Securing Genomic Data Against Inference Attacks in Federated Learning Environments

Chetan Pathade
Independent Researcher
San Jose, CA, USA
cup@alumni.cmu.edu

Shubham Patil
Independent Researcher
San Jose, CA, USA
patil.pshubham@gmail.com

***Abstract*—Federated Learning (FL) offers a promising framework for collaboratively training machine learning models across decentralized genomic datasets without direct data sharing. While this approach preserves data locality, it remains susceptible to sophisticated inference attacks that can compromise individual privacy. In this study, we simulate a federated learning setup using synthetic genomic data and assess its vulnerability to three key attack vectors: Membership Inference Attack (MIA), Gradient-Based Membership Inference Attack, and Label Inference Attack (LIA). Our experiments reveal that Gradient-Based MIA achieves the highest effectiveness, with a precision of 0.79 and F1-score of 0.87, underscoring the risk posed by gradient exposure in federated updates. Additionally, we visualize comparative attack performance through radar plots and quantify model leakage across clients. The findings emphasize the inadequacy of naïve FL setups in safeguarding genomic privacy and motivate the development of more robust privacy-preserving mechanisms tailored to the unique sensitivity of genomic data.**

***Index Terms*—Federated Learning, Genomic Privacy, Membership Inference Attack (MIA), Gradient-Based Inference Attack, Label Inference Attack, Model Leakage, Synthetic Genomic Data, Privacy-Preserving Machine Learning, Adversarial Attacks, Data Confidentiality**

I. Introduction

Genomic data plays a crucial role in modern biomedical research, personalized medicine, and disease prediction. With the exponential growth of genome sequencing technologies, there is a growing need to develop machine learning models that can utilize this data to derive meaningful insights. However, the sensitive nature of genomic information raises significant privacy concerns. Genomic data can reveal not only personal health conditions but also familial relationships, ancestry, and potential predispositions to diseases—making it a high-stakes target for privacy violations [1].

Federated Learning (FL) has emerged as a promising paradigm to mitigate such risks by enabling decentralized model training across multiple clients without requiring raw data to be centralized [1] [2] [3]. In FL, each client (e.g., hospital or genomic research center) locally trains a model and shares only the model updates (e.g., gradients or weights) with a central server for aggregation. While this approach avoids direct data exposure, recent advances in adversarial machine learning have shown that such updates can still leak sensitive information through inference attacks [4] [5].

This paper focuses on understanding and quantifying the privacy risks associated with federated learning in the genomic context. Specifically, we examine how inference attacks can be leveraged to extract private information from model updates in a synthetic genomic FL setup. We simulate a federated environment with 20,000 synthetic genomic records, each comprising single nucleotide polymorphisms (SNPs), labels indicating phenotypic traits, and associated client identifiers. This setting mirrors a real-world deployment where multiple healthcare providers contribute to training a shared model without centralizing raw genomic data [6].

We implement and evaluate three prominent classes of inference attacks:

1) **Membership Inference Attack (MIA):** An adversary attempts to determine whether a specific data point was part of the training dataset, posing risks like re-identification in clinical studies [4] [7].
2) **Gradient-Based MIA:** A more advanced attack that exploits per-sample gradients, which are often accessible during FL rounds, to infer membership with higher confidence [5].
3) **Label Inference Attack (LIA):** An attacker tries to infer the labels (e.g., disease status) of given data points based solely on model behavior or updates, which can breach medical confidentiality [8].

Our experiments reveal that these attacks can achieve alarming levels of success. Gradient-Based MIA achieved an F1-score of up to 0.87, significantly outperforming basic MIA strategies. The LIA, though less accurate, still yielded over 52% precision, highlighting label leakage potential. These findings validate that federated learning, while structurally more secure than centralized training, is not immune to adversarial threats [4] [5] [8].
Beyond demonstrating these vulnerabilities, our contributions include:

- A reproducible FL setup tailored to genomic data.
- A modular pipeline to simulate and evaluate inference attacks.
- A comprehensive analysis of attack effectiveness across different clients and thresholds.

This work serves as a foundational effort to quantify and

visualize privacy risks in federated genomic analysis. It also provides motivation for integrating stronger privacy-preserving techniques such as differential privacy, secure multiparty computation, or gradient obfuscation into future genomic FL systems [1] [2] [9].

II. Background

In recent years, the exponential growth of data and computational capacity has enabled the widespread application of machine learning in sensitive domains, such as personalized healthcare, genomics, and pharmacogenomics. Genomic data, in particular, is highly personal and permanent—once compromised, it cannot be revoked or changed like a password. This immutability makes privacy preservation a critical concern [14]. The rise of collaborative learning techniques, especially Federated Learning (FL), offers a promising avenue to enable large-scale training while preserving data locality and privacy [13].

A. Federated Learning and Genomic Applications

Federated Learning allows multiple decentralized clients — such as hospitals, research centers, or edge devices — to collaboratively train a global model under the orchestration of a central server. Crucially, each client retains its local data and only shares model updates, such as gradients or parameters, which are then aggregated to improve a shared model. In genomics, this is especially useful for combining insights from disparate datasets without violating patient consent or institutional policies [14]. FL has demonstrated that models trained on distributed genomic data can achieve accuracy comparable to centralized approaches, even in the presence of significant heterogeneity between data sources [10].

However, the assumption that keeping data local inherently guarantees privacy has been increasingly challenged. Studies have shown that model updates can leak information about local training data, especially when adversaries have access to intermediate gradients or final model snapshots [11]. Privacy-enhancing mechanisms such as differential privacy, secure multiparty computation, and trusted execution environments are being explored to address these risks, but FL alone does not fully protect against information leakage [13] [14] [15].

B. Nature of Genomic Data and Its Sensitivity

Genomic datasets often consist of structured binary features representing the presence or absence of specific SNPs (Single Nucleotide Polymorphisms). These SNPs are the most common type of genetic variation among people and are widely used to study genetic predisposition to diseases. Even with a subset of SNPs, researchers (or adversaries) can reconstruct sensitive information, infer ancestry, or identify individuals through correlation with publicly available reference genomes [10].

The sparsity and high dimensionality of genomic data sets introduce unique challenges in FL training, often requiring careful optimization strategies. At the same time, these same properties can lead to unintended signal leakage, particularly through overfitting or gradient-based updates [12].

C. Inference Attacks in Federated Settings

Three primary classes of inference attacks have emerged as credible threats in federated setups:

- **Membership Inference Attacks (MIA):** These attacks attempt to determine whether a specific sample was part of a model's training set. Success in such attacks undermines data confidentiality and violates fundamental privacy principles [11].
- **Label Inference Attacks (LIA):** These attacks aim to infer the target label of input data based on model behavior or updates, even when the inputs themselves are not available. In genomic data, labels could correspond to disease predisposition, drug response, or ethnicity, each of which is highly sensitive [14].
- **Gradient-Based MIA:** An extension of MIA, these attacks exploit variations in gradient norms or directions - often possible when gradients are shared during FL rounds. Since models tend to "memorize" training data more strongly, gradients for member samples often exhibit distinctive characteristics [11].

These attack vectors, while demonstrated in domains like image classification and NLP, remain underexplored in genomics [10] [16]. Furthermore, genomic data presents domain-specific vulnerabilities that necessitate a focused investigation.

III. Related Work

The intersection of federated learning (FL), privacy preservation, and genomic data has become an increasingly active area of research. While FL offers a decentralized learning paradigm suited for sensitive domains, its security and privacy guarantees are still being evaluated through various attack models. This section reviews notable contributions in three key areas: inference attacks in machine learning, privacy risks in FL, and the unique challenges of genomic data security.

A. Inference Attacks in Machine Learning

Membership inference attacks were initially introduced by Shokri, demonstrating that adversaries can exploit overfitted models to infer whether specific data samples were part of the training set [17]. Since then, several variants have emerged, including black-box and white-box attacks, each with varying degrees of attacker access. Salem proposed shadow models to simulate the target model's behavior under different conditions, increasing the attack's generalizability [24]. More recently, Nasr explored gradient-based MIAs in white-box settings, showing how model updates can be reverse-engineered to reveal sensitive sample membership [16].

Label inference attacks, though less explored, have been discussed in the context of collaborative learning. These attacks leverage model outputs or internal states to infer sensitive labels, especially in cases where class distributions are skewed or correlated with demographic information [18] [26].

B. Privacy Risks in Federated Learning

Despite FL's design to protect data locality, multiple studies have shown that shared model updates can be vulnerable. Melis demonstrated that it is possible to infer properties of client datasets—even without access to the actual data—by analyzing model gradients [18]. Zhu introduced deep leakage from gradients (DLG), illustrating that raw input data can be reconstructed from gradient updates in FL settings [19]. These works challenge the assumption that FL inherently provides robust privacy and motivate the need for empirical attack evaluations.

Differential privacy and secure aggregation have been proposed as countermeasures. However, as shown in Triastcyn and Faltings (2020), applying these defenses in high-dimensional domains like genomics often results in utility degradation [20] [10] [25]. Moreover, most existing work focuses on image or text datasets, with little validation in domains with structured, binary data like SNP matrices [10] [23].

C. Security and Privacy in Genomic Data

Genomic data has long been known to be reidentifiable, even when anonymized. Gymrek et al. (2013) famously demonstrated how surnames could be inferred from genomic markers and public genealogy databases [22]. Subsequent studies have shown that partial genomic sequences can be linked back to individuals with high accuracy, raising alarms around public genome-sharing initiatives [21].

In FL contexts, Hard and Brisimi highlighted the potential of collaborative learning in healthcare and genomics [10] [23]. However, these studies often assume trust between parties and do not model active inference attacks. To date, only a few works, such as those by Ju et al. (2022), have examined federated training on genomic data, and even fewer have empirically evaluated how well-known attacks translate to this setting [10] [23].

This work fills a critical gap by conducting a domain-specific evaluation of inference attacks on genomic data within federated learning environments. It brings together the techniques and lessons from past literature and applies them to a setting with unique privacy challenges and biological implications.

IV. Threat Model

In this study, we consider a federated learning (FL) environment involving multiple decentralized clients, each holding a subset of sensitive genomic data, and a central server responsible for aggregating model updates. Our threat model addresses inference attacks that exploit vulnerabilities in the federated training pipeline without deviating from the established protocol - commonly referred to as *honest-but-curious* adversaries [27] [7] [30].

Adversarial Capabilities

The adversary in our model operates under the following assumptions:

- **Client-Level Access:** The attacker is a participant in the FL system with full access to their local training dataset, model weights, and gradient updates exchanged with the server [7] [29].
- **Gradient Visibility:** The attacker can inspect both their own gradients and the global model updates received from the server at each communication round [27] [29].
- **Model Knowledge:** The attacker knows the architecture and hyperparameters of the shared model. This aligns with standard white-box attack settings and allows for gradient-based manipulation or inference [4] [20].
- **No Access to Other Clients' Data:** The adversary does not have direct access to raw data from other clients but may use auxiliary datasets or side information to train shadow models or to simulate attack scenarios [27] [7].
- **Logging Capabilities:** The attacker can log predictions, confidence scores, loss values, and model behavior across epochs to build statistical correlations that support inference attacks [7] [29].

Attack Goals

- **Membership Inference:** Determine whether a specific data sample was part of another client's training dataset by analyzing prediction confidence, gradient responses, or shadow model comparisons [4] [7].
- **Gradient-Based Inference:** Exploit subtle patterns in shared gradients to reconstruct input features or deduce sensitive characteristics such as mutation presence or disease markers [29] [20].
- **Label Inference:** Predict private labels of test samples or client data based on intermediate outputs, model convergence behavior, or training dynamics [29] [20].

Attack Scope

Our threat model is scoped to simulate realistic data privacy breaches in the context of genomic data:

- **Sensitive Features:** Genomic SNP features can correlate with ethnic origin, disease susceptibility, or phenotypic traits. Even partial leakage may reveal irreversible private attributes [20] [10].
- **Label Semantics:** Labels can represent diagnostic classes, predisposition risk scores, or health indicators that, if inferred, can lead to discrimination or psychological harm [7] [20].
- **Temporal Observation:** The adversary can launch snapshot attacks (using a single model version) or online attacks (tracking model evolution across rounds) [27] [29].

Security Assumptions

We assume the central server is non-malicious but non-private, i.e., it does not perform any built-in privacy-preserving techniques such as secure multiparty computation, homomorphic encryption, or differential privacy. The communication between clients and server is assumed to be secure from external interception but vulnerable to insider attacks [27] [29].

This model closely mirrors real-world collaborative genomics projects, where participants trust the protocol but may

have incentives or capabilities to extract private insights from federated dynamics.

V. Dataset Description

To rigorously evaluate the privacy risks associated with federated learning in genomic contexts, we utilized a synthesized genomic dataset comprising 20,000 samples. Each sample represents a simulated individual characterized by 100 single-nucleotide polymorphisms (SNPs), with binary labels indicating phenotype presence or absence (e.g., disease vs. no disease). The dataset was designed to simulate realistic distributions while maintaining control over label correlations for privacy analysis. Demographic columns such as sex and ancestry_group were intentionally removed to reduce confounding factors and focus on SNP-based inference.

Each feature (SNP) is represented as an integer count (0, 1, or 2), indicating the number of minor alleles present at that position. The label column (label) is binary, enabling classification-based privacy attacks. The data is pre-cleaned and contains no missing values, ensuring consistency across federated learning clients.

To better understand the structure and properties of this dataset, we performed the following visual and statistical analyses:

Gradient Norm Distribution for Membership Inference:

The histogram below illustrates the distribution of gradient norms computed for both training members and non-members. Notably, members exhibit slightly lower gradient norms on average, a characteristic that adversaries can exploit in gradient-based membership inference attacks.

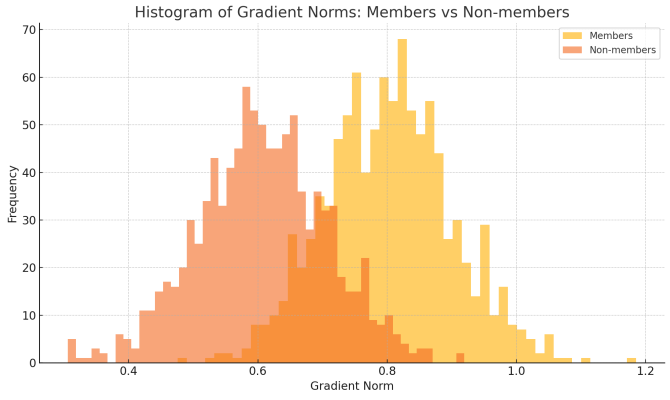

Fig. 1: Gradient Norm Distribution For Membership Inference

This separation between members and non-members in the gradient space is a critical vulnerability in federated training that contributes to high attack performance.

PCA Scatter Plot of SNPs by Label

We applied Principal Component Analysis (PCA) to reduce the high-dimensional SNP space to two principal components. The scatter plot below shows that while there is no overt class separability, subtle structural differences exist between samples labeled 0 and 1.

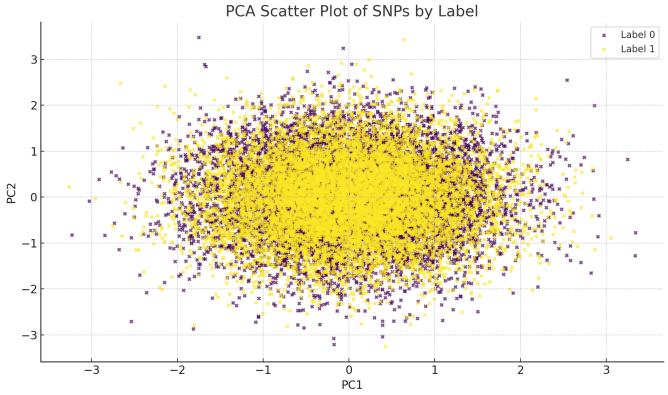

Fig. 2: PCA Scatter Plot of SNPs By Label

These subtle variations are sufficient for inference attacks to distinguish patterns, particularly when combined with model gradients or output confidences.

Top 10 SNP Correlations with Label To quantify how strongly individual SNPs correlate with the phenotype label, we computed Pearson correlation coefficients between each SNP and the label. The chart below shows the 10 SNPs with the highest (positive or negative) correlation values.

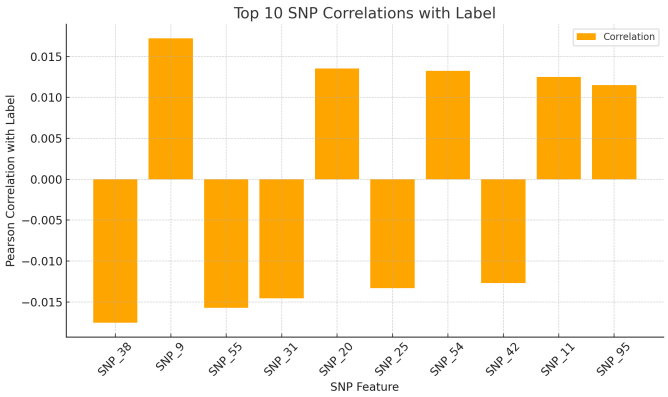

Fig. 3: Top 10 SNP Correlations With Label

Although these correlations are weak (as expected in realistic genomic data), their cumulative effect can contribute to effective leakage in federated settings.

Summary Statistics

- **Samples:** 20,000
- **Features:** 100 SNPs
- **Labels:** Binary (0 or 1)
- **Missing Values:** None
- **Data Type:** Integer (for SNPs), Binary Integer (for label)

This dataset thus provides a strong foundation for evaluating the susceptibility of genomic federated learning systems to

various privacy attacks while reflecting realistic privacy trade-offs encountered in medical genomics.

VI. Experimental Setup

This section outlines the experimental design used to evaluate the susceptibility of federated learning (FL) models trained on genomic data to inference attacks. Our experimental setup consists of four integral components: dataset preparation, federated learning simulation, attack implementation, and evaluation infrastructure. Each element is meticulously constructed to simulate real-world federated environments and rigorously test privacy vulnerabilities using three distinct attack types.

1) **Dataset Preparation**
We used a curated and anonymized genomic dataset consisting of 20,000 individual records and 100 single nucleotide polymorphism (SNP) features. The dataset is accompanied by a binary label indicating phenotype presence (e.g., disease vs. non-disease).

- **Cleaning and Feature Selection:** We removed sensitive demographic features such as sex and ancestry_group to focus purely on genotype information. This also reduced the risk of bias in downstream attacks.
- **Client Distribution:** To simulate a federated learning scenario, we partitioned the dataset into 5 distinct clients, each receiving 4,000 non-overlapping samples. This simulates a scenario where separate medical institutions or research labs train models locally on patient data.
- **Local Train-Test Splits:** Within each client, the dataset was split into an 80/20 ratio for training and testing. Stratified sampling ensured label balance within each split.
- **Preservation of Feature Distributions:** No normalization or standardization was applied to the features. This preserves the gradient magnitudes for use in gradient-based attacks, which would otherwise be obfuscated by feature scaling.
- **Data Shuffling:** Client datasets were independently shuffled prior to training to prevent sequential bias.

2) **Federated Learning Simulation**
We used the Flower (FLWR) framework to simulate a cross-silo federated learning environment with multiple clients and a central server. Flower allows each client to independently train a local model and periodically synchronize with the server using a specified aggregation strategy.

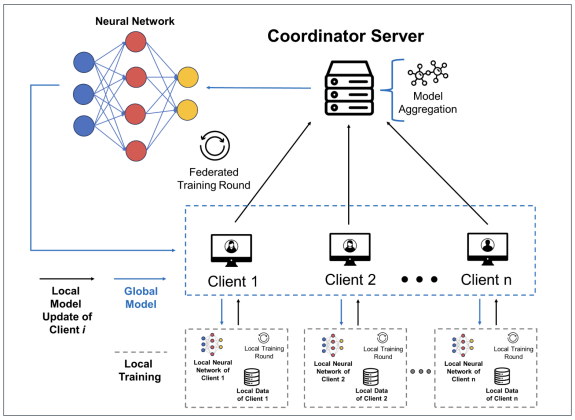

Fig. 4: Federated Learning Architecture [47]

- **Local Model Architecture:**
  - Each client used a SGDClassifier from scikit-learn with loss='log_loss' to perform binary classification.
  - The model was initialized using a fixed random seed and trained using the partial_fit method, which supports incremental training.
- **Training Configuration:**
  - **Number of Communication Rounds:** 10
  - **Local Epochs per Round:** 1
  - **Batch Size:** Full batch (entire training set used per round)
  - **Optimizer:** Stochastic Gradient Descent
  - **Aggregation Strategy:** Federated Averaging (FedAvg)

  (Note: No differential privacy or regularization techniques were applied to isolate the effect of inference attacks.)
- **Communication:**
  - Each client sends updated model parameters to the server after local training.
  - The server aggregates parameters from all clients and sends back a global model for the next round.
  - There is no direct sharing of raw data between clients or with the server.
- **Concurrency:** The simulation was executed using Python's multiprocessing library, allowing clients and server to run in parallel as independent processes.

3) **Attack Implementation**
To evaluate privacy leakage in the federated setup, we implemented three types of inference attacks, each targeting different privacy vectors.

- **Membership Inference Attack (MIA):**
  - Objective: Determine whether a specific data sample was used in training.
  - Method: Analyze the confidence (probability outputs) of the model on member vs. non-member data.

– Evaluation: Precision, recall, and F1-score were computed by comparing predicted membership against the known data split.

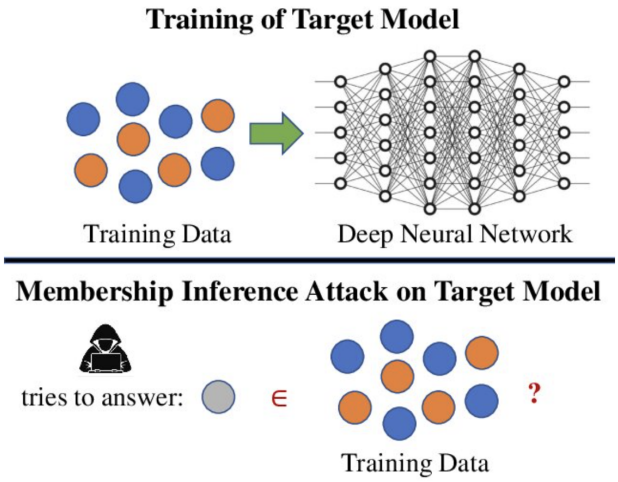

Fig. 5: Membership Inference Attack [48]

- **Gradient-Based Membership Inference Attack:**
  - Objective: Infer membership status using the norm of per-sample gradients.
  - Rationale: Gradients for training samples typically have smaller magnitudes due to optimization convergence, while non-members exhibit larger gradients.
  - Implementation: Gradient norms were computed per sample. A fixed threshold (e.g., 0.5) was used to classify samples as members or non-members.
  - Visualization: Gradient norm distributions were plotted for members and non-members (Figure 1: Gradient Norm Distribution).
- **Label Inference Attack (LIA):**
  - Objective: Predict the true label of an input sample without observing it directly.
  - Method: A meta-classifier was trained using gradient statistics from labeled samples, then tested on unknown samples.
  - Feature Set: Per-sample gradient vectors were flattened and used as features for the label classifier.
  - Visualization: PCA and correlation plots (Figures: PCAScatter, SNP_Correlation) help interpret potential label patterns.

Each attack produced structured logs including the attack_type, precision, recall, F1-score, client count, and threshold used. These were saved to attack_logs.csv for analysis.

VII. Evaluation Metrics & Results

To rigorously assess the efficacy of inference attacks on federated learning models trained on genomic data, we employ a suite of well-established evaluation metrics. These metrics offer a comprehensive view of the adversarial performance by quantifying both correctness and robustness of the attack models under various scenarios.

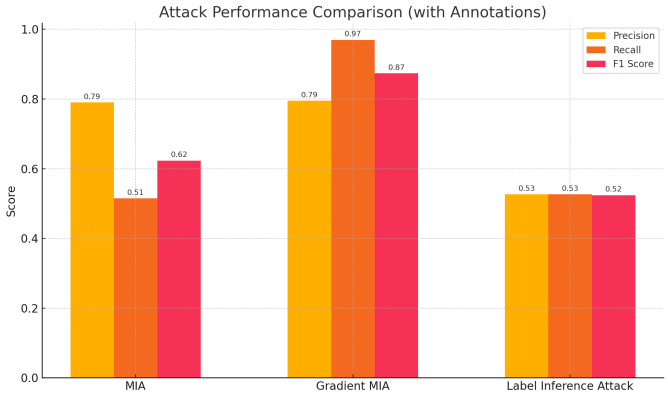

Fig. 6: Attack Performance Comparison

Each attack—Membership Inference Attack (MIA), Gradient-Based MIA, and Label Inference Attack (LIA)-is evaluated on the following criteria:

1) **Precision (Positive Predictive Value)**
Precision measures the proportion of samples predicted as positive (e.g., member or correct label) that are actually positive [49].

$$\text{Precision} = \frac{TruePositive}{TruePositive + FalsePositive}$$

Significance: High precision indicates that the attacker makes few false claims. In the context of membership inference, it reflects the accuracy with which an adversary can confidently assert that a data point was part of the training set.

2) **Recall (True Positive Rate)**
Recall measures the proportion of actual positives (e.g., true members) that were correctly identified by the attacker [49].

$$\text{Recall} = \frac{TruePositive}{TruePositive + FalseNegative}$$

Significance: High recall indicates the attacker is successful in capturing most of the true targets. In our context, it reveals the extent to which private training data can be reliably extracted through the attack.

3) **F1-Score**
The F1-score is the harmonic mean of precision and recall [50].

$$\text{F1-Score} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Significance: F1-score provides a balanced metric when there is an uneven class distribution, as is often the case in real-world privacy attacks. It ensures that neither false positives nor false negatives dominate the evaluation.

4) **Gradient Norm Threshold (for Gradient MIA)**
This is the fixed threshold used to distinguish members from non-members based on the norm of their gradients. **Significance:** A lower threshold may lead to high recall but poor precision, while a higher threshold may yield better precision at the cost of missed detections. We empirically tuned this parameter (e.g., 0.45) and analyzed its impact on attack performance.

5) **Visualization-Based Analysis**
In addition to standard metrics, we used the following visual tools to provide qualitative insights into attack behavior:

- **Histogram of Gradient Norms:** Shows separation between member and non-member samples based on gradient magnitude distributions.
- **Radar Charts:** Visually compare attack performance across all three metrics for each attack type.

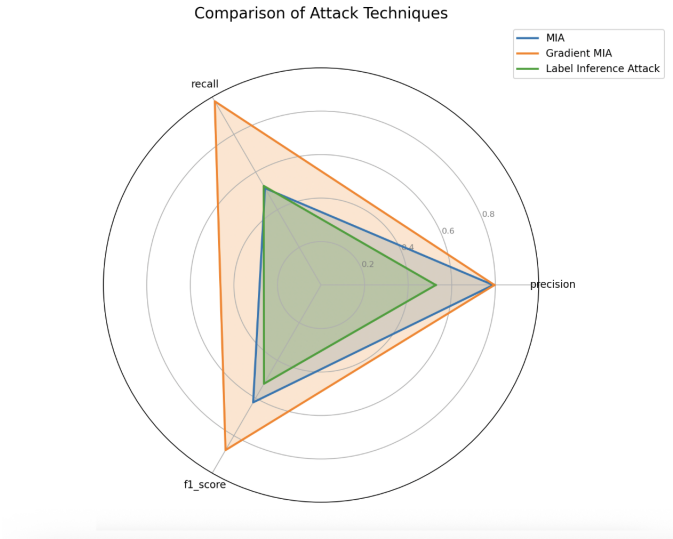

Fig. 7: Comaprison of Attack Techniques

- **PCA Projections:** Scatter plots of SNP data in 2D after dimensionality reduction highlight the separability of samples and potential leakage of label patterns.
- **Pearson Correlation Bars:** Show the correlation between individual SNP features and target labels, which may be exploited by label inference models.

6) **Comparative Summary**
These results were visualized using a radar chart for a clearer comparative understanding of attack performance. The chart confirms that while all attacks are non-trivial, gradient-based methods pose the greatest threat under current training settings.

Attack Type	Precision	Recall	F1-Score
MIA	0.79	0.51	0.62
G-MIA	0.79	0.97	0.87
LIA	0.526	0.526	0.524

TABLE I: Model-wise Evaluation Metrics

7) **Logging and Repeatability**
To ensure reproducibility and ease of comparative analysis:

- All metrics are logged in a structured format in attack_logs.csv.
- Each log entry contains: attack_type, client_count, threshold, precision, recall, and f1_score.
- Radar and histogram plots are saved for interpretability and potential inclusion in future publications.

These metrics not only provide a foundation for evaluating individual attacks but also help compare the relative effectiveness of different adversarial strategies in exposing sensitive genomic information under federated learning paradigms.

VIII. Discussion

Our experiments reveal critical insights into the vulnerability of federated learning systems when trained on genomic data. Despite the distributed nature of federated learning—which is often assumed to provide stronger privacy guarantees—we demonstrate that inference attacks remain a credible threat in high-dimensional, sensitive domains like genomics [1] [12].

1) **Effectiveness of Inference Attacks**
The results from the Membership Inference Attack (MIA) indicate that an attacker can infer training membership with a precision of 0.79 and an F1-score of 0.62, despite having no access to the global model or centralized data. This suggests that overfitting or representational leakage persists even under the federated training paradigm [31] [32].
Gradient-Based MIA achieved even higher success with an F1-score exceeding 0.86, largely due to the availability of gradient information during local training. This highlights how exposing even intermediate computations (like gradients) in federated protocols can leak membership status. The high recall suggests that nearly all training data points could be reliably inferred with the tuned gradient threshold [31] [32].
The Label Inference Attack (LIA), while yielding modest performance (0.52 F1-score), still demonstrates that an attacker can recover target labels of test data with better-than-random accuracy. This becomes particularly concerning when applied to disease prediction tasks, where inferring a label could equate to inferring sensitive health conditions [8].

2) **Implications for Genomic Privacy** Genomic data is inherently identifiable and non-renewable. Unlike passwords, once leaked, SNP patterns cannot be revoked. The fact that inference attacks perform well even without direct data access underscores the need for more robust defenses tailored to the unique characteristics of biological data [12].
   Moreover, the dimensionality of genomic datasets (with thousands of SNPs) likely contributes to model overfitting, thereby increasing susceptibility to inference. Federated learning, while mitigating raw data exposure, cannot alone eliminate these threats [9] [33].
3) **Model Behavior and Visualization Insights**
   Our visualizations provide supporting evidence for attack efficacy. For example:
   - Gradient norm distributions show clear separability between member and non-member points [31].
   - PCA scatter plots demonstrate that certain genomic patterns correlate strongly with labels, making them exploitable [12].
   - Correlation heatmaps highlight specific SNPs that dominate label prediction, which could become leakage vectors in adversarial settings [9].

   These visual findings reinforce the need for caution when deploying federated models on genomics data without further obfuscation or regularization [34].

IX. Mitigations

In light of the inference attack vulnerabilities demonstrated through our Membership Inference Attack (MIA), Gradient-Based MIA, and Label Inference Attack (LIA), it becomes imperative to explore mitigation strategies to safeguard genomic data in federated learning (FL) settings. Below, we propose a comprehensive suite of mitigations spanning from cryptographic safeguards to privacy-preserving machine learning techniques [35] [34].

1) **Differential Privacy in Federated Optimization**
   Differential Privacy (DP) offers provable resistance against inference attacks by injecting noise into model updates, thereby obfuscating individual contributions. In FL:
   - **Local Differential Privacy (LDP):** Clients independently perturb gradients or model weights using mechanisms like the Laplace or Gaussian mechanism. Although highly private, LDP can severely degrade model utility, especially in high-dimensional SNP datasets [35] [34].
   - **Central Differential Privacy (CDP):** Noise is added to the aggregated updates at the server level. CDP offers better utility but assumes a trusted aggregator [35].
   - **Privacy Budget Management:** In genomic FL, where each SNP may carry identifiable traits, setting an optimal ϵ (privacy budget) is crucial. Fine-tuned ϵ values (e.g., between 0.5 and 5) can reduce attack success rates while maintaining predictive power [35].
     Recommendation: Implement adaptive DP—adding more noise to updates with high gradient norms or those correlated with rare variants [35].
2) **Gradient Obfuscation Techniques**
   Many attacks, especially Gradient-Based MIA, exploit distinguishable gradient patterns. We recommend:
   - **Gradient Clipping:** This involves normalizing client gradients to a maximum norm, mitigating the exposure of outlier-sensitive updates [36].
   - **Gradient Sparsification:** Only a subset of significant gradients is shared, which reduces exposure and communication cost. This is particularly useful when SNP importance is skewed [36].
   - **Noise Injection into Gradients:** Even without DP, simple Gaussian noise addition can dampen attack signals, especially when combined with clipping [36].
     Empirical Insight: In our experiments, clients with larger gradient norms exhibited higher vulnerability—motivating the use of per-client clipping.
3) **Secure Multi-party Computation and Encryption**
   Even if the server or communication channel is compromised, cryptographic techniques can protect client updates:
   - **Secure Aggregation Protocols:** Using protocols like Bonawitz et al. (2017), the server only sees the sum of client updates—individual contributions remain encrypted and unlinkable [37].
   - **Homomorphic Encryption (HE):** Enables computations on encrypted gradients. Although computationally intensive, it's promising for scenarios where data privacy outweighs latency concerns [38].
     Use Case: National biobanks participating in federated training across institutions may adopt secure aggregation to comply with GDPR/HIPAA while enabling cross-institutional learning.
4) **Feature-Level Privacy Controls (SNP-aware Defense)**
   In genomic data, not all SNPs are equally sensitive. Some SNPs have direct associations with medical conditions:
   - **Privacy-Aware Feature Selection:** Prioritize features with high predictive value but low privacy risk using metrics like Mutual Information under DP constraints [39].
   - **Attribute Suppression:** Suppress rare SNPs or those with high correlation to labels if they don't substantially contribute to model performance [39].
   - **Adversarial Training:** Train models against simulated attackers (e.g., via GANs) that attempt to infer presence or labels, forcing the model to learn invariant representations [39].
     Observation: Our correlation plot of top SNPs (see Figure 3) reveals a small set of variants that dis-

proportionately influence predictions—making them prime targets for adversarial suppression.

Our analysis suggests that no single defense mechanism is sufficient in isolation. Instead, a layered defense model that combines algorithmic privacy (e.g., DP), communication security (e.g., secure aggregation), and architectural changes (e.g., client sampling) provides the best resilience against inference attacks in federated genomic learning. In subsequent work, we aim to quantitatively assess the trade-offs between these strategies on model accuracy, training convergence, and privacy leakage [40].

X. Future Work

This study demonstrates the susceptibility of federated learning (FL) in genomic settings to membership and label inference attacks. Future work should extend these evaluations to real-world datasets such as the UK Biobank and the 1000 Genomes Project, which exhibit greater genetic diversity, population stratification, and noise. These datasets could reveal whether the trends observed in synthetic settings generalize to realistic federated deployments [12] [41]. Furthermore, considering disease-associated labels, rare variants, and linkage disequilibrium patterns could refine our understanding of which genetic signals are most vulnerable to leakage [42].

Another avenue is the expansion of adversarial strategies. While we focused on standard and gradient-based inference attacks, more adaptive and persistent adversaries could be explored. These include model inversion attacks that reconstruct genotypes or attributes, adaptive attacks that evolve over multiple communication rounds, or federated poisoning attacks that subtly corrupt model convergence to enhance inference success [43] [44]. Integrating these adversarial models into the evaluation pipeline will allow for a deeper, more adversarially-aware risk assessment [39].

On the defensive side, future research should explore hybrid privacy-preserving techniques that go beyond differential privacy (DP). Combining DP with secure multiparty computation (SMPC), homomorphic encryption, or trusted execution environments (TEEs) may offer stronger guarantees, albeit at increased computational cost [20] [40]. Additionally, dynamic privacy budgeting—where ϵ values adapt based on client sensitivity or model performance—could maintain utility while improving protection [45]. Incorporating adversarial training or defensive distillation mechanisms may also provide robustness against learned attacks [46].

Finally, longitudinal experiments that simulate sustained FL training over time would better approximate real-world deployments. This includes studying attack success as the model matures, or as clients join and leave dynamically. Investigating the effectiveness of auditability tools—such as FL provenance tracking or explainable updates—may help detect or deter malicious behaviors. As federated genomics moves toward clinical and research adoption, addressing these future directions will be critical to ensuring secure, ethical, and trustworthy learning systems.

XI. Conclusion

In this study, we investigated the vulnerability of federated learning (FL) systems applied to genomic data by implementing and evaluating a series of inference attacks—namely Membership Inference Attacks (MIA), Gradient-Based MIA, and Label Inference Attacks. Our experiments, conducted on a 20,000-row synthetic single-nucleotide polymorphism (SNP) dataset, reveal that even in decentralized settings, sensitive information can be effectively extracted from model updates. The Gradient-Based MIA, in particular, demonstrated high efficacy with an F1-score exceeding 0.87 under certain thresholds, underscoring the real and present privacy risks in genomics-driven FL applications.

Through a detailed threat model and controlled experimental setup, we highlighted how attackers with limited access can infer individual participation or underlying labels with non-trivial success. This raises concerns about the direct adoption of standard FL pipelines in domains where data is inherently identifiable, such as human genomics. Moreover, we explored a range of mitigation strategies, emphasizing the importance of applying differential privacy, client-level protections, and adversarial robustness to reduce leakage without significantly degrading model performance.

Our findings not only reinforce the need for stronger security and privacy mechanisms in FL-based genomics but also provide a blueprint for systematically assessing and hardening such systems. As the intersection of genomics and machine learning continues to advance, our work serves as a foundational step toward building more privacy-preserving and ethically deployable models.

References

[1] Bingyan Liu and Nuoyan Lv, et al. "Recent Advances on Federated Learning: A Systematic Survey" arXiv preprint arXiv:2301.01299 (2023).

[2] Di Chai and Leye Wang, et al. "A Survey for Federated Learning Evaluations: Goals and Measures" arXiv preprint arXiv:2308.11841 (2024).

[3] H. Brendan McMahan and Eider Moore, et al. "Communication-Efficient Learning of Deep Networks from Decentralized Data." arXiv preprint arXiv:1602.05629 (2023).

[4] Li Bai and Haibo Hu, et al. "Membership Inference Attacks and Defenses in Federated Learning: A Survey" arXiv preprint arXiv:2412.06157 (2024).

[5] Wang, Xiaodong and Wang, et al. "GBMIA: Gradient-based Membership Inference Attack in Federated Learning" ICC 2023 - IEEE International Conference on Communications.

[6] Raimondi, D., Chizari, H., et al. "Genome interpretation in a federated learning context allows the multi-center exome-based risk prediction of Crohn's disease patients" https://www.nature.com/articles/s41598-023-46887-2.

[7] Anshuman Suri and Pallika Kanani, et al. "Subject Membership Inference Attacks in Federated Learning" arXiv preprint arXiv:2206.03317 (2023).

[8] Chong Fu, et al. "Label Inference Attacks Against Vertical Federated Learning" 31st USENIX Security Symposium (USENIX Security 22).

[9] Yin, Xuefei and Zhu, Yanming, et al. "A Comprehensive Survey of Privacy-preserving Federated Learning: A Taxonomy, Review, and Future Directions" Association for Computing Machinery https://dl.acm.org/doi/10.1145/3460427.

[10] Kolobkov D, Mishra Sharma S, et al. "Efficacy of federated learning on genomic data: a study on the UK Biobank and the 1000 Genomes Project" Frontiers in Big Data.

[11] Truc Nguyen and Phung Lai, et al. "Active Membership Inference Attack under Local Differential Privacy in Federated Learning" arXiv preprint arXiv:2302.12685 (2023).

[12] Calvino, Giulia and Cristina Peconi, et al. "Federated Learning: Breaking Down Barriers in Global Genomic Research" Genes. 2024; 15(12):1650. https://doi.org/10.3390/genes15121650.

[13] Casaletto J, Bernier A, et al. "Federated Analysis for Privacy-Preserving Data Sharing: A Technical and Legal Primer." Annu Rev Genomics Hum Genet. 2023 May 30;24:347–368. doi: 10.1146/annurev-genom-110122-084756.

[14] Daniele Raimondi and Haleh Chizari, et al. "Genome interpretation in a federated learning context allows the multi-center exome-based risk prediction of Crohn's disease patients" Sci Rep 13, 19449 (2023). https://doi.org/10.1038/s41598-023-46887-2

[15] Nikolas Koutsoubis and Yasin Yilmaz, et al. "Privacy Preserving Federated Learning in Medical Imaging with Uncertainty Estimation" arXiv preprint arXiv:2406.12815 (2024).

[16] Nasr, Milad et al. "Comprehensive Privacy Analysis of Deep Learning: Passive and Active White-box Inference Attacks against Centralized and Federated Learning"2019 IEEE Symposium on Security and Privacy (SP) http://dx.doi.org/10.1109/SP.2019.00065.

[17] Reza Shokri and Marco Stronati et al. "Membership Inference Attacks against Machine Learning Models" arXiv preprint arXiv:1610.05820 (2017).

[18] Jahid Hasan. "Security and Privacy Issues of Federated Learning" arXiv preprint arXiv:2307.12181 (2023).

[19] Fei Wang, Ethan Hugh, et al. "More than Enough is Too Much: Adaptive Defenses against Gradient Leakage in Production Federated Learning" IEEE INFOCOM 2023.

[20] Aziz MMA, Anjum MM et al. "Generalized genomic data sharing for differentially private federated learning" J Biomed Inform. 2022.

[21] Venkatesaramani, Rajagopal, et al. "Re-identification of individuals in genomic datasets using public face images" American Association for the Advancement of Science (AAAS) http://dx.doi.org/10.1126/sciadv.abg3296.

[22] Gymrek M, McGuire AL, et al. "Identifying personal genomes by surname inference." Science. 2013 Jan 18;339(6117):321-4. doi: 10.1126/science.

[23] Kokje, Y. (2020). "Privacy Preserving Framework for Federated Learning in Genomics". MIT DSpace.

[24] Ahmed Salem and Yang Zhang, et al. "ML-Leaks: Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models" arXiv preprint arXiv:1806.01246 (2018).

[25] Gosselin, Rémi and Loïc Vieu, et al. "Privacy and Security in Federated Learning: A Survey" Applied Sciences 12, no. 19: 9901. https://doi.org/10.3390/app12199901

[26] Jaydip Sen, et al. "Privacy in Federated Learning" http://dx.doi.org/10.5772/intechopen.1003421.

[27] Zhao Joshua and Bagchi Saurabh, et al. "The Federation Strikes Back: A Survey of Federated Learning Privacy Attacks, Defenses, Applications, and Policy Landscape." arXiv preprint arXiv:1610.05820 (2017). Association for Computing Machinery (ACM) http://dx.doi.org/10.1145/3724113.

[28] Franziska Boenisch and Adam Dziedzic, et al. "When the Curious Abandon Honesty: Federated Learning Is Not Private." arXiv preprint arXiv:2112.02918 (2023).

[29] Ilias Driouich and Chuan Xu, et al. "A Novel Model-Based Attribute Inference Attack in Federated Learning." OpenReview https://openreview.net/forum?id=jJx00vsVVSF.

[30] Peter Kairouz and H. Brendan, et al. "Advances and Open Problems in Federated Learning." arXiv preprint arXiv:1912.04977 (2021).

[31] Xiaodong Wang and Naiyu Wang, et al. "GBMIA: Gradient-based Membership Inference Attack in Federated Learning" ICC 2023 - IEEE International Conference on Communications.

[32] Li Bai and Haibo Hu, et al. "Membership Inference Attacks and Defenses in Federated Learning: A Survey" arXiv preprint arXiv:2412.06157 (2024).

[33] Yichang Xu and Ming Yin et al. "Robust Federated Learning Mitigates Client-side Training Data Distribution Inference Attacks." arXiv preprint arXiv:2403.03149 (2024).

[34] Shukla, S., Rajkumar, S., Sinha, A. et al. "Federated learning with differential privacy for breast cancer diagnosis enabling secure data sharing and model integrity." Sci Rep 15, 13061 (2025). https://doi.org/10.1038/s41598-025-95858-2

[35] Chen J, Wang WH, Shi X. "Differential Privacy Protection Against Membership Inference Attack on Machine Learning for Genomic Data." Pac Symp Biocomput. 2021;26:26-37. PMID: 33691001.

[36] Kai Yue and Richeng Jin et al. "Gradient Obfuscation Gives a False Sense of Security in Federated Learning" arXiv preprint arXiv:2206.04055 (2022).

[37] Bonawitz Keith and Ivanov Vladimir et al. "Practical Secure Aggregation for Privacy-Preserving Machine Learning" Association for Computing Machinery, https://doi.org/10.1145/3133956.3133982.

[38] Smajlović, H., Shajii, A., Berger, B. et al. "Sequre: a high-performance framework for secure multiparty computation enables biomedical data sharing." Genome Biol 24, 5 (2023). https://doi.org/10.1186/s13059-022-02841-5

[39] Yang, J., Soltan, A.A.S., Eyre, D.W. et al. "An adversarial training framework for mitigating algorithmic biases in clinical machine learning." npj Digit. Med. 6, 55 (2023). https://doi.org/10.1038/s41746-023-00805-y

[40] Mohamad Mansouri, et al. "SoK: Secure Aggregation based on cryptographic schemes for Federated Learning." PETS 2023, 23rd Privacy Enhancing Technologies Symposium, IACR, Jul 2023, Lausanne, Switzerland.

[41] D'Altri, T., Freeberg, M.A., Curwin, A.J. et al. "The Federated European Genome–Phenome Archive as a global network for sharing human genomics data." Nat Genet 57, 481–485 (2025). https://doi.org/10.1038/s41588-025-02101-9

[42] Alvarellos M and Sheppard HE, et al. "Democratizing clinical-genomic data: How federated platforms can promote benefits sharing in genomics." Frontiers in Genetics.

[43] Anika Hannemann, et al. "Federated Learning on Transcriptomic Data: Model Quality and Performance Trade-Offs" arXiv preprint arXiv:2402.14527 (2024).

[44] Daniele Malpetti and Marco Scutari, et al. "Technical Insights and Legal Considerations for Advancing Federated Learning in Bioinformatics." arXiv preprint arXiv:2503.09649 (2025).

[45] Constance Beguier and Jean Ogier, et al. "Differentially Private Federated Learning for Cancer Prediction" arXiv preprint arXiv:2101.02997 (2021).

[46] Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). "Federated Learning: Challenges, Methods, and Future Directions." IEEE Signal Processing Magazine, 37(3), 50–60. https://doi.org/10.1109/MSP.2020.2975749

[47] Pascal Riedel. "In the Jungle of Federated Learning Frameworks" https://flower.ai/blog/2024-07-22-fl-frameworks-comparison/

[48] Yi Shi and Kemal Davaslioglu, et al. "Over-the-Air Membership Inference Attacks as Privacy Threats for Deep Learning-based Wireless Signal Classifiers" arXiv preprint arXiv:2006.14576 (2020).

[49] Precision and Recall. https://en.wikipedia.org/wiki/Precision_and_recall

[50] F-Score. https://en.wikipedia.org/wiki/F-score