The Genomic Privacy Paradox

Scaling Diversity Without Centralizing Sensitive Data

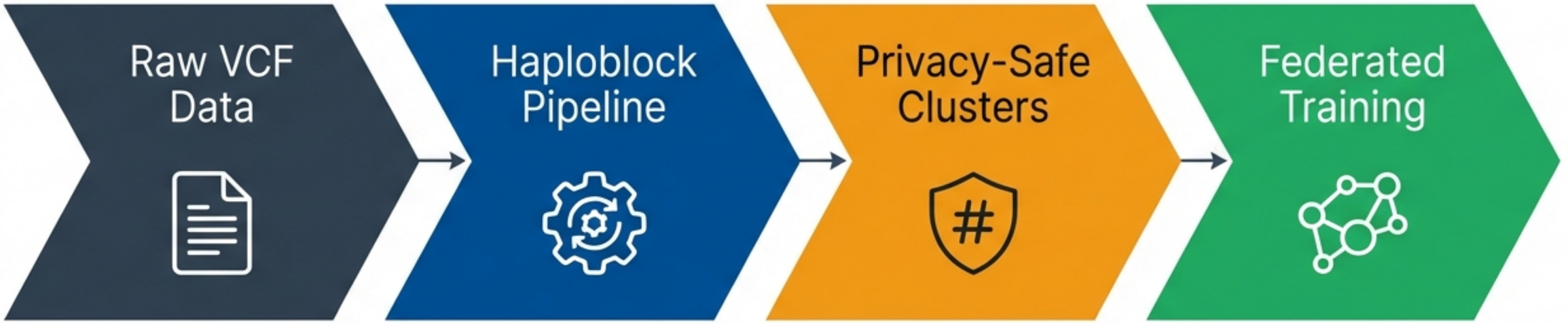
Barrier: Privacy Laws & Ethics

Need: Massive Diverse Datasets

The Problem: Population stratification requires large datasets to correct allele frequency discrepancies across ancestries. **The Barrier:** Traditional PCA and raw genotype sharing violate strict privacy laws (GDPR) **The Barrier:** Traditional PCA and raw genotypi sharing violate strict privacy laws (GDPR) that demand data sovereignty. **The Goal:** Connect isolated biobanks without moving the raw VCF files.

Med_SNP_Deconvolution: Privacy-Preserving Inference

Transforming Raw Genotypes into Safe, Abstract Features



Core Mechanism: Instead of sharing individual SNPs, we share recombination-defined genomic hashes. These discrete categorical features preserve population structure while obscuring individual-level variation.

Validation: Tested on Chromosome 6 (2,288 haploblocks) using 2,548 individuals from 1000 Genomes Project Phase 3.

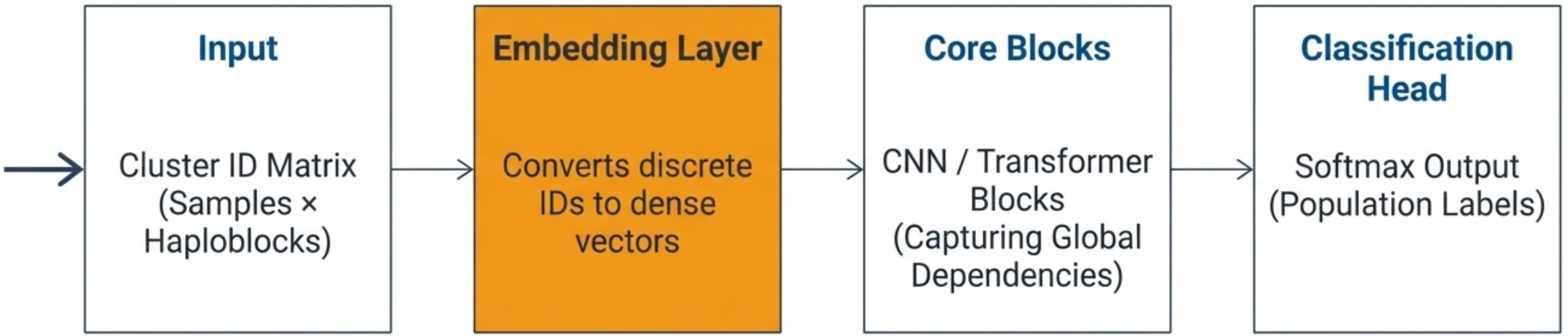
Unified Model Abstraction Layer

Interchangeable Backends for Speed or Complexity

Strategy A: XGBoost (Baseline)			Strategy B: Deep Learning (Advanced)
- **Input:** Sparse Genotype Matrix or Cluster IDs - **Method:** Histogram-based Gradient Boosting (GPU) - **Strengths:** Fast execution, handles sparsity efficiently, provides interpretable feature importance (Haploblock ranking).			- **Input:** Cluster ID Matrix (Samples × Haploblocks) - **Method:** Attention-Based Architecture - **Strengths:** Captures long-range genomic patterns via embedding layers; learns complex non-linear representations of ancestry.

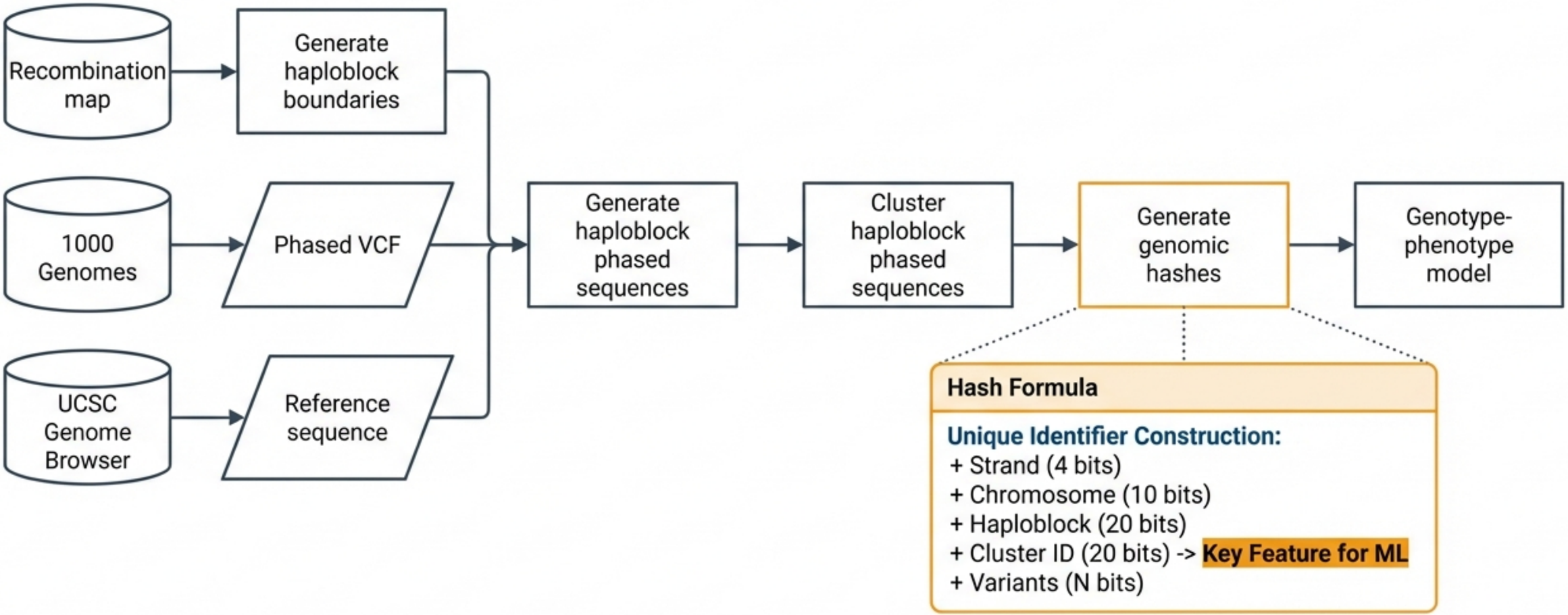
Deep Learning Architecture for Genomic Hashes

Deep learning model structure for processing genomic data.



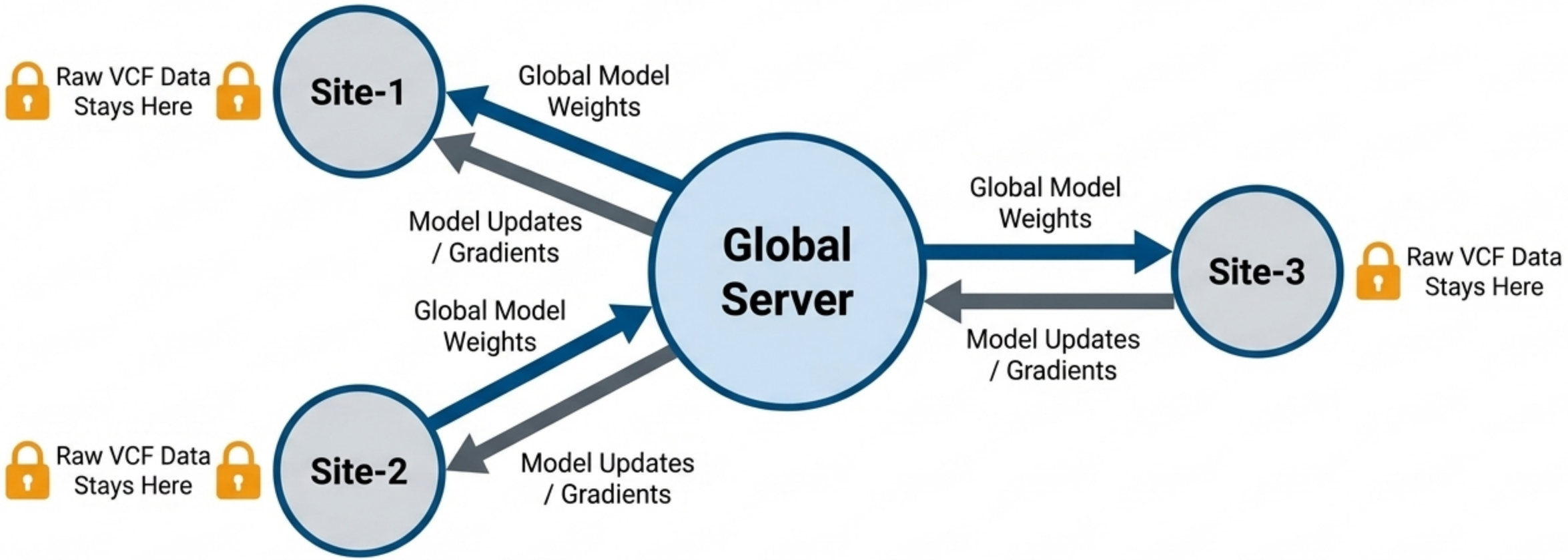
Tech Stack: PyTorch Lightning integrated with NVIDIA FLARE for seamless federated deployment.

The Anonymization Engine: Generating Genomic Hashes



Distributed Training with NVIDIA FLARE

Compute-to-Data Paradigm



Workflow: Clients download the model, train locally on private genomic hashes, and upload updates. Raw data never leaves the secure executor.

Experimental Strategy: Simulating Real-World Heterogeneity

Algorithms Tested

- **FedAvg**: Standard Federated Averaging.
- **FedProx**: Handles system heterogeneity.
- **Scaffold**: Controls for client drift.

Data Distribution Scenarios

IID (Ideal)

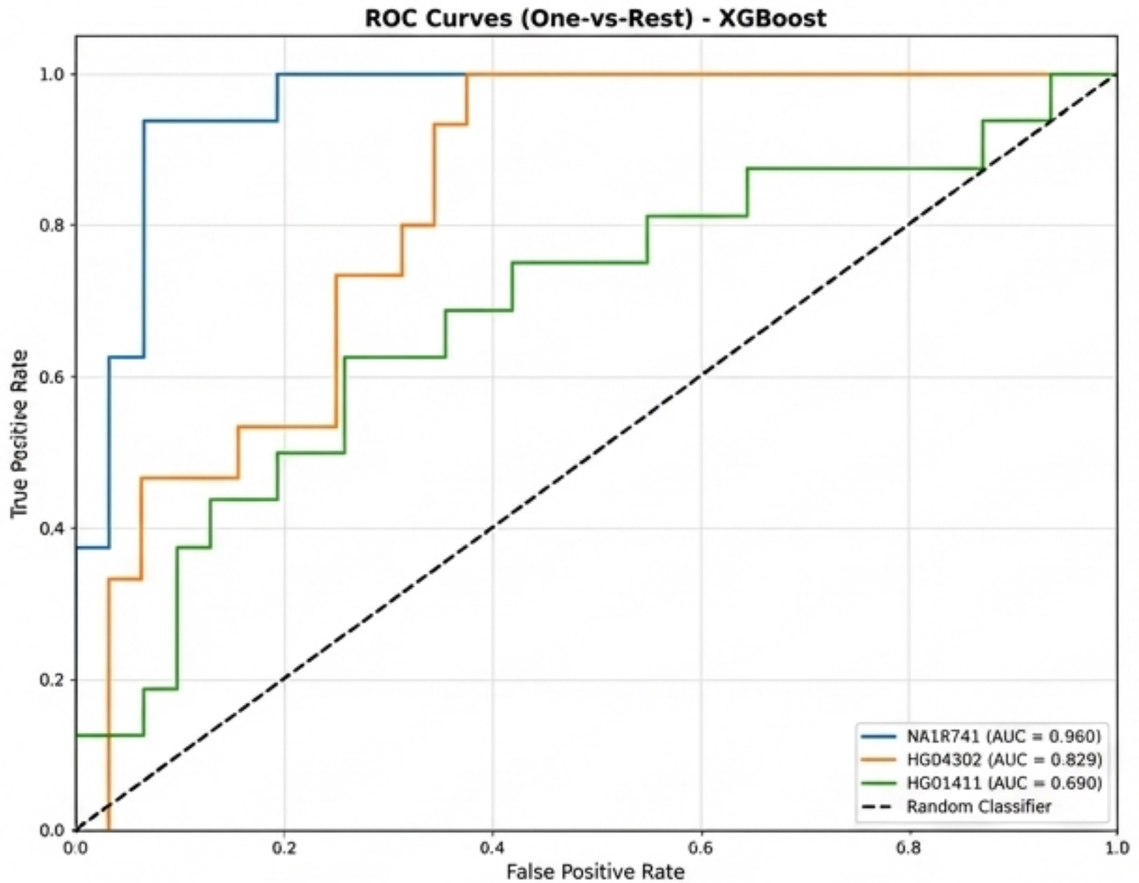
Data distributed evenly across sites.

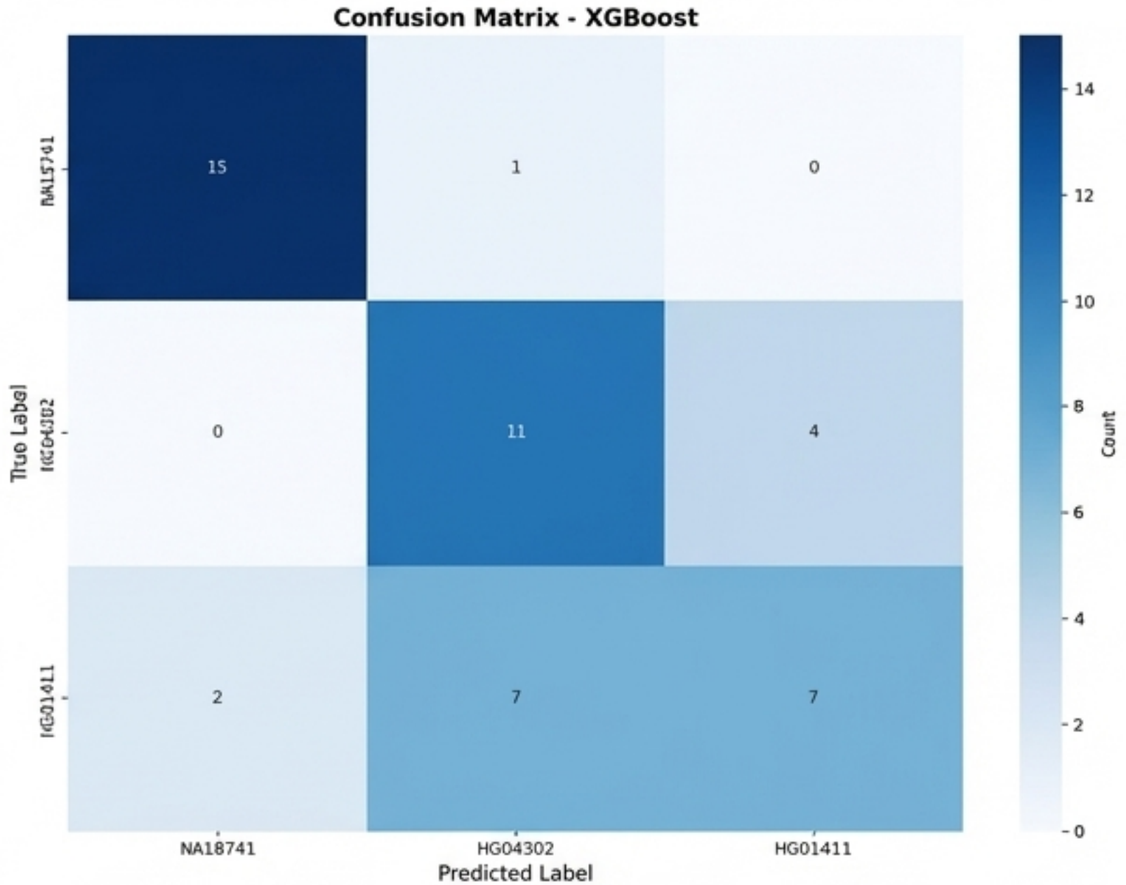
Dirichlet / Non-IID (Real-World)

Data skew simulates population-specific biobanks.

Baseline Validation: Centralized XGBoost

Proving the Haploblock Method on Centralized Data

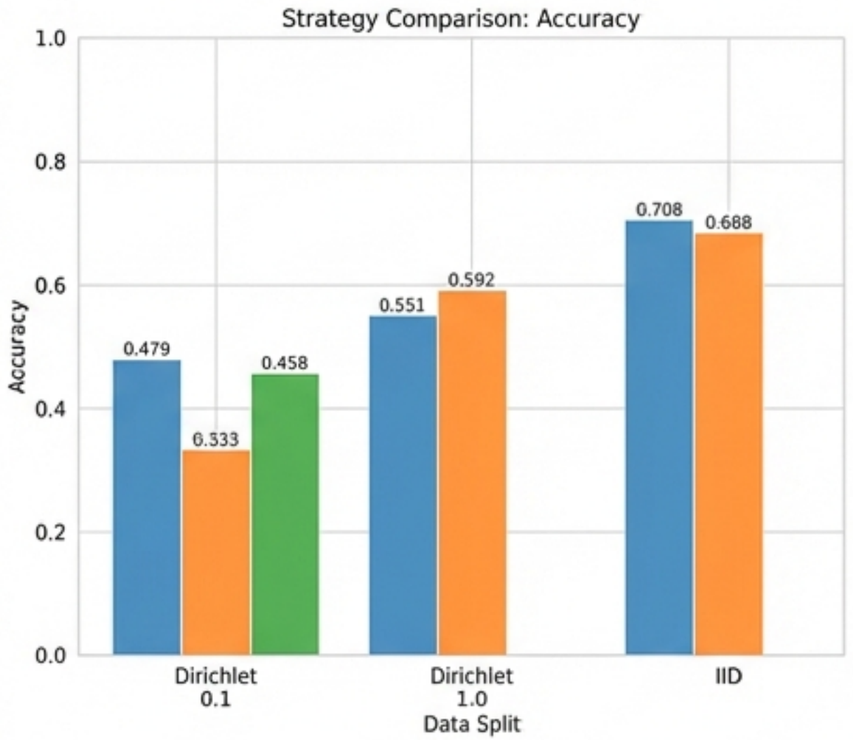


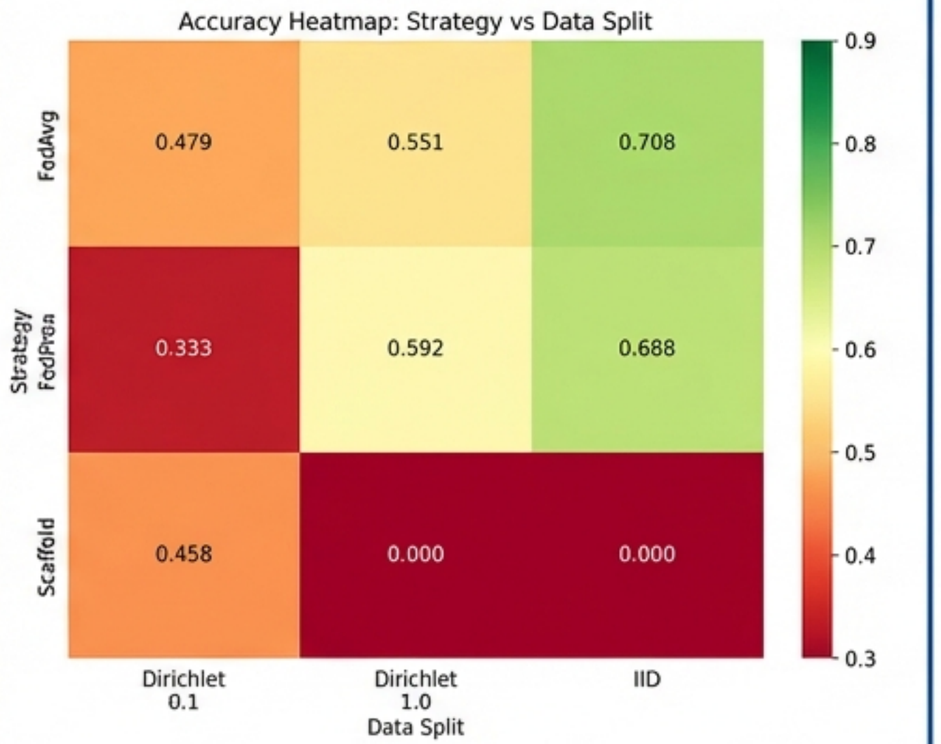


Result: High discrimination performance (AUC ~0.96) proves that genomic hash features retain sufficient biological signal for ancestry inference.

Federated Deep Learning Performance

Strategy Comparison Across Data Splits

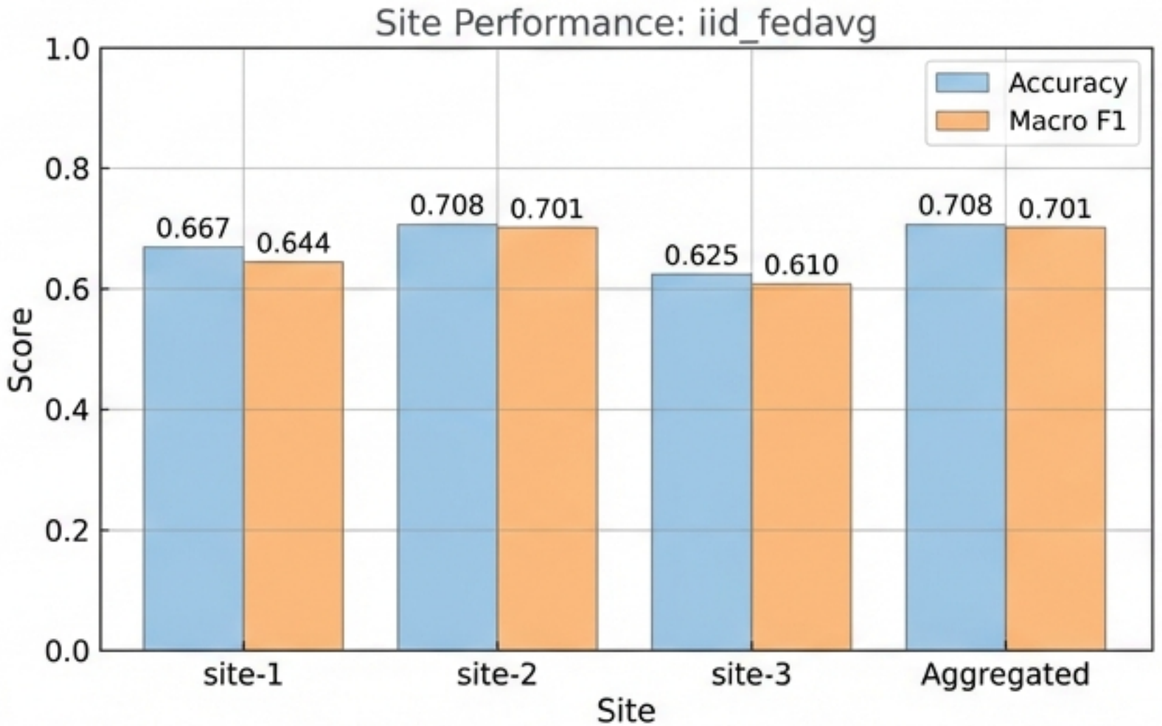




Analysis:

- **IID Setting:** FedAvg performs robustly (Accuracy ~0.70+), approaching centralized baselines.
- **Non-IID Challenge:** Performance drops in high-skew settings (Dirichlet 0.1), highlighting the challenge of 'Client Drift' in genomic federation.

Site-Level Contribution & Conclusion



Final Verdict

Conclusion:

1. **Privacy:** Successfully preserved via Haploblock Clusters.
2. **Utility:** Validated classification signal across sites.
3. **Scalability:** Proven deployment via NVFlare.

Result: A viable path for privacy-preserving collaborative genomic research.