**Final Year Project Report (Draft)**

**Title**

**Inference Attack Detection Using Gradient Boosting on DeepFake Metadata**

**1. Introduction**

The rise of DeepFake technology has raised serious privacy and security concerns. In privacy research, Membership Inference Attacks (MIA) are a major threat, where an attacker can infer whether a specific data sample was used to train a machine learning model. This project aims to simulate and analyze the feasibility of such inference attacks using metadata features extracted from DeepFake datasets, applying machine learning techniques to detect membership leakage.

**2. Objective**

To develop and evaluate a Membership Inference Attack (MIA) framework using Gradient Boosting (XGBoost) to detect whether a data sample was part of the training dataset of a model, using metadata from a DeepFake dataset (SDFVD).

**3. Dataset Used**

* **Dataset Name**: Small-scale Deepfake Forgery Video Dataset (SDFVD)
* **Source**: [Open-source academic dataset]
* **Description**: The dataset consists of labeled metadata for real and fake videos, including face attributes and confidence scores.

**4. Methodology**

**4.1 Data Preprocessing**

* Downloaded and unzipped SDFVD dataset.
* Simulated data distribution among multiple clients to represent a decentralized or federated learning environment.
* Converted tabular metadata into suitable format for model training.

**4.2 Victim Model Training**

* Trained a Logistic Regression model using one client’s data (victim client).
* The model learns to classify "real" vs "fake" samples based on metadata.

**4.3 Attack Simulation**

* Collected model outputs (confidence scores, losses, prediction correctness).
* Built two datasets:
  + **Member set**: Data used to train the victim model.
  + **Non-member set**: Data from other clients.

**4.4 Attack Model**

* Trained an XGBoost classifier using the following features:
  + Confidence Score of prediction
  + Log Loss for each prediction
  + Whether the prediction was correct (0/1)
* Labelled samples as "1" (member) and "0" (non-member).

**4.5 Evaluation**

* Evaluated the attack using ROC Curve and AUC (Area Under Curve).
* Visualized model behavior and attack success.

**5. Results**

* **ROC AUC Score**: *~0.85* (indicative of strong attack performance).
* The attack model could distinguish members from non-members with significant accuracy.
* Adding multiple features improved the attack success compared to using only confidence score.

**6. Tools and Technologies**

* Python 3.10
* Scikit-learn
* XGBoost
* Pandas, NumPy, Matplotlib
* Google Colab (Cloud Training)

**7. Conclusion**

This project successfully demonstrates that machine learning models are vulnerable to Membership Inference Attacks. Using metadata features and gradient boosting, we were able to infer whether a sample was used in model training. This highlights the importance of privacy-preserving techniques in AI systems, especially in sensitive domains like DeepFake detection.

**8. Future Work**

* Apply the same attack methodology on larger, real-world DeepFake datasets (e.g., DFDC).
* Extend to federated learning environments using frameworks like Flower or FedML.
* Implement privacy-preserving techniques like Differential Privacy to mitigate attacks.

**9. References**

1. [Abdellah Elmrini, decAttack GitHub Repo](https://github.com/AbdellahElmrini/decAttack)
2. [XGBoost Documentation](https://xgboost.readthedocs.io/)
3. [SDFVD Dataset Source Link]
4. OpenMined, PySyft for Privacy-preserving Machine Learning