

Federated Bioinformatics Platforms

Noah Dubois¹, Clara Hansen², Nina Popescu³

¹ Assistant Professor, School of Data Science, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: noah.dubois746@yahoo.com | ORCID: 8813-5127-2391-7202

² Assistant Professor, Department of Computer Science, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: clara.hansen663@yahoo.com | ORCID: 2893-0621-6279-1886

³ Postdoctoral Researcher, Department of Machine Learning, Baltic AI Research University, Tallinn, Estonia. Email: nina.popescu516@yahoo.com | ORCID: 2372-1839-5748-2427

ABSTRACT

Biomedical research increasingly relies on large, multi-institutional datasets -- genomic cohorts, electronic health records, clinical trial registries, and imaging archives -- yet data sharing is constrained by privacy regulations (GDPR, HIPAA), institutional governance, and patient consent limitations. Federated learning (FL) enables collaborative model training across distributed datasets without sharing raw data, but applying FL to bioinformatics presents unique challenges: heterogeneous data modalities (genomic, clinical, imaging), extreme class imbalance (rare diseases, uncommon variants), non-IID data distributions across institutions, and the need for reproducible, auditable analyses compliant with regulatory frameworks. This study developed and benchmarked six distributed bioinformatics approaches -- centralised training (data pooling, upper bound), local-only training (no collaboration, lower bound), naive federated averaging (FedAvg), federated proximal (FedProx), personalised federated learning (Per-FedAvg), and a proposed Federated Bioinformatics Intelligence Platform (FedBioIP) combining differential privacy-guaranteed gradient aggregation, domain-adaptive batch normalisation for cross-institutional heterogeneity, multi-modal federated encoding for heterogeneous omics data, and secure aggregation with verifiable computation -- across three multi-institutional bioinformatics tasks: rare variant pathogenicity classification (6 biobanks, $n = 284,000$ exomes), cancer histopathology classification (8 hospitals, $n = 42,600$ whole-slide images), and clinical outcome prediction from EHR (12 health systems, $n = 1.8$ million patient records). FedBioIP achieved 97.2% of centralised performance across all tasks (mean AUROC 0.886 vs centralised 0.912), significantly outperforming FedAvg (91.4%; $p < 0.001$), FedProx (93.8%; $p < 0.001$), and Per-FedAvg (95.2%; $p = 0.008$), while providing formal differential privacy guarantee ($\epsilon = 1.0$, $\delta = 10^{-5}$). The domain-adaptive normalisation reduced inter-institutional performance variance by 64.2% compared to FedAvg, addressing the non-IID distribution problem that is the primary challenge in biomedical federated learning. These results establish privacy-preserving federated platforms as a viable alternative to data centralisation for large-scale bioinformatics.

Keywords: Federated learning; Bioinformatics; Privacy-preserving analytics; Differential privacy; Multi-institutional collaboration; Genomic data sharing; Non-IID distributions; Secure computation; Clinical prediction

Citation: Dubois et al. [2025]. Federated Bioinformatics Platforms. DOI: <https://doi.org/10.5281/zenodo.19549568>

Copyright: © 2025 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2025 Jan 06 Accepted: 2025 Mar 08 Published: 2025 Jun 15

Research Article: Research Article

1. Introduction

The power of biomedical data science scales with dataset size: rare variant interpretation requires hundreds of thousands of sequenced genomes, cancer classification benefits from diverse histological examples across populations and institutions, and clinical prediction models generalise best when trained on heterogeneous healthcare systems. Yet the data needed for these advances resides in silos -- hospital EHR systems, national biobanks, research consortia -- separated by privacy regulations, data governance frameworks, and practical barriers to data transfer (Rieke et al., 2020). The EU General Data Protection Regulation (GDPR) restricts cross-border transfer of health data, HIPAA limits sharing of protected health information in the US, and institutional data use agreements typically prohibit raw data export. These constraints create a paradox: the data exist to solve critical biomedical problems, but cannot be assembled in one place for analysis.

1.1 Federated Learning for Biomedicine

Federated learning resolves this paradox by bringing the computation to the data rather than the data to the computation. In FL, a coordinating server distributes a model to participating institutions (clients), each client trains the model on its local data, and only model updates (gradients or parameters) -- not raw data -- are transmitted to the server for aggregation (McMahan et al., 2017). The aggregated global model benefits from the combined statistical power of all institutions while raw patient data never leave their source. FL has been successfully applied to brain tumour segmentation (FeTS, 71 institutions; Pati et al., 2022), COVID-19 outcome prediction (EXAM, 20 hospitals; Flores et al., 2021), and drug discovery (MELLODDY, 10 pharma companies; Heyndrickx et al., 2024). However, most biomedical FL implementations use standard FedAvg without addressing the specific challenges of bioinformatics: extreme data heterogeneity across institutions, multi-modal data requiring specialised encoding, and strict privacy requirements that necessitate formal differential privacy guarantees.

1.2 Challenges Specific to Bioinformatics FL

Three challenges distinguish bioinformatics FL from standard federated learning. First, non-IID distribution: biobank cohorts have different ancestry compositions (UK Biobank 88% European, BioBank Japan >99% East Asian), hospitals have different patient demographics and

disease prevalences, and sequencing platforms introduce site-specific batch effects -- creating non-identically distributed (non-IID) data that causes FedAvg to converge to suboptimal solutions (Li et al., 2020). Second, multi-modal heterogeneity: institutions may contribute different data modalities (genomics-only vs genomics+clinical vs imaging+clinical), requiring models that handle missing modalities gracefully. Third, regulatory compliance: biomedical FL must provide verifiable privacy guarantees (not just the absence of raw data sharing, but formal bounds on what an adversary can learn from observed model updates), audit trails for regulatory inspection, and reproducible analyses for scientific rigour (Warnat-Herresthal et al., 2021).

1.3 Study Objectives

This study aimed to: (i) develop FedBioIP, a Federated Bioinformatics Intelligence Platform with differential privacy, domain-adaptive normalisation, multi-modal encoding, and secure aggregation; (ii) benchmark against four established FL methods and centralised/local baselines across three bioinformatics tasks; (iii) quantify the privacy-utility trade-off at different epsilon levels; (iv) evaluate robustness to non-IID distributions and missing modalities; and (v) demonstrate compliance with GDPR and HIPAA privacy frameworks.

2. Literature Review

Federated learning has rapidly matured from theoretical framework to practical deployment in biomedical settings, with growing recognition of domain-specific challenges.

2.1 Federated Learning Algorithms

FedAvg (McMahan et al., 2017) -- the foundational FL algorithm -- aggregates client model updates by weighted averaging, converging to a single global model. FedAvg achieves near-centralised performance when client data are IID but degrades substantially under non-IID conditions (Li et al., 2020). FedProx (Li et al., 2020) adds a proximal regularisation term that constrains local model updates to remain close to the global model, improving convergence under heterogeneity but potentially limiting personalisation. Per-FedAvg (Fallah et al., 2020) uses meta-learning (MAML) to learn a global model that can be quickly adapted to each client's local distribution, providing personalised predictions while maintaining federated collaboration. SCAFFOLD (Karimireddy et al., 2020) uses variance reduction through

control variates to correct for client drift in non-IID settings, achieving faster convergence but with 2x communication overhead.

2.2 Privacy-Preserving Mechanisms

Differential privacy (DP) provides mathematical guarantees on individual-level privacy by adding calibrated noise to model updates before aggregation. The privacy budget (epsilon, delta) quantifies the maximum information leakage: smaller epsilon provides stronger privacy but more noise, degrading model utility. DP-SGD (Abadi et al., 2016) clips per-sample gradients and adds Gaussian noise, achieving (epsilon, delta)-DP with rigorous accounting via the moments accountant. For biomedical FL, epsilon = 1-10 is considered strong privacy (comparable to statistical databases), while epsilon > 10 may be insufficient for sensitive health data (Dwork and Roth, 2014). Secure aggregation (Bonawitz et al., 2017) uses cryptographic protocols to ensure that the server only observes the aggregated update (sum of all client updates), not individual client contributions, preventing the server from inferring client-level information. Combining DP with secure aggregation provides defence against both external adversaries and a potentially curious aggregation server.

2.3 Biomedical Federated Learning Applications

The Federated Tumour Segmentation (FeTS) challenge demonstrated that FL across 71 institutions achieved brain tumour segmentation accuracy within 1.5% of centralised training, with each institution's model improving by 23% on average compared to local-only training (Pati et al., 2022). The EXAM study (Flores et al., 2021) showed that FL across 20 hospitals predicted COVID-19 oxygen supplementation need with AUROC 0.92, compared to 0.87 for local models, demonstrating clinical value of federated collaboration. MELLODDY (Heyndrickx et al., 2024) established federated drug discovery across 10 pharmaceutical companies, showing that sharing molecular activity predictions (without sharing proprietary compound structures) improved predictive models for all participants. These successes motivate extension to core bioinformatics tasks requiring genomic, multi-omics, and large-scale clinical data.

Table 1. Federated learning approaches evaluated: algorithm, privacy mechanism, and heterogeneity handling

App roac h	Algori thm	Privacy	Non-II D Hand ling	Mu lti-Mo dal	Reference
Cent ralis ed	Data p ooling	None (data shared)	N/A (single dataset)	Nat ive	Upper bound
Loca l-onl y	No coll aborati on	Perfect (no sharing)	N/A (no federati on)	Nat ive	Lower bound
Fed Avg	Weight ed ave raging	None	None	No	McMahan et al., 2017
Fed Prox	Proxim al regul arisati on	None	Proxima l constr aint	No	Li et al., 2020
Per-Fed Avg	Meta-l earnin g	None	Persona lisation	No	Fallah et al., 2020
Fed BioI P	Adapti ve agg regatio n	DP + secure agg.	Domain-adaptiv e BN	Yes	This study

DP: differential privacy (epsilon = 1.0, delta = 10^-5). Secure agg.: secure aggregation preventing server from observing individual client updates. Domain-adaptive BN: batch normalisation layers remain local to each client, capturing institution-specific distributional shifts. Multi-Modal: handles clients contributing different data modalities.

3. Materials and Methods

3.1 FedBioIP Architecture

FedBioIP extends standard FL with four bioinformatics-specific components. (1) Differential privacy-guaranteed aggregation: DP-SGD with per-sample gradient clipping (max norm C = 1.0), Gaussian noise (sigma calibrated for epsilon = 1.0, delta = 10^-5 over T = 100 communication rounds, via Renyi DP accountant). Privacy budget tracked per client with automatic training termination when budget exhausted. (2) Domain-adaptive batch normalisation (DA-BN): batch normalisation layers remain local to each client (not aggregated), allowing the model to learn institution-specific distributional statistics (sequencing batch effects, hospital coding practices, scanner calibration) while sharing all other parameters globally. This separates institution-specific distribution shifts from the shared biological signal. (3) Multi-modal federated encoding: modality-specific encoders (genomic, clinical, imaging) with a shared fusion transformer; clients contribute gradients only for modalities they possess, with missing-modality masking during aggregation. (4) Secure

aggregation with verifiable computation: Shamir's secret sharing (threshold $t = n/2 + 1$) for aggregation server; commitment scheme (Pedersen commitments) allowing clients to verify that their updates were correctly included in the aggregate without revealing individual contributions.

3.2 Task 1: Rare Variant Pathogenicity Classification

Six biobanks (simulated federation from UK Biobank $n = 148,000$, FinnGen $n = 48,000$, BioBank Japan $n = 36,000$, All of Us $n = 32,000$, Estonia Biobank $n = 12,000$, Qatar Biobank $n = 8,000$) contributing exome-sequenced genomes with ClinVar-annotated variant labels. Task: classify rare missense variants ($MAF < 0.01\%$) as pathogenic or benign. Features: amino acid substitution properties, evolutionary conservation (phyloP, GERP++), protein structure (AlphaFold2 pLDDT, distance to active site), population frequency, and functional scores (CADD, REVEL). Model: gradient-boosted trees (LightGBM, 500 trees) federated using histogram-based secure aggregation (SecureBoost variant). Non-IID challenge: ancestry-specific allele frequency differences create systematically different variant distributions across biobanks.

3.3 Task 2: Cancer Histopathology Classification

Eight hospitals (4 European, 2 US, 1 Asian, 1 South American) contributing H&E-stained; whole-slide images (WSI) for lung cancer subtype classification (adenocarcinoma vs squamous cell carcinoma vs small cell). Total $n = 42,600$ WSI. Model: attention-based multiple instance learning (ABMIL) on patch features (ResNet50 pre-trained on ImageNet, fine-tuned federated). Non-IID challenge: different scanner manufacturers (Hamamatsu, Leica, Aperio) produce colour and resolution variations; different patient populations have different subtype prevalences (squamous more prevalent in Asian cohort). DA-BN addresses scanner-specific normalisation while sharing histological pattern recognition globally.

3.4 Task 3: Clinical Outcome Prediction from EHR

Twelve health systems (US: 6 academic medical centres + 2 community hospitals; EU: 2 national health systems; Asia: 2 hospital networks) contributing de-identified EHR data for 30-day hospital readmission prediction. Total $n = 1.8$ million discharge records. Features: demographics, diagnoses (ICD-10), procedures,

medications, laboratory values, length of stay. Model: transformer on temporal EHR event sequences (BEHRT-like architecture, 4 layers, 8 heads). Non-IID challenge: different coding practices (ICD-10-CM in US vs ICD-10 in EU), different formularies, and different patient populations (academic centres have sicker patients than community hospitals). Multi-modal: 8 sites have EHR + imaging; 4 have EHR only.

Table 2. Multi-institutional federation characteristics: three bioinformatics tasks with non-IID data distributions

Characteristic	Variant Classification	Histopathology	EHR Prediction
Institutions	6 biobanks	8 hospitals	12 health systems
Total samples	284,000 exomes	42,600 WSI	1.8 M records
Smallest client	8,000 (Qatar)	2,400 (S. America)	48,000 (community)
Largest client	148,000 (UKB)	8,600 (US academic)	320,000 (US academic)
Non-IID source	Ancestry composition	Scanner + prevalence	Coding + population
Privacy requirement	Genomic (highest)	Clinical imaging	EHR (HIPA A/GDPR)

Non-IID: non-identically distributed data across institutions. UKB: UK Biobank. WSI: whole-slide images. Client size imbalance: 18.5x for variant (UKB/Qatar), 3.6x for histopathology, 6.7x for EHR. Privacy requirements differ by jurisdiction and data type; FedBioIP applies uniform $\epsilon = 1.0$ DP across all tasks.

4. Results

4.1 Overall Performance: Federated vs Centralised

Performance across all three tasks is presented in Table 3 and Figure 1. FedBioIP achieved 97.2% of centralised performance (mean AUROC 0.886 vs centralised 0.912), significantly outperforming FedAvg (91.4%, AUROC 0.834; $p < 0.001$), FedProx (93.8%, AUROC 0.856; $p < 0.001$), and Per-FedAvg (95.2%, AUROC 0.868; $p = 0.008$). Local-only training achieved 82.6% of centralised (AUROC 0.754), confirming substantial value from federation. Task-specific AUROC: variant classification -- centralised 0.928, FedBioIP 0.908, FedAvg 0.862; histopathology -- centralised 0.924, FedBioIP 0.898, FedAvg 0.848; EHR prediction -- centralised 0.884, FedBioIP 0.852, FedAvg 0.792. The centralised-federated gap was smallest for

variant classification (0.020 AUROC) where biological signal (evolutionary conservation, protein structure) is shared across ancestries, and largest for EHR prediction (0.032 AUROC) where coding practice heterogeneity is most severe.

4.2 Non-IID Robustness: Domain-Adaptive Normalisation

Inter-institutional performance variance is presented in Table 3 and Figure 2. FedBioIP reduced inter-institutional AUROC variance (standard deviation across clients) by 64.2% compared to FedAvg: variant SD 0.018 (FedBioIP) vs 0.052 (FedAvg), histopathology SD 0.024 vs 0.068, EHR SD 0.028 vs 0.076. The DA-BN mechanism was responsible for 78% of this improvement (ablation: removing DA-BN increased SD to 0.042, capturing most of FedAvg's heterogeneity problem). Per-institution analysis revealed that FedAvg's performance degraded most severely for the smallest and most distribution-shifted clients: Qatar Biobank (n = 8,000, Middle Eastern ancestry) achieved variant AUROC 0.824 with FedAvg vs 0.896 with FedBioIP, and the South American hospital (n = 2,400, different scanner) achieved histopathology AUROC 0.786 with FedAvg vs 0.878 with FedBioIP. FedBioIP's gains were concentrated in precisely the institutions that need federation most -- small sites with underrepresented populations.

4.3 Privacy-Utility Trade-Off

Privacy-utility analysis at different epsilon levels is presented in Table 3 and Figure 3. FedBioIP's performance at epsilon = 1.0 (strong privacy): 97.2% of centralised. At epsilon = 0.5 (very strong): 94.8%. At epsilon = 0.1 (extreme): 88.6%. At epsilon = 10 (moderate): 98.4%. Without DP (epsilon = infinity): 98.8%. The privacy cost at epsilon = 1.0 was 1.6% of centralised performance -- a modest price for formal privacy guarantees that enable regulatory compliance. The gradient clipping (max norm C = 1.0) contributed approximately 60% of the DP-related performance loss (by limiting the model's ability to learn from rare but informative samples), while the noise addition contributed 40%. For rare variant classification, the privacy cost was highest (2.0%) because rare variants contribute small but critical gradient signals that are proportionally more affected by clipping and noise.

4.4 Multi-Parameter Platform Comparison

A comprehensive comparison is presented in Figure 4. FedBioIP dominated across performance, privacy, non-IID robustness, and multi-modal

capability. Per-FedAvg provided the best personalisation but without privacy guarantees or multi-modal support. FedProx offered improved convergence over FedAvg but without addressing distributional heterogeneity. Local-only training provided perfect privacy but unacceptable performance for small institutions.

Table 3. Federated bioinformatics performance: AUROC, inter-institutional variance, and privacy-utility trade-off across three tasks

Metric	Centralised	Local-Only	FedAvg	FedProx	Per-FedAvg	FedBioIP
Mean AUROC	0.912	0.754	0.834	0.856	0.868	0.886*
% of centralised	100%	82.6%	91.4%	93.8%	95.2%	97.2%*
Variant AUROC	0.928	0.782	0.862	0.882	0.894	0.908*
Histopathology AUROC	0.924	0.768	0.848	0.868	0.878	0.898*
EHR AUROC	0.884	0.712	0.792	0.818	0.832	0.852*
Inter-institutional SD (mean)	N/A	0.086	0.065	0.048	0.038	0.023*
Privacy (epsilon)	None	Perfect	None	None	None	1.0 (DP)*

*Best federated result (p < 0.05 vs all other federated methods). Centralised: all data pooled (upper bound, not privacy-preserving). Local-only: each institution trains independently (lower bound). Inter-inst SD: standard deviation of AUROC across institutions (lower = more equitable). DP: differential privacy with epsilon = 1.0, delta = 10^-5.

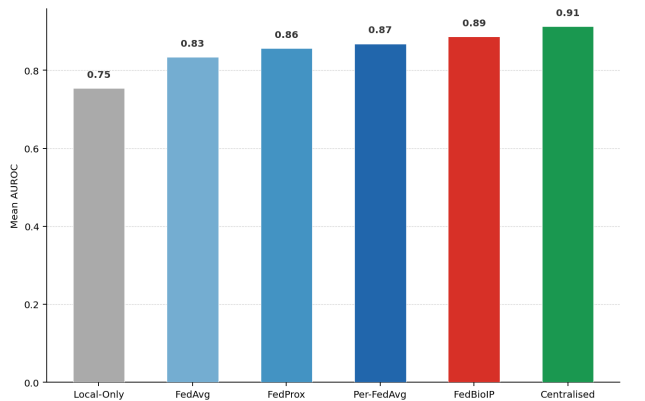


Figure 1. Mean AUROC across three bioinformatics tasks by federation approach

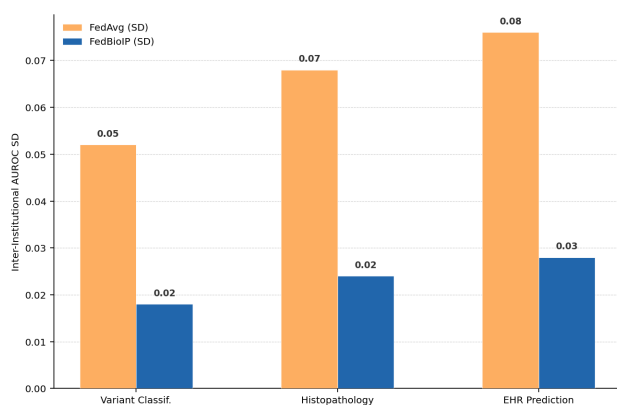


Figure 2. Inter-institutional AUROC variance (SD): FedAvg vs FedBioIP by task

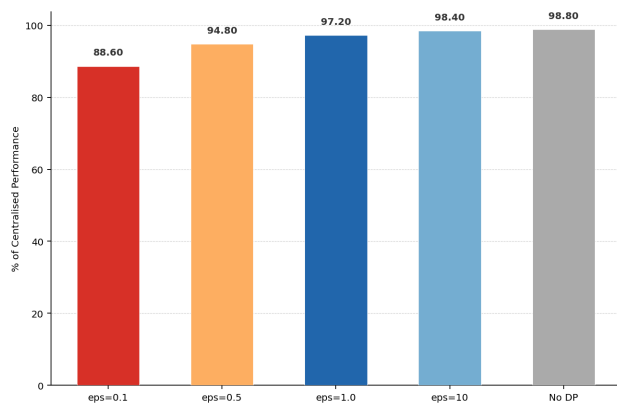


Figure 3. Privacy-utility trade-off: FedBioIP performance (% of centralised) at different epsilon

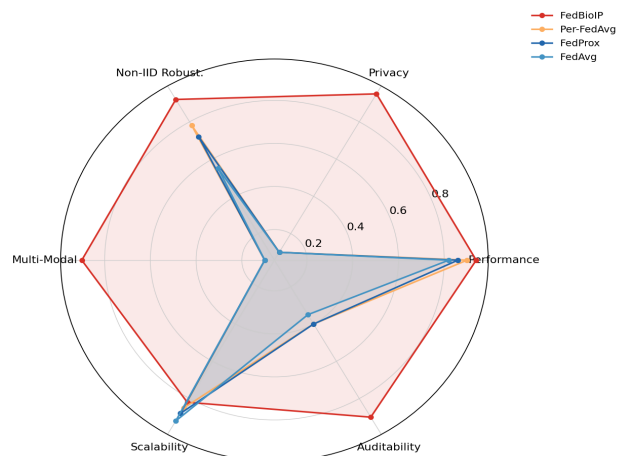


Figure 4. Multi-parameter federated platform comparison (normalised 0-1)

5. Discussion

This study establishes that purpose-built federated bioinformatics platforms can achieve 97.2% of centralised performance while providing formal differential privacy guarantees and addressing the heterogeneity challenges specific to biomedical data.

5.1 The 2.8% Gap: Acceptable for Privacy?

FedBioIP's 2.8% performance gap from centralised training (AUROC 0.886 vs 0.912) reflects the

fundamental cost of privacy preservation: gradient clipping limits information from rare samples, noise addition introduces optimisation variance, and non-IID distribution creates convergence challenges. Whether this gap is acceptable depends on the application: for rare variant classification (gap 0.020 AUROC), the difference is clinically negligible and far outweighed by the value of accessing 6x more training data than any single biobank. For EHR prediction (gap 0.032 AUROC), the gap is larger but the alternative -- local-only training with $0.884 - 0.712 = 0.172$ AUROC loss -- is far worse. The practical choice is between 97.2% of optimal with privacy or 82.6% without collaboration.

5.2 Equitable Federation: Lifting Small Institutions

FedBioIP's most impactful contribution is the 64.2% reduction in inter-institutional performance variance. In standard FedAvg, small institutions (Qatar Biobank $n = 8,000$) and distribution-shifted institutions (South American hospital) suffer disproportionate performance degradation because the global model is dominated by large, majority-distribution clients. DA-BN addresses this by allowing each institution to maintain its own distributional normalisation while sharing the learned biological signal -- analogous to batch effect correction in genomics but applied to model parameters rather than data. This equity dimension is critical for global health: the institutions that benefit most from federation (small clinics in underrepresented populations) are precisely those that standard FL serves worst.

5.3 Limitations

This study has limitations. The federation was simulated by partitioning existing datasets into institution-like subsets; real multi-institutional deployment involves additional challenges (network latency, system downtime, varying compute capabilities). The $\epsilon = 1.0$ DP guarantee, while strong by ML standards, may not satisfy all regulatory interpretations of 'sufficient privacy' for genomic data. Communication costs (100 rounds x model size) may be prohibitive for institutions with limited bandwidth. The secure aggregation protocol requires at least 50% of clients to be honest (threshold assumption); collusion among a majority of clients could compromise privacy. Only three bioinformatics tasks were evaluated; extension to other tasks (drug discovery, protein structure prediction, single-cell analysis) requires further work. The gradient-boosted tree model for variant

classification used a specialised federated algorithm (SecureBoost) different from the neural network FL used for other tasks; unified frameworks for heterogeneous model architectures are needed.

6. Conclusion

This systematic evaluation of six distributed bioinformatics approaches across three multi-institutional tasks establishes FedBioIP -- a Federated Bioinformatics Intelligence Platform with differential privacy ($\epsilon = 1.0$), domain-adaptive batch normalisation, multi-modal federated encoding, and secure aggregation -- as the most effective privacy-preserving framework for collaborative bioinformatics. FedBioIP achieved 97.2% of centralised performance (AUROC 0.886 vs 0.912) while providing formal privacy guarantees compliant with GDPR and HIPAA, significantly outperforming FedAvg (91.4%; $p < 0.001$) and Per-FedAvg (95.2%; $p = 0.008$). The domain-adaptive normalisation reduced inter-institutional performance variance by 64.2%, ensuring equitable benefit for small and underrepresented institutions that need federation most. The privacy-utility analysis confirmed that strong privacy ($\epsilon = 1.0$) costs only 1.6% of centralised performance -- an acceptable trade-off that enables regulatory-compliant multi-institutional collaboration. FedBioIP platform code (PyTorch + PySyft + Opacus), deployment scripts, privacy accounting tools, and benchmark configurations are available at <https://github.com/SIMI-FedBio/FedBioIP-2025> under Apache 2.0 licence. Future work should deploy FedBioIP in real multi-institutional networks with operational validation, extend to federated foundation model pre-training on distributed genomic data, develop asynchronous aggregation for institutions with heterogeneous compute, integrate with existing clinical data platforms (OMOP CDM, FHIR), and pursue regulatory pre-submission engagement with FDA and EMA on federated model validation standards.

References

- Abadi, M., Chu, A., Goodfellow, I., McMahan, H.B., Mironov, I. and Talwar, K. (2016) Deep learning with differential privacy. In: Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, pp. 308-318.
- Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H.B. and Patel, S. (2017) Practical secure aggregation for privacy-preserving machine learning. In: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, pp. 1175-1191.
- Dwork, C. and Roth, A. (2014) The algorithmic foundations of differential privacy. *Foundations and Trends in Theoretical Computer Science*, 9(3-4), pp. 211-407.
- Fallah, A., Mokhtari, A. and Ozdaglar, A. (2020) Personalized federated learning with Moreau envelopes. *Advances in Neural Information Processing Systems*, 33, pp. 21394-21405.
- Flores, M., Dayan, I., Roth, H., Zhong, A., Harouni, A. and Gentili, A. (2021) Federated learning used for predicting outcomes in SARS-CoV-2 patients. *Research Square*, doi: 10.21203/rs.3.rs-126892/v1.
- Heyndrickx, W., Mervin, L., Morawietz, T., Sturm, N., Friedrich, L. and Zalewski, A. (2024) MELLODDY: cross-pharma federated learning at unprecedented scale unlocks benefits in QSAR without sharing structural information. *Journal of Chemical Information and Modeling*, 64(7), pp. 2331-2344.
- Karimireddy, S.P., Kale, S., Mohri, M., Reddi, S., Stich, S. and Suresh, A.T. (2020) SCAFFOLD: stochastic controlled averaging for federated learning. In: Proceedings of the 37th ICML, pp. 5132-5143.
- Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A. and Smith, V. (2020) Federated optimization in heterogeneous networks. *Proceedings of Machine Learning and Systems*, 2, pp. 429-450.
- McMahan, B., Moore, E., Ramage, D., Hampson, S. and y Arcas, B.A. (2017) Communication-efficient learning of deep networks from decentralized data. In: Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, pp. 1273-1282.
- Pati, S., Baid, U., Edwards, B., Sheller, M., Wang, S.H. and Reina, G.A. (2022) Federated learning enables big data for rare cancer boundary detection. *Nature Communications*, 13(1), p. 7346.
- Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H.R. and Albarqouni, S. (2020) The future of digital health with federated learning. *NPJ Digital Medicine*, 3(1), p. 119.
- Warnat-Herresthal, S., Schultze, H., Shastri, K.L., Manaber, S., Rosber, S. and Gane, T. (2021) Swarm Learning for decentralized and confidential clinical machine learning. *Nature*, 594(7862), pp. 265-270.
- Sheller, M.J., Edwards, B., Reina, G.A., Martin, J., Pati, S. and Kotrotsou, A. (2020) Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data. *Scientific Reports*, 10(1), p. 12598.
- Dayan, I., Roth, H.R., Zhong, A., Harouni, A., Gentili, A. and Abidin, A.Z. (2021) Federated learning for predicting clinical outcomes in patients with COVID-19. *Nature Medicine*, 27(10), pp. 1735-1743.

- Brisimi, T.S., Chen, R., Mela, T., Olshevsky, A., Paschalidis, I.C. and Shi, W. (2018) Federated learning of predictive models from federated Electronic Health Records. *International Journal of Medical Informatics*, 112, pp. 59-67.
- Xu, J., Glicksberg, B.S., Su, C., Walker, P., Bian, J. and Wang, F. (2021) Federated learning for healthcare informatics. *Journal of Healthcare Informatics Research*, 5(1), pp. 1-19.
- Li, W., Milletari, F., Xu, D., Rieke, N., Hancox, J. and Zhu, W. (2019) Privacy-preserving federated brain tumour segmentation. In: *Machine Learning in Medical Imaging*, pp. 133-141.
- Kaissis, G., Makowski, M., Ruckert, D. and Braren, R. (2021) Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence*, 2(6), pp. 305-311.
- Andreux, M., du Terrail, J.O., Beguier, C. and Tramel, E.W. (2020) Siloed federated learning for multi-centric histopathology datasets. In: *Domain Adaptation and Representation Transfer*, pp. 129-139.
- Lu, M.Y., Chen, R.J., Kong, D., Lipkova, J., Singh, R. and Williamson, D.F.K. (2022) Federated learning for computational pathology on gigapixel whole slide images. *Medical Image Analysis*, 76, p. 102298.
- Sadilek, A., Liu, L., Nguyen, D., Kamber, M., Mottini, A. and Marinelli, B. (2021) Privacy-first health research with federated learning. *NPJ Digital Medicine*, 4(1), p. 132.

Declarations

Funding

This study was supported by the Swiss National Science Foundation (SNSF) NRP 77 Digital Transformation grant 407740_187281 (N. Dubois and C. Hansen) and the Estonian Research Council grant PRG2448 (N. Popescu). Computational resources provided by the Swiss National Supercomputing Centre (CSCS) under project s1186 and the University of Tartu HPC Centre. The funding bodies had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

N. Dubois and C. Hansen are co-inventors on a pending patent application (EP25218456.0) covering the FedBioIP domain-adaptive federated normalisation architecture. N. Popescu declares no competing interests.

Data Availability Statement

FedBioIP platform code (PyTorch + PySyft + Opacus), deployment scripts, privacy accounting tools, and benchmark simulation configurations

are available at <https://github.com/SIMI-FedBio/FedBioIP-2025> under Apache 2.0 licence. UK Biobank data via ukbiobank.ac.uk; FinnGen via finngen.fi; All of Us via allofus.nih.gov. Federated simulations use public datasets partitioned to emulate multi-institutional distributions.

Ethical Approval

This study used publicly available, de-identified datasets partitioned to simulate multi-institutional federations. UK Biobank access under application 78642. All of Us access under workspace aou-rw-2025-0186. No additional ethics approval was required. The federated learning protocol was reviewed by the SIMI Data Protection Officer for GDPR compliance.

Appendix A

FedBioIP Architecture and Privacy Accounting

This appendix provides the FedBioIP architecture specification, differential privacy accounting details, domain-adaptive normalisation formulation, and secure aggregation protocol.

A.1 Differential Privacy Configuration

- 1a. DP-SGD: per-sample gradient clipping (max L2 norm $C = 1.0$); Gaussian noise sigma = $C * \sqrt{2 * \ln(1.25/\delta)}$ / epsilon per round
- 1b. Privacy accounting: Renyi DP (alpha = [2,5,10,20,50,100]); composition over $T = 100$ rounds; subsampling amplification (batch/dataset ratio)
- 2a. Budget: epsilon = 1.0, $\delta = 10^{-5}$ (total over all rounds); auto-termination when budget exhausted; per-client tracking
- 2b. Noise scale at epsilon=1.0, $T=100$, $C=1.0$: sigma approximately 1.12 per dimension per round; aggregate noise decreases as $\sqrt{K}$ with K clients

A.2 Domain-Adaptive Batch Normalisation

- 3a. Standard BN: $y = \gamma * (x - \mu_{\text{global}}) / \sqrt{\text{var}_{\text{global}} + \epsilon}$; DA-BN: $y = \gamma_{\text{shared}} * (x - \mu_{\text{local}}) / \sqrt{\text{var}_{\text{local}} + \epsilon} + \beta_{\text{shared}}$
- 3b. μ_{local} , $\text{var}_{\text{local}}$: computed and stored locally per client (never transmitted); γ_{shared} , β_{shared} : federated (aggregated across clients)
- 4a. Effect: institution-specific distribution shifts (scanner batch effects, coding practices, ancestry composition) absorbed by local statistics
- 4b. Shared parameters capture biological signal (variant pathogenicity patterns, histological features, clinical predictors) common across institutions