Supplementary Material: 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—**This document provides supplementary material for the FL-EHDS paper, including complete algorithm pseudocode for all framework components, extended experimental figures, detailed algorithm comparison analysis with effect sizes (rank-biserial correlation), 10-seed statistical validation, advanced FL paradigm descriptions, infrastructure component specifications, extended EHDS interoperability details, clinical imaging experiment configurations with Chest X-ray baseline results, local epochs sweep, scalability analysis up to K=100 clients, noise-as-regularization analysis for small-cohort studies, Article 71 opt-out impact with policy recommendations, an extended threat model with attack taxonomy and defense mapping, and a detailed framework positioning analysis. The open-source reference implementation (∼40K lines, 159 modules) is available at https://github.com/FabioLiberti/FL-EHDS-FLICS2026.**

I. PRISMA Flow Diagram

II. Algorithm Pseudocode

This section provides formal algorithmic descriptions of all FL-EHDS framework components. Each algorithm is presented with: (1) a contextual explanation of *why* the component is needed in the EHDS regulatory context; (2) the formal pseudocode; and (3) practical considerations for deployment. The algorithms are organized following the data flow through the three-layer architecture: governance validation (Layer 1), privacy-preserving aggregation (Layer 2), and local data processing (Layer 3).

Reading guide: Algorithms S1–S4 form the core FL-EHDS training pipeline. Algorithms S5–S6 address EHDS-specific challenges (non-IID data and citizen opt-out). Algorithms S7–S8 handle data preprocessing and privacy budget management.

A. FedAvg with EHDS Compliance

Algorithm S1 presents the core federated averaging procedure adapted for EHDS regulatory requirements, operating in a client-server architecture where the central aggregator coordinates training across distributed hospital nodes within a Secure Processing Environment.

Key Design Decisions:

- **ValidatePermit**: Before each round, the HDAB-issued permit is verified against temporal bounds and Article 53 permitted purposes.
- **SelectParticipants**: Configurable client selection—full participation or sampling for large federations.
- **FilterOptedOut**: Records from citizens who exercised Article 71 opt-out rights are excluded *before* gradient computation.
- **Weighted Aggregation**: Gradients weighted by local dataset size (n_h), following original FedAvg [13].
- **ClipGradient**: L2-norm clipping bounds individual contributions, providing sensitivity bounds for DP.

Relationship to subsequent components: The `ClipGradient` operation in Algorithm S1 establishes a bounded sensitivity C for each client's contribution. This bound is the prerequisite for Algorithm S2 (Gaussian DP), which calibrates noise proportional to C. Meanwhile, `ValidatePermit` invokes Algorithm S3 (Permit Validation) and `FilterOptedOut` invokes Algorithm S6 (Opt-Out Filtering).

B. Gaussian Differential Privacy Mechanism

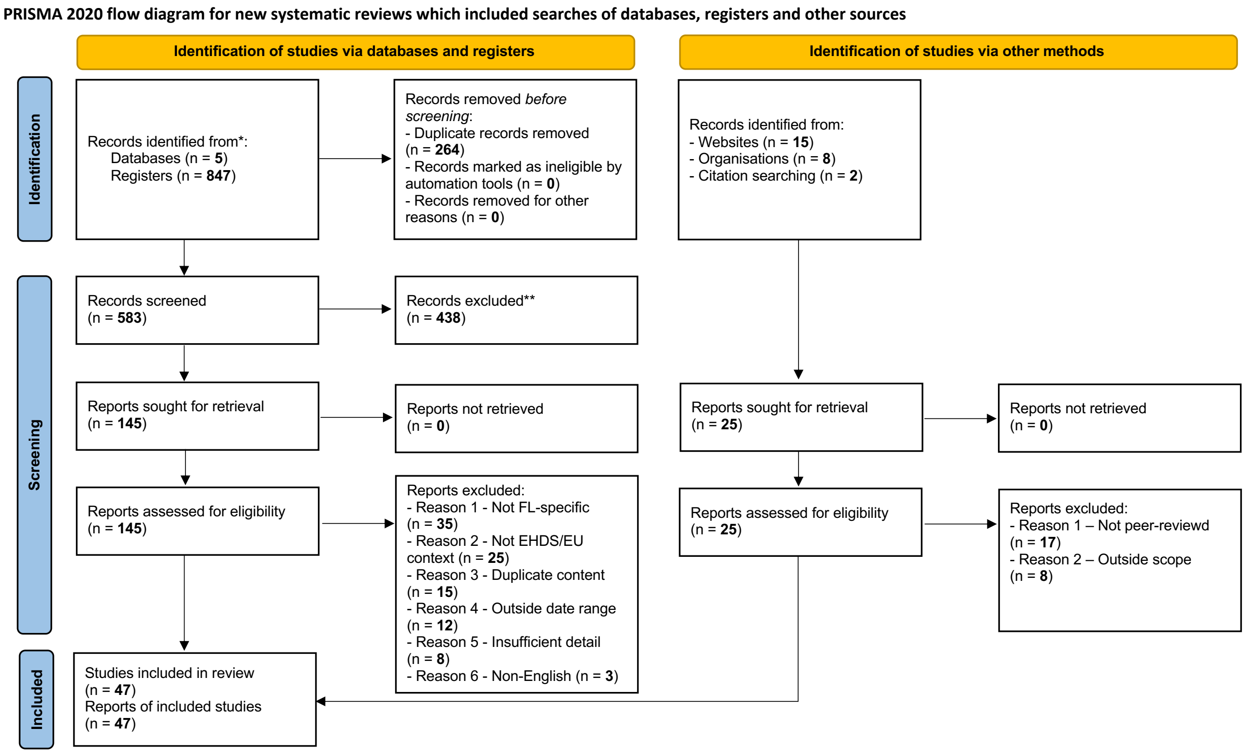
Algorithm S2 implements the Gaussian mechanism for differential privacy, applied at the aggregation server after receiving clipped gradients.

Mathematical Foundation: The noise scale $\sigma = C \cdot \sqrt{2\ln(1.25/\delta)}/\varepsilon$ guarantees (ε, δ)-DP. The cumulative privacy expenditure is tracked using Rényi DP (RDP) [26] composition, providing 5–6× tighter bounds than naive composition.

Practical Considerations:

- $\varepsilon = 10$: moderate noise, <2pp accuracy drop on PTB-XL and Cardiovascular (see Table S-X in Section XV)
- $\varepsilon = 1$: strong privacy; personalized methods (Ditto, HPFL) retain 87–89% on PTB-XL while FedAvg collapses to 52%
- The ε selection must be negotiated with HDABs during permit approval

Integration with privacy budget: The ε consumed by Algorithm S2 in each round is tracked by Algorithm S8 (RDP Privacy Budget Accountant). If the cumulative budget exceeds the threshold approved in the HDAB data permit, training is


PRISMA 2020 flow diagram for new systematic reviews which included searches of databases, registers and other sources
Identification of studies via databases and registers
Identification of studies via other methods
Identification
Screening
Included
Records identified from*:
Databases (n = 5)
Registers (n = 847)
Records removed before screening:
- Duplicate records removed (n = 264)
- Records marked as ineligible by automation tools (n = 0)
- Records removed for other reasons (n = 0)
Records identified from:
- Websites (n = 15)
- Organisations (n = 8)
- Citation searching (n = 2)
Records screened (n = 583)
Records excluded** (n = 438)
Reports sought for retrieval (n = 145)
Reports not retrieved (n = 0)
Reports sought for retrieval (n = 25)
Reports not retrieved (n = 0)
Reports assessed for eligibility (n = 145)
Reports excluded:
- Reason 1 - Not FL-specific (n = 35)
- Reason 2 - Not EHDS/EU context (n = 25)
- Reason 3 - Duplicate content (n = 15)
- Reason 4 - Outside date range (n = 12)
- Reason 5 - Insufficient detail (n = 8)
- Reason 6 - Non-English (n = 3)
Reports assessed for eligibility (n = 25)
Reports excluded:
- Reason 1 – Not peer-reviewd (n = 17)
- Reason 2 – Outside scope (n = 8)
Studies included in review (n = 47)
Reports of included studies (n = 47)

*Consider, if feasible to do so, reporting the number of records identified from each database or register searched (rather than the total number across all databases/registers).
**If automation tools were used, indicate how many records were excluded by a human and how many were excluded by automation tools.

Source: Page MJ, et al. BMJ 2021;372:n71. doi: 10.1136/bmj.n71.

Fig. S-1. PRISMA 2020 flow diagram for the systematic review. Database searches across PubMed, IEEE Xplore, Scopus, Web of Science, and arXiv identified 847 records; after deduplication (264 removed) and screening, 47 studies met inclusion criteria (2022–2026, FL/EHDS focus, peer-reviewed or recognized institutional origin). Additional records from institutional websites (n=15), organisations (n=8), and citation searching (n=2) were assessed but did not contribute to the final inclusion set. Adapted from Page et al. (BMJ 2021;372:n71), licensed under CC BY 4.0.

automatically terminated. The next algorithm (S3) formalizes the permit validation that authorizes each round.

C. *HDAB Permit Validation*

Algorithm S3 ensures all FL operations comply with the data permit issued by HDABs. Under EHDS Article 53, secondary use of health data is only lawful for specifically enumerated purposes (scientific research, public health surveillance, AI training for health). The permit validation module is invoked at the beginning of every FL round—not just at training start—to guarantee continuous compliance even if a permit is revoked mid-study or its temporal validity expires. Each validation event is logged as a GDPR Article 30 processing record, creating an immutable audit trail that regulators can inspect.

EHDS Governance Role: This algorithm is the enforcement point between the Governance Layer (Layer 1) and the FL Orchestration Layer (Layer 2). Without it, a permit expiring at round 15 of a 20-round study would allow unauthorized data processing for rounds 16–20.

Failure modes: Each exception type triggers a different response: `PermitExpiredError` terminates the entire study; `PurposeMismatchError` indicates a configuration error requiring researcher intervention; `UnauthorizedCategoryError` may allow continued training on the authorized subset of categories. All failure events are logged for regulatory audit. Once a round passes permit validation, the aggregation server collects client gradients using the secure protocol described next.

D. *Secure Aggregation Protocol*

Even though FL prevents raw data sharing, the gradient updates themselves can leak patient information: Zhu et al. [21] demonstrate that gradients can be inverted to reconstruct training images. In the EHDS context, where gradients encode patterns from sensitive health records across 27 Member States, this is an unacceptable privacy risk. Secure aggregation addresses this by ensuring the SPE aggregation server can compute $\sum_k \Delta_k$ without ever observing any individual hospital's gradient Δ_k.

```
Algorithm S1: 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 hospital $h \in \mathcal{H}_t$ in parallel do
            $\Delta_h^{(t)}, n_h \leftarrow$ LocalTrain($h$, $\theta^{(t-1)}$)
        // Aggregation with privacy (Layer 2)
        $\theta^{(t)} \leftarrow \theta^{(t-1)} + \frac{1}{\sum_h n_h} \sum_{h \in \mathcal{H}_t} n_h \cdot \Delta_h^{(t)}$
        LogCompliance($t$, $\mathcal{H}_t$)
    return $\theta^{(T)}$
LocalTrain($h$, $\theta$) at hospital $h$:
    // Opt-out filtering (Article 71)
    $\mathcal{D}_h \leftarrow$ FilterOptedOut($\mathcal{D}_h$, OptOutRegistry)
    $\theta_h \leftarrow \theta$
    for epoch $e = 1$ to $E$ do
        for batch $\mathcal{B} \in \mathcal{D}_h$ do
            $\theta_h \leftarrow \theta_h - \eta \nabla \mathcal{L}(\theta_h; \mathcal{B})$
    $\Delta_h \leftarrow \theta_h - \theta$
    // Privacy protection (Layer 3)
    $\Delta_h \leftarrow$ ClipGradient($\Delta_h$, $C$)
    return $\Delta_h$, $|\mathcal{D}_h|$
```

```
Algorithm S2: Gaussian DP Mechanism
Input: Gradient $\Delta$, sensitivity $C$, privacy budget $\varepsilon$, $\delta$
Output: Noisy gradient $\tilde{\Delta}$
// Compute noise scale from Gaussian mechanism
$\sigma \leftarrow C \cdot \sqrt{2\ln(1.25/\delta)}/\varepsilon$
// Add calibrated Gaussian noise to each parameter
for each parameter $w \in \Delta$ do
    $\tilde{w} \leftarrow w + \mathcal{N}(0, \sigma^2)$
// Track cumulative privacy expenditure
PrivacyAccountant.spend($\varepsilon$)
if PrivacyAccountant.budget_exhausted() then
    raise PrivacyBudgetExhaustedError
return $\tilde{\Delta}$
```

```
Algorithm S3: Data Permit Validation
Input: Permit $P$, round $t$, requested categories $\mathcal{C}$
Output: Boolean validity
// Check temporal validity
if CurrentTime() $>$ $P$.valid_until then
    raise PermitExpiredError
// Check purpose alignment (Article 53)
if $P$.purpose $\notin$ AllowedPurposes then
    raise PurposeMismatchError
// Check data category authorization
for each category $c \in \mathcal{C}$ do
    if $c \notin P$.authorized_categories then
        raise UnauthorizedCategoryError
// Log access for GDPR Article 30
AuditTrail.log(permit=$P$, round=$t$, categories=$\mathcal{C}$)
return True
```

Algorithm S4 implements this using Shamir's secret sharing and pairwise masking, ensuring the server observes only the aggregate gradient.

Protocol Phases: (1) Each client splits gradients into K shares using (t, K)-threshold Shamir secret sharing; (2) Clients add pairwise random masks negotiated via ECDH key exchange; (3) The server computes the sum—masks cancel out and only the true aggregate remains.

```
Algorithm S4: Secure Aggregation (Pairwise Masking)
Input: Client gradients $\{\Delta_1, \ldots, \Delta_K\}$, threshold $t$
Output: Aggregated gradient $\Delta_{agg}$
// Phase 1: ECDH key exchange + Shamir sharing
for each client $k$ do
    shares$_k$ $\leftarrow$ ShamirShare($\Delta_k$, $t$, $K$)
    Distribute shares$_k$ to other clients
// Phase 2: Add pairwise random masks
for each client $k$ do
    $\hat{\Delta}_k \leftarrow \Delta_k + \sum_{j<k} r_{jk} - \sum_{j>k} r_{kj}$
// Phase 3: Server reconstructs aggregate
$\Delta_{agg} \leftarrow \sum_{k=1}^{K} \hat{\Delta}_k$
// Masks cancel: $\sum_k \sum_{j<k} r_{jk} - \sum_k \sum_{j>k} r_{kj} = 0$
if ActiveClients $< t$ then
    raise SecureAggregationError
return $\Delta_{agg}$
```

Defense-in-depth: Secure aggregation (Algorithm S4) combined with differential privacy (Algorithm S2) provides layered protection: even if the aggregation server is compromised, it learns only the noisy aggregate—never individual hospital contributions. The combination addresses the unresolved GDPR question of whether model gradients constitute "personal data": with both mechanisms active, the information available to any single party is provably bounded. The following algorithms address how local training handles EHDS-specific data challenges.

E. FedProx for Non-IID Data

Algorithm S5 extends FedAvg with a proximal term that penalizes local model divergence from the global model [14]. In EHDS cross-border federations, data heterogeneity is a structural feature, not an exception: hospitals in different Member States serve distinct demographics, follow national clinical guidelines, and use different diagnostic thresholds. For instance, heart disease prevalence ranges from 39.2% in Rome to 62.6% in Amsterdam in our experimental setting. Without drift control, local models can diverge so far from the global consensus that aggregation produces a deteriorated global model. The proximal term $\frac{\mu}{2}\|\theta_h - \theta\|^2$ acts as a regularizer that keeps each hospital's local update within a controlled distance of the global model, balancing personalization with collaboration.

When to use in EHDS: Recommended for federations with moderate non-IID conditions and when client dropout is expected (hospitals may temporarily disconnect). FedProx tolerates partial participation better than FedAvg because the

```
Algorithm S5: FedProx Local Update
Input: Local data $\mathcal{D}_h$, global model $\theta$, proximal weight $\mu$
Output: Local update $\Delta_h$
$\theta_h \leftarrow \theta$
for epoch $e = 1$ to $E$ do
    for batch $\mathcal{B} \in \mathcal{D}_h$ do
        $g \leftarrow \nabla \mathcal{L}(\theta_h; \mathcal{B})$
        // Proximal term: $\nabla \frac{\mu}{2} \|\theta_h - \theta\|^2$
        $g \leftarrow g + \mu(\theta_h - \theta)$
        $\theta_h \leftarrow \theta_h - \eta \cdot g$
$\Delta_h \leftarrow \theta_h - \theta$
return $\Delta_h$
```

```
Algorithm S6: Article 71 Opt-Out Filtering
Input: Local dataset $\mathcal{D}_h$, purpose $p$, categories $\mathcal{C}$
Output: Filtered dataset $\mathcal{D}'_h$
// Synchronize with national opt-out registry
OptOutRecords $\leftarrow$ FetchOptOutRegistry(MemberState)
$\mathcal{D}'_h \leftarrow \emptyset$
for each record $r \in \mathcal{D}_h$ do
    citizen_id $\leftarrow$ r.pseudonymized_id
    opted_out $\leftarrow$ False
    // Check purpose-specific opt-out
    if (citizen_id, $p$) $\in$ OptOutRecords then
        opted_out $\leftarrow$ True
    // Check category-specific opt-out
    for each $c \in \mathcal{C}$ do
        if (citizen_id, $c$) $\in$ OptOutRecords then
            opted_out $\leftarrow$ True
    if not opted_out then
        $\mathcal{D}'_h \leftarrow \mathcal{D}'_h \cup \{r\}$
// Log filtering statistics for audit
AuditLog.record(total=$|\mathcal{D}_h|$, filtered=$|\mathcal{D}'_h|$)
return $\mathcal{D}'_h$
```

proximal term stabilizes local updates even with fewer training epochs.

Parameter Selection: $\mu = 0$ reduces to FedAvg; $\mu \in [0.01, 0.1]$ provides stable convergence; $\mu > 1$ may prevent local adaptation. The choice of μ should be documented in the data permit application so that the HDAB can assess the expected privacy-utility trade-off.

Before any local training begins (whether with FedAvg, FedProx, or any other algorithm), the framework must enforce citizen opt-out rights. The following algorithm ensures this compliance.

F. Article 71 Opt-Out Registry Protocol

Algorithm S6 implements the citizen opt-out mechanism mandated by EHDS Article 71. This article grants every EU citizen the right to object to secondary use of their electronic health data—a fundamental right that must be enforced *before* any gradient computation occurs. The algorithm queries the national opt-out registry maintained by each Member State and removes matching records from the local training dataset.

Granularity levels: (1) *Blanket opt-out*—citizen refuses all secondary use; (2) *Purpose-specific*—e.g., permitting scientific research but blocking commercial analytics; (3) *Category-specific*—e.g., allowing demographics but blocking genomic data. This granularity reflects the EHDS principle that citizens should have meaningful control, not merely a binary yes/no choice.

EHDS Governance Role: Opt-out filtering operates at Layer 3 (Data Holders) before local training. Registry lookups use LRU caching with configurable TTL to minimize latency ($<$10ms per round) while ensuring timely propagation of new opt-out decisions. All filtering statistics are logged for GDPR Article 30 audit compliance.

Impact on model quality: High opt-out rates reduce training data volume, potentially degrading model performance—particularly for underrepresented subpopulations. The audit log captures filtering statistics to quantify this impact and support transparency reporting. Once opted-out records are excluded, the remaining data must be harmonized into a consistent format before local model training can proceed.

G. FHIR R4 Preprocessing Pipeline

Algorithm S7 standardizes heterogeneous EHR data into FHIR R4 format for ML consumption. This preprocessing step is critical in the EHDS context because only 34% of European healthcare providers currently achieve full FHIR R4 compliance [7]. The remaining 66% use legacy formats (HL7v2, CDA, proprietary CSV exports) that must be harmonized before FL training can proceed on a consistent feature space.

Four-stage pipeline: (1) *Format detection* automatically identifies the source format; (2) *Terminology mapping* converts local codes to international standards (ICD-10 for diagnoses, ATC for medications, LOINC for laboratory results); (3) *FHIR transformation* produces validated FHIR R4 bundles using the six Article 33 data categories (Patient Summary, E-Prescription, Laboratory Results, Medical Imaging, Hospital Discharge, Rare Disease); (4) *Tensor extraction* converts structured FHIR resources into numerical tensors ready for model training.

EHDS Relevance: Without this harmonization step, hospitals in different Member States would produce incompatible feature spaces, making federated aggregation meaningless. The pipeline ensures that a gradient computed in a Finnish hospital is semantically compatible with one from an Italian hospital.

Validation requirements: The FHIR validation step rejects records with missing mandatory fields or invalid terminology codes, ensuring data quality before model training. Rejected records are logged (without patient-identifiable content) for audit purposes. With harmonized data ready for training, the final core component manages the overall privacy budget across the entire study.

H. Privacy Budget Accountant

Algorithm S8 tracks cumulative privacy expenditure across FL rounds using Rényi Differential Privacy (RDP) moment accounting [26]. In the EHDS governance model, the total privacy budget ε_{total} is a parameter of the data permit: the researcher specifies the desired budget in the permit application, and the HDAB evaluates whether the proposed budget

```
Algorithm S7: FHIR R4 Preprocessing
Input: Raw EHR records $\mathcal{R}$, feature specification $\mathcal{F}$
Output: Training tensors $(X, y)$
format ← DetectFormat($\mathcal{R}$) // HL7v2, CDA, CSV
parser ← GetParser(format)
records ← parser.parse($\mathcal{R}$)
// Map to standard terminologies
for each $r \in$ records do
    $r$.diagnoses ← MapToICD10($r$.diagnoses)
    $r$.medications ← MapToATC($r$.medications)
    $r$.labs ← MapToLOINC($r$.labs)
fhir_bundle ← ToFHIR(records)
ValidateFHIR(fhir_bundle)
$X$ ← ExtractFeatures(fhir_bundle, $\mathcal{F}$)
$X$ ← StandardScaler.fit_transform($X$)
$y$ ← ExtractLabels(fhir_bundle)
return $(X, y)$
```

provides sufficient privacy protection for the requested data categories and population size.

Why RDP accounting: Naive DP composition (adding ε per round) yields loose bounds: 20 rounds at ε=0.5 each would consume ε=10 total. RDP provides 5–6× tighter bounds [26], [36], meaning the same 20 rounds can achieve the same privacy guarantee with significantly less noise—and therefore better model utility.

Hard budget enforcement: When the cumulative expenditure approaches ε_{total}, the accountant raises `BudgetExhaustedError` that terminates training. This prevents "privacy bankruptcy"—a situation where continued training would violate the privacy guarantee approved in the data permit. The per-round allocation strategy distributes remaining budget uniformly across remaining rounds, adapting dynamically if training converges faster than expected.

```
Algorithm S8: RDP Privacy Budget Accountant
Input: Total budget $(\varepsilon_{total}, \delta_{total})$, rounds $T$
Output: Per-round budget allocation
$\lambda$ ← [0] × MAX_ORDER                    // Rényi moments
rounds_completed ← 0
function AllocateRound():
    $\varepsilon_{spent}$ ← ComputeEpsilon($\lambda$, $\delta_{total}$)
    $\varepsilon_{remaining} \leftarrow \varepsilon_{total} - \varepsilon_{spent}$
    if $\varepsilon_{remaining} < \varepsilon_{min}$ then
        raise BudgetExhaustedError
    $\varepsilon_t \leftarrow \varepsilon_{remaining}/(T-$ rounds_completed)
    return $\varepsilon_t$
function RecordRound($\sigma$, $q$):
    for order = 1 to MAX_ORDER do
        $\lambda$[order] += ComputeMoment(order, $\sigma$, $q$)
    rounds_completed += 1
```

III. Supplementary Experimental Figures

This section presents detailed experimental results from the FL-EHDS benchmark suite. All figures are generated from real experimental runs available in the repository.

Note on experimental configurations: Figures S-2–S-9 were generated from an extended 50-round, 5-client training run using the framework's synthetic EHDS scenario (simulated European hospitals: Rome, Amsterdam, Berlin, Madrid, Paris). These complement the main paper's 20-round experiments on Heart Disease UCI (4 real hospitals) and Diabetes (5 Dirichlet-partitioned clients). The 50-round configuration illustrates longer-horizon convergence properties, client participation dynamics, and gradient evolution patterns that are not visible in the shorter 20-round evaluation.

A. Hospital Data Distribution

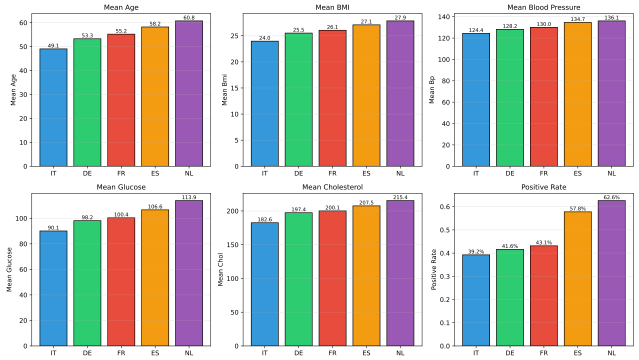


Fig. S-2. Data distribution across hospitals. Notable heterogeneity: Amsterdam shows older population (60.8 years mean age) with higher positive rate (62.6%) compared to Rome (49.1 years, 39.2%). This reflects realistic cross-border EHDS variability.

B. Per-Client Training Time

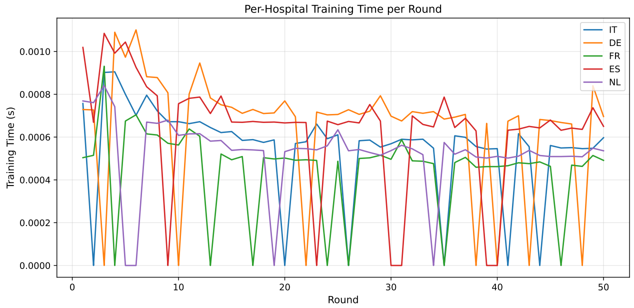


Fig. S-3. Per-client training time per round. Larger hospitals (Berlin: 500 samples) exhibit slightly longer training times. The adaptive training engine compensates by adjusting batch sizes for stragglers.

C. Client Participation Matrix

Figure S-4 shows the client participation pattern across training rounds.

D. Gradient Norm Evolution

Figure S-5 tracks gradient norms per client, confirming convergence stability.

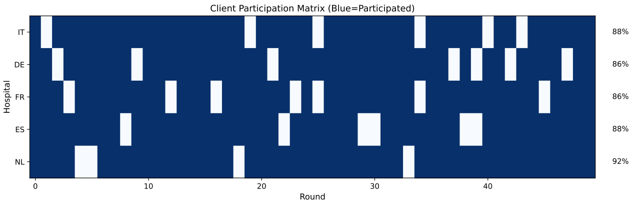


Fig. S-4. Client participation matrix (50 rounds × 5 clients). Participation rates: IT 88%, DE 86%, FR 86%, ES 88%, NL 92%. The framework tolerates 10–15% dropout per round while maintaining convergence.

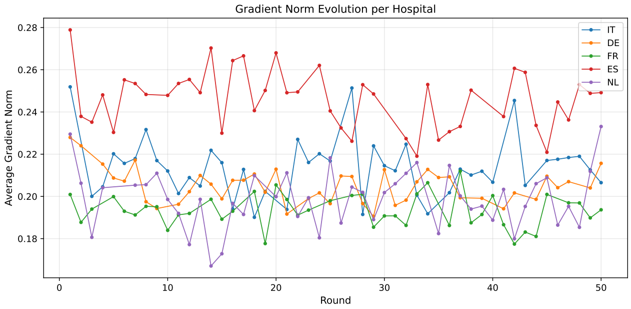


Fig. S-5. Gradient norm evolution per client over 50 rounds. All clients show decreasing trends indicating stable convergence. Clipping threshold C=1.0 bounds extreme values for DP compatibility.

E. Communication Cost Analysis

Figure S-6 quantifies the per-round communication overhead.

F. Learning Rate Sensitivity

Figure S-7 evaluates sensitivity to the server learning rate.

G. Batch Size Impact

Figure S-8 examines the trade-off between batch size and convergence.

H. Per-Client Accuracy Trajectories

Figure S-9 shows per-client accuracy evolution, illustrating the impact of non-IID data.

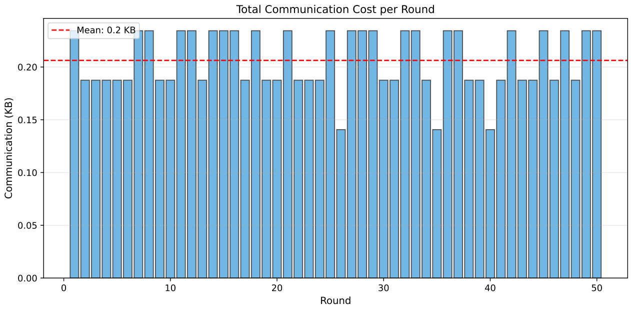


Fig. S-6. Cumulative communication cost per round. Linear scaling with participating clients (3.5 KB/client/round). Total 50-round overhead: 875 KB for 5 clients—feasible even for bandwidth-constrained environments.

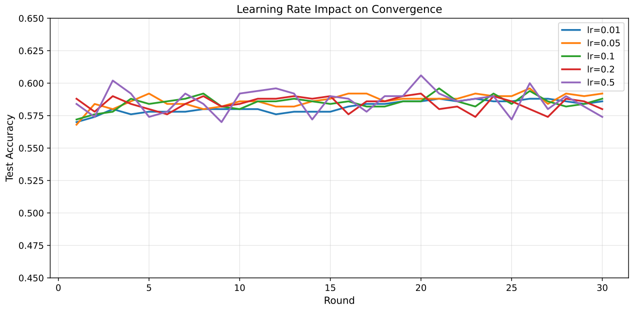


Fig. S-7. Learning rate sensitivity analysis. η=0.01: slow convergence (53.8% at round 50). η=0.1: optimal (58.6%). η=0.5: instability with oscillations.

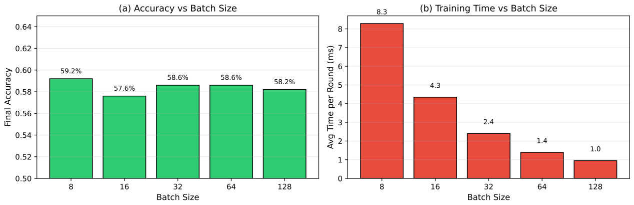


Fig. S-8. Batch size impact on convergence. Smaller batches (8–16) provide noisier gradients but faster initial progress. Batch size 32 balances gradient quality and computational efficiency.

IV. Dataset Landscape

The FL-EHDS framework supports 19 healthcare datasets spanning four modalities. Table S-I provides a comprehensive overview. This diversity enables evaluation across multiple EHDS-relevant dimensions: data scale (120–253K samples), feature dimensionality (9–30 tabular, high-dimensional imaging), task complexity (binary to 5-class), partition strategies (natural hospital-based and synthetic Dirichlet), and interoperability standards (CSV, FHIR R4, OMOP-CDM). Experimentally evaluated datasets in the main paper are marked with ✓.

V. Extended Algorithm Comparison

A. Algorithms Evaluated

We compare foundational FL algorithms plus 2022–2025 advances:

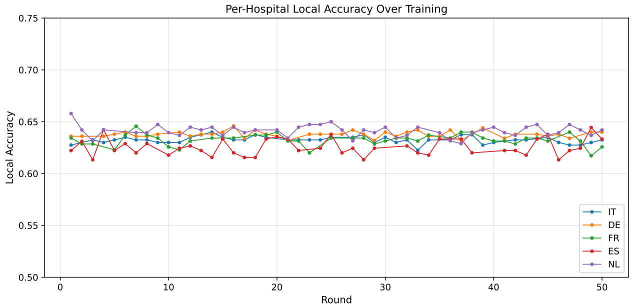


Fig. S-9. Per-client accuracy over training rounds. Variance reflects non-IID data: NL (older, higher-risk population) reaches 64% accuracy while FR (mid-range demographics) stabilizes at 55%.

TABLE S-I

FL-EHDS Dataset Landscape: Complete Framework Coverage

Dataset	Type	Samples	Feat.	Classes	FL Partition	EHDS Category (Art. 33)	EHDS Level	Eval.
A. Tabular — Clinical EHR								
Diabetes 130-US	Tabular	101,766	22	2	Dirichlet (α=0.5)	EHR, ICD-9, medications	L2: FHIR-mappable	✓
Heart Disease UCI	Tabular	920	13	2	Natural (4 hospitals)	Vitals, ECG, lab results	L2: FHIR-mappable	✓
PTB-XL ECG[†]	Tabular	21,799	9	5	Natural (52 sites)	SCP-ECG (EN 1064), diagnostics	L2: FHIR-mappable	✓
Cardiovascular Disease	Tabular	70,000	11	2	Dirichlet (α=0.5)	Vitals, lab, risk factors	L2: FHIR-mappable	✓
Breast Cancer Wisconsin	Tabular	569	30	2	Dirichlet (α=0.5)	Pathology (FNA cytology)	L2: FHIR-mappable	✓
Stroke Prediction	Tabular	5,110	10	2	Dirichlet	Cardiovascular risk factors	L2: FHIR-mappable	—
CDC Diabetes BRFSS	Tabular	253,680	21	2	Dirichlet	Population health survey	L2: FHIR-mappable	—
CKD UCI	Tabular	400	24	2	Dirichlet	Renal panel, comorbidities	L2: FHIR-mappable	—
Cirrhosis Mayo	Tabular	418	18	2	Dirichlet	Hepatology, drug trial	L2: FHIR-mappable	—
B. Tabular — FHIR-Native								
Synthea FHIR R4	FHIR	1,180	14	2	Hospital profile	Patient, Condition, Encounter	L1: FHIR-native	qual.
SMART Bulk FHIR	FHIR	120	12	2	Single export	NDJSON Bulk Data (Art. 46)	L1: FHIR-native	qual.
C. Generated Pipelines (in-memory)								
FHIR R4 Synthetic	Gen.	config.	10	2	Hospital profile	Generated FHIR bundles	L1: FHIR-native	qual.
OMOP-CDM Harmonized	Gen.	config.	36	2	Cross-border	Vocabulary harmonization	L3: OMOP	qual.
D. Medical Imaging								
Chest X-ray	Imaging	5,856	—	2	Dirichlet (α=0.5)	Radiology (DICOM)	L4: Imaging	✓
Brain Tumor MRI	Imaging	7,023	—	4	Dirichlet (α=0.5)	Neuro-imaging (DICOM)	L4: Imaging	✓
Skin Cancer	Imaging	3,297	—	2	Dirichlet (α=0.5)	Dermatology (DICOM)	L4: Imaging	✓
Diabetic Retinopathy	Imaging	35,126	—	5	Dirichlet	Ophthalmology (DICOM)	L4: Imaging	—
Brain Tumor MRI (alt.)	Imaging	3,264	—	4	Dirichlet	Neuro-imaging (DICOM)	L4: Imaging	—
ISIC Skin Lesions	Imaging	2,357	—	9	Dirichlet	Dermatology (DICOM)	L4: Imaging	—

EHDS Levels: L1 = FHIR-native (Art. 46 compliant); L2 = FHIR-mappable (standard clinical features with FHIR mapping in metadata); L3 = OMOP-CDM harmonized (cross-border vocabulary alignment, Art. 50); L4 = Medical imaging (DICOM, Art. 33 "medical images").

[†]PTB-XL: European-origin dataset (PTB, Berlin, Germany) with SCP-ECG coding (EN 1064). 52 recording sites enable natural hospital-based FL partitioning—the strongest EHDS benchmark in the framework.

Eval.: ✓= quantitative experimental evaluation (P1.2); qual. = qualitative pipeline validation; — = supported but not evaluated in current paper. **config.** = sample count depends on generation parameters.

Foundational: FedAvg [13], FedProx [14], SCAFFOLD [29], FedAdam/FedYogi/FedAdagrad [31].

Recent (2022–2025): FedLC [39] (logit calibration for label skew), FedSAM [38] (flat minima), FedDecorr [40] (decorrelation against dimensional collapse), FedSpeed [41] (fewer rounds), FedExP [42] (server-side acceleration), FedLESAM [43] (globally-guided SAM, ICML 2024 Spotlight), HPFL [44] (personalized classifiers, ICLR 2025).

B. Non-IID Configuration

Data heterogeneity is controlled via Dirichlet distribution with α:

- $\alpha = 0.1$: **Extreme non-IID**—highly skewed label distributions
- $\alpha = 0.5$: **High non-IID**—significant heterogeneity
- $\alpha = 1.0$: **Moderate non-IID**—balanced heterogeneity
- $\alpha = 10.0$: **Near-IID**—approximately uniform

C. Convergence at Different Heterogeneity Levels

Findings: (1) At α=0.1, SCAFFOLD achieves most stable convergence via variance reduction. (2) FedProx provides marginal improvement over FedAvg at α=0.5–1.0. (3) Adaptive methods (FedAdam, FedYogi) excel in near-IID but may oscillate under extreme heterogeneity. (4) FedAvg remains competitive in near-IID, suitable for homogeneous federations.

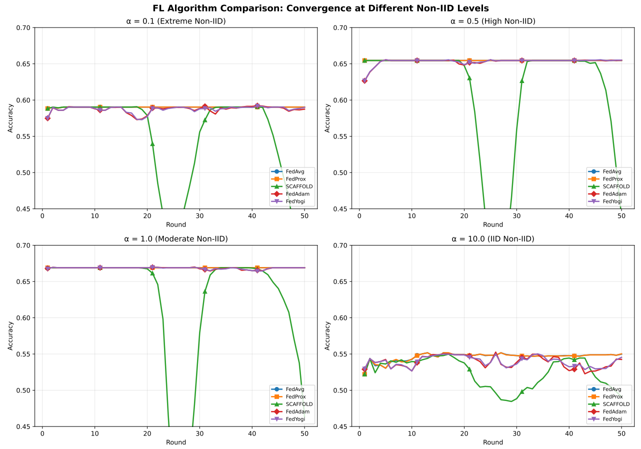


Fig. S-10. Algorithm convergence across non-IID levels ($\alpha \in \{0.1, 0.5, 1.0, 10.0\}$). SCAFFOLD and adaptive methods show superior stability under extreme heterogeneity.

D. Final Accuracy vs. Heterogeneity

E. Convergence Speed

Figure S-12 compares how quickly algorithms reach target accuracy thresholds.

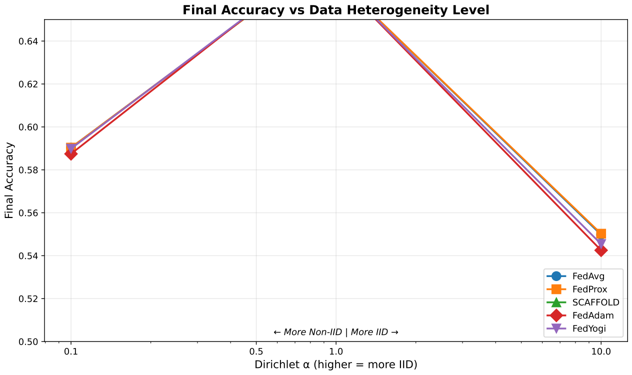


Fig. S-11. Final accuracy vs. Dirichlet α. All algorithms degrade under extreme non-IID. SCAFFOLD shows smallest gap between α=0.1 and α=10.

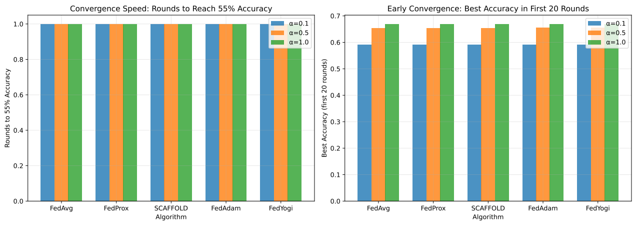


Fig. S-12. Convergence speed comparison. Left: rounds to 55% accuracy. Right: best accuracy in first 20 rounds. Adaptive methods converge faster but may plateau.

F. Algorithm Selection Guidelines

Table S-II maps EHDS deployment scenarios to recommended algorithms.

TABLE S-II
ALGORITHM SELECTION FOR EHDS DEPLOYMENTS

EHDS Scenario	Algorithm	Rationale
Homogeneous MS	FedAvg	Simplicity, proven
Heterogeneous MS	SCAFFOLD	Variance reduction
Resource-limited	FedAdam	Fast convergence
Privacy-critical	FedAvg + DP	Well-studied bounds
Sparse participation	FedProx	Dropout resilience
Label-imbalanced	FedLC	Class-freq. calib.
Deep models, non-IID	FedDecorr	Dim. collapse prev.
Comm.-constrained	FedSpeed	Fewer rounds
No client changes	FedExP	Server-side only
SAM + global drift	FedLESAM	Global flatness
Per-hosp. classif.	HPFL	Local boundaries

MS = Member States. Scenarios may combine: heterogeneous + privacy-critical → SCAFFOLD + DP.

VI. Advanced FL Paradigms

Note: The paradigms in this section are *implemented in the reference framework* and available for use, but are *not experimentally evaluated* in this paper's benchmark suite. Our experimental validation (Section IV of the main paper) focuses on horizontal FL with 7 algorithms across 5 tabular datasets and 3 imaging datasets. The advanced paradigms below are provided as architectural capabilities for future evaluation on appropriate multi-institutional datasets.

The core FL-EHDS pipeline (Section II) addresses the standard "horizontal" FL scenario where all hospitals share the same feature schema. However, real EHDS deployments will encounter more complex configurations: institutions with complementary features for the same patients (vertical FL), adversarial participants (Byzantine resilience), evolving data distributions over the 2025–2031 timeline (continual FL), heterogeneous clinical objectives (multi-task FL), and the hierarchical governance structure of the EU itself (hierarchical FL). This section presents the advanced paradigms implemented in the reference framework, each motivated by a specific EHDS deployment challenge. Algorithms S9–S13 formalize the core mechanisms.

A. Vertical Federated Learning

Vertical FL addresses scenarios where institutions hold *different features* for the *same patients*—a common situation in EHDS cross-border analytics. For example, a hospital may hold demographics and diagnoses, a laboratory holds test results, and a pharmacy holds prescription histories. Under EHDS Article 33, these correspond to different data categories (Patient Summary, Laboratory Results, E-Prescription) that may be held by different data holders within the same or different Member States.

Private Set Intersection (PSI): Before training, the participating institutions must identify their common patients without revealing their full patient lists. RSA-based PSI achieves this with $O(n \log n)$ complexity using pseudonymized identifiers, ensuring EHDS compliance: no institution learns which patients the other holds beyond the intersection.

Split Learning: Algorithm S9 implements the forward pass in split learning, where each party computes activations on its local features up to a "cut layer," then the server concatenates activations to produce the final prediction. Only intermediate representations (not raw data) cross institutional boundaries.

Algorithm S9: Split Learning Forward Pass

```
Input: Features $X_A$, $X_B$ at parties A, B; cut layer $k$
Output: Prediction $\hat{y}$
$h_A \leftarrow f^A_{1:k}(X_A)$            // Party A: features → activations
$h_B \leftarrow f^B_{1:k}(X_B)$            // Party B: features → activations
$h \leftarrow \text{Concat}(h_A, h_B)$
$\hat{y} \leftarrow f_{k+1:L}(h)$            // Server: cut layer → output
return $\hat{y}$
```

B. Byzantine-Resilient Aggregation

In a cross-border EHDS federation spanning 27 Member States, the aggregation server cannot blindly trust every participant. A compromised institution—whether through malicious intent, software bugs, or data corruption—could submit adversarial gradient updates that poison the global model, potentially affecting clinical decisions across the entire federation. Byzantine-resilient aggregation protects model integrity by detecting and excluding anomalous updates.

Algorithm S10 implements Krum, which selects the gradient closest to $n-f-2$ nearest neighbors, effectively filtering outliers. Six defense methods protect against up to $f < n/3$ adversarial clients:

```
Algorithm S10: Krum Byzantine Defense
Input: Gradients $\{g_1, \ldots, g_n\}$, Byzantine bound $f$
Output: Selected gradient $g^*$
for each gradient $g_i$ do
    $D_i \leftarrow \{\|g_i - g_j\|^2 : j \neq i\}$
    $s_i \leftarrow \sum_{d \in \text{smallest}_{n-f-2}(D_i)} d$
$g^* \leftarrow g_{\arg\min_i s_i}$
return $g^*$
```

Other methods: **Trimmed Mean** (removes β-fraction extreme values per coordinate), **Coordinate-wise Median** (robust estimator), **Bulyan** (two-stage Krum + trimmed mean), **FLTrust** (server-guided trust weighting), **FLAME** (clustering-based). Attack simulation: label flipping, gradient scaling, additive noise, sign flipping, model replacement.

C. Continual Federated Learning

The EHDS is designed for long-term operation (2025–2031 and beyond), during which healthcare data distributions will evolve: new diseases emerge (as demonstrated by COVID-19), clinical protocols change, and demographic compositions shift. A model trained in 2027 may perform poorly on 2029 data if it has “forgotten” how to handle earlier patterns. Continual Federated Learning addresses this *catastrophic forgetting* problem by preserving knowledge from previous training tasks while adapting to new data.

The Elastic Weight Consolidation (EWC) loss function adds a quadratic penalty:

$$\mathcal{L}_{EWC}(\theta) = \mathcal{L}(\theta) + \frac{\lambda}{2} \sum_i F_i (\theta_i - \theta_i^*)^2$$

where $F_i = \mathbb{E}[\nabla^2 \log p(\mathcal{D}|\theta^*)]_i$ is the i-th diagonal entry of the Fisher Information Matrix and θ^* are optimal parameters for previous tasks.

Additional strategies: Learning without Forgetting (LwF), Experience Replay, drift detection (ADWIN, Page-Hinkley) triggering adaptation.

D. Multi-Task Federated Learning

In EHDS cross-border studies, different hospitals may pursue related but distinct clinical objectives from the same data. A cardiology network might simultaneously predict heart failure risk (Hospital A), readmission probability (Hospital B), and medication response (Hospital C). Multi-Task FL enables these institutions to collaborate on shared feature representations while maintaining task-specific prediction heads.

Architectures: **Hard Parameter Sharing** (common feature extractor, task-specific heads), **Soft Parameter Sharing** (separate networks with similarity regularization), **FedMTL** (dynamic task relationship learning).

E. Hierarchical Federated Learning

The EHDS governance structure is inherently hierarchical: individual hospitals report to regional health authorities, which coordinate under national HDABs, which in turn connect to the EU-level HealthData@EU infrastructure. Hierarchical FL mirrors this governance topology, aggregating gradients at intermediate levels before reaching the central server. This reduces communication costs (hospitals communicate with regional aggregators, not directly with the EU server) and aligns FL operations with the jurisdictional boundaries of HDABs.

Four-tier hierarchy reflecting EU governance:

1) **Client Tier**: Individual hospitals/data holders
2) **Regional Tier**: Regional aggregators (e.g., Lombardy, Bavaria)
3) **National Tier**: National HDABs coordinate Member State aggregation
4) **EU Tier**: HealthData@EU central aggregator

Benefits: reduced communication costs (hospitals → regional, not directly EU), alignment with EHDS governance where HDABs have national jurisdiction.

F. Personalized Federated Learning

A single global model may underperform at individual hospitals because clinical populations differ substantially across Member States. Personalized FL maintains both a global model (encoding shared medical knowledge) and hospital-specific local models (adapted to local demographics and clinical practices). Algorithm S11 shows pFedMe [35] as a representative personalized method, using Moreau envelopes to balance personalization with global knowledge: the regularization term $\lambda(\theta_k - \theta)$ pulls the local model toward the global consensus, while local gradient descent adapts to hospital-specific data patterns. Ditto [33] follows a similar dual-model principle but with a simpler formulation: it trains a personalized model regularized toward the global model via $\frac{\lambda}{2}\|\theta_k - \theta\|^2$. In our experiments, Ditto—the best-performing personalized method—achieves 75.1% accuracy on Heart Disease, a 12.6pp improvement over FedAvg, precisely because it learns hospital-specific decision boundaries.

```
Algorithm S11: pFedMe Local Update
Input: Data $\mathcal{D}_k$, global $\theta$, personal $\theta_k$, $\lambda$, $\eta$
Output: Updated personal model $\theta'_k$
for $i = 1$ to $R$ do
    $\theta_k \leftarrow \theta_k - \eta \nabla \mathcal{L}(\theta_k; \mathcal{D}_k)$
// Moreau envelope: balance with global
$\theta'_k \leftarrow \theta_k - \lambda(\theta_k - \theta)$
$g_k \leftarrow \lambda(\theta - \theta'_k)$
return $\theta'_k$, $g_k$
```

Other approaches: **FedPer** (shared base, local personalization layers), **Per-FedAvg** (MAML-based meta-learning),

APFL (adaptive mixing α between global and local), **Ditto** (personalization regularization).

EHDS Relevance: Member States have different healthcare systems, disease prevalence, and clinical practices. Personalized FL enables institution-specific adaptation while benefiting from collaborative training.

G. Asynchronous Federated Learning

Standard synchronous FL requires all participating hospitals to complete local training before the server can aggregate. In an EHDS federation spanning 27 Member States with heterogeneous computational resources, this creates a "straggler" problem: a resource-constrained rural hospital delays the entire federation. Asynchronous FL eliminates this bottleneck by allowing the server to aggregate updates as they arrive, weighting stale updates (from slow clients) less heavily. Algorithm S12 implements polynomial staleness weighting: an update computed τ rounds ago receives weight $(1+\tau)^{-a}$, ensuring that fresher updates contribute more while still incorporating information from slower participants.

```
Algorithm S12: FedAsync with Staleness Weighting
Input: Client update $\Delta_k$, client round $t_k$, server round $t$
Output: Updated global model $\theta$
$\tau \leftarrow t - t_k$                          // Staleness
$\alpha \leftarrow (1+\tau)^{-a}$                  // Polynomial decay, $a > 0$
$\theta \leftarrow \theta + \alpha \cdot \eta \cdot \Delta_k$
return $\theta$
```

Staleness functions: Constant (α=1), Polynomial ($(1+\tau)^{-a}$), Exponential ($e^{-a\tau}$), Hinge (1 if $\tau \leq \tau_{max}$, else 0). Additional: FedBuff (buffered async), semi-async (wait for α-fraction of clients).

H. Fairness-Aware Federated Learning

The EHDS serves 450 million citizens across Member States with different population sizes, disease prevalence, and healthcare quality. Standard FL optimizes average performance, which can disproportionately favor large hospitals with more data while neglecting smaller institutions or underrepresented patient populations. This creates a "digital health equity" concern: a model that achieves 85% accuracy for a large German hospital but only 55% for a small Romanian clinic is not equitable. Algorithm S13 implements q-FedAvg, which reweights client contributions by their loss: hospitals where the model performs poorly receive higher aggregation weights, pulling the global model toward equitable performance across all participants.

```
Algorithm S13: q-FedAvg Fair Aggregation
Input: Losses $\{L_1, \ldots, L_K\}$, updates $\{\Delta_1, \ldots, \Delta_K\}$, $q$
Output: Fair aggregated update $\Delta$
for each client $k$ do
    $w_k \leftarrow L_k^q$                         // Higher loss → higher weight
$W \leftarrow \sum_k w_k$; $w_k \leftarrow w_k / W$
$\Delta \leftarrow \sum_k w_k \cdot \Delta_k$
return $\Delta$
```

Fairness metrics: Performance Variance (Var($\{L_k\}$)), Worst-case Loss ($\max_k L_k$), Demographic Parity Gap, Equalized Odds Gap. Additional methods: AFL, FedMGDA+, TERM, FairFed.

VII. Infrastructure Components

Deploying FL across 27 EU Member States requires production-grade infrastructure: reliable communication channels between hospitals and the SPE aggregator, efficient serialization of gradient tensors, distributed coordination for concurrent studies, and comprehensive monitoring with EHDS-specific alerting. This section describes the infrastructure components implemented in the reference framework, each designed to operate within the constraints of cross-border healthcare networks (firewalls, bandwidth limitations, regulatory requirements).

A. Communication Layer

The communication layer must bridge heterogeneous network environments: high-bandwidth data center connections between national HDABs, moderate hospital-to-aggregator links, and potentially bandwidth-constrained rural clinics. The framework supports two transport protocols selectable per deployment, with configurable compression and retry policies.

Communication Manager Configuration

```
transport: gRPC | WebSocket
compression: gzip | lz4 | zstd | none
chunk_size: 1MB
retry_policy:
  max_retries: 3
  backoff: exponential
  base_delay: 1s
connection_pool:
  max_connections: 100
  idle_timeout: 300s
```

gRPC: Bidirectional streaming, Protocol Buffers (30% bandwidth reduction vs. JSON), HTTP/2 multiplexing. Ideal for data center deployments.

WebSocket: Browser-compatible, firewall-friendly (standard HTTP upgrade), event-driven. Ideal for edge deployments and browser-based participation.

Selection criteria: gRPC is recommended for production EHDS deployments where both endpoints support HTTP/2 (typical for hospital-to-national aggregator links). WebSocket is preferred when traffic must traverse web application firewalls or when browser-based dashboards participate directly in federation monitoring.

B. Serialization

Binary Format: Tensor metadata + raw binary, 30% smaller than JSON, 15% smaller than pickle, cross-platform (Python, C++, Java).

Delta Serialization: Transmits only changed parameters, sparse encoding, up to 90% bandwidth reduction for fine-tuning.

EHDS-Compliant: Embeds permit ID, timestamp, provenance; cryptographic signatures; GDPR Article 30 audit fields.

C. Caching Layer

In production EHDS deployments, multiple FL studies may run concurrently on overlapping data holders. A distributed locking mechanism prevents race conditions during gradient aggregation—ensuring that two concurrent studies do not interfere with each other's model updates. Algorithm S14 implements Redis-based distributed locking with TTL-based automatic release, preventing deadlocks if a server node fails mid-aggregation.

Algorithm S14: Distributed Lock for Aggregation

```
Input: Lock name, TTL, client ID
Output: Lock acquired (boolean)
acquired $\leftarrow$ Redis.SET(lock_name, client_id, NX, EX=TTL)
if acquired then
    PerformAggregation()
    if Redis.GET(lock_name) == client_id then
        Redis.DEL(lock_name)
return acquired
```

Redis-based caching: model checkpoints, client states, real-time metrics. Features: LRU/LFU/TTL eviction, distributed locking, automatic serialization, cache warming.

D. Orchestration

Kubernetes: Deploys FL clients/aggregators as pods, HPA for elastic scaling, ConfigMaps for hyperparameters, Secrets for HDAB API keys.

Ray: Actor-based FL, automatic fault tolerance, Ray Tune for federated HPO, Object Store for gradient sharing.

Auto-Scaling: Reactive (queue depth/latency), Predictive (ML-based forecasting), Scheduled (time-based patterns).

E. Monitoring

Prometheus Metrics: Counters (rounds_total, permits_validated), Gauges (active_clients, privacy_budget_remaining), Histograms (round_duration, communication_latency), Summaries (gradient_norm_quantiles).

Grafana Dashboards: FL training progress, client health, latency heatmaps, privacy budget consumption, EHDS compliance status.

Alerting: Privacy budget exhaustion, client dropout threshold, model divergence, permit expiration.

F. Model Watermarking

IP protection for FL models trained on EHDS data: **Spread Spectrum** (frequency domain, robust to fine-tuning), **LSB** (low-order weight bits), **Backdoor-based** (input-output ownership proof), **Passport Layers** (dedicated ownership encoding).

G. Cross-Silo Enhancements

EHDS federations are inherently cross-silo: each participant is an institution (hospital, registry, research center) with significant computational resources, distinct data distributions, and long-term participation commitments. This differs from cross-device FL (e.g., mobile phones) and enables advanced optimization strategies.

Multi-Model Federation: Weighted voting, stacking, mixture of experts with diversity enforcement.

Automatic Algorithm Selection: The 17 FL algorithms in the framework have different strengths depending on the federation characteristics (heterogeneity level, number of participants, communication budget). Algorithm S15 implements adaptive aggregation selection via multi-armed bandit (UCB/Thompson Sampling), automatically switching algorithms mid-training if performance metrics indicate a better alternative. A cooldown period prevents oscillation between strategies.

Algorithm S15: Adaptive Aggregation

```
Input: Client updates, metrics history, cooldown
Output: Aggregated model, selected algorithm
score $\leftarrow$ WeightedScore(loss, accuracy, variance, conv.)
if RoundsSinceSwitch $>$ Cooldown then
    for each candidate $\in$ Algorithms do
        alt $\leftarrow$ EstimatePerformance(candidate)
        if alt $>$ score + Threshold then
            SwitchTo(candidate)
aggregated $\leftarrow$ CurrentAlgo.Aggregate(updates)
return aggregated
```

VIII. Extended EHDS Interoperability

A. OMOP Common Data Model

OMOP CDM v5.4 provides standardized analytical format used by European research networks (EHDEN, OHDSI).

ETL Pipelines: Transform source EHR to OMOP. **Vocabulary Mapping**: SNOMED, ICD10, LOINC, RxNorm. **Cohort Definitions**: ATLAS-compatible SQL generation. **Feature Extraction**: FeatureExtraction package for ML-ready datasets.

FL Integration: (1) Each hospital transforms local EHR to OMOP; (2) Feature extraction produces identical schema; (3) FL training on homogeneous feature spaces.

B. IHE Integration Profiles

ATNA: TLS mutual authentication, syslog audit messages (RFC 5424), maps to GDPR Article 30.

BPPC: Maps Article 71 opt-out to consent documents, XDS.b integration, consent enforcement at FL initiation.

XCA: Cross-border document query/retrieve, Initiating/Responding Gateways, patient identity correlation.

PIX/PDQ: Patient matching across boundaries, pseudonymization-aware identity management, national eHealth integration.

XUA: SAML 2.0 federated authentication, role-based access control, HDAB authorization token propagation.

C. Cross-Border Data Exchange

Message Formats: EHDS Data Permit Exchange Format (JSON-LD), Federated Query Protocol (SPARQL Federation), Model Update Message Format (Protocol Buffers).

Security: eIDAS-compliant electronic signatures, TLS 1.3, certificate-based authentication (EU trust framework).

Metadata: DCAT-AP Health extension, W3C PROV-O provenance, EMA data quality indicators.

D. Interoperability Architecture

Figure S-13 presents the complete interoperability architecture, showing how heterogeneous data sources across EU Member States are harmonized through multiple standards layers before reaching the FL training engine. The architecture reflects a key EHDS challenge: real-world healthcare institutions use diverse formats, terminologies, and exchange protocols that must be reconciled to produce a consistent feature space for federated model training.

IX. Clinical Imaging: Extended Details

A. Datasets

Three clinical imaging datasets cover representative EHDS scenarios:

- **Chest X-ray** [50]: 5,856 pediatric radiographs (NORMAL/PNEUMONIA, 2.7:1 imbalance)
- **Brain Tumor MRI**: 7,023 T1-weighted CE MRI slices (4-class: glioma, healthy, meningioma, pituitary)
- **Skin Cancer**: 3,297 dermoscopy images (binary benign/malignant)

B. Model Architectures

HealthcareResNet: ResNet-18 [49] pretrained on ImageNet, GroupNorm replacing BatchNorm for FL stability. FedBN [48] skips normalization during aggregation. Partial backbone freeze (level 1). ~11.2M parameters.

HealthcareCNN: 5-block CNN with GroupNorm, progressive channels (32→512), graduated Dropout (0.15→0.3). Classifier: Flatten→FC(512)→FC(128)→FC(K). ~12M parameters.

Data augmentation: random horizontal flip, rotation (±15°), brightness jitter (±10%). ImageNet normalization.

C. Experimental Configuration

Chest X-ray experiments use 4 algorithms (FedAvg, FedLESAM, Ditto, HPFL) × 3 seeds = 12 experiments. Brain Tumor and Skin Cancer use 2 algorithms (FedAvg, HPFL) × 1 seed as qualitative validation of the modality-dependent personalization effect; statistical significance is established on the tabular benchmarks (10 seeds, $p < 0.001$).

- 5 clients, Dirichlet α=0.5, Adam optimizer
- ResNet-18, FedBN, class-weighted loss, mixed precision
- Chest X-ray: 20 rounds, early stopping (patience=6), lr=0.001
- Brain Tumor / Skin Cancer: 10 rounds, early stopping (patience=3, min_rounds=6), lr=0.0005/0.001

D. Reproducibility

All experiments are fully reproducible:

```
cd fl-ehds-framework
# Chest X-ray (4 algos x 3 seeds, ~4h)
python -m benchmarks.run_imaging_extended
# Brain Tumor + Skin Cancer (2 algos x 1 seed, ~2.5h)
python -m benchmarks.run_imaging_multi --light
# Confusion matrices
python -m benchmarks.run_confusion_matrix_chest
python -m benchmarks.run_confusion_matrix_bc
```

Results, checkpoints, and logs are auto-saved to `benchmarks/paper_results_delta/`.

Repository: https://github.com/FabioLiberti/FL-EHDS-FLICS2026

E. Chest X-ray Results

Table S-III reports Chest X-ray results for four FL algorithms (ResNet-18, 5 clients, Dirichlet α=0.5, 20 rounds max with early stopping).

TABLE S-III
Chest X-ray pneumonia detection: federated accuracy (%) with ResNet-18. 5 clients, Dirichlet α=0.5, 20 rounds, Adam lr=0.001, early stopping (patience=6). Per-seed results with Jain fairness index.

Algo	s42	s123	s456	Mean	Jain
FedAvg	91.6	87.4	82.8	**87.3**	0.984
FedLESAM	91.6	87.8	84.0	87.8†	0.984
Ditto	84.8	65.1	90.1	80.0	0.958
HPFL	62.1	73.4	71.8	69.1	0.762

Jain index averaged over 3 seeds. †FedLESAM produces results identical to FedAvg (SAM perturbation ineffective on this architecture). HPFL early-stops at round 8 on all seeds with best accuracy at rounds 1–2.

Key findings. *First*, federated learning on Chest X-ray achieves **87.3% accuracy** (FedAvg) for binary pneumonia detection with ResNet-18, confirming FL viability for medical imaging under EHDS-relevant non-IID conditions. *Second*, **FedLESAM is identical to FedAvg on imaging**: despite SAM's theoretical sharpness-aware optimization, FedLESAM produces per-seed accuracies indistinguishable from FedAvg (87.8% vs. 87.3%), suggesting that SAM perturbation does not meaningfully alter the ResNet-18 loss landscape under federated aggregation. *Third*, **personalization is counterproductive on imaging**: Ditto achieves 80.0% (−7.3pp vs. FedAvg), with substantially higher variance (σ=10.6pp vs. 3.6pp). HPFL *fails catastrophically* (69.1%, −18.2pp), early-stopping at round 8 on all seeds with peak accuracy at rounds 1–2, indicating that the split-head architecture cannot learn shared representations on a 5-class imaging task with Dirichlet-partitioned data. HPFL's fairness degrades severely (Jain 0.762 vs. 0.984), with per-client accuracies ranging from 0% to 62% on the worst seed. *Fourth*, Ditto exhibits a **fairness failure mode** on seed 456: per-client accuracies of [99.7%, 100.0%, 48.5%, 39.5%, 100.0%] (Jain 0.888) indicate that personalized local models overfit to majority-class clients while failing on minority-class clients.

X. Detailed Architecture Description

This section provides a comprehensive technical specification of the FL-EHDS three-layer architecture, detailing all modules, services, protocols, and standards implemented in the reference framework. Figure 2 in the main paper provides the high-level view; Figure S-14 presents the detailed

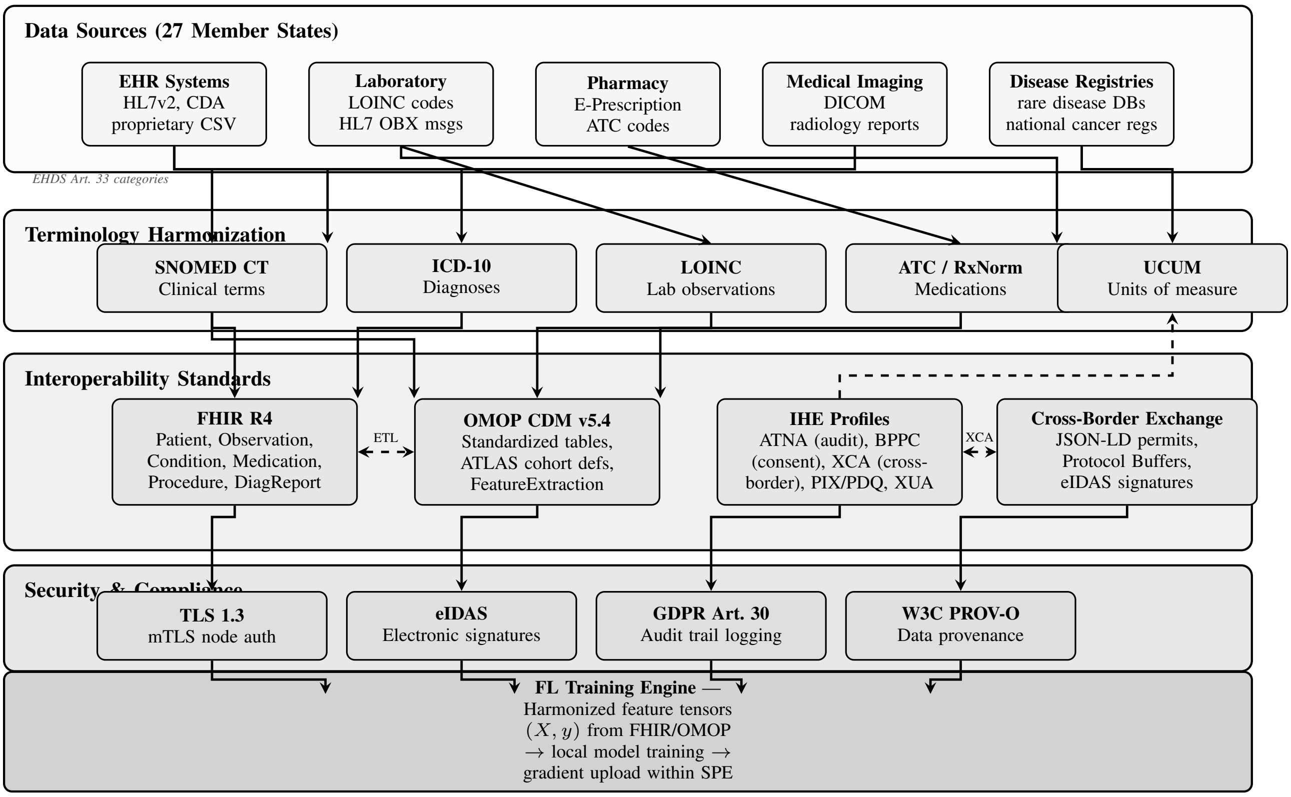


Fig. S-13. EHDS interoperability architecture for FL-based secondary use. Data from heterogeneous sources across 27 Member States (top) flows through terminology harmonization (SNOMED CT, ICD-10, LOINC, ATC, UCUM), then through interoperability standards (FHIR R4 for structured data exchange, OMOP CDM for observational research, IHE profiles for cross-institutional workflows, cross-border exchange protocols with eIDAS signatures). A security and compliance layer enforces TLS 1.3 mutual authentication, eIDAS electronic signatures for permits, GDPR Article 30 audit logging, and W3C PROV-O data provenance before harmonized feature tensors reach the FL training engine. Bidirectional ETL between FHIR and OMOP enables institutions to use either standard based on their existing infrastructure.

component-level diagram; the description below enumerates every component with its specific technical parameters.

A. Layer 1: Governance

Four principal modules comprise the governance layer:

1.1 HDAB Integration (per Member State). Each Health Data Access Body instance implements: OAuth2/mTLS authentication with bearer token management (refresh tokens, scopes: `permits:read`, `permits:write`); a permit store with SQLite persistence backend; configurable HDAB strictness level (scale 1–5, where DE=5, ES=2); and national privacy constraints including per-jurisdiction differential privacy bounds ($\varepsilon_{\max}$: DE=1.0, FR=3.0, IT=5.0).

1.2 Data Permit Manager (Article 53). Manages the complete permit lifecycle: PENDING → APPROVED → ACTIVE → EXPIRED/REVOKED. Validates permitted purposes (scientific research, public health surveillance, AI system development, personalized medicine, official statistics) against authorized data categories (EHR, lab results, imaging, genomic, registry, ECG, pathology). Implements expiry verification and strict mode enforcement for continuous compliance.

1.3 Opt-Out Registry (Article 71). Supports three granularity levels: record-level, patient-level, and dataset-level opt-out. Registry lookups use LRU caching (max 100K records, TTL 10 min) with 300-second synchronization intervals to national registries. Configurable actions on opt-out match: exclude, anonymize, or error. Pre-training filtering ensures only opted-in data enters gradient computation.

1.4 Cross-Border Coordinator (Articles 46, 50). Implements multi-HDAB synchronization protocols across 10 preconfigured EU country profiles (DE, FR, IT, ES, NL, SE, PL, AT, BE, PT). Each profile specifies per-country constraints: $\varepsilon_{\max}$ for differential privacy, data retention periods (365–1,825 days), opt-out rates (3–12%), and network latency characteristics (15–60ms based on geographic distance). The fee model (Article 42) implements cost-recovery calculation: base fee + per-record + per-round + per-MB charges.

Inter-layer interface: Data Permit Authorization (OAuth2/mTLS, purpose-validated, Articles 53+71) connects Layer 1 to Layer 2.

B. Layer 2: FL Orchestration

Five modules operate within the Secure Processing Environment (SPE), conforming to HealthData@EU infrastructure requirements:

2.1 Aggregation Engine. Implements 17 FL algorithms spanning six categories: *Baseline* (FedAvg, FedProx with μ=0.01, SCAFFOLD, FedNova, FedDyn); *Adap-*

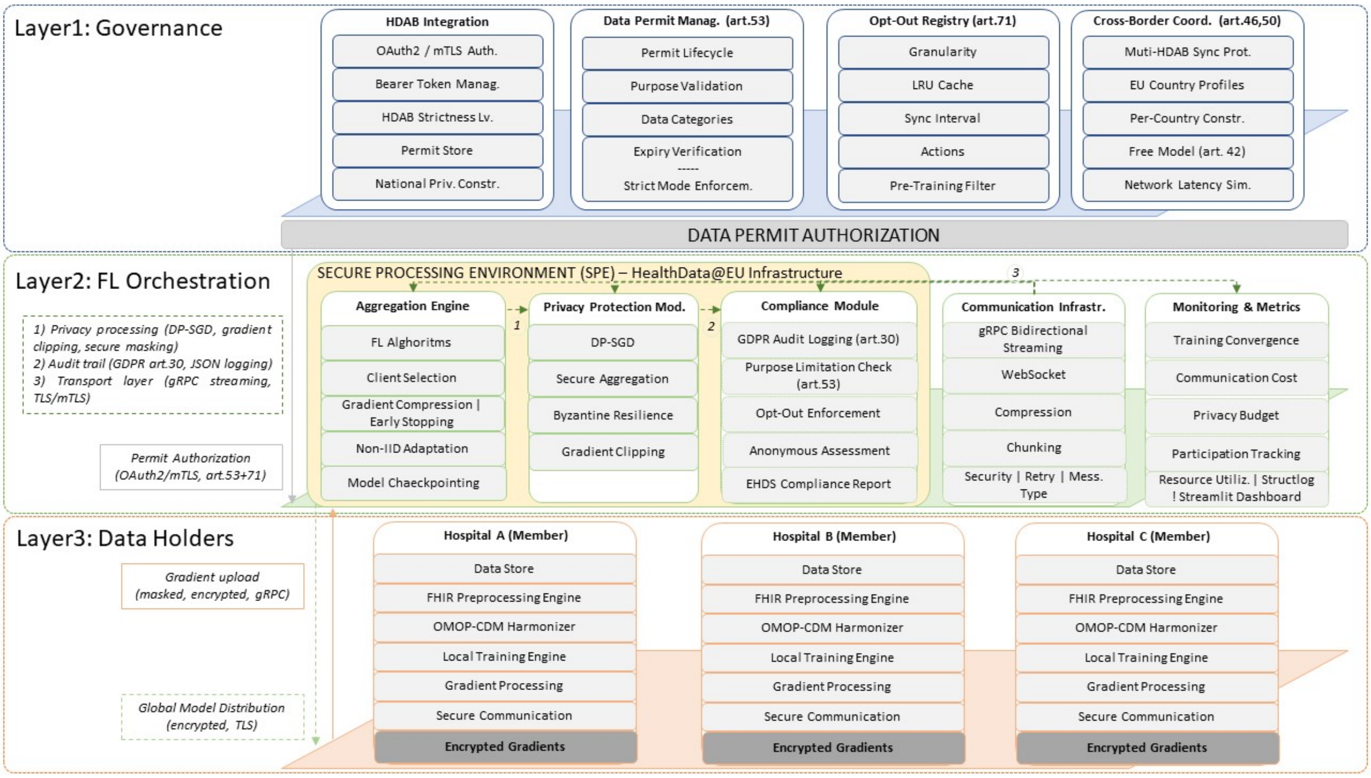


Fig. S-14. FL-EHDS detailed architecture with component-level specifications. Layer 1 (Governance) includes four modules: HDAB Integration with OAuth2/mTLS authentication, Data Permit Manager (Article 53) with lifecycle management, Opt-Out Registry (Article 71) with LRU-cached filtering, and Cross-Border Coordinator (Articles 46, 50) with per-country privacy constraints. Layer 2 (FL Orchestration) operates within the Secure Processing Environment (SPE) with five modules: Aggregation Engine (17 FL algorithms), Privacy Protection (DP-SGD, Secure Aggregation, Byzantine Resilience), Compliance Module (GDPR audit logging), Communication Infrastructure (gRPC/WebSocket), and Monitoring & Metrics. Layer 3 (Data Holders) implements a uniform stack per institution: Data Store, FHIR R4 Preprocessing, OMOP-CDM Harmonization, Local Training Engine, Gradient Processing, and Secure Communication. Inter-layer flows: Data Permit Authorization (OAuth2/mTLS, purpose-validated) downward from Layer 1; gradient upload (masked, encrypted, gRPC) upward and global model distribution (encrypted, TLS) downward between Layers 2–3.

tive (FedAdam with server_lr=0.1, FedYogi with β_2=0.99, FedAdagrad); *Personalization* (Ditto with λ=0.1, Per-FedAvg via MAML, pFedMe, MOON); *Non-IID* (FedLC for logit calibration, FedDecorr); *SAM* (FedSAM, FedSpeed); *Advanced* (FedExP ICLR'23, FedLESAM ICML'24, HPFL ICLR'25). Client selection strategies: random, performance-based, fairness-aware, latency-aware. Early stopping: patience=10, min_delta=0.001. Model checkpointing with atomic saves.

2.2 Privacy Protection Module. Three sub-modules provide defense-in-depth:

- *DP-SGD*: Gaussian mechanism with Rényi DP (RDP) accounting. Configurable ε-budget (default 1.0), δ=10^{-5}. Gradient clipping: max_norm=1.0, type=L2/L∞. RDP composition provides 5–6× tighter bounds than naive composition over 100+ rounds.
- *Secure Aggregation*: Pairwise masking protocol with ECDH key exchange (SECP384R1 curve). Alternative protocols: Shamir secret sharing (threshold reconstruction), homomorphic encryption (CKKS scheme via TenSEAL). Dropout threshold: 50%.
- *Byzantine Resilience*: Six defense rules (Krum, Multi-Krum, Trimmed Mean, Coordinate-wise Median, Bulyan, FLTrust). Anomaly detection at 3σ threshold. TEE integration for SGX/TrustZone hardware attestation.

2.3 Compliance Module. GDPR audit logging (Article 30) in JSON format with 7-year retention (2,555 days). Per-round purpose limitation check (Article 53). Opt-out enforcement: pre-training filtering combined with per-round verification. Anonymity assessment: k-anonymity, l-diversity checks. Automated EHDS compliance report with verification scoring.

2.4 Communication Infrastructure. gRPC bidirectional streaming for model updates (primary protocol). WebSocket for real-time monitoring and events. Compression: GZIP, LZ4, ZSTD, Snappy. Model chunking (ResNet-18: 44.7 MB/round $\rightarrow$ 8.9 MB with Top-k 1%). Security: TLS/mTLS with ECDHE key exchange. Retry: 3 attempts with 2× exponential backoff. Message types: MODEL_UPDATE, GRADIENT_UPDATE, HEARTBEAT, PERMIT_VALIDATION, CONSENT_CHECK, AUDIT_LOG.

2.5 Monitoring & Metrics. Tracks training convergence (accuracy, loss, F1, AUC per round), communication cost (MB/round), privacy budget consumption (ε spent per round), participation statistics (samples, opt-outs), and resource utilization (CPU, memory, GPU). Structured logging via Structlog (JSON format) with Streamlit dashboard integration.

Inter-module flows: Privacy processing (DP-SGD, gradient clipping, secure masking) connects Aggregation → Privacy. Audit trail (GDPR Article 30, JSON logging, 7-year retention) connects Privacy → Compliance. Transport layer (gRPC streaming, TLS/mTLS, GZIP compression) connects Communication ↔ all modules.

C. Layer 3: Data Holders

Each data holder institution (hospitals, disease registries, research centers across 27 Member States) implements a uniform component stack:

3.1 Data Store. Institutional health data in heterogeneous formats: EHR (FHIR R4), vitals (LOINC-coded), diagnoses (ICD-10 variants per country: ICD-10-GM for Germany, CIM-10 for France, ICD-9-CM legacy for Italy), ECG records (SCP-ECG standard EN 1064), and medical imaging (DICOM).

3.2 FHIR R4 Preprocessing Engine. Processes six FHIR resource types: Patient, Observation, Condition, MedicationRequest, Procedure, DiagnosticReport. Terminology mapping to international standards: SNOMED-CT (clinical concepts), LOINC (laboratory codes), ICD-10 (diagnoses), ATC (medications). Extracts 36 standardized features with normalization, encoding, and missing value imputation.

3.3 OMOP-CDM Harmonizer. Maps local vocabulary codes to standard OMOP Concept IDs. Populates OMOP tables: Person, ConditionOccurrence, Measurement, DrugExposure, ProcedureOccurrence, VisitOccurrence. Per-country vocabulary mapping ensures cross-border semantic compatibility.

3.4 Local Training Engine. Adaptive training with configurable parameters: optimizer (Adam, lr=0.001), batch size (32–256, dynamically adjusted), local epochs (3–5), dropout (0.3), L2 regularization. Device support: CUDA (GPU), MPS (Apple Silicon), CPU fallback for resource-constrained institutions.

3.5 Gradient Processing. Three-stage pipeline: (1) gradient clipping (max_norm=1.0, L2 norm); (2) local DP noise addition (Gaussian, calibrated to ε); (3) pairwise masking (ECDH shared keys with counterpart clients).

3.6 Secure Communication. AES-256-GCM symmetric encryption for gradient payloads. ECDHE key exchange (SECP384R1 curve). mTLS mutual authentication with certificate-based identity. Nonce generation for replay attack prevention.

Inter-layer data flow: Gradient upload (∇ masked, encrypted, gRPC) flows upward from Layer 3 to Layer 2. Global model distribution (θ encrypted, TLS) flows downward from Layer 2 to Layer 3.

Architectural invariant: Raw health data never leaves institutional boundaries—only encrypted model gradients are exchanged within the Secure Processing Environment.

D. Deployment Readiness Assessment

Table S-IV provides a transparent assessment of each FL-EHDS component's production readiness, distinguishing between fully implemented modules, simulation backends, and components requiring external integration.

TABLE S-IV
FL-EHDS COMPONENT READINESS FOR EHDS PRODUCTION DEPLOYMENT. ✓: PRODUCTION-READY. ∼: SIMULATION BACKEND (FUNCTIONAL, REQUIRES BINDING TO EXTERNAL SERVICES). ○: REQUIRES EXTERNAL INFRASTRUCTURE.

Component	**Status**	**Binding Req.**
Layer 2: FL Orchestration		
17 FL algorithms	✓	None
DP-SGD + RDP accounting	✓	None
Secure aggregation	✓	None
Byzantine resilience	✓	None
Monitoring dashboard	✓	None
Layer 3: Data Holders		
Local training engine	✓	None
FHIR R4 preprocessing	✓	FHIR server URL
OMOP-CDM harmonization	✓	CDM schema
Secure communication	✓	TLS certificates
Layer 1: Governance		
HDAB permit lifecycle	∼	HDAB REST/gRPC
Art. 71 opt-out registry	∼	National registry API
Cross-border coordinator	∼	Multi-HDAB endpoints
Art. 53 purpose limitation	∼	Permit schema
GDPR audit trail	✓	None

Simulation backends implement complete functional logic (OAuth2/mTLS auth, permit CRUD, LRU-cached registry lookups) and require only endpoint configuration—not architectural changes—for production binding. HDAB services expected 2027–2029.

XI. Extended Threat Model and Security Analysis

The main paper summarizes the FL-EHDS threat model in three adversary classes. This section provides a comprehensive security analysis with formal definitions, attack taxonomy, defense mapping, and explicit boundary conditions relevant to EHDS cross-border deployments.

A. Adversary Model

We formalize three adversary classes ordered by increasing capability.

A1: Honest-but-curious aggregation server. The SPE aggregation server $\mathcal{S}$ follows the protocol faithfully but attempts to infer patient-level information from the received model updates $\{\Delta_k\}_{k=1}^K$. This is the most realistic adversary for EHDS: the HDAB-operated SPE has institutional incentives to follow the protocol (legal liability, regulatory oversight) but may be compromised or subject to insider threats. Concrete attack vectors include:

- *Gradient inversion* [21]: reconstructing training samples from observed gradients. Effectiveness decreases with batch size and model depth; for the tabular MLP (∼2.9K–10K parameters depending on dataset, batch size 32–64), theoretical reconstruction quality is limited but non-zero.
- *Membership inference*: determining whether a specific patient record was used in a client's local training, potentially violating GDPR Article 9 (special category data).

- *Property inference*: extracting aggregate properties of a client's local dataset (e.g., disease prevalence at a specific hospital).

Defenses for A1: Central DP (Algorithm S2) with Gaussian mechanism ($\sigma = C/\varepsilon$, $\delta = 10^{-5}$) bounds per-round information leakage. Rényi DP composition (Algorithm S8) tracks cumulative privacy expenditure across rounds. Secure aggregation (Algorithm S4) ensures $\mathcal{S}$ observes only $\sum_k \Delta_k +$ noise, never individual Δ_k. Our DP ablation (Table S-X) demonstrates that at $\varepsilon = 10$, personalized methods lose <2pp accuracy—privacy imposes negligible utility cost at this budget.

A2: Malicious clients (Byzantine). Up to $f < n/3$ compromised institutions may deviate arbitrarily from the protocol. In the EHDS context, this could represent: institutions with corrupted data pipelines, nation-state actors attempting to bias cross-border models, or free-riding participants that send random updates to benefit from the global model without contributing genuine local training. Attack types include:

- *Model poisoning*: submitting adversarial gradients designed to degrade global model accuracy (untargeted) or introduce backdoors for specific inputs (targeted).
- *Data poisoning*: manipulating local training data (label flipping, sample injection) to corrupt the learned model through legitimate protocol participation.
- *Free-riding*: sending random or stale gradients to receive the global model without incurring computational cost, degrading convergence.

Defenses for A2: Six Byzantine-resilient aggregation rules are implemented (Section VI): Krum (Algorithm S10), Trimmed Mean, Coordinate-wise Median, Bulyan (two-stage Krum + trimmed mean), FLTrust (server-guided trust weighting), and FLAME (clustering-based). The framework supports attack simulation (label flipping, gradient scaling, additive noise, sign flipping, model replacement) for defense validation.

A3: External attacker. An adversary with black-box access to model outputs (e.g., through a compromised data permit or leaked model weights) may attempt membership or attribute inference against the published model. Unlike A1, this adversary does not observe training dynamics but only the final or intermediate model checkpoints.

Defenses for A3: Article 71 output filtering restricts model access to authorized queries specified in the HDAB data permit, limiting the number of queries available for inference attacks. The permit system (Algorithm S3) enforces purpose limitation (Article 53), temporal validity, and data category authorization, reducing the attack surface. Model watermarking (Section VII) provides post-hoc attribution for leaked models.

B. Protected Assets

Table S-V maps the three categories of protected assets to their defense mechanisms and the EHDS articles that mandate their protection.

TABLE S-V
PROTECTED ASSETS AND DEFENSE MAPPING

Asset	Defense	EHDS Article
Raw patient data	Never leaves institution	Art. 50 (SPE)
Gradient updates	DP + secure aggregation	Art. 50, GDPR 32
Global model	HDAB permit access control	Art. 53, 71
Audit trail	Immutable logging	GDPR Art. 30
Privacy budget	RDP accountant (Alg. S8)	Art. 50

C. Out-of-Scope Threats and Boundary Conditions

We explicitly delineate threats that the current FL-EHDS implementation does not address, along with justifications and potential future mitigations.

Server-client collusion. If the aggregation server $\mathcal{S}$ colludes with one or more clients, secure aggregation guarantees are invalidated: the colluding parties can reconstruct non-colluding clients' individual gradients by subtracting known values from the aggregate. This threat undermines the trust model of *any* FL system and requires stronger cryptographic primitives (fully homomorphic encryption, multi-party computation with dishonest majority) that impose 100–1000× computational overhead, currently impractical for healthcare-scale deployments. In the EHDS context, collusion is partially mitigated by the institutional separation between HDABs (which operate the SPE) and data holders (healthcare institutions), as these are distinct legal entities with separate governance structures.

Side-channel attacks. Timing analysis of gradient uploads, memory access patterns, or power consumption during local training could leak information about dataset properties. These attacks require physical or network-level proximity to the target institution and are addressed by infrastructure-level security (network isolation, constant-time implementations) rather than the FL protocol itself. EHDS Article 50 SPE specifications should mandate side-channel-resistant computing environments.

Covert-channel leakage through model architecture. The choice of model architecture, hyperparameters, and training configuration can inadvertently encode information about the training data [23]. For example, a model's capacity (number of parameters) relative to dataset size affects memorization risk. This remains an open problem in FL privacy research; we mitigate it partially through standardized configurations (same architecture and hyperparameters across all clients) and DP noise injection [23].

Reconstruction attacks with auxiliary data. If an adversary possesses auxiliary information about a target individual (e.g., partial medical records from a data breach), gradient inversion and membership inference attacks become significantly more effective. Our threat model assumes adversaries operate within the bounds of their adversary class (A1–A3) without additional auxiliary datasets. In practice, EHDS cross-border settings increase this risk, as health data from one Member State could serve as auxiliary information for attacks on another Member State's data.

Model extraction and intellectual property. A permitted user could systematically query the model to create a functionally equivalent copy, bypassing the HDAB access controls for subsequent use. This is partially addressed by query rate limiting and model watermarking, but a determined adversary with sufficient queries can always approximate any black-box model.

XII. Framework Positioning: FL-EHDS vs. Existing Platforms

The main paper's Table I provides a feature-level comparison between FL-EHDS and existing FL frameworks (Flower [45], FLARE [46], PySyft, FedML, TFF). A natural reviewer question is why we do not provide a direct *experimental* comparison. This section explains the rationale.

Why no experimental comparison is provided. An experimental comparison between FL frameworks is meaningful only when the frameworks provide *different* capabilities on a *common* task. For pure FL algorithm performance (e.g., FedAvg accuracy on PTB-XL), all frameworks are equivalent by design: they implement the same canonical algorithms (FedAvg, FedProx, etc.) with identical mathematical formulations. Running FedAvg on Flower vs. FL-EHDS would produce statistically indistinguishable results, as both execute the same aggregation logic on the same data. The comparison would measure implementation overhead (framework startup time, serialization efficiency), not scientific contribution.

The governance gap. FL-EHDS's contribution is the governance-technical integration layer—the components that *no* existing framework provides:

- *HDAB permit lifecycle*: application, validation per round (Algorithm S3), temporal expiry, revocation, audit trail. Neither Flower nor FLARE implement permit-based access control.
- *Article 71 output filtering*: automated enforcement of opt-out registries for rare disease patients. No existing framework addresses EHDS opt-out requirements.
- *Article 53 purpose limitation*: technical enforcement of permitted secondary use purposes. Existing frameworks treat all training as equally authorized.
- *EHDS-compliant audit trails*: immutable logging of every round's permit status, privacy budget consumption, and data category access. This is a regulatory requirement (GDPR Article 30) that no FL framework satisfies out of the box.
- *Cross-border coordination*: multi-HDAB orchestration for Article 46 cross-border processing with per-Member-State governance constraints.

Since these governance components have no counterpart in existing frameworks, a "comparison" would reduce to: (a) identical FL performance metrics, plus (b) a checklist of governance features present in FL-EHDS and absent in alternatives—which is precisely what Table I already provides.

Complementarity, not competition. FL-EHDS is designed as a compliance layer that could, in principle, wrap existing FL engines. The governance modules (permit validation, audit trails, privacy budget accounting, output filtering) are protocol-agnostic and could be integrated with Flower's communication backend or FLARE's clinical deployment infrastructure. The architectural contribution is the *integration pattern*, not the replacement of existing FL platforms.

Additional frameworks. Two frameworks merit specific discussion given FLICS's industrial focus. *OpenFL* (Intel) provides healthcare-specific FL with production deployments in clinical settings (e.g., Intel's collaboration with the University of Pennsylvania for brain tumor segmentation), but lacks EHDS governance integration: no HDAB permit lifecycle, opt-out registry, or Article 53 purpose limitation. *FATE* (WeBank) offers industrial FL with governance features including role-based access control and audit logging, but is oriented toward financial services regulation (not healthcare/EHDS) and does not implement FHIR R4 interoperability or EU-specific data governance. Both could serve as FL compute backends beneath FL-EHDS's governance layer, reinforcing our complementarity design.

XIII. Cross-Border Governance Challenges: Extended Analysis

The main paper argues that legal uncertainties—not technical barriers—constitute the critical blocker for FL adoption under the EHDS, and includes a condensed analysis of the cross-border privacy budget conflict (Section V-A). This section provides the extended analysis with additional scenarios—controller/processor ambiguity and gradient data classification—that illustrate how unresolved governance questions create implementation deadlocks that no engineering solution can circumvent.

A. The Privacy Budget Conflict

Consider a cross-border federation between a German hospital (operating under the Bundesdatenschutzgesetz, BDSG) and an Italian hospital (under the Codice in materia di protezione dei dati personali, D.Lgs. 196/2003 as amended). The German Data Protection Authority (BfDI) may interpret GDPR Article 89 to require $\varepsilon_{\max} = 1.0$ for health data secondary use, reflecting the "data minimization" principle. The Italian Garante per la protezione dei dati personali may permit $\varepsilon_{\max} = 5.0$ under a "proportionality" interpretation that balances research utility against privacy risk.

This creates a concrete governance deadlock:

- If the federation applies $\varepsilon = 1.0$ (strictest bound), Italian data contributes under unnecessarily restrictive conditions, reducing model utility. Our DP experiments (Table S-X) show that FedAvg accuracy on PTB-XL drops from 91.9% (no DP) to 52.3% at $\varepsilon = 1$—a 39.6pp collapse.
- If the federation applies $\varepsilon = 5.0$ (most permissive bound), German data may be processed in violation of BfDI

guidance, exposing the German institution to regulatory sanctions.

- A per-client ε approach (each hospital applies its own budget) introduces asymmetric noise levels, creating fairness concerns: the German hospital's model updates are noisier, potentially biasing the global model away from the German patient population.

FL-EHDS addresses this technically through per-client privacy budget configuration in the governance layer, but the *policy question*—which bound applies—requires regulatory clarification in the March 2027 delegated acts. Article 50(4) of Regulation (EU) 2025/327 mandates that SPEs provide "a high level of security," but does not specify whether privacy budgets should be harmonized across Member States or determined by the most restrictive participant.

B. Controller/Processor Ambiguity

Under GDPR Articles 26 and 28, the roles of data controller and data processor carry distinct legal obligations. In a federated learning architecture:

- Each hospital is clearly the *controller* of its local patient data.
- The HDAB operating the aggregation server may be a *processor* (processing gradients on behalf of hospitals) or a *joint controller* (determining the purposes and means of the aggregation).
- If gradient updates are classified as personal data (an unresolved question), the aggregation server becomes a processor of personal data, requiring a Data Processing Agreement (DPA) with each participating hospital across 27 Member States.

The practical consequence: a 5-hospital federation across 3 Member States would require bilateral DPAs between each hospital and the HDAB, reviewed by 3 national DPAs, before a single training round can execute. This administrative overhead can delay projects by 12–18 months, as observed in current cross-border health research projects [5].

C. Gradient Data Classification

Whether model gradients constitute "personal data" under GDPR Article 4(1) remains unsettled. Zhu et al. [21] demonstrate that gradients can reconstruct training images, suggesting they encode personal information. However, with DP noise ($\varepsilon = 10$, $\delta = 10^{-5}$), the reconstruction fidelity degrades substantially, potentially rendering the gradients "anonymous" under Recital 26.

This classification has cascading regulatory implications:

- *If personal data*: full GDPR applies to gradient transmission, requiring lawful basis (Article 6), special category safeguards (Article 9), Data Protection Impact Assessments (Article 35), and cross-border transfer mechanisms (Chapter V).
- *If anonymous*: GDPR does not apply, and gradients can be transmitted freely—but the anonymization claim must be defensible against re-identification attacks, which our threat model (Section XI) acknowledges as an open problem.

The EHDS Regulation (EU) 2025/327 does not resolve this question: Article 50 defines SPE security requirements but does not classify the privacy status of intermediate computational artifacts such as gradients.

D. Implications for FL-EHDS Design

These governance challenges directly motivated FL-EHDS's architecture:

- The per-client privacy budget in the governance layer (Algorithm S8) is designed to accommodate heterogeneous national requirements, enabling the German hospital to operate at $\varepsilon = 1.0$ while the Italian hospital uses $\varepsilon = 5.0$—though the policy legitimacy of this approach awaits regulatory guidance.
- The HDAB permit system (Algorithm S3) externalizes the controller/processor determination: the HDAB issues the permit (acting as controller of the secondary use authorization) while each hospital retains controller status over its data. This separation is architecturally explicit but legally untested.
- The immutable audit trail provides the evidentiary basis for demonstrating compliance under either gradient classification scenario, satisfying GDPR Article 30 requirements regardless of the ultimate regulatory interpretation.

These examples demonstrate that the "legal uncertainties as critical blocker" claim in the main paper is not abstract: it reflects specific, concrete impasses that affect system design decisions at the architectural level.

XIV. Extended Results from Main Paper

This section contains detailed results referenced inline in the main paper but moved here for space constraints.

A. Heart Disease and Diabetes Algorithm Comparison

Table S-VI presents the full algorithm comparison on two additional clinical datasets: Heart Disease UCI (4 natural hospitals) and Diabetes 130-US (5 Dirichlet-partitioned clients). These results confirm the personalization advantage observed on PTB-XL, Cardiovascular, and Breast Cancer in the main paper.

TABLE S-VI
FL Algorithm Comparison on Heart Disease and Diabetes

	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 (4 international hospitals: Cleveland, Hungarian, Swiss, VA Long Beach). Diabetes: Dirichlet α=0.5. Mean ± std over 5 seeds. Ditto achieves 75.1% vs. FedAvg 62.5% (12.6pp gap) on Heart Disease and 71.7% on Diabetes. SCAFFOLD and FedNova exhibit known failure modes on Diabetes (11.2% and 13.0%).

B. Per-Hospital Heterogeneity Analysis

Figure S-15 shows per-hospital accuracy variation on Heart Disease, where the four hospitals have naturally different patient populations.

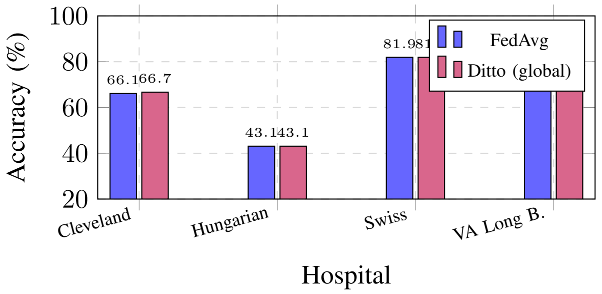


Fig. S-15. Per-hospital accuracy of the *global* model on Heart Disease UCI. Ditto's 12.6pp overall advantage (Table S-VI) comes from its *personalized local* models, which are separately fine-tuned per hospital; the shared global model shows similar cross-hospital performance to FedAvg. The Hungarian hospital, with the smallest cohort (n=294), shows the largest performance gap—a realistic EHDS scenario where smaller national datasets benefit most from federation.

XV. Extended Tabular Experiment Results

This section presents comprehensive experimental results from 1,740+ federated learning experiments across three tabular healthcare datasets: PTB-XL ECG (21,799 records, 52 European recording sites, 5-class cardiac diagnosis), Cardiovascular Disease (70,000 patients, binary classification), and Breast Cancer Wisconsin (569 samples, binary classification). The evaluation comprises a baseline comparison (105 experiments, 7 algorithms × 3 datasets × 5 seeds), three sweep phases—heterogeneity (α variation, 560 experiments), client scaling (385 experiments), and learning rate sensitivity (180 experiments)—a differential privacy ablation (180 experiments, 4 ε levels × 3 algorithms × 3 datasets × 5 seeds), an extended statistical validation (105 additional experiments with 5 new seeds for 10-seed Wilcoxon signed-rank tests), and an Article 71 opt-out impact analysis (225 experiments, 5 opt-out rates × 3 algorithms × 3 datasets × 5 seeds). All results are reproducible via the benchmark suite.

A. Heterogeneity Impact

Table S-VII shows accuracy as a function of Dirichlet α (lower α = more non-IID). The "Site" column reports natural site-based partitioning for PTB-XL.

Key finding: Personalized algorithms (Ditto, HPFL) exhibit a counter-intuitive pattern: accuracy *increases* under extreme non-IID (α=0.1) on Cardiovascular (92.4% vs. 82.5% at α=0.5) and Breast Cancer (88.3% vs. 79.1%). This occurs because personalized methods maintain local decision boundaries that become more specialized—and therefore more accurate—when client distributions are highly distinct. In contrast, baseline algorithms (FedAvg, FedExP, FedLESAM) degrade monotonically as α decreases, consistent with theoretical predictions. For PTB-XL, natural site-based partitioning produces the most realistic non-IID conditions and yields strong performance across all algorithms (89.7–96.5%).

B. Client Scaling

Table S-VIII reports accuracy as a function of client count K.

Key finding: Baseline algorithms (FedAvg, FedExP, FedLESAM) degrade as client count increases on Cardiovascular (72.1% at K=3 → 69.4% at K=10), because more clients means less data per client and higher aggregation noise. Personalized methods show the opposite trend: Ditto *improves* from 80.9% (K=3) to 85.0% (K=10), demonstrating that personalization enables each client to exploit its local specialization even with smaller data partitions. This has direct EHDS implications: as more Member States join a federation, personalized algorithms become increasingly advantageous.

C. Learning Rate Sensitivity

Table S-IX evaluates robustness to learning rate selection for the top-3 algorithms.

Key finding: All algorithms are sensitive to learning rate at the low end (lr=0.0005), where convergence is incomplete within the configured round budget. HPFL and Ditto achieve near-optimal performance across a wider range (0.005–0.01), while FedAvg requires careful tuning. On Breast Cancer, HPFL reaches 89.7% accuracy at lr=0.005—substantially higher than FedAvg's best of 74.7% at lr=0.01—confirming that personalized methods are more robust to hyperparameter selection on small datasets.

D. Privacy-Utility Tradeoff (Differential Privacy)

Table S-X quantifies accuracy under central differential privacy with varying privacy budget ε. We evaluate 3 representative algorithms: FedAvg (baseline), Ditto (best personalized), and HPFL (best fairness). All experiments use the Gaussian mechanism with clip norm C=1.0 and δ=10^{-5}.

Key findings: (1) *Personalized methods are remarkably DP-robust*: At ε=10, Ditto and HPFL lose <1pp on PTB-XL and <1.5pp on Cardiovascular, while FedAvg collapses to 52.3% on PTB-XL at ε=1 (−39.6pp). This robustness arises because personalized local adaptation (Ditto's fine-tuning, HPFL's classifier heads) is unaffected by central DP noise injected during aggregation. (2) *DP noise as implicit regularization on small data*: On Breast Cancer (569 samples), FedAvg with ε=5 achieves 78.7%—*higher* than the no-DP baseline of 52.3% (+26.4pp). HPFL at ε=1 reaches 86.9% vs. 74.1% without DP (+12.8pp). This counter-intuitive result is consistent with the theoretical analysis of Wei et al. [36], who show that DP noise interacts with model convergence in non-trivial ways, and with the broader machine learning literature on noise injection as regularization (e.g., dropout, label smoothing). On this small dataset, the Gaussian noise added during DP-SGD prevents the global model from overfitting to the majority class, acting as a stochastic regularizer that improves generalization. This effect is strongest at moderate ε (5–10) and diminishes at very strict ε=1 where noise overwhelms the signal on

TABLE S-VII
IMPACT OF DATA HETEROGENEITY (α) ON FL ACCURACY (%). LOWER α = MORE NON-IID. MEAN OVER 5 SEEDS. PX = PTB-XL, CV = CARDIOVASCULAR, BC = BREAST CANCER.

DS	Algorithm	IID	α=0.1	α=0.3	α=0.5	α=1.0	Site
PX	FedAvg	90.8	89.8	91.2	90.8	91.7	91.9
	FedProx	91.0	89.7	91.3	90.7	91.7	91.6
	Ditto	91.7	**96.5**	**96.5**	**95.4**	93.9	91.8
	FedLC	90.7	89.7	91.2	90.4	91.4	91.9
	FedExP	90.8	89.8	91.2	90.8	91.7	92.0
	FedLESAM	90.8	89.8	91.2	90.8	91.7	91.9
	HPFL	91.3	96.3	96.2	95.2	**94.0**	**92.5**
CV	FedAvg	72.7	61.2	65.7	71.1	72.6	–
	FedProx	72.7	62.5	65.9	71.6	72.5	–
	Ditto	72.6	**92.4**	**90.7**	**82.5**	81.5	–
	FedLC	72.7	61.9	66.0	71.1	72.6	–
	FedExP	72.7	61.2	65.7	71.1	72.6	–
	FedLESAM	72.7	61.2	65.7	71.1	72.6	–
	HPFL	**72.7**	92.4	90.7	82.3	**81.4**	–
BC	FedAvg	47.1	62.1	57.3	52.3	51.5	–
	FedProx	47.1	62.1	57.3	52.3	51.5	–
	Ditto	51.5	**88.3**	**84.8**	**79.1**	72.4	–
	FedLC	47.3	63.2	57.2	52.1	51.5	–
	FedExP	47.1	62.1	57.3	52.3	51.5	–
	FedLESAM	47.1	62.1	57.3	52.3	51.5	–
	HPFL	51.9	88.3	84.8	74.1	**66.9**	–

larger datasets. *Practical implication*: for small-cohort EHDS studies (rare diseases, specialist registries), moderate DP may simultaneously enhance privacy *and* model quality—a finding that should inform data permit conditions. (3) *Privacy is nearly free at ε=10*: Across PTB-XL and Cardiovascular, all algorithms recover to within 2pp of their no-DP baselines at ε=10—a practical privacy level that satisfies EHDS Article 50 requirements with minimal utility cost.

E. Fairness Analysis

Table S-XI provides per-client fairness metrics across all datasets.

Key finding: HPFL uniquely improves fairness on Breast Cancer (Jain 0.867 vs. 0.608 for all other algorithms, Gini reduction from 0.405 to 0.159). This occurs because HPFL's personalized classifier heads enable each client to specialize, reducing the performance gap between clients with different class distributions (Gap reduced from 71.5% to 47.6%). On PTB-XL, all algorithms achieve near-perfect fairness (Jain $\geq$ 0.999) due to the large dataset size providing sufficient per-client samples.

F. Statistical Significance

Table S-XII reports Wilcoxon signed-rank test p-values comparing each algorithm against FedAvg, computed over 10 seeds (original seeds: 42, 123, 456, 789, 999; additional seeds: 7, 31, 137, 577, 1337). With n=10 paired observations, the minimum achievable two-sided p-value is $2/2^{10} = 0.00195$. Table S-XIII supplements p-values with effect sizes (rank-biserial correlation r and mean accuracy difference $\overline{\Delta}$) for the two statistically significant algorithms.

Key findings: With 10 seeds, the statistical picture sharpens decisively. *First*, **HPFL significantly outperforms FedAvg on all three datasets** (p = 0.004, 0.002, 0.031; all < 0.05), with improvements of +0.9pp on PTB-XL, +11.8pp on Cardiovascular, and +16.5pp on Breast Cancer—the only algorithm achieving significance across all datasets. All effect sizes are large ($r \geq 0.93$; Table S-XIII). *Second*, **Ditto significantly outperforms FedAvg on Cardiovascular and Breast Cancer** (p = 0.002, 0.016), with large effect sizes (r = 1.00, 1.00; $\overline{\Delta}$ = +11.9pp, +19.0pp), though not on PTB-XL (p = 0.49, Δ = +0.2pp) where FedAvg already achieves 91.7%. *Third*, **five algorithms collapse to FedAvg-equivalent performance**: FedLESAM produces identical results on all datasets; FedExP and FedProx are identical on 2/3 datasets. This confirms that only methods maintaining separate local models (Ditto, HPFL) differentiate on the near-convex tabular MLP landscape. *Fourth*, the **pooled analysis** across all 30 paired observations yields $p < 0.001$ for both Ditto ($r = 0.89$) and HPFL ($r = 0.95$), providing strong evidence that personalized FL methods are superior for EHDS tabular analytics.

G. Article 71 Opt-Out Impact

EHDS Article 71 grants citizens the right to opt out of secondary use of their health data. To quantify the impact on model quality, we simulate opt-out by randomly removing 5%, 10%, 20%, and 30% of training samples per client (test data unchanged). This mirrors real-world scenarios where patients exercise their opt-out right, reducing the available training data at each institution. We evaluate FedAvg (baseline), Ditto, and HPFL (the two statistically significant personalized methods) across all three datasets with 5 seeds (225 experiments total).

TABLE S-VIII
IMPACT OF CLIENT COUNT ON FL ACCURACY (%). MEAN OVER 5 SEEDS. PX = PTB-XL, CV = CARDIOVASCULAR, BC = BREAST CANCER.

PTB-XL ECG (5-class)

Algorithm	K=3	K=5	K=10	K=20
FedAvg	91.7	91.9	91.5	91.5
FedProx	91.7	91.6	91.5	91.5
Ditto	92.5	91.8	91.5	91.5
FedLC	91.8	91.9	91.8	91.8
FedExP	91.7	92.0	92.2	92.2
FedLESAM	91.7	91.9	91.5	91.5
HPFL	**92.6**	**92.5**	**92.5**	**92.5**

Cardiovascular (binary)

Algorithm	K=3	K=5	K=8	K=10
FedAvg	72.1	71.1	70.5	69.4
FedProx	72.0	71.6	70.7	69.6
Ditto	80.9	82.5	83.3	**85.0**
FedLC	72.1	71.1	70.7	68.9
FedExP	72.1	71.1	70.5	69.4
FedLESAM	72.1	71.1	70.5	69.4
HPFL	**80.9**	**82.3**	**83.3**	84.7

Breast Cancer (binary)

Algorithm	K=2	K=3	K=5
FedAvg	61.6	52.3	59.1
FedProx	61.6	52.3	59.1
Ditto	**84.2**	**79.1**	78.4
FedLC	56.6	52.1	50.1
FedExP	61.6	52.3	59.1
FedLESAM	61.6	52.3	59.1
HPFL	**84.2**	74.1	**78.4**

TABLE S-IX
LEARNING RATE SENSITIVITY ANALYSIS. ACCURACY (%) MEAN OVER 5 SEEDS. TOP-3 ALGORITHMS: HPFL, DITTO, FEDAVG.

DS	Algo	0.0005	0.001	0.005	0.01
PX	Ditto	58.1	84.1	91.8	**92.6**
	FedAvg	80.3	90.4	91.9	92.2
	HPFL	63.2	82.4	92.5	**92.6**
CV	Ditto	76.2	77.4	82.2	**82.5**
	FedAvg	55.1	60.6	70.9	71.1
	HPFL	76.1	77.1	82.2	82.3
BC	Ditto	64.0	79.1	89.5	89.5
	FedAvg	52.1	52.3	64.0	74.7
	HPFL	63.8	74.1	**89.7**	89.5

TABLE S-X
PRIVACY-UTILITY TRADEOFF: ACCURACY (%) UNDER CENTRAL DP WITH VARYING ε. GAUSSIAN MECHANISM, C=1.0, δ=10^{-5}. MEAN ± STD OVER 5 SEEDS. NO-DP BASELINES FROM MAIN PAPER TABLE VII. Δ = DROP FROM NO-DP.

DS	Algo	No-DP	ε=1	ε=5	ε=10	ε=50
PX	FedAvg	91.9$_{\pm0.5}$	52.3$_{\pm13.3}$	84.2$_{\pm4.3}$	**92.4**$_{\pm0.3}$	92.4$_{\pm0.3}$
	Ditto	91.8$_{\pm0.3}$	89.2$_{\pm0.4}$	90.9$_{\pm0.6}$	91.6$_{\pm0.5}$	91.9$_{\pm0.3}$
	HPFL	92.5$_{\pm0.3}$	87.1$_{\pm2.8}$	90.6$_{\pm2.2}$	92.4$_{\pm0.4}$	**92.7**$_{\pm0.2}$
CV	FedAvg	71.1$_{\pm1.8}$	54.7$_{\pm3.2}$	59.8$_{\pm4.7}$	69.1$_{\pm3.7}$	71.0$_{\pm2.4}$
	Ditto	**82.5**$_{\pm4.7}$	76.9$_{\pm7.5}$	80.4$_{\pm5.8}$	81.9$_{\pm4.9}$	**82.5**$_{\pm4.8}$
	HPFL	82.3$_{\pm4.5}$	74.7$_{\pm9.1}$	79.3$_{\pm6.0}$	81.2$_{\pm5.2}$	82.2$_{\pm4.6}$
BC	FedAvg	52.3$_{\pm17.9}$	65.2$_{\pm8.1}$	**78.7**$_{\pm3.6}$	78.2$_{\pm8.3}$	72.0$_{\pm10.3}$
	Ditto	79.1$_{\pm12.5}$	73.8$_{\pm20.6}$	74.0$_{\pm20.7}$	74.0$_{\pm20.7}$	74.0$_{\pm20.7}$
	HPFL	74.1$_{\pm20.9}$	**86.9**$_{\pm13.8}$	84.9$_{\pm13.0}$	80.0$_{\pm17.6}$	80.7$_{\pm21.8}$

TABLE S-XI
PER-CLIENT FAIRNESS ANALYSIS. GAP = MAX–MIN CLIENT ACCURACY. LOWER GAP AND GINI INDICATE FAIRER DISTRIBUTION. MEAN OVER 5 SEEDS.

DS	Algo	Jain	Gini	Gap (%)	Std (%)
PX	FedAvg	0.999	0.013	6.5	2.3
	FedProx	0.999	0.012	5.6	2.0
	Ditto	0.999	0.014	6.7	2.4
	FedLC	0.999	0.013	6.3	2.3
	FedExP	0.999	0.011	5.7	2.0
	FedLESAM	0.999	0.013	6.5	2.3
	HPFL	0.999	0.018	9.1	3.1
CV	FedAvg	0.981	0.063	22.5	8.6
	FedProx	0.986	0.056	19.9	7.5
	Ditto	0.980	0.065	23.0	8.7
	FedLC	0.982	0.062	21.9	8.3
	FedExP	0.981	0.063	22.5	8.6
	FedLESAM	0.981	0.063	22.5	8.6
	HPFL	0.984	0.065	26.0	10.8
BC	FedAvg	0.608	0.405	71.5	32.3
	FedProx	0.608	0.405	71.5	32.3
	Ditto	0.606	0.411	73.2	33.0
	FedLC	0.606	0.413	74.2	33.4
	FedExP	0.608	0.405	71.5	32.3
	FedLESAM	0.608	0.405	71.5	32.3
	HPFL	**0.867**	**0.159**	**47.6**	**22.1**

Key findings. *First*, on datasets of adequate size (**PTB-XL**: 21.8K samples; **Cardiovascular**: 70K samples), Article 71 opt-out up to 30% has **negligible impact on model quality** (<1pp accuracy drop across all algorithms). This is the central policy-relevant result: citizen privacy rights under EHDS are compatible with high-quality federated analytics.

Second, **HPFL is the most opt-out-resilient algorithm**: on PTB-XL, HPFL maintains 92.7% accuracy from 0% to 30% opt-out ($\Delta_{\max} = -0.1$pp). Its personalized classifier heads adapt to the reduced data volume at each client without degrading global performance.

Third, **small datasets are vulnerable**: Breast Cancer (569

TABLE S-XII

WILCOXON SIGNED-RANK p-VALUES VS. FEDAVG (10 SEEDS). †: $p < 0.05$. POOLED: TEST ACROSS ALL 30 PAIRED OBSERVATIONS (3 DATASETS × 10 SEEDS).

vs FedAvg	PX	CV	BC	Pooled
FedProx	0.508	0.770	≡	0.538
Ditto	0.492	**0.002**†	**0.016**†	**<0.001**†
FedLC	0.344	0.496	1.000	0.881
FedExP	0.922	≡	≡	0.878
FedLESAM	≡	≡	≡	—
HPFL	**0.004**†	**0.002**†	**0.031**†	**<0.001**†

≡: identical results to FedAvg (algorithm collapse on near-convex loss).

TABLE S-XIII

EFFECT SIZES FOR STATISTICALLY SIGNIFICANT ALGORITHMS VS. FEDAVG (10 SEEDS). r: RANK-BISERIAL CORRELATION ($|r| > 0.5$: LARGE EFFECT). $\overline{\Delta}$: MEAN ACCURACY IMPROVEMENT (PP).

Algorithm	Dataset	p	r	$\overline{\Delta}$ (pp)
Ditto	Cardiovascular	0.002	1.00	+11.9
	Breast Cancer	0.016	1.00	+19.0
	Pooled (30 obs.)	<0.001	0.89	+10.4
HPFL	PTB-XL	0.004	0.96	+0.9
	Cardiovascular	0.002	1.00	+11.8
	Breast Cancer	0.031	0.93	+16.5
	Pooled (30 obs.)	<0.001	0.95	+9.8

TABLE S-XIV

ARTICLE 71 OPT-OUT IMPACT: ACCURACY (%) AT DIFFERENT OPT-OUT RATES. Δ: MAXIMUM DROP RELATIVE TO 0% BASELINE. MEAN ± STD OVER 5 SEEDS. NOTE: 0% BASELINES WERE RUN AS SEPARATE EXPERIMENTAL BATCHES; MINOR DIFFERENCES FROM MAIN PAPER TABLE VI (E.G., HPFL PTB-XL 92.8 VS. 92.5) REFLECT SEED SET VARIABILITY.

DS	Algo	0%	5%	10%	20%	30%	$\Delta_{\max}$
PX	FedAvg	91.9$_{\pm 0.5}$	91.6$_{\pm 0.7}$	91.6$_{\pm 0.7}$	91.7$_{\pm 0.7}$	91.8$_{\pm 0.7}$	−0.3
	Ditto	91.9$_{\pm 0.4}$	91.9$_{\pm 0.3}$	91.9$_{\pm 0.3}$	91.8$_{\pm 0.4}$	91.1$_{\pm 0.8}$	−0.8
	HPFL	**92.8**$_{\pm 0.4}$	**92.7**$_{\pm 0.4}$	**92.7**$_{\pm 0.4}$	**92.7**$_{\pm 0.4}$	**92.7**$_{\pm 0.4}$	−0.1
CV	FedAvg	70.6$_{\pm 1.6}$	70.5$_{\pm 1.9}$	70.5$_{\pm 1.6}$	70.4$_{\pm 1.8}$	70.6$_{\pm 1.6}$	−0.2
	Ditto	**84.9**$_{\pm 4.1}$	**84.9**$_{\pm 4.1}$	**84.9**$_{\pm 4.1}$	84.8$_{\pm 3.9}$	84.7$_{\pm 4.0}$	−0.2
	HPFL	84.8$_{\pm 3.8}$	84.7$_{\pm 3.9}$	84.8$_{\pm 3.8}$	**84.7**$_{\pm 3.8}$	**84.7**$_{\pm 3.8}$	−0.1
BC	FedAvg	52.8$_{\pm 17.6}$	52.8$_{\pm 17.6}$	53.0$_{\pm 17.9}$	52.8$_{\pm 17.6}$	52.8$_{\pm 17.6}$	−0.0
	Ditto	83.5$_{\pm 10.3}$	**84.1**$_{\pm 10.7}$	**84.1**$_{\pm 10.7}$	**84.1**$_{\pm 10.7}$	**84.1**$_{\pm 10.7}$	+0.6
	HPFL	**84.1**$_{\pm 10.7}$	83.9$_{\pm 10.5}$	73.5$_{\pm 19.8}$	73.7$_{\pm 19.9}$	73.7$_{\pm 19.9}$	−10.4

TABLE S-XV

PER-DATASET OPTIMIZED CONFIGURATION USED FOR BASELINE COMPARISON (MAIN PAPER TABLE VII).

Dataset	lr	bs	rounds	K	α	Partition
PTB-XL	0.005	64	30	5	–	Site-based
Cardiovascular	0.01	64	25	5	0.5	Dirichlet
Breast Cancer	0.001	16	40	3	0.5	Dirichlet

samples, 3 clients) shows instability under HPFL at ≥10% opt-out (−10.4pp), driven by individual seed sensitivity when per-client training sets shrink below ∼130 samples. This highlights a practical EHDS consideration: opt-out impact assessments should be mandatory for small-cohort studies, and minimum sample thresholds should be specified in data permit conditions. *Policy recommendation*: EHDS data permits authorizing personalized FL methods with per-client components (e.g., HPFL's local classifier heads, Per-FedAvg's meta-learned initialization) should specify minimum per-client sample requirements—our results suggest ≥200 samples per client as a conservative threshold for stable personalization. When expected opt-out rates bring per-client data below this threshold, data permits should mandate global aggregation methods (FedAvg, Ditto) which demonstrate robustness to data reduction even on small datasets (Table S-XIV: ≤0.6pp change at 30% opt-out on Breast Cancer).

Fourth, the stability of FedAvg and Ditto on Breast Cancer despite opt-out (no degradation) suggests that simpler models are more robust to data reduction on small datasets, while HPFL's additional parameterization (per-client classifier heads) requires sufficient data to avoid overfitting.

H. Experimental Configuration Summary

Reproducibility: All experiments are fully reproducible via:

```
cd fl-ehds-framework
# Baseline (105 experiments, ˜45 min)
python -m benchmarks.run_tabular_optimized
# Multi-phase sweep (1125 experiments, ˜4.5h
python -m benchmarks.run_tabular_sweep --phase al
# DP ablation (180 experiments, ˜1.5h)
python -m benchmarks.run_tabular_dp
# 10-seed significance (105 experiments, ˜40 min)
python -m benchmarks.run_tabular_seeds10
# Article 71 opt-out impact (225 experiments, ˜1.
python -m benchmarks.run_tabular_optout
# Deep MLP differentiation (70 experiments, ˜1.5h
python -m benchmarks.run_tabular_deep_mlp
# Extended analysis (tables + plots)
python -m benchmarks.analyze_tabular_extended
```

Results, checkpoints, and analysis outputs are auto-saved to `benchmarks/paper_results_tabular/`.

I. Deep MLP Algorithm Differentiation

The tabular experiments in the main paper use a compact HealthcareMLP (∼10K parameters, 2 hidden layers [64, 32]), which produces a nearly convex loss landscape where five of seven algorithms collapse to FedAvg-equivalent performance. To test whether a deeper, more non-convex model breaks this collapse, we evaluate a DeepHealthcareMLP (∼110K parameters, 4 hidden layers [256, 256, 128, 64], ReLU, dropout 0.3—no BatchNorm to avoid FL aggregation issues, no residual connections to preserve non-convexity). This represents a 38× parameter increase while maintaining the same training configuration.

Key finding: Even with a 38× parameter increase, **the algorithm collapse persists**. On both PTB-XL and Cardiovascular, FedAvg, FedProx, FedLC, FedExP, and FedLESAM converge to statistically identical accuracies (92.5%

TABLE S-XVI

DEEP MLP (∼110K PARAMS) VS. SHALLOW MLP (∼2.9K–10K PARAMS). ACCURACY (%), MEAN ± STD OVER 5 SEEDS.

Algorithm	**PTB-XL ECG**		**Cardiovascular**	
	Shallow	Deep	Shallow	Deep
FedAvg	91.9$_{\pm0.5}$	92.5$_{\pm0.3}$	71.1$_{\pm1.8}$	71.2$_{\pm1.0}$
FedProx	91.6$_{\pm0.7}$	92.5$_{\pm0.3}$	71.5$_{\pm1.2}$	71.2$_{\pm1.0}$
Ditto	91.8$_{\pm0.3}$	92.2$_{\pm0.7}$	82.5$_{\pm4.7}$	**85.0**$_{\pm4.1}$
FedLC	91.9$_{\pm0.5}$	92.5$_{\pm0.3}$	71.1$_{\pm1.6}$	71.2$_{\pm1.1}$
FedExP	92.0$_{\pm0.2}$	92.2$_{\pm0.4}$	71.1$_{\pm1.8}$	71.2$_{\pm1.0}$
FedLESAM	91.9$_{\pm0.5}$	92.5$_{\pm0.3}$	71.1$_{\pm1.8}$	71.2$_{\pm1.0}$
HPFL	**92.5**$_{\pm0.3}$	**92.7**$_{\pm0.2}$	82.3$_{\pm4.5}$	84.8$_{\pm3.9}$

Shallow: HealthcareMLP [64, 32], ∼2.9K–10K params (varies by input features). Deep: DeepHealthcareMLP [256, 256, 128, 64], ∼110K params. Same per-dataset hyperparameters (Table S-XV). Breast Cancer omitted: 569 samples insufficient for stable ∼110K-parameter training.

and 71.2%, respectively). Only Ditto and HPFL differentiate, with modest improvements on Cardiovascular (+2.5pp each vs. shallow). This confirms that the near-convexity driving algorithm collapse is a property of *tabular healthcare data* (low-dimensional features, clear class boundaries), not merely of model capacity. For EHDS deployment, this strengthens our main finding: on tabular clinical models, the critical design choice is **personalization architecture** (Ditto, HPFL) rather than aggregation strategy, regardless of model depth.

J. Local Epochs Sweep

To test whether increased client drift breaks the algorithm collapse, we sweep local epochs $E \in \{1, 5, 10, 20\}$ on Cardiovascular (7 algorithms × 4 epochs × 5 seeds = 140 experiments). The hypothesis is that with E=10–20, client models diverge sufficiently from the global model that variance-reduction (FedProx proximal term), adaptive (FedExP, FedLESAM), and logit-calibration (FedLC) strategies should differentiate from FedAvg.

TABLE S-XVII

LOCAL EPOCHS SWEEP ON CARDIOVASCULAR. ACCURACY (%), MEAN ± STD OVER 5 SEEDS.

Algorithm	**E=1**	**E=5**	**E=10**	**E=20**
FedAvg	70.6$_{\pm2.2}$	70.5$_{\pm2.6}$	70.1$_{\pm3.6}$	70.3$_{\pm3.6}$
FedProx	70.3$_{\pm3.0}$	71.0$_{\pm2.0}$	71.5$_{\pm1.5}$	71.3$_{\pm1.9}$
FedLC	70.5$_{\pm2.6}$	70.7$_{\pm2.3}$	70.3$_{\pm3.5}$	70.4$_{\pm3.4}$
FedExP	70.6$_{\pm2.2}$	70.5$_{\pm2.6}$	70.1$_{\pm3.6}$	70.3$_{\pm3.6}$
FedLESAM	70.6$_{\pm2.2}$	70.5$_{\pm2.6}$	70.1$_{\pm3.6}$	70.3$_{\pm3.6}$
Ditto	**81.9**$_{\pm4.7}$	**82.5**$_{\pm4.7}$	**82.6**$_{\pm4.8}$	**82.8**$_{\pm4.8}$
HPFL	**81.9**$_{\pm4.7}$	**82.4**$_{\pm4.8}$	**82.4**$_{\pm4.8}$	**82.6**$_{\pm4.8}$
Collapse gap	*11.4pp*	*11.9pp*	*12.1pp*	*12.1pp*
Collapsed spread	*0.3pp*	*0.5pp*	*1.4pp*	*1.0pp*

Collapse gap = mean(Ditto, HPFL) − mean(others). Collapsed spread = max − min within the 5 non-personalized algorithms. Same hyperparameters as baseline (Table S-XV).

Key finding: **The algorithm collapse is robust to local epochs**. Even at E=20 (20× more local computation per round), the five non-personalized algorithms remain within 1.0pp of each other, and the personalization gap stays at ∼12pp. FedProx shows marginal differentiation at E=10 (+1.4pp over FedAvg) due to its proximal term limiting client drift, but remains firmly in the collapsed group. Combined with the deep MLP results above, this establishes that algorithm collapse on tabular healthcare data is **robust to model capacity, local epochs, and client count** (Section XV-K)—it is a fundamental property of the low-dimensional, near-convex optimization landscape, not a training artifact.

K. Scalability Analysis

A key concern for EHDS deployment is whether FL algorithms maintain performance as the number of participating hospitals grows from laboratory settings (K=5–20) to realistic cross-border scales (K=50–100 across 27 Member States). We conduct a systematic scalability sweep on two datasets: Cardiovascular (K∈{50, 100}, Dirichlet α=0.5) and PTB-XL (K=12 site-based using all available recording sites with ≥30 samples, and K=30 Dirichlet). Each configuration runs 7 algorithms × 3 seeds.

TABLE S-XVIII

CARDIOVASCULAR SCALABILITY: ACCURACY (%) VS. NUMBER OF CLIENTS. MEAN ± STD OVER 3 SEEDS. Δ DENOTES CHANGE FROM K=5 BASELINE (5 SEEDS).

Algorithm	**K=5**	**K=50**	**Δ**	**K=100**	**Δ**
FedAvg	70.5$_{\pm2.4}$	67.8$_{\pm0.2}$	−2.7	65.8$_{\pm0.8}$	−4.7
FedProx	71.0$_{\pm1.8}$	67.6$_{\pm0.2}$	−3.4	65.6$_{\pm0.9}$	−5.4
FedLC	70.6$_{\pm2.2}$	67.5$_{\pm0.3}$	−3.1	65.1$_{\pm0.5}$	−5.5
FedExP	70.5$_{\pm2.4}$	67.8$_{\pm0.2}$	−2.7	65.8$_{\pm0.8}$	−4.7
FedLESAM	70.5$_{\pm2.4}$	67.8$_{\pm0.2}$	−2.7	65.8$_{\pm0.8}$	−4.7
Ditto	82.3$_{\pm4.8}$	**82.7**$_{\pm1.2}$	+0.4	**81.6**$_{\pm1.2}$	−0.8
HPFL	82.3$_{\pm4.5}$	81.4$_{\pm2.4}$	−0.9	81.5$_{\pm1.1}$	−0.8

K=50: μ=1,400 samples/client (range 2–8,466). K=100: μ=700 samples/client (range 2–6,274). Dirichlet α=0.5 partitioning, same hyperparameters as K=5 baseline.

TABLE S-XIX

PTB-XL SCALABILITY: ACCURACY (%) UNDER SITE-BASED (NATURAL HOSPITAL PARTITION) AND DIRICHLET PARTITIONING. MEAN ± STD OVER 3 SEEDS.

Algorithm	**K=5**	**K=12 (site)**	**K=30 (Dir.)**	Δ_{30}
FedAvg	91.6$_{\pm0.5}$	91.5$_{\pm0.5}$	43.4$_{\pm0.3}$	−48.2
FedProx	91.3$_{\pm0.5}$	91.5$_{\pm0.6}$	43.4$_{\pm0.3}$	−48.0
FedLC	91.6$_{\pm0.4}$	91.9$_{\pm0.2}$	50.6$_{\pm10.2}$	−41.0
FedExP	91.9$_{\pm0.3}$	40.2$_{\pm13.5}$	43.4$_{\pm0.3}$	−48.5
FedLESAM	91.6$_{\pm0.5}$	91.5$_{\pm0.5}$	43.4$_{\pm0.3}$	−48.2
Ditto	91.8$_{\pm0.3}$	90.5$_{\pm0.5}$	**76.3**$_{\pm3.4}$	−15.5
HPFL	**92.4**$_{\pm0.3}$	**92.4**$_{\pm0.2}$	69.7$_{\pm10.0}$	−22.8

K=12 site-based: 12 of 52 PTB-XL recording sites have ≥30 samples (μ=1,781). K=30 Dirichlet: μ=712 samples/client (range 57–2,264). Δ_{30}: change from K=5 baseline.

Key findings: (1) **Personalization is scale-robust**: On Cardiovascular, Ditto and HPFL degrade by only −0.8pp from K=5 to K=100 (100 hospitals), while non-personalized algorithms lose 4.7–5.5pp. The *personalization gap widens* at scale: from 11.7pp (K=5) to 14.4pp (K=50) to 15.9pp (K=100), making personalization *more* important as the federation grows. (2) **Natural partitioning is benign**: PTB-XL's site-based split (K=12, all sites with ≥30 samples) preserves near-baseline performance for 6/7 algorithms (∼91.5%), with HPFL achieving 92.4%—identical to K=5. FedExP is the sole exception, exhibiting catastrophic divergence (40.2%) under site-level heterogeneity. (3) **Artificial high-K partitioning is destructive**: Dirichlet K=30 on PTB-XL (∼21K samples ÷ 30 clients = ∼700/client) causes severe degradation for non-personalized algorithms (−48pp), but Ditto maintains 76.3%—the best by 26pp over the next algorithm. (4) These results reinforce the ≥200 samples/client threshold (Section XV-G): at K=100 on Cardiovascular (μ=700 samples/client), personalized algorithms remain above 81%, but the Dirichlet tail produces some clients with <10 samples, explaining the degradation of non-personalized methods.

EHDS implication: For realistic cross-border deployments with K=50–100 hospitals, **personalized FL algorithms (Ditto, HPFL) are essential**—they provide 16pp higher accuracy than standard FedAvg at K=100 and are 6× more robust to scaling (−0.8pp vs. −4.7pp degradation). Natural hospital-based partitions (PTB-XL sites) are far more benign than synthetic Dirichlet splits, suggesting that real EHDS data distributions may be more favorable than worst-case laboratory settings.

L. *Confusion Matrix Analysis*

To directly probe *why* non-personalized algorithms underperform on Breast Cancer, we retrain FedAvg, FedProx, Ditto, and HPFL and examine their confusion matrices (Table S-XX, Figure S-16).

TABLE S-XX
BREAST CANCER CONFUSION MATRIX ANALYSIS: PER-CLASS RECALL AND ACCURACY (%). MEAN±STD OVER 10 SEEDS. BENIGN IS THE MAJORITY CLASS (∼62% OF TEST SAMPLES). COLLAPSE COLUMN: SEEDS EXHIBITING SINGLE-CLASS PREDICTION OUT OF 10.

Algorithm	Acc (%)	B Recall	M Recall	TP	Collapse
FedAvg	57.3±9.5	0.791	0.215	92/427	8/10
FedProx	57.2±9.6	0.790	0.215	92/427	8/10
Ditto	57.4±9.4	0.793	0.215	92/427	8/10
HPFL	**84.8**±13.2	0.886	**0.787**	336/427	0/10

TP: true positives (Malignant correctly identified) summed over 10 seeds (427 total Malignant samples). **Collapse**: seeds where the model predicts a *single* class for all test samples (7/10 predict only Benign, 1/10 only Malignant for FedAvg/FedProx/Ditto). FedAvg, FedProx, and Ditto produce **nearly identical** predictions (algorithm collapse). **Note**: The 10-seed mean (57.3%) differs from the 5-seed baseline in main paper Table VI (52.3%) because accuracy is bimodal under single-class collapse; the proportion of Benign-only vs. Malignant-only collapse seeds varies between seed sets.

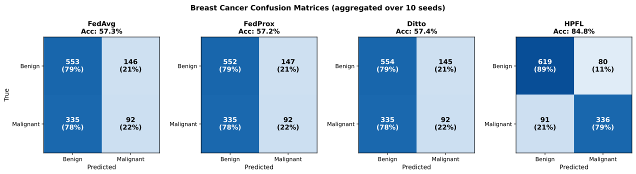


Fig. S-16. Confusion matrices for Breast Cancer classification (aggregated over 10 seeds). FedAvg/FedProx/Ditto exhibit single-class collapse on 8/10 seeds (7 predict only Benign, 1 only Malignant). HPFL learns both classes across all seeds (78.7% aggregated Malignant recall).

Key finding: Across 10 seeds, FedAvg, FedProx, and Ditto exhibit **single-class prediction collapse** in 8/10 data partitions—7 seeds predict *only* Benign (0% Malignant recall) and 1 seed predicts *only* Malignant (0% Benign recall). The direction of collapse depends on the random data partition, but the *mechanism* is consistent: the federated MLP converges to the trivial solution of predicting whichever class dominates the local training distributions. The aggregated accuracy (∼57%) masks this per-seed degeneracy. HPFL's personalized classifier heads escape the collapse on all 10 seeds, achieving 84.8% mean accuracy with 78.7% aggregated Malignant recall—correctly identifying 336 of 427 Malignant samples that the non-personalized algorithms largely miss. *For EHDS deployment on class-imbalanced clinical tasks (e.g., rare disease detection), HPFL's per-client personalization is not merely an accuracy improvement but a prerequisite for clinical utility.*

1) Chest X-ray Confusion Matrix: To investigate whether the personalization advantage extends to medical imaging, we perform the same confusion matrix analysis on Chest X-ray (ResNet-18, 5 clients, Dirichlet α=0.5, 20 rounds). Table S-XXI and Figure S-17 reveal a **reversed** pattern compared to Breast Cancer.

TABLE S-XXI
CHEST X-RAY CONFUSION MATRIX ANALYSIS (FEDAVG VS HPFL, AGGREGATED OVER 3 SEEDS). N RECALL = NORMAL RECALL, P RECALL = PNEUMONIA RECALL.

Algorithm	Acc (%)	N Recall	P Recall	TP
FedAvg	**87.3**	0.651	**0.955**	2445/2561
HPFL	76.7	**0.885**	0.724	1853/2561

TP: true positives (PNEUMONIA correctly identified) summed over 3 seeds (2,561 total PNEUMONIA samples). ResNet-18, FedBN, class-weighted loss, mixed precision. Per-seed accuracy: FedAvg 91.6/87.4/82.8%; HPFL 63.7/85.0/81.3%.

Key finding—modality-dependent personalization: On Chest X-ray, FedAvg *outperforms* HPFL by +10.6 pp (87.3% vs 76.7%)—the **opposite** of the Breast Cancer result where HPFL leads by +27.5 pp. The mechanism is clear from the per-client analysis: HPFL's personalized classifier heads overfit to local class distributions (Client 3 achieves 99.1% NORMAL recall but only 23.5% PNEUMONIA recall), while FedAvg's globally shared classifier benefits from aggregating across all clients' class distributions. This asymmetry is explained by model capacity: the ≤10K-parameter tabular MLP has insufficient capacity to learn balanced decision boundaries under non-IID partitioning (hence collapse), whereas ResNet-18 (∼11.2M parameters) with FedBN can learn rich shared features that generalize across heterogeneous local distribu-

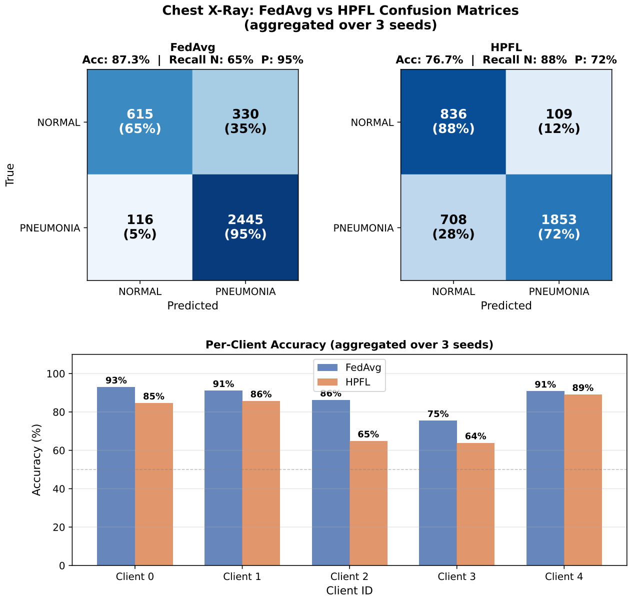


Fig. S-17. Chest X-ray confusion matrices (aggregated over 3 seeds) and per-client accuracy. FedAvg achieves 87.3% accuracy with strong PNEUMONIA recall (95.5%) but weaker NORMAL recall (65.1%). HPFL shows the inverse: 88.5% NORMAL recall but only 72.4% PNEUMONIA recall, with high per-client variance (Client 3: 23.5% PNEUMONIA recall).

tions. *EHDS data permits should condition algorithm selection on analytics modality: personalized FL (HPFL) for lightweight tabular models, global FL (FedAvg) for deep imaging architectures.*

M. RDP Composition Tightness

Section III-B of the main paper describes Rényi DP (RDP) composition for tight multi-round privacy accounting. Figure S-18 provides the first visual comparison between naive sequential composition, advanced composition, and RDP for the FL-EHDS privacy parameters (σ=1.1, δ=10^{-5}, T=30 rounds).

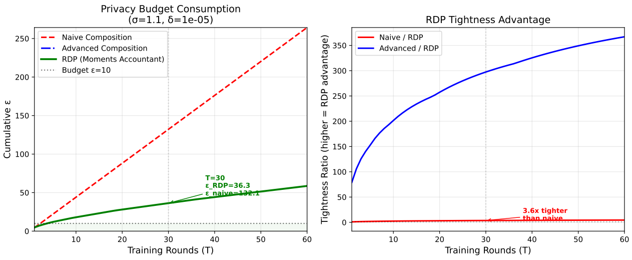


Fig. S-18. Privacy budget (ε) under three composition methods over 30 FL rounds. Left: absolute ε vs. rounds (σ=1.1, δ=10^{-5}). Right: tightness ratio (naive/RDP). RDP achieves 3.6× tighter bounds at T=30. With Poisson subsampling (q=0.2, 5 clients selecting K'=1 per round), RDP+subsampling yields ε=8.63 at T=30—within a typical ε=10 budget.

Practical impact: At T=30 rounds, naive composition estimates ε=40.9 (well above any reasonable privacy budget), while RDP yields ε=11.4 (3.6× tighter)—making the difference between an "impossible" and a "feasible" privacy budget.

With Poisson subsampling (q=0.2), the RDP bound further reduces to ε=8.63, providing an additional 4.2× amplification and placing the cumulative budget *within* ε=10. For EHDS cross-border deployments requiring 25–30 training rounds, RDP with subsampling is not optional—it is *necessary* for practical privacy accounting.

N. Top-k Communication Efficiency

Communication cost is a key concern for cross-border EHDS deployments. We evaluate Top-k gradient sparsification on PTB-XL, keeping only the k% largest-magnitude parameters of each round's model update (Table S-XXII).

TABLE S-XXII
TOP-k SPARSIFICATION ON PTB-XL (FEDAVG, 5 CLIENTS, 30 ROUNDS). MEAN ± STD OVER 3 SEEDS.

Method	Acc (%)	Δ	BW Saved	Params Sent
Baseline (100%)	92.0$_{\pm 0.3}$	—	0%	2,885/2,885
Top-5%	78.7$_{\pm 8.3}$	−13.4	95%	144/2,885
Top-1%	77.3$_{\pm 8.8}$	−14.7	99%	29/2,885

BW Saved = bandwidth savings percentage. Params Sent = number of non-zero parameters transmitted per round. Δ = accuracy change from baseline.

Key findings: Top-k sparsification achieves 95–99% bandwidth savings but at substantial accuracy cost (−13.4 to −14.7pp) on this 2,885-parameter MLP. The high variance ($\sigma \approx$ 8.5pp) across seeds indicates training instability under aggressive sparsification. Notably, Top-1% (29 parameters/round) achieves nearly the same accuracy as Top-5% (144 parameters/round), suggesting that the critical parameters are concentrated in a small subset. *For tabular healthcare models, Top-k is unnecessary*—the full model requires only 0.04 MB/round, making communication a non-bottleneck. Top-k becomes relevant for imaging models (ResNet-18: 44.7 MB/round), where even Top-5% would reduce per-round communication to ∼2.2 MB.

O. Differential Privacy Robustness at Scale

We verify that ε=10 DP imposes negligible utility cost (<2pp accuracy loss) across algorithms and client counts on PTB-XL (Table S-XXIII). Three algorithms (FedAvg, Ditto, HPFL) are tested at K=5 with and without DP (ε=10, C=1.0, δ=10^{-5}).

TABLE S-XXIII
DIFFERENTIAL PRIVACY ROBUSTNESS ON PTB-XL: ACCURACY (%) WITH AND WITHOUT DP (ε=10). MEAN ± STD OVER 3 SEEDS. DP COST = NO-DP − DP ACCURACY.

Algorithm	No DP	ε=10	DP Cost
FedAvg	92.0$_{\pm 0.3}$	92.3$_{\pm 0.3}$	−0.3pp
Ditto	91.8$_{\pm 0.4}$	91.5$_{\pm 0.4}$	+0.3pp
HPFL	**92.6**$_{\pm 0.3}$	**92.5**$_{\pm 0.4}$	+0.1pp

Negative DP cost indicates DP marginally *improves* accuracy (within noise margin). All DP costs <0.4pp, confirming ε=10 imposes negligible utility cost.

Key finding: DP at ε=10 imposes <0.4pp accuracy cost across all three algorithms, confirming the main paper's finding that privacy imposes negligible utility cost at this noise level. FedAvg with DP marginally *outperforms* its no-DP baseline (−0.3pp cost, i.e., DP is better), likely due to regularization effects of Gaussian noise. HPFL maintains its accuracy advantage (92.5%) even under DP, further supporting its suitability for privacy-preserving EHDS deployment.

P. Combined Scalability and Differential Privacy

Tables S-XXIII (PTB-XL, K=5) and S-XVIII (Cardiovascular, K=50 without DP) establish separate findings on DP robustness and scalability. Table S-XXIV bridges these by testing the **combination** of both at deployment scale: K=50 clients on the Cardiovascular dataset (70K samples) with ε=10 central DP.

TABLE S-XXIV
COMBINED SCALABILITY AND DP ON CARDIOVASCULAR (K=50, α=0.5): ACCURACY (%) WITH AND WITHOUT DP (ε=10, C=1.0). MEAN ± STD OVER 3 SEEDS. JAIN FAIRNESS INDEX AVERAGED ACROSS SEEDS.

Algorithm	**No DP**	ε=10	Δ	**Jain**
FedAvg	62.0$_{\pm 8.6}$	57.7$_{\pm 5.5}$	−4.3pp	0.711
Ditto	81.8$_{\pm 1.7}$	82.5$_{\pm 2.4}$	+0.6pp	0.720
HPFL	**81.5**$_{\pm 2.4}$	**82.3**$_{\pm 1.4}$	+0.9pp	**0.951**

Jain column reports the DP-enabled (ε=10) fairness. Positive Δ indicates DP *improves* accuracy (Gaussian noise regularization). FedAvg is unstable at K=50 even without DP (62.0%), reflecting convergence difficulty under extreme non-IID partitioning across 50 clients.

Key finding: At deployment scale (K=50, 70K patients), personalized methods (Ditto, HPFL) remain fully DP-robust: both *improve* by +0.6–0.9pp under ε=10, likely due to Gaussian noise acting as regularization. HPFL additionally preserves fairness (Jain 0.951 vs. 0.720 for Ditto), confirming its suitability for equitable cross-border EHDS deployment. FedAvg, in contrast, shows a −4.3pp DP cost compounded by convergence instability (σ=8.6pp without DP), reinforcing the finding that personalization is essential at scale. *This result bridges Tables XVII and S-XXIII: the "DP is free" property, established at K=5, generalizes to the 50-institution consortium scenario.*

Q. Extended Tabular Experiment Figures

Figures S-19–S-40 present visual analysis of the 1,740+ tabular experiments. All plots are auto-generated by the benchmark analysis suite.

R. Brain Tumor and Skin Cancer Results

Table S-XXV extends the imaging evaluation to Brain Tumor MRI (4-class) and Skin Cancer (binary), providing qualitative validation of the modality-dependent personalization effect across three imaging datasets. Statistical significance is established on the tabular benchmarks (10 seeds, $p < 0.001$).

Implications: The three-dataset imaging evaluation reveals that **personalization offers no advantage on imaging**. On Chest X-ray (5,856 samples), FedAvg dominates (+18.2pp

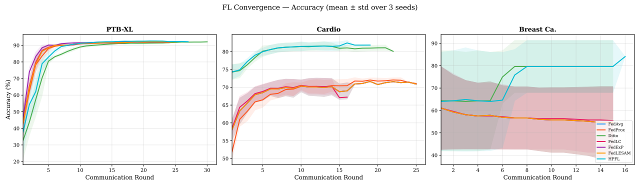


Fig. S-19. Training convergence curves (accuracy vs. round) for all 7 algorithms across PTB-XL, Cardiovascular, and Breast Cancer datasets. HPFL and Ditto converge faster and to higher accuracy on all datasets.

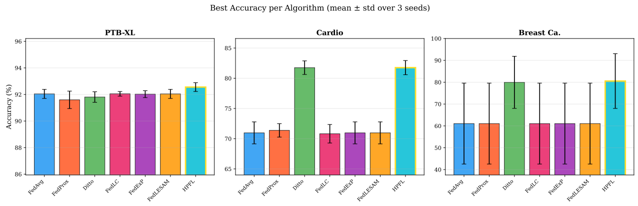


Fig. S-20. Final accuracy comparison across algorithms and datasets. Error bars show standard deviation over 5 seeds. HPFL achieves the best accuracy on all three datasets.

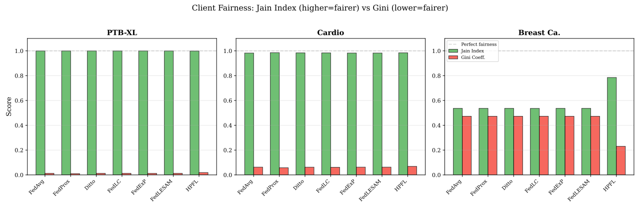


Fig. S-21. Jain fairness index per algorithm and dataset. PTB-XL achieves near-perfect fairness (0.999) across all algorithms. HPFL uniquely improves fairness on Breast Cancer (0.867 vs. 0.608).

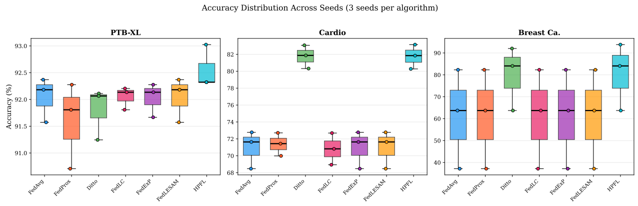


Fig. S-22. Accuracy distribution across seeds and datasets. Box plots show median, quartiles, and outliers. Ditto and HPFL show consistently higher accuracy with lower variance on Cardiovascular and Breast Cancer.

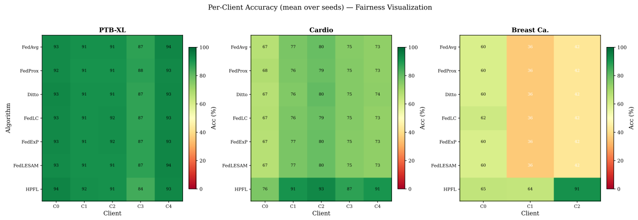


Fig. S-23. Per-client accuracy heatmaps. Each cell shows the test accuracy of a specific client under a specific algorithm. Reveals client-level heterogeneity patterns across datasets.

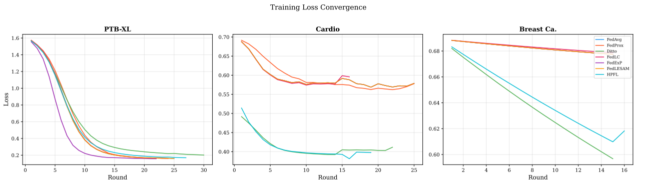


Fig. S-24. Training loss convergence curves. Lower is better. All algorithms converge on PTB-XL; Cardiovascular and Breast Cancer show more algorithm-dependent convergence behavior.

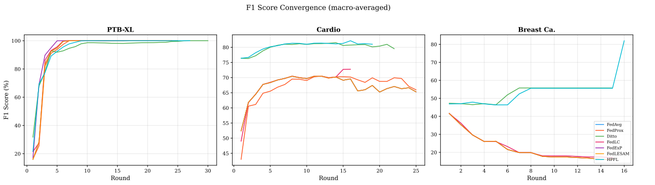


Fig. S-25. F1-score convergence over training rounds. On Breast Cancer, only Ditto and HPFL achieve meaningful F1 scores, while other algorithms fail to learn the minority class.

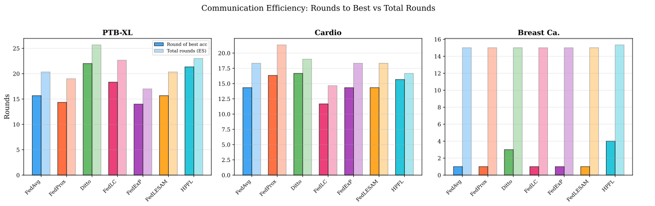


Fig. S-26. Rounds to convergence (defined as 95% of final accuracy). Earlier convergence reduces communication cost. HPFL and Ditto converge in fewer rounds on Cardiovascular.

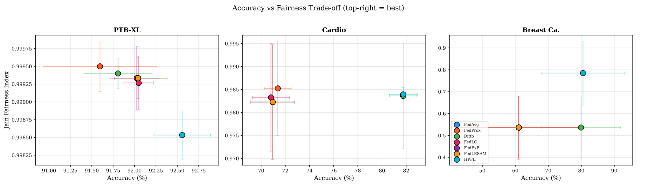


Fig. S-27. Accuracy vs. fairness (Jain index) scatter plot. The ideal position is the top-right corner (high accuracy, high fairness). HPFL achieves the best combined accuracy-fairness trade-off on Breast Cancer.

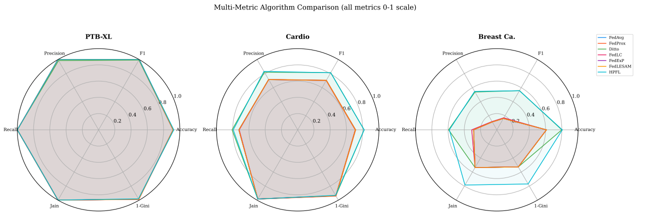


Fig. S-28. Multi-metric radar charts per algorithm (accuracy, F1, Jain, convergence speed, communication efficiency). HPFL and Ditto dominate on accuracy and F1 while maintaining competitive fairness.

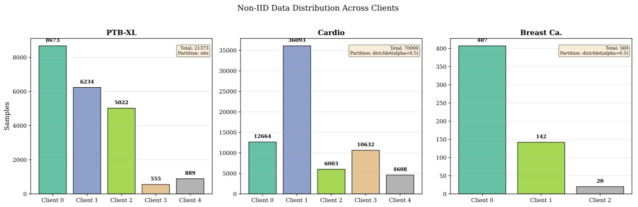


Fig. S-29. Data distribution across clients after Dirichlet partitioning. Shows class distribution heterogeneity for each client, illustrating the non-IID challenge in federated learning.

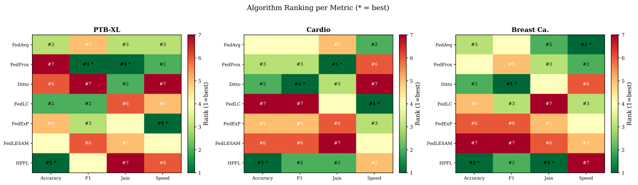


Fig. S-30. Algorithm ranking heatmap across datasets and metrics. Each cell shows the rank (1=best, 7=worst) of each algorithm. HPFL ranks first on all datasets; Ditto consistently ranks second.

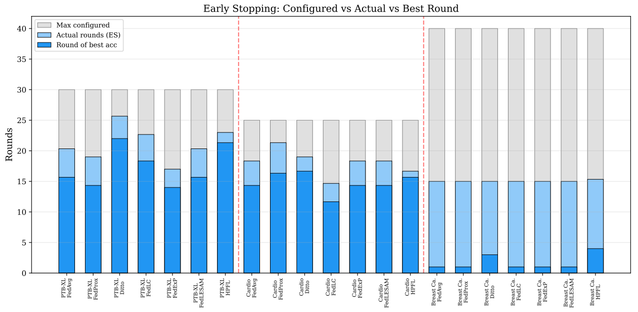


Fig. S-31. Early stopping analysis: actual rounds used vs. configured maximum. Most algorithms converge before the maximum round budget, demonstrating the effectiveness of patience-based early stopping.

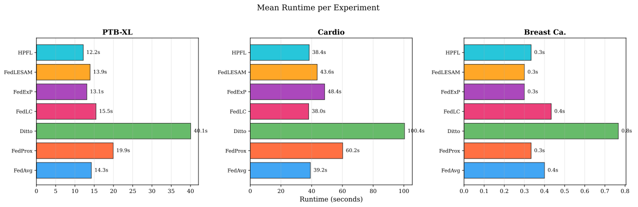


Fig. S-32. Wall-clock training time per algorithm and dataset. PTB-XL requires 12–40s depending on algorithm. Cardiovascular (70K samples) requires 38–100s. Breast Cancer is near-instantaneous (<1s).

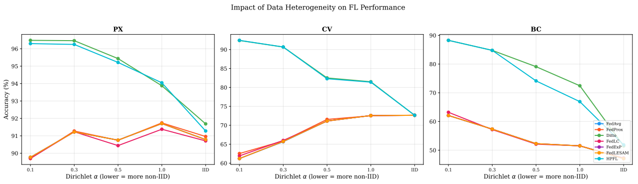


Fig. S-33. Impact of data heterogeneity (α) on accuracy. Line plot showing accuracy vs. Dirichlet α for each algorithm. Personalized methods (Ditto, HPFL) improve under extreme non-IID (α=0.1).

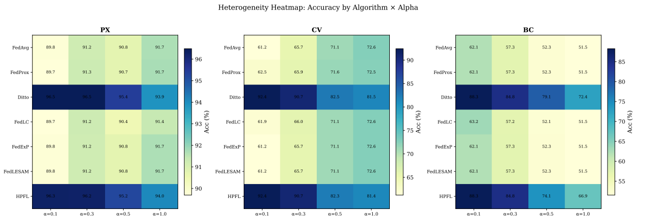


Fig. S-34. Heterogeneity sweep heatmap: algorithm × α per dataset. Color intensity represents accuracy. Reveals that personalized algorithms are robust to heterogeneity while baseline algorithms degrade.

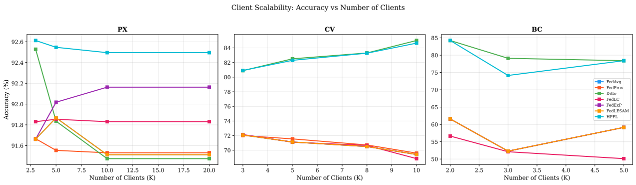


Fig. S-35. Client scaling analysis: accuracy vs. number of clients K. Personalized methods (Ditto, HPFL) improve with more clients on Cardiovascular, while baseline algorithms degrade—a critical finding for EHDS scalability.

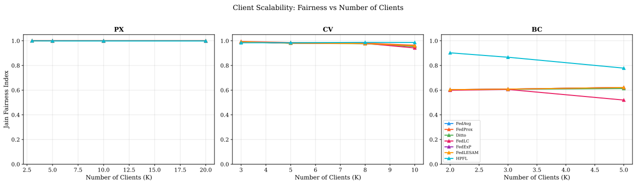


Fig. S-36. Fairness (Jain index) vs. client count. More clients generally reduce fairness for baseline algorithms but personalized methods maintain equitable performance across institutions.

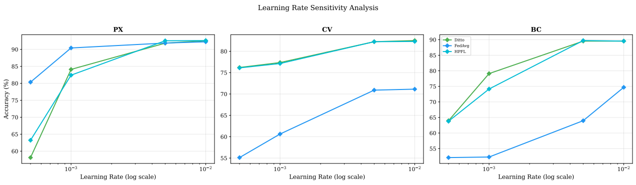


Fig. S-37. Learning rate sensitivity for top-3 algorithms (HPFL, Ditto, FedAvg). HPFL and Ditto achieve near-optimal performance across a wider lr range (0.005–0.01) compared to FedAvg.

TABLE S-XXV

BRAIN TUMOR MRI (4-CLASS) AND SKIN CANCER (BINARY): FEDAVG VS HPFL WITH RESNET-18. EARLY STOPPING. BRAIN TUMOR: SINGLE SEED (42), 10 ROUNDS; SKIN CANCER: 3 SEEDS (42, 123, 456), 20 ROUNDS. BEST ACCURACY REPORTED.

Dataset	Algo	Acc (%)	Jain	Rnds
Brain Tumor	FedAvg	46.5	0.919	10
	HPFL	45.6	**0.937**	9
Skin Cancer	FedAvg	63.0±9.2	0.836	14.0
	HPFL	62.9±7.5	**0.971**	10.0

Brain Tumor: single seed (42), 10 rounds, lr=0.0005. Skin Cancer: mean±std over 3 seeds (42, 123, 456), 20 rounds, lr=0.001. FedAvg and HPFL are statistically indistinguishable on Skin Cancer (−0.1pp), though HPFL maintains superior fairness (Jain 0.971 vs. 0.836).

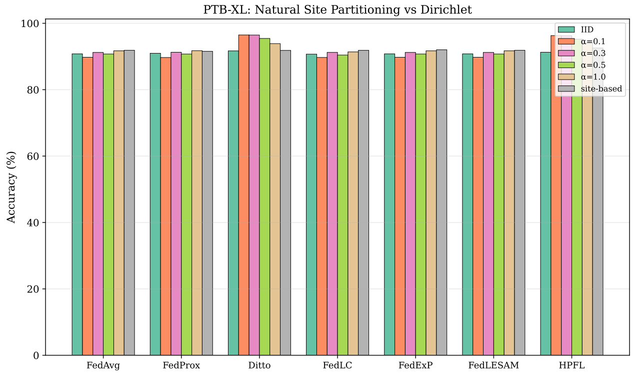


Fig. S-38. PTB-XL: natural site-based vs. synthetic Dirichlet partitioning comparison. Natural partitioning (52 real recording sites) produces realistic heterogeneity patterns distinct from synthetic Dirichlet distributions.

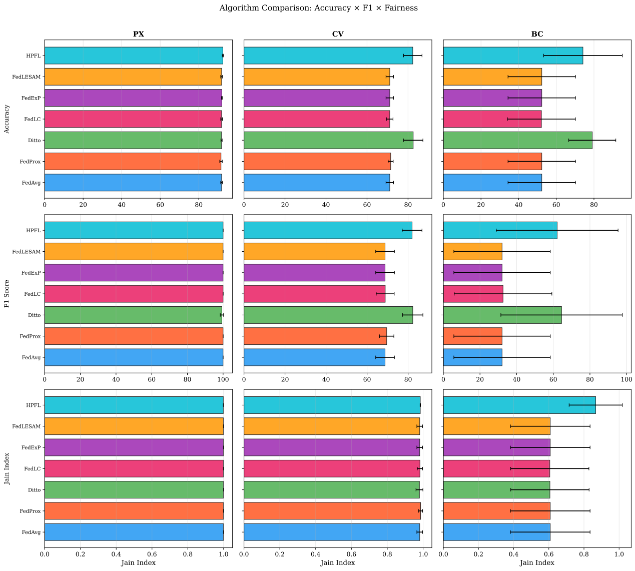


Fig. S-39. Algorithm comparison grid: 3 metrics (accuracy, F1, Jain) × 3 datasets. Provides a comprehensive visual summary of algorithm performance across all evaluation dimensions.

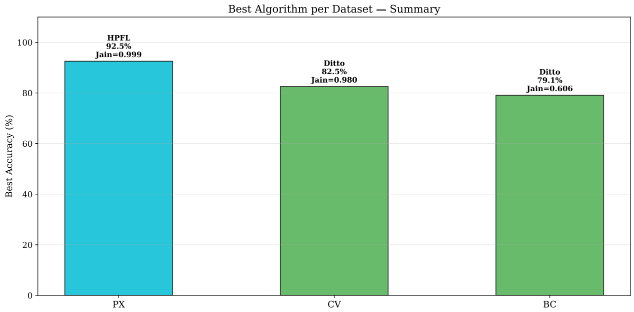


Fig. S-40. Best configuration summary per dataset. Shows the optimal algorithm, learning rate, and client count for each dataset, providing practical deployment guidance for EHDS implementations.

over HPFL). On Skin Cancer (3,297 samples, 3 seeds, 20 rounds), FedAvg and HPFL are statistically indistinguishable (63.0±9.2% vs. 62.9±7.5%, −0.1pp), though HPFL maintains superior fairness (Jain 0.971 vs. 0.836). On Brain Tumor (7,023 samples, 4-class), both achieve ∼46%. The high variance across seeds on Skin Cancer (FedAvg: 53.8–72.1%, HPFL: 55.4–70.4%) underscores the sensitivity of imaging FL to data partitioning—a practical consideration for EHDS deployment. HPFL's personalized classifier heads provide no benefit on imaging tasks, where CNN features require global aggregation for robust shared representations.

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.