Data Poisoning Attack Defense and Evolutionary Domain Adaptation for Federated Medical Image Segmentation (Appendix)

The appendix of this study provides comprehensive details that support the main framework, methodology, and experimental results presented in the paper. Below is a summary of each section: Section A provides a detailed explanation of the core algorithms of AdaShield-FL, including an overview of the framework, malicious client purification, evolutionary domain adaptation, and attack detection and aggregation. Section B presents additional quantitative and qualitative results for segmentation, disentangling encoder, independence loss, and ablation studies. Section C presents qualitative segmentation results of AdaShield-FL and other methods. Section D describes how external attackers inject adversarial perturbations into clients' local datasets to degrade the global model's performance. Section E provides the limitations of the study and potential directions for future work.

A Method Algorithms

A.1 Overview

Algorithm 1 provides a comprehensive overview of the AdaShield-FL framework, encompassing key components such as malicious client purification, evolutionary domain adaptation, and attack detection and aggregation.

A.2 Malicious Client Purification

Algorithm 2 illustrates the purification process for mitigating the influence of malicious clients by refining their local datasets, ensuring the patient data diversity.

A.3 Evolutionary Domain Adaptation

Algorithm 3 represents the evolutionary domain adaptation process, which enhances model robustness across diverse and heterogeneous client data distributions.

A.4 Attack Detection and Aggregation

Algorithm 4 shows the process of identifying malicious clients and aggregating model weights and covariance matrices.

Algorithm 1 AdaShield-FL Overall Algorithm

Input: Number of clients N, local datasets I_i, global model weights in r-th round θ^r, malicious status λ_i^r;
Output: Final global model weights θ^R;

for each round $r = 1$ to R **do**
  for each client $i \in \{1, \ldots, N\}$ **do**
    Download θ_{global}^r, C_{global}^r, λ_i^r
    if $\lambda_i^r = 1$ **then**
      $I_i^{'} \leftarrow$ **Malicious Client Purification** (I_i, θ_{global}^r)
      $(\theta_i^{r+1}, C_i^{r+1}, L_i^{r+1}) \leftarrow$ **Evolutionary Domain Adaptation** $(I_i^{'}, \theta_i^r, C_{global}^r)$
    else
      $(\theta_i^{r+1}, C_i^{r+1}, L_i^{r+1}) \leftarrow$ **Evolutionary Domain Adaptation** $(I_i, \theta_i^r, C_{global}^r)$
    end if
    Upload θ_i^{r+1}, C_i^{r+1}, L_i^{r+1} to the server
  end for
  $(\theta_{global}^{r+1}, C_{global}^{r+1}, \lambda_{1,2,\ldots,N}^{r+1}) \leftarrow$ **Attack Detection and Aggregation** $(\theta_{1,2,\ldots,N}^{r+1}, C_{1,2,\ldots,N}^{r+1}, L_{1,2,\ldots,N}^{r+1})$
  Broadcast θ_{global}^{r+1}, C_{global}^{r+1}, and $\lambda_{1,2,\ldots,N}^{r+1}$ to all clients
end for
return Final global model weights θ^R

Algorithm 2 Malicious Client Purification

Input: Local data I_i, global model weights θ_{global};
Output: Purified data I_i';

Initialize DRS using θ_{global}
Input I_i into the DRS
for each iteration $l \in \{0, \ldots, L-1\}$ **do**
  Transform l-th purification data I_i^l to k-space using FFT:

$$s(k_x, k_y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} I(x, y) e^{-i2\pi(xk_x + yk_y)} dx\, dy$$

  Extract f_R and f_S using disentangling encoder in the k-space
  Transform f_R and f_S to $f_R^{'}$ and $f_S^{'}$ using IFFT
  Obtain $(l+1)$-th purification data I_i^{l+1} using $f_R^{'}$ and reconstruction decoder
end for
$I_i' \leftarrow I_i^L$
return I_i'

Algorithm 3 Evolutionary Domain Adaptation

Input: Local dataset I_i, global covariance matrix C^r_{global}, global model weights θ^r_{global};
Output: Updated model weights θ^{r+1}_i, covariance matrix C^{r+1}_i, and training loss sequence L^{r+1}_i;

Initialize θ^r_{global} to segmentation model
for each training iteration $t \in \{1, \ldots, T\}$ **do**
  Define total loss $\mathcal{L}_{total}$:

$$\mathcal{L}_{total} = (1 - \beta^t)\mathcal{L}_{dice} + \beta^t \mathcal{L}_{cm}$$

  for each generation $g \in \{0, \ldots, G-1\}$ **do**
    Sample candidate solutions using:

$$x^{g+1}_k = m^g + \sigma^g \cdot \mathcal{N}(0, \mathbf{\Sigma}^g), \quad k = 1, \ldots, 20$$

    Compute total loss and select top-10 candidates using $\mathcal{L}_{total}(x^{g+1}_k)$:
    Update mean m^{g+1} using top-10 candidates.
    if $\|m^{g+1} - m^g\| \leq \epsilon_{cma}$ **then**
      Set $\beta^t = m^{g+1}$ and stop evolution
      Update segmentation model using $\mathcal{L}_{total}$
    end if
    Update covariance matrix $\mathbf{\Sigma}^{g+1}$ using top-10 candidates
    Sample new candidate solutions with $\mathbf{\Sigma}^{g+1}$ and m^{g+1}
  end for
end for
Obtain final segmentation model weight θ^{r+1}_i
Compute covariance matrix C^{r+1}_i using disentangled segmentation features f_S
Compute training dice loss sequence L^{r+1}_i
return θ^{r+1}_i, C^{r+1}_i, L^{r+1}_i

Algorithm 4 Attack Detection and Aggregation

Input: Training dice loss sequence L^r_i, local model weights $\theta^r_{1,2,\ldots,N}$, local covariance matrices $C^r_{1,2,\ldots,N}$;
Output: Updated global model weights θ^r, attack status $\lambda^r_{1,\ldots,N}$;

Extract gradient latent code $\Delta L^r_{1,2,\ldots,N}$
Extract curvature latent code $\Delta^2 L^r_{1,2,\ldots,N}$
Concatenate differential loss patterns
Obtain the malicious status $\lambda^r_{1,2,\ldots,N}$ by applying anomaly detection to the concatenated differential loss patterns.
Aggregate global model weights θ^r using $\theta^r_{1,2,\ldots,N}$ and $\lambda^r_{1,2,\ldots,N}$:

$$\theta^r \leftarrow \theta^{r-1} - \eta \frac{1}{N - \sum_{i=1}^N \lambda^r_i} \sum_{i=1}^N \frac{g^{r-1}_i \cdot (1 - \lambda^r_i)}{\sqrt{G^{r-1}} + \epsilon}$$

Aggregate global covariance matrix C^r_{global} using $C^r_{1,2,\ldots,N}$ and $\lambda^r_{1,2,\ldots,N}$:

$$C^r_{global} = \frac{1}{\sum_{i=1}^N \rho_i (1 - \lambda^r_i)} \Bigg[\sum_{i=1}^N \rho_i C^r_i (1 - \lambda^r_i) + \sum_{i=1}^N \rho_i (\mu^r_i - \mu^r)(\mu^r_i - \mu^r)^T (1 - \lambda^r_i) \Bigg]$$

return Malicious status $\lambda^r_{1,\ldots,N}$, global covariance C^r_{global}, and updated weights θ^r to clients

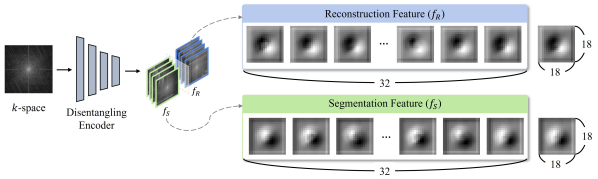

Figure 1: Visualization of disentangled reconstruction feature (f_R) and segmentation feature (f_S) extracted our disentangling encoder in k-space.

	Dice Score			
	$L = 1$	$L = 2$	$L = 3$	$L = 4$
AdaShield-FL	78.2	80.3	**82.7**	81.1

Table 1: Impact of iterations (L) on dice score in DRS on the M&Ms dataset under the FGSM attack.

B Additional Experiments

B.1 Disentangling Encoder

In the DRS module, a disentangling encoder [Zhou *et al.*, 2021] is composed of four residual blocks, with each block consisting of two convolutional layers, a batch normalization layer, and a Leaky ReLU activation function. In our framework, the reconstruction and segmentation decoder mirror this structure symmetrically; however, its input vector is half the size of the encoder output.

Figure 1 illustrates the disentangled features obtained from k-space data through a disentangling encoder. It shows that the disentangled features exhibit distinct spatial distributions with the reconstruction feature (f_R) focusing on specific regions, while the segmentation feature (f_S) concentrates on distinct areas, indicating that the two features are directed toward different parts of k-space data. This directional separation indicates that the encoder effectively disentangles the features by learning task-specific representations tailored for reconstruction and segmentation. This separation highlights the effectiveness of feature disentanglement in facilitating distinct objectives in medical imaging.

B.2 the Number Iterations L

In addition, Table 1 demonstrates the impact of the number of iterations (L) on the dice score in the DRS module, evaluated on the M&Ms [Campello *et al.*, 2021] dataset under a fast gradient sign method (FGSM) [Goodfellow *et al.*, 2014] attack. AdaShield-FL achieves its optimal performance when L is set to 3, providing the best balance between reconstruction and segmentation accuracy. In Fig. 2, as L increases, the segmentation maps progressively improve, with false-negative cases decreasing significantly. Early iterations (L=0 and L=1) show notable incorrect segmentation in regions such as the left ventricle and myocardium. However, as the purification process continues, these false-negative cases are gradually corrected, leading to segmentation maps that closely align with the ground truth. Notably, when $L = 3$, the segmentation map closely resembles the ground truth, indicating that three iterations yield the most balanced and accurate results in terms of segmentation quality. This result

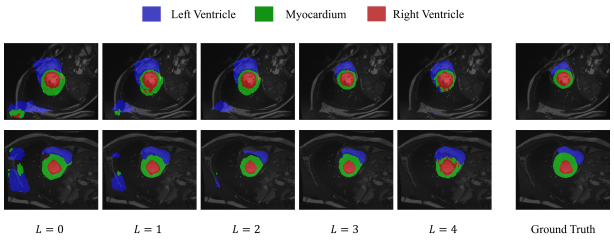

Figure 2: Qualitative examples of the segmentation map during L iterations in disentangled reconstruction and segmentation on ACDC dataset perturbed by PGD in "hypertrophic cardiomyopathy" patients.

Model	Dice Score		
	FGSM	PGD	C&W
FedAvg [McMahan *et al.*, 2017]	55.3	54.8	55.1
[Yi *et al.*, 2024]	61.7	63.1	62.5
IOS [Wu *et al.*, 2023]	62.3	61.5	60.9
[Karimireddy *et al.*, 2021]	61.1	60.8	61.6
[Li *et al.*, 2020]	60.3	60.7	60.9
FedRBN [Hong *et al.*, 2023]	61.2	62.3	61.0
AdaShield-FL	**73.1**	**72.8**	**73.5**

Table 2: Comparison with prior methods on the M&Ms validation set in terms of dice score. Each FL method trained with **3 clients** on the perturbed data generated by each adversarial attack.

Model	Dice Score		
	FGSM	PGD	C&W
FedAvg [McMahan *et al.*, 2017]	47.3	47.1	46.9
[Yi *et al.*, 2024]	50.3	52.8	51.7
IOS [Wu *et al.*, 2023]	52.0	51.9	51.3
[Karimireddy *et al.*, 2021]	51.3	51.7	51.1
[Li *et al.*, 2020]	50.9	51.3	50.4
FedRBN [Hong *et al.*, 2023]	50.1	51.7	50.3
AdaShield-FL	**61.3**	**63.7**	**61.9**

Table 3: Comparison with prior methods on the M&Ms validation set in terms of dice score. Each FL method trained with **4 clients** on the perturbed data generated by each adversarial attack.

underscores the appropriateness of using three iterations to achieve optimal segmentation performance while maintaining robustness under adversarial attacks.

B.3 Number of Malicious Clients

Tables 2 and 3 demonstrate the performance of AdaShield-FL with 3 and 4 malicious clients on the M&Ms dataset. This analysis specifically measures the impact of increasing malicious clients on model accuracy, providing insights into the effectiveness of purification and detection mechanisms. Despite the increase in the number of malicious clients, AdaShield-FL shows an ability to maintain performance stability, which is a critical indicator of the model's robustness against data poisoning attacks in federated learning (FL) environments.

Feature Variation	Adversarial Attacks		
	FGSM	PGD	C&W
Δf_R	3.182	2.674	2.945
Δf_R (w/o $\mathcal{L}_{ind}$)	4.866	5.131	5.298
Δf_S	9.233	8.415	8.193
Δf_S (w/o $\mathcal{L}_{ind}$)	7.162	7.064	6.883

Table 4: MAE between features extracted from clean and perturbed data, with and without the **independence loss** $\mathcal{L}_{ind}$ on the M&Ms dataset.

Model	Dice Score		
	FGSM	PGD	C&W
FedAvg [McMahan *et al.*, 2017]	51.2	50.3	50.6
[Yi *et al.*, 2024]	60.2	60.3	60.6
IOS [Wu *et al.*, 2023]	59.3	60.1	59.8
[Karimireddy *et al.*, 2021]	59.5	59.9	60.5
[Li *et al.*, 2020]	61.3	60.7	60.0
FedRBN [Hong *et al.*, 2023]	59.7	61.2	59.8
AdaShield-FL	**68.3**	**68.2**	**68.7**

Table 5: Comparison with prior methods on the **CTICH dataset** in terms of dice score. Each FL method trained on the perturbed data generated by each adversarial attack.

B.4 Independence Loss

Table 4 shows the mean absolute error (MAE) between two features, f_R and f_S, extracted from clean and perturbed data, both with and without the $\mathcal{L}_{ind}$ loss term. The results demonstrate that incorporating $\mathcal{L}_{ind}$ reduces perturbations affecting f_R, indicating enhanced adversarial robustness. This underscores the effectiveness of $\mathcal{L}_{ind}$ in improving model robustness by minimizing feature variation in f_R while directing perturbations toward f_S.

B.5 CT Image Segmentation for other Modality

In Table 5, we utilized the "Computed Tomography Images for Intracranial Haemorrhage Detection and Segmentation" (CTICH) dataset [Hssayeni *et al.*, 2020], which is publicly available on PhysioNet. The dataset comprises 82 computed tomography (CT) scans from patients with traumatic brain injuries, including 36 scans where intracranial hemorrhage was diagnosed. Each CT scan consists of approximately 30 slices per patient. The average age of the patients is 27.8 years with a standard deviation of 19.5 years; 46 patients are male, and 36 are female. Moreover, Table 5 demonstrates the superior performance of AdaShield-FL compared to other FL methods on the CTICH dataset, a different modality involving CT images. The results indicate that AdaShield-FL consistently achieves superior Dice scores across all adversarial attacks, including FGSM [Goodfellow *et al.*, 2014], PGD [Madry *et al.*, 2017], and C&W [Carlini and Wagner, 2017], highlighting its robustness and generalizability to diverse medical imaging modalities. This result underscores the adaptability and broad applicability of AdaShield-FL to various imaging modalities beyond the original training domain.

B.6 Ablation Study

Table 6 presents an additional ablation study of the individual contributions of each module in the AdaShield-FL frame-

Differential Loss-based Attack Detection	Evolutionary DA	Malicious Client Purification	Dice Score
✓	✓	✓	**82.6**
✓	✓		74.9
✓		✓	81.3
	✓		71.8
✓			75.5
			67.8

Table 6: **Additional ablation study** for AdaShield-FL on the ACDC dataset perturbed by PGD attack in terms of dice coefficient score.

Differential Loss-based Attack Detection		Precision			Recall		
First-order Derivative Extractor	Second-order Derivative Extractor	FGSM	PGD	C&W	FGSM	PGD	C&W
✓		0.88	0.85	0.89	0.84	0.82	0.85
	✓	0.82	0.83	0.88	0.90	0.83	0.81
✓	✓	**0.98**	**0.99**	**0.98**	**0.95**	**0.96**	**0.94**

Table 7: Ablation study for differential loss-based attack detection performance in terms of recall and precision for detecting malicious clients on the M&Ms datasets perturbed by each adversarial attack.

work. The results demonstrate that each module significantly enhances the overall performance of AdaShield-FL. Furthermore, the combination of these modules operates synergistically to maximize segmentation performance, effectively addressing challenges posed by data poisoning attacks and data heterogeneity.

In addition, Table 7 compares the precision and recall of the differential loss-based attack detection framework under different adversarial attack scenarios. The results are presented for two configurations of the differential loss-based attack detection module: one using the first-order derivative extractor and the other combining the first-order and second-order derivative extractors. The configuration that utilizes both first-order and second-order derivative extractors demonstrates superior performance across all metrics. In contrast, the first-order-only configuration exhibits lower performance, with precision and recall. These results indicate that incorporating second-order derivatives enhances the robustness and accuracy of AdaShield-FL's attack detection mechanism. Moreover, these results highlight that abnormal patterns such as oscillations and slow convergence speed are more likely to be caused by malicious clients than by the diversity of patients, vendors, and diseases.

C Visual Segmentation Results

Figures 3−6 provide qualitative segmentation results of segmentation results achieved by AdaShield-FL under diverse scenarios, highlighting its robustness and accuracy across different patient conditions.

Specifically, Figure 3 depicts segmentation results for patients with "hypertrophic cardiomyopathy". The results of AdaShield-FL exhibit exceptional segmentation, similar to the ground truth, accurately capturing the thickness of myocardial regions without predicting false-positive segmentation in unrelated areas. This precise segmentation facilitates the accurate diagnosis of hypertrophic cardiomyopathy, a disease characterized by the thickening of the myocardium. In comparison, [Li *et al.*, 2020] and [Karimireddy *et al.*, 2021] may lead to incorrect diagnoses as a result of inaccurate segmentation of the myocardial region.

Figure 4 also presents cases of "normal" cardiac patients. AdaShield-FL provides clean segmentation maps, preserving anatomical boundaries with high precision. FedAvg and [McMahan *et al.*, 2017] often provide artifacts from noise or fail to clearly delineate boundaries, resulting in false positive regions and [Yi *et al.*, 2024] also provide incorrect right ventricle regions, potentially diagnosing a normal patient as a patient who has another disease. AdaShield-FL avoids these issues to maintain robust performance.

Moreover, Figure 5 illustrates segmentation results for patients with an "abnormal right ventricle," characterized by significant enlargement of the right ventricle. Unlike other methods, AdaShield-FL accurately captures the thickened right ventricle and the thinned myocardium regions. IOS [Wu *et al.*, 2023] and [Li *et al.*, 2020] are particularly inaccurate in predicting the right ventricle and myocardial regions, making it challenging to assess the thickness of these regions, which is a key characteristic of the disease. As a result, this method leads to inaccurate diagnoses, unlike AdaShield-FL.

Furthermore, Figure 6 illustrates segmentation maps from "dilated cardiomyopathy" patients, where the left ventricle is significantly enlarged. AdaShield-FL accurately provides the left ventricle without false-negative segmentation, unlike other methods. For instance, methods like FedRBN [Hong *et al.*, 2023] or IOS [Wu *et al.*, 2023] often struggle with false-negative segmentation of the left ventricle, incorrectly predicting part of the ventricle as background. This incorrect prediction of left ventricle regions has potentially diagnosed a patient with "dilated cardiomyopathy" as a normal patient. In contrast, AdaShield-FL ensures more reliable patient diagnoses.

Additionally, as shown in the Fig 7, we provide qualitative segmentation results for an additional modality. We note that CT images are transformed into the frequency domain using fast Fourier transform (FFT), where the transformed frequency domain corresponds to the k-space in magnetic resonance imaging (MRI). This domain representation allows AdaShield-FL to effectively separate the data, applying the same mechanisms as those used for k-space. Figure 7 illustrates segmentation results on the CTICH dataset for "traumatic brain injury" patients perturbed by FGSM. AdaShield-FL achieves accurate segmentation of intracranial hemorrhage regions, closely matching the ground truth, even under adversarial perturbations. FedAvg [McMahan *et al.*, 2017], [Yi *et al.*, 2024] and IOS [Wu *et al.*, 2023], which fail to consistently identify hemorrhage regions and often misclassify critical areas as background or noise. Moreover, FedRBN [Hong *et al.*, 2023] often makes incorrect predictions or misses parts of the hemorrhage, potentially leading to diagnostic errors. In contrast, AdaShield-FL avoids false positives and negatives by focusing on key hemorrhage regions, demonstrating robust segmentation performance and reliability in challenging adversarial conditions. This ensures greater accuracy and dependability for medical diagnoses in clinical scenarios involving traumatic brain injuries.

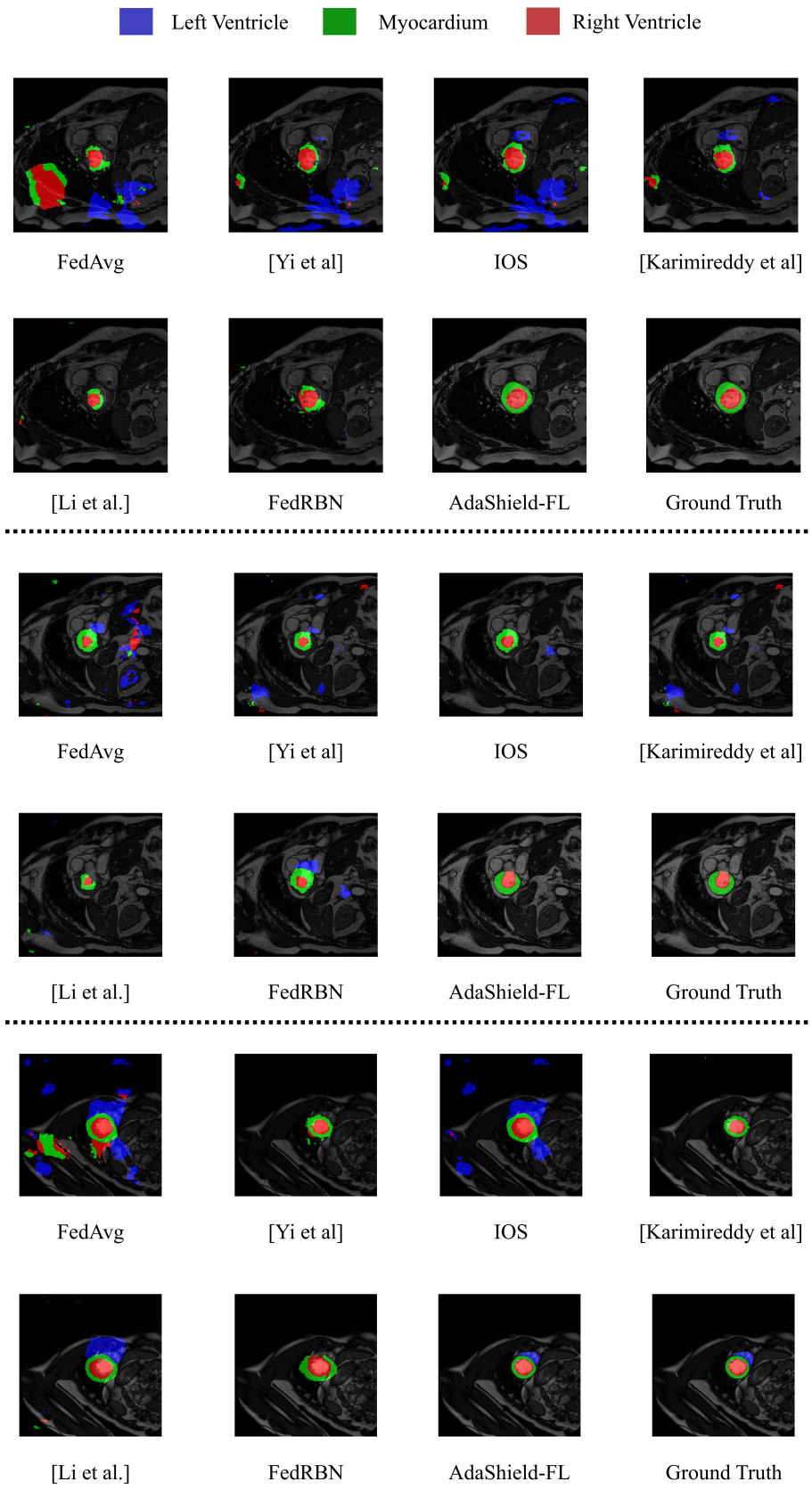

Figure 3: Qualitative segmentation results on M&Ms dataset perturbed by FGSM in **"hypertrophic cardiomyopathy"** patients.

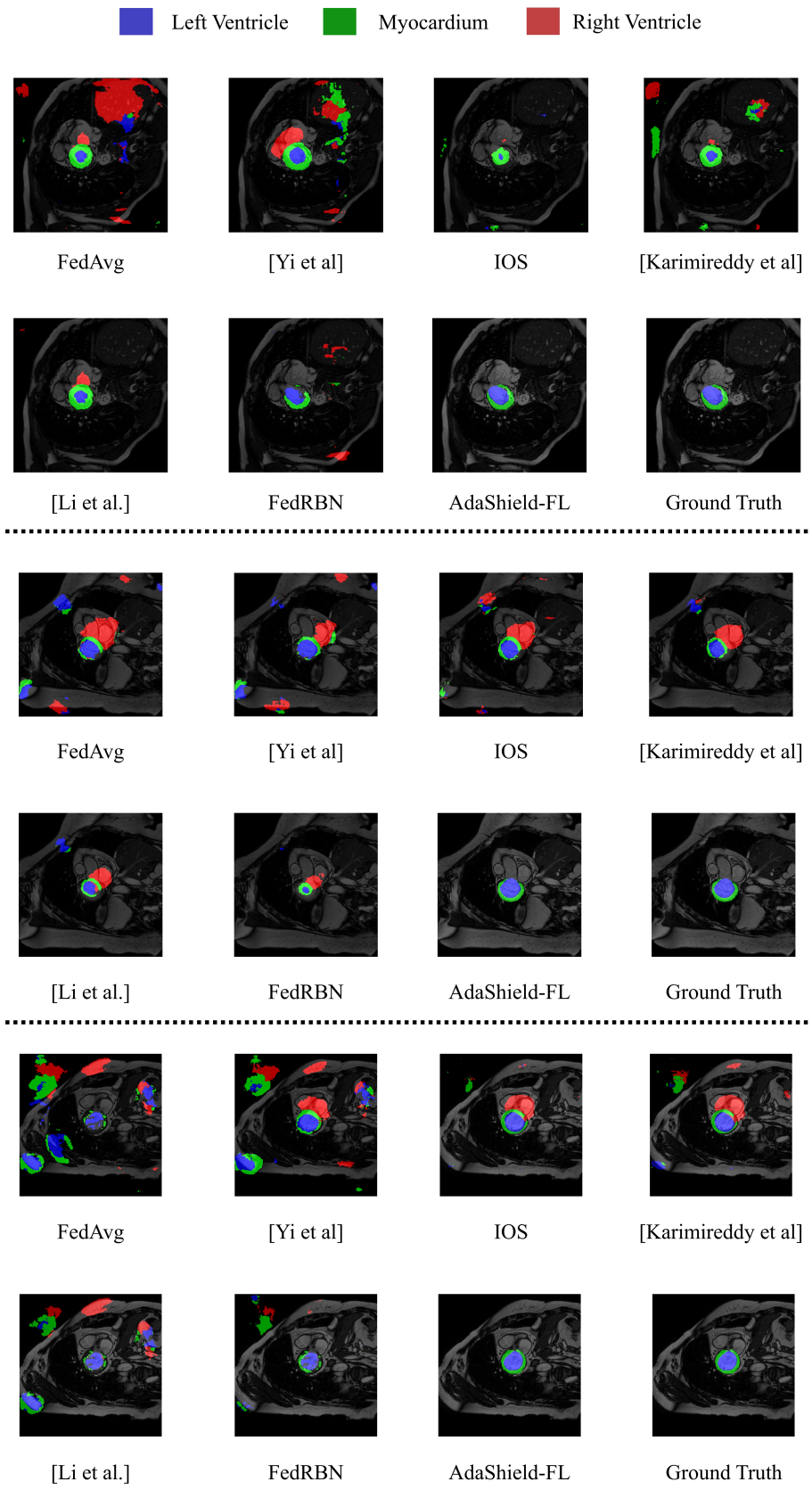

Figure 4: Qualitative segmentation results on M&Ms dataset perturbed by PGD in **"normal"** patients.

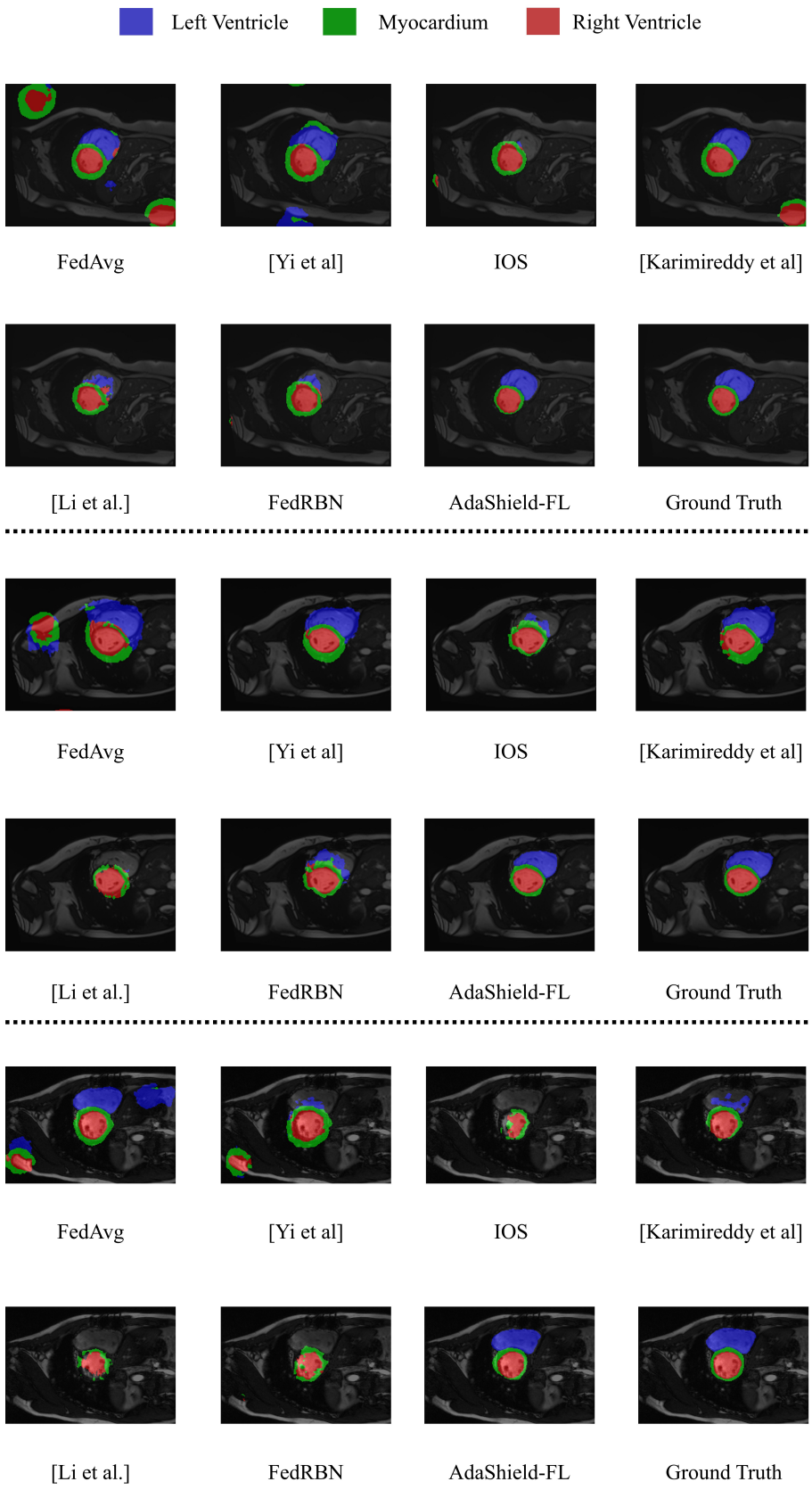

Figure 5: Qualitative segmentation results on ACDC dataset perturbed by C&W in **"abnormal right ventricle"** patients.

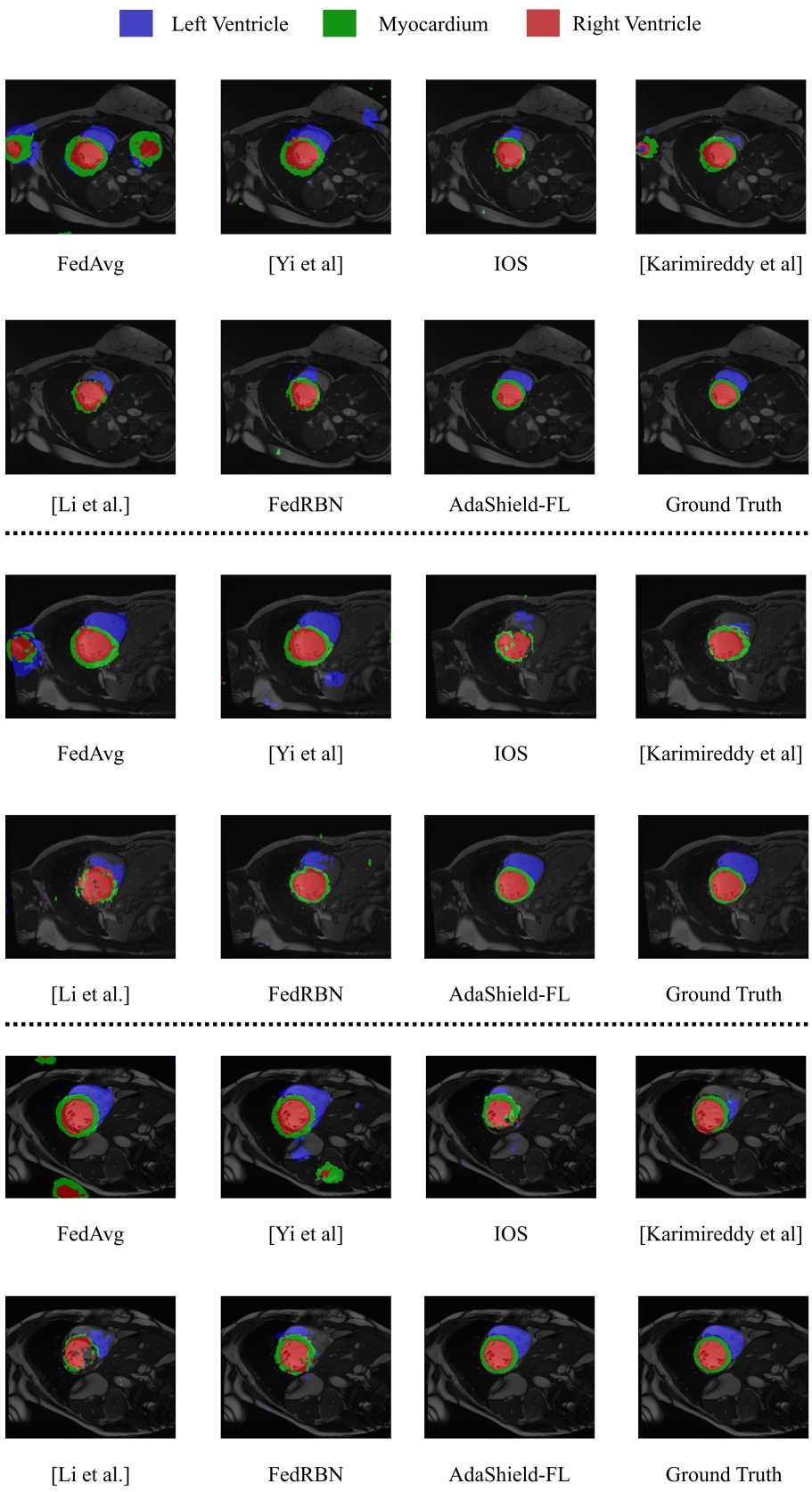

Figure 6: Qualitative segmentation results on ACDC dataset perturbed by PGD in **"dilated cardiomyopathy"** patients.

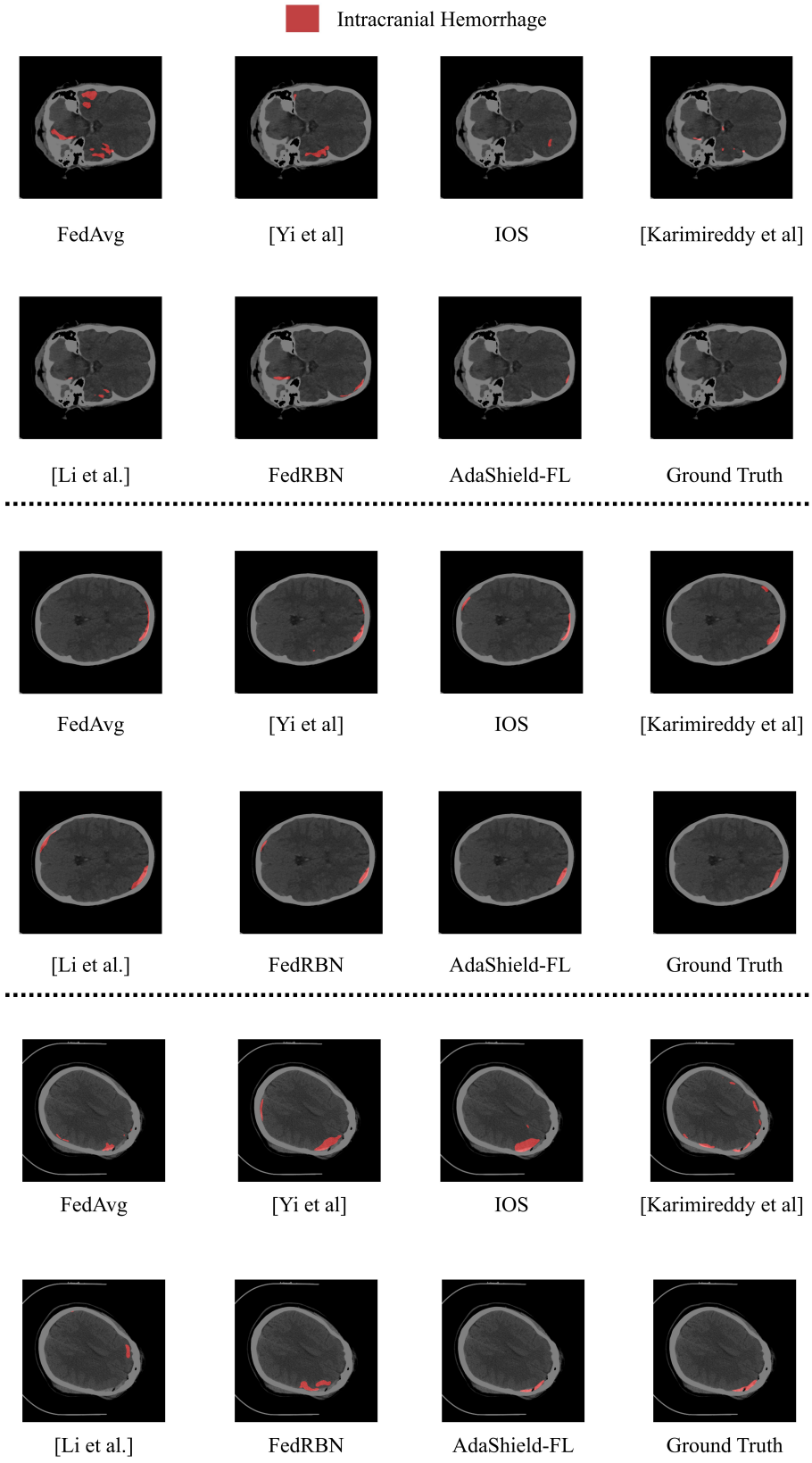

Figure 7: Qualitative segmentation results on **CTICH dataset** perturbed by FGSM in **"traumatic brain injury"** patients.

D Adversarial Scenario

D.1 Attacker's Goal

The external attacker generates adversarial examples by applying small, imperceptible perturbations to input images. These perturbations are generated using techniques such as FGSM [Goodfellow *et al.*, 2014], projected gradient descent (PGD) [Madry *et al.*, 2017], or Carlini and Wagner attack (C&W) [Carlini and Wagner, 2017], and are intended to result in incorrect segmentation maps. Moreover, the attacker exploits knowledge of the model's parameters and gradients (assuming they have partial access to these) to optimize the perturbations for maximum impact. The result of these attacks is a reduced segmentation performance of the global model, which directly affects its clinical utility. In medical applications, such a degradation can led to failure in identifying critical conditions, delayed treatment, or even financial fraud in insurance claims.

D.2 Attacker's Capability

The white-box scenario assumes that the attacker has unauthorized access to both clinical data and model parameters of certain clients. This access allows the attacker to inject data poisoning attacks. Specifically, the attacker can inject poisoned data, either by modifying input images or by manipulating output labels. This poisoned data corrupts the local training process and, when aggregated, negatively impacts the global model. The adversarial scenario assumes that the attacker can compromise more than 25% of the total client number.

D.3 Attacker's Background Knowledge

This work distinguishes between two types of clients. Benign clients are uncompromised clients that contribute clean data and trustworthy updates to the global model. They represent the majority of clients in the FL system and are crucial for maintaining the model's integrity. Malicious clients have been manipulated by the external attacker and are responsible for injecting poisoned data or adversarial examples into the training process. The attacker controls these clients and uses them to execute attacks without being directly involved in the FL process. The attacker has access to the local datasets of compromised clients, which are used to generate poisoned data or adversarial examples. Since medical datasets are typically highly sensitive and diverse, this access poses a significant threat to the integrity of the global model.

E Limitation and Future Work

E.1 Parameter Reduction in Reconstruction Decoder

The reconstruction decoder in AdaShield-FL, while effective, involves a slightly higher number of parameters. Future work will focus on reducing the parameter count in the reconstruction decoder by employing a knowledge distillation approach. Specifically, the global reconstruction decoder will serve as the teacher model, while the student model will be designed as a lightweight local reconstruction decoder, tailored to specific client data distributions. Through this method, the student model can mimic the performance of the teacher model while utilizing significantly fewer parameters, thereby optimizing efficiency and reducing computational overhead. However, this approach is not applicable to the segmentation decoder, as maintaining high segmentation performance is the primary priority. Reducing parameters in the segmentation decoder through knowledge distillation or similar methods could compromise its accuracy, making it unsuitable for this component of the framework. Thus, parameter optimization efforts will be focused exclusively on the reconstruction decoder while preserving the segmentation decoder's robust performance.

E.2 Multimodal Data Heterogeneity and Label-free Adaptation

In FL, multimodal data heterogeneity remains a significant challenge, particularly when dealing with datasets derived from diverse imaging modalities, such as CT, MRI, and X-ray, each possessing distinct acquisition protocols and anatomical variations. This study is limited to clients using a single modality, and addresses data heterogeneity within that scope by aggregating local covariances to form a global covariance. However, this approach does not account for the more complex multimodal heterogeneity arising in FL scenarios with clients from multiple modalities. In future work, we aim to develop a framework where clients utilize independent encoders for each modality and a shared decoder, enabling the aggregation of covariances from these encoders. We will focus on aligning multimodal regions by matching between the global covariance and each local covariance, addressing the challenges of multimodal heterogeneity effectively. Furthermore, we will explore self-supervised learning techniques, such as contrastive learning, to extract meaningful representations from unlabeled data. These methods can help the model train invariant features across multimodal datasets, further reducing the impact of data heterogeneity. By addressing multimodal heterogeneity and enabling label-free adaptation, AdaShield- FL can expand its applicability to real-world federated medical imaging challenges.

References

[Campello *et al.*, 2021] Victor M Campello, Polyxeni Gkontra, Cristian Izquierdo, Carlos Martin-Isla, Alireza Sojoudi, Peter M Full, Klaus Maier-Hein, Yao Zhang, Zhiqiang He, Jun Ma, et al. Multi-centre, multi-vendor and multi-disease cardiac segmentation: the m&ms challenge. *IEEE Transactions on Medical Imaging*, 40(12):3543–3554, 2021.

[Carlini and Wagner, 2017] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017.

[Goodfellow *et al.*, 2014] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

[Hong *et al.*, 2023] Junyuan Hong, Haotao Wang, Zhangyang Wang, and Jiayu Zhou. Federated robustness propagation: sharing adversarial robustness in heterogeneous federated learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 7893–7901, 2023.

[Hssayeni *et al.*, 2020] Murtadha Hssayeni, M Croock, A Salman, H Al-khafaji, Z Yahya, and B Ghoraani. Computed tomography images for intracranial hemorrhage detection and segmentation. *Intracranial hemorrhage segmentation using a deep convolutional model. Data*, 5(1):14, 2020.

[Karimireddy *et al.*, 2021] Sai Praneeth Karimireddy, Lie He, and Martin Jaggi. Learning from history for byzantine robust optimization. In *International Conference on Machine Learning*, pages 5311–5319. PMLR, 2021.

[Li *et al.*, 2020] Suyi Li, Yong Cheng, Wei Wang, Yang Liu, and Tianjian Chen. Learning to detect malicious clients for robust federated learning. *arXiv preprint arXiv:2002.00211*, 2020.

[Madry *et al.*, 2017] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[McMahan *et al.*, 2017] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics*, pages 1273–1282. PMLR, 2017.

[Wu *et al.*, 2023] Zhaoxian Wu, Tianyi Chen, and Qing Ling. Byzantine-resilient decentralized stochastic optimization with robust aggregation rules. *IEEE transactions on signal processing*, 2023.

[Yi *et al.*, 2024] Yuhao Yi, Ronghui You, Hong Liu, Changxin Liu, Yuan Wang, and Jiancheng Lv. Near-optimal resilient aggregation rules for distributed learning using 1-center and 1-mean clustering with outliers. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16469–16477, 2024.

[Zhou *et al.*, 2021] Hong-Yu Zhou, Jiansen Guo, Yinghao Zhang, Lequan Yu, Liansheng Wang, and Yizhou Yu. nnformer: Interleaved transformer for volumetric segmentation. *arXiv preprint arXiv:2109.03201*, 2021.