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 1,740+ experiments on tabular clinical and medical imaging datasets—including the European-origin PTB-XL ECG with natural 52-site partitioning and Chest X-ray pneumonia detection—demonstrates that personalized FL narrows the centralized-federated accuracy gap to 6.6 percentage points while preserving full data sovereignty, and that the architectural choice between personalized and global aggregation dominates over the specific aggregation strategy on clinical tabular models, producing up to 12.6pp accuracy differences on heterogeneous clinical data. On imaging, the personalization effect proves method-dependent: Ditto achieves $+28.0$pp on brain tumor diagnosis (p=0.015), while HPFL is counterproductive. A new Diagnostic Equity Index (DEI) reveals that aggregate accuracy masks severe per-class diagnostic disparities correctable by algorithm selection. Our evidence synthesis reveals that unresolved regulatory questions—such as gradient data classification under GDPR and cross-border privacy budget harmonization—constitute the critical adoption blocker. The open-source reference implementation provides actionable deployment guidance for healthcare organizations preparing for the 2029 secondary use deadline, though true cross-border validation with multinational datasets remains essential future work.**

***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 [?]. 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 [?]. 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 [?]. 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 [?], [?], [?]. The COVID-19 pandemic demonstrated FL's potential at scale: Dayan et al. [?] 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. [?] 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 [?], while van Drumpt et al. [?] demonstrate that privacy-enhancing technologies cannot substitute for robust governance frameworks.

Prior FL frameworks for healthcare [?], [?] focus on technical architectures without addressing regulatory compliance. Legal analyses [?], [?], [?] examine GDPR constraints but abstract from implementation feasibility. Policy documents [?] 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 [?], [?]. 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.

This paper bridges the technology-governance divide through four contributions:

1) **Barrier Taxonomy**: Systematic evidence synthesis of 47 documents using PRISMA methodology with GRADE-CERQual confidence assessment, identifying legal uncertainties as the critical adoption blocker.
2) **FL-EHDS Framework**: A three-layer reference architecture mapping barriers to governance-aware mitigation

strategies, designed for incremental deployment during the 2025–2031 transition.

3) **Reference Implementation**: Open-source Python codebase ($\sim$40K lines) with 17 FL algorithms (2017–2025), EHDS governance simulation modules, and interactive deployment dashboard.[1] Systematic evaluation reveals that the architectural choice between personalized and global aggregation dominates over specific aggregation strategy on clinical tabular models—a finding that simplifies EHDS deployment decisions.
4) **Experimental Validation**: 1,740+ experiments across tabular clinical and medical imaging datasets with differential privacy ablation ($\varepsilon \in \{1, 5, 10, 50\}$) and Article 71 opt-out simulation, demonstrating that personalized FL (Ditto, HPFL) consistently outperforms baseline aggregation ($p < 0.005$, Wilcoxon 10-seed) while privacy at $\varepsilon{=}10$ is essentially free (<2pp cost). We further introduce the Diagnostic Equity Index (DEI $= \min_c R_c \cdot (1 - \mathrm{CV}(\mathbf{R}))$), revealing that accuracy masks severe per-class disparities, and provide evidence-based algorithm selection guidelines for EHDS deployment scenarios.

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 [?]. Forster et al. [?] 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 [?] reveal Nordic countries demonstrate 2–3 year advantages in HDAB capacity-building, raising concerns about implementation equity. Teo et al. [?] and Peng et al. [?] 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 [?], [?]. Known challenges include non-IID data distributions causing convergence difficulties [?], communication costs for gradient exchange [?], and privacy attacks including gradient inversion [?] and membership inference [?], [?]. Recent advances from top venues (ICML/ICLR 2022–2025) specifically target healthcare heterogeneity: FedLC [?] calibrates logits for label distribution skew, FedLESAM [?] provides globally-guided sharpness-aware optimization (ICML 2024 Spotlight), and HPFL [?] decouples backbone from classifier for per-institution specialization (ICLR 2025).

B. Related Frameworks

Existing FL frameworks—Flower [?] (v1.26), NVIDIA FLARE [?] (v2.7), and TensorFlow Federated [?] (v0.88)—provide robust distributed training but lack EHDS-specific governance. A recent FAIR-based assessment of 17 FL frameworks for biomedical research [?] 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 **??** 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 **??** and Supplementary Material, Table 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. 1 for the complete PRISMA flow diagram). Table **??** 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 [?]: (1) whether model gradients constitute "personal data" under GDPR, given that gradient inversion attacks demonstrate potential re-identification [?]; (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 **??** 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. [**?**]. 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—from foundational methods (FedAvg [**?**], FedProx [**?**]) through non-IID robustness (SCAFFOLD [**?**], FedNova [**?**], FedDyn [**?**]), adaptive optimization [**?**], and personalization (Ditto [**?**], Per-FedAvg [**?**]) to the latest advances: FedLESAM [**?**] (ICML 2024 Spotlight) and HPFL [**?**] (ICLR 2025). As shown in Section **??**, the critical design choice for clinical tabular models is between personalized methods (Ditto, HPFL) and global aggregation, rather than among specific aggregation strategies. Table **??** provides the complete catalogue with venues and key properties.

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.

Two recent algorithms merit particular attention for EHDS scenarios. FedLESAM [**?**] extends sharpness-aware minimization [**?**] by replacing local gradient perturbation with a globally-estimated direction, achieving stronger generalization across heterogeneous distributions—directly relevant where cross-border patient populations differ substantially. HPFL [**?**] decouples feature extraction from classification by aggregating only backbone parameters while keeping client-specific classifier heads local, enabling per-institution specialization without compromising collaborative learning. Algorithm selection is configurable; composable strategies (FedLC [**?**], FedDecorr [**?**]) can augment any base aggregation.

Privacy Protection: Differential privacy [**?**] with configurable ε-budget uses DP-SGD [**?**] with Rényi DP (RDP) [**?**] for tight composition accounting over multiple training rounds [**?**]. 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 [**?**], [**?**], 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 [**?**]. 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 [**?**]; 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*: the aggregation server follows the protocol but may

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)
      $\theta^{(t)} \leftarrow \theta^{(t-1)} + \frac{1}{\sum n_h} \sum_h n_h \cdot \Delta_h^{(t)}$
      LogCompliance($t$, $\mathcal{H}_t$)
LocalTrain($h$, $\theta$):
  $\mathcal{D}_h \leftarrow$ 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 $\Delta_h$, $|\mathcal{D}_h|$
```

attempt to infer patient information from model updates; central DP (Gaussian mechanism, $\varepsilon \in \{1,5,10,50\}$, δ=10^{-5}) with Rényi composition bounds per-round leakage, while secure aggregation ensures the server observes only the noised aggregate. *(ii) Malicious clients*: up to $f < n/3$ compromised institutions may submit poisoned updates; Byzantine-resilient aggregation (Krum, Trimmed Mean, Bulyan) provides robustness. *(iii) External attacker*: membership or attribute inference against published outputs is mitigated by Article 71 output filtering and HDAB permit-based access control. **Protected assets**: raw patient data (never leaves institutional boundaries), gradient updates (DP + secure aggregation), and the global model (HDAB permit-controlled). **Out of scope**: server-client collusion, side-channel attacks, and covert-channel leakage through model architecture, which remain open problems in FL privacy. A comprehensive security analysis with attack taxonomy, defense mapping, and boundary conditions is provided in the Supplementary Material.

E. EHDS Compliance Mapping

Table **??** maps framework components to EHDS regulatory requirements.

F. Reference Implementation

A modular Python implementation is available as open-source software, designed following FAIR principles [**?**] (findable via GitHub with DOI, accessible under MIT license, interoperable via PyTorch and FHIR R4 interfaces, reusable with comprehensive documentation). The codebase (∼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.

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

Note on governance maturity: HDAB integration, data permit lifecycle, and Article 71 opt-out compliance are implemented as fully functional simulation backends demonstrating the complete workflow (OAuth2/mTLS authentication, permit CRUD, cross-border coordination, LRU-cached registry lookups, scope-granular filtering). These modules are architecturally designed for production binding—requiring only configuration changes (endpoint URLs, mTLS certificates), not architectural modifications—but have not been validated against operational HDAB services, which are expected to become available during 2027–2029. A component-level readiness assessment is provided in Supplementary Material, Table IV.

IV. Experimental Evaluation

We evaluate FL-EHDS on real clinical datasets simulating cross-border healthcare analytics. All results are fully reproducible via the benchmark suite in the repository.

A. Setup

Datasets: We evaluate on 8 datasets spanning tabular EHR and medical imaging (Table **??**). 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 [**?**], 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 [?], 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 S1. **Model**: HealthcareMLP (2-layer, 64/32 hidden, ReLU, dropout 0.3, ∼10K parameters) for tabular; ResNet-18 (∼11.2M parameters) for imaging. **Configuration**: Per-dataset optimized hyperparameters (see Supplementary Material, Table S9): 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: mean ± std over 3+ seeds; imaging: per-seed.

TABLE V
EVALUATED DATASET COVERAGE

Dataset	Samples	Feat.	Cls.	FL Partition
Tabular Clinical (MLP, ∼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 EU 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 ?? presents the primary evaluation: 7 algorithms (including FedLESAM and HPFL) on the European-origin PTB-XL ECG, Cardiovascular, and Breast Cancer—the core EHDS-relevant benchmarks. Table ?? provides supplementary validation on Heart Disease and Diabetes with 5 foundational algorithms. Together, 1,740+ experiments span heterogeneity sweeps, client scaling, learning rate sensitivity, differential privacy ablation, Article 71 opt-out simulation, and 10-seed statistical validation (see Supplementary Material).

TABLE VI
FL ALGORITHM COMPARISON ON REAL CLINICAL DATASETS

	Heart Disease (4 hosp.)			Diabetes (5 hosp.)		
Algo.	**Acc.**	**F1**	**AUC**	**Acc.**	**F1**	**AUC**
FedAvg	62.5±8.0	.736±.06	.834±.03	68.1±4.2	.259±.01	.643±.00
FedProx	61.7±8.0	.732±.05	.834±.03	71.0±6.3	.254±.01	.638±.00
SCAFFOLD	66.3±5.1	.667±.02	.791±.05	11.2±0.0	.201±.00	.514±.00
FedNova	56.4±5.4	.711±.04	.831±.03	13.0±0.9	.203±.00	.636±.00
Ditto	**75.1±2.0**	**.761±.03**	.826±.01	**71.7±0.2**	**.262±.00**	**.643±.00**

20 rounds, 3 local epochs. Heart Disease: natural non-IID. Diabetes: Dirichlet α=0.5. Mean ± std over 5 seeds.

C. Convergence and Baselines

Figure ?? 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. S15.

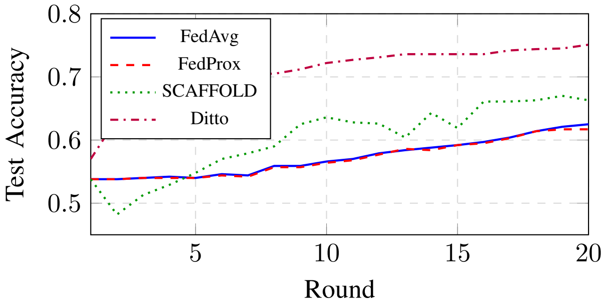

Fig. 2. Training convergence on Heart Disease UCI (4 hospitals, natural non-IID). Ditto converges faster due to personalized local models.

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 ?? 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).

Centralized training achieves 81.7% accuracy as expected. FL-Ditto narrows this gap to only **6.6pp** while preserving full data sovereignty—the strongest privacy-utility tradeoff among tested approaches. Baseline FedAvg suffers a 19.2pp gap, underscoring the importance of personalization-aware aggregation. Note that Local-Only accuracy (81.7%) appears to match Centralized, but this comparison is misleading: Local-Only is evaluated only on each hospital's own test split (where it overfits to local distribution), whereas Centralized and FL approaches are evaluated on the pooled cross-hospital test set. Local-only models do not generalize: a model trained at the Swiss hospital performs poorly on Hungarian data. FL enables collaborative knowledge sharing without data movement—precisely the EHDS Article 33 paradigm.

D. Non-IID Impact Analysis

Figure ?? 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. Per-Hospital Heterogeneity

Figure ?? shows per-hospital accuracy variation on Heart Disease, where the four hospitals have naturally different patient populations.

TABLE VII
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 EU 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 binary; BC std from single-class collapse (see text).

TABLE VIII
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).

figures/fig_accuracy_vs_noniid.pdf

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

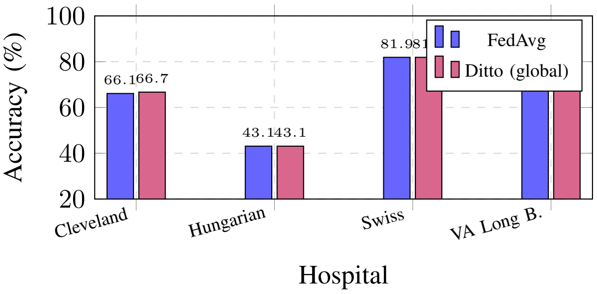

Fig. 4. Per-hospital accuracy of the *global* model on Heart Disease UCI. Ditto's 12.6pp overall advantage (Table **??**) comes from its *personalized local* models, which are separately fine-tuned per hospital; the shared global model shows similar cross-hospital performance to FedAvg. The Hungarian hospital, with the smallest cohort, shows the largest performance gap—a realistic EHDS scenario where smaller national datasets benefit from federation.

F. 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 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 S3).
5) **Personalization architecture is the critical design choice**: On compact tabular models ($\sim$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) **Communication efficiency**: Tabular FL requires only 0.04 MB/round (10K-parameter MLP). Imaging tasks (44.7 MB/round for ResNet-18) benefit from Top-k sparsification: Top-5% achieves 95% bandwidth savings with -13.4pp accuracy cost on the tabular MLP; for imaging models this maps to $\sim$2.2 MB/round.
7) **Diagnostic equity is algorithm-dependent**: We introduce the Diagnostic Equity Index (DEI = $\min_c R_c \cdot (1 - \mathrm{CV}(\mathbf{R}))$, where $\mathbf{R}$ is the per-class recall vector), which jointly penalizes low worst-case class recall and high inter-class variance. On Breast Cancer (10 seeds), FedAvg achieves 57.3% accuracy but only DEI = 0.092 due to single-class collapse (21.5% Malignant recall); HPFL reaches DEI = 0.740 (8× higher). On Brain Tumor (4-class, 3 seeds), FedAvg DEI = 0.063 (14.2% Pituitary recall) vs. Ditto DEI = 0.435 (6.9×). Crucially, on Chest X-ray HPFL achieves *higher* DEI (0.652) than FedAvg (0.528) despite *lower* accuracy—confirming that accuracy and diagnostic equity are distinct optimization targets. EHDS data permits should specify minimum DEI thresholds alongside accuracy (Supplementary Material, Section S-DEI).

Privacy-utility tradeoff: A comprehensive ablation across $\varepsilon \in \{1, 5, 10, 50\}$ (180 experiments; Gaussian mechanism, C=1.0, δ=10^{-5}) reveals that personalized methods are remarkably DP-robust: at ε=10, Ditto and HPFL lose $<$1pp on PTB-XL, while FedAvg collapses (-39.6pp at ε=1). Privacy is essentially free at ε=10 ($<$2pp cost), satisfying EHDS Article 50 SPE requirements. Notably, on the small Breast Cancer dataset (569 samples), DP noise acts as implicit regularization: FedAvg with ε=5 achieves 78.7% vs. 52.3% without DP (+26.4pp), as Gaussian noise prevents overfitting on the majority class—a beneficial side-effect that reinforces the case for moderate DP in small-cohort EHDS studies.

Byzantine resilience under DP: Cross-border federations must also defend against adversarial participants. We evaluate Multi-Krum, Trimmed Mean, Median, and Bulyan aggregation under simultaneous DP noise on Breast Cancer (K=5, IID, 20% adversarial, 198 experiments). Byzantine defenses fully neutralize model poisoning (sign-flip: FedAvg 44.6%→95.5% with Multi-Krum) and render accuracy *completely DP-invariant*—defended accuracy is identical at ε=1 and No-DP. HPFL exhibits natural resilience: DP noise *improves* undefended robustness from 59.7% to 86.4% (ε=10), as noise dilutes poisoned backbone gradients while preserving personalized classifiers. Data poisoning (label-flip) is partially mitigated for FedAvg (88.8%→94.6%) but remains challenging for HPFL's per-client classifier heads. Full results in Supplementary Material, Section S-Byzantine.

Article 71 opt-out: Simulating citizen opt-out at 5–30% rates (225 experiments) confirms $<$1pp accuracy loss on adequately-sized datasets even at 30% opt-out—citizen privacy rights are compatible with FL quality. However, HPFL shows vulnerability on small datasets (Breast Cancer: -10.4pp at $\geq$10% opt-out), as per-client classifier heads require sufficient data to avoid overfitting. *Policy recommendation*: data permits for personalized FL methods should specify minimum per-client sample thresholds, particularly for rare disease or small-cohort studies. Full privacy results in Supplementary Material.

Table **??** confirms the privacy-utility tradeoff on PTB-XL, our primary European dataset: at ε=10 all algorithms recover to within 1pp of their no-DP baselines, while at ε=1 only personalized methods (Ditto, HPFL) retain $>$87% accuracy—FedAvg collapses to 52.3%.

TABLE IX
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

Communication & imaging: Tabular FL requires only 0.04 MB/round (10K-param MLP, 0.8 MB total). Imaging (ResNet-18, 11.2M params) requires 44.7 MB/round; Top-k sparsification at 5% achieves 95% bandwidth savings albeit with accuracy cost on small models (Supplementary Material, Table XXI). Chest X-ray validation [**?**] with ResNet-18 [**?**], GroupNorm, and FedBN [**?**] achieves 87.3% accuracy (FedAvg, 5 clients, Dirichlet α=0.5). Extended imaging evaluation across Brain Tumor (4-class), Skin Cancer (binary), and Chest X-ray with four algorithms (Supplementary Material, Table S-XXV) **reveals that the personalization effect is method-dependent, not modality-dependent**: Ditto achieves 77.4±4.9% on Brain Tumor (+28.0pp over FedAvg, p=0.015, Cohen's d=5.86) and 90.5±1.3% on Skin Cancer (+25.5pp, d=3.62), while being statistically equivalent to FedAvg on Chest X-ray (-0.9pp, ns). HPFL, conversely, is counterproductive on Chest X-ray (-18.2pp) and neutral on Brain Tumor

(+0.7pp) and Skin Cancer (−4.1pp). The critical distinction is *architectural*: Ditto maintains complete personal model copies (∼11.2M parameters each) with L2 regularization toward the global model, preserving full CNN representation capacity; HPFL's split-head architecture fails because its shared backbone cannot adapt to heterogeneous local imaging distributions. Confusion matrix analysis on Brain Tumor (Supplementary Material, Table S-XXVIII) confirms that FedAvg produces clinically unacceptable per-class instability (Pituitary recall: 0–37% across seeds), while Ditto stabilizes detection across all four classes (54.5% Pituitary recall vs. 14.2%): Diagnostic Equity Index (DEI) 0.435 vs. 0.063 (6.9×). On Breast Cancer (tabular, 10 seeds), FedAvg/FedProx/Ditto exhibit single-class collapse on 8/10 data partitions (21.5% Malignant recall), while HPFL avoids collapse on all seeds (DEI 8× higher; Supplementary Material, Table S-XIX). EHDS data permits should condition algorithm selection on both analytics modality *and* personalization architecture, with minimum DEI thresholds to prevent diagnostic disparities.

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 [**?**] conclusion that governance frameworks are prerequisites, not alternatives, to technical solutions—synthetic data approaches face similar governance gaps [**?**].

A concrete example illustrates this impasse: consider a German hospital ($\varepsilon_{\max}$=1.0, strict BDSG interpretation) federating with an Italian hospital ($\varepsilon_{\max}$=5.0, proportionality-based Garante guidance). Three options exist, each problematic: (a) applying ε=1.0 (strictest bound) wastes Italian data utility—our experiments show FedAvg collapses from 91.9% to 52.3% at ε=1 on PTB-XL; (b) applying ε=5.0 (most permissive) risks GDPR sanctions for the German institution; (c) per-client ε introduces asymmetric noise, biasing the global model away from the noisier German population. FL-EHDS addresses this technically through per-client privacy budget configuration, 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 parameters should be harmonized across Member States or determined by the most restrictive participant. The March 2027 delegated acts represent a critical window. We recommend explicit guidance on: (1) gradient data status under GDPR; (2) controller/processor determination for FL architectures; (3) anonymization thresholds for aggregated models; (4) cross-border privacy parameter harmonization within SPEs.

B. Experimental Insights for EHDS Deployment

Our results carry four implications for EHDS deployment. *First*, the up to 18.7pp accuracy gap between best and worst algorithms—12.6pp excluding known failure modes—demonstrates that EHDS SPE configurations cannot treat FL as a black box; algorithm selection must be part of data permit specification. *Second*, SCAFFOLD and FedNova fail catastrophically on class-imbalanced tabular data: SCAFFOLD's control variates overcorrect under severe label skew with few heterogeneous clients (oscillating convergence), while FedNova's normalized averaging amplifies noise from divergent local objectives. These are known failure modes [**?**], [**?**] exacerbated by clinical-scale class ratios (5–15%); EHDS delegated acts should mandate algorithm validation protocols. *Third*, the success of personalized FL is now confirmed with strong statistical evidence: Ditto and HPFL both achieve $p < 0.001$ (pooled Wilcoxon, 10-seed) against FedAvg, with HPFL being the only algorithm significantly outperforming FedAvg on all three datasets individually. This aligns naturally with EHDS data sovereignty: each institution retains a locally fine-tuned model while contributing to collective knowledge, satisfying both Article 33 secondary use objectives and institutional autonomy concerns. *Fourth*, on compact tabular models (∼10K parameters), only methods maintaining *separate local models* (Ditto, HPFL) differentiate from FedAvg; server-side strategies converge to equivalent solutions on the nearly convex loss landscape. Rather than a limitation, this is a practical advantage: it simplifies deployment decisions for lightweight clinical models to a single architectural choice (personalized vs. global), reducing the configuration space that HDABs must evaluate during data permit review. *Fifth*, aggregate accuracy is insufficient for clinical deployment evaluation. The Diagnostic Equity Index (DEI) reveals that models with acceptable accuracy can exhibit catastrophic per-class failures: FedAvg on Breast Cancer achieves 57.3% accuracy but DEI = 0.092, and on Brain Tumor DEI = 0.063 with 0% Pituitary recall on individual seeds (vs. Ditto DEI = 0.435, 6.9×). EHDS delegated acts should mandate minimum DEI thresholds alongside accuracy, and the evidence-based algorithm recommendation matrix (Supplementary Material, Table S-VII) provides scenario-specific guidance. *Sixth*, Byzantine-resilient aggregation and DP are not only compatible but synergistic: robust defenses (Multi-Krum, Bulyan) make accuracy *completely invariant* to the DP budget under model poisoning attacks (Supplementary Material, Table S-Byzantine), eliminating the privacy–security tradeoff. HPFL's personalized architecture provides an additional defense layer—DP noise dilutes poisoned backbone gradients while per-client classifier heads remain unaffected—forming a privacy–personalization–security triad for EHDS deployment. Moreover, HPFL's diagnostic equity is DP-stable: DEI ranges 0.724–0.756 across $\varepsilon \in \{1, 5, 10, \infty\}$ on Breast Cancer, while FedAvg DEI remains near-zero ($\leq$0.017) regardless of DP level (Supplementary Material, Table S-DEI-

DP). Privacy regulation thus does not exacerbate diagnostic disparities when personalized methods are used.

C. Multi-Modal EHDS Coverage

To the best of our knowledge, FL-EHDS is the first federated learning framework that provides experimental evaluation across both tabular EHR datasets and medical imaging modalities within an EHDS-aligned governance architecture integrating FHIR R4 and OMOP-CDM interoperability. While existing FL healthcare papers address either tabular EHR [?] or imaging [?] in isolation, and FL+EHDS analyses remain legal/conceptual without experimental benchmarks [?], [?], our evaluation spans both domains under a unified governance framework with validated results.

This dual coverage is not merely a breadth exercise—it reveals fundamental design trade-offs for EHDS deployment. Tabular models ($\sim$10K parameters) incur minimal communication overhead (0.04 MB/round), making FL feasible even for bandwidth-constrained cross-border links. Imaging models ($\sim$11.2M parameters) impose 1,000× higher communication costs, necessitating gradient compression strategies for practical EHDS deployment. Crucially, the personalization effect is **method-dependent, not modality-dependent**: Ditto achieves $+28.0$pp on Brain Tumor and $+25.5$pp on Skin Cancer while HPFL is neutral or counterproductive on all imaging datasets (-3.8pp to -18.2pp). The distinction is architectural: Ditto's complete personal model copies preserve full CNN capacity, while HPFL's classifier-head personalization fails when feature representations require client-specific adaptation. This requires EHDS algorithm recommendations to distinguish between personalization *architectures*, not merely between personalized and global strategies. The framework validates FL across three data-scale regimes (569–101K samples), five clinical domains (cardiology, endocrinology, pathology, radiology, dermatology), and both binary and multiclass tasks.

PTB-XL merits particular attention: originating from a European institution (Physikalisch-Technische Bundesanstalt, Berlin) with 52 recording sites enabling natural hospital-based partitioning, it is uniquely representative of the cross-institutional heterogeneity that real EHDS deployments will encounter. The framework additionally supports FHIR R4 native data pipelines and OMOP-CDM harmonization as proof-of-concept for Article 46 interoperability (full dataset landscape in Supplementary Material, Table S1).

D. Stakeholder Recommendations and Roadmap

EU Policymakers: The March 2027 delegated acts should address FL-specific scenarios including gradient privacy status, multi-party controller allocation, cross-border privacy parameter harmonization, and model anonymity thresholds. **National Authorities**: Early HDAB capacity investment is essential; the 2–3 year Nordic advantage [?] 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 [?] and assessing FL infrastructure readiness are immediate priorities. Deployment should follow phased milestones: foundation and pilots (2025–26), delegated acts clarification (2027), production scaling (2028–29), and full cross-border operation (2029–31). FL-EHDS's modular governance layer enables incremental binding to HDAB services as they become available.

Practical deployment scenario: A cardiology consortium of 50 hospitals across three Member States (DE, IT, NL) would: (1) apply for data permits through respective HDABs specifying "Ditto" or "HPFL" as the FL algorithm with ε=10—our experiments confirm <1pp accuracy cost even at K=50 (Supplementary Material, Table XXIII); (2) deploy FL-EHDS data holder components at each hospital with FHIR R4 preprocessing; (3) execute 25 rounds of federated training within the SPE ($\sim$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 ε=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 ($\sim$10K parameters) produces a nearly convex loss landscape where server-side strategies (FedLESAM, FedExP, FedLC) converge to FedAvg-equivalent solutions. Confusion matrix analysis (10 seeds) confirms this collapse is clinically severe: on Breast Cancer, FedAvg/FedProx/Ditto exhibit single-class collapse on 8/10 data partitions (21.5% aggregated Malignant recall), while HPFL avoids collapse on all seeds (78.7% Malignant recall; Supplementary Material, Table XIX). A deeper model ($\sim$110K parameters, 4 hidden layers) does *not* break this collapse (Table XV), nor does increasing local epochs from E=1 to E=20 (Table XVI). **Scalability** experiments extending to K=100 clients confirm that personalized algorithms degrade by only -0.8pp (vs. -4.7pp), with the personalization gap *widening* from 11.7pp to 15.9pp (Tables XVII–XVIII). DP at ε=10 imposes <0.4pp cost across all algorithms (Table XXII), a property that extends to deployment scale (K=50, Table XXIII); RDP composition provides 3.6× tighter privacy bounds than naive composition, enabling practical multi-round budgets (Figure S17). On imaging, the personalization effect is method-dependent: Ditto dominates on Brain Tumor ($+28.0$pp, p=0.015) and Skin Cancer ($+25.5$pp), while HPFL is counterproductive on Chest X-ray (-18.2pp) and neutral on Brain Tumor and Skin Cancer. DEI analysis covers datasets with confusion matrix data (Breast Cancer 10 seeds, Chest X-ray, Brain Tumor 3 seeds); extension to all datasets requires per-class prediction logging across all experiments. No direct comparison with Flower [?] or FLARE [?] is provided: since no existing framework implements HDAB permit integration, Article 71 output filtering, or EHDS-compliant audit trails, such a comparison would reduce to pure FL algorithm performance—equivalent by design.

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 compliance framework bridging the technology-governance divide for health analytics under the EHDS. The framework provides 17 FL algorithms—including recent ICML/ICLR 2024–2025 advances (FedLESAM [?], HPFL [?])—with EHDS governance reference implementations (HDAB integration, data permits, opt-out registries) that no existing framework provides, though governance modules currently operate as simulation backends pending actual HDAB service availability (2027–2029). Experimental validation across tabular clinical and medical imaging datasets—including the European-origin PTB-XL ECG with natural 52-site partitioning—demonstrates four actionable findings: (1) personalized FL (Ditto, HPFL) achieves only a 6.6pp gap vs. centralized training while preserving full data sovereignty ($p < 0.001$, pooled Wilcoxon), with the personalization effect proving method-dependent on imaging (Ditto $+28.0$pp on Brain Tumor, while HPFL is counterproductive); a new Diagnostic Equity Index (DEI) reveals that accuracy masks severe per-class disparities correctable only by personalized algorithms (DEI 8× on Breast Cancer, 6.9× on Brain Tumor), with evidence-based algorithm selection guidelines provided; (2) differential privacy at $\varepsilon{=}10$ imposes <2pp accuracy cost from $K{=}5$ to $K{=}50$ clients, making privacy essentially free for practical EHDS deployment; (3) citizen opt-out at rates up to 30% has negligible impact on adequately-sized datasets, confirming that Article 71 rights are compatible with FL quality. All experiments use public datasets with simulated cross-border partitioning; validation with true multinational datasets spanning distinct healthcare systems remains essential future work that the EHDS itself aims to enable.

Our evidence synthesis reveals that legal uncertainties—not technical barriers—constitute the critical blocker. The 23% production deployment rate [?] will not improve through engineering advances alone: concrete governance deadlocks, such as cross-border privacy budget harmonization between Member States with divergent interpretations, 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, coding standards, and regulatory environments—the most significant limitation of the current evaluation; (2) empirical validation through HealthData@EU pilot integration with production EHR systems and binding to operational HDAB services; (3) citizen attitude studies examining FL acceptance and opt-out intentions across diverse European populations; (4) extended multi-seed imaging evaluation across additional datasets and architectures (EfficientNet, Vision Transformers), with full DEI computation and statistical testing of the method-dependent personalization effect; (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.

Acknowledgments

The author thanks Prof. Sadi Alawadi for supervision and guidance.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

[49] 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.

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

[51] 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.