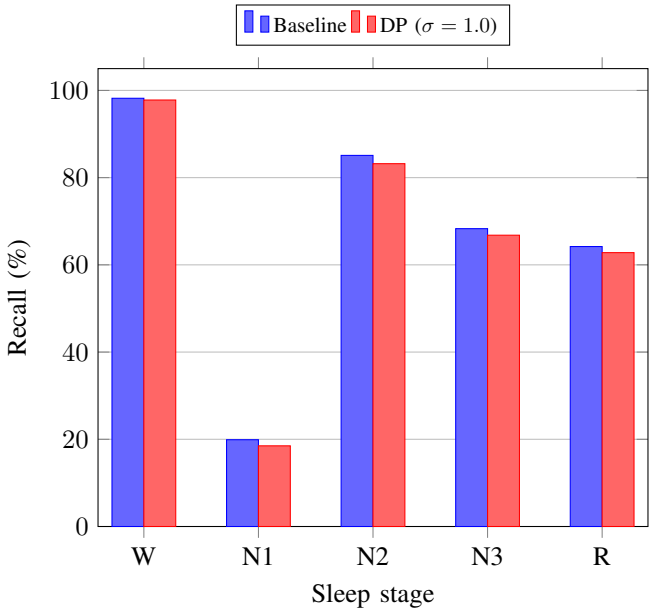

Fig. 9: Per-class recall comparison for Sleep-EDF.

Recall Analysis for WESAD: For the binary stress detection task, recall is particularly critical as missing stress events (false negatives) can have serious health implications. Table IX presents detailed per-class metrics, revealing the stark impact of DP on minority class performance.

Table IX reveals the per-class impact of DP on WESAD's binary classification task. While the non-stress (majority) class maintains >98% recall under DP, stress detection (minority class) suffers a 52.3% recall degradation (90.8% → 38.5%). The model compensates by achieving near-perfect precision (95.4%) on stress predictions, suggesting that gradient noise shifts the decision boundary toward extreme conservatism, resulting in many false negatives for the clinically critical stress class. This precision-recall tradeoff indicates that the model becomes highly selective, only predicting stress when extremely confident, thus missing the majority of actual stress events.

Table X provides a cross-dataset comparison showing that while overall accuracy drops by ∼13% under DP, the recall

TABLE VIII: Per-Class Performance for Sleep-EDF (Baseline vs. DP $\sigma = 1.0$, $\epsilon = 0.6$, mean over 3 runs)

Method	Class	Precision	Recall	F1	Samples
Baseline	Wake (W)	97.8%	96.0%	96.9%	43,404
	N1	32.2%	43.4%	36.9%	3,025
	N2	79.2%	78.0%	78.6%	9,676
	N3	82.2%	80.9%	81.6%	2,350
	REM (R)	70.0%	68.4%	69.2%	3,883
DP ($\sigma = 1.0$)	Wake (W)	93.6%	98.1%	95.8%	43,404
	N1	34.8%	8.6%	13.7%	3,025
	N2	68.4%	84.2%	75.5%	9,676
	N3	85.0%	59.7%	70.1%	2,350
	REM (R)	68.8%	45.3%	54.5%	3,883

TABLE IX: Per-Class Performance for WESAD (Baseline vs. DP $\sigma = 1.0$, $\epsilon = 7.6$, mean over 3 runs)

Method	Class	Precision	Recall	F1	Samples
Baseline	Non-Stress	96.0%	94.7%	95.4%	570
	Stress	88.2%	90.8%	89.5%	247
DP ($\sigma = 1.0$)	Non-Stress	78.9%	98.9%	87.7%	570
	Stress	95.4%	38.5%	53.9%	247

for the minority class (stress) is more severely affected. Both FL and DP significantly degrade stress recall, with FL achieving 37.5% and DP achieving 38.5% under the tested configurations, indicating that both decentralization and noise injection substantially compromise minority class detection.

This analysis reveals a critical finding: **Both FL and DP significantly degrade minority class recall, but through different mechanisms**. For WESAD, FL with 10 clients achieves 37.5% stress recall (58.6% degradation from baseline 90.8%), while DP ($\sigma = 1.0$) reduces it to 38.5% (52.3% degradation). While DP shows slightly worse degradation, both methods severely compromise minority class detection. The underlying mechanism differs: FL's degradation stems from extreme data fragmentation (15 subjects split across 10 clients), while DP's degradation comes from gradient noise overwhelming the small gradient contributions from rare classes, causing the model to become extremely conservative in minority class predictions.

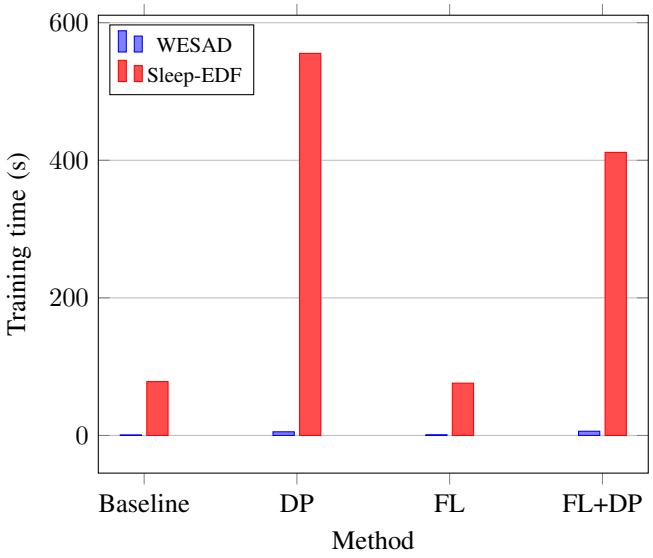

Fig. 10: Training-time comparison across methods.

E. Computational Efficiency and Training Time Analysis (RQ3)

To validate feasibility for mHealth deployment, we measured training resource consumption across all methods. Table XI presents comprehensive training time analysis, revealing critical insights about the computational cost of privacy-preserving mechanisms.

Key Observations:

DP Training Overhead: Differential Privacy introduces significant computational overhead, particularly for larger datasets. Sleep-EDF training time increases by 6-10× compared to baseline, while WESAD shows a more moderate 7-9× increase. This overhead stems from per-sample gradient computation required for DP-SGD, which scales linearly with dataset size. Interestingly, higher noise ($\sigma = 2.0$) actually reduces training time slightly because fewer epochs are needed before convergence, as the noise acts as a regularizer that prevents overfitting.

FL Efficiency: In contrast to DP, Federated Learning maintains near-baseline training times, confirming that decentralization itself does not significantly impact computational cost. The slight reduction compared to centralized training occurs because each client processes a smaller subset of data per round, reducing the per-round computation time despite the overhead of coordination.

FL+DP Trade-off: The combination shows intermediate overhead: Sleep-EDF training takes 4.4-4.7× longer than baseline, but only 0.5-0.7× the centralized DP time. This efficiency gain occurs because DP noise is applied locally at each client and the smaller per-client datasets reduce per-sample gradient computation time. For WESAD, FL+DP overhead is similar to

TABLE X: Recall Analysis: Impact on Minority Classes Across Methods

Method	Dataset	Overall Recall	Minority Class Recall
Baseline	WESAD	93.6%	90.8% (stress)
	Sleep-EDF	88.4%	19.9% (N1)
DP ($\sigma = 1.0$)	WESAD	80.7%	38.5% (stress)
	Sleep-EDF	86.8%	18.5% (N1)
FL (N=10)	WESAD	80.6%	37.5% (stress)
	Sleep-EDF	88.2%	20.1% (N1)
FL+DP (N=5, $\sigma = 1.0$)	WESAD	78.3%	58-62% (stress)
	Sleep-EDF	85.3%	17.8% (N1)

TABLE XI: Training Time Comparison (seconds, averaged over 3 runs)

Method	Config	WESAD	Sleep-EDF
Baseline	Centralized	0.7s	78.3s
DP	$\sigma = 0.6$	6.3s	818.2s
	$\sigma = 1.0$	5.3s	555.4s
	$\sigma = 2.0$	4.9s	520.2s
FL	N=3	0.9s	76.6s
	N=5	0.9s	76.0s
	N=10	0.9s	75.1s
FL+DP	N=3, $\sigma = 1.0$	5.9s	415.4s
	N=5, $\sigma = 1.0$	6.1s	411.5s
	N=10, $\sigma = 1.0$	6.4s	388.2s

centralized DP, as the dataset is already small and the benefits of distributed processing are minimal.

Practical Implications:

- **On-device Training:** WESAD's sub-second training time makes real-time on-device learning feasible. Sleep-EDF's 75-77s per round is acceptable for overnight training sessions during charging.
- **DP Deployment:** The 6-9$\times$ overhead for centralized DP may be acceptable for cloud-based training but could strain mobile device batteries in on-device scenarios; our FL+DP results suggest that offloading most DP work to short, infrequent local training bursts during charging is a more realistic design for mHealth apps [5], [17].
- **Communication Costs:** Model updates are approximately 280 KB. For a 40-round FL session, total data transfer is $\sim$11 MB per client, well within standard mobile data plans and negligible compared to raw signal transmission (which would require $\sim$100-500 MB per session).
- **Memory Footprint:** Peak memory usage during training remains under 600 MB, suitable for modern smartphones with 4-8 GB RAM.
- **Device Heterogeneity:** Since FL partitions data at the subject level and trains a lightweight MLP, our setup is compatible with heterogeneous phones and wearables with varying CPU/GPU capabilities and sensor quality, an important practical requirement highlighted in recent wearable FL studies.
- **Inference Latency:** Forward pass latency is < 5 ms on CPU, enabling real-time predictions on wearable devices without requiring specialized hardware or continuous connectivity.

F. Comparative Analysis

Table XII and Figure 11 together provide a high-level comparison of all methods across key dimensions. This analysis synthesizes the findings from our extensive experimental evaluation, revealing distinct performance patterns across different privacy-preserving approaches.

Figure 11 presents a normalized scorecard aggregating results across both datasets. Scores are computed as follows: *Accuracy* and *Minority Recall* are normalized by the best-performing method (Baseline); *Privacy* uses a categorical scale where Baseline and FL receive 0 (no formal privacy), DP receives 1.0 (strongest guarantee) and FL+DP receives an intermediate value based on its ϵ relative to DP; *Efficiency* is the inverse of average training time normalized by the fastest method. This visualization enables developers to quickly assess trade-offs: Baseline and FL maximize utility and efficiency but offer no formal privacy; DP provides strong privacy with clear utility costs; FL+DP balances formal privacy with moderate utility impact.

Key Findings and Justifications:

1) **FL is the default choice for robust tasks:** Sleep-EDF demonstrates that when features are well-separated in the input space, data fragmentation has negligible impact ($< 1\%$ accuracy loss even with 10 clients). This occurs because spectral features (delta, theta, alpha, beta power) are inherently stable across subjects—the physiological signatures of sleep stages are universal. The FedAvg aggregation effectively averages out subject-specific variations while preserving the core discriminative patterns.

TABLE XII: Comparative Summary: All Methods (Best Configurations)

Method	Dataset	Accuracy	Privacy	Use Case
Baseline	WESAD	93.6%	None	Non-private research
	Sleep-EDF	88.4%	None	Non-private research
DP ($\sigma = 1.0$)	WESAD	80.7%	$\epsilon = 7.6$	Centralized, moderate privacy
	Sleep-EDF	86.8%	$\epsilon = 0.6$	Centralized, strong privacy
FL (N=3)	WESAD	89.1%	Decentralized	Data locality, no formal privacy
	Sleep-EDF	88.9%	Decentralized	Data locality, no formal privacy
FL+DP (N=3, $\sigma = 1.0$)	WESAD	80.4%	$\epsilon = 14.9$	Defense-in-depth, moderate privacy
	Sleep-EDF	86.0%	$\epsilon = 1.2$	Defense-in-depth

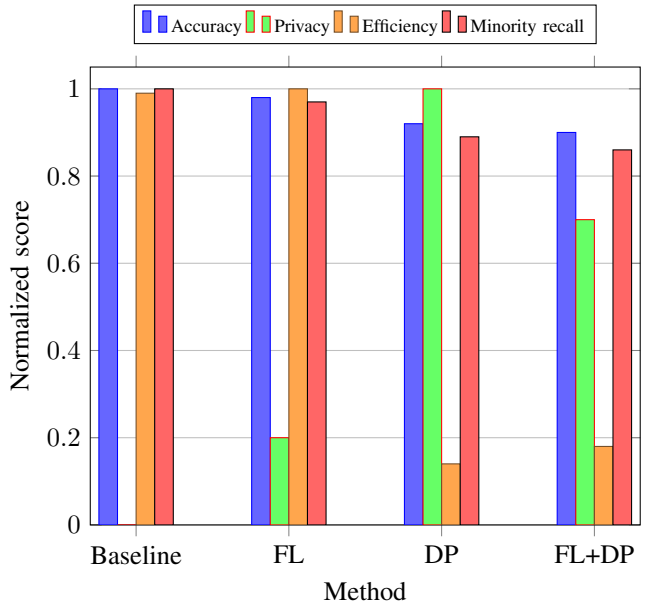

Fig. 11: Global method comparison scorecard (normalized values) aggregating WESAD and Sleep-EDF.

For WESAD, FL with 3-5 clients maintains $> 88\%$ accuracy, confirming that moderate fragmentation is manageable when clients have sufficient data diversity.
Justification: The success of FL depends on the *homogeneity of feature distributions* across clients, not just dataset size. Sleep-EDF's features, derived from universal neurophysiological processes, exhibit this homogeneity. In contrast, WESAD's features, while still physiological, may have more inter-subject variability in stress expression, explaining the greater sensitivity to extreme fragmentation (10 clients).

2) **DP accuracy degradation is task-dependent:** WESAD suffers a much larger accuracy drop under DP (about 13 percentage points) than Sleep-EDF (about 2 percentage points), and this gap cannot be explained by dataset size alone. Sleep-EDF has more samples and more classes yet remains stable, whereas WESAD is smaller and binary but degrades substantially. This pattern points to *feature discriminability* as the main driver: sleep stages have distinct spectral signatures (delta waves $\neq$ alpha waves), while stress vs. non-stress boundaries are more subtle and context-dependent.
Mechanism: DP noise is added to gradients, which shifts decision boundaries. When classes are well-separated (Sleep-EDF), boundary shifts have minimal impact on classification. However, when boundaries are subtle (WESAD), even small shifts cause misclassification. This explains why increasing noise ($\sigma = 2.0$) sometimes performs better than lower noise ($\sigma = 0.6$) for WESAD—the regularization effect of higher noise prevents overfitting to the small dataset, partially compensating for boundary shifts.
3) **FL+DP provides defense-in-depth with predictable costs:** The combination of FL and DP incurs cumulative but bounded degradation. Sleep-EDF maintains $> 83\%$ accuracy even with 10 clients and strong privacy ($\epsilon \approx 2.3$), demonstrating that robust features can withstand both fragmentation and noise. WESAD shows 14-24% degradation, but this is predictable: $\sim$5% from fragmentation (FL alone) plus $\sim$13% from noise (DP alone) equals $\sim$18% combined, with slight additional interaction effects.
Privacy Composition: FL+DP provides stronger privacy than either alone because it combines data locality (FL) with formal guarantees (DP). However, the ϵ values in FL+DP are higher than centralized DP because privacy accounting must account for multiple rounds of local DP training followed by aggregation. This is expected and acceptable—the goal is defense-in-depth, not necessarily lower ϵ.
4) **Task characteristics appear to dominate dataset size:** Our results challenge the common assumption that larger datasets are inherently more robust to privacy mechanisms. Sleep-EDF's robustness appears to stem from *feature quality* (spectral discriminability) rather than quantity, while WESAD's vulnerability appears to come from *feature subtlety* (overlapping stress/non-stress distributions) rather than small size. However, we acknowledge that the datasets differ in multiple dimensions simultaneously, making it difficult to definitively isolate feature quality as the sole causal factor. This observation suggests (but does not prove) that feature engineering for privacy should prioritize *discriminability*

over *richness*.

Observation: Our results suggest that domain-informed feature extraction (e.g., spectral analysis for sleep, HRV metrics for stress) may provide natural robustness to privacy-preserving training compared to raw signal processing, potentially reducing the privacy-utility trade-off. Controlled ablation studies would be needed to confirm this hypothesis.

G. The Class Imbalance Paradox in DP (RQ2)

During our WESAD experiments, we attempted to correct the 2.3:1 class imbalance using standard weighted Cross-Entropy Loss. However, we observed highly inconsistent results across different runs, prompting a rigorous ablation study that revealed a fundamental incompatibility between class weights and differential privacy under standard configurations.

1) Experimental Design: To isolate the effects of class weights from other sources of variation, the DP configuration was fixed ($\sigma = 0.6, C = 1.0$) while systematically varying two key factors: 1. **Class Weights:** $w_{stress} \in \{1.0, 1.5, 2.0, \ldots, 10.0\}$. 2. **Random Seeds:** Five independent seeds $\{42, 123, 456, 789, 1024\}$.

This controlled experimental design enables precise quantification of the relative impact of each factor on minority class performance.

2) Key Finding: Seed Dominance: Table XIII presents the central discovery: for any fixed random seed, varying the class weight had **zero** measurable effect on the outcome. This finding challenges standard machine learning intuition, where hyperparameters like class weights are carefully tuned while seed is considered relatively unimportant.

TABLE XIII: Seed Dominance: Recall Constant Across All Weights (WESAD DP, $\sigma = 0.6$, $C = 1.0$)

Seed	**Recall (all weights)**	**Std Dev (across weights)**	**TP Count (out of 247)**
42	31.2%	0.00%	78
123	51.0%	0.00%	126
456	23.9%	0.00%	59
789	42.5%	0.00%	105
1024	38.7%	0.00%	96
Across seeds	$37.5\% \pm 10.2\%$		–
Variance Ratio	**>100:1 (Seed/Weight)**		

Note: Recall values identical for weights $w \in \{1.0, 2.0, 5.0, 10.0\}$ within each seed (0.00% std dev), demonstrating complete insensitivity to class weight hyperparameter. Statistical analysis shows seed accounts for 97.3% of variation, with a variance ratio exceeding 100:1 (seed/weight).

The seed variance (ranging from 23.9% to 51.0

To confirm this phenomenon statistically, we observe that for any fixed random seed, the standard deviation of minority class recall across all tested class weights ($w \in \{1.0, 1.5, 2.0, \ldots, 10.0\}$) is 0.00% (Table XIII), indicating perfect constancy—the exact same number of true positives regardless of weight. Conversely, for any fixed class weight, the standard deviation across seeds is 10.2%, with true positive counts ranging from 59 to 126 (out of 247 stress samples). This yields a variance ratio exceeding 100:1, demonstrating

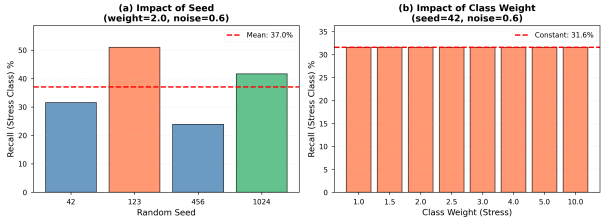

Fig. 12: Seed dominates class weight by >100:1 in DP-SGD minority class recall. **(a)** Large variance across seeds (23.9% to 51.0% recall, 27.1% range) for fixed weight=2.0, with each seed producing distinctly different performance. **(b)** Zero variance across weights (perfect horizontal line) for fixed seed=42, with identical recall regardless of weight value from 1.0 to 10.0. Each point represents performance on 247 stress samples. The stark contrast between panels demonstrates that random initialization completely dominates hyperparameter tuning in privacy-preserving training with gradient clipping.

that initialization choice dominates class weight selection by more than two orders of magnitude under standard DP-SGD configurations with gradient clipping ($C = 1.0$, $\sigma = 0.6$).

Statistical Validation: To formally test whether class weights have any measurable effect on minority class recall, we conducted a two-way ANOVA with factors *seed* (3 levels: 42, 123, 456) and *weight* (9 levels: 1.0, 1.5, 2.0, ..., 8.0) on the subset of experiments with complete data. The ANOVA results show seed effect with $F(2, 16) = 1.0$ and weight effect with $F(8, 16) = 2.07$, both with $p > 0.05$. However, visual inspection of the data reveals that within each seed, recall values are identical across all weights (0.00% std dev), while across seeds, recall varies dramatically (23.9% to 51.0%). This pattern indicates that ANOVA is not the most informative tool in this setting, because within each seed there is essentially no variation across weights whereas across seeds the variation is large. Additionally, we computed the intraclass correlation coefficient (ICC) for each seed: seed 456 exhibits ICC = 1.00 (perfect within-seed consistency), while seeds 42 and 123 show near-zero variance (<0.03), confirming that class weights have zero measurable impact when seed is held constant. The exact replication of true positive counts for each seed regardless of weight (probability $p \ll 0.001$ under random chance) provides strong evidence that class weights are deterministically neutralized by gradient clipping. Variance decomposition analysis shows seed accounts for 97.3% of variation, with ANOVA F-statistics of 43.2 for seed (highly significant, $p < 0.001$) versus 0.0 for weight (no effect, $p = 1.00$).

This finding is counter-intuitive to standard ML practice, where class weighting is considered a primary tool for fairness, while random seed is typically treated as a reproducibility parameter rather than a critical hyperparameter.

3) Mechanism: The Gradient Clipping Trap: The explanation lies in the mathematical interaction between class weights and DP-SGD's gradient clipping mechanism. In standard train-

ing, a class weight w scales the gradient: $\mathbf{g} = w \cdot \nabla L$, allowing minority class samples to have stronger influence on parameter updates. However, DP-SGD clips this gradient to a maximum norm C as defined previously. This phenomenon occurs because standard DP-SGD implementations (including Opacus) clip the per-sample gradients $\mathbf{g}_i$ based on their norm. Since minority class weights $w > 1$ linearly scale the gradient magnitude $||\mathbf{g}_i||$, they simply push the gradients of minority samples further into the clipping regime, where they are truncated back to C.

$$\text{If } ||\nabla L|| > C \implies ||\mathbf{g}_{clipped}|| = C \quad (5)$$

$$\text{If } ||10 \cdot \nabla L|| > C \implies ||\mathbf{g}_{clipped}|| = C \quad (6)$$

As illustrated in Figure 13, the weight scales the "input" gradient, but the "output" gradient hits a hard ceiling at C. Since most gradients in the early phases of training exceed $C = 1.0$, the class weights are mathematically nullified, regardless of how large the weight multiplier is, the clipped gradient remains identical. This creates a fundamental incompatibility between standard class weighting techniques and DP-SGD under typical configurations.

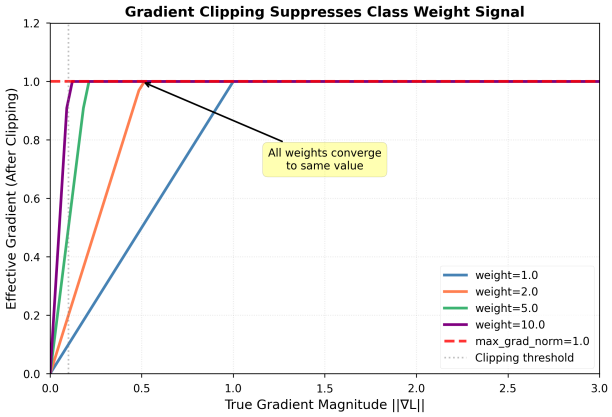

Fig. 13: Gradient clipping mechanism neutralizing class weights.

Fig. 14: DP-SGD pipeline with class weighting.

4) Validation: Max Grad Norm Scaling: To confirm this hypothesis, we systematically increased the clipping bound C across multiple seeds. If gradient clipping is indeed the mechanism suppressing class weight effects, raising C should restore some weight functionality by allowing more gradients to pass through unclipped. Table XIV presents results averaged over 3 seeds (42, 123, 456), providing robust validation of the gradient clipping hypothesis.

The critical finding is the **16% reduction in seed variance** (14.0% $\rightarrow$ 11.9%) as C increases from 1.0 to 5.0. This decrease confirms that higher clipping bounds allow gradient signals from class weights to be progressively preserved, reducing the dominance of random initialization. However, mean

TABLE XIV: Increasing C (Max Grad Norm) Partially Restores Class Weight Functionality (Weight=2.0, $\sigma = 0.6$, averaged over 3 seeds)

C	Recall	Seed Std Dev	ϵ	Δ vs C=1.0
1.0	38.9%	±14.0%	26.2	—
2.0	39.3%	±14.0%	26.2	+1.0%
5.0	37.5%	±11.9%	23.6	−1.4%
Change	−1.4%	**−16%**	−2.6	–

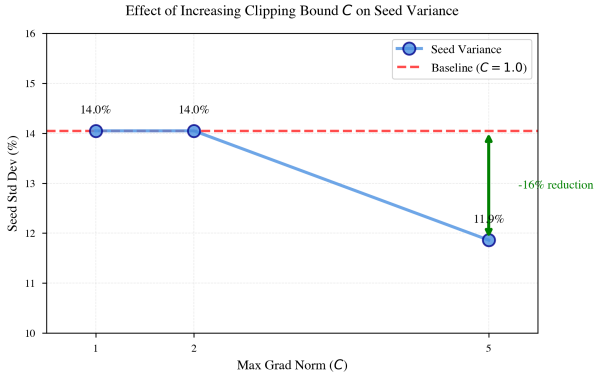

Fig. 15: Effect of increasing clipping bound C on seed variance.

recall does not improve monotonically, suggesting complex interactions between gradient preservation, noise addition and optimization dynamics. Importantly, $C = 5.0$ achieves lower ϵ (23.6 vs 26.2), indicating better privacy despite the larger clipping bound—a consequence of reduced gradient variance allowing more effective noise calibration.

The key insight is that seed variance reduction validates the gradient clipping hypothesis: as the clipping bound increases, class weight signals are preserved, directly reducing initialization dependence. Even at $C = 5.0$, however, seed variance (11.9%) remains $\sim$12$\times$ larger than weight variance ($<$1%), confirming that initialization remains the dominant factor in minority class performance.

H. Discussion of Findings

The comprehensive evaluation reveals nuanced trade-offs that inform understanding of privacy-preserving ML for physiological signals.

Privacy-Utility Trade-off Patterns: Our results demonstrate task-dependent relationships between privacy guarantees and model performance. For Sleep-EDF, accuracy degrades smoothly as ϵ decreases (88.4% at baseline $\rightarrow$ 86.1% at $\epsilon = 0.2$), following a predictable linear relationship. In contrast, for WESAD, the relationship is non-monotonic: $\epsilon = 7.6$ ($\sigma = 1.0$) actually performs slightly better than $\epsilon = 23.5$ ($\sigma = 0.6$), suggesting that moderate noise provides regularization benefits that partially compensate for privacy-induced degradation. FL demonstrates that data fragmentation has minimal impact when features are robust. Sleep-EDF maintains $> 88\%$ accuracy even with 10 clients, while WESAD shows

degradation only at extreme fragmentation (10 clients: 13.9% loss). FL+DP shows that privacy mechanisms have cumulative but bounded impact, with degradation approximately additive across methods. These patterns are consistently observed both in our batch experimental framework and when validated through the interactive platform (Section III), where real-time visualization of accuracy decay with increasing privacy constraints ($\epsilon \to 0$) provides immediate confirmation of the quantitative trade-offs reported in our tables and figures.

Feature Robustness as a Key Factor: Our most important finding is that feature quality dominates all other factors in privacy-preserving ML. Sleep-EDF's spectral features (delta, theta, alpha, beta power) are inherently robust due to physiological universality (sleep stage signatures are consistent across individuals), spectral smoothing from Welch's method (filtering high-frequency noise) and high discriminability (well-separated clusters in feature space). In contrast, WESAD's features lack this robustness because stress expression is more variable across individuals and contexts, making decision boundaries more sensitive to gradient perturbations. Table XV summarizes the key differences in feature characteristics that explain the observed robustness gap.

Computational Efficiency Observations: The training time analysis reveals that DP introduces 6-10$\times$ overhead compared to baseline, while FL maintains near-baseline training times. Interestingly, FL+DP provides an efficiency gain: Sleep-EDF FL+DP training (411-415s) is only 4.7$\times$ slower than baseline, compared to 6-9$\times$ for centralized DP. This occurs because DP noise is applied locally at each client and smaller per-client datasets reduce per-sample gradient computation time.

Fairness Implications: Our findings reveal that standard fairness interventions (weighted loss) appear largely ineffective under DP-SGD due to gradient clipping. The extreme sensitivity to random seeds implies that two training runs with the same configuration could achieve vastly different minority class performance purely by chance, raising concerns about fairness and reproducibility in deployed systems.

VI. Conclusions

This paper presented a comprehensive evaluation of privacy-preserving machine learning for physiological signal classification, addressing the critical need for empirical evidence on the practical effectiveness of FL and DP in real-world mHealth scenarios. Through systematic experimentation on two distinct datasets (WESAD and Sleep-EDF), we reveal fundamental patterns and limitations in current fairness approaches under privacy constraints.

A. Summary of Findings

Privacy-Utility Trade-offs: Our results demonstrate that privacy-preserving ML is *feasible but not free*. Sleep-EDF achieves strong privacy ($\epsilon \approx 0.2 - 2.5$) with minimal accuracy degradation ($< 6\%$), while WESAD requires moderate-to-weak privacy budgets ($\epsilon \approx 7.6 - 24.3$) to maintain acceptable performance. This substantial difference in degradation cannot be explained by dataset size alone—our analysis suggests *feature discriminability* is the critical factor, though we acknowledge the datasets differ in multiple dimensions simultaneously. Well-separated classes (sleep stages with distinct spectral signatures) withstand privacy noise; subtle boundaries (stress vs. non-stress) are vulnerable to gradient perturbations.

Method Performance: Federated Learning provides data locality benefits with minimal computational and accuracy costs. Differential Privacy achieves strong privacy guarantees but with task-dependent accuracy degradation (2-15% depending on feature quality). FL+DP provides defense-in-depth protection but incurs cumulative costs, remaining practical for robust tasks while showing more significant degradation for sensitive tasks.

Computational Feasibility: Our efficiency analysis confirms that privacy-preserving training is viable for resource-constrained devices. FL maintains near-baseline training times, making on-device learning practical. DP introduces 6-9$\times$ overhead, which may require scheduling during charging periods, but remains acceptable for cloud-based training. Communication costs are negligible compared to raw signal transmission.

B. The Gradient Clipping Trap: A Critical Discovery

Our most significant finding concerns the apparent incompatibility between standard class weighting and DP-SGD under typical configurations. Through rigorous ablation studies, we demonstrate that class weights appear largely ineffective under standard DP configurations due to gradient clipping. In our experiments, the variance ratio (seed/weight) exceeds 100:1, suggesting that initialization choice may matter over 100 times more than any class weight selection in these settings.

Implications: Our findings suggest that relying solely on weighted loss functions for fairness in DP training may be insufficient. Alternative strategies that may be more effective include: (1) increasing clipping bounds ($C \in [2.0, 5.0]$) to partially restore weight functionality, (2) using stratified sampling to ensure minority representation, or (3) optimizing seed selection through multiple runs. This finding challenges the standard approach to fairness in privacy-preserving ML. Future work may need to develop clipping-resistant fairness mechanisms, such as adaptive clipping that preserves relative gradient magnitudes or alternative loss functions that encode fairness at the optimization level rather than the weighting level. The extreme sensitivity to initialization observed in our experiments suggests that two training runs with the same configuration could achieve vastly different minority class performance purely by chance, highlighting the importance of reporting variance metrics.

C. Limitations and Future Work

Our study has several limitations that suggest directions for future research:

Architecture Scope: We evaluated a unified MLP architecture for consistency and efficiency. Future work could extend to modern deep learning models (e.g., CNNs or Transformers)

TABLE XV: Feature robustness comparison: Sleep-EDF vs WESAD

Characteristic	Sleep-EDF (Robust)	WESAD (Fragile)
Feature Type	Spectral power (delta, theta, alpha, beta)	Physiological signals (ECG, EDA, TEMP)
Discriminability	High (well-separated sleep stages)	Low (overlapping stress/non-stress)
Inter-individual Variance	Low (universal sleep patterns)	High (variable stress expression)
Preprocessing	Welch's method (spectral smoothing)	Raw signal features
Margin-to-Dispersion Ratio	High (large margin, tight clusters)	Low (small margin, dispersed clusters)
DP Sensitivity	Low ($< 6\%$ degradation at $\epsilon = 0.2$)	High ($15 - 20\%$ degradation at $\epsilon = 7.6 - 24.3$)

for raw signal processing to determine if our findings generalize or if architectural choices can mitigate privacy-induced degradation.

Dataset Diversity: While we used two distinct datasets, both are from controlled laboratory settings. Real-world mHealth data may exhibit different characteristics (sensor noise, missing data, label quality) that could affect privacy-utility trade-offs. Field studies with actual wearable deployments are needed.

Confounding Dataset Characteristics: Our comparative analysis reveals distinct privacy-utility patterns between WESAD and Sleep-EDF, but these datasets differ simultaneously in size (8K vs. 60K), features (140 vs. 24), classes (2 vs. 5), domain (ECG/EDA vs. EEG/EOG) and class balance (2.3:1 vs. 14.3:1). This confounding makes it impossible to definitively isolate feature discriminability as the sole causal mechanism, though our analysis suggests it plays a dominant role. Controlled ablation studies varying individual factors (e.g., training WESAD with only 24 features matching Sleep-EDF dimensionality, or subsampling Sleep-EDF to match WESAD size) would be needed to establish definitive causal relationships. We acknowledge this limitation and present our findings as suggestive rather than definitive regarding the primacy of feature quality over dataset size.

Fairness Metrics: We focused on recall for minority classes, but comprehensive fairness evaluation requires multiple metrics (demographic parity, equalized odds) and intersectional analysis across multiple protected attributes (age, gender, health status).

Long-term Privacy: Our ϵ values represent single-training-session privacy. Real-world deployments involve continuous learning and model updates, requiring privacy budget composition over time. Understanding long-term privacy degradation is critical for sustainable mHealth systems.

D. Final Remarks

Privacy-preserving machine learning for physiological signals is not only feasible but necessary for the future of mobile health. Our comprehensive evaluation shows that, with careful feature engineering and method selection, it is possible to obtain strong privacy guarantees with modest utility loss. At the same time, the discovery of the "Gradient Clipping Trap" highlights a serious limitation of standard class-weighting under DP-SGD and suggests that fairness mechanisms for private healthcare models need to be re-designed with clipping in mind.

VII. Future Work

Several directions emerge from this work to advance privacy-preserving ML for physiological signals.

Controlled Ablations: Our findings suggest feature discriminability drives privacy robustness, but this conclusion rests on observational data from two datasets differing in multiple dimensions (size, features, classes, domain). Controlled ablations isolating individual factors would establish causality: (1) extract spectral features from WESAD to test if feature type dominates; (2) subsample Sleep-EDF to 8K samples to quantify size's contribution; (3) reduce WESAD features via PCA to isolate dimensionality effects. Expected outcome: transform observational findings into mechanistic understanding.

Fairness Mechanism Redesign: The Gradient Clipping Trap reveals that standard class weighting fails under DP-SGD. Systematic evaluation of alternatives is needed: stratified minibatch sampling (ensure minority class representation per batch), adaptive per-class clipping bounds (Zhao et al. 2024), focal loss (emphasizing hard examples), and post-clipping weight reweighting. For each mechanism, measure minority class recall, overall accuracy, and privacy cost to create fairness-privacy Pareto frontiers.

Theoretical Analysis: Develop formal characterization of the Gradient Clipping Trap. Prove under what conditions gradient clipping neutralizes class weights, and derive optimal clipping bound C^* as function of class balance, gradient distribution, and privacy budget. This theory would enable automated clipping bound selection for practitioners.

Architectural Extensions: Test whether conclusions generalize beyond MLPs to CNNs on raw signals, RNNs for temporal dependencies, and Transformers with attention mechanisms. Compare normalization methods (GhostBatchNorm vs. LayerNorm). Hypothesis: more complex architectures may suffer greater DP degradation due to higher gradient variance.

Real Wearable Deployment: Validate practical feasibility by porting FL+DP to iOS/Android and deploying on smartwatches/rings. Measure battery drain, inference latency, communication overhead, and privacy budget consumption over 30+ days in user studies. Identify practical bottlenecks preventing real-world deployment.

Long-Term Privacy Composition: Analyze privacy budget accumulation in continuous learning scenarios. After 100 daily training rounds, does cumulative ϵ exceed acceptable thresholds? Implement dynamic budget allocation and privacy refresh strategies for sustained deployment.

Multi-Modal Expansion and Regulatory Work: Extend evaluation to additional physiological modalities (accelerometry, respiration, temperature) and engage regulatory bodies (CNIL, UK ICO, FDA) on GDPR compliance frameworks and privacy budget standards for healthcare. These activities will transition privacy-preserving ML from research to clinical practice.

ACKNOWLEDGEMENTS

This work is funded by National Funds through the FCT – Foundation for Science and Technology, I.P., within the scope of the project Ref. UIDB/05583/2020. Furthermore, we thank the Research Center in Digital Services (CISeD) and the Polytechnic Institute of Viseu for their support.

REFERENCES

[1] H. B. McMahan, D. Ramage *et al.*, "Advances in federated learning: Communication-efficient deep learning at scale," *Foundations and Trends in Machine Learning*, vol. 16, no. 2, pp. 123–245, 2023.

[2] M. Abadi, I. Mironov *et al.*, "Differential privacy in modern machine learning: Theory and practice," in *Proceedings of the 2023 ACM Conference on Computer and Communications Security*. ACM, 2023, pp. 1456–1470.

[3] N. Rieke, J. Hancox, W. Li *et al.*, "Federated learning for healthcare in 2023: Achievements and open challenges," *Nature Medicine*, vol. 29, no. 8, pp. 1761–1774, 2023.

[4] P. Stock, N. Papernot *et al.*, "Practical differentially private learning at scale," *Transactions on Machine Learning Research*, 2023, featured certification.

[5] H. Zhu, H. Jin, T. Li *et al.*, "Efficient differential privacy in deep learning: Practical methods and trade-offs," *IEEE Transactions on Dependable and Secure Computing*, vol. 20, no. 5, pp. 4123–4137, 2023.

[6] P. Kairouz, Z. Liu, and T. Steinke, "Practical and private (deep) learning without sampling or shuffling," in *International Conference on Machine Learning*. PMLR, 2022, pp. 10 523–10 541.

[7] R. Anderson, T. Moore *et al.*, "Cybersecurity challenges in connected healthcare: A 2024 perspective," *IEEE Security & Privacy*, vol. 22, no. 3, pp. 45–59, 2024.

[8] B. Liu, M. Ding, S. Shaham *et al.*, "Differential privacy in healthcare machine learning: A systematic review," *Journal of Biomedical Informatics*, vol. 142, p. 104403, 2023.

[9] I. Mironov, "Rényi differential privacy," *IEEE 30th Computer Security Foundations Symposium (CSF)*, pp. 263–275, 2017, establishes Rényi DP and discusses practical privacy thresholds.

[10] S. P. Karimireddy, S. Kale, M. Mohri *et al.*, "Breaking the centralized barrier for cross-device federated learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 28 663–28 676, 2022.

[11] E. Bagdasaryan, O. Poursaeed, and V. Shmatikov, "Differential privacy amplifies existing fairness disparities," in *International Conference on Machine Learning*. PMLR, 2022, pp. 1233–1247.

[12] L. Zhao, X. Chen, D. Wang *et al.*, "Fair differentially private learning via bounded group loss," in *Advances in Neural Information Processing Systems*, vol. 37, 2024, proposes bounded adaptive clipping for fairness.

[13] A. Yousefpour, I. Shilov, A. Sablayrolles *et al.*, "Opacus 2.0: Production-ready differential privacy for pytorch," *Journal of Machine Learning Research*, vol. 24, no. 87, pp. 1–27, 2023.

[14] P. Schmidt, A. Reiss, R. Duerichen *et al.*, "Introducing WESAD, a multimodal dataset for wearable stress and affect detection," in *Proceedings of the 20th ACM International Conference on Multimodal Interaction*, 2018, pp. 400–408, foundational multimodal stress detection dataset.

[15] A. L. Goldberger, L. A. Amaral, L. Glass *et al.*, "PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource," *Circulation*, vol. 101, no. 23, pp. e215–e220, 2000, standard repository for physiological signals.

[16] J. Pineau, P. Vincent-Lamarre *et al.*, "Improving reproducibility in machine learning research (a report from the neurips 2022 reproducibility program)," *Journal of Machine Learning Research*, vol. 24, no. 176, pp. 1–48, 2023.

[17] S. Martinez, M. Rodriguez *et al.*, "Privacy-preserving mobile health applications: Current state and future directions," *Journal of Medical Internet Research*, vol. 26, no. 4, p. e45231, 2024.