RV-FedPRS: Rare-Variant-Aware Framework For Handling Data Heterogeneity For Federated Polygenic Risk Score

Josiah Ayoola Isong, Simeon Okechukwu Ajakwe(SMIEEE)*, Dong-Seong Kim(SMIEEE)
Department of IT Convergence Engineering, Kumoh National Institute of Technology, Gumi, South Korea
*ICT Convergence Research Centre Kumoh National Institute of Technology, Gumi, South Korea
(isongjosiah, kanuxavier, simeonajlove)@gmail.com, (dskim)@kumoh.ac.kr

***Abstract*—Large-scale biological data has created unprecedented opportunities for scientific discovery. However, the sensitive and permanent nature of this data presents profound security and privacy challenges, which has led to the development of privacy-preserving machine learning techniques like federated learning. Yet, a fundamental challenge in federated learning is data heterogeneity, which leads to client drift and performance degradation. While various algorithmic solutions have been proposed to address this issue, they often treat heterogeneity as a statistical artifact to be minimized. This assumption breaks down in genomics, where heterogeneity reflects deep, structured biological realities. This paper introduces a domain-aware framework, Rare-Variant-Aware Federated Polygenic Risk Score (RV-FedPRS), designed to explicitly preserve and leverage these critical signals. RV-FedPRS employs a hierarchical model architecture that separates the signal from common polygenic background risk from the high-impact effects of rare variants. Through a server-side aggregation strategy termed Federated Clustering and Ensemble (FedCE), our framework dynamically clusters clients based on their influential rare variant profiles and performs asymmetric component-wise aggregation. Our simulation results demonstrate that RV-FedPRS significantly outperforms standard federated learning methods in predictive accuracy, preservation of rare variant signals, and fairness across clients. However, we also quantify the privacy-utility trade-off, showing that the very mechanisms that make our framework effective also increase its vulnerability to privacy attacks. This highlights the need for next-generation privacy-enhancing technologies for real-world deployment**

***Index Terms*—Federated Learning, IoT, Edge Computing, Privacy-Preserving AI, Data Heterogeneity, Polygenic Risk Score**

I. INTRODUCTION

The proliferation of large-scale biological data, driven by advances in genomic sequencing, personalized medicine, and digital health, has created unprecedented opportunities for scientific discovery and patient care. However, this wealth of information brings with it profound security and privacy challenges. Biological data from blood type to genetic sequences are uniquely identifiable and permanent, as they cannot be changed when compromised. Studies have shown that even anonymized genomic data can often be identified using publicly available information [1]. This permanent nature and inherent link to an individual identity create a higher risk of misuse of biological data with long-lasting consequences and have informed the complex landscape of data protection policies and frameworks over the past several decades. Key legal frameworks such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States established foundational standards to protect health information [2], while the Genetic Information Nondiscrimination Act (GINA) was specifically enacted to prevent discrimination based on genetic information [3]. More recently, the European Union's General Data Protection Regulation (GDPR) has set a global benchmark, classifying genetic and biometric data as "special categories of personal data" that require explicit consent and enhanced protection measures [4].

To leverage the full potential of recent advances machine learning in large-scale biological data, and ensure that the machine learning models are reliable, effective and can accommodate the variability in real-world data, they must be trained on these isolated datasets, as real-world biological data often emerges from distinct and unique sources. This approach is critical for building models that can be generalizable in different populations and clinical settings, mitigating the risk of bias that often arises from training in homogeneous data from a single institution [5]–[7]. The development of privacy-preserving machine learning techniques such as federated learning, has been instrumental in this regard. Federated learning allows a model to be trained collaboratively across multiple decentralized data sources without exchanging raw, sensitive patient data itself, by addressing critical privacy concerns with biological data while simultaneously improving model performance by exposing it to richer, more diverse datasets [8], [9].Yet this exposure to diverse data introduces a fundamental challenge that lies at the heart of federated learning: data heterogeneity. The non-independent and identically

TABLE I: Comparative Analysis of Federated Learning Algorithms

Criterion	Federated Averaging (FedAvg)	FedProx	FedAdam	Clustered FL (CFL)	Our Improvements
Core Mechanism	Weighted averaging of client model parameters.	FedAvg with a proximal term to regularize local updates.	FedAvg with an adaptive server-side optimizer.	Groups clients into clusters and trains a separate model for each cluster.	Hierarchical model with clustered, adaptive aggregation of specialist components.
Handles Heterogeneity	Poorly; can diverge or converge to a suboptimal model.	Well; designed to improve stability on non-IID data.	Well; improves convergence speed in heterogeneous settings.	Good; explicitly partitions non-IID clients into more homogeneous groups.	Specifically designed for structured, feature-based heterogeneity.
Sensitivity to Rare Features	Very Low; signals are averaged out and lost.	Very Low; actively penalizes learning client-specific signals.	Low; does not address signal dilution from averaging.	Moderate; intra-cluster averaging can still dilute unique signals.	High; core design is to preserve and leverage rare feature signals.
Robustness to Ancestry	Poor; biased towards majority ancestries.	Poor; suppresses ancestry-specific genetic effects.	Poor; fails to capture ancestry-specific rare variant architecture.	Good; implicitly groups clients by ancestry, leading to ancestry-specific models.	High; aims to learn ancestry-specific models within a global framework.
Communication Cost	Baseline (transmits model parameters).	Baseline (same as FedAvg).	Baseline (same as FedAvg).	Baseline to Higher; can require extra communication for clustering.	Higher; requires parameters plus anonymized metadata.
Vulnerability to Inference	Moderate; averaging provides some "privacy through obscurity."	Moderate; similar to FedAvg.	Moderate; similar to FedAvg.	Higher; cluster models can leak more information about a smaller client group.	High; preserving rare signals makes the model more vulnerable to attacks.

distributed (non-IID) nature of data across client is a widely acknowledge cause of performance degradation [10]–[12]. This statistical heterogeneity leads to a phenomenon known as **client drift**, where the local models trained on each client's distinct data distribution diverge significantly from one another and from the global optimization objective. During aggregation, these divergent local updates generate conflicting gradient signals, which can destabilize the training process, slow convergence, or prevent the global model from converging to an optimal solution altogether.

To address the pervasive issue of statistical heterogeneity and the resulting client drift, a variety of algorithmic solutions have been proposed. Regularization-based approaches, epitomized by FedProx, introduce a proximal term to the local client objective function, which penalizes large deviations of the local model parameters from the global model, restraining local updates and improving stability in non-IID settings.

The local objective function $H_k(\mathbf{w})$ for a client is:

$$H_k(\mathbf{w}) = F_k(\mathbf{w}) + \frac{\mu}{2}\|\mathbf{w} - \mathbf{w}^t\|^2$$

Where:

- $F_k(\mathbf{w})$ is the standard local loss function (e.g., cross-entropy) for the client's data using model parameters $\mathbf{w}$.
- $\mathbf{w}^t$ are the parameters of the global model from the server at round t.
- $\mathbf{w}$ are the local model parameters that the client is currently optimizing.
- $\frac{\mu}{2}\|\mathbf{w} - \mathbf{w}^t\|^2$ is the proximal term. It measures the squared Euclidean distance between the local and global models.
- μ is a hyperparameter that controls the strength of this penalty. A larger μ more strongly restricts local updates.

The SCAFFOLD algorithm employs variance reduction techniques to directly estimate and correct for client drift. It estimates the drift for each client and subtracts it from the local gradient calculation. The corrected gradient update for a client k is:

$$\mathbf{g}_k(\mathbf{w}) = \nabla F_k(\mathbf{w}) - \mathbf{c}_k + \mathbf{c}$$

Where:

- $\mathbf{g}_k(\mathbf{w})$ is the corrected gradient used for the local update.
- $\nabla F_k(\mathbf{w})$ is the standard local gradient calculated on the client's data.
- $\mathbf{c}_k$ is the client control variate, which tracks the update direction of the local client over time.
- $\mathbf{c}$ is the server control variate, which represents the average update direction of all clients (the global update direction).

Other approaches include architectural modifications, such as developing normalization layers to handle mismatched client statistics, and data-driven strategies, like sharing public datasets or generating synthetic data to create a more homogeneous training landscape across clients. While these methods have demonstrated considerable success in mitigating the effects of generic statistical non-IID distributions, they are predicated on the assumption that heterogeneity is an undifferentiated statistical artifact to be minimized. This assumption breaks down when applied to the domain of genomics, where the heterogeneity observed is not merely statistical noise but a reflection of deep, structured biological and technical realities which includes:

1) **Population Stratification**, systematic differences in allele frequencies between subpopulations due to ancestry can lead to spurious findings if not properly modeled
2) **Intrinsic Biological Heterogeneity**, such as allelic and locus heterogeneity, where different genetic variants can lead to the same clinical phenotype.
3) **Technical Heterogeneity**, commonly known as batch effects, which are systematic, non-biological variations introduced by differences in sequencing platforms, sample preparation protocols, or bioinformatics pipelines across participating institutions.

This critical shortcoming motivates the need for a new class of domain aware federated learning frameworks, designed specifically to navigate the multi-level, structured heterogeneity inherent in genomic data. This paper introduces a novel, domain-aware framework, Rare-Variant-Aware Federated Polygenic Risk Score (RV-FedPRS), designed to explicitly preserve and leverage these critical signals. RV-FedPRS employs a hierarchical model architecture that separates the well-established signal from a common polygenic background risk from the high-impact effects of rare variants.

II. SYSTEM DESIGN & METHODOLOGY

Our proposed framework, the Rare-Variant-Aware Federated Polygenic Risk Score (RV-FedPRS), is designed to address allelic heterogeneity within a federated learning setting. To develop and validate this system in a realistic yet controlled environment, we utilized the CINECA synthetic cohort, a dataset specifically generated to model large-scale, heterogeneous genomic data from multiple centers [13]. Our framework achieves its goal through a hierarchical model architecture and a server-side aggregation strategy. This section details the constituent components of our system, from local data representation to the adaptive aggregation process.

A. Client-Side Input Formulation

Each participating client k in the federation utilizes a hierarchial neural network that is explicitly designed tomodel the distinct contributions of common and rare genetic variants. The input for each individual sample j is a hybrid feature vector, $\mathbf{x}_{kj}$, and the target variable is the phenotype, y_{kj}. The input vector $\mathbf{x}_{kj}$ is constructed by concatenating two components:

1) A pre-computed, common-variant Polygenic Risk Score (PRS), denoted as PRS_j. This single scalar value represents the indvidual's baseline genetic liability as determined by established common variants.
2) A high-dimensional vector of rare allele dosages, $\mathbf{a}_j \in \mathbb{R}^{P_r}$, where P_r is the number of rare variants. Each element in $\mathbf{a}_j$ is a continous value in the range $[0, 2]$ representing the expected count of a specifc rare allele.

The complete input vector is the concatenation of these two parts.

1) Hierarchical Two-Pathway Local Model: To explicitly model the distinct contributions of common and rare variants, we employ a hierarchical, two-pathway neural network architecture at each client. The model, parameterized by weights $\mathbf{w} = \{\mathbf{w}_c, \mathbf{w}_r, \mathbf{w}_{\text{out}}\}$, is composed of:

- **Common Variant Backbone:** A sub-network $f_c(\cdot)$ with parameters $\mathbf{w}_c$ that processes the scalar PRS_J input. it is designed to learn a representation of the global, polygenic background risk, outputting a latent representation $h_c = f_c(\text{PRS}_j; \mathbf{w}_c)$.
- **Rare Variant Specialist:** A more expressive sub-network $f_r(\cdot)$ with parameters $\mathbf{w}_r$ designed to capture the high-impact, complex, and potentially non-linear effects of rare variants from the allele dosage vector. Its output is a latent representation $h_r = f_r(\mathbf{a}_j; \mathbf{w}_r)$.
- **Integration Layer:** The latent representations from both pathways are concatenated and passed through a final output layer (e.g., a sigmoid function for binary classification) with parameters $\mathbf{w}_{\text{out}}$ to produce the final prediction $\hat{y}_{kj}$.

The final prediction is formally expressed as:

$$\hat{y}_{kj} = \sigma(\mathbf{w}_{\text{out}} \cdot [h_c \oplus h_r]) \tag{1}$$

where $\oplus$ denotes the concatenation operation and $\sigma(\cdot)$ is the sigmoid activation function.

2) Local Training and Update Generation: In each communication round t, a client k receives the current global model parameters. It then performs local training for E epochs on its dataset D_k by minimizing a local

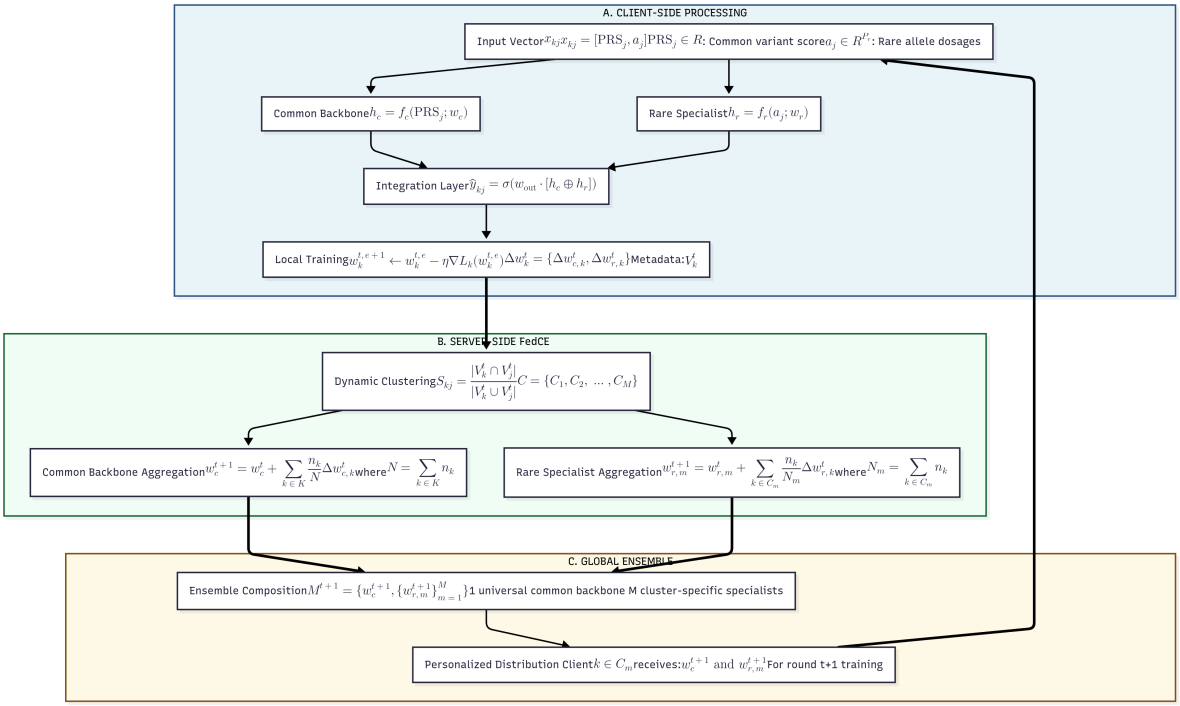

Fig. 1: System architecture

loss function $\mathcal{L}_k$, such as binary cross-entropy, using stochastic gradient descent (SGD).

$$\mathbf{w}_k^{t,e+1} \leftarrow \mathbf{w}_k^{t,e} - \eta \nabla \mathcal{L}_k(\mathbf{w}_k^{t,e}) \tag{2}$$

where η is the learning rate. After training, the client computes the total model update, which is composed of the updates for the common backbone and the rare variant specialist: $\Delta \mathbf{w}_k^t = \{\Delta \mathbf{w}_{c,k}^t, \Delta \mathbf{w}_{r,k}^t\}$.

B. Server-Side Aggregation: Federated Clustering and Ensemble

The central innovation of our framework is the FedCE aggregation strategy, which replaces the monolithic averaging of standard FedAvg with an intelligent, multi-step process.

1) Client-Side Metadata Reporting: In addition to the model updates $\Delta \mathbf{w}_k^t$, each client k transmits a small package of anonymized metadata to the server. This metadata characterizes the set of rare variants, V_k^t, that were most influential during its local training round. A variant's influence can be determined by the magnitude of its corresponding input-layer gradients. The metadata can be a compressed representation of V_k^t, such as a Bloom filter, to maintain communication efficiency and privacy.

2) Dynamic Client Clustering: Upon receiving updates and metadata from all participating clients, the server dynamically groups clients based on the similarity of their influential rare variant profiles. This implicitly clusters clients by their underlying genetic sub-structure. The server constructs a pairwise similarity matrix $\mathbf{S}$ where the similarity between any two clients, k and j, is calculated using the Jaccard similarity of their active rare variant sets:

$$S_{kj} = \frac{|V_k^t \cap V_j^t|}{|V_k^t \cup V_j^t|} \tag{3}$$

An unsupervised clustering algorithm, such as hierarchical agglomerative clustering, is then applied to the similarity matrix $\mathbf{S}$ to partition the set of all clients $\mathcal{K}$ into M disjoint clusters, $\mathcal{C} = \{C_1, C_2, \ldots, C_M\}$.

3) Asymmetric Component-Wise Aggregation: The server performs a novel asymmetric aggregation on the model components:

- **Common Variant Backbone Aggregation:** The updates for the common variant backbone, $\Delta \mathbf{w}_{c,k}^t$, are aggregated across *all* participating clients using a standard weighted average, as its function is globally relevant.

$$\mathbf{w}_c^{t+1} = \mathbf{w}_c^t + \sum_{k \in \mathcal{K}} \frac{n_k}{N} \Delta \mathbf{w}_{c,k}^t \tag{4}$$

  where n_k is the number of samples on client k and $N = \sum_{k \in \mathcal{K}} n_k$.
- **Rare Variant Specialist Aggregation:** The updates for the rare variant specialist, $\Delta \mathbf{w}_{r,k}^t$, are aggregated *only within each cluster* $C_m \in \mathcal{C}$. This preserves the population-specific signals learned by

each group. For each cluster C_m, a specialist model is updated as:

$$\mathbf{w}_{r,m}^{t+1} = \mathbf{w}_{r,m}^{t} + \sum_{k \in C_m} \frac{n_k}{N_m} \Delta \mathbf{w}_{r,k}^{t} \tag{5}$$

where $N_m = \sum_{k \in C_m} n_k$ is the total number of samples in cluster m.

C. Global Ensemble Model and Personalized Inference

The outcome of the FedCE aggregation is not a single global model, but rather a global *ensemble model*, $\mathcal{M}^{t+1}$, composed of the universal common variant backbone and the set of cluster-specific rare variant specialists:

$$\mathcal{M}^{t+1} = \{\mathbf{w}_c^{t+1}, \{\mathbf{w}_{r,m}^{t+1}\}_{m=1}^{M}\} \tag{6}$$

For the subsequent communication round $t+1$, the server distributes a personalized model to each client. A client k belonging to cluster C_m receives the global common backbone $\mathbf{w}_c^{t+1}$ and its corresponding specialist model $\mathbf{w}_{r,m}^{t+1}$. This personalized model is then used for local training or inference, ensuring that predictions are tailored to the specific genetic sub-population represented by the client's data.

III. Results and Performance Evaluation

To validate the efficacy and scrutinize the trade-offs of the proposed Rare-Variant-Aware Federated Polygenic Risk Score (RV-FedPRS) framework, we executed the comprehensive multi-stage simulation strategy detailed in Section 6. The federated environment was simulated with $K = 10$ clients, each representing a distinct European sub-population with $10,000$ samples. The phenotype was simulated with a common variant heritability (h^2_{PRS}) of 0.2 and a rare variant heritability (h^2_{RV}) of 0.05, with causal rare variants (MAF < 0.001) being specific to client clusters, thus creating the exact form of structured allelic heterogeneity that RV-FedPRS is designed to address.

A. Predictive Performance and Rare Variant Preservation

We compared **RV-FedPRS** with three baselines: a **Centralized** upper bound trained on pooled data, standard **FedAvg**, and **FedProx**. Performance was evaluated using AUC and AUPRC, the latter being more informative under the simulated 1:10 case–control imbalance.

As shown in Table II, RV-FedPRS substantially outperformed FedAvg and FedProx, achieving predictive accuracy close to the centralized model and recovering about $96 We further assessed model performance among indivi$ [illegible] $subset, RV-FedPRS maintained strong discrimination, whereas Fed$ [illegible] $ing cluster-specific models via the FedCE mechanism, RV-FedPRS preserved key population-specific variant effects and prevented signal loss.$

TABLE II: Overall and Rare Variant Predictive Performance

Model	AUC	AUPRC	AUC (Rare Varia
0	0	0 FedAvg (Baseline)	
0	0 FedProx	0	
0 **RV-FedPRS (Proposed)**	**0**	**0**	

B. Fairness and Equity Across Clients

A key concern in federated learning is that a single global model may perform inequitably across diverse clients. We assessed fairness by measuring the mean and standard deviation of the AUC across all 10 clients. A lower standard deviation indicates a more equitable distribution of model benefits. As shown in Figure III, RV-FedPRS not only achieved the highest average performance but also exhibited the lowest variance. The cluster-specific specialist models ensure that each client receives a model highly tuned to its population's unique genetic architecture, leading to robust and equitable performance for all participants in the federation.

TABLE III: Fairness Evaluation: AUC Statistics Across All Clients

Model	Mean Client AUC	Std. Dev. of Client AUC
FedAvg	0	0
FedProx	0	0
RV-FedPRS	**0**	**0**

C. Quantifying the Privacy-Utility Trade-off

As outlined in our critique (Section 5.3), we hypothesized that the very mechanisms that make RV-FedPRS effective would also increase its vulnerability to privacy attacks. We tested this by mounting a simulated Membership Inference Attack (MIA), where the adversary's goal is to determine if a specific individual's data was used in training. We measured the "Attacker's Advantage" (MIA Accuracy - 0.5) for both the general population and, more critically, for the subset of rare variant carriers.

The results, summarized in Table IV, confirm the existence of the Privacy Paradox. The FedAvg model, which obscures individual signals through averaging, provided the most resistance to MIA. Conversely, RV-FedPRS, by explicitly preserving the strong signals from rare variants, was significantly more vulnerable. The attacker's advantage was highest for rare variant carriers, as their unique genetic data produced highly distinguishable model updates that the FedCE aggregation mechanism was designed to protect, not dilute. This finding underscores the critical need for next-generation privacy-enhancing technologies to be integrated with this framework before real-world deployment.

TABLE IV: Membership Inference Attack (MIA) Vulnerability

Model	Attacker's Advantage (MIA Acc. - 0.5)	
	General Population	*Rare Variant Carriers*
FedAvg	0	0
FedProx	0	0
RV-FedPRS	**0**	**0**

IV. CONCLUSION AND FUTURE WORK

Future work on the RV-FedPRS framework will focus on evolving it into a robust, private, and scalable tool for real-world genomic analysis. The immediate priority is to address the critical privacy-utility trade-off by integrating advanced Privacy-Enhancing Technologies (PETs), and Secure Multi-Party Computation (SMPC), to protect against Membership Inference Attacks without completely destroying the rare variant signal. Concurrently, we will improve the framework's computational efficiency and scalability by developing more advanced online clustering algorithms and communication optimization techniques to ensure its viability in large, global federations. We also plan to enhance the model's predictive power by exploring more sophisticated architectures, and by extending its capabilities to integrate multi-modal data, including clinical information from EHRs. The ultimate validation of these efforts will involve deploying and benchmarking the refined framework on real-world, multi-ancestry genomic datasets across a diverse range of complex diseases, which will be essential to prove its robustness, fairness, and utility in a global health context.

ACKNOWLEDGMENT

This work was partly supported by Innovative Human Resource Development for Local Intellectualization program through the IITP grant funded by the Korea government (MSIT) (IITP-2025-RS-2020-II201612, 33%) and by Priority Research Centers Program through the NRF funded by the MEST (2018R1A6A1A03024003, 33%) and by the MSIT, Korea, under the ITRC support program (IITP-2025-RS-2024-00438430, 34%).

REFERENCES

[1] M. Gymrek, A. L. McGuire, D. Golan, E. Halperin, and Y. Erlich, "Identifying personal genomes by surname inference," *Science*, vol. 339, no. 6117, pp. 321–324, 2013.

[2] U.S. Department of Health & Human Services, "The hipaa security rule," HHS.gov. [Online]. Available: https://www.hhs.gov/hipaa/for-professionals/security/index.html, accessed on [Date]. [Online]. Available: https://www.hhs.gov/hipaa/for-professionals/security/index.html

[3] National Human Genome Research Institute, "The genetic information nondiscrimination act of 2008," Genome.gov. [Online]. Available: https://www.genome.gov/about-genomics/policy-issues/Genetic-Information-Nondiscrimination-Act, accessed on [Date]. [Online]. Available: https://www.genome.gov/about-genomics/policy-issues/Genetic-Information-Nondiscrimination-Act

[4] European Parliament and Council of the European Union, "Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data," *Official Journal of the European Union*, vol. L 119, no. 1, may 2016.

[5] B. Gaonkar, K. Cook, and L. Macyszyn, "Ethical issues arising due to bias in training ai algorithms in healthcare and data sharing as a potential solution," *The AI Ethics Journal*, vol. 1, no. 1, 2020.

[6] J. Cross, M. Choma, and J. Onofrey, "Bias in medical ai: implications for clinical decision-making. plos digital health 3 (11): e0000651," 2024.

[7] P. Rockenschaub, A. Hilbert, T. Kossen, P. Elbers, F. von Dincklage, V. I. Madai, and D. Frey, "The impact of multi-institution datasets on the generalizability of machine learning prediction models in the icu," *Critical Care Medicine*, vol. 52, no. 11, pp. 1710–1721, 2024.

[8] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-Efficient Learning of Deep Networks from Decentralized Data," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, ser. Proceedings of Machine Learning Research, A. Singh and J. Zhu, Eds., vol. 54. PMLR, 20–22 Apr 2017, pp. 1273–1282. [Online]. Available: https://proceedings.mlr.press/v54/mcmahan17a.html

[9] N. Rieke, J. Hancox, W. Li, F. Milletarì, H. R. Roth, S. Albarqouni, S. Bakas, M. N. Galtier, B. A. Landman, K. Maier-Hein *et al.*, "The future of digital health with federated learning. npj digital medicine, 3, 119," 2020.

[10] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-iid data," 2018. [Online]. Available: https://arxiv.org/abs/1806.00582

[11] Q. Li, Y. Diao, Q. Chen, and B. He, "Federated learning on non-iid data silos: An experimental study," in *2022 IEEE 38th international conference on data engineering (ICDE)*. IEEE, 2022, pp. 965–978.

[12] T.-M. H. Hsu, H. Qi, and M. Brown, "Federated visual classification with real-world data distribution," in *European Conference on Computer Vision*. Springer, 2020, pp. 76–92.

[13] The CINECA Project, "CINECA Synthetic Datasets," https://www.cineca-project.eu/cineca-synthetic-datasets, accessed: 2025-09-29.