Proceedings of the 12th International Conference of Control Systems, and Robotics (CDSR 2025)
July 13, 2025 - July 15, 2025 | Imperial College London Conference Center, London, United Kingdom
Paper No.129
DOI: 10.11159/cdsr25.129

Improved Federated Learning with Differential Privacy for YOLO Detector on Phase Images

Qianyu Chen[1,2,*], Kerem Delikoyun[1,2], Oliver Hayden[1,2], Klaus Diepold[1,2,*]
[1]Technical University of Munich
Arcisstraße 21, 80333 München, Germany
[2]TUMCREATE
1 Create Wy, #10-02 CREATE Tower, Singapore
qianyu.chen@tum-create.edu.sg; kerem.delikoyun@tum-create.edu.sg; oliver.hayden@tum.de; kldi@tum.de

Abstract - Federated learning (FL) offers a decentralized paradigm for privacy-preserving medical imaging, yet static differential privacy (DP) often degrades model utility. We propose FL-YOLO-DADP, integrating dynamic adaptive DP (DADP) with YOLOv8n for hematological cell detection in digital holographic microscopy. Unlike static DP-FL, DADP adaptively scales noise and clipping based on training phases, maintaining high accuracy while guaranteeing (ϵ, δ)-DP. Our experiments on a multi-client dataset demonstrate that FL-YOLO-DADP significantly outperforms both centralized training on synthetic images and naive federated baselines, with 46.8% higher mAP50 than the GANs-generated train set. Further, Eigen-CAM visualizations confirm that critical morphological features are retained, showcasing the framework's robustness in small-data and heterogeneous settings.

Keywords**:** Federated Learning, YOLO, Dynamic Adaptive Differential Privacy, Phase Image

1. Introduction

The growing demand for privacy-preserving machine learning in medical imaging necessitates frameworks that balance data confidentiality with model performance. Deploying deep learning for hematological cell detection (e.g., leukocytes, erythrocytes, and platelets) faces two competing demands: preserving patient privacy under regulations like GDPR and maintaining diagnostic accuracy. Centralized training on pooled data risks exposing sensitive patient information, while synthetic data generated by diffusion models [1] or Generative Adversarial Networks (GANs) [2] introduces distribution shifts that degrade detection performance. For instance, while diffusion models can generate high-quality synthetic images, studies show ([3-4]) they do not always enhance downstream segmentation performance when substituting real medical images, and may even introduce distortions that degrade detection accuracy.

Still, generative models like encoder-decoder networks, GANs, Denoising Diffusion Probabilistic Models (DDPMs), and ControlNet-latent-img2img models are often proposed to replace raw images to mitigate privacy risks in small-scale settings. Recent studies [5-8] have demonstrated the feasibility of generative models in medical imaging tasks, including segmentation and synthetic data generation. For instance, encoder-decoder frameworks have been utilized to conceal sensitive information in medical images, enabling data sharing without compromising patient privacy [5]. Similarly, GAN-based models have been leveraged for medical image segmentation, effectively addressing domain adaptation and data scarcity challenges [6]. Moreover, DDPMs have been applied to generate high-resolution volumetric medical images, facilitating privacy-preserving data sharing without compromising clinical utility [7]. Additionally, advanced generative models have been proposed for medical image synthesis, enhancing data availability while preserving patient confidentiality. In [8], the authors discuss the integration of ControlNet with various external condition generation methods to enhance its image synthesis capabilities.

Federated learning (FL) circumvents data sharing by training models locally on client devices and aggregating model updates globally [9]. However, naive FL remains vulnerable to gradient inversion attacks [10], where adversaries reconstruct training images from shared gradients. While prior DP-FL (Differential Privacy) approaches [11] apply fixed noise scales, gradient statistics evolve non-linearly during training, rendering static clipping norms suboptimal. To address this, we propose FL-YOLO-DADP, a framework for integrating dynamic adaptive differential privacy (DADP) into federated YOLOv8 training, ensuring minimal utility loss while enforcing rigorous privacy guarantees. Fig. 1 overviews the FL-

YOLO-DADP workflow, illustrating how each client's model updates and makes noises before sending it to an FL Server Aggregator employing a FedMedian operation. In contrast to generative models, FL preserves raw data fidelity while enforcing privacy at the algorithm level, avoiding the pitfalls of distributional mismatch in synthetic data.

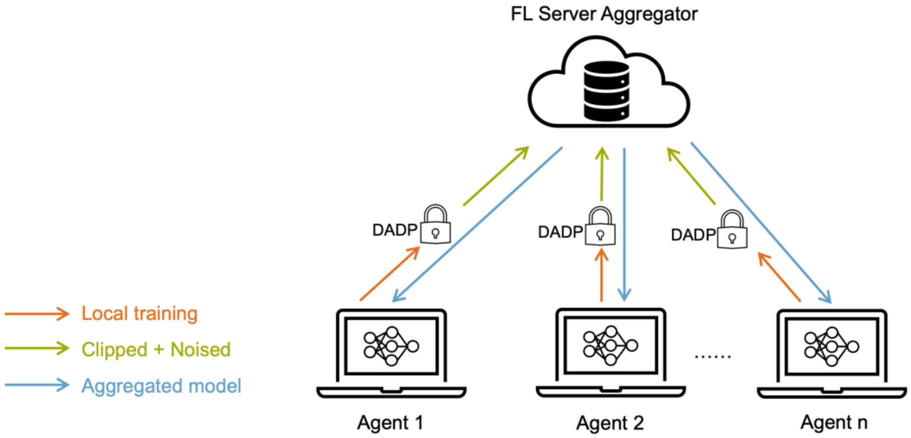

Fig. 1: FL-YOLO-DADP workflow.

Digital Holographic Microscopy (DHM) is a label-free imaging modality that measures refractive index variations within cellular structures in information of cells. In this study, we employed DHM [1] blood cell imaging. The microscope employs a 528 nm SLED (PowerStar, Oslon) under partially coherent Köhler illumination in a transmission setup. The laboratory prototype is equipped with a Nikon CFI LWD 40× objective (NA = 0.55), providing in particular phase resolution for visualizing subtle cell morphologies in high-throughput flow cytometry conditions.

Notably, popular differential privacy (DP) mechanisms like DP-SGD are not under our consideration, as they require clipping each sample's gradient and injecting noise into the aggregated gradients. However, wrapping the optimizer with DP-SGD during the client's training phase in YOLO requires tremendous modification on the Ultralytics package. Therefore, after training on each client, our methodology adds noise to the model parameters after training on each client before sending them to the server called Local Differential Privacy (LDP). We inject LDP to FedAvg and FedMedian for vanilla federation.

Our approach addresses key challenges in medical imaging, where privacy, heterogeneous data distribution, and small sample sizes constrain traditional centralized learning. By adaptively clipping and injecting noise based on training phases, FL-YOLO-DADP achieves near-centralized accuracy while ensuring stringent (ϵ, δ)-differential privacy guarantees. Comparative experiments on a phase image dataset show substantial improvements over synthetic data baselines. Using a multi-client FL setup, we evaluate our DADP-FL framework against a YOLOv8n baseline and four generative models. Metrics include mean average precision (mAP50) and inference speed, with FL configurations optimized to minimize communication overhead. Experiments on a multi-client hematology dataset validate our framework's comparability in privacy-sensitive medical applications with small-data regimes.

Our contributions are:

1) By adaptively clipping and injecting noise over training phases, FL-YOLO-DADP achieves better accuracy than centralized training on generative images while ensuring (ϵ, δ)-differential privacy.
2) Comparative experiments on a phase-image dataset show substantial improvements over synthetic-data baselines.
3) Visualization of model explainability and proof of convergence under adaptive noise support theoretical guarantees.

2. Related Work

[1] Ovizio Imaging Systems (Belgium)

2.1. General Comments

To comprehensively prove the feasibility of employing FL-based methods for developing a trustworthy data protection protection system, we also utilize generative models to synthesize artificial images with conditional labels for feeding to to YOLO model to classify cells for object detection performance comparison, as the original feature information can be be fully encrypted on synthesized images. The generated images retain clinical relevance and anatomical correctness, as validated by expert assessments in prior studies [12]. Such a framework can augment limited datasets and enhance downstream diagnostic models without exposing sensitive patient information [13].

2.2. Benchmark

YOLO (You Only Look Once) has become a state-of-the-art object detection family for real-time applications, performing bounding-box regression and classification in a single stage [14]. In [15], the authors applied YOLOv8x-p2 for cell classification tasks on phase images and achieved high accuracy. We adopt YOLOv8n for its improved backbone and detect-head structure, which is well-suited to small or transparent objects such as blood cells.

2.3. Generative Models

2.3.1 Encoder-Decoder – Our model extends a simplified U-Net [16] that downsamples the conditional mask with 3 channels through convolution and pooling layers, then upsamples via transposed convolutions with skip connections. This setup preserves spatial information crucial for synthesizing biologically realistic outputs.

Given the need to handle non-iid data across decentralized institutions, the encoder-decoder framework incorporates domain-specific adaptations at the decoder level to address site-specific imaging variances, following strategies such as modality-specific decoders[17]. U-Net is the foundational architecture for many popular generative models.

2.3.2 Generative Adversarial Networks (GANs) – In this work, a conditional GAN framework is applied, where a U-Net-based generator takes bounding box masks as input and synthesizes corresponding RGB images. The generator architecture is inspired by the Pix2Pix [18] model, incorporating downsampling and upsampling layers for effective feature extraction and image reconstruction [19]. The discriminator is based on the PatchGAN architecture [20], which classifies image patches as real or fake, encouraging the generator to produce high-fidelity images at the local level.

During training, the generator aims to minimize the L1 reconstruction loss, which ensures that the generated images are structurally like the real images, while the discriminator aims to minimize a binary cross-entropy loss to distinguish between real and synthetic images. The combination of adversarial loss and reconstruction loss encourages the generator to produce realistic images that capture both global and local features [21]. The generator's encoder-downsampler and decoder-upsampler blocks follow the pix2pix design, enhanced by skip connections that preserve spatial detail. The discriminator uses consecutive convolution as LeakyReLU layers, leading to a final map where each patch's output signals real or fake.

2.3.3 Denoising Diffusion Probabilistic Models (DDPM) – Unlike adversarial training, which can suffer from instability and mode collapse, DDPMs iteratively transform random noise into structured images through a series of learned denoising steps [22], leveraging a U-Net-based architecture to predict the noise at each time step.

The forward process introduces Gaussian noise to the real images over a series of time steps, progressively converting them into pure noise. The model learns to reverse this process by predicting the added noise at each step, conditioned on the input mask. The implemented U-Net model concatenates the noisy image and the mask as input and employs skip connections to preserve spatial information. In testing, we sample random noise and then apply the learned reverse steps conditioned on the bounding-box mask to yield synthetic images.

2.3.4 ControlNet Latent Img2Img – The proposed framework combines latent diffusion models (LDMs) with ControlNet for conditional image synthesis. LDMs leverage the latent space of a pretrained variational autoencoder, ControlNet introduces additional conditioning inputs for fine-grained spatial control over the generation process [23].

The model consists of the following components: Autoencoder (VAE), U-Net Backbone, ControlNet Module, and Noise Scheduler. VAE is Used to encode input images into a compressed latent representation and reconstruct high-quality outputs. U-Net Backbone is a denoising neural network trained to iteratively remove noise from the latent space. The ControlNet module processes bounding box masks as spatial conditioning inputs and injects structured feature maps into the U-Net via residual connections [24]. Noise scheduler implements a discrete noise distribution based on a scaled linear schedule.

3. Federated Learning

3.1. Overview

In this study, we implemented a FedMedian aggregation strategy within all federated YOLOv8 object detection frameworks to detect cellular structures in microscopic medical images. This will be described in the following subsections. The FedMedian strategy enhances robustness to outliers by calculating the median of local client models, thereby mitigating the influence of adversarial or biased updates from individual clients [25].

3.2. FL-YOLO

Let there be N clinical sites (clients), each holding a local dataset $\mathcal{D}_k$ of medical for $k \in \{1, \dots, N\}$. We define $\mathcal{D}_k$ as $\{(x_{k,i}, y_{k,i}) | i = 1, \dots, |\mathcal{D}_k|\}$, where $x_{k,i}$ is the i-th image on client k and $y_{k,i}$ includes bounding boxes and class labels. For local model updates at round r, the server broadcasts the global parameters $\mathbf{w}^{(r-1)}$. Each client k trains locally for one or more epochs on its dataset $\mathcal{D}_k$. We denote the new local model parameters by $\mathbf{w}^{(r)}$. The clients' updates are determined as

$$\Delta\mathbf{w}_k^{(r)} = \mathbf{w}_k^{(r)} - \mathbf{w}^{(r)}. \tag{1}$$

In FedMedian aggregation, after collecting $\Delta\mathbf{w}_{\boldsymbol{k}}^{(\boldsymbol{r})}$ from k clients, the server aggregates using FedMedian

$$\mathbf{m}^{(r)} = \text{median}(\Delta\mathbf{w}_1^{(r)}, \Delta\mathbf{w}_2^{(r)}, \dots, \Delta\mathbf{w}_k^{(r)}), \tag{2}$$

where "median" is taken coordinate by coordinate. The new global model is updated by

$$\mathbf{w}^{(r)} = \mathbf{w}^{(r-1)} + \mathbf{m}^{(r)}. \tag{3}$$

3.3. FL-YOLO-Dynamic Adaptive Differential Privacy

3.3.1 Overview – Our framework combines federated YOLOv8 training with dynamic adaptive DP, where noise and clipping norms adjust based on real-time gradient statistics and training phase. The system comprises: Client-Side for local training with adaptive gradient clipping and phase-dependent noise and Server-Side as robust FedMedian aggregation and privacy budget tracking.

3.3.2 (ϵ, δ)-Differential Privacy – A randomized algorithm $\mathcal{M}$ satisfies (ϵ, δ)-DP [26] if, for any two neighbouring datasets $\mathcal{D}$ and $\mathcal{D}'$ differing in a single record and for all measurable subsets S,

$$\Pr[\mathcal{M}(D) \in S] \leq e^{\epsilon} \Pr[\mathcal{M}(D') \in S] + \delta. \tag{4}$$

In federated learning, each client's local update $\Delta\mathbf{w}_k$ is clipped and noised before transmission, making the effective update a differentially private mechanism. Then the server aggregates the differentially private updates to produce $\mathbf{w}^{(r)}$.

3.3.3 Adaptive Clipping Scheduling – In practice, a static noise level σ across all training rounds may be suboptimal. We propose a dynamic scheduling strategy that decays the noise in early rounds and increases it slightly in later rounds if needed, mitigating catastrophic performance drops while maintaining overall privacy. We define:

$$\sigma^{(R)} = \begin{cases} \sigma_0[\alpha + (1 - \alpha)exp(-\omega r)], & r < R/2 \text{ (Exploration)}, \\ \sigma_0[1 + \beta(r - R/2)], & r \geq R/2 \text{(Exploitation)}, \end{cases} \tag{5}$$

where σ_0 is an initial scaling, and α, ω controls decay rate, β governs late-stage noise growth. A central server coordinates training over R rounds. At round r, each client uses $\sigma^{(r)}$ for the noise addition step, balancing privacy budget usage across training.

3.3.4 Per-Client Differential Privacy Update – During round r, client k computes the $\mathcal{L}_{\mathbf{2}}$ norm of update vector $\Delta\mathbf{w}_k^{(r)}$. The clipping norm $C^{(r)}$ is updated via exponential moving average (EMA). With θ stabilizes the norm against outliers

$$C^{(r)} = \theta C^{(r-1)} + (1-\theta) \parallel \Delta\mathbf{w}_k^{(r)} \parallel_2, \tag{6}$$

3.3.5 Process Steps – We perform the following processing steps:

1) Clip globally: Rescale $\Delta\mathbf{w}^{(r)}$ so that $\parallel \Delta\mathbf{w}_k^{(r)} \parallel_2 \leq C$, where C is the global clipping norm

$$\Delta\mathbf{w}_k^{(r)} \leftarrow \Delta\mathbf{w}_k^{(r)} \cdot \min\left(1, \frac{C^{(r)}}{\parallel \Delta\mathbf{w}_k^{(r)} \parallel_2}\right), \tag{7}$$

2) Add noise: Sample isotropic Gaussian noise $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \sigma^2 C^2 \mathbf{I})$ and construct

$$\Delta\widetilde{\mathbf{w}_k}^{(r)} = \Delta\mathbf{w}_k^{(r)} + \mathbf{z}, \qquad \text{where } z \sim \mathcal{N}(\mathbf{0}, \sigma^2 C^2 \mathbf{I}), \tag{8}$$

The parameter σ is the dynamic noise multiplier controlling the trade-off between utility and privacy.

3) Update local model parameters

$$\mathbf{w}^{(r)} = \mathbf{w}^{(r-1)} + \text{median}(\Delta\widetilde{\mathbf{w}_1}^{(r)}, \Delta\widetilde{\mathbf{w}_2}^{(r)}, \ldots, \Delta\widetilde{\mathbf{w}_k}^{(r)}). \tag{9}$$

Under bounded gradients and adaptive noise, FL-YOLO-DADP converges to a stationary point of the federated loss.

4. Experiments

4.1. Dataset and Setup

We use immunomagnetic cell separation to build a DHM database for calibration. The produced dataset was randomly split ~~into with 80% for training, 10% for validation, and 10% for testing~~for training, validation, and testing in the ratio of 10:1:1.

During local training, the model was evaluated on the validation set at regular intervals to monitor its accuracy and adjust hyperparameters accordingly. Finally, the models were assessed on the test set to eliminate bias. We generated a lightweight dataset for 3 agents, each comprises a training set with 2,400 frames and a validation set with 240 frames, and an ultimate test set with 240 frames for evaluating all models. Each frame, measuring 384x512 pixels, contains up to 30 cells. Benchmark and generative models were trained for 20 epochs in our experiment settings, with specific hyperparameters set to optimize their performance.

For generative models, we simulated the scenario of synthesizing images from all agents locally and send to the central server for YOLOv8n training. In FL-based models, each local YOLOv8n was trained 1 epoch in each round and aggregated for 20 rounds to mimic the case of each client can update the model per round as in the centralized learning environment.

Therefore, in total all 7,200 images were trained for generative models and FL-based models, either by synthesized or federated. The learning rate for both local and centralized YOLOv8n was set to 1e-02, momentum to 0.937, and weight decay to 5e-04, with batch sizes of 32. These hyperparameters were selected based on a combination of default settings and hyperparameter tuning.

All experiments were conducted using Python 3.10.13 and PyTorch 2.0.1+cu118 on an NVIDIA Tesla V100S GPU with 32GB of memory, CUDA version 11.6, to sustain robust and consistent training environments. We used Python packages: Ultralytics YOLOv8.1.34, flwr 1.11.1, for setting up Federated Learning.

4.2. Comparative Experiments

4.2.1 Evaluation Metrics – To evaluate the performance of object detection like YOLO, for local and central cases, we utilized macro-averaged precision, macro-averaged recall, and mAP50 as our primary metrics. Macro-averaged precision averages the precision scores across all classes, giving equal weight to each class. Macro-averaged recall averages the recall scores across all classes.

The mean Average Precision mAP50 at 50% Intersection over Union (IoU) threshold measures the model's accuracy in predicting bounding boxes around objects. The quantity mAP50 considers various IoU thresholds, starting at 0.5. With TP_i and FP_i as true positives and false positives for class i, N as the number of classes, AP_i as the average precision for class I at IoU of 0.5, we calculate the equations as

$$\text{Macro Precision} = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FP_i}\ , \qquad \text{Macro Recall} = \frac{1}{N}\sum_{i=1}^{N}\frac{TP_i}{TP_i + FN_i}\ , \qquad \mathrm{AP}_{50} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{AP}_i\ . \tag{10}$$

4.2.2 Comparative Analysis – We conducted a comparative analysis of various object detection models in our experimental setup. YOLOv8n was chosen for its established performance and efficiency in various object detection tasks. Four generative models were included to explore whether we could find incremental improvements in speed and accuracy for data privacy protection compared with FL framework. FedMedian was selected as a fundamental weighting approach for federation.

All local models were trained from scratch. We evaluated the performance of all different models with the same computational configuration on our customized dataset, summarized in Table 1. This test set contains annotated 240 phase images. The quantitative comparisons of four models on cell classification are in terms of macro-averaged precision, macro-averaged recall, macro-averaged AP50 for all classes, macro-averaged AP_{50} for the erythrocyte class, macro-averaged AP50 for the leukocyte class, macro-averaged AP50 for the platelet class, image size (pixel), and inference time (milliseconds).

Table 1: Comparative experimental results on our customized dataset

Model	**Precision**	**Recall**	**AP_{50}^{val}-all**	**AP_{50}^{val}-RBC**	**AP_{50}^{val}-WBC**	**AP_{50}^{val}-PLT**
Benchmark						
YOLOv8n	0.693	0.85	0.856	0.81	0.959	0.799
Generative Models						
Encoder-decoder	0.358	0.629	0.518	0.207	0.906	0.441
GANs	0.856	0.305	0.327	0.003377	0.871	0.106
DDPM	0.911	0.281	0.315	0.0264	0.813	0.107
Controlnet-img2img	0.383	0.556	0.459	0.445	0.838	0.0928
Federated Learning-Based Models						
FedAvg-YOLO	0.547	0.7	0.672	0.612	0.868	0.536
FedMedian-YOLO	0.555	0.688	0.673	0.605	0.84	0.574
Ours						
FL-YOLO-DADP	**0.375**	**0.501**	**0.48**	**0.323**	**0.718**	**0.399**

For the performance evaluation, we averaged the statistical metrics over the 3 agents. Our proposed FL-YOLO-DADP model achieved comparable accuracy of macro-averaged AP50 with training on small dataset, up to 72% for single class. The accuracy is much higher than generative models like GANs, DDPM, and ControlNet-latent-img2img in mAP50-all, particularly for mAP50-PLT. It improves 46.79% from GANs in mAP50-all, and 276.42% in in mAP50-PLT. Encoder-decoder-generated images got slightly better testing results than ours.

We then applied the model explainability tool Eigen-CAM, based on class activation maps (CAM), focusing on making sense of what a model learns from the visual data to arrive at its predictions. The visualizations of our experimental models are displayed in Fig. 2. It shows Eigen-CAM explains most accurately for YOLOv8n trained on an exemplary original phase image for correctly locating and tightly masking cells, even for small objects as platelets in our case in subfigure 2). In 4), the edges of the map are blurrier than the explanation of the benchmark and vanilla FedMedian-YOLO, which is within our expectations. It shows it retained focus on cell boundaries despite the noise, outperforms the generative models. This suggests that its capability to grasp significant features in the model explainability space is aligned with how humans generally comprehend vision.

To investigate the general effectiveness, the noise multiplier σ and clipping norm C are set as various groups to ensure (ϵ, δ)-DP guarantees. We tracked when $\sigma = \{0.001, 0.0001\}$ and $C = \{1, 10\}$ in initial settings for the noise scale and gradient clipping bound, how would they control the trade-off between utility and privacy. The macro-averaged AP50 is plotted in Fig. 3. Our developed FL-YOLO-DADP is generally outperforming the vanilla FL-YOLO, and the DADP mechanism particularly achieves better convergence performance in FedMedian than in FedAvg. It was noticed of slightly lower mAP50 in our approach with $\sigma = 0.0001$ and $C = 10$ on our test set, meaning adding the least noise as less distortion to the gradients for faster convergence, and preserving the most gradient information for higher sensitivity. It indicates that FL-YOLO-DADP can improve model performance while increasing privacy guarantees.

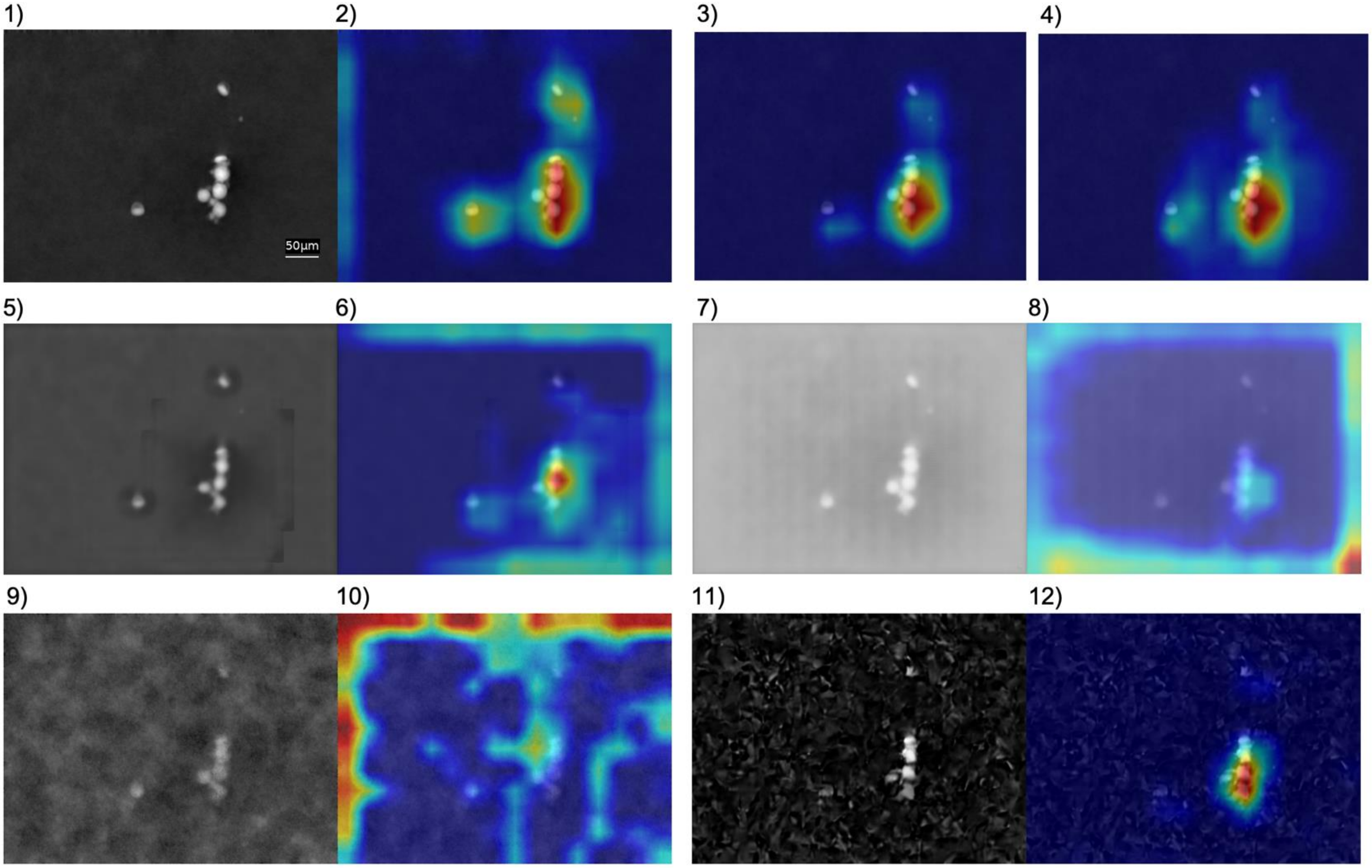

Fig. 2: 1) Original phase image, 2) Eigen-CAM explains YOLOv8n trained on original phase images, 3) Eigen-CAM explains FedMedian-YOLO trained on original phase images, 4) Eigen-CAM explains FL-YOLO-DADP trained on original phase images, 5) Encoder-decoder-generated image, 6) Eigen-CAM explains YOLOv8n trained on encoder-decoder-generated images, 7) GANs-generated image, 8) Eigen-CAM explains YOLOv8n trained on GANs-generated images, 9) DDPM-generated image, 10) Eigen-CAM

explains YOLOv8n trained on DDPM-generated images, 11) ControlNet-latent-img2img-generated image, 12) Eigen-CAM explains YOLOv8n trained on ControlNet-latent-img2img-generated images.

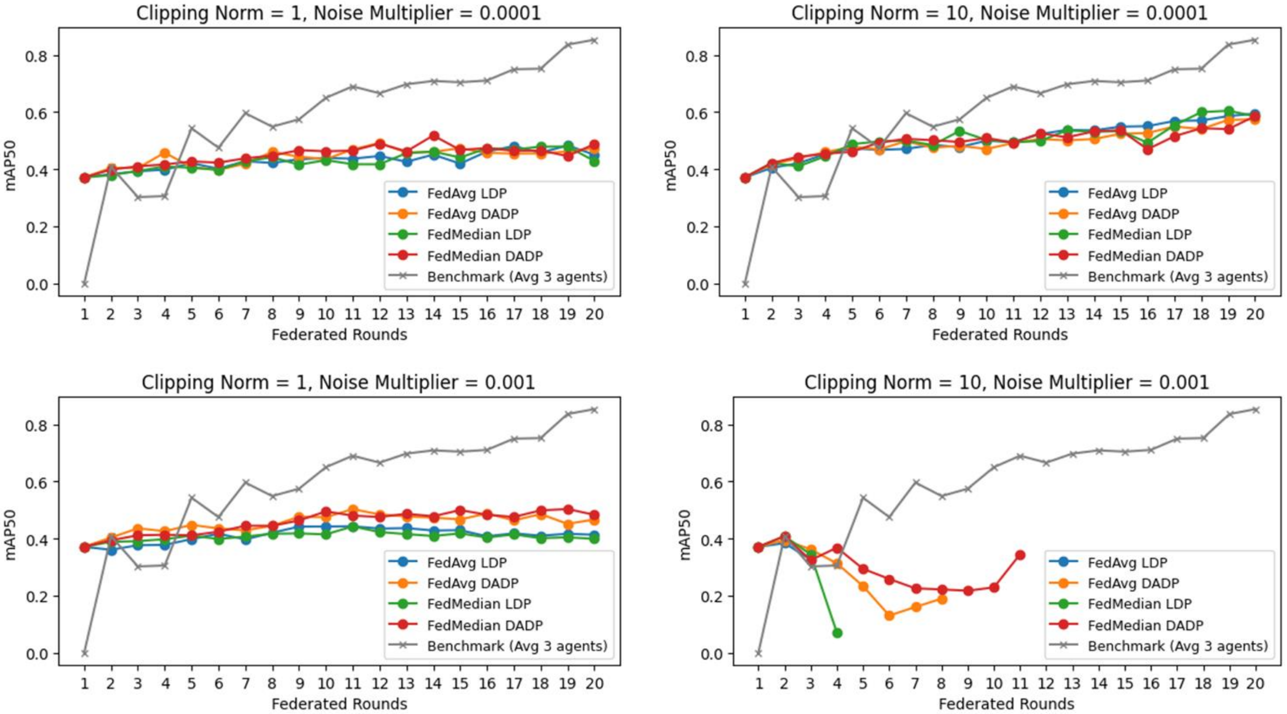

Fig. 3: Eigen-CAM: 1) encoder-decoder, 2) GANs, 3) denoising-diffusion-probabilistic-model, 4) controlNet-latent-img2img model.

5. Conclusion

FL-YOLO-DADP advances privacy-preserving medical imaging by dynamically adapting DP parameters to training phases, achieving a principled balance between data privacy protocol compliance and diagnostic accuracy. On a multi-client digital holographic microscopy dataset, the framework outperforms generative models like GANs by 46.8% in mAP50 crossing all classes and 276.4% in mAP50-PLT, while guaranteeing (ε, δ)-DP. Eigen-CAM visualizations confirm retained sensitivity to fine cellular textures, addressing distribution shifts in small-data regimes.

In future work, we will extend the approach to real-world multi-center study in a hospital environment.

Acknowledgements

This research is supported by the Singapore National Research Foundation (No. NRF2019-THE002-0008). This study is approved by the National Healthcare Group Domain Specific Review Board (Reference numbers 2021/00930 and 2021/01130). We thank Si Ko Myo from TUMCREATE for his contributions to configure the server.

References

[1] R. Corvi, D. Cozzolino, G. Zingarini, G. Poggi, K. Nagano, and L. Verdoliva, “On The Detection of Synthetic Images Generated by Diffusion Models,” in *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Rhodes Island, Greece: IEEE, Jun. 2023, pp. 1–5. doi: 10.1109/ICASSP49357.2023.10095167.

[2] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative Adversarial Networks,” 2014, *arXiv*. doi: 10.48550/ARXIV.1406.2661.

[3] D. G. Saragih, A. Hibi, and P. N. Tyrrell, “Using diffusion models to generate synthetic labeled data for medical image segmentation,” *Int. J. Comput. Assist. Radiol. Surg.*, vol. 19, no. 8, pp. 1615–1625, Jun. 2024, doi: 10.1007/s11548-024-03213-z.

[4] Z. Huang, Q. Yang, M. Tian, and Y. Gao, “Synthesizing with Diffusion model for improving medical image segmentation performance,” in *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Istanbul, Turkiye: IEEE, Dec. 2023, pp. 1977–1982. doi: 10.1109/BIBM58861.2023.10385456.

[5] J. Wu, W. Ren, X. Zhang, and X. Zheng, “An Encoder-Based Framework for Privacy-Preserving Machine Learning,” in *Algorithms and Architectures for Parallel Processing*, vol. 15256, T. Zhu, J. Li, and A. Castiglione, Eds., in Lecture Notes in Computer Science, vol. 15256. , Singapore: Springer Nature Singapore, 2025, pp. 37–46. doi: 10.1007/978-981-96-1551-3_4.

[6] H. Guan and M. Liu, “Domain Adaptation for Medical Image Analysis: A Survey,” *IEEE Trans. Biomed. Eng.*, vol. 69, no. 3, pp. 1173–1185, Mar. 2022, doi: 10.1109/TBME.2021.3117407.

[7] H. Shibata, S. Hanaoka, T. Nakao, T. Kikuchi, Y. Nakamura, Y. Nomura, T. Yoshikawa, and O. Abe, “Practical Medical Image Generation with Provable Privacy Protection Based on Denoising Diffusion Probabilistic Models for High-Resolution Volumetric Images,” *Appl. Sci.*, vol. 14, no. 8, p. 3489, Apr. 2024, doi: 10.3390/app14083489.

[8] Y. Wang, H. Xu, X. Zhang, Z. Chen, Z. Sha, Z. Wang, and Z. Tu, “OmniControlNet: Dual-stage Integration for Conditional Image Generation,” in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, Seattle, WA, USA: IEEE, Jun. 2024, pp. 7436–7448. doi: 10.1109/CVPRW63382.2024.00739.

[9] G. Kaissis, A. Ziller, J. Passerat-Palmbach, T. Ryffel, D. Usynin, A. Trask, I. Lima, J. Mancuso, M. Jungmann, M. Steinborn, A. Saleh, M. Makowski, D. Rueckert, and R. Braren, “End-to-end privacy preserving deep learning on multi-institutional medical imaging,” *Nat. Mach. Intell.*, vol. 3, no. 6, pp. 473–484, May 2021, doi: 10.1038/s42256-021-00337-8.

[10]N. Koutsoubis, Y. Yilmaz, R. P. Ramachandran, M. Schabath, and G. Rasool, “Privacy Preserving Federated Learning in Medical Imaging with Uncertainty Estimation,” 2024, *arXiv*. doi: 10.48550/ARXIV.2406.12815.

[11]F. A. Hölzl, D. Rueckert, and G. Kaissis, “Bridging the Gap: Differentially Private Equivariant Deep Learning for Medical Image Analysis,” 2022, *arXiv*. doi: 10.48550/ARXIV.2209.04338.

[12]M. Dombrowski, H. Reynaud, M. Baugh, and B. Kainz, “Foreground-Background Separation through Concept Distillation from Generative Image Foundation Models,” in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France: IEEE, Oct. 2023, pp. 988–998. doi: 10.1109/ICCV51070.2023.00097.

[13]I. Ktena, O. Wiles, I. Albuquerque, S. Rebuffi, R. Tanno, A. G. Roy, S. Azizi, D. Belgrave, P. Kohli, T. Cemgil, A. Karthikesalingam, and S. Gowal, “Generative models improve fairness of medical classifiers under distribution shifts,” *Nat. Med.*, vol. 30, no. 4, pp. 1166–1173, Apr. 2024, doi: 10.1038/s41591-024-02838-6.

[14]Jocher, G., Qiu, J., & Chaurasia, A. (2023). Ultralytics YOLO (Version 8.0.0) [Computer software]. https://github.com/ultralytics/ultralytics

[15] Q. Chen, M. E. Cove, K. Delikoyun, K. Diepold, O. Hayden, W. S. Kuan, W. Liu, and J. Soong, “AI-Enhanced Detection of Cellular Aggregate Biomarkers For Point-of-Care Using Digital Holographic Microscopy,” in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Lisbon, Portugal: IEEE, Dec. 2024, pp. 5610–5615. doi: 10.1109/BIBM62325.2024.10821749.

[16]O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional Networks for Biomedical Image Segmentation,” in *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015*, vol. 9351, N. Navab, J. Hornegger, W. M. Wells, and A. F. Frangi, Eds., in Lecture Notes in Computer Science, vol. 9351. , Cham: Springer International Publishing, 2015, pp. 234–241. doi: 10.1007/978-3-319-24574-4_28.

[17]Q. Dai, D. Wei, H. Liu, J. Sun, L. Wang, and Y. Zheng, “Federated Modality-Specific Encoders and Multimodal Anchors for Personalized Brain Tumor Segmentation,” *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 2, pp. 1445–1453, Mar. 2024, doi: 10.1609/aaai.v38i2.27909.

[18]T. Rahman, A. Bilgin, and S. D. Cabrera, “Asymmetric decoder design for efficient convolutional encoder-decoder architectures in medical image reconstruction,” in *Multimodal Biomedical Imaging XVII*, F. S. Azar, X. Intes, and Q. Fang, Eds., San Francisco, United States: SPIE, Mar. 2022, p. 19. doi: 10.1117/12.2610084.

[19]P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, “Image-to-Image Translation with Conditional Adversarial Networks,” in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI: IEEE, Jul. 2017, pp. 5967–5976. doi: 10.1109/CVPR.2017.632.

[20] J. Shim, E. Kim, H. Kim, and E. Hwang, “Enhancing Image Representation in Conditional Image Synthesis,” in *2023 IEEE International Conference on Big Data and Smart Computing (BigComp)*, Jeju, Korea, Republic of: IEEE, Feb. 2023, pp. 203–210. doi: 10.1109/BigComp57234.2023.00041.

[21] H. Yang and P. Qian, “GAN-Based Medical Images Synthesis: A Review,” *Int. J. Health Syst. Transl. Med.*, vol. 1, no. 2, pp. 1–9, Jul. 2021, doi: 10.4018/IJHSTM.2021070101.

[22] J. Ho, A. Jain, and P. Abbeel, “Denoising Diffusion Probabilistic Models,” 2020, *arXiv*. doi: 10.48550/ARXIV.2006.11239.

[23] L. Zhang, A. Rao, and M. Agrawala, “Adding Conditional Control to Text-to-Image Diffusion Models,” in *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France: IEEE, Oct. 2023, pp. 3813–3824. doi: 10.1109/ICCV51070.2023.00355.

[24] M. Li, T. Yang, H. Kuang, J. Wu, Z. Wang, X. Xiao, and C. Chen, “ControlNet$$++$$: Improving Conditional Controls with Efficient Consistency Feedback: Project Page: liming-ai.github.io/ControlNet_Plus_Plus,” in *Computer Vision – ECCV 2024*, vol. 15065, A. Leonardis, E. Ricci, S. Roth, O. Russakovsky, T. Sattler, and G. Varol, Eds., in Lecture Notes in Computer Science, vol. 15065. , Cham: Springer Nature Switzerland, 2025, pp. 129–147. doi: 10.1007/978-3-031-72667-5_8.

[25] J. Zhang, J. Zhou, J. Guo, and X. Sun, “Visual Object Detection for Privacy-Preserving Federated Learning,” *IEEE Access*, vol. 11, pp. 33324–33335, 2023, doi: 10.1109/ACCESS.2023.3263533.

[26] M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang, “Deep Learning with Differential Privacy,” in *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, Vienna Austria: ACM, Oct. 2016, pp. 308–318. doi: 10.1145/2976749.2978318.