

Adversarial Attacks on Medical Image Segmentation in Federated Learning

Tess Roovers
10633626

Bachelor thesis
Credits: 18 EC

Bachelor *Kunstmatige Intelligentie*



University of Amsterdam
Faculty of Science
Science Park 900
1098 XH Amsterdam

Supervisor
dr. N. Awasthi

Informatics Institute
Faculty of Science
University of Amsterdam
Science Park 900
1098 XH Amsterdam

Semester 1, 2023-2024

Abstract

In this study, the impact of adversarial attacks on medical image segmentation in federated learning was researched. Experiments were conducted through training a variety of U-Net segmentation models, using the same chest x-ray dataset. Using different parameter values to create perturbations in one out of three clients' training data, global models appear to be robust against these attacks. Regularisation effects occurred for almost all model configurations, with only Grad-CAM Gaussian noise indicating adversarial attacks negatively impacting the global model performance. Future research into targeted attacks is required to gain a better understanding of adversarial attacks in federated learning.

Keywords: federated learning, medical image segmentation, U-Net, adversarial attack, poisoning attack, convolutional neural network, Gaussian noise, salt and pepper noise, Grad-CAM, attention maps

Contents

1	Introduction	3
2	Theoretical framework	4
2.1	Medical image segmentation	4
2.2	Federated learning	5
2.3	Adversarial attacks	5
3	Method	7
3.1	Data processing	7
3.2	U-Net: model architecture	7
3.3	Federated learning setup	7
3.4	Adversarial attacks implementation	7
3.4.1	Gaussian noise	8
3.4.2	Salt and pepper noise	8
3.4.3	Grad-CAM attention maps	9
3.4.4	Noise configurations	10
3.5	Evaluation	11
4	Results and analysis	12
4.1	Standard model	14
4.2	Gaussian noise	14
4.3	Salt and pepper noise	15
4.4	Gaussian noise with Grad-CAM	15
4.5	Salt and pepper noise with Grad-CAM	15
5	Conclusion and discussion	17
	References	18

1 Introduction

Medical image segmentation is an essential part of modern healthcare applications since it enables high precision identification of regions of interest (ROIs) within medical images. Recently, significant advancements have been made with the use of convolutional neural networks (CNNs) (Li, Huang, Xu, & Lu, 2022; Milletari, Navab, & Ahmadi, 2016).

Enabling data owners to collaborate on model training without sharing private data, federated learning (FL) has gained more attention as a decentralised machine learning approach. Research conducted on the potential vulnerabilities and security threats of FL indicate a dichotomy in perspectives. Shejwalkar, Houmansadr, Kairouz, and Ramage (2022) found that FL is very robust in real-world scenarios, even with simple defence measures. On the other hand, a variety of publications also show the potential for adversarial attacks on medical image segmentation models. According to Zhang et al. (2023), adversarial training can significantly affect the test accuracy in case of non-independent and -identically distributed (IID) data.

The intersection of federated learning and medical image segmentation presents a promising perspective for collaborative healthcare analysis. In order to identify the vulnerabilities and potential threats within medical image segmentation models in federated learning, this project aims to evaluate the robustness and security by introducing adversarial attacks. Intending to cause a well-trained machine learning model to make incorrect predictions, adversarial attacks deliberately introduce small perturbations to the input data (Shejwalkar et al., 2022). The security risk of adversarial attacks in deep learning models, particularly those used for medical image analysis, is a subject of concern that requires more extensive research.

Recently developed attack strategies for medical image segmentation involve adaptable models. Ozbulak, Van Messem, and De Neve (2019) introduced the Adaptive Segmentation Mask Attack (ASMA) algorithm, capable of creating targeted adversarial examples that include imperceptible perturbations. Li et al. (2022) presented the first query-based black-box attack, incorporating an improved gradient estimation of loss and a learnable variance of adaptive distribution.

The current project intends to further expand this work by introducing adversarial attacks to the domain of medical image segmentation models within a FL framework. Evaluating the model robustness and security against adversarial attacks serves to answer the following research question:

How can we devise a potent adversarial attack strategy to challenge existing federated learning defence mechanisms, with a specific focus on unveiling vulnerabilities in segmentation models within the federated learning framework?

Since the available literature indicates conflicting perspectives on the robustness of FL in medical image segmentation, this study undertakes an exploratory approach without presenting a formal hypothesis. The objective is to conduct comprehensive and unbiased research.

2 Theoretical framework

2.1 Medical image segmentation

Medical image segmentation involves identification of the boundaries of regions of interest (ROIs) through pixel-based classification in medical images. In order to reliably identify segmentation objects within the medical domain, model training often requires a considerable amount of data. Given the sensitive nature of medical images, developing trustworthy models can thus be unfeasible for smaller organisations, such as local hospitals (Sandhu, Gorji, Tavakolian, Tavakolian, & Akhbardeh, 2023).

Convolutional neural networks (CNNs) have contributed to significant advancements within the last few years (Li et al., 2022). CNNs are deep learning architectures that incorporate multiple convolutional layers, enabling them to learn hierarchical representations from the input automatically (Lecun, Bottou, Bengio, & Haffner, 1998). Due to their strong ability to capture spatial hierarchies, CNNs are widely used for image analysis, including image segmentation.

One example of such a CNN architecture for semantic image segmentation is U-Net. Its U-shaped model architecture is included in **Fig. 1** and involves both a contracting and expansive path, intended for context extraction and spatial delineation, respectively. U-Net is capable of capturing structures with high detail due to its utilisation of skip connections, making it well-suited for tasks involving medical image segmentation (Ronneberger, Fischer, & Brox, 2015).

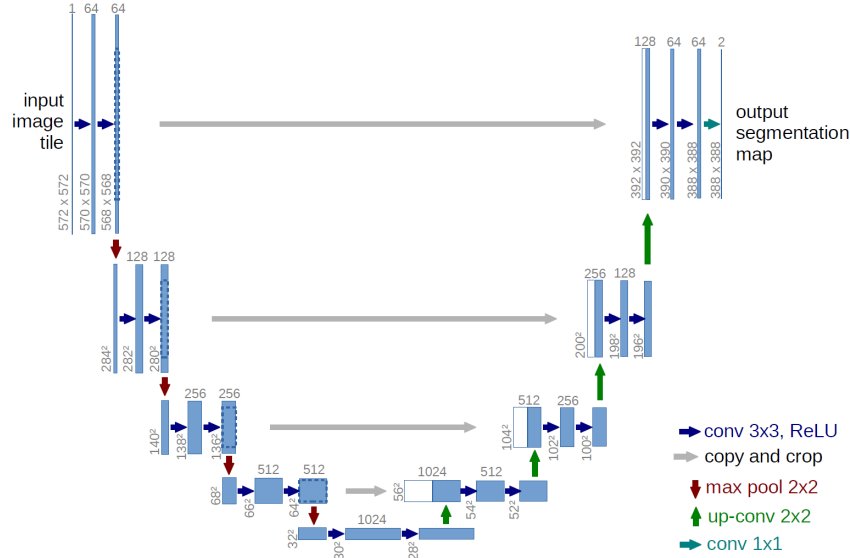


Figure 1: Standard U-Net model architecture for 32x32 pixels (Ronneberger et al., 2015). Operations are represented as arrows, multi-channel feature maps as blue boxes (with the number specifying the amount of channels) and copied feature maps as white boxes.

2.2 Federated learning

The decentralised machine learning approach of federated learning (FL) enables data owners to collaborate in model training without sharing their private data (McMahan, Moore, Ramage, Hampson, & y Arcas, 2017). Because of this, even smaller organisations with a limited amount of data can contribute to global model training and thus benefit from collaborative efforts between mutually untrusted clients (Shejwalkar et al., 2022).

The diagram in **Fig. 2** portrays a basic setup for FL with three clients. During training, the global model is distributed to either a randomly selected subset of, or all clients, who separately train a local copy of that model on their training data. After that, the local model weights are collected on the server, where the global model is updated based on an aggregation algorithm. Repeating this process for multiple iterations or epochs, clients are able to train a model together, without the server collecting their private data (Shejwalkar et al., 2022).

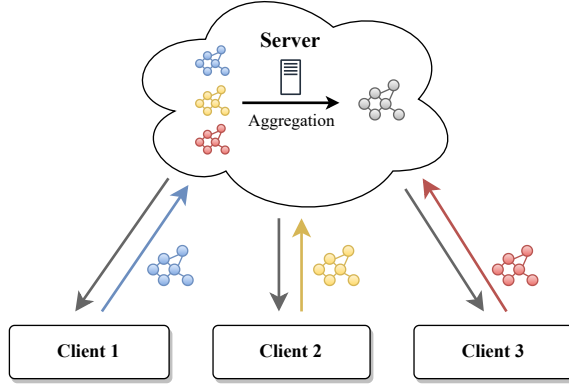


Figure 2: Setup for federated learning with three clients.

The most frequently used aggregation algorithm within the context of FL is the average aggregation rule. Here, η represents the learning rate, k the current client, n the total amount of samples across all clients, n_k the amount of samples in the current client’s dataset and w_t the current model. First, the average gradient on the local data is computed as g_k for each client, after which the central server conducts the update outlined in **Eq. 1** (McMahan et al., 2017).

$$w_{t+1} \leftarrow w_t - \eta \sum_{k=1}^K \frac{n_k}{n} g_k \quad (1)$$

2.3 Adversarial attacks

Recent studies show that the adding noise to training images in a federated learning environment can be used to impair a model’s performance, but also to improve it (You, Ye, Li, Xu, & Wang, 2019). Ozbulak et al. (2019) address the security risk of adversarial examples in deep learning models that are specifically used for medical image analysis. Focusing on image segmentation, the authors present the novel Adaptive Segmentation Mask Attack (ASMA) algorithm. The ASMA

algorithm can be used to create targeted adversarial examples that are found to be effective and include perturbations that are mostly invisible to the human eye (Ozbulak et al., 2019).

Szegedy et al. (2013) find that deep neural networks are significantly capable of learning discontinuous input-output mappings. The implication of this finding is that introducing imperceptible perturbations to an image can lead to the network misclassifying it. Furthermore, these perturbations are consistent across different networks trained on varying subsets.

In their article, Shejwalkar et al. (2022) present a critical evaluation of poisoning attacks on production federated learning. The authors state that potential vulnerabilities of federated learning to poisoning attacks by compromised clients have been identified through research. According to them, however, the true impact on real-world production federated learning systems is uncertain. As opposed to their expectations, Shejwalkar et al. (2022) find that federated learning is very robust in real-world scenarios, even with simple and cost-effective defence measures.

According to Goodfellow, Shlens, and Szegedy (2014), neural networks are vulnerable to adversarial perturbations because of their linear nature, but training a model on a combination of adversarial examples and clean images can also be beneficial and even lead to regularisation.

Within the context of adversarial attacks, two main types can be distinguished. White box attacks are based on the assumption that the adversary has complete access to the target model (Li et al., 2022; Ozbulak et al., 2019). On the other hand, black box attacks operate under the assumption that the adversary has no direct access to the target model and needs to be queried instead. Among black box attacks are transfer-based attacks, which provide an approach that is based on the transferability of adversarial examples between different models (Szegedy et al., 2013). More reliable and efficient, query-based attacks have recently gained more attention as an alternative approach that uses black-box optimisation algorithms in order to locate extrema of loss, which often correspond to adversarial examples (Li et al., 2022).

Tramèr et al. (2017) further classify adversarial attack strategies as single-step or iterative, with the first one requiring a single gradient computation and the latter computing multiple. Iterative attacks enable the adversary to alter perturbations, thus being able to adapt and craft adversarial examples better calibrated to the model. In single-step adversarial attacks, perturbations are added to the training data once, after which they remain constant (Tramèr et al., 2017).

3 Method

3.1 Data processing

For this study, the JSRT dataset¹ (Shiraishi et al., 2000) consisting of 247 greyscale human chest radiographs was used. In accordance with the methodology of Li et al. (2022), the images were resized to (256, 256) and their pixel values were rescaled to a range of [0, 255].

Annotations from the SCR database² (van Ginneken, Stegmann, & Loog, 2006) were consulted to generate corresponding ground truth masks. The masks were also resized to (256, 256) and consist of binary pixel values within a range of [0, 1]. Similarly to the research conducted by Li et al. (2022), the heart was selected as segmentation object, although ground truth masks were also generated for both lungs and clavicles (left and right), intended to facilitate further research.

For this study, the dataset was divided into 198 training and 49 test images.

3.2 U-Net: model architecture

The U-Net model as portrayed in **Fig. 1** was trained as an image segmentation model for the chest x-ray dataset through leveraging a pre-trained model³ as starting point. Adaptations were then made to ensure compatibility with the dataset and to allow for the inclusion of adversarial attacks in a FL configuration. The final code base used in this research is available on GitHub⁴.

3.3 Federated learning setup

In line with the diagram in **Fig. 2**, the experiments in this study were conducted with $K = 3$ clients. The 198 training images were randomised and divided among them equally, resulting in $n_k = 66$ local training images for all clients $k \in \{1, \dots, K\}$. The 49 test images were used for global model testing at the end of every global epoch.

Fixed parameter values were chosen to be $C = 1000$ (number of global epochs), $E = 1$ (number of local epochs), $\eta = 0.00001$ (learning rate) and $B = 1$ (batch size). The average aggregation rule from **Eq. 1** was chosen as aggregation method and model training was conducted for varying configurations in parallel (on Snellius GPU).

3.4 Adversarial attacks implementation

Common practices within related research involve the assumption that adversaries can compromise approximately 25% (or even 50%) of the clients (Shejwalkar et al., 2022). Although arguments opposing this have been presented, this research follows the common practice, randomly selecting $K_a = \frac{K}{3}$ adversarial clients. Since clients have no access to each other’s local models and are only able to manipulate their local models in attempting to affect the global model, this setup can be considered a black box attack method. In order to compare the influence of different attack methods, each model variant was instantiated using the same client and data split, trained alongside a baseline model without adversarial clients.

¹JSRT dataset: <http://db.jsrt.or.jp/eng.php>

²SCR dataset: <https://zenodo.org/records/7056076>

³Pytorch-UNet: <https://github.com/milesial/Pytorch-UNet> - GNU General Public License v3.0

⁴Complete code base: <https://github.com/TessRoovers/FL-Medical-Attack>

Four different noise configurations were selected for this study, as explained in the next subsections. Attempting to analyse the impact of adversarial attacks on a global model in FL, these noise configurations were each implemented as a single-step attack approach (Tramèr et al., 2017); Perturbations were added to the adversarial client’s training data once, to be used throughout the entire global training phase.

3.4.1 Gaussian noise

One common form of perturbations in images is Gaussian noise, as defined in **Eq. 2**. It involves the addition of a pixel-based value randomly selected from a normal distribution, with a specific mean (μ) and standard deviation (σ).

$$f(x, y) = I(x, y) + \mathcal{N}(\mu, \sigma) \quad (2)$$

In this study, the value of μ was fixed at 0, whereas models were trained with seven different values for $\sigma \in \{0.01, 0.02, 0.03, 0.04, 0.05, 0.1, 0.2\}$. Perturbations were generated for the adversarial client once, with examples of varying intensities portrayed in **Fig. 3**.

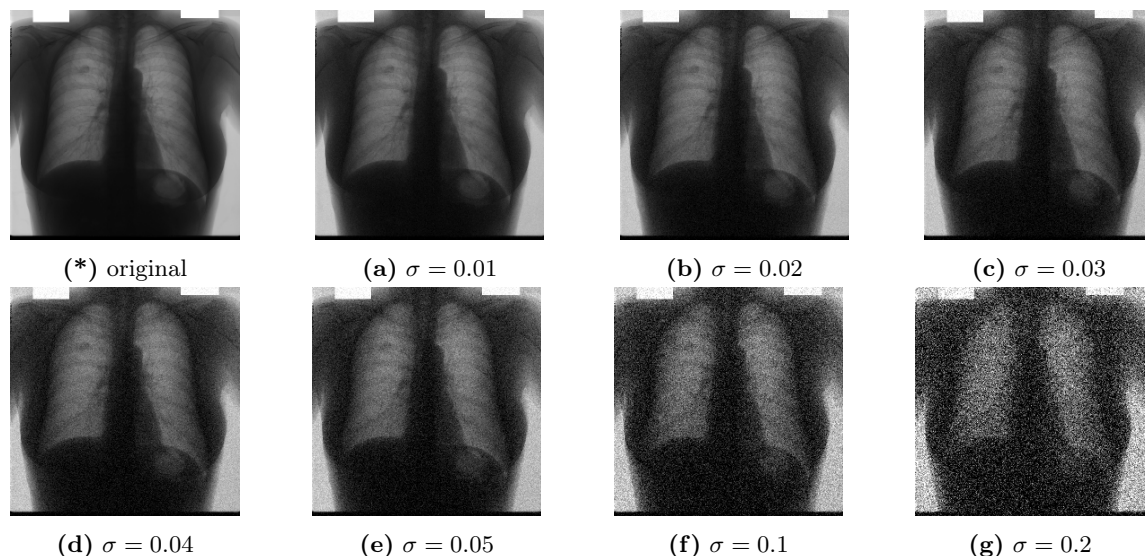


Figure 3: Gaussian noise for different standard deviations (with $\mu = 0$).

3.4.2 Salt and pepper noise

Another common occurrence of noise in images is salt and pepper (or impulse) noise, as defined in **Eq. 3**. Salt and pepper noise involves the addition of white pixels in dark regions and black pixels in bright regions of greyscale images. It is executed through random selecting pixels within the image dimensions with a specific probability (p) and changing their values to 0 (black) or 1 (white).

$$f(x, y) = \begin{cases} 1.0 & \text{with probability } \frac{p}{2} \\ 0.0 & \text{with probability } \frac{p}{2} \\ I(x, y) & \text{with probability } 1 - p \end{cases} \quad (3)$$

Different noise configurations with $p \in \{0.001, 0.005, 0.01, 0.02, 0.03, 0.04, 0.05\}$ were used to train models in this study, with p being the noise probability, equally distributed for salt and pepper noise. The noise added to the adversarial client's training images is visualised in **Fig. 4** for each configuration.

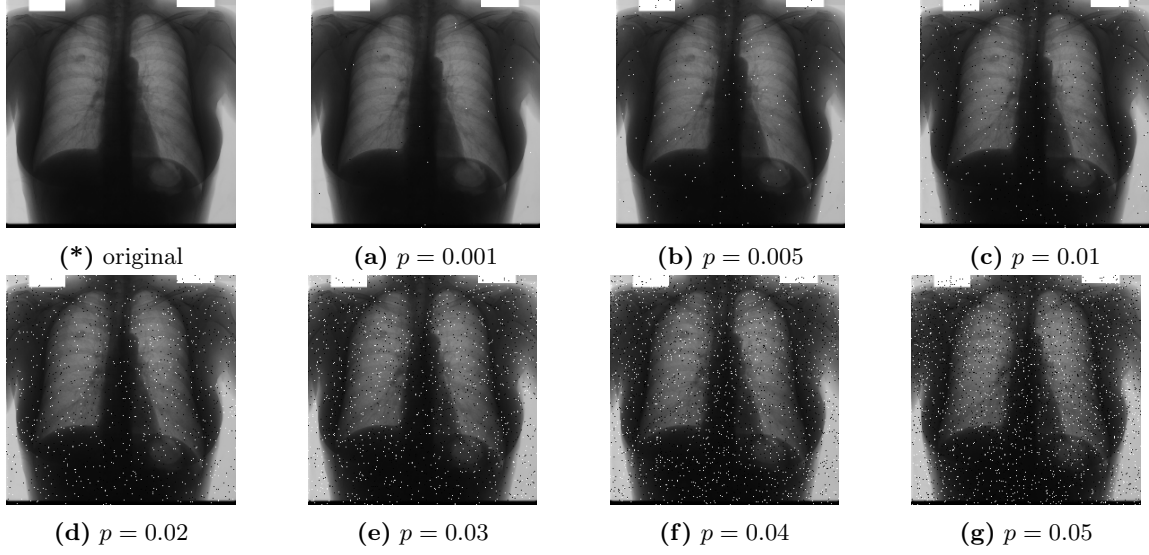


Figure 4: Salt and pepper noise for different probabilities.

3.4.3 Grad-CAM attention maps

Another noise addition approach included in this study is Gradient-Weighted Class Activation Mapping (Grad-CAM), through code adapted from (Gildenblat, 2021). The aim of incorporating Grad-CAM is to determine which pixels are important for the accuracy of a predicted mask. Grad-CAM uses gradients of the segmentation object flowing into the final convolutional layer to create a localisation map, thus highlighting the important regions of the image (Selvaraju et al., 2017). The Grad-CAM implementation used in this research is defined in **Eq. 4** and **5**, with $L_{Grad-CAM}^c$ representing the Grad-CAM layer output for class c , α_m^c the weight for channel m and class c , A_m^m the attention map for channel m and Z representing a normalisation term.

$$L_{Grad-CAM}^c = ReLU \left(\sum_m \alpha_m^c A_m^m \right) \quad (4)$$

$$\alpha_m^c = \frac{1}{Z} \sum_i \sum_j \frac{\partial y^c}{\partial A_{ij}^m} \quad (5)$$

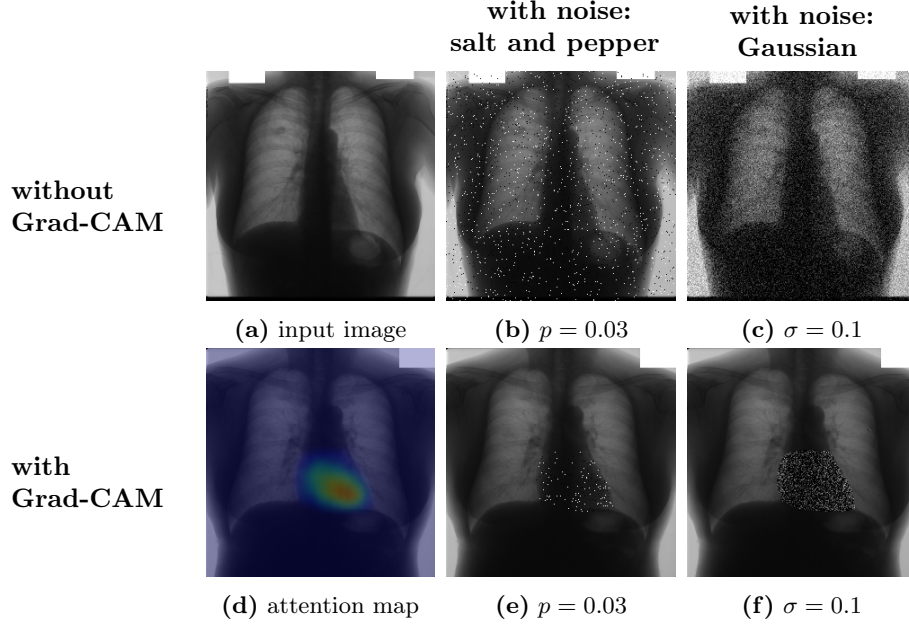


Figure 5: Noise with and without Grad-CAM attention maps.

For every input image, attention maps were generated locally, to be consulted at the noise addition phase. Perturbations with Grad-CAM were computed through first creating a Gaussian or salt and pepper perturbed image, as detailed in the previous subsections. Pixels with a value lower than a specified threshold ($\tau = 0.5$) were restored to the input image’s pixel values, ensuring that perturbations were only preserved for locations considered to be important for segmentation.

Visualisations of the resulting perturbed images are included in **Fig. 5** for similar values of p and σ in the creation of salt and pepper and Gaussian noise, respectively. The corresponding Grad-CAM generated attention map as portrayed in (d) clearly highlights the segmentation object (heart), where Gaussian filtering was applied for visibility purposes.

3.4.4 Noise configurations

Model training was conducted with seven noise configurations, with their corresponding parameter values included in **Table 1**.

	A	B	C	D	E	F	G
σ	0.01	0.02	0.03	0.04	0.05	0.1	0.2
p	0.001	0.005	0.01	0.02	0.03	0.04	0.05

Table 1: Overview of parameter values used in noise configuration A-G, with σ used for (Grad-CAM) Gaussian noise and p for (Grad-CAM) salt and pepper noise.

3.5 Evaluation

The Dice coefficient (Dice, 1945; Sorensen, 1948) is a widely used evaluation metric that provides insight into the similarity between two image samples. For a predicted segmentation mask \hat{y} and the corresponding ground truth mask y , the Dice score is given as:

$$D(\hat{y}, y) = \frac{1}{K} \sum_k \frac{\sum_i^H \sum_j^W \hat{y}_{ij}^k \cdot y_{ij}^k}{\sum_i^H \sum_j^W \hat{y}_{ij}^k + \sum_i^H \sum_j^W y_{ij}^k} \quad (6)$$

In this equation, summations are carried out across the image height (H) and width (W), with K representing the number of instances. For each pixel at location (i, j) , a predicted label is represented as $\hat{y}_{ij}^k \in \{0, 1\}$ and its ground truth label as $y_{ij}^k \in \{0, 1\}$, with 0 indicating it does not belong to the segmentation object and 1 signalling it does. After every global epoch, the aggregated model will be evaluated through averaging the Dice scores obtained from **Eq. 6** for all test images. Furthermore, the Structural Similarity Index (SSIM) will be used to measure the similarity between the predicted masks and the ground truth masks, given by the following equation:

$$SSIM(\hat{y}, y) = \frac{(2\mu_{\hat{y}}\mu_y + c_1)(2\sigma_{\hat{y}y} + c_2)}{(\mu_{\hat{y}}^2 + \mu_y^2 + c_1)(\sigma_{\hat{y}}^2 + \sigma_y^2 + c_2)} \quad (7)$$

In **Eq. 7**, \hat{y} and y represent the predicted and ground truth masks, $\mu_{\hat{y}}$ and μ_y are their respective means, $\sigma_{\hat{y}}$ and σ_y the corresponding standard deviations, $\sigma_{\hat{y}y}$ their covariance and c_1 and c_2 are constants introduced to prevent division by zero. This metric will be used as an additional evaluation metric to compare the predicted segmentation masks to the ground truth masks of every test image.

4 Results and analysis

The registered maximum and minimum Dice scores for every model configuration across 1000 epochs are included in **Table 2** and **3**, respectively. Based on this, the best performing noise model configurations for each type are $\sigma = 0.03$ (C) for Gaussian noise, $p = 0.005$ (B) for salt and pepper noise and $\sigma = 0.01$ (A) and $p = 0.001$ (A) for their Grad-CAM counterparts.

As the results indicate, the standard model (S) is outperformed by at least one configuration of each noise type at some point during training. Although some models display a minimum Dice score approximating zero, all models seem to converge to similar performance levels.

	S	G	SP	GC-G	GC-SP
A	0.92502	0.92620	0.91432	0.92946	0.92565
B		0.92161	0.92560	0.92010	0.92304
C		0.92913	0.91793	0.92399	0.91034
D		0.91904	0.91837	0.92119	0.90259
E		0.91641	0.91176	0.91991	0.91128
F		0.91877	0.91989	0.92403	0.91683
G		0.89194	0.90328	0.91390	0.91375

Table 2: Maximum Dice score for each noise configuration A-G for models: standard (S), Gaussian (G), salt and pepper (SP), Grad-CAM Gaussian (GC-G) and Grad-CAM salt and pepper (GC-SP).

	S	G	SP	GC-G	GC-SP
A	0.61526	0.54677	0.58726	0.70138	0.73099
B		0.73936	0.67522	0.56666	0.51372
C		0.63538	0.70512	0.51795	0.00000*
D		0.62070	0.62398	0.02102	0.00000*
E		0.63481	0.63638	0.00000*	0.00000*
F		0.65538	0.60861	0.57996	0.75200
G		0.60690	0.68712	0.00000*	0.00000*

* 1.74528×10^{-10}

Table 3: Minimum Dice score for each noise configuration A-G for models: standard (S), Gaussian (G), salt and pepper (SP), Grad-CAM Gaussian (GC-G) and Grad-CAM salt and pepper (GC-SP).

Providing more detailed insight into each model’s learning behaviour during training, **Fig. 6** displays the test Dice scores averaged across 5 epochs. As a baseline for comparison, the standard model scores are included as well. While some models portray more fluctuations than others, convergence seems to occur at a value of approximately 0.9 (Dice score).

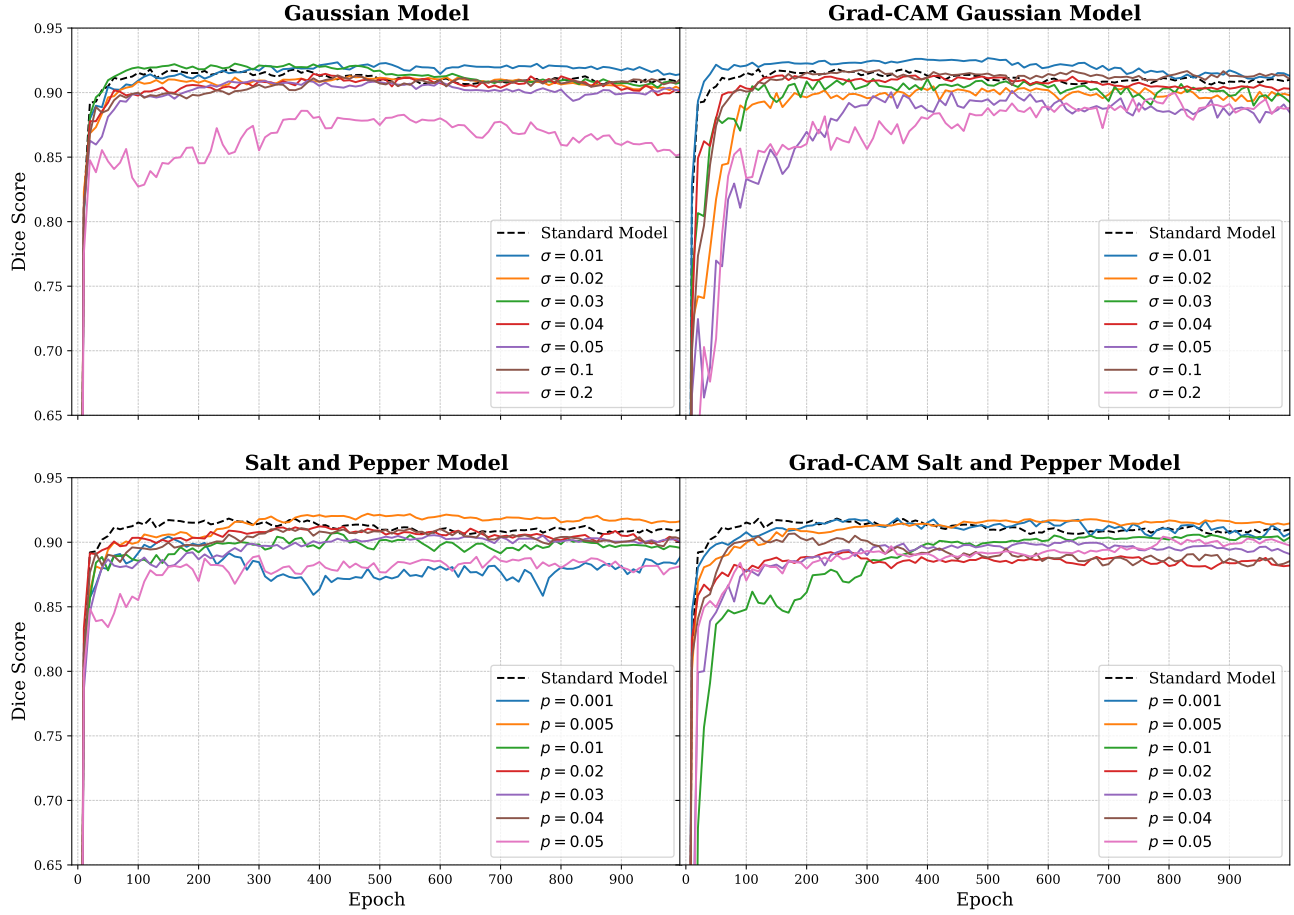


Figure 6: Dice score comparison for different parameter values within each model configuration, averaged across five epochs.

The best performing model configurations were saved during training and used to generate visual output for predicted masks on the test set. In **Table 4**, a sample of the predicted output for one input image is included. For every model type, a specific analysis will be included in the following subsections.

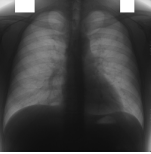






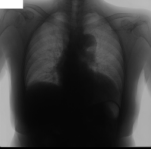






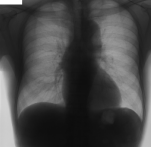






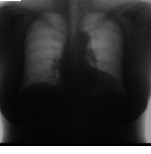

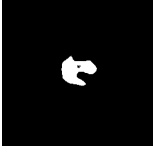




Original image	Ground truth	Predictions				
		S	G	SP	GC-G	GC-SP
						
						
						
						

Table 4: Predicted vs. ground truth masks for the standard model (S) and best-scoring noise model configurations: $\sigma = 0.03$ (G), $p = 0.005$ (SP), $\sigma = 0.01$ (GC-G) and $p = 0.001$ (GC-SP).

4.1 Standard model

The baseline model without any adversarial clients achieves a maximum Dice score of 0.92502 and reports a minimum score of 0.61526. Although it is outperformed by noise models at certain points during the 1000 epochs, it shows a consistent high performance. Furthermore, it is one of the fastest converging models, stabilising at around 100 epochs. The model has an average SSIM on the test set images of 0.9667.

4.2 Gaussian noise

Out of all noise configurations, the best performance for Gaussian noise is achieved with $\sigma = 0.03$, with a Dice score of 0.92913, outperforming the baseline (S) model. The worst performance in terms of maximum Dice score (0.89194) can be found for $\sigma = 0.2$, which is the configuration with highest Gaussian noise intensity; As seen in **Table 2**, this is the lowest maximum Dice score reported out of all models. However, the lowest score for Gaussian noise is reported for $\sigma = 0.01$ (**Table 3**). **Fig. 6** indicates that the Gaussian noise model with $\sigma = 0.2$ indeed struggles to converge, which is unique in comparison to the other models. With most Gaussian noise models (including $\sigma = 0.03$) failing to consistently achieve higher Dice scores than the baseline model (S), only $\sigma = 0.01$ seems to converge with a better performance at the 500-1000 epochs range. It achieves an average SSIM

of 0.9644, whereas $\sigma = 0.03$ results in an average SSIM of 0.9664, indicating slightly more similarity between the ground truth and predicted masks.

In order to determine the effectiveness of adversarial attacks, visibility of the perturbations added is an essential aspect to consider. Most of the variations of noise intensity as displayed in **Fig. 3** are noticeable, especially with values of $\sigma = 0.04$ or higher. Based on the results, only the model with $\sigma = 0.01$ and potentially also $\sigma = 0.03$ seem to achieve promising regularisation results, with $\sigma = 0.2$ affecting the global model performance negatively. However, the visibility of perturbations in this model might be too noticeable.

4.3 Salt and pepper noise

The best performing noise configuration for salt and pepper noise is found for $p = 0.005$, which corresponds to an average SSIM of 0.9650. This model also seems to outperform the standard model consistently after training for around 300 epochs. Even though most of the model configurations seem to converge well, the ones with the highest ($p = 0.05$) and lowest ($p = 0.001$) probabilities seemingly struggle the most.

As visible in **Fig. 4**, salt and pepper noise is considerably noticeable, even for lower values of p . Taking this into consideration in evaluating the effectiveness of salt and pepper noise on the global model performance, this method might attract unwanted attention, making it less suited than other, less noticeable types of perturbations. However, it might be a valid method for regularisation, since at least one configuration outperforms the standard model.

4.4 Gaussian noise with Grad-CAM

As compared to the regular Gaussian noise models, their Grad-CAM counterparts have fewer perturbed pixels. Because of this, adversarial examples could be expected to be less noticeable. However, the Grad-CAM perturbations included in **Fig. 5** clearly display noise in a confined area (inside the boundaries of the segmentation object), creating a visual contrast between the non-noisy and noisy part in case of high intensity perturbations.

The best performing configuration for Grad-CAM Gaussian noise is found for $\sigma = 0.01$, being the best performing model in this entire research. With only minor perturbations, the adversarial examples generated in this configuration are more likely to remain unnoticed than its regular Gaussian noise counterpart. The best performing model has an average SSIM of 0.9641, which is not significantly higher than the other models' scores.

Due to the high performance, Grad-CAM Gaussian noise models could be useful for regularisation for lower values of σ . However, the Dice scores included in **Fig. 6** also indicate that global model performance could be negatively affected for higher intensity perturbations, such as $p = 0.05$ and $p = 0.2$. Comparing these results with the regular Gaussian noise models, at least one model seems to require longer to converge, indicating that Grad-CAM Gaussian noise might be interesting to explore further in the context of adversarial attacks.

4.5 Salt and pepper noise with Grad-CAM

Similarly to Grad-CAM Gaussian noise, Grad-CAM salt and pepper noise has less perturbed pixels in the image as compared to its regular counterpart. Since salt and pepper is fairly noticeable, the contrast between the perturbed and non-perturbed regions stand out (**Fig. 5**). In order to

prevent an adversary from being noticed in their attempt to attack the global model, lower intensity perturbations might be considered over higher intensity perturbations.

The best performing Grad-CAM salt and pepper model is the one with the lowest probability, where $p = 0.001$. The maximum and minimum Dice scores reported in **Table 2** and **3** indicate that the different noise configurations adapt well after a number of epochs, but also include four minimum values approximating zero. The subplot in **Fig: 6** reveal that these lower values occur near the beginning, confirming the models' adaptability. The average SSIM for the best performing configuration is 0.9648, which is comparable to the other model types.

With all noise configurations converging and the potential visibility of salt and pepper noise, this Grad-CAM salt and pepper model does not seem to be suitable for adversarial attacks.

5 Conclusion and discussion

In this research, the following research question was explored:

How can we devise a potent adversarial attack strategy to challenge existing federated learning defence mechanisms, with a specific focus on unveiling vulnerabilities in segmentation models within the federated learning framework?

In order to answer this question, one baseline U-Net model was trained on the dataset, alongside four main attack model setups. Analysing the impact of varying noise intensities for each setup, it was discovered that a significant amount of the models are able to converge well, thus demonstrating robustness and adaptability when concerned with adversarial attacks. According to Kwon (2021), medical image segmentation models can become more robust through adversarial training, which aligns with these findings.

The achieved results indicate that U-Net as an image segmentation model is quite robust to adversarial attacks involving Gaussian-, salt and pepper, Grad-CAM Gaussian or Grad-CAM salt and pepper noise. Many models even achieve a better performance than the baseline model, confirming the occurrence of regularisation effects.

Given the fact that ground truth masks were generated for other segmentation objects within the same dataset used in this study, experiments could be repeated for other segmentation objects, besides the heart. This study was conducted with a total of three clients in the federated learning setup, thus only including one adversary. Expanding this to include a bigger pool of clients, while maintaining the 1 : 3 ratio of adversarial clients could provide more insight into the impact of poisoning attacks on medical image segmentation in federated learning.

With research presenting conflicting results on the impact of adversarial attacks on FL, some suggestions for future research will be made. In this research, global model aggregation was conducted through a basic approach of average aggregation. In order to gain insight into the robustness of federated learning against adversarial attacks, other aggregation algorithms should be studied as well. Recently, Zhao, Zhou, and Wan (2024) presented SuperFL as a more robust federated learning setup that uses two servers for global model aggregation, thus better protecting privacy and offering more resistance to poisoning attacks. Other algorithms include clipped average-, secure-, momentum-, Bayesian-, adversarial-, quantisation-, hierarchical-, personalised- and ensemble bases aggregation (Moshawrab, Adda, Bouzouane, Ibrahim, & Raad, 2023). With an explicit emphasis on fraud detection through outlier rejection, secure enclaves and model-based anomaly detection, the adversarial aggregation approach should be considered in future research.

Other perspectives to consider include the investigation of other, less visible noise methods such as speckle noise (multiplication by random values from Gaussian distribution), Poisson noise (commonly used in medical image segmentation), gradient noise (introduces random variations based on local intensity gradients). Furthermore, instead of examining the impact of noise types alone, further research should further explore adaptive attack setups such as the Adaptive Segmentation Mask Attack (ASMA) algorithm (Ozbulak et al., 2019), in which adversarial examples are specifically crafted based on perturbations proven to be effective.

References

- Dice, L. R. (1945). Measures of the amount of ecologic association between species. *Ecology*, 26(3), 297–302.
- Gildenblat, J. (2021). *Pytorch library for cam methods*. <https://github.com/jacobgil/pytorch-grad-cam>. GitHub.
- Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.
- Kwon, H. (2021). Medicalguard: U-net model robust against adversarially perturbed images. *Security and Communication Networks*, 2021, 1–8.
- Lecun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324. doi: 10.1109/5.726791
- Li, S., Huang, G., Xu, X., & Lu, H. (2022). Query-based black-box attack against medical image segmentation model. *Future Generation Computer Systems*, 133, 331–337. Retrieved from <https://www.sciencedirect.com/science/article/pii/S0167739X2200084X> doi: <https://doi.org/10.1016/j.future.2022.03.008>
- McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. In *Artificial intelligence and statistics* (pp. 1273–1282).
- Milletari, F., Navab, N., & Ahmadi, S.-A. (2016). V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3d vision (3dv)* (pp. 565–571). doi: <https://doi.org/10.1109/3DV.2016.79>
- Moshawrab, M., Adda, M., Bouzouane, A., Ibrahim, H., & Raad, A. (2023). Reviewing federated learning aggregation algorithms; strategies, contributions, limitations and future perspectives. *Electronics*, 12(10). Retrieved from <https://www.mdpi.com/2079-9292/12/10/2287> doi: 10.3390/electronics12102287
- Ozbulak, U., Van Messem, A., & De Neve, W. (2019). Impact of adversarial examples on deep learning models for biomedical image segmentation. In *Medical image computing and computer assisted intervention—miccai 2019: 22nd international conference, shenzhen, china, october 13–17, 2019, proceedings, part ii 22* (pp. 300–308). Retrieved from <https://arxiv.org/pdf/1907.13124>
- Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention—miccai 2015: 18th international conference, munich, germany, october 5–9, 2015, proceedings, part iii 18* (pp. 234–241).
- Sandhu, S. S., Gorji, H. T., Tavakolian, P., Tavakolian, K., & Akhbardeh, A. (2023). Medical imaging applications of federated learning. *Diagnostics*, 13(19), 3140.
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the ieee international conference on computer vision* (pp. 618–626).
- Shejwalkar, V., Houmansadr, A., Kairouz, P., & Ramage, D. (2022). Back to the drawing board: A critical evaluation of poisoning attacks on production federated learning. In *2022 ieee symposium on security and privacy (sp)* (pp. 1354–1371). Retrieved from <https://arxiv.org/pdf/2108.10241>
- Shiraishi, J., Katsuragawa, S., Ikezoe, J., Matsumoto, T., Kobayashi, T., Komatsu, K., . . . Doi, K. (2000). Development of a digital image database for chest radiographs with and without a

- lung nodule: receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *American Journal of Roentgenology*, 174(1), 71–74.
- Sorensen, T. (1948). A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. *Biologiske skrifter*, 5, 1–34.
- Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*. doi: <https://doi.org/10.48550/arXiv.1312.6199>
- Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., & McDaniel, P. (2017). Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*.
- van Ginneken, B., Stegmann, M. B., & Loog, M. (2006). Segmentation of anatomical structures in chest radiographs using supervised methods: a comparative study on a public database. *Medical Image Analysis*, 10(1), 19–40. Retrieved from <https://www.sciencedirect.com/science/article/pii/S1361841505000368> doi: <https://doi.org/10.1016/j.media.2005.02.002>
- You, Z., Ye, J., Li, K., Xu, Z., & Wang, P. (2019). Adversarial noise layer: Regularize neural network by adding noise. In *2019 IEEE International Conference on Image Processing (ICIP)* (p. 909–913). doi: 10.1109/ICIP.2019.8803055
- Zhang, J., Li, B., Chen, C., Lyu, L., Wu, S., Ding, S., & Wu, C. (2023). Delving into the adversarial robustness of federated learning. *arXiv preprint arXiv:2302.09479*.
- Zhao, Y., Zhou, H., & Wan, Z. (2024). Superfl: Privacy-preserving federated learning with efficiency and robustness. *Cryptology ePrint Archive*.