### **Project Plan**

#### **1. Project Overview**

* **Project Title:** Privacy Benchmarking Tool for Secure & Accurate Federated Learning
* **Primary Objective:** To validate and extend the previous finding that selective encryption of early layers in a CNN is the most efficient method for protecting model parameters in Federated Learning against privacy attacks.
* **Key Question:** Does the "sweet spot" of encrypting only the first 1-2 layers hold true for different model architectures and under more sophisticated attack scenarios?
* **Success Criteria:**
  + Reproduce the previous group's core results on ResNet18/CIFAR-100.
  + Successfully test at least one additional model architecture (e.g., VGG, a simple custom CNN).
  + Introduce and evaluate a new, more complex attack strategy (e.g., based on the future work's "randomized parameters").
  + Provide clear, data-backed recommendations on optimal encryption strategies.

#### **2. Scope**

* **In-Scope:**
  + **Models:** ResNet18 (for replication), and **one new model** (e.g., VGG16 or a simpler CNN to see if the effect is architecture-dependent).
  + **Dataset:** CIFAR-100 (for consistency and comparison).
  + **Tools:** TenSEAL (for HE), Breaching library (for attacks), PyTorch/TensorFlow (for FL simulation).
  + **Experiments:** Re-running "Yin", "Random", and "None" attacks while varying the number and location of encrypted layers.
  + **New Element:** Designing a simple "confusion" strategy where some client updates are intentionally noisy or randomized.
* **Out-of-Scope:**
  + Testing on very large-scale datasets (e.g., ImageNet).
  + Developing a new encryption library or a new attack method from scratch.
  + A full-blown, multi-institution FL deployment. This remains a simulation.

#### **3. Milestones, Deliverables & Timeline (Example for a 12-week project)**

|  |  |  |
| --- | --- | --- |
| Week | Milestone | Deliverable |
| **3-4** | **Project Setup & Environment Reproducibility** | A working code environment that can run the previous group's baseline (ResNet18, no encryption, under attack). |
| **5-7** | **Milestone 1: Core Validation** | A report and graphs confirming the previous results (PSNR/SSIM vs. # of encrypted layers for ResNet18). |
| **8-9** | **Milestone 2: Architectural Exploration** | Results and analysis for the new model architecture (e.g., VGG16), showing how the privacy/overhead trade-off differs. |
| **10-11** | **Milestone 3: Advanced Attack Scenario** | Results from testing the "confusion" strategy or other advanced attack simulations. |
| **11-12** | **Milestone 4: Synthesis & Analysis** | A consolidated analysis comparing all results. Identification of the "optimal" encryption strategy. |
| **12-14** | **Final Report & Presentation** | A comprehensive final report and presentation slides summarizing findings, limitations, and future work. |

#### **4. Team Roles & Responsibilities**

* **Role 1:** Focused on reproducibility and running experiments on ResNet18.
* **Role 2:** Focused on integrating and testing the new model architecture (VGG16/Custom CNN).
* **Role 3:**
* **Role 4:**
* **Role 5:**
* **Role 6:**
* **Role 7:**
* **Role 8:**
* **Role 9:**
* **Role 10:**
* **Role 11:**
* **All Researchers:** Collaborate on designing the new attack scenario, analyzing results, and writing the final report.

#### **5. Technical Approach & Methodology**

This is the core of ourplan.

**Phase 1: Baseline Reproduction**

1. Set up the FL simulation with a single "malicious" server using the Breaching library.
2. Train ResNet18 on CIFAR-100 in a federated setting **without any encryption**.
3. Launch the "Yin" and "Random" attacks to establish baseline attack success (low PSNR/SSIM).

**Phase 2: Selective Encryption Testing**

1. Integrate TenSEAL into the training loop.
2. Define a clear strategy for which layers to encrypt (e.g., Encrypt Layer 1 only -> Layers 1&2 -> Layers 1,2,3 -> ... -> Full model).
3. For each configuration:
   1. Run the FL training.
   2. Launch the same set of attacks.
   3. Record the **PSNR** and **SSIM** of the reconstructed data.
   4. Measure and record the **computational time and communication overhead**.

**Phase 3: Model Generalization**

1. Repeat Phase 1 and Phase 2 with the new model architecture (e.g., VGG16). Does the "first layers are most critical" rule still hold?

**Phase 4: Enhanced Attack Simulation**

1. Modify the FL simulation so that a percentage of client updates include added noise or are partially randomized before being sent to the server.
2. Re-run key encryption configurations against this new, more complex environment to test the robustness of the selective encryption strategy.

#### **6. Evaluation Metrics**

* **Primary Metrics (Privacy):** PSNR (Peak Signal-to-Noise Ratio) and SSIM (Structural Similarity Index) of the attacker-reconstructed images. **Lower PSNR/SSIM = Better Privacy.**
* **Primary Metrics (Performance):**
  + **Computational Overhead:** Total training time for each encryption configuration.
  + **Communication Overhead:** Size of the encrypted model updates vs. plaintext.
  + **Model Utility:** Final accuracy of the global FL model on a test set (to ensure encryption isn't harming learning).
* **Secondary Metric:** Number and proportion of encrypted parameters (as a straightforward measure of cost).

#### **7. Potential Risks & Mitigation**

* **Risk:** Inability to reproduce previous group's results due to missing code or environment details.
  + **Mitigation:** Start this immediately. Document all issues. If reproduction fails, use the methodology description to implement your own baseline.
* **Risk:** TenSEAL library is complex and causes integration headaches or excessive computation times.
  + **Mitigation:** Start with a "Hello World" example in TenSEAL. Plan for extra time in the schedule for debugging.
* **Risk:** Experiments take too long to run, blocking progress.
  + **Mitigation:** Use a small subset of the data for initial testing. Ensure you have access to sufficient computational resources (e.g., GPU).

#### **8. Future Work & Conclusion (To be filled after your work)**

This section in the final report will directly build on our findings. It might include:

* Testing on a real-world, non-image dataset (e.g., medical records).
* Exploring hybrid privacy techniques (e.g., HE + Differential Privacy).
* A deeper dive into the regulatory implications (GDPR, HIPAA) based on your quantitative results.