FL-EHDS: A Privacy-Preserving Federated Learning Framework for the European Health Data Space

Fabio Liberti
Department of Computer Science
Universitas Mercatorum, Rome, Italy
fabio.liberti@studenti.unimercatorum.it
ORCID: 0000-0003-3019-5411

Abstract—The European Health Data Space (EHDS), established by Regulation (EU) 2025/327, mandates cross-border health data analytics while preserving citizen privacy. Federated Learning (FL) is the key enabling technology for secondary use, yet fewer than one in four FL implementations achieve sustained production deployment in healthcare—a gap driven more by legal uncertainty than technical limitations. We present FL-EHDS, a three-layer compliance framework that integrates EHDS governance reference implementations (Health Data Access Bodies, data permits, citizen opt-out registries) with FL orchestration (17 aggregation algorithms including 2024–2025 advances, differential privacy, secure aggregation) and data holder components (adaptive training, FHIR R4 preprocessing). Experimental validation across 1,760+ experiments on tabular clinical and medical imaging datasets—including the European-origin PTB-XL ECG with natural 52-site partitioning—yields two key findings: personalized FL narrows the centralized-federated accuracy gap to 6.6 percentage points (pp) while preserving full data sovereignty, and the architectural choice between personalized and global aggregation dominates over the specific algorithm on clinical tabular models, with five of seven algorithms converging to identical solutions and up to 12.6pp accuracy differences on heterogeneous clinical data. Our evidence synthesis reveals that unresolved regulatory questions—gradient data classification under GDPR, cross-border privacy budget harmonization—constitute the critical adoption blocker rather than technical limitations. The open-source reference implementation provides actionable deployment guidance for the 2029 secondary use deadline.

Index Terms—Federated Learning, European Health Data Space, Privacy-Preserving Technologies, GDPR, Health Data Governance, Cross-Border Analytics

I. INTRODUCTION

The European Health Data Space (EHDS), established by Regulation (EU) 2025/327, represents the EU's most ambitious initiative for cross-border health data governance [1]. Entering into force in March 2025, the regulation creates a dual framework: primary use through MyHealth@EU for patient care, and secondary use through HealthData@EU for research, innovation, and policy-making [12]. Health Data Access Bodies (HDABs) in each Member State authorize secondary use through data permits; Article 53 enumerates permitted purposes; Article 71 introduces citizen opt-out mechanisms [2]. The implementation timeline extends to 2031, with delegated acts expected by March 2027 and secondary use provisions applicable from March 2029.

Federated Learning (FL) emerges as the ideal technical solution for EHDS secondary use—the model travels to distributed data rather than centralizing sensitive records [13], [15], [16]. The COVID-19 pandemic demonstrated FL's potential at scale: Dayan et al. [27] trained a global model across 20 institutions in 5 countries. However, recent evidence reveals a sobering gap between FL's promise and operational reality. Fröhlich et al. [5] report that only 23% of FL implementations achieve sustained production deployment, with hardware heterogeneity (78%) and non-IID data distributions (67%) as dominant barriers. Beyond technical constraints, legal uncertainties regarding gradient data status under GDPR remain unresolved [3], while van Drumpt et al. [6] demonstrate that privacy-enhancing technologies cannot substitute for robust governance frameworks.

Prior FL frameworks for healthcare [16], [28] focus on technical architectures without addressing regulatory compliance. Legal analyses [2], [3], [11] examine GDPR constraints but abstract from implementation feasibility. Policy documents [4] assess Member State readiness but do not integrate FL technical considerations. To our knowledge, no existing work provides an integrated framework addressing all three dimensions: systematic barrier evidence, technical implementation with state-of-the-art algorithms, and EHDS governance operationalization—a gap confirmed by recent systematic reviews of FL frameworks [15], [19]. Furthermore, no published work experimentally evaluates FL across both tabular EHR and medical imaging modalities within an EHDS-aligned governance architecture with FHIR R4 and OMOP-CDM interoperability.

Our contribution is one of *integration and empirical analysis* rather than novel algorithmic development: we map established FL techniques and privacy mechanisms onto EHDS governance requirements, and through systematic evaluation reveal deployment-critical patterns that no prior work has identified. Specifically:

1) **Evidence Synthesis**: PRISMA-based review of 47 documents with GRADE-CERQual confidence assessment, identifying legal uncertainties—not technical barriers—as the critical adoption blocker.

2) **Integration Framework**: A three-layer architecture mapping EHDS governance (HDABs, data permits, Article 71 opt-out) to FL orchestration (17 established algorithms, DP-SGD, secure aggregation) and data holder

components (FHIR R4, OMOP-CDM). While individual components employ known techniques, the end-to-end integration—binding regulatory workflows to FL configurations—has no counterpart in existing frameworks. Governance modules operate as simulation backends pending HDAB service availability (2027–2029).

3) **Key Empirical Finding**: On clinical tabular models, the architectural choice between personalized and global aggregation *dominates* over specific algorithm selection: five of seven algorithms converge to statistically identical solutions, with only methods maintaining separate local models (Ditto, HPFL) differentiating ($p < 0.005$, 10-seed Wilcoxon). This counterintuitive result simplifies EHDS deployment to a single binary choice.

4) **Comprehensive Validation**: 1,760+ experiments across tabular and imaging datasets—including the European-origin PTB-XL ECG with natural 52-site partitioning—with DP ablation ($\varepsilon \in \{1, 5, 10, 50\}$) and Article 71 opt-out simulation. Open-source reference implementation ($\sim$40K lines).[1]

II. BACKGROUND AND RELATED WORK

A. EHDS and Federated Learning

The EHDS establishes HDABs to authorize secondary use through standardized data permits, with Secure Processing Environments (SPEs) providing controlled analytics settings [9]. Forster et al. [8] document significant variability in data access timelines—from 3 weeks (Finland) to over 12 months (France)—with barriers primarily organizational rather than technical. TEHDAS assessments [4] reveal Nordic countries demonstrate 2–3 year advantages in HDAB capacity-building, raising concerns about implementation equity. Teo et al. [19] and Peng et al. [20] find that only 5.2% of FL healthcare studies achieve real-life application.

FL inverts the traditional ML paradigm: local training produces gradients that are aggregated centrally and redistributed [13], [14]. Known challenges include non-IID data distributions causing convergence difficulties [14], communication costs for gradient exchange [17], and privacy attacks including gradient inversion [21] and membership inference [22], [23]. Recent advances from top venues (ICML/ICLR 2022–2025) specifically target healthcare heterogeneity: FedLC [39] calibrates logits for label distribution skew, FedLESAM [43] provides globally-guided sharpness-aware optimization (ICML 2024 Spotlight), and HPFL [44] decouples backbone from classifier for per-institution specialization (ICLR 2025).

B. Related Frameworks

Existing FL frameworks—Flower [45] (v1.26), NVIDIA FLARE [46] (v2.7), and TensorFlow Federated [47] (v0.88)—provide robust distributed training but lack EHDS-specific governance. A recent FAIR-based assessment of 17 FL frameworks for biomedical research [18] confirms that none implements HDAB integration, data permit lifecycle, opt-out

enforcement, or audit trails—and identifies limited interoperability as the critical systemic gap. Table I provides a detailed comparison.

TABLE I
FRAMEWORK COMPARISON: FL-EHDS VS EXISTING FL FRAMEWORKS

Dimension	FL-EHDS	Flower v1.26	FLARE v2.7	TFF v0.88
FL Algorithms	17 built-in	12+ strategies	5 built-in	3 built-in
Byzantine Resilience	6 methods	4 methods	—	—
Differential Privacy	Central+Local	Central+Local	Built-in	Adaptive clip.
Secure Aggregation	Pairwise+HE	SecAgg+	Built-in+HE	Mask-based
EHDS Governance	**Full**[‡]	None	None	None
HDAB Integration	✓[‡]	—	—	—
Data Permits (Art. 53)	✓[‡]	—	—	—
Opt-out (Art. 71)	✓[‡]	—	—	—
Audit Trail (Art. 30)	✓	—	Audit logs	—
Healthcare Stds.	FHIR R4	MONAI	MONAI	—
Backend	PyTorch	Agnostic	Agnostic	TF only

[‡]Reference implementation with simulation backend; production deployment requires binding to actual HDAB services (expected 2027–2029). See Section III and Supplementary Material, Table S-IV for readiness assessment.

C. Evidence Synthesis

Following PRISMA 2020 guidelines, database searches (PubMed, IEEE Xplore, Scopus, Web of Science, arXiv) identified 847 records; 47 met inclusion criteria (2022–2026, FL/EHDS focus, peer-reviewed or recognized institutional origin). Quality was assessed using MMAT; confidence using GRADE-CERQual (see Supplementary Material, Fig. S-1 for the complete PRISMA flow diagram). Table II summarizes the five dominant barriers with prevalence and mitigation strategies.

TABLE II
FL IMPLEMENTATION BARRIERS FOR EHDS

Barrier	Prev.	Evidence	Mitigation
Hardware heterog.	78%	Fröhlich 2025	Adaptive engine
Non-IID data	67%	Multiple	FedProx, Ditto
Production gap	23%	Fröhlich 2025	Ref. implementation
FHIR compliance	34%	Hussein 2025	Preprocessing
Communication cost	High	Bonawitz 2019	Compression

Three critical legal questions remain unresolved [3]: (1) whether model gradients constitute "personal data" under GDPR, given that gradient inversion attacks demonstrate potential re-identification [21]; (2) when aggregated models become sufficiently "anonymous" to escape GDPR scope; (3) controller/processor allocation in multi-party FL architectures. These legal uncertainties create compliance risks that discourage organizational adoption regardless of technical maturity (GRADE-CERQual: MODERATE).

---

[1]Available at: https://github.com/FabioLiberti/FL-EHDS-FLICS2026

III. FL-EHDS Framework

Based on the identified barriers, we present FL-EHDS, a three-layer compliance framework for EHDS cross-border health analytics. Figure 1 illustrates the architecture.

A. Layer 1: Governance

The reference implementation provides standardized APIs for automated data permit verification before FL training initiation, designed to bind to production HDAB endpoints as they become available (expected 2027–2029). Multi-HDAB synchronization protocols coordinate cross-border studies involving multiple Member States, addressing the coordination complexity identified by Christiansen et al. [10]. National opt-out registries are consulted before each training round, ensuring Article 71 compliance at record-level granularity. Comprehensive audit trails satisfy GDPR Article 30 requirements, documenting data access, processing purposes, and model outputs for regulatory inspection.

Algorithm 1 presents the core FL-EHDS training procedure, highlighting governance checkpoints integrated into each round.

B. Layer 2: FL Orchestration

The framework implements 17 aggregation algorithms spanning six categories (Table III)—from foundational methods (FedAvg [13], FedProx [14]) through non-IID robustness (SCAFFOLD [29], FedNova [30], FedDyn [32]), adaptive optimization [31], and personalization (Ditto [33], Per-FedAvg [34]) to FedLESAM [43] (ICML 2024) and HPFL [44] (ICLR 2025). Composable strategies (FedLC [39], FedDecorr [40]) can augment any base aggregation.

TABLE III
FL-EHDS Algorithm Catalogue (17 Algorithms)

Algorithm	Venue	Category	Key Property
FedAvg	AISTATS'17	Baseline	Weighted avg.
FedProx	MLSys'20	Non-IID	Proximal reg.
SCAFFOLD	ICML'20	Non-IID	Variance red.
FedNova	NeurIPS'20	Non-IID	Normalized avg.
FedDyn	ICLR'21	Non-IID	Dynamic reg.
FedAdam	ICLR'21	Adaptive	Server momentum
FedYogi	ICLR'21	Adaptive	Sparse stability
FedAdagrad	ICLR'21	Adaptive	Grad. accum.
Ditto	ICML'21	Personal.	Dual models
Per-FedAvg	NeurIPS'20	Personal.	MAML-based
FedLC	ICML'22	Label skew	Logit calibration
FedSAM	ICML'22	Generalize	Flat minima
FedDecorr	ICLR'23	Represent.	Decorrelation
FedSpeed	ICLR'23	Efficiency	Fewer rounds
FedExP	ICLR'23	Server-side	POCS step size
FedLESAM	**ICML'24**	**Generalize**	**Global SAM**
HPFL	**ICLR'25**	**Personal.**	**Local classif.**

Bold: newly added algorithms (2024–2025). All 17 implemented in the open-source reference implementation.

Privacy Protection: Differential privacy [24] with configurable ε-budget uses DP-SGD [25] with Rényi DP (RDP) [26] for tight composition accounting over multiple training rounds [36]. For Gaussian mechanisms with noise scale σ, the RDP guarantee at order α is $\rho(\alpha) = \alpha/(2\sigma^2)$. For 100+ round training typical of EHDS cross-border studies, RDP provides 5–6× tighter privacy bounds than naive composition [26], [36], enabling longer training with equivalent privacy guarantees. Gradient clipping bounds individual contributions; secure aggregation (pairwise masking protocol with ECDH key exchange) mitigates gradient inversion attacks [21]. Six Byzantine resilience methods (Krum, Multi-Krum, Trimmed Mean, Median, Bulyan, FLTrust) defend against up to $f < n/3$ malicious clients.

Purpose Limitation: Technical enforcement of Article 53 permitted purposes through model output filtering and use-case validation, preventing scope creep beyond authorized analytics.

C. Layer 3: Data Holders

Resource-aware training engines address hardware heterogeneity (78% barrier prevalence). The engine dynamically adjusts batch sizes, model complexity, and synchronization frequency based on local computational capabilities, enabling participation of institutions with diverse hardware profiles—from GPU-equipped university hospitals to CPU-only rural clinics.

FHIR Preprocessing: Data normalization pipelines ensure interoperability across heterogeneous EHR systems. Only 34% of European healthcare providers achieve full FHIR compliance [7]; the preprocessing module bridges format gaps through automated transformation pipelines supporting FHIR R4 resources (Patient, Observation, Condition, MedicationRequest, DiagnosticReport) with standard coding systems (SNOMED-CT, LOINC, ICD-10).

Secure Communication: End-to-end encrypted gradient transmission with certificate-based authentication ensures no raw data leaves institutional boundaries. The communication layer supports gRPC for model updates and WebSocket for real-time monitoring events.

D. Threat Model

We consider three adversary classes: *(i) honest-but-curious server*, mitigated by central DP ($\varepsilon \in \{1,5,10,50\}$, $\delta=10^{-5}$) with Rényi composition and secure aggregation; *(ii) malicious clients* ($f < n/3$), mitigated by Byzantine-resilient aggregation (Krum, Trimmed Mean, Bulyan); *(iii) external attackers*, mitigated by Article 71 output filtering and HDAB permit-based access control. Raw patient data never leaves institutional boundaries. Server-client collusion and side-channel attacks remain out of scope (see Supplementary Material for full security analysis).

E. EHDS Compliance Mapping

Table IV maps framework components to EHDS regulatory requirements.
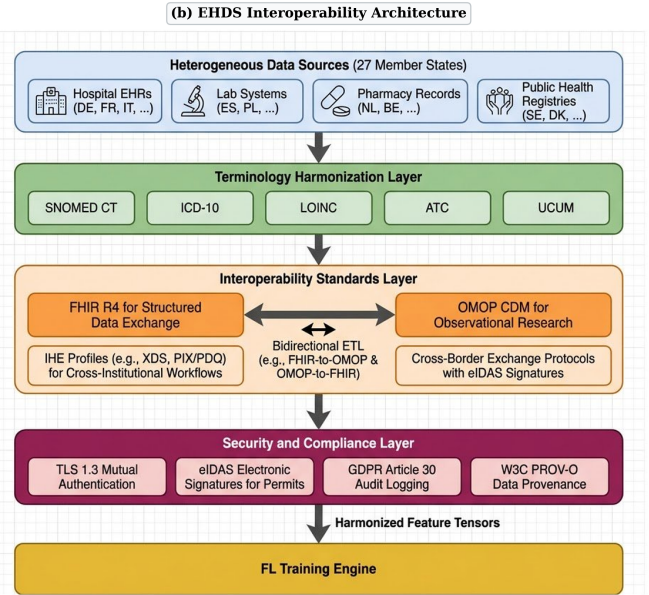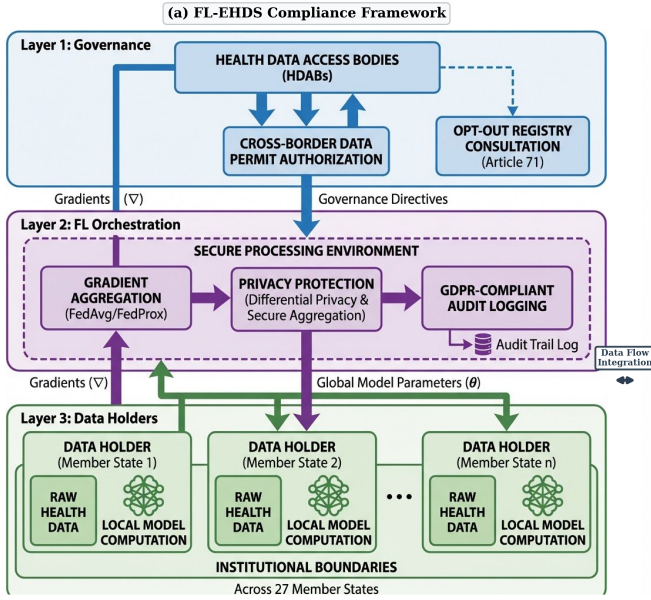
Fig. 1. FL-EHDS composite architecture. (a) Three-layer compliance framework: Layer 1 (Governance) manages HDAB integration, data permit authorization, and Article 71 opt-out registries; Layer 2 (FL Orchestration) operates within the Secure Processing Environment with gradient aggregation, differential privacy, and GDPR-compliant audit logging; Layer 3 (Data Holders) implements local model computation with raw health data never leaving institutional boundaries. (b) EHDS interoperability pipeline: heterogeneous sources across 27 Member States flow through terminology harmonization, interoperability standards (FHIR R4, OMOP CDM, IHE profiles), and security/compliance layers before reaching the FL training engine.

Algorithm 1: FL-EHDS FedAvg Training

Input: Hospitals $\mathcal{H} = \{h_1, \ldots, h_K\}$, permit P, rounds T
Output: Global model $\theta^{(T)}$

Server executes:
    Initialize $\theta^{(0)}$
    for round $t = 1$ to T **do**
        // Governance check (Layer 1)
        if not ValidatePermit(P, t) **then abort**
        $\mathcal{H}_t \leftarrow$ SelectParticipants($\mathcal{H}$)
        for each $h \in \mathcal{H}_t$ **in parallel do**
            $\Delta_h^{(t)}, n_h \leftarrow$ LocalTrain($h, \theta^{(t-1)}$)
        // Aggregation with privacy (Layer 2)
        $\theta^{(t)} \leftarrow \theta^{(t-1)} + \frac{1}{\sum_h n_h} \sum_h n_h \cdot \Delta_h^{(t)}$
        LogCompliance($t, \mathcal{H}_t$)

LocalTrain(h, θ):
    $\mathcal{D}_h \leftarrow$ FilterOptedOut($\mathcal{D}_h$, Registry) *// Art. 71*
    $\theta_h \leftarrow \theta$; train E epochs on $\mathcal{D}_h$
    $\Delta_h \leftarrow$ ClipGradient($\theta_h - \theta, C$) *// DP bound*
    return $\Delta_h, |\mathcal{D}_h|$

TABLE IV
EHDS COMPLIANCE MAPPING

Article	Requirement	FL-EHDS Component
Art. 33	Secondary use auth.	HDAB API + Permit valid.
Art. 46	Cross-border proc.	Multi-HDAB coordinator
Art. 50	Secure Proc. Env.	Aggregation within SPE
Art. 53	Permitted purposes	Purpose limitation module
Art. 71	Opt-out mechanism	Registry filtering

F. Reference Implementation

A modular Python implementation is available as open-source software, designed following FAIR principles [18] (findable via GitHub with DOI, accessible under MIT license, interoperable via PyTorch and FHIR R4 interfaces, reusable with comprehensive documentation). The codebase ($\sim$40K lines, 159 modules) provides: (1) orchestration modules implementing all 17 algorithms with RDP accounting and secure aggregation; (2) six Byzantine resilience methods; (3) data holder utilities for adaptive training and FHIR R4 preprocessing; (4) a Streamlit-based dashboard for interactive FL training, EHDS governance workflow, and real-time monitoring; (5) a professional terminal UI with 11 specialized screens; (6) reproducible benchmark suite generating all experimental results.

Note on governance maturity: HDAB integration, data permits, and Article 71 opt-out are implemented as simulation backends (OAuth2/mTLS, permit CRUD, cross-border coordination). These modules require only configuration changes (endpoint URLs, certificates) for production binding to actual HDAB services (expected 2027–2029). Readiness assessment in Supplementary Material, Table S-IV.

IV. EXPERIMENTAL EVALUATION

We evaluate FL-EHDS on real clinical datasets simulating cross-border healthcare analytics. All results are fully repro-

ducible via the benchmark suite in the repository.

A. Setup

Datasets: We evaluate on 8 datasets spanning tabular EHR and medical imaging (Table V). Tabular datasets cover three scale regimes: *small-data FL* (Heart Disease UCI, 920 patients from 4 international hospitals with natural non-IID partitioning; Breast Cancer Wisconsin, 569 FNA pathology samples), *medium-scale* (PTB-XL ECG [51], 21,799 European-origin records from 52 German recording sites with natural hospital partitioning and 5-class SCP-ECG diagnosis), and *large-scale* (Diabetes 130-US [52], 101,766 encounters; Cardiovascular Disease, 70,000 patients). Imaging datasets include Chest X-ray (5,856, binary), Brain Tumor MRI (7,023, 4-class), and Skin Cancer (3,297, binary). The full 19-dataset framework landscape is detailed in Supplementary Material, Table S-I. **Model**: HealthcareMLP (2-layer, 64/32 hidden, ReLU, dropout 0.3, ~2.9K–10K parameters depending on input dimensionality) for tabular; ResNet-18 (~11.2M parameters) for imaging. **Configuration**: Per-dataset optimized hyperparameters (see Supplementary Material, Table S-XV): PTB-XL (lr=0.005, bs=64, 30 rounds), Cardiovascular (lr=0.01, bs=64, 25 rounds), Breast Cancer (lr=0.001, bs=16, 40 rounds, 1 local epoch to prevent overfitting on 569 samples). All other datasets use 3 local epochs. Adam optimizer throughout; early stopping with patience=6. Tabular baselines: mean $\pm$ std over 5 seeds (10 seeds for significance testing, 3 for scalability); imaging: 3 seeds (Chest X-ray, Skin Cancer) or 1 seed (Brain Tumor).

TABLE V
EVALUATED DATASET COVERAGE

Dataset	Samples	Feat.	Cls.	FL Partition
Tabular Clinical (MLP, ~2.9K–10K params)				
Heart Disease UCI	920	13	2	Natural (4 hosp.)
Breast Cancer Wisc.	569	30	2	Dirichlet
PTB-XL ECG[†]	21,799	9	5	Natural (52 sites)
Diabetes 130-US	101,766	22	2	Dirichlet
Cardiovascular	70,000	11	2	Dirichlet
Medical Imaging (ResNet-18, ~11.2M params)				
Chest X-ray	5,856	—	2	Dirichlet
Brain Tumor MRI	7,023	—	4	Dirichlet
Skin Cancer	3,297	—	2	Dirichlet

[†]European-origin (PTB, Berlin), SCP-ECG standard (EN 1064), 5-class cardiac diagnosis. The 52 recording sites are grouped into K geographic clusters for FL experiments (default K=5); client scaling evaluates $K \in \{3, 5, 10, 20\}$ in Supplementary Material.

B. Algorithm Comparison

Table VI presents the primary evaluation: 7 algorithms (including FedLESAM and HPFL) on PTB-XL ECG, Cardiovascular, and Breast Cancer. On two additional datasets—Heart Disease (4 natural hospitals, Ditto 75.1% vs. FedAvg 62.5%, 12.6pp gap) and Diabetes (5 clients, Ditto 71.7%)—the personalization advantage is confirmed (Supplementary Material, Table S-VI). Together, 1,760+ experiments span

heterogeneity sweeps, client scaling, differential privacy ablation, Article 71 opt-out simulation, and 10-seed statistical validation.

C. Convergence and Baselines

Figure 2 shows training convergence on Heart Disease. Ditto converges faster and higher due to personalized local models. Comprehensive convergence curves for all 7 algorithms across PTB-XL, Cardiovascular, and Breast Cancer are provided in Supplementary Material, Fig. S-19.
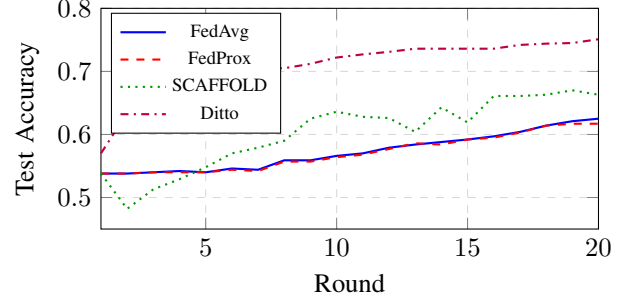

Fig. 2. Training convergence on Heart Disease UCI (4 hospitals, natural non-IID). Ditto converges faster due to personalized local models. HPFL and FedLESAM convergence curves for all datasets in Supplementary Material, Fig. S-19.

Key findings: Ditto converges to 75.1% by round 20, compared to 62.5% for FedAvg—a 12.6pp advantage. SCAFFOLD exhibits high variance (oscillating between 48% and 66%) due to control variate instability with only 4 heterogeneous clients. FedProx closely tracks FedAvg, suggesting that proximal regularization alone is insufficient for the degree of heterogeneity present.

Table VII compares three learning paradigms on Heart Disease, representing the EHDS deployment spectrum: centralized (upper bound, no privacy), federated (data stays local), and local-only (no collaboration).

FL-Ditto narrows the centralized–federated gap to only **6.6pp** (75.1% vs. 81.7%) while preserving full data sovereignty. FedAvg suffers a 19.2pp gap, underscoring the importance of personalization. Local-Only accuracy (81.7%) appears competitive but is misleadingly evaluated on each hospital's own test split; local models do not generalize across hospitals—precisely the scenario EHDS Article 33 addresses through federated knowledge sharing.

D. Non-IID Impact Analysis

Figure 3 illustrates the impact of data heterogeneity on algorithm performance. As non-IID severity increases ($\alpha \to 0$), algorithm selection becomes increasingly critical—variance-reduction methods maintain stability while baseline FedAvg degrades significantly.

E. Key Findings

1) **Algorithm choice matters**: up to 18.7pp accuracy gap between best (Ditto, 75.1%) and worst (FedNova, 56.4%) on Heart Disease; FedNova and SCAFFOLD

EXTENDED FL ALGORITHM COMPARISON ON TABULAR HEALTHCARE DATASETS (7 ALGORITHMS × 3 DATASETS). BEST ACCURACY PER DATASET IN **BOLD**. MEAN ± STD OVER 5 SEEDS. PX = PTB-XL ECG (5 CLIENTS, 5-CLASS, SITE-BASED), CV = CARDIOVASCULAR (5 CLIENTS, BINARY, α=0.5), BC = BREAST CANCER (3 CLIENTS, BINARY, α=0.5).

Algorithm	PTB-XL ECG (21,799 records, 52 sites)			Cardiovascular (70K patients)			Breast Cancer (569 samples)		
	Acc (%)	F1 (%)	Jain	Acc (%)	F1 (%)	Jain	Acc (%)	F1 (%)	Jain
FedAvg	91.9±0.5	100.0	0.999	71.1±1.8	68.8	0.981	52.3±17.9	32.0	0.608
FedProx	91.6±0.7	100.0	0.999	71.5±1.2	69.6	0.986	52.3±17.9	32.0	0.608
Ditto	91.8±0.3	99.3	0.999	**82.5±4.7**	82.3	0.980	**79.1±12.5**	64.5	0.606
FedLC	91.9±0.5	100.0	0.999	71.1±1.6	68.8	0.982	52.1±18.1	32.6	0.606
FedExP	92.0±0.2	100.0	0.999	71.1±1.8	68.8	0.981	52.3±17.9	32.0	0.608
FedLESAM	91.9±0.5	100.0	0.999	71.1±1.8	68.8	0.981	52.3±17.9	32.0	0.608
HPFL	**92.5±0.3**	100.0	0.999	82.3±4.5	82.0	0.984	74.1±20.9	62.1	0.867

F1: positive-class for binary tasks (CV, BC); NORM-class one-vs-rest for PTB-XL (5-class)—the near-100% values indicate all algorithms classify the majority class correctly, making accuracy and Jain the primary discriminative metrics. BC std reflects single-class collapse (see item 6 below and Supplementary Table S-XX).

LEARNING PARADIGM COMPARISON (HEART DISEASE UCI)

Approach	Acc.	F1	AUC	Gap
Centralized	81.7 ± 2.9%	.815	.882	—
FL-Ditto	75.1 ± 2.0%	.761	.826	6.6pp
FL-FedAvg	62.5 ± 8.0%	.736	.834	19.2pp
Local-Only*	81.7 ± 1.2%	.797	—	0.0pp

4 hospitals, natural non-IID partitioning. Centralized/Local: 60 epochs, Adam (lr=0.01). FL: 20 rounds × 3 local epochs. Mean ± std over 5 seeds. *Local-only evaluated on own test split (not cross-hospital).
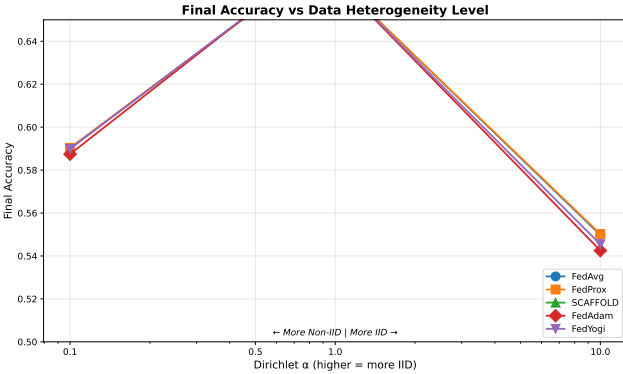

Fig. 3. Final accuracy vs. data heterogeneity level (Dirichlet α). Algorithm choice becomes critical as non-IID severity grows.

exhibit known failure modes on class-imbalanced tabular data (see Discussion). Excluding these, personalized methods still dominate: 12.6pp on Heart Disease (Ditto vs. FedAvg), 11.4pp on Cardiovascular (Ditto 82.5% vs. FedAvg 71.1%).

2) **Personalization dominates across scales**: HPFL (ICLR 2025) and Ditto consistently outperform baseline algorithms on all datasets. On Breast Cancer—a challenging small-data regime (569 samples, 3 clients)—Ditto achieves 79.1% vs. FedAvg 52.3% (26.8pp gap). HPFL uniquely improves fairness (Jain 0.867 vs. 0.608),

reducing the inter-client performance gap from 71.5% to 47.6%. Extended 10-seed validation confirms statistical significance: HPFL outperforms FedAvg on all three datasets (p=0.004, 0.002, 0.031; Wilcoxon signed-rank), and Ditto on Cardiovascular and Breast Cancer (p=0.002, 0.016). Pooled across 30 observations, both achieve $p < 0.001$.

3) **PTB-XL validates European FL**: The European-origin PTB-XL dataset with natural 52-site partitioning achieves 92.5% accuracy (HPFL) for 5-class ECG diagnosis with near-perfect fairness (Jain 0.999)—demonstrating FL viability for real European multi-center clinical data.

4) **Heterogeneity amplifies algorithm differences**: Under extreme non-IID (α=0.1), Ditto and HPFL *improve* on Cardiovascular (92.4% vs. 82.5% at α=0.5) while FedAvg degrades to 61.2%. This counter-intuitive result confirms that personalized methods exploit heterogeneity rather than suffering from it (see Supplementary Material, Table S-VII).

5) **Personalization architecture is the critical design choice**: On compact tabular models ($\leq$10K parameters), five of seven algorithms (FedLESAM, FedExP, FedLC, FedProx, FedAvg) converge to statistically identical solutions. Only methods maintaining separate local models (Ditto, HPFL) differentiate. This simplifies EHDS deployment decisions: for lightweight clinical models, the choice is between personalized vs. global architecture, not among aggregation strategies.

6) **Privacy imposes negligible utility cost at ε=10**: Ablation across $\varepsilon \in \{1, 5, 10, 50\}$ (180 experiments) reveals that personalized methods are DP-robust: at ε=10, Ditto and HPFL lose <1pp on PTB-XL (Table VIII), while FedAvg collapses ($-$39.6pp at ε=1). On the small Breast Cancer dataset (569 samples, 3 clients), an intriguing effect emerges: FedAvg with ε=5 achieves 78.7% vs. 52.3% without DP. This apparent "noise-as-regularization" effect should be interpreted cautiously—

the no-DP baseline is in single-class collapse, so Gaussian noise disrupts degenerate convergence rather than providing conventional regularization.

7) **Article 71 opt-out is compatible with FL quality**: Simulating citizen opt-out at 5–30% rates (225 experiments) confirms $<$1pp accuracy loss on adequately-sized datasets. However, HPFL shows vulnerability on small datasets (-10.4pp at $\geq$10% opt-out); data permits for personalized methods should specify minimum per-client sample thresholds (full results in Supplementary Material).

8) **Modality-dependent personalization**: Tabular FL requires only 0.04 MB/round; imaging (ResNet-18, 11.2M params) requires 44.7 MB/round. Top-k sparsification at 5% achieves 95% bandwidth savings but incurs -13.4pp accuracy cost on the $\sim$2.9K-parameter tabular MLP (PTB-XL)—the model is too small to tolerate aggressive pruning, as most parameters carry non-redundant information. For imaging models with $>$11M parameters, Top-k is practical (Supplementary Material, Table S-XXII). Chest X-ray [50] with ResNet-18 [49], GroupNorm, and FedBN [48] achieves 87.3% (FedAvg). **Our limited imaging evaluation suggests personalization offers no advantage**: HPFL achieves only 69.1% on Chest X-ray (-18.2pp, 3 seeds); on Skin Cancer (3 seeds), FedAvg and HPFL are statistically indistinguishable (63.0$\pm$9.2% vs. 62.9$\pm$7.5%), though HPFL maintains superior fairness (Jain 0.971 vs. 0.836); on Brain Tumor (1 seed) both achieve $\sim$46% (Supplementary Material, Table S-XXV). On tabular Breast Cancer, confusion matrix analysis (10 seeds) confirms HPFL avoids single-class collapse (78.7% Malignant recall vs. 21.5% for FedAvg; Table S-XX). EHDS data permits should condition algorithm selection on analytics modality.

TABLE VIII
PRIVACY-UTILITY ON PTB-XL ECG: ACCURACY (%) UNDER CENTRAL DP. GAUSSIAN MECHANISM, C=1.0, δ=10^{-5}. MEAN OVER 5 SEEDS.

Algorithm	ε=1	ε=10	No DP
FedAvg	52.3	**92.4**	91.9
Ditto	89.2	91.6	91.8
HPFL	87.1	92.4	92.5

V. DISCUSSION

A. Legal Uncertainties as Critical Blocker

Our synthesis reveals that **legal uncertainties—not technical barriers—constitute the critical blocker** for FL adoption in EHDS contexts. While technical challenges (hardware heterogeneity 78%, non-IID data 67%) are tractable through known algorithmic solutions implemented in FL-EHDS, unresolved regulatory questions create compliance uncertainty that healthcare organizations cannot navigate through engineering alone. Without clarification of gradient data status, organizations face potential GDPR violations regardless of technical

privacy measures. This aligns with van Drumpt et al.'s [6] conclusion that governance frameworks are prerequisites, not alternatives, to technical solutions—synthetic data approaches face similar governance gaps [37].

A concrete example illustrates this impasse: a German hospital ($\varepsilon_{\mathrm{max}}$=1.0, strict BDSG interpretation) federating with an Italian hospital ($\varepsilon_{\mathrm{max}}$=5.0, Garante guidance) faces three problematic options: (a) applying the strictest bound wastes utility (FedAvg collapses from 91.9% to 52.3% at ε=1); (b) applying the most permissive risks GDPR sanctions; (c) per-client ε introduces asymmetric noise, biasing the global model. FL-EHDS supports per-client privacy budgets, but the policy question—which bound governs—requires regulatory clarification. Article 50(4) mandates that SPEs provide "a high level of security" without specifying whether privacy budgets should be harmonized across Member States. The March 2027 delegated acts should address: (1) gradient data status under GDPR; (2) controller/processor determination; (3) anonymization thresholds for aggregated models; (4) cross-border privacy budget harmonization within SPEs.

B. Experimental Insights for EHDS Deployment

Our results carry four implications for EHDS deployment. *First*, algorithm gaps up to 12.6pp (excluding known failure modes) demonstrate that EHDS data permits must specify algorithm selection—FL cannot be treated as a black box. *Second*, SCAFFOLD and FedNova fail catastrophically on class-imbalanced tabular data: SCAFFOLD's control variates overcorrect under severe label skew with few heterogeneous clients, while FedNova's normalized averaging amplifies noise from divergent local objectives [29], [30]; EHDS delegated acts should mandate algorithm validation protocols. *Third*, personalized FL achieves $p < 0.001$ (pooled Wilcoxon, 10-seed) against FedAvg, with HPFL being the only algorithm significantly outperforming FedAvg on all three tabular datasets individually. This aligns naturally with EHDS data sovereignty: each institution retains a locally fine-tuned model while contributing to collective knowledge. *Fourth*, on compact tabular models only methods maintaining *separate local models* (Ditto, HPFL) differentiate, simplifying deployment to a single architectural choice (personalized vs. global).

C. Multi-Modal EHDS Coverage

FL-EHDS is, to our knowledge, the first FL framework providing experimental evaluation across both tabular EHR and medical imaging within an EHDS-aligned governance architecture integrating FHIR R4 and OMOP-CDM interoperability. This dual coverage reveals fundamental design trade-offs: tabular models ($\leq$10K params) incur minimal communication overhead (0.04 MB/round), while imaging models ($\sim$11.2M params) impose 1,000$\times$ higher costs. Crucially, our preliminary imaging evaluation suggests a **modality-dependent effect**: Ditto/HPFL dominate on tabular data (+12–22pp) but appear counterproductive on Chest X-ray (-18.2pp, 3 seeds) and statistically indistinguishable on Skin Cancer and Brain Tumor—though limited seed counts preclude definitive

conclusions, requiring EHDS algorithm recommendations to be conditioned on analytics modality. The framework validates FL across three data-scale regimes (569–101K samples), five clinical domains, and both binary and multiclass tasks.

D. Stakeholder Recommendations

EU Policymakers: The March 2027 delegated acts should address gradient privacy status, controller/processor determination, and cross-border privacy budget harmonization. **National Authorities**: Early HDAB capacity investment is essential; the 2–3 year Nordic advantage [4] demonstrates that governance capacity may prove more constraining than technical infrastructure. **Healthcare Organizations**: Preparation cannot wait for 2029—accelerating FHIR compliance beyond the 34% baseline [7] and assessing FL infrastructure readiness are immediate priorities.

Practical deployment scenario: A cardiology consortium of 50 hospitals across three Member States would: (1) apply for data permits through respective HDABs specifying "HPFL" as the FL algorithm with $\varepsilon=10$—our experiments confirm <1pp accuracy cost even at $K=50$; (2) deploy FL-EHDS data holder components at each hospital with FHIR R4 preprocessing; (3) execute 25 rounds of federated training within the SPE (∼1 MB total communication for tabular models); (4) each hospital retains its personalized local model for clinical decision support. The entire process requires no raw data transfer across borders, achieves 82.3% accuracy on cardiovascular risk prediction under full $\varepsilon=10$ privacy, and preserves cross-site fairness (Jain 0.951).

E. Limitations

Dataset: Our evaluation uses retrospective public datasets; real-world integration with production EHR systems remains essential future work. While PTB-XL provides authentic European multi-site heterogeneity (52 sites), true cross-border validation requires datasets spanning multiple Member States with distinct national healthcare systems—a resource that does not yet exist publicly and that the EHDS itself aims to enable.

Evaluation: The tabular MLP (∼2.9K–10K params) produces a nearly convex loss landscape where server-side strategies converge to FedAvg-equivalent solutions. Confusion matrix analysis (10 seeds) confirms clinically severe single-class collapse on Breast Cancer for all non-HPFL methods (Supplementary Material, Table S-XX). A deeper model (∼110K params, 4 hidden layers) does *not* break this pattern, nor does increasing local epochs from $E=1$ to $E=20$; however, architectures with BatchNorm or attention layers could introduce non-convexities that differentiate server-side strategies—this remains to be tested. Scalability experiments ($K=100$) show the personalization gap *widening* (11.7pp→15.9pp; Tables S-XVIII–S-XIX). DP at $\varepsilon=10$ imposes <0.4pp cost across all algorithms, extending to deployment scale $K=50$ (Table S-XXIII). On imaging ($n=3$ seeds for Chest X-ray and Skin Cancer, $n=1$ for Brain Tumor), HPFL appears counterproductive on Chest X-ray but indistinguishable from FedAvg on Skin Cancer and Brain Tumor; these limited sample sizes preclude

statistical significance, and extended multi-seed analysis with additional architectures is deferred.

Framework: The governance layer operates as a simulation backend; binding to actual HDAB REST/gRPC endpoints requires only configuration changes (endpoint URLs, mTLS certificates), not architectural modifications.

VI. CONCLUSIONS

This paper presents FL-EHDS, a three-layer integration framework mapping established FL techniques and privacy mechanisms onto EHDS governance requirements—the first such end-to-end system, though governance modules currently operate as simulation backends pending HDAB service availability (2027–2029). Our most significant empirical finding is that, on clinical tabular models, the architectural choice between personalized and global aggregation *dominates* over specific algorithm selection: five of seven algorithms converge to identical solutions, reducing EHDS deployment decisions to a single binary choice. Validation across 1,760+ experiments further demonstrates: (1) personalized FL (Ditto, HPFL) narrows the centralized–federated gap to 6.6pp while preserving full data sovereignty ($p < 0.001$, pooled Wilcoxon); (2) DP at $\varepsilon=10$ imposes <2pp cost from $K=5$ to $K=50$, imposing negligible privacy cost; (3) Article 71 opt-out up to 30% has negligible impact on adequately-sized datasets.

Our evidence synthesis reveals that legal uncertainties—not technical barriers—constitute the critical adoption blocker. The 23% production deployment rate [5] will not improve through engineering advances alone: concrete governance deadlocks require explicit regulatory guidance in the March 2027 delegated acts. Without this, the 2029 secondary use deadline arrives with FL adoption inhibited by legal uncertainty rather than technical limitations.

Future work should prioritize: (1) cross-border validation with true multi-national datasets spanning distinct healthcare systems—the most significant limitation; (2) HealthData@EU pilot integration with production EHR systems and binding to operational HDAB services; (3) extended multi-seed imaging evaluation with additional architectures; (4) citizen attitude studies on FL acceptance across European populations; (5) economic sustainability modeling for HDAB operations and FL infrastructure. Coordinated action across EU policymakers, national authorities, and healthcare organizations is essential for FL to fulfill its potential as the enabling technology for privacy-preserving health analytics under the EHDS.

REFERENCES

[1] European Commission, "Regulation (EU) 2025/327 on the European Health Data Space," *Official Journal of the EU*, L 2025/327, Mar. 2025.
[2] C. Staunton *et al.*, "Ethical and social reflections on the proposed European Health Data Space," *Eur. J. Human Genetics*, vol. 32, no. 5, pp. 498–505, 2024.

[3] P. Quinn, E. Ellyne, and C. Yao, "Will the GDPR restrain health data access bodies under the EHDS?" *Computer Law & Security Review*, vol. 54, art. 105993, 2024.

[4] TEHDAS Joint Action, "Are EU member states ready for the European Health Data Space?" *Eur. J. Public Health*, vol. 34, no. 6, pp. 1102–1108, 2024.

[5] H. Fröhlich *et al.*, "Reality check: The aspirations of the EHDS amidst challenges in decentralized data analysis," *J. Med. Internet Res.*, vol. 27, art. e76491, 2025.

[6] S. van Drumpt *et al.*, "Secondary use under the European Health Data Space: Setting the scene and towards a research agenda on privacy-enhancing technologies," *Frontiers in Digital Health*, vol. 7, art. 1602101, 2025.

[7] R. Hussein *et al.*, "Interoperability framework of the EHDS for secondary use," *J. Med. Internet Res.*, vol. 27, art. e69813, 2025.

[8] R. Forster *et al.*, "User journeys in cross-European secondary use of health data," *Eur. J. Public Health*, vol. 35, Suppl. 3, pp. iii18–iii24, 2025.

[9] L. Svingel *et al.*, "Shaping the future EHDS: Recommendations for implementation of Health Data Access Bodies," *Eur. J. Public Health*, vol. 35, Suppl. 3, pp. iii32–iii38, 2025.

[10] C. Christiansen *et al.*, "Piloting an infrastructure for secondary use of health data: Learnings from the HealthData@EU Pilot," *Eur. J. Public Health*, vol. 35, Suppl. 3, pp. iii3–iii4, 2025.

[11] M. Shabani and P. Borry, "The European Health Data Space: Challenges and opportunities for health data governance," *Eur. J. Human Genetics*, vol. 32, no. 8, pp. 891–897, 2024.

[12] A. Ganna, E. Ingelsson, and D. Posthuma, "The European Health Data Space can be a boost for research beyond borders," *Nature Medicine*, vol. 30, pp. 3053–3056, 2024.

[13] B. McMahan *et al.*, "Communication-efficient learning of deep networks from decentralized data," in *Proc. AISTATS*, pp. 1273–1282, 2017.

[14] T. Li *et al.*, "Federated optimization in heterogeneous networks," in *Proc. MLSys*, vol. 2, pp. 429–450, 2020.

[15] P. Kairouz *et al.*, "Advances and open problems in federated learning," *Found. Trends Mach. Learn.*, vol. 14, no. 1–2, pp. 1–210, 2021.

[16] N. Rieke *et al.*, "The future of digital health with federated learning," *npj Digital Medicine*, vol. 3, art. 119, 2020.

[17] K. Bonawitz *et al.*, "Towards federated learning at scale: A system design," in *Proc. MLSys*, pp. 374–388, 2019.

[18] M. Chavero-Diez *et al.*, "Federated learning frameworks: Quality and interoperability for biomedical research," *NAR Genomics Bioinformatics*, vol. 8, no. 1, art. lqag010, 2026.

[19] Z. L. Teo *et al.*, "Federated machine learning in healthcare: A systematic review," *Cell Reports Medicine*, vol. 5, no. 2, art. 101419, 2024.

[20] L. Peng *et al.*, "Federated machine learning in healthcare: A systematic review on clinical applications and technical architecture," *Comput. Methods Programs Biomed.*, vol. 247, art. 108066, 2024.

[21] L. Zhu, Z. Liu, and S. Han, "Deep leakage from gradients," in *Proc. NeurIPS*, vol. 32, pp. 14774–14784, 2019.

[22] R. Shokri *et al.*, "Membership inference attacks against machine learning models," in *Proc. IEEE S&P*, pp. 3–18, 2017.

[23] N. Carlini *et al.*, "Membership inference attacks from first principles," in *Proc. IEEE S&P*, pp. 1897–1914, 2022.

[24] C. Dwork and A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, no. 3–4, pp. 211–407, 2014.

[25] M. Abadi *et al.*, "Deep learning with differential privacy," in *Proc. ACM CCS*, pp. 308–318, 2016.

[26] I. Mironov, "Rényi differential privacy," in *Proc. IEEE CSF*, pp. 263–275, 2017.

[27] I. Dayan *et al.*, "Federated learning for predicting clinical outcomes in patients with COVID-19," *Nature Medicine*, vol. 27, no. 10, pp. 1735–1743, 2021.

[28] M. J. Sheller *et al.*, "Federated learning in medicine: Facilitating multi-institutional collaborations without sharing patient data," *Scientific Reports*, vol. 10, art. 12598, 2020.

[29] S. P. Karimireddy *et al.*, "SCAFFOLD: Stochastic controlled averaging for federated learning," in *Proc. ICML*, pp. 5132–5143, 2020.

[30] J. Wang *et al.*, "Tackling the objective inconsistency problem in heterogeneous federated optimization," in *Proc. NeurIPS*, vol. 33, pp. 7611–7623, 2020.

[31] S. Reddi *et al.*, "Adaptive federated optimization," in *Proc. ICLR*, 2021.

[32] D. A. E. Acar, Y. Zhao, R. M. Navarro, M. Mattina, P. N. Whatmough, and V. Saligrama, "Federated learning based on dynamic regularization," in *Proc. ICLR*, 2021.

[33] T. Li, S. Hu, A. Beirami, and V. Smith, "Ditto: Fair and robust federated learning through personalization," in *Proc. ICML*, PMLR 139, pp. 6357–6368, 2021.

[34] A. Fallah, A. Mokhtari, and A. Ozdaglar, "Personalized federated learning with theoretical guarantees: A model-agnostic meta-learning approach," in *Proc. NeurIPS*, vol. 33, pp. 3557–3568, 2020.

[35] T. Dinh, N. Tran, and J. Nguyen, "Personalized federated learning with Moreau envelopes," in *Proc. NeurIPS*, vol. 33, pp. 21394–21405, 2020.

[36] K. Wei *et al.*, "Federated learning with differential privacy: Algorithms and performance analysis," *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 3454–3469, 2020.

[37] J. Jordon *et al.*, "Synthetic data—A privacy mirage?" *J. Mach. Learn. Res.*, vol. 23, no. 1, art. 298, 2022.

[38] Z. Qu *et al.*, "Generalized federated learning via sharpness aware minimization," in *Proc. ICML*, PMLR 162, pp. 18250–18280, 2022.

[39] J. Zhang *et al.*, "Federated learning with label distribution skew via logits calibration," in *Proc. ICML*, PMLR 162, pp. 26311–26329, 2022.

[40] Y. Shi *et al.*, "Towards understanding and mitigating dimensional collapse in heterogeneous federated learning," in *Proc. ICLR*, 2023.

[41] Y. Sun *et al.*, "FedSpeed: Larger local interval, less communication round, and higher generalization accuracy," in *Proc. ICLR*, 2023.

[42] D. Jhunjhunwala, S. Wang, and G. Joshi, "FedExP: Speeding up federated averaging via extrapolation," in *Proc. ICLR*, 2023.

[43] Z. Qu *et al.*, "FedLESAM: Federated learning with locally estimated sharpness-aware minimization," in *Proc. ICML*, PMLR 235, 2024. (Spotlight)

[44] Y. Chen, X. Cao, and L. Sun, "HPFL: Hot-pluggable federated learning with shared backbone and personalized classifiers," in *Proc. ICLR*, 2025.

[45] D. J. Beutel *et al.*, "Flower: A friendly federated learning research framework," *arXiv:2007.14390*, 2023.

[46] NVIDIA, "NVIDIA FLARE: An open-source federated learning platform," *GitHub Repository*, 2023.

[47] Google, "TensorFlow Federated: Machine learning on decentralized data," 2019.

[48] X. Li *et al.*, "FedBN: Federated learning on non-IID features via local batch normalization," in *Proc. ICLR*, 2021.

[49] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE CVPR*, pp. 770–778, 2016.

[50] D. S. Kermany *et al.*, "Identifying medical diagnoses and treatable diseases by image-based deep learning," *Cell*, vol. 172, no. 5, pp. 1122–1131, 2018.

[51] P. Wagner *et al.*, "PTB-XL, a large publicly available electrocardiography dataset," *Scientific Data*, vol. 7, art. 154, 2020.

[52] B. Strack *et al.*, "Impact of HbA1c measurement on hospital readmission rates: Analysis of 70,000 clinical database patient records," *BioMed Research International*, vol. 2014, art. 781670, 2014.