# Exploring Privacy Risks: Reconstructing Medical Images from Obfuscated Gradients in Federated Learning

## Project Proposal

### Motivation and Objectives:

Federated learning offers collaborative machine learning while preserving data privacy. However, concerns persist regarding privacy in medical image analysis. Leveraging a reconstruction attack framework proposed by Yue et al., I aim to assess gradient obfuscation techniques' effectiveness in safeguarding patient privacy. With medical images being highly sensitive, I'll evaluate methods like gradient noise injection, compression, and quantization. Through experiments with real-world medical datasets, I'll quantify potential privacy leaks and assess the impact on model utility. Balancing privacy with model performance is crucial in healthcare. This project aims to contribute to more robust privacy protection mechanisms, enhancing patient confidentiality in medical image analysis within federated learning.

### Related Work

In my project, I will review prior work, particularly the study by Zhu et al., to assess its applicability to Federated Averaging (FedAvg) settings, focusing on its limitations concerning image reconstruction. This critical review will inform our research by highlighting gaps in existing methodologies and guiding the development of more effective privacy protection mechanisms for FedAvg and similar federated learning frameworks.

### Proposed Work

In this project, I aim to adapt an existing reconstruction framework, as outlined in the research paper [1], to work with medical image datasets. Unlike the original application of the framework, which focused on ImageNet datasets, my adaptation will address the unique challenges posed by medical images, containing highly sensitive patient information. I will modify this framework to suit medical image datasets and undertake data preprocessing to ensure anonymity and compliance with regulations. Following this, I will attempt to reconstruct the images protected by federated learning techniques like Federated Averaging (FedAvg). By doing so, I seek to assess the framework's effectiveness in preserving privacy and maintaining data security within federated

learning settings. This approach differs from existing work, which primarily focused on reconstruction techniques for ImageNet datasets. My project extends this research to address the specific privacy concerns and requirements associated with medical image analysis within federated learning frameworks. Through this adaptation and evaluation, I aim to contribute to the development of more robust privacy protection mechanisms tailored to the healthcare domain.

## Evaluation

For evaluating the effectiveness of the adapted reconstruction framework on medical image datasets, I plan to employ similar evaluation techniques used in the research paper. The primary evaluation metric will be the Learned Perceptual Image Patch Similarity (LPIPS) score, which has been demonstrated to effectively emulate human perception and evaluate the quality of reconstructed images. LPIPS offers advantages over traditional metrics such as mean squared error (MSE), peak signal-to-noise ratio (PSNR), and structural similarity index measure (SSIM) by capturing high-level semantics in raw images more accurately. Additionally, I will explore other methods to evaluate semantic privacy loss.

## Plan of Action

**Week 1:**

Review the research paper and understand the reconstruction framework. Gather medical image datasets and preprocess them for anonymity and compliance. Set up the development environment and necessary tools for implementation. Understand the code for the reconstruction framework (given in the research paper) on GitHub.

**Week 2:**

Work with the code for the reconstruction framework provided in the GitHub repository. Implement data preprocessing techniques to ensure compliance with regulations.

**Week 3:**

Adapt the reconstruction framework to work with medical image datasets. Begin model training using federated learning techniques like Federated Averaging (FedAvg).

**Week 4:**

Continue model training and optimization. Refine data preprocessing techniques as necessary. Explore initial evaluations using LPIPS and other relevant metrics.

**Week 5:**

Complete model training and evaluation. Conduct further evaluations and refine evaluation metrics. Analyze results and identify any areas for improvement. Begin drafting project report and documentation.

**Week 6:**

Finalize project report, including methodology, results, and conclusions. Prepare presentation slides for project presentation. Practice project presentation.

**Week 7:**

Reflect on the project process and outcomes. Prepare for any additional presentation or demonstration requirements.

## Bibliography

[1] Kai Yue, Richeng Jin, Chau-Wai Wong, Dror Baron, and Huaiyu Dai Gradient Obfuscation Gives a False Sense of Security in Federated Learning

[2] Ligeng Zhu, Zhijian Liu, and Song Han. Deep leakage from gradients. In Advances in Neural Information Processing Systems, 2019

[3] Rui Zhang, Song Guo, Junxiao Wang, Xin Xie, and Dacheng Tao. A survey on gradient inversion: Attacks, defenses and future directions.