MICCAI2025
28th International Conference on Medical Image Computing and Computer Assisted Intervention
23-27 September 2025
Daejeon Convention Center
Daejeon REPUBLIC OF KOREA

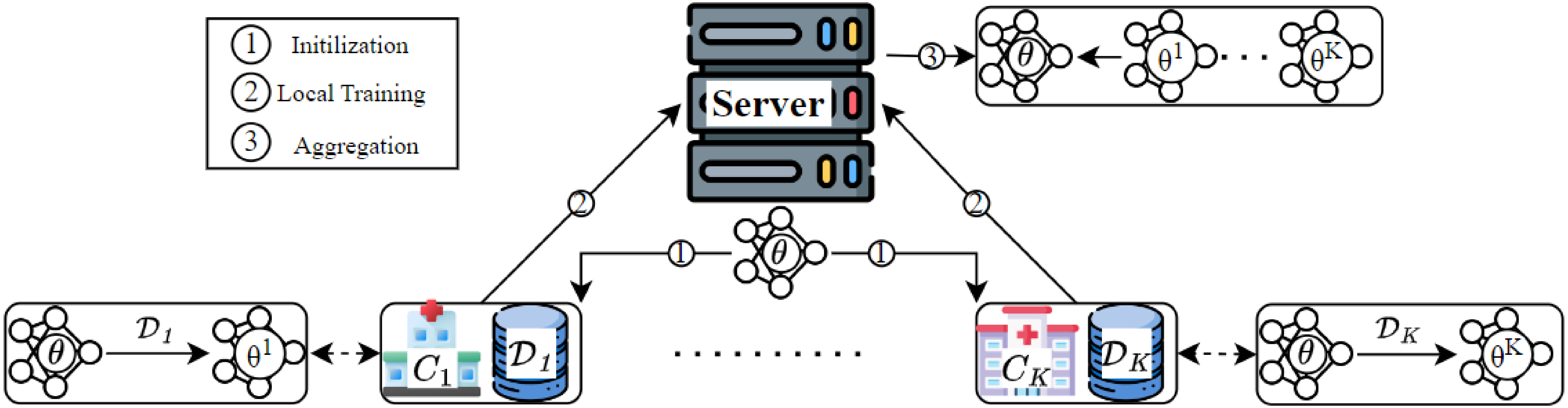
UNIVERSITI MALAYA

# Maverick: *Collaboration-free* Federated Unlearning for Medical Privacy

**Win Kent Ong**, Chee Seng Chan

Universiti Malaya, Malaysia

# Federated Learning

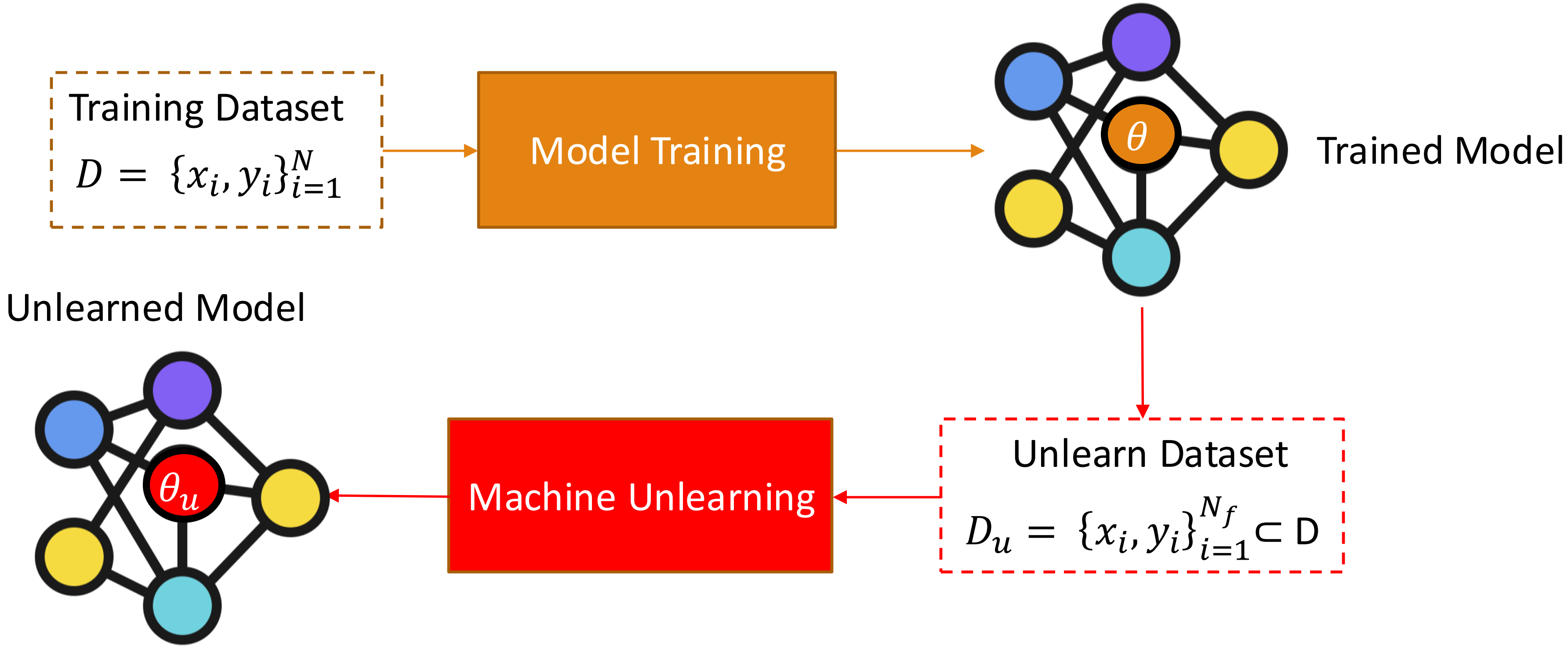



Machine Learning algorithm enables multiple parties to collaboratively train a model

- Without sharing private data, only sharing trained weights
- Better data privacy protection, reducing the risk of privacy leakage

# Machine Unlearning

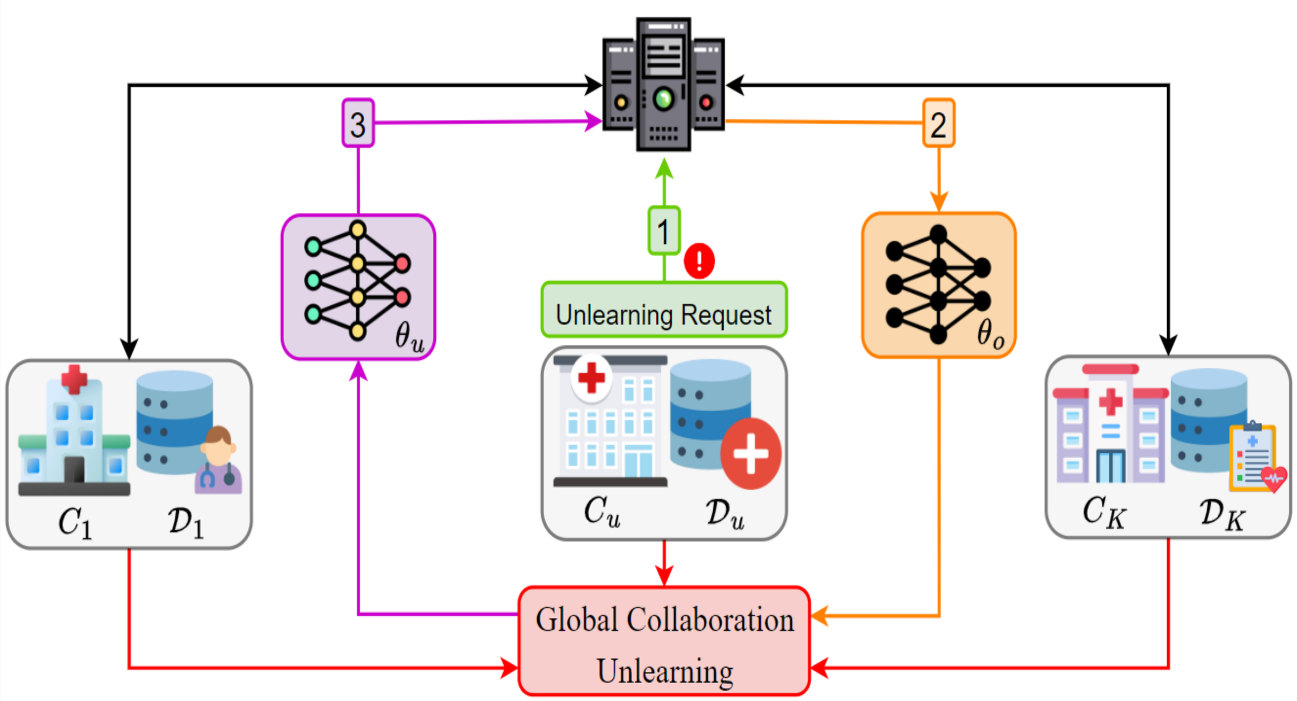
- Remove the influence of a subset of its training dataset from the trained neural network without retraining.



# Machine Unlearning

- Privacy Regulation Laws
  - California Consumer Privacy Act (CCPA)
  - General Data Protection Regulation (GDPR)
  - Consumer Privacy Protection Act (CPPA)
  - Secure *the right to be forgotten* of data provider
- Security and Privacy Protection
  - Prevents *leaking sensitive or private information* that was unintentionally memorized during training (e.g., passwords, medical details).
- Error Correction
  - To remove mislabeled samples, or corrupted records
  - Allows selective removal of *biased or unfair features* so that the model's predictions become more equitable.

# Motivation




**Existing** ***Federated Unlearning*** **framework:**

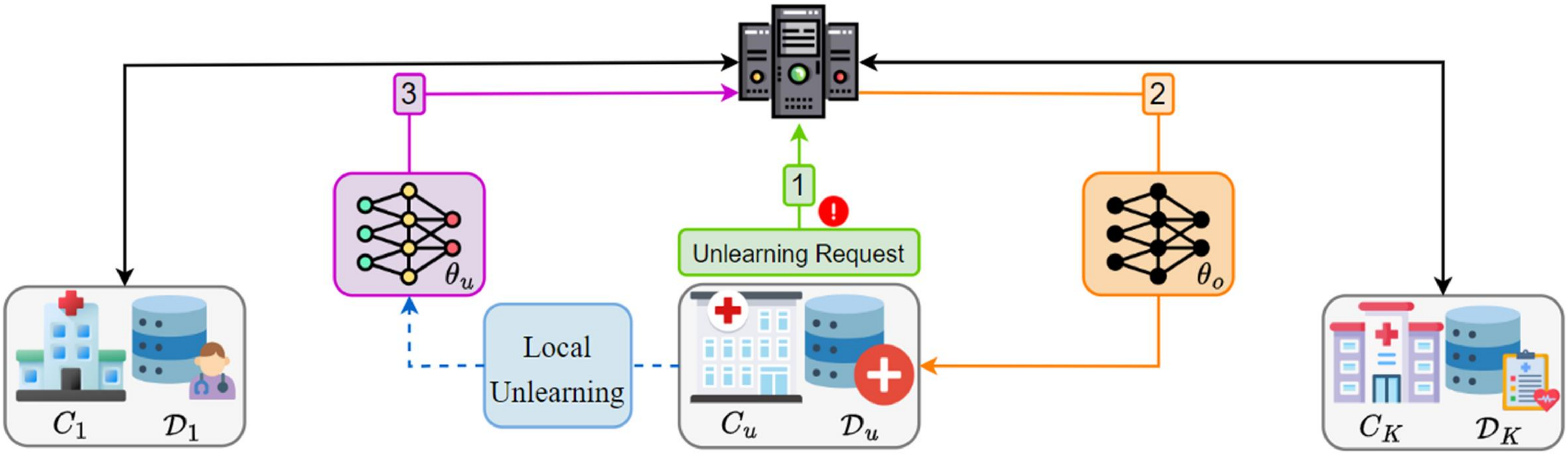
1. Require global client collaboration.
2. Increasing privacy risks
3. High computational burden.
4. Lack of a unified unlearning solution that can works across sample, class and client unlearning.

# Objectives

To develop a ***Federated Unlearning*** algorithm for medical AI that:

1. Works without global client collaboration
2. Preserves privacy
3. Improves efficiency

# Proposed Method - *Maverick*




Enables ***local unlearning*** at the target client ***without requiring collaboration from other clients***, ensuring privacy, effectiveness and efficiency.

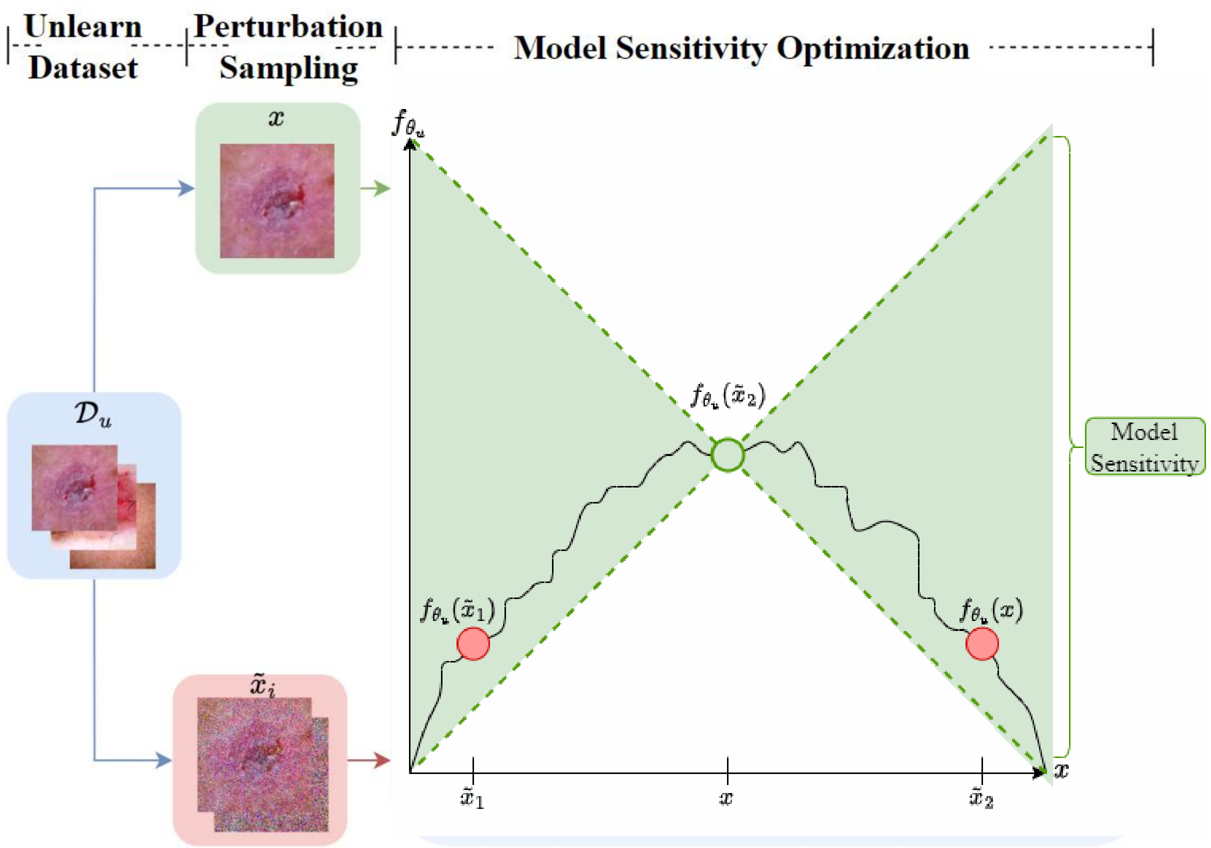
# Model Sensitivity

The model sensitivity ***s*** of the model $f_\theta$ with respect to the sample ***x*** is defined as:

$$\text{Model sensitivity, } \boldsymbol{s} = \frac{\|f(x)-f(\bar{x})\|}{\|(x)-(\bar{x})\|} = \frac{\|f(x)-f(x+\delta)\|}{\|(x)-(x+\delta)\|} = \frac{\|f(x)-f(x+\delta)\|}{\|\delta\|}$$

- Quantifies the rate of change in the model's output relative to input perturbations.
- A smaller value of ***s*** indicates that $f_\theta$ exhibits minimal memorization of sample ***x***.
- Averages output variation over perturbations $\delta$, eliminating dependence on the entire dataset.

# Local Unlearning




1. ***Perturbation Sampling***: Add Gaussian noise to input samples

$$\tilde{x} = x + \delta, where\ \delta \sim \mathcal{N}(0, \sigma^2)$$

2. ***Model Sensitivity Approximation***: Monte Carlo sensitivity estimation.

$$\mathbb{E}_\delta \frac{\|f_{\theta_o}(x) - f_{\theta_o}(\tilde{x})\|_2}{\|x - \tilde{x}\|_2} \sim \frac{1}{N}\sum_{i=1}^{N} \frac{\|f_{\theta_o}(x) - f_{\theta_o}(\tilde{x}_i)\|_2}{\|\delta_i\|_2}$$

3. ***Local Optimization***: Reduce model's output response to the target data.

$$\theta_u = argmin \mathbb{E}_{(x,y)\in D_u} \frac{1}{N}\sum_{i=1}^{N} \frac{\|f_{\theta_o}(x) - f_{\theta_o}(\tilde{x}_i)\|_2}{\|\delta_i\|_2}$$

# Experimental Setup

**Federated Learning Simulation**

- Horizontal FL setup with ***K* = 10 clients**
- IID setting: each client receives **10% of the dataset**

**Unlearning Scenarios**

- **Sample unlearning**: 40% of a client's data removed using backdoor-based techniques
- **Class unlearning**: Class 1 removed from the client's dataset
- **Client unlearning**: Entire client dataset removed

**Model & Medical Datasets**

- Datasets:
    1. Colorectal Cancer Histology Slides (Path)
    2. Pigmented Skin Lesions (Derma)
    3. Blood Cells (Blood)
- Backbone: **ResNet18**

# Baselines

- **Baseline**
  - Original model before unlearning.
- **Retrain**
  - Model training without the presence of unlearn feature.
- **Fine-tune (FT)**
  - Fine-tuning baseline model with the retain dataset.
- **FedCDP**
  - Class unlearning via Term Frequency Inverse Document Frequency (TF-IDF) guided channel pruning.
- **FedRecovery**
  - Sample and client unlearning visa client gradient submission.

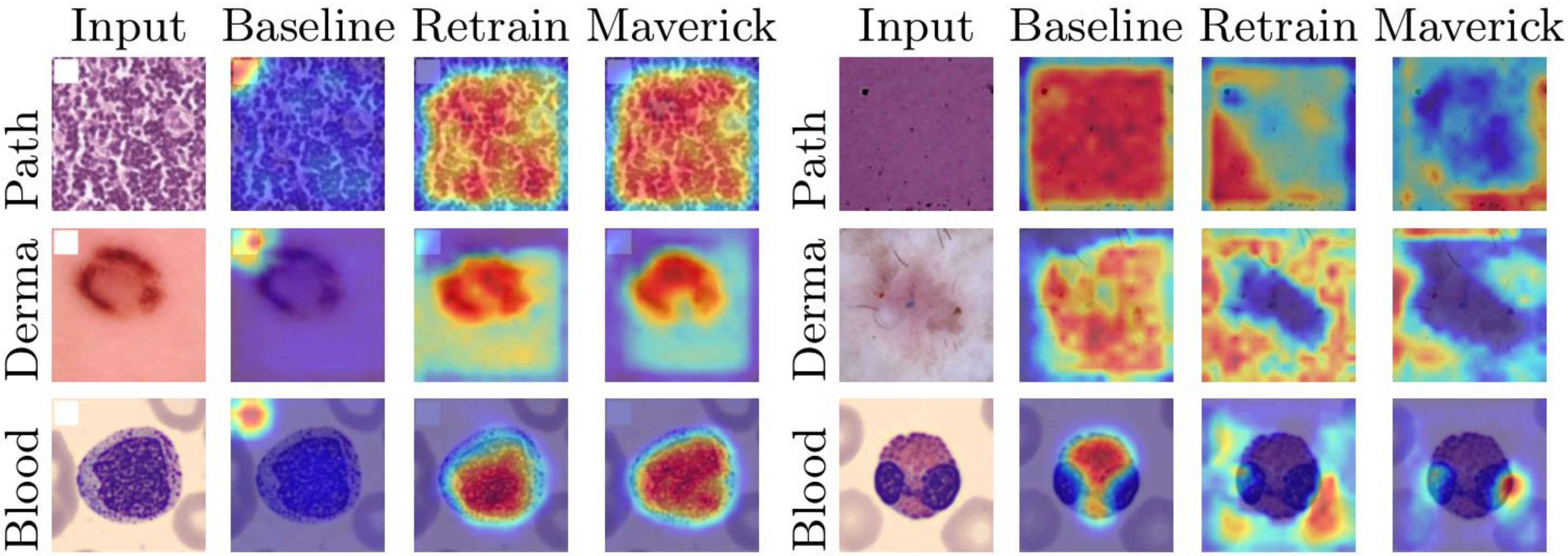
# Metrics

- **Fidelity**
  - ➢ Accuracy of retain dataset $D_r$
- **Effectiveness**
  - ➢ Accuracy of unlearn dataset $D_u$
- **Privacy**
  - ➢ Membership inference attack (MIA)
- **Efficiency**
  - ➢ Runtime in seconds

# Effectiveness and Utility

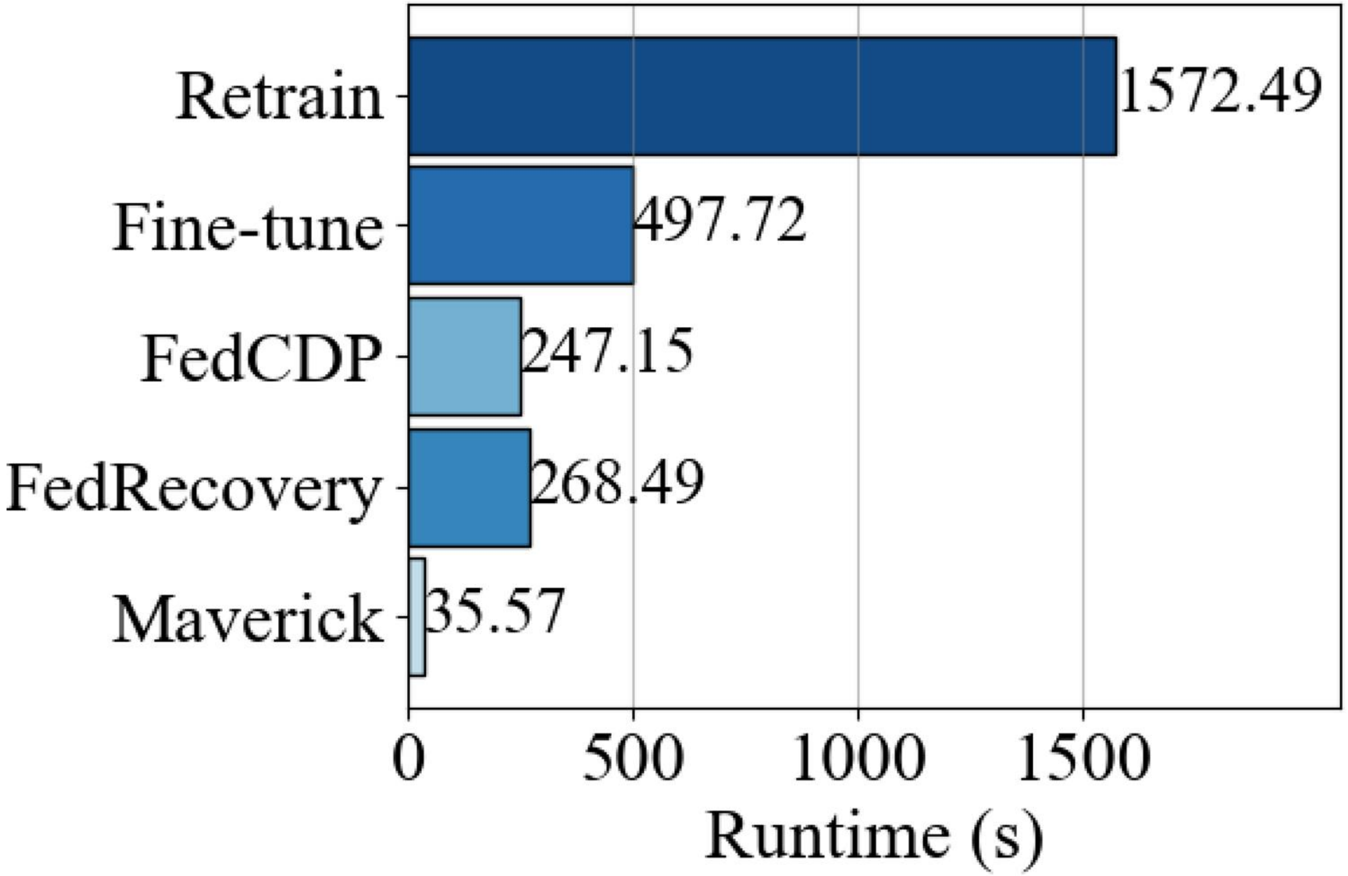
| Scenarios | Datasets | Metrics | Accuracy (%) | | | | | |
|---|---|---|---|---|---|---|---|---|
| | | | **Baseline** | **Retrain** | **FT** | **FedCDP** | **FedRecovery** | **Maverick** |
| Sample | Path | $D_r$ | 91.37 | 92.50 | **93.04** | 70.19 | 90.14 | 89.43 |
| | | $D_u$ | 90.48 | 0.00 | 46.13 | 22.61 | 2.35 | **0.71** |
| | | MIA | 92.51 | 8.69 | 55.49 | 38.05 | 13.43 | **10.04** |
| Class | Derma | $D_r$ | 82.52 | 80.39 | **81.38** | 79.31 | 55.51 | 79.18 |
| | | $D_u$ | 80.88 | 0.00 | 53.69 | 0.51 | 31.40 | **0.18** |
| | | MIA | 90.62 | 2.60 | 40.44 | 5.17 | 34.16 | **0.49** |
| Client | Blood | $D_r$ | 91.21 | 91.90 | **93.38** | 79.58 | 89.54 | 88.33 |
| | | $D_u$ | 92.83 | 0.00 | 43.38 | 25.29 | 1.95 | **0.53** |
| | | MIA | 96.71 | 5.78 | 52.57 | 39.85 | 10.95 | **6.73** |

# Attention Map




(a) Sample unlearning.    (b) Class unlearning.

# Time Efficiency



# Conclusion

- Maverick is the *first* collaboration-free federated unlearning method for medical AI.
- Enables local unlearning without disturbing other clients.
- Demonstrates strong results in privacy, efficiency, and fidelity.
- Well-suited for real-world healthcare and privacy-critical domains.

# Thank you for listening!

Paper

Code