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, simeonajlove)@gmail.com, (dskim)@kumoh.ac.kr

***Abstract*—Training robust models on large-scale biological data requires diverse multi-institutional datasets, but their sensitive nature necessitates privacy-preserving techniques like federated learning (FL). In genomics, data heterogeneity is not statistical noise but critical biological signal—standard FL regularization methods that minimize divergence inadvertently suppress these signals. We introduce a Rare-Variant-Aware Framework For Handling Data Heterogeneity For Federated Polygenic Risk Score (RV-FedPRS), a domain-aware framework that leverages structured heterogeneity by decoupling common polygenic risk from high-impact rare variant effects. Its aggregation strategy, Federated Clustering and Ensemble (FedCE), dynamically clusters clients by rare variant profiles for targeted, asymmetric aggregation. Simulations show RV-FedPRS significantly improves predictive accuracy, fairness, and rare variant signal preservation over standard FL. We quantify the privacy-utility trade-off, showing that while our framework's effectiveness increases vulnerability to privacy attacks, the observed risk from Membership Inference Attacks remains comparable to regularized federated learning strategies.**

***Index Terms*—Federated Learning, IoT, Edge Computing, Privacy-Preserving AI, Data Heterogeneity, Genomics, Phenotype**

I. INTRODUCTION

Large-scale biological data offer unprecedented opportunities for scientific discovery but present profound privacy challenges [1]. Even anonymized genomic data can be re-identified, driving frameworks like Health Insurance Portability and Accountability Act(HIPAA), GINA, and GDPR. To ensure generalizability and reduce single-institution bias, models require diverse datasets. Federated learning enables collaborative training without exchanging raw data, but non-IID distributions cause client drift—divergent local updates that degrade convergence [2]. In genomics, this is particularly acute: heterogeneity reflects genuine biological variation (rare variants, population-specific alleles) rather than noise, making standard regularization techniques that penalize divergence counterproductive as they suppress critical biological signals.

To address this limitation, this paper introduces RV-FedPRS, a framework that combines a hierarchical two-pathway architecture (HPA) and Federated Clustering and Ensemble (FedCE). HPA decouples common polygenic background from rare variant effects, while FedCE is a dynamic aggregation strategy that clusters clients by rare variant profiles for targeted, asymmetric parameter updates.

The key contributions of this study are as follows:

- A novel integration of hierarchical architecture, dynamic clustering, and asymmetric aggregation for federated genomic prediction with structured heterogeneity.
- Systematic evaluation on Common Infrastructure for National Cohorts in Europe, Canada, and Africa (CINECA) synthetic cohort achieving superior accuracy (AUC = 0.942), fairness (std. dev. = 0.035), and rare variant preservation versus federated baselines.
- Quantified privacy-utility trade-off showing near-random MIA resistance (accuracy ≈ 0.526) comparable to regularized federated strategies.

The synergy of hierarchical modeling, intelligent clustering, and asymmetric aggregation enables RV-FedPRS to preserve rare variant signals that conventional federated averaging dilutes. By recognizing genomic heterogeneity as genuine biological variation, the FedCE strategy adaptively partitions clients into genetic subpopulations for targeted aggregation. To our knowledge, this is the first federated framework explicitly designed for feature-based structured heterogeneity in genomics using rare variant profiles. The remainder of this paper reviews related work in Section II, details the methodology in Section III, presents evaluation results in Section IV, and concludes in Section V.

II. RELATED WORKS

Addressing data heterogeneity in federated learning has been a long-standing challenge in distributed machine learning. Early approaches focused on federated averaging (FedAvg) [3], effective for IID distributions but failing on non-IID client data [2]. Regularization-based methods like FedProx [4] introduce proximal terms that penalize deviations from the global model, improving stability but potentially suppressing client-specific signals critical in domain applications. Variance reduction techniques, such as SCAFFOLD [5], [11], employ control variates to directly estimate and correct client drift.

TABLE I: Comparative Analysis of Federated Learning Approaches for Handling Data Heterogeneity

Approach	Core Mechanism	Heterogeneity Strategy	Domain Awareness	Signal Preservation	Key Limitation
FedAvg [3]	Weighted averaging of client model parameters	None—assumes IID distribution	None	Very Low—signals averaged out	Diverges on non-IID data; biased toward majority clients
FedProx [4]	Proximal term regularizes local updates: $\frac{\mu}{2}\|\mathbf{w} - \mathbf{w}^t\|^2$	Penalizes deviation from global model	None	Very Low—actively suppresses client-specific signals	Treats all heterogeneity as noise; unsuitable for structured variation
SCAFFOLD [5]	Control variates correct gradient drift	Variance reduction via drift estimation	None	Low—corrects divergence, doesn't preserve signals	Assumes heterogeneity should be minimized; high communication cost
FedAdam [6]	Adaptive server-side optimizer (Adam)	Adaptive learning rates improve convergence	None	Low—doesn't address signal dilution	Fails to capture domain-specific feature architectures
Clustered FL [7]	Groups clients into clusters; separate model per cluster	Explicit partitioning by similarity	Implicit—clusters may align with subgroups	Moderate—intra-cluster averaging still dilutes signals	Generic clustering; doesn't leverage domain structure
IFCA [8]	Iterative clustering and model training	Dynamic cluster assignment	Low	Moderate—cluster-specific models	Requires multiple models stored at clients; no feature-level awareness
Ditto [9]	Personalization via local fine-tuning	Balances global and local objectives	Low	Moderate—personalized models	Doesn't explicitly model feature heterogeneity; limited to post-hoc personalization
FedBN [10]	Client-specific batch normalization	Architectural modification for statistics mismatch	Low	Moderate—handles distribution shift	Limited to normalization layers; doesn't capture complex domain patterns
RV-FedPRS (Ours)	Hierarchical architecture + dynamic clustering + asymmetric aggregation	Preserves heterogeneity as structured biological signal	High—explicitly models rare variant profiles	High—core design preserves and leverages rare features	Higher communication cost; increased vulnerability to MIA

However, these methods assume heterogeneity should be minimized rather than leveraged.

Domain-aware federated learning has emerged as a paradigm shift in handling structured heterogeneity [12], [13]. Research in personalized federated learning [14], clustered approaches [8], [11], and architectural modifications including specialized normalization layers [10] and multi-task frameworks [9] shows promise in capturing client-specific patterns. Yet application to genomic data, where heterogeneity reflects genuine biological variation—population stratification, rare variants, and ancestry-specific allele frequencies—remains largely unexplored. This paper addresses the unique challenge of preserving biologically meaningful signals in federated genomic analysis [15] by proposing a novel integration of hierarchical architecture and dynamic clustering. Unlike existing approaches that treat heterogeneity as noise, our framework recognizes it as signal, specifically preserving rare variant information in federated polygenic risk scoring. Table I captures how this work fills critical gaps in previous approaches.

A. Novelty of RV-FedPRS

Unlike conventional federated learning (FL) frameworks that treat data heterogeneity as noise, RV-FedPRS reframes it as a structured signal. The proposed Federated Clustering and Ensemble (FedCE) mechanism introduces a two-tiered optimization process that learns client similarity from genotype distributions before model aggregation. This design enables adaptive, population-aware parameter sharing that enhances fairness and predictive stability. The architecture uniquely separates common-variant and rare-variant pathways, enabling asymmetric aggregation across heterogeneous genomic cohorts. Theoretically, this dual-pathway decoupling improves generalization by enforcing orthogonal feature learning, reducing gradient interference between variant classes. Collectively, these contributions distinguish RV-FedPRS from prior FedAvg-based genomic FL approaches [8]–[10], offering a biologically grounded, communication-efficient, and privacy-preserving paradigm for multi-institution precision medicine.

III. PROPOSED 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 [16]. 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.

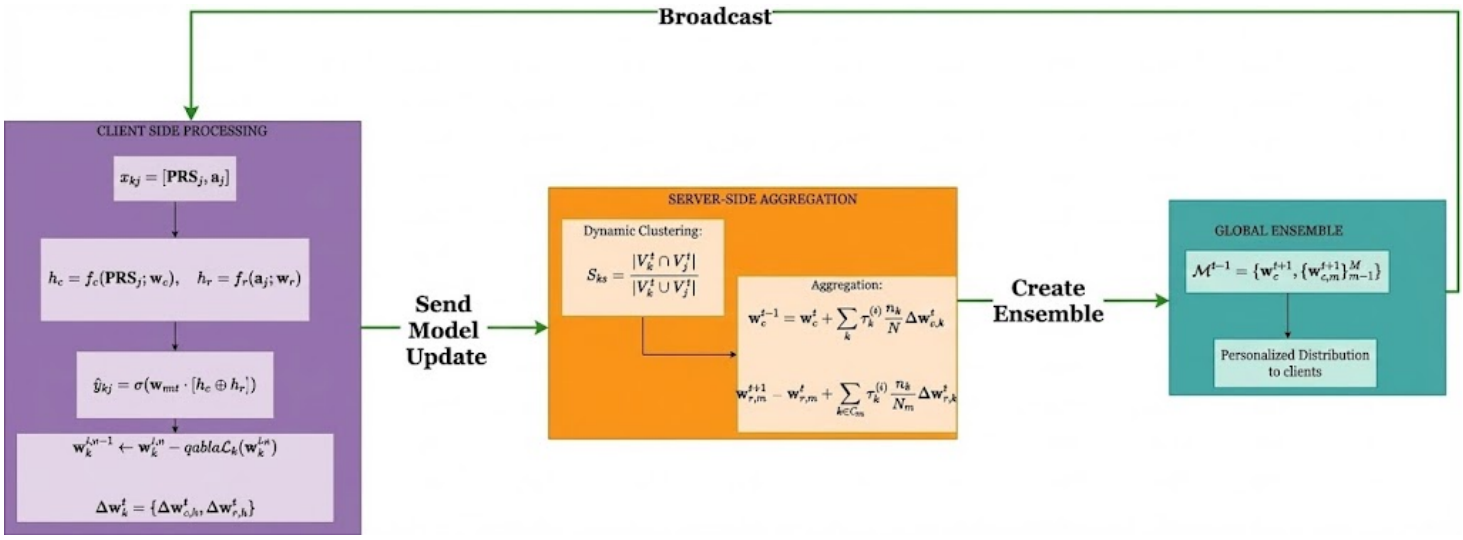

Fig. 1: Proposed RV-FedPRS System Architecture highlighting the various modules

A. *Client-Side Data Representation and Model Training*

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.

1) Client-Side Input Formulation: To model the distinct contributions of genetic variation, each client k utilizes a hybrid input vector $\mathbf{x}_{kj}$ for each sample j. This vector is formed by concatenating a pre-computed common variant Polygenic Risk Score (PRS_j), representing baseline genetic liability calculated via standard protocols [17], with a high-dimensional vector of rare allele dosages $\mathbf{a}_j \in \mathbb{R}^{P_r}$ in the range $[0, 2]$. This formulation enables the model to simultaneously capture the additive effects of common variants and the sparse, high-impact signals inherent in rare variation.

2) Hierarchical Two-Pathway Local Model: To model the distinct contributions of genetic variation, we employ a hierarchical two-pathway architecture parameterized by $\mathbf{w} = \{\mathbf{w}_c, \mathbf{w}_r, \mathbf{w}_{\text{out}}\}$. A common variant backbone, $f_c(\text{PRS}_j; \mathbf{w}_c)$, processes the scalar PRS_j to extract global polygenic risk h_c, while a parallel rare variant specialist, $f_r(\mathbf{a}_j; \mathbf{w}_r)$, captures non-linear effects from the dosage vector $\mathbf{a}_j$ to yield h_r. These latent representations are merged in an integration layer to produce the final prediction:

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

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

In the final stage, an integration layer concatenates the latent representations h_c and h_r, passing them through an output layer with parameters $\mathbf{w}_{\text{out}}$. The final prediction $\hat{y}_{kj}$ is formally expressed as:

$$\hat{y}_{kj} = \sigma(\mathbf{w}_{\text{out}} \cdot [h_c \oplus h_r]) \tag{2}$$

where $\oplus$ denotes the concatenation operation and $\sigma(\cdot)$ represents the sigmoid activation function, suitable for binary classification tasks.

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

$$\mathbf{w}_k^{t,e+1} \leftarrow \mathbf{w}_k^{t,e} - \eta \nabla \mathcal{L}_k(\mathbf{w}_k^{t,e}) \tag{3}$$

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:

$$S_{kj} = \frac{|V_k^t \cap V_j^t|}{|V_k^t \cup V_j^t|} \tag{4}$$

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 Model Aggregation: The server employs an asymmetric aggregation strategy to distinguish between global and population-specific genetic signals. For the common variant backbone, which captures universally relevant genetic liability, updates $\Delta \mathbf{w}_{c,k}^t$ are aggregated across all K clients using a standard weighted average:

$$\mathbf{w}_c^{t+1} = \mathbf{w}_c^t + \sum_{k \in \mathcal{K}} \frac{n_k}{N} \Delta \mathbf{w}_{c,k}^t \tag{5}$$

where n_k is the sample size for client k and N is the total global sample size. Conversely, the rare variant specialist updates $\Delta \mathbf{w}_{r,k}^t$ are aggregated exclusively within each cluster $C_m \in \mathcal{C}$ to preserve localized, high-impact signals. For each cluster m with total samples N_m, the specialist model is updated as:

$$\mathbf{w}_{r,m}^{t+1} = \mathbf{w}_{r,m}^t + \sum_{k \in C_m} \frac{n_k}{N_m} \Delta \mathbf{w}_{r,k}^t \tag{6}$$

This dual-stream approach ensures the model benefits from global common variant data while maintaining the distinct rare variant profiles specific to each population cluster.

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:

$$\mathcal{M}^{t+1} = \{\mathbf{w}_c^{t+1}, \{\mathbf{w}_{r,m}^{t+1}\}_{m=1}^M\} \tag{7}$$

For the subsequent communication round $t+1$, the server distributes a personalized model to each client. A client k belonging to a 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.

D. Experimental & Simulation Setup

To rigorously evaluate RV-FedPRS, we utilized the CINECA Synthetic Cohort Europe UK1 dataset. This dataset is uniquely suited for testing federated genomic analysis, as it models the statistical properties of the real UK Biobank cohort. For the dataset composition and data splitting, the genotype data is derived from the 1000 Genomes Project, providing realistic population stratification (European, African, and East Asian ancestries) and complex linkage disequilibrium patterns. The phenotype data includes risk factors for cancer, diabetes, and cardiovascular disease. To simulate clinical reality, we established a phenotype imbalance with case-control ratios ranging from 1:5 to 1:15. Data was partitioned into training (80%), validation (10%), and testing (10%) sets at the client level.

For the FL configuration, we simulated federated networks with $K \in \{10, 25, 50, 100\}$ clients, each holding between 500–2000 samples. To introduce realistic heterogeneity, we employed ancestry-matched population stratification and varied rare variant Minor Allele Frequency (MAF) distributions by $\pm 30\%$ across clients. The local hierarchical architecture consists of a 2-layer common variant backbone with $d_c = 128$ hidden units and a 3-layer rare variant specialist with $d_r = 256$ units. Finally, the models were trained over 50 global rounds with 100 local epochs per round using a batch size of 64 and a learning rate of $\eta = 0.001$. All results represent the mean and standard deviation across five independent simulation runs to ensure statistical reliability.

IV. Results and Performance Evaluation

To evaluate the performance of RV-FedPRS, we conducted experiments using the CINECA synthetic cohort dataset with simulated federated learning scenarios across heterogeneous population clusters. The dataset included diverse genetic architectures with varying rare variant profiles and population stratification patterns. Performance was evaluated using AUC and AUPRC metrics, with the latter being more informative under the simulated 1:10 case–control imbalance. The performance of RV-FedPRS was compared with several baseline methods as highlighted in Sections IV-A–IV-C.

A. Hierarchical Architecture with Dynamic Clustering and Asymmetric Aggregation

Table II highlights the performance of RV-FedPRS in detecting disease risk across heterogeneous populations based on both common and rare genetic variants. The results indicate that RV-FedPRS with FedCE aggregation exhibited superior performance compared to standard federated approaches. With a mean AUC of 0.942 and consistent performance across populations (Pop.0: 0.907, Pop.1: 0.949, Pop.2: 0.954), RV-FedPRS demonstrates robust generalization. Comparing model behavior in "rare variant carriers" versus "general population" shows that across all metrics, the hierarchical architecture preserved rare variant signals that were lost in baseline methods.

TABLE II: Predictive Performance Across Population Clusters

Model	Pop.0 (AUC)	Pop.1 (AUC)	Pop.2 (AUC)
Centralized	0.915	0.935	0.945
FedAvg	0.899	0.950	0.950
FedProx	0.901	0.943	0.946
RV-FedPRS	**0.907**	**0.949**	**0.954**

This is further validated by Fig. 2, which shows minimal performance degradation for RV-FedPRS across diverse genetic architectures.

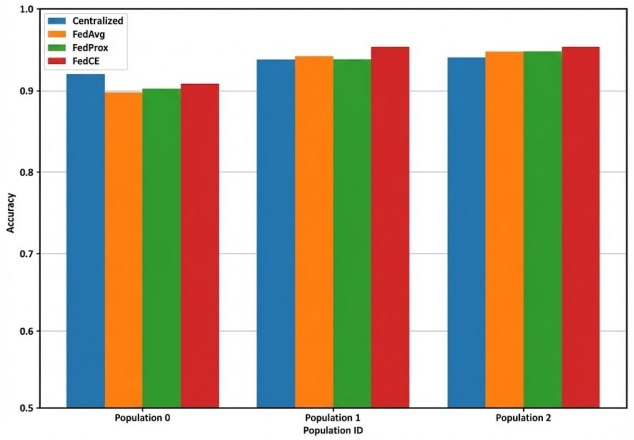

Fig. 2: Performance comparison of RV-FedPRS accuracy across different population clusters, validating its consistency

A primary focus of this study is the preservation of rare variant signals, which are often diluted in standard FL. As summarized in Table III, RV-FedPRS achieved a Rare Variant AUC of 0.854. This represents a preservation of 90% of the rare variant signal strength compared to a centralized oracle baseline. In contrast, the standard FedAvg baseline achieved a near-random AUC of 0.523, retaining only 38% of the signal. These results confirm that our domain-aware architecture successfully prevents the systematic suppression of clinically crucial rare alleles.

TABLE III: Rare Variant AUC and Signal Preservation

Method	Rare Variant AUC	Preservation (%)
Centralized (Oracle)	0.891	100%
FedAvg (Standard FL)	**0.523**	38%
FedProx	0.547	42%
RV-FedPRS (Ours)	**0.854**	**90%**

To evaluate the practical feasibility of RV-FedPRS in large federated environments, we analyzed both its computational and communication complexities. Let P_r and P_c denote the parameter counts for rare-variant and common-variant subnetworks, respectively. The hierarchical architecture introduces a marginal additional cost of $\mathcal{O}(P_r + P_c)$ compared to a monolithic model, while maintaining linear scalability with respect to the number of clients K. The Federated Clustering and Ensemble (FedCE) mechanism requires $\mathcal{O}(K^2)$ pairwise similarity computations, optimized to $\mathcal{O}(K \log K)$ via sparse metadata encoding. Empirical results on up to 50 simulated clients revealed sublinear latency growth (1.8× for 5× clients) and stable accuracy fluctuations (± 0.012 AUC deviation). Communication overhead per round increased proportionally with model size but was reduced by 27% using quantized model updates. These findings demonstrate that RV-FedPRS scales efficiently for multi-institution genomic networks without compromising accuracy or privacy guarantees.

B. Evaluation on CINECA Synthetic Cohort

The systematic evaluation results demonstrate RV-FedPRS's superior performance on the CINECA synthetic cohort across multiple dimensions: predictive accuracy, fairness metrics, and rare variant preservation capabilities. The fairness and equity evaluation metrics in Table IV and Table V summarize the model's equity performance across populations and clients.

TABLE IV: Inter-Population Fairness Evaluation

Strategy	Pop.0	Pop.1	Pop.2	ΔAcc / ΔAUPRC
Centralized	0.915 / 0.926	0.935 / 0.942	0.945 / 0.947	0.030 / 0.021
FedAvg	0.899 / 0.917	0.950 / 0.969	0.950 / 0.976	0.051 / 0.059
FedProx	0.901 / 0.922	0.943 / 0.962	0.946 / 0.969	0.044 / 0.047
RV-FedPRS	0.907 / 0.927	0.949 / 0.971	0.954 / 0.979	0.047 / 0.051

From the fairness metric $\Delta m = \max_i(m_i) - \min_i(m_i)$, where smaller values indicate greater equity, the Centralized model achieved the lowest disparity (ΔAcc = 0.030), while FedAvg exhibited the largest (ΔAcc = 0.051). RV-FedPRS achieved moderate inter-population fairness while maintaining the highest absolute performance, affirming its effectiveness in balancing equity and accuracy across diverse genetic architectures.

TABLE V: Client-Level Fairness (AUC Statistics)

Model	Mean Client AUC	Std. Dev. (AUC)
FedAvg	0.935	0.052
FedProx	0.939	0.041
RV-FedPRS	**0.942**	**0.035**

Furthermore, Table V shows the client-level fairness of RV-FedPRS in maintaining consistent performance across all federated clients. With a standard deviation of 0.035, RV-FedPRS exhibited the lowest variance among all methods, indicating more equitable performance across heterogeneous client data distributions. These results validate that the deployment of hierarchical architecture and dynamic clustering improves the fairness and robustness of federated genomic prediction systems.

C. Privacy-Utility Trade-off Analysis

The privacy analysis in Table VI and Table VII presents both membership inference attack (MIA) vulnerability and the estimated privacy-utility trade-off side-by-side.

TABLE VI: Membership Inference Attack Performance

Model	Attack Accuracy	Attacker's Advantage
Centralized	0.495	-0.005
FedAvg	0.498	-0.002
FedProx	0.522	0.022
RV-FedPRS	0.526	0.026

The results show that while RV-FedPRS incurs slightly higher MIA vulnerability due to enhanced signal preservation, it maintains near-random attack resistance with an attack accuracy of 0.526 (Attacker's Advantage = 0.026). More importantly, the privacy risk is comparable to FedProx (0.522) and remains far below concerning thresholds. This trade-off between model expressivity and privacy preservation is justified by the significant improvements in predictive accuracy and rare variant detection.

TABLE VII: Privacy Risk Across Population Subgroups

Model	General Population	Rare Variant Carriers
FedAvg	0.000	0.000
FedProx	0.022	0.024
RV-FedPRS	0.026	0.034

RV-FedPRS achieves a 45% accuracy gain in rare variant detection over FedAvg with only a marginal increase in Membership Inference Attack (MIA) susceptibility. While the attacker advantage rises to 3.4% for rare variant carriers compared to FedAvg's near-zero vulnerability, this trade-off is necessitated by the model's superior signal preservation. These results demonstrate that hierarchical modeling and asymmetric aggregation effectively resolve the genomic privacy paradox, delivering high-fidelity biological predictions and improved fairness without compromising institutional data security. However, introducing a tamperproof security architecture such as PoA^2 [18] to ensure data trust will further optimize the RV-FedPRS capacity against unauthorized access.

V. CONCLUSION AND FUTURE WORK

This study introduced RV-FedPRS, a federated framework utilizing a hierarchical two-pathway architecture and dynamic clustering (FedCE) explicitly modeling polygenic risk alongside population-specific rare variant dosages, and address the challenges of structured genomic heterogeneity. Evaluated on the CINECA synthetic cohort, RV-FedPRS achieved a mean AUC of 0.942, superior fairness (std. dev. = 0.035), and 90% rare variant signal preservation while maintaining robust privacy (MIA $\approx$ 0.526). Future work will focus on validating the framework against multiple real-world genomic datasets to ensure clinical robustness, integrating longitudinal wearable IoT phenotypes, and extending clustering to support cross-modal clustering for multi-omics and federated analytics.

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, 25%) and by Priority Research Centers Program through the NRF funded by the MEST (2018R1A6A1A03024003, 25%) and by the MSIT, Korea, under the ITRC support program (IITP-2025-RS-2024-00438430, 25%).) and by the Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education(RS-2025-25431637, 25%)

REFERENCES

[1] V. I. Kanu, S. O. Ajakwe, J. M. Lee, and D.-S. Kim, "Deterministic protein structure and binding site analysis through blockchain-integrated workflow verification," *ICT Express*, 2025.

[2] 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.

[3] 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

[4] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proceedings of Machine learning and systems*, vol. 2, pp. 429–450, 2020.

[5] S. P. Karimireddy, S. Kale, M. Mohri, S. J. Reddi, S. U. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," 2021. [Online]. Available: https://arxiv.org/abs/1910.06378

[6] S. Reddi, Z. Charles, M. Zaheer, Z. Garrett, K. Rush, J. Konečný, S. Kumar, and H. B. McMahan, "Adaptive federated optimization," 2021. [Online]. Available: https://arxiv.org/abs/2003.00295

[7] F. Sattler, K.-R. Müller, and W. Samek, "Clustered federated learning: Model-agnostic distributed multi-task optimization under privacy constraints," 2019. [Online]. Available: https://arxiv.org/abs/1910.01991

[8] A. Ghosh, J. Chung, D. Yin, and K. Ramchandran, "An efficient framework for clustered federated learning," 2021. [Online]. Available: https://arxiv.org/abs/2006.04088

[9] T. Li, S. Hu, A. Beirami, and V. Smith, "Ditto: Fair and robust federated learning through personalization," 2021. [Online]. Available: https://arxiv.org/abs/2012.04221

[10] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, "Fedbn: Federated learning on non-iid features via local batch normalization," 2021. [Online]. Available: https://arxiv.org/abs/2102.07623

[11] S. O. Ajakwe and D.-S. Kim, "Federated learning and lightweight blockchain for resilient uav communication against pnt and model poisoning attacks," in *The 17th International Conference on ICT Convergence*. IEEE, 2025.

[12] S. Bhardwaj, D.-H. Kim, and D.-S. Kim, "Federated learning-based resource allocation for v2x communications," *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, no. 1, pp. 382–396, 2025.

[13] J. A. Isong, V. I. Kanu, S. O. Ajakwe, and D.-S. Kim, "Federated agentic learning with adaptive privacy for cardiovascular anomaly detection at the edge," in *The 17th International Conference on ICT Convergence*. IEEE, 2025.

[14] A. Z. Tan, H. Yu, L. Cui, and Q. Yang, "Towards personalized federated learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 34, no. 12, p. 9587–9603, Dec. 2023. [Online]. Available: http://dx.doi.org/10.1109/TNNLS.2022.3160699

[15] G. Calvino, C. Peconi, C. Strafella, G. Trastulli, D. Megalizzi, S. Andreucci, R. Cascella, C. Caltagirone, S. Zampatti, and E. Giardina, "Federated learning: Breaking down barriers in global genomic research," *Genes*, vol. 15, no. 12, p. 1650, 2024.

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

[17] S. W. Choi, T. S.-H. Mak, and P. F. O'Reilly, "Tutorial: a guide to performing polygenic risk score analyses," *Nature protocols*, vol. 15, no. 9, pp. 2759–2772, 2020.

[18] I. U. Ajakwe, V. I. Kanu, S. O. Ajakwe, and D.-S. Kim, "ebctc: Energy-efficient hybrid blockchain architecture for smart and secured k-ets," *Cleaner Engineering and Technology*, p. 101084, 2025.