This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3676321

FedGA: Genetic Algorithm-Guided Federated Learning for Medical Image Segmentation with Non-IID Features

Faisal Ahmed, Rodrigo Moreno, David Sánchez, *Senior Member, IEEE*, Zouhair Haddi, and Josep Domingo-Ferrer, *Fellow, IEEE*

***Abstract*— Federated learning (FL) enhances data privacy and compliance with data regulations by enabling multiple decentralized parties to collaboratively train machine learning models without sharing their data. This makes it an ideal paradigm for the healthcare domain. Despite its inherent privacy-by-design, FL faces performance and convergence challenges when dealing with non-independent and identically distributed (non-IID) data. Although previous studies have primarily addressed the challenges of skewed label distribution across clients, which are more effective for classification, we focus in this paper on the much less explored challenge of multi-domain FL in medical image segmentation tasks, where client data originate from different domains with varying feature distributions. To address this problem, we propose FedGA, which performs heuristic, gradient-free optimization on the server side after local model aggregation in order to optimize the global model by using a genetic algorithm. Empirical results on breast lesion segmentation from ultrasound images and prostate segmentation from T2w MRI images show the potential of FedGA in FL schemes. Specifically, FedGA improves the segmentation precision in critical boundary regions compared to existing approaches, accelerates global model convergence, reduces the total number of communication rounds required to achieve optimal performance, and offers better overall efficiency.**

***Index Terms*— Data Heterogeneity, Data Privacy, Federated Learning, Genetic Algorithms, Medical Imaging, Non-IID Data, Segmentation, Statistical Heterogeneity**

I. Introduction

Manuscript received XXX; revised XXX. Support to this work has been received from Marie Skłodowska-Curie Doctoral Networks Actions (HORIZON-MSCA-2021-DN-01-01; 101073222), Cancerfonden (22-2389 Pj), the Government of Catalonia (ICREA Acadèmia Prizes to J. Domingo-Ferrer and to D. Sánchez) and MCIN/AEI under grant PID2024-157271NB-I00 "CLEARING-IT" (*Corresponding author: Faisal Ahmed.*)

Faisal Ahmed and Zouhair Haddi are with NVISION Systems and Technologies SL, Barcelona, Catalonia, e-mail: {faisal.ahmed,zouhair.haddi}@nvision.es

Rodrigo Moreno is with KTH Royal Institute of Technology, Department of Biomedical Engineering and Health Systems, Hälsovägen 11C, 14157, Huddinge, Stockholm, Sweden, e-mail: rodmore@kth.se

Faisal Ahmed, David Sánchez and Josep Domingo-Ferrer are with Universitat Rovira i Virgili, Dept. of Computer Engineering and Mathematics, ComSCIAM-Center for Computational Science and Applied Mathematics, CYBERCAT-Center for Cybersecurity Research of Catalonia, Av. Països Catalans 26, 43007 Tarragona, Catalonia, e-mail: {david.sanchez,josep.domingo}@urv.cat

IN medical imaging, it is common that highly sensitive medical images and the corresponding metadata are distributed across multiple entities (*e.g.*, hospitals, clinics). Due to their sensitivity, and under the umbrella of data protection regulations such as the GDPR, the HIPAA, or the European Union AI Act, these data cannot be freely shared for, *e.g.*, building machine learning models.

Federated learning (FL) offers a distributed alternative to centralized machine learning by enabling multiple clients holding sensitive data to collaboratively train a machine learning model without sharing their data, thereby mitigating privacy risks and enhancing compliance with privacy regulations. However, FL struggles when dealing with non-independent and identically distributed (non-IID) data across participating entities. Unfortunately, this is the case for many real-world applications, in which data from different parties can be heterogeneous in terms of *feature distribution*, *class distribution*, *temporal distribution*, *quantity*, and *patterns* [1]. Data heterogeneity is even more pronounced in the medical imaging domain as multiple types of heterogeneity coexist due to the variety of imaging scanners, acquisition protocols, and patient demographic characteristics [2].

In the FL setting, each client updates its local model by using its own data without accounting for the overall data distribution across all clients; for this reason, the local objectives may conflict and diverge from the global model's objective. This could slow the convergence and prevent FL from achieving an optimal global model [3], [4]. Several solutions have been proposed to mitigate the heterogeneity of data in the label and quantity distribution in FL settings, but they are only applicable to medical image classification tasks, where the primary challenge is the imbalanced distribution of image labels and variations of data set size between hospitals. In classification tasks, variations in local imaging feature distributions have minimal or no impact on the learning process. This can be attributed to the *encoder-only* architectures of classification models (*e.g.*, ResNet [5], EfficientNet [6], etc.), which focus on extracting high-level image representationsto assign a label to the image without the need to account for localized variations in feature distribution.

Segmentation is another equally critical task in the medical imaging domain. Unlike classification, variations in local feature distributions might have an impact in FL settings in which

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3676321

the model must capture and analyze fine-grained spatial and structural information embedded in the images. Segmentation models leverage an *encoder-decoder* architecture (*e.g.*, U-Net [7], SegNet [8], TransUNet [9], DeepLab [10], etc.), in which the decoder upscales high-level representations learned by the encoder to reconstruct spatial details with pixel-level annotations. Due to feature heterogeneity among clients, local models struggle to converge to a consensus in pixel-wise label assignment. As a result, existing FL solutions that address data heterogeneity in label and quantity distribution struggle with medical image segmentation tasks.

Several approaches have been designed for medical image segmentation. However, they focus on learning models that generalize only to the non-IID data distributions of the participating institutions, and thus, their efficacy is often constrained by domain shift. This limitation reduces their robustness to out-of-distribution (OOD) data, as the models may not perform as well on data sets from non-participating hospitals. This raises concerns for clinical deployment, where reliability across diverse and unseen data sources is essential to ensure patient safety and regulatory compliance.

To address this research gap, we propose an FL method for medical image segmentation with three primary goals: (1) learning the segmentation model using non-IID data from the participating hospitals; (2) generalizing the model across the collective data distributions of all hospitals; and (3) making the model robust to external, out-of-distribution data.

We address the feature heterogeneity problem in federated medical image segmentation tasks with a novel perspective based on metaheuristic optimization of the global model. Specifically, our contributions are as follows.

1) We propose FedGA (Genetic Algorithm-Guided Federated Learning), a solution to find the optimal global model from a suboptimal aggregated global model by using metaheuristic optimization. FedGA initializes a set of random models near the aggregated global model in the solution space via Monte Carlo simulation and subsequently employs a genetic algorithm with an evolutionary strategy to thoroughly explore the solution space and identify the optimal global model.
2) We curate multiple data sets that capture real-world data heterogeneity arising from variations in imaging protocols, patient demographics, and imaging equipment from different manufacturers.
3) We perform extensive experiments on medical image segmentation tasks using two image modalities and two model architectures. Our results demonstrate that FedGA not only enhances segmentation precision compared to existing approaches, but also facilitates faster convergence of the global model, reduces the total number of communication rounds required to reach optimal performance, and offers better overall efficiency.

Section II provides background on FL. Section III discusses related work that tackles data heterogeneity in FL. Section V presents FedGA. Section VI details the experimental setup. Section VII reports, discusses, and compares empirical results. Section IX draws conclusions and presents potential future research directions.

II. Background on Federated Learning

A typical FL system consists of a central server S and a set of N participants, known as clients. Each client i has a (potentially sensitive) data set D_i, which is a subset of the global data set D. In some FL settings, the server may also hold public or private data sets, similar to the clients. The goal of FL is to train a global model based on the data of all clients, but without requiring them to share their data sets.

In local premises, each client i optimizes a local objective $\mathcal{L}_i$, called the loss function, under the model parameters θ_i as

$$\min_{\theta_i} \mathcal{L}_i(\theta_i; D_i). \tag{1}$$

The global objective $\mathcal{L}$ is generally formulated as the weighted average of the local optimization objectives, that is,

$$\min_{\hat{\theta}} \mathcal{L}(\hat{\theta}; D) = \sum_{i=1}^{N} \frac{|D_i|}{|D|} \mathcal{L}_i(\hat{\theta}; D_i), \tag{2}$$

where $|\cdot|$ denotes the number of elements in the set and $\hat{\theta}$ represents the parameters of the global model.

Personalized Federated Learning (pFL) is a variant of FL that aims to learn customized models for each client rather than a single global model. These personalized models are tailored to the specific characteristics of each client's data while still leveraging the benefits of collaborative training across all clients. Formally, the goal of pFL is to learn a set of models $\{\theta_i\}_{i=1}^{N}$ such that each client i minimizes the local loss $\mathcal{L}_i(\theta_i; \mathcal{D}_i)$, while taking into account the data of all other clients $\{\mathcal{D}_j\}_{j \neq i,\, j \in \{1,\ldots,N\}}$.

III. Related Work

State-of-the-art solutions to address data heterogeneity in FL either *regularize local training*, *improve aggregation efficiency* or *combine local regularization with efficient aggregation*.

FedProx [4] uses the squared Euclidean distance between the current local and global models as a regularization term in the local objective function to prevent local updates from deviating too much from the global model. FedDyn [3] introduces a dynamic regularization term in local training. This term updates the client's risk objective at each step to ensure that the local optima are asymptotically consistent with the stationary points of the global empirical loss. Unlike FedProx, FedDyn considers the history of local updates and global model changes in the regularization terms. To keep track of the update history, clients must maintain additional memory. However, while both methods perform well under marginal label heterogeneity, they experience significant performance degradation as heterogeneity increases or statistical heterogeneity in feature distribution arises.

FedNova [11] normalizes and scales each local update by the number of local steps performed during local training before updating the global model on the server. Therefore, each client must share with the server information about the number of local steps it performs during each training round. This not only increases communication overhead, but also opens the door to gradient-based privacy attacks [12].

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3676321

FedMGC [13] replaces the cross-entropy loss with focal loss during local training to amplify the loss impact of minority classes, thereby mitigating the model's bias towards majority classes. However, focal loss is incompatible with segmentation models that require loss functions directly optimizing overlap metrics, such as intersection over the union (IoU) or Dice Similarity Coefficient (DSC).

SCAFFOLD [14] addresses data heterogeneity by applying variance reduction techniques to mitigate client drift. It introduces control variates at both the server and client sides, which are used to estimate the update directions for the global model and individual client models, respectively. Their difference is treated as drift in local training, which is used to correct local updates on the client side [1]. However, SCAFFOLD requires clients to share their control variates with the server at each communication round, in addition to the local model updates. This doubles the communication length compared to Federated Averaging (FedAvg) [15]. Moreover, since the control variate encapsulates gradient history, it is vulnerable to gradient-based privacy attacks.

FedDecorr [16] argues that high label heterogeneity in data causes *feature dimension collapse*, where global model feature representations shift toward a lower-dimensional space rather than the ambient space derived from local models. To mitigate this, FedDecorr applies a regularization term during local training that encourages the different dimensions of the representation to be uncorrelated. This requires access to a one-dimensional feature vector immediately preceding the output layer. However, in segmentation models, the feature representation module (encoder) is located far from the output layer and generates multiple feature maps. As a result, implementing FedDecorr in segmentation models is challenging, and the regularization effect diminishes before it reaches the encoder module during backpropagation.

FedBN [17] is the only method (to our knowledge) specifically designed to address *non-IID features* among clients. It updates the *batch normalization* layers locally without communicating and aggregating them on the server. However, its applicability is limited to neural network architectures that include batch normalization layers. Many state-of-the-art segmentation models do not incorporate batch normalization due to its poor performance, thereby limiting the applicability of the solution. Moreover, *batch normalization* assumes that the training data are derived from a single distribution, allowing the mean and variance to be calculated from each mini-batch [18]. This assumption does not hold for medical images, even within a single hospital, because different imaging protocols may be used for patients according to their disease conditions. Moreover, FedBN is designed exclusively for personalized FL and does not generate any global model.

MOON [19] requires changes to the architecture to incorporate a projection head for feature alignment, which is not feasible for legacy or proprietary models. The projection head improves the effectiveness of both contrastive learning and model contrastiveness, as empirically demonstrated in [20] and MOON papers, respectively. Consequently, when the latent features extracted by the encoder are used directly to implement the MOON method without a projection head, the segmentation performance is expected to decrease.

FedCross [21] conducts local model training in a round-robin sequential fashion among participating clients. This sequential processing inherently limits the parallelization of training and results in substantially increased training time. In addition, it is prone to catastrophic forgetting, a phenomenon in which a model's performance on previously encountered client data deteriorates if those data have not been experienced for a long time. On the other hand, FedCross is more vulnerable to privacy attacks and breaches than conventional FL because, after local training on a client's private data, the resulting model is shared directly with another client.

FedDP [22] is a pFL technique tailored for self-attention networks (vision transformers) in segmentation tasks. It first personalizes the long-term dependency modeling module, with each site retaining its query vector parameters of the self-attention module and sharing the remaining model parameters with the server. Subsequently, each site collects all local models in a single communication round and fine-tunes its local parameters, guided by inter-site prediction inconsistencies. The architecture-specific nature of FedDP makes it inapplicable to CNN-based segmentation models, such as the highly efficient nnU-Net. Moreover, it opens the door to successful privacy attacks by directly exposing client-specific and dependency-personalized local models to all participating clients, where some of them can be potential adversaries.

FedCE [23] is explicitly designed to optimize both collaborative fairness (rewarding clients based on their contribution) and performance fairness (achieving uniform performance across clients) simultaneously. FedCE estimates the contribution of each client in both the gradient and the data spaces by tracking gradient direction differences among clients and measuring prediction errors on each client's data using an auxiliary model. Based on these estimates, it determines the aggregation weights for the global model. However, by enforcing the objective of fairness, it can strongly limit the performance of certain clients. Moreover, its design objective does not account for the robustness of the global model to out-of-domain data distributions, restricting its clinical applicability.

FedCLAM [24] integrates client-adaptive momentum terms derived from each client's loss reduction during local training, along with a personalized dampening factor to mitigate overfitting. In addition, it introduces a novel intensity alignment loss that aligns predicted and ground-truth foreground distributions to account for heterogeneous images with varying intensity profiles across institutions and devices. Similar to FedCE, FedCLAM's learning objectives do not include explicit components or mechanisms to improve generalization to out-of-domain data, which limits its applicability in clinical settings.

A common limitation of existing methods is that they were explicitly designed to tackle problems of quantity and label skewness in classification tasks. To our knowledge, these methods for addressing data heterogeneity in FL have not been tested with segmentation models. Whereas classification models comprise an encoder component for feature representation, followed by a classification layer, segmentation models use encoder-decoder architectures. The encoder module learns feature representations as in classification models, while the

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3676321

decoder module performs the upsampling task to generate the segmentation mask. Consequently, existing methods either disrupt the learning process of the decoder component when attempting to align the encoder in the local model with the global model, or their effects on the encoder component get diminished since they first backpropagate through the decoder module and then reach the encoder part.

IV. Problem Statement

We address the problem of data heterogeneity in FL that originates from *feature distribution shifts* and *skewness in the number of images* between different hospitals for a supervised image segmentation task. We assume that the data for each hospital are different, with variations in imaging systems, acquisition protocols, patient demographics, and data set sizes. In other words, data across hospitals exhibit heterogeneity in both feature distributions and sample quantities. In our setting, we assume that the server represents another hospital with enough computing power (*i.e.*, the server is not resource-constrained) that holds a small set of medical images with the corresponding segmentation masks. Since the server also acts as a data-contributing institution, similar to the client hospitals participating in federated training, both the server and client hospitals are considered mutually trusted entities. This setup inherently mitigates potential privacy concerns. Furthermore, even if this assumption does not fully hold, the server can alternatively employ a public or synthetically generated data set for fitness evaluation, thereby ensuring that data privacy is preserved.

Our goal is to learn a parameterized segmentation model in a supervised manner on the aggregation of the data sets of all hospitals. The model should generalize across the collective data distributions of all participating hospitals, as well as to external, out-of-distribution data, thereby enabling accurate and general segmentation.

V. FedGA: Genetic Algorithm-Guided Federated Learning

The overall workflow of FedGA is shown in Figure 1. It follows the client-server interaction of standard FL, along with an end-to-end model flow. The server coordinates the training process by broadcasting the global model θ_t to clients in training round t. The clients use the global model as a fresh starting point and perform local optimization on their private data to achieve local objectives. After completing local updates, each client sends its updates θ_t^k to the server, which aggregates the collected information from all participants into a new global model version θ_{t+1}^* for the federated training round $t+1$, which is sub-optimal due to the data heterogeneity across the client hospitals. Subsequently, the server performs metaheuristic optimization using a genetic algorithm (GA) to produce a optimal global model θ_{t+1}. The new optimal global model θ_{t+1} is broadcast again by the server until T training rounds are completed (or alternatively, until the successive learned global models converge to a stable model).

A. Server Side Optimization

The high heterogeneity in both the feature and quantity distributions across hospitals causes them to train local segmentation models with conflicting gradients, leading to local objectives that significantly deviate from the global objective. When local models with conflicting gradients are aggregated on the server, they often yield a suboptimal global model that fails to generalize well to both in-distribution and out-of-distribution data sets.

We hypothesize that this issue can be mitigated by searching for the optimal global model from a suboptimal one through the integration of Monte Carlo simulation with metaheuristic optimization on the server at each communication round. Based on this hypothesis, our solution incorporates a metaheuristic optimization process on the server on top of federated averaging (FedAvg) at each training round to search for the optimal global model from a suboptimal one in the solution space. More specifically, at training round t, the server first aggregates each client hospital's updates θ_t^k to produce an intermediate global model θ_{t+1}^* for the next federated training round $t+1$ using the FedAvg [15] algorithm as

$$\theta_{t+1}^* = \sum_{k=1}^{K} \frac{n_k}{n} \theta_t^k, \tag{3}$$

where $n_k = |D_i|$ represents the number of data samples held by client k and $n = |D|$ denotes the total number of samples across all participating clients. The server then performs metaheuristic optimization by using a GA to find the optimal global model θ_{t+1} from the sub-optimal aggregated model θ_{t+1}^*. We chose GA because it is easy to implement, requires a small amount of data, and can converge to an acceptable solution close to the optimum in a reasonable amount of time [25]. GA works in several steps: *forming a population of candidate solutions*, *evaluating their fitness*, *selecting the fittest individuals*, and *creating new offspring through crossover and mutation operations*. These are described below:

Forming initial populations. In this step, a set of candidate models is generated in the search space via Monte Carlo simulation from the averaged global model. As the initial candidate set, we consider ten random models in the neighboring space of the averaged global model, and implement the Monte Carlo simulation by the Gaussian noise mechanism.

A neighboring model $\hat{g}$ is generated from the suboptimal averaged global model g by adding noise to its parameters:

$$\hat{g} = g + \eta, \eta \sim N(0, \sigma^2), \tag{4}$$

where η is the Gaussian noise sampled from $N(0, \sigma^2)$ and σ^2 is the variance of the noise. The original suboptimal global model and the neighboring models generated in this step comprise the initial population of metaheuristic optimization in the next step. Inclusion of the suboptimal global model in the set of candidate models ensures that the solution will not perform worse than it.

Fitness evaluation. In this step, the fitness of each candidate model is evaluated based on its precision in segmenting unseen

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3676321

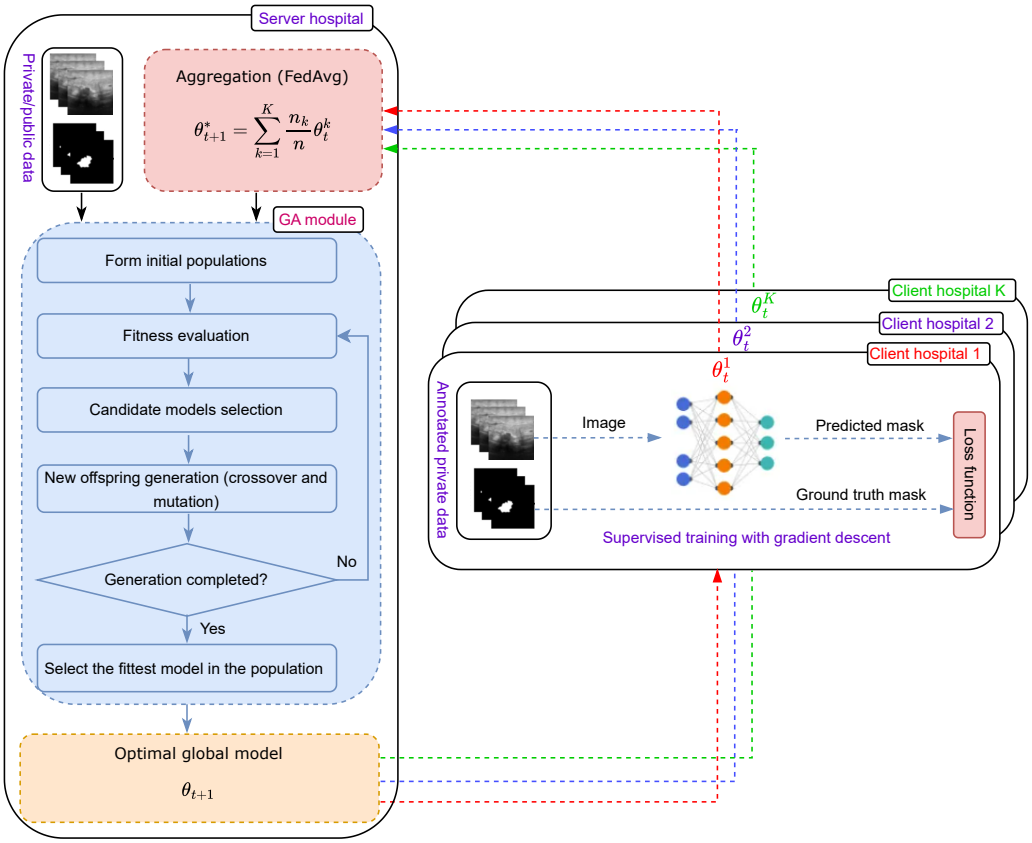


Fig. 1: Overall workflow of the proposed FedGA framework.

images held by the coordinating hospital. This unseen data set is expected to have a feature distribution that differs from those of the participating hospitals, primarily due to variations in patient demographics, imaging protocols, or imaging systems. DSC is used as the fitness function. Formally, the fitness F of a candidate model M on the unseen data set $\mathcal{D}_{\text{unseen}} = \{(x_i, y_i)\}_{i=1}^{m}$ with x_i and y_i being the image and the corresponding ground-truth segmentation mask is defined as

$$F(M) = \frac{1}{m} \sum_{i=1}^{m} \frac{2 \cdot |M(x_i) \cap y_i|}{|M(x_i)| + |y_i|}, \tag{5}$$

where $M(x_i)$ is the predicted segmentation mask for the input x_i, $|\cdot|$ denotes the cardinality (e.g., the number of pixels in the mask), and $\cap$ represents the intersection of the predicted and the ground-truth masks in $\mathcal{D}_{\text{unseen}}$.

Candidate model selection. In this step, models from the current population are chosen to become the parent models for the next generation to improve the overall population. We adopt the elitism selection strategy [26], where the best performing models are directly passed on to the next generation. More specifically, we sort the models in the current population according to their DSC such that $F(M_1) \geq F(M_2) \geq \cdots \geq F(M_N)$, and the best $N/2$ models are selected for the next generation, where N is the size of the population, which is 10 in our case.

New offspring generation. In this step, two genetic operations *crossover* and *mutation* are employed to generate new and better candidate models from the candidate models of the previous generation.

Crossover explores the solution space by combining information from two parent models to create one or more offspring models. It preserves, propagates, and exploits the best parameter components present in the parent models in the offspring models. We consider a uniform crossover in which each of the weights in the offspring model is independently chosen from one of the parent models. Formally, a new offspring model O from the two-parent models P_1 and P_2 is defined as

$$O(i,j) = \begin{cases} P_1(i,j), & \text{if } r < 0.5, \\ P_2(i,j), & \text{if } r \geq 0.5, \end{cases} \tag{6}$$

where $O(i,j)$, $P_1(i,j)$, and $P_2(i,j)$ are, respectively for models O, P_1, and P_2, the weight between the i-th neuron in layer $l-1$ and the j-th neuron in layer l, and $r \in [0,1]$ is a random value generated independently for each weight.

In turn, *mutation* helps to explore the solution space more thoroughly by introducing diversity in the offspring models and preventing the algorithm from getting stuck in local optima. This diversity is implemented by adding Gaussian noise to some randomly selected parameters of the offspring model O with a probability of α as follows:

$$O'(i,j) = \begin{cases} O(i,j) + \mathcal{N}(0, \sigma^2), & \text{with probability } \alpha, \\ O(i,j) & \text{otherwise}, \end{cases} \tag{7}$$

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3676321

where $O'(i,j)$ is the updated weight at position (i,j) after mutation, $\mathcal{N}(0,\sigma^2)$ is Gaussian noise with mean 0 and variance σ^2, and α is the probability of mutation that controls the fraction of weights to which mutation is applied. We used $\alpha = 0.1$ in the experiments.

In this way, the crossover and mutation operations balance exploration (searching new areas) and exploitation (refining known good areas) in the solution space, helping to find (near) optimal solutions during a particular training generation.

B. Client Side Optimization

Each client hospital downloads the new global model θ_t at training round t and subsequently updates the model parameter by minimizing the loss function $\mathcal{L}$ on their respective privately held data set, using stochastic gradient descent as follows:

$$\theta_t^k = \theta_t - \eta \nabla \mathcal{L}(\theta_t; D_k), \tag{8}$$

where $\mathcal{L}(\theta_t; D_k)$ is the weighted dice and cross-entropy loss function, $\nabla\mathcal{L}(\theta_t; D_k)$ is the gradient computed on the local data set D_k, and η is the local learning rate.

The complete FedGA workflow is formalized in Algorithm 1, where the GA-based post-aggregation optimization module, along with its associated inputs, is highlighted in blue to distinguish it from conventional FL approaches. FedGA performs additional optimization only on the server side, while the client-side remain the same as in vanilla FL. Moreover, this optimization is conducted on the distorted versions of the aggregated global model generated by adding Gaussian noise. Therefore, FedGA either improves or at least offers the same level of privacy and robustness against inversion attacks on the clients as vanilla FL.

Algorithm 1 FedGA: Genetic Algorithm-Guided FL

Input: Number of clients N, minibatch size B, local epochs E, learning rate η, population size P, generations $G_{\max}$, crossover probability P_c, mutation probability P_m
Output: θ_T, optimized global model after T rounds

Server Executes
Initialize global model θ_0
for each round $t = 0, \ldots, T-1$ **do**
  for each client $k \in \{1, \ldots, K\}$ **in parallel do**
    $\theta_t^k \leftarrow \text{LocalUpdate}(\theta_t, B, E, \eta)$
  end for
  $\theta_{t+1}^* \leftarrow \sum_{k=1}^{K} \frac{n_k}{n} \theta_t^k$ ▷ FedAvg aggregation
  $\mathcal{P} \leftarrow \{\theta_{t+1} + \Delta_k \mid \Delta_k \sim \mathcal{N}(0, \sigma^2),\ p = 1, \ldots, P\}$
  for generation $g = 1, \ldots, G_{\max}$ **do**
    Evaluate fitness $f(M)$ for each $M \in \mathcal{P}$ using server data
    Sort $\mathcal{P}$ by fitness and retain the top $P/2$ as parents
    Generate offspring $\mathcal{O}$ via crossover with probability P_c
    Mutate $\mathcal{O}$ with probability P_m
    $\mathcal{P} \leftarrow \mathcal{P} \cup \mathcal{O}$
  end for
  $\theta_{t+1} \leftarrow \arg\max_{M \in \mathcal{P}} f(M)$
end for
return θ_T
Client Executes
function LocalUpdate(θ, B, E, η)
  for epoch $e = 1, \ldots, E$ **do**
    for each minibatch b of size B **do**
      $\theta \leftarrow \theta - \eta \nabla \mathcal{L}(\theta; b)$
    end for
  end for
  return θ
end function

VI. Experimental Setup

A. Tasks and data sets

FedGA was evaluated in two medical image segmentation tasks of two different modalities. For both tasks, the data associated with the simulated clinical centers exhibited heterogeneous distributions, primarily due to variations in imaging conditions such as scanner manufacturers, magnetic field strengths, acquisition protocols, and patient populations. Moreover, these two data sets are the largest available image sets that match the scenario addressed in this paper.

Prostate segmentation: For this task, we collected Magnetic Resonance Imaging (MRI) data sets from three public repositories: *NCI-ISBI13* [27], *I2CVB* [28] and *PROMISE12* [29]. Some of these data sets are multimodal; however, for the purposes of our experiments, we exclusively used the T2-weighted (T2w) modality. Since the *NCI-ISBI13* and *PROMISE12* data sets contain data from multiple sources, we separated each source and treated it as an individual client in the federated scenario, resulting in six federated sites. Specifically, across the six sites, samples from Sites A and B are from the *NCI-ISBI13* data set; samples from Site C are from the *I2CVB* data set; and samples from Sites D, E, and F are from the *PROMISE12* data set. Table I illustrates the details of the scanners, imaging protocols, data sources, and number of slices for the sites. For all sites, we adopted the same image preprocessing pipeline outlined in [30]. All images were preprocessed to ensure a consistent field of view centered on the prostate region and resized to dimensions of 384×384 in the axial plane. Intensity values were subsequently normalized to have zero mean and unit variance.

Breast lesion segmentation: For this task, we curated four breast ultrasound (US) image data sets from public sources: *BUS-BRA* [31], *BUSI* [32], and *QAMEBI* [33]. The BUS-BRA data set comprises images acquired from multiple US scanners with distinct hardware configurations and imaging protocols. To preserve data heterogeneity and reflect real-world cross-silo scenarios, we partitioned the data according to the originating scanner, treating each subset as an independent client within the FL scenario. This resulted in six federated sites (A-F). More specifically, the imaging samples from sites D and E were derived from the BUSI and QAMEBI data sets, respectively, and the samples from sites A, B, C, and F originated from the BUS-BRA data set. Table II gives details on US systems, imaging protocols, data sources, and the number of images per site. The same preprocessing pipeline was applied to the images across all sites to ensure data consistency. Both images and segmentation masks were converted to grayscale to enforce a single-channel representation. Intensity values were normalized to reduce variability between different US systems, with each pixel intensity re-scaled based on the minimum and maximum values in the image to confine them within the range [0, 1]. Finally, all images and labels were resized to a standardized spatial resolution of (384, 384) to enable consistent batch processing and efficient model training.

As shown in Tables I and II, the data of each site for a

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3676321

specific task are different from those of other sites in terms of MRI or US systems, imaging protocols, patient demographics, and the number of slices or images.

To quantify feature heterogeneity across images originating from different sites for a given task, we compute the Fréchet Inception Distance (FID) between each pair of site-specific data sets associated with that task. FID evaluates the dissimilarity between two sets of images by comparing their feature distributions, specifically capturing differences in both the first-order (mean) and second-order (covariance) statistics of features extracted by a pre-trained deep neural network. A higher FID score reflects greater divergence in feature representations between data sets. The FID scores for the prostate and breast lesion segmentation tasks are visualized as heatmaps in Figures 2 and 3, respectively. As shown, both tasks exhibit consistently elevated FID values between data sets from different sites, underscoring substantial cross-site feature heterogeneity. This naturally introduces both feature and quantity distribution heterogeneity, creating a more realistic and challenging setting for FL.

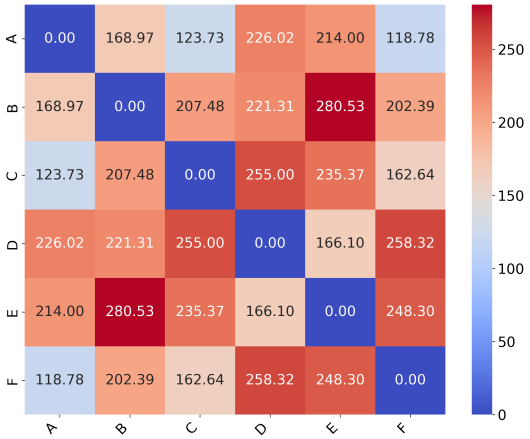


Fig. 2: FID values between data sets from the sites in Table I for the prostate segmentation task.

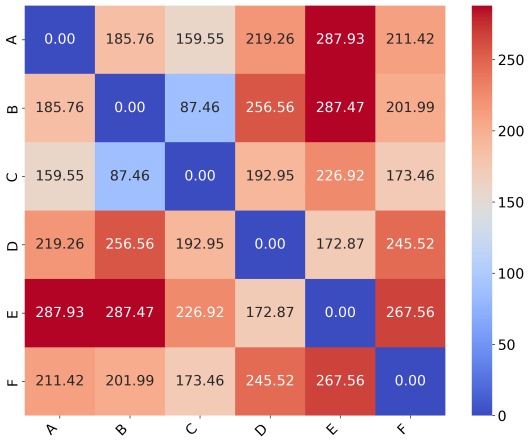


Fig. 3: FID values between data sets from the sites in Table II for the breast lesion segmentation task.

B. Segmentation Models

We adopted a 2D residual U-Net architecture for the breast lesion segmentation task. The model takes single-channel 2D images as input and outputs a two-channel segmentation mask corresponding to the lesion and background classes. The network consists of five encoding-decoding levels with channel dimensions set to [16, 32, 64, 128, 256], and downsampling is performed using strides of [2, 2, 2, 2] at each level. Each convolutional block in both the encoder and decoder paths comprises two residual units, which facilitate improved gradient flow and training stability. Skip connections are used between corresponding encoder and decoder levels to preserve spatial information. The model outputs per-pixel class probabilities, which are subsequently post-processed using softmax for final mask prediction.

On the other hand, we used a 2D Attention U-Net architecture for the prostate segmentation task because the segmentation was modeled as slice-based segmentation, following previous FL studies [23], [24] on this data set, to ensure computational feasibility in the federated setting.. The input and output channels, as well as the encoder-decoder structure, were kept consistent with those used in the 2D Residual U-Net for the breast lesion segmentation task. Unlike the Residual U-Net, this model integrates attention gates into the skip connections between the encoder and decoder pathways. These gates leverage an attention mechanism that enables the network to focus on salient features relevant to prostate anatomy while suppressing irrelevant or noisy activations.

It is worth noting that FedGA can be extended to any model architecture because it is model-agnostic and operates solely on the server side using aggregated model parameters. In particular, it does not require any modifications to the model architecture or training procedure on the client side, which makes it versatile.

C. Evaluation Criteria and Metrics

We split each client's data into training and validation sets in an 80/20 ratio. The training subset was used for local model updates, while the validation subset was used to enforce early stopping regularization and choose the best local model across local epochs. The server data were split into two mutually exclusive segments: one was used for server-side optimization via a GA, and the other part was kept reserved for global model performance evaluation.

For evaluation and comparison, we used DSC, which assesses the similarity between the predicted and ground-truth segmentation masks:

$$\mathrm{DSC}(X, Y) = \frac{2|X \cap Y|}{|X| + |Y|}, \tag{9}$$

where X and Y denote the predicted segmentation and the ground-truth mask, respectively. Additionally, the 95th percentile Hausdorff Distance (HD95) was employed to evaluate boundary consensus between ground-truth and predicted masks. HD95 can be calculated as follows:

$$\mathrm{HD95}(X, Y) = PC_{95}(h(X, Y), h(Y, X)), \tag{10}$$

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3676321

TABLE I: Overview of data sets from different federated sites for the prostate segmentation task.

data set	Institute	Manufacturer	Field Strength (T)	Resolution (in-plane/ through-plane) in mm	Endorectal Coil	# Cases	# Slices
Site A [27]	RUNMC	Siemens	3	0.6–0.625/3.6–4.0	No	30	421
Site B [27]	BMC	Philips	1.5	0.4/3	Endorectal	30	384
Site C [28]	I2CVB	Siemens	3	0.67-0.79/1.25	No	19	468
Site D [29]	BIDMC	GE	3	0.25/2.2–3	Endorectal	12	261
Site E [29]	HK	Siemens	1.5	0.625/3.6	Endorectal	12	158
Site F [29]	UCL	Siemens	1.5 and 3	0.325–0.625/3–3.6	No	13	175

TABLE II: Overview of data sets from different federated sites for the breast lesion segmentation task.

data set	US System	Probe Type	Frequency Range (MHz)	# Patients	# Images
Site A [31]	Toshiba Aplio 300	–	12–14	72	902
Site B [31]	GE Logiq 5	–	10–12	490	1064
Site C [31]	GE Logiq 7	–	10–14	488	809
Site D [32]	LOGIQ E & LOGIQ E9 Agile	Matrix Linear probe	1–5	–	697
Site E [33]	AirPlorer Ultimate	Linear probe	5–18	–	232
Site F [31]	GE U-Systems	–	12–14	14	25

where

$$h(X,Y) = \max_{x\in X}\min_{y\in Y}\|x-y\|,$$
$$h(Y,X) = \max_{y\in Y}\min_{x\in X}\|y-x\|, \tag{11}$$

and PC_{95} is the 95th percentile. The higher DSC and the lower HD95, the better.

D. Baseline Methods

We compared our method with the FedBN state-of-the-art FL method [17], which is the only method designed to address *non-IID features* among clients. As reference, we also evaluated the vanilla FL method *FedAvg* and two other baseline methods designed to address *non-IID label distributions* or *label distribution skewness* between clients: FedProx [4] and SCAFFOLD [14]. Finally, we compare FedGA with FedCLAM [24], which is explicitly designed for medical image segmentation with non-IID data.

E. Implementation Details

Federated training was conducted over 20 communication rounds, and each client performed 10 local epochs per round. The model architecture and hyperparameters were the same across all clients to ensure system homogeneity. The parameters of the initial global model were initialized using the default initialization schemes provided by the PyTorch library. Local models were optimized using the Adam optimizer, configured with momentum parameters $\beta_1 = 0.9$ and $\beta_2 = 0.99$, a learning rate of 1×10^{-3} and a batch size of 128.

The optimization was used to minimize the weighted combination of DSC and Cross-Entropy losses, with equal weighting assigned to both components. This approach aimed to maximize class overlap via the DSC and pixel-wise classification accuracy through the Cross-Entropy loss.

To avoid overfitting and improve computational efficiency, early stopping regularization was applied to local model updates in each communication round to terminate the training process if the DSC did not improve for three consecutive epochs. For the FedGA method, we set the crossover and mutation probabilities to 0.5 and 0.1, respectively, for both tasks, as these values have been shown to perform optimally in the literature. In Section VII-G we report an ablation study on these parameters that further reinforce our choice. To initialize the candidate population, we sampled 10 random models in the local neighborhood of the averaged global model using a Gaussian-noise–based Monte Carlo sampling mechanism. For both tasks, the GA was executed for 20 generations.

Segmentation models, image preprocessing, and local training pipelines were implemented using the MONAI framework, built on top of PyTorch with CUDA support for GPU acceleration. The overall federated training pipeline used Python 3.10.12, with Scikit-learn 2.7 leveraged for auxiliary machine learning utilities and data handling. All experiments were executed on a system equipped with an Intel Xeon Silver 4114 processor, 64 GB of RAM, and an NVIDIA A10 GPU with 24 GB of VRAM, running on Ubuntu 22.04.3 LTS operating system.

VII. Empirical Results

A. Misalignment of Local Models Induced by Feature Heterogeneity

This section empirically validates and illustrates how heterogeneity in feature distributions and the number of images across hospitals gives rise to hospital-specific segmentation models with inherently conflicting gradient directions.

For this illustration, we trained five independent segmentation models for each segmentation task outlined in Section VI-A, using data from five different hospitals associated with each task. Subsequently, each model was evaluated on data from an external hospital corresponding to the same task but exhibiting a markedly different feature distribution. We used the same network architecture, initial parameters, and hyperparameters to train each of the five models corresponding to a task.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3676321

Specifically, for breast lesion segmentation from US images, we trained five independent models using data from hospitals labeled A through E and evaluated their performance on data from hospital F, as detailed in Table II. Similarly, five models were trained on data from hospitals A to E for prostate segmentation from MRI images and tested on the data set from hospital F, as shown in Table I.

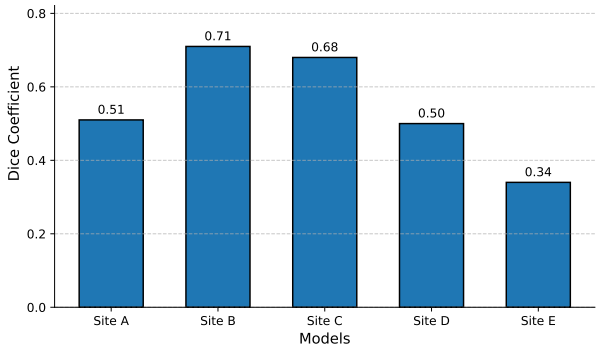


Fig. 4: Model performance across sites on a common out-of-distribution and unseen data set for breast lesion segmentation.

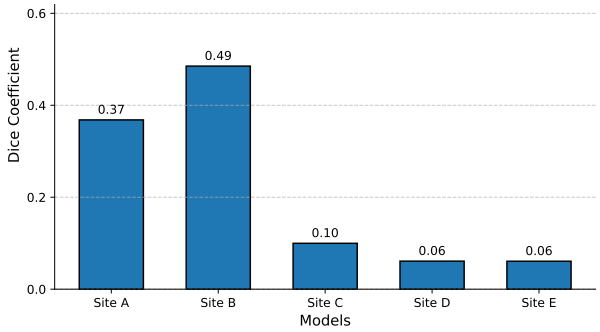


Fig. 5: Model performance across sites on a common out-of-distribution and unseen data set for prostate segmentation.

As Figure 4 shows, the DSC of the models from different sites varies from 0.34 to 0.71 for the breast lesion segmentation task on the external data set. Similarly, prostate segmentation models in MRI follow the same pattern, with a DSC ranging from 0.06 to 0.49, as shown in Figure 5. This validates our assumption: the gradients of individual models might conflict with each other and not align in a consistent direction due to the non-IID feature distribution. When such conflicting gradients are aggregated, they produce a global model that converges to a local optimum in the search space rather than a global optimum.

B. Validation of FedGA Across Different FL Settings

We next validate the stability of FedGA across both segmentation tasks under varying FL settings. Specifically, we assess the consistency and stability of each client, where both the participating clients and the datasets used for GA-based optimization on the server side differ.

To simulate different FL settings for a given segmentation task, we designate one hospital associated with that task as the server and assign the remaining hospitals as participating clients. This process is repeated until every hospital has acted as the server. Consequently, across different FL settings, at least one participating client changes, and both the number and set of images used for fitness evaluation vary accordingly.

For example, in the first FL configuration of the **breast lesion segmentation** task, site A was designated as the server, while sites B–F (as listed in Table II) served as clients. The GA-based model optimization employed 451 images, corresponding to 50% of site A's patient data, for the fitness evaluation. In the final FL configuration, site F acted as the server, with sites A–E as clients. Here, 12 mutually exclusive images (50% of patient data from site F) were used for fitness evaluation. For each client, we evaluated all global models in terms of DSC and HD95 within FL settings that include that client, using its local test set. We then calculated the mean and standard deviation (SD) of DS and HD95 along with a 95% bootstrap confidence interval (CI). The results are summarized in Tables III and IV for the breast lesion and prostate segmentation tasks, respectively. The small SDs and narrow bootstrap CIs observed across all sites for both segmentation tasks indicate that FedGA is robust to variations in FL settings, including the subset of participating clients and the image used for metaheuristic optimization at the server. FedGA remains stable under heterogeneous FL configurations by searching for the optimal global model in each communication round, thereby ensuring that the learning dynamics converge toward a balanced equilibrium.

TABLE III: DSC and HD95 statistics (mean, standard deviation, and 95% confidence interval) per site across different federated settings for the **breast lesion segmentation** task. The 95% confidence interval was estimated via bootstrapping the scores across federated settings.

Sites	DSC	HD95 (mm)
Site A	79.35 ± 1.93 [77.84, 80.70]	45.85 ± 1.16 [44.97, 46.83]
Site B	86.74 ± 1.79 [85.40, 88.16]	28.23 ± 1.20 [27.16, 29.11]
Site C	85.45 ± 1.86 [84.11, 86.99]	35.99 ± 1.40 [35.01, 37.26]
Site D	75.01 ± 1.84 [73.61, 76.37]	56.52 ± 1.70 [54.97, 57.66]
Site E	82.67 ± 1.74 [81.39, 84.11]	47.71 ± 1.44 [46.56, 48.70]
Site F	81.61 ± 1.26 [80.68, 82.65]	51.37 ± 1.18 [50.48, 52.38]

TABLE IV: DSC and HD95 statistics (mean, standard deviation, and 95% confidence interval) per site across different federated settings for the **prostate segmentation** task. The 95% confidence interval was estimated via bootstrapping the scores across federated settings.

Sites	DSC	HD95 (mm)
Site A	84.06 ± 1.31 [82.90, 84.88]	13.16 ± 0.33 [12.85, 13.40]
Site B	83.96 ± 0.86 [83.28, 84.65]	11.08 ± 0.27 [10.89, 11.29]
Site C	36.00 ± 1.73 [34.55, 37.42]	138.57 ± 3.45 [135.87, 141.28]
Site D	85.04 ± 1.40 [83.98, 86.09]	15.97 ± 0.76 [15.43, 16.65]
Site E	86.07 ± 1.26 [85.10, 86.96]	09.50 ± 0.40 [09.18, 09.80]
Site F	80.94 ± 1.70 [79.54, 82.14]	33.60 ± 1.53 [32.45, 34.72]

C. Comparison for Out-of-Domain Generalization

We next evaluate how well FedGA performs compared to existing FL methods when handling data distributions not seen during training (e.g., data from a completely new site).

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3676321

For this experiment, we designated *Site F* as the server, and the remaining sites as clients for both tasks. We evaluated the performance of the best global model across all federated training rounds on the server data, which are not directly exploited in the training process. To rigorously evaluate performance differences between FedGA and competing baseline methods, we also conducted statistical hypothesis testing using per-test image DSC and HD95 scores. Specifically, we evaluated each method on the test set, recording the DSC and HD95 scores for each image in the data sets, and then performed a pairwise comparison between FedGA and each of the baseline methods based on the resulting DSC and HD95 values. Before significance testing on the difference, we applied the Shapiro–Wilk test (with a significance threshold of $p = 0.05$) to determine whether the DSC and HD95 scores followed a normal distribution. If both methods yielded p-values greater than 0.05, we used a paired t-test. Otherwise, we employed the non-parametric Wilcoxon signed-rank test to ensure robust statistical inference. Table V summarizes the quantitative performance (mean ± standard deviation) of FedGA and the comparative baseline methods across two segmentation tasks. It also reports the p-values from statistical tests conducted between each baseline method and FedGA.

At first glance, it is clear that, for both tasks, our method consistently outperformed state-of-the-art approaches in terms of DSC and HD95. The performance improvements are statistically significant, as indicated by p-values < 0.001 for all baseline methods across both tasks and metrics. This improvement can be attributed to the GA's server-side metaheuristic optimization, which searches for the best model from the suboptimal global models in each training round.

FedAvg and FedProx show suboptimal performance, with inconsistent effectiveness across imaging modalities. FedProx outperforms FedAvg in ultrasound imaging, while FedAvg achieves better results in MRI segmentation. FedAvg's suboptimal performance is expected, as it lacks mechanisms to mitigate inter-client data heterogeneity, including label and feature distribution skewness. On the other hand, FedProx adds a proximal term to prevent client updates from straying too far from the global model. Although this helps mitigate label distribution skew in classification models, it can be overly restrictive for segmentation tasks where pixel-level classification is required. This restriction prevents clients from adequately adapting to their local data for segmentation tasks, leading to underfitting. In addition, the feature encoder stays far behind the output layer in segmentation models. As a result, the impact of regularization for feature harmonization diminishes during backpropagation.

FedBN and SCAFFOLD exhibit the worst performance in out-of-domain segmentation. Despite being designed to address feature heterogeneity, FedBN performs poorly because it follows a personalized federated learning approach: instead of producing a single global model, it generates client-specific local models by fine-tuning the global model on local data to better capture unique distributions and client-specific patterns. Since it does not aggregate batch statistics across clients, it fails to harmonize features among them, which limits its ability to generalize to unseen out-of-domain data sets.

SCAFFOLD's poor performance in segmentation tasks is due to its inability to accurately capture spatial relationships and structures within input images and masks. SCAFFOLD was originally developed to address the skewness of the label distribution in classification tasks, where each input is reduced to a single class label. In contrast, segmentation performs pixel-wise annotations, where spatial relationships and boundary and edge information are crucial. However, SCAFFOLD's variance correction mechanism (control variates) distorts spatial structures, boundary, and edge information, and introduces artifacts in the predicted masks, ultimately degrading segmentation performance. The performance of FedCLAM is worse than that of FedAvg and FedProx, as its design objective focuses on producing a well-fitted model for the data distribution of the participating clients in federation, while completely neglecting out-of-domain generalization.

D. Comparison as a Personalized Approach

This section evaluates FedGA as a *personalized approach*. Instead of building a global model, each client produces a customized model that adapts to its data distribution but that benefits from the knowledge aggregated from other clients' data. To implement this approach, we fine-tune the global model on each client's training set (80% of the data) and evaluate its performance on the validation set (20% of the data). The results are presented in Tables VI and VII for the breast lesion and prostate segmentation tasks, respectively.

In the breast lesion segmentation task, four personalized models achieved the highest DSC, and three attained the highest HD95 out of five personalized models produced by FedGA across five sites. For the remaining model, the DSC or HD95 of FedGA are comparable to those of the best-performing method, FedAvg, with no significant drop in score. The other three baseline methods fail to produce personalized models that outperform FedGA due to the reasons outlined in the previous section.

For the prostate MRI segmentation task, FedGA followed a similar trend, achieving the highest HD95 for three out of five personalized models. However, unlike breast lesion segmentation, FedBN outperformed FedGA in the remaining two models. All personalized models produced by FedBN were expected to deliver the best performance for both tasks, as its design explicitly targets delivering personalized models. Nevertheless, many of FedGA's personalized models across different sites offer the best high-precision segmentation for both tasks, further indicating FedGA's effectiveness.

A notable observation is that DSC drops to 35.86 for Site C in the prostate segmentation task, as reported in Table VII. This drop can be explained by the clearly visible and distinct tumor size distributions observed at Site C, as illustrated in the violin plot in Figure 6. The prostate size distribution at Site C is much wider and is shifted higher, with a median size much larger than at the other sites. A Kruskal-Wallis test also revealed a statistically significant difference in prostate size distributions ($H = 274.298$, p-value $= 3.340 \times 10^{-57}$). Post-hoc Mann-Whitney U tests with Holm correction confirmed that Site C differed significantly from all other sites (all

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3676321

TABLE V: Generalization performance of baseline methods (FedAvg, FedProx, SCAFFOLD, FedBN, FedCLAM) and our proposed method (FedGA) on out-of-distribution domains across two tasks: prostate segmentation and breast lesion segmentation. The best values are highlighted in bold.

Method	Breast Lesion Segmentation				Prostate Segmentation			
	DSC ↑	p-value	HD95 (mm) ↓	p-value	DSC ↑	p-value	HD95 (mm) ↓	p-value
FedAvg [15]	81.77±2.42	1.906e-04	59.14±2.67	1.976e-07	64.64±2.97	1.451e-03	83.23±3.24	1.936e-12
FedProx [4]	82.62±1.48	4.903e-04	60.62±1.98	6.960e-14	58.47±2.90	6.327e-11	97.28±2.73	4.200e-24
SCAFFOLD [14]	72.06±2.31	4.397e-22	83.22±3.27	8.644e-24	24.44±2.44	1.263e-31	150.40±2.40	1.967e-38
FedBN [17]	65.25±2.78	6.760e-22	110.34±2.63	9.471e-32	20.95±3.85	5.056e-27	75.52±3.58	3.175e-03
FedCLAM [24]	78.83±1.91	1.487e-11	62.37±1.52	1.961e-16	57.98±1.98	1.194e-14	89.21±1.93	6.782e-22
FedGA	**84.52 ±1.58**	–	**54.16±1.92**	–	**66.29±1.52**	–	**72.13±1.43**	–

TABLE VI: Personalized performance across federated sites for baseline methods (FedAvg, FedProx, SCAFFOLD, FedBN, FedCLAM) and the proposed method (FedGA) evaluated on site-specific test data for the **breast lesion segmentation** task. The best values are highlighted in bold, while the second-best values are underlined.

Method	DSC ↑					HD95 (mm) ↓				
	Site A	Site B	Site C	Site D	Site E	Site A	Site B	Site C	Site D	Site E
FedAvg [15]	76.90	<u>87.13</u>	83.66	**74.87**	<u>82.70</u>	**45.98**	30.59	<u>42.27</u>	**56.37**	<u>50.22</u>
FedProx [4]	75.79	86.61	83.34	74.11	80.99	54.69	30.31	42.99	61.69	57.32
SCAFFOLD [14]	67.38	82.87	81.26	63.37	74.19	56.76	62.78	45.56	83.71	66.40
FedBN [17]	<u>77.75</u>	84.04	80.17	63.58	78.06	52.85	48.67	51.33	83.59	59.20
FedCLAM [24]	76.87	85.49	<u>83.90</u>	71.89	82.63	55.94	<u>29.79</u>	40.87	58.11	52.17
FedGA	**78.49**	**87.94**	**84.69**	<u>74.85</u>	**83.55**	<u>46.08</u>	**28.36**	**36.29**	<u>56.90</u>	**48.25**

TABLE VII: Personalized performance across federated sites for baseline methods (FedAvg, FedProx, SCAFFOLD, FedBN, FedCLAM) and the proposed method (FedGA) evaluated on site-specific test data for the **prostate segmentation** task. The best values are highlighted in bold, while the second-best values are underlined.

Method	DSC ↑					HD95 (mm) ↓				
	Site A	Site B	Site C	Site D	Site E	Site A	Site B	Site C	Site D	Site E
FedAvg [15]	83.02	83.80	35.69	83.32	<u>86.28</u>	14.54	14.49	141.25	17.16	11.21
FedProx [4]	<u>83.53</u>	84.50	33.67	82.87	84.93	<u>13.17</u>	<u>12.65</u>	154.51	17.93	<u>10.97</u>
SCAFFOLD [14]	70.26	77.37	<u>54.62</u>	74.76	74.57	29.65	27.09	142.25	24.80	34.70
FedBN [17]	82.29	**85.44**	**72.50**	<u>84.41</u>	**88.47**	17.66	16.51	**110.10**	**13.66**	14.27
FedCLAM [24]	83.11	83.03	36.87	83.24	84.11	15.33	16.51	153.76	17.09	16.85
FedGA	**84.25**	<u>84.80</u>	35.86	**84.84**	85.75	**13.05**	**14.08**	<u>143.22</u>	<u>15.92</u>	**9.63**

TABLE VIII: Leave-one-site-out validation results for the breast lesion segmentation. The best values are highlighted in bold.

Method	DSC ↑							HD95 (mm) ↓						
	Site A	Site B	Site C	Site D	Site E	Site F	Mean	Site A	Site B	Site C	Site D	Site E	Site F	Mean
FedAvg [15]	71.19	**80.59**	**80.18**	**68.89**	81.61	81.77	77.37	64.20	**60.65**	52.84	81.67	53.97	59.14	62.08
FedGA	**73.35**	78.79	80.03	68.64	**82.22**	**84.52**	**77.93**	**55.98**	64.83	**48.16**	**74.27**	**52.28**	**54.16**	**58.28**

TABLE IX: Leave-one-site-out validation results for the prostate segmentation task. The best values are highlighted in bold.

Method	DSC ↑							HD95 (mm) ↓						
	Site A	Site B	Site C	Site D	Site E	Site F	Mean	Site A	Site B	Site C	Site D	Site E	Site F	Mean
FedAvg [15]	67.41	66.40	35.75	**73.06**	78.73	64.64	64.33	**36.32**	49.38	136.97	45.49	**19.27**	83.23	61.78
FedGA	**69.30**	**67.04**	**38.06**	72.80	**79.09**	**66.29**	**65.43**	36.80	**31.64**	**132.74**	**33.63**	19.33	**72.13**	**54.38**

corrected p-values $< 3.039083 \times 10^{-6}$), with large effect sizes (Cliff's $\delta > 0.496$). This deviation may arise from differences in patient demographics, segmentation protocols, or annotation inconsistencies. On the other hand, the DSC scores for the breast lesion segmentation task remain consistent between sites, as presented in Table VI, due to the lack of substantial variation in the tumor size distribution among them, as shown in Figure 7. The performance decline for Site C is consistent with our expectations, given that FedGA is primarily designed to deal with feature heterogeneity across sites rather

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3676321

than inconsistency in the segmentation protocol or annotation quality. Future work will aim to model and explicitly mitigate these additional sources of domain shift.

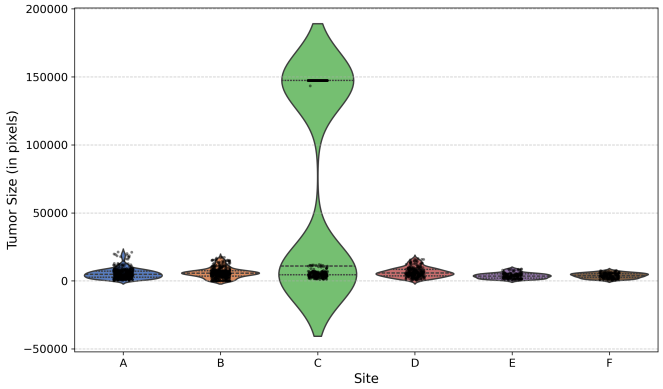


Fig. 6: Distribution of prostate sizes across sites in the prostate segmentation task.

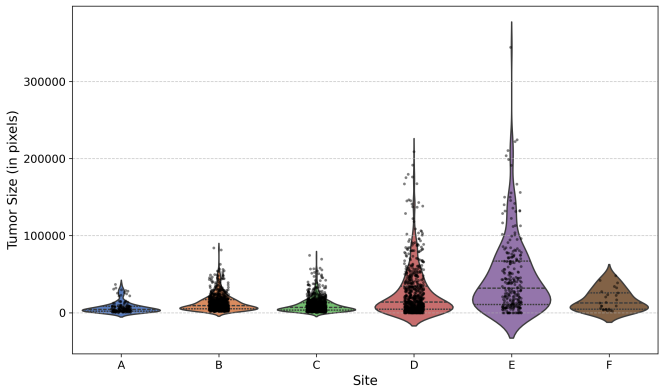


Fig. 7: Distribution of tumor sizes across sites in the breast lesion segmentation task.

E. Performance on Leave-One-Site-Out Validation

We also used a leave-one-site-out evaluation protocol to test the generalizability of FedGA on both tasks, following the approach outlined in [30]. This protocol iteratively designates one site as the FL server while using the remaining sites as clients for model training. We also implemented the FedAvg baseline using the same protocol for a fair comparison.

More specifically, this protocol evaluated the generalization of FedGA on two segmentation tasks using two distinct model architectures (Residual UNet for breast lesion segmentation and Attention UNet for prostate segmentation) and different FL settings in terms of the client hospitals, the number and selection of images used for metaheuristic optimization on the server. For example, for the unseen site A, hospitals B–F joined FL (as shown in Table I). The fitness function was evaluated using 451 images, corresponding to 50% of patients from site A. In contrast, for the unseen site F, the participating hospitals were sites A–E. Here, 12 exclusive images, representing 50% of site F's patients, were used for fitness evaluation. Thus, for each unseen site, at least one participating client differed and both the number and the image set for evaluation changed. In addition, two different network architectures were used for the two segmentation tasks. As shown in Table VIII and Table IX, FedGA outperforms FedAvg on both metrics on average in both tasks, across different model architectures and diverse FL settings. This demonstrates its generalizability.

FedGA also greatly improves the HD95 score, indicating more accurate and precise tumor boundary delineation. This precision in the boundary region ensures that radiation is applied to the tumor while protecting healthy tissue, reducing adverse effects, and improving patient outcomes. This also helps to decide the best intervention approach, whether surgery, radiation therapy, or targeted drug delivery in the case of cancer treatment.

This experiment protocol also provides a thorough validation of FedGA under varying numbers of images used for fitness evaluation, thereby addressing the question of how many unseen images are necessary for this purpose. In the breast lesion segmentation task, 451, 511, 397, 384, 50, 116 and 12 unseen images were employed for fitness evaluation, corresponding to 50% of the total images (a patient-level partition was performed whenever patient information was available) from the unseen sites A–H, respectively. Similarly, in the prostate segmentation task, 210, 192, 221, 130, 79 and 78 unseen images were used at the leave-out sites A–H, also representing 50% of the data from each site. The results presented in Table VIII and Table IX demonstrate that FedGA remains robust with respect to the number of unseen images used for fitness evaluation, with as few as 12 unseen images proving sufficient for fitness evaluation.

F. Computational and Communication Costs, Memory Usage, and Scalability Analysis

We empirically evaluate FedGA's overall computational and communication efficiency in comparison to FedAvg by analyzing their learning dynamics and training time. Specifically, we report the training time overhead introduced by FedGA on the server side and track the evolution of the DSC of the global models for both methods on unseen data throughout the training process for the breast lesion segmentation task. The same experimental setup described in Section VII-C is used here for the breast lesion segmentation task.

Table X presents the training times for both methods for a single FL round (client-side model training and server-side model aggregation and optimization) on average, computed over 20 rounds, and the percentage of computational overhead incurred by FedGA. At first glance, FedGA appears to introduce 15.04% computational overhead due to the server-side metaheuristic optimization. However, the learning dynamics shown in Figure 8 indicates that FedGA reached its highest DSC in the 13th communication round, outperforming FedAvg in both accuracy and communication efficiency. In comparison, FedAvg achieved its highest DSC in the 19th round. Converging six rounds earlier, FedGA reduces the computational burden on both the client and server sides: it saves 30 client-side model updates and 6 server-side aggregations and

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3676321

broadcasts, which account for 20.29% savings of computation, as shown in Table XI. Therefore, the additional server-side overhead imposed by FedGA is more than compensated for by the cumulative computational savings between clients. In addition, the reduction in client-side updates and communication rounds results in lower energy consumption and a reduced carbon footprint, making FedGA not only more efficient but also more environmentally sustainable, aligning well with the operational realities and ethical priorities FL deployments.

FedGA incurs additional memory usage only on the server side for processing candidate solutions during GA-based optimization. Memory usage depends on the model size used for segmentation, but it should have minimal impact on server performance, as server environments are usually equipped with enough memory for large-scale data processing.

In this context, FedGA is well-suited for typical healthcare scenarios. In such cases, developing high-quality and trustworthy models using limited and heterogeneous data sources while minimizing communication rounds and carbon footprint is more critical and desirable than reducing server-side computational overhead or memory consumption, as healthcare servers are not expected to be resource-constrained.

Regarding scalability with an increasing number of clients, FedGA generates a fixed number of initial candidate solutions (models) from the aggregated global model for GA-based optimization, irrespective of the number of clients. Consequently, the server-side workload remains constant as additional clients join the federation. In other words, the computational complexity of the GA module is $O(1)$. Therefore, FedGA is expected to maintain consistent performance regardless of the number of participating clients.

TABLE X: Average runtime (in seconds) per training round for FedAvg and FedGA, and percentage of runtime overhead introduced by FedGA, calculated over 20 training rounds.

FedAvg (s)	FedGA (s)	Overhead (%)
904.18	1040.18	+15.04%

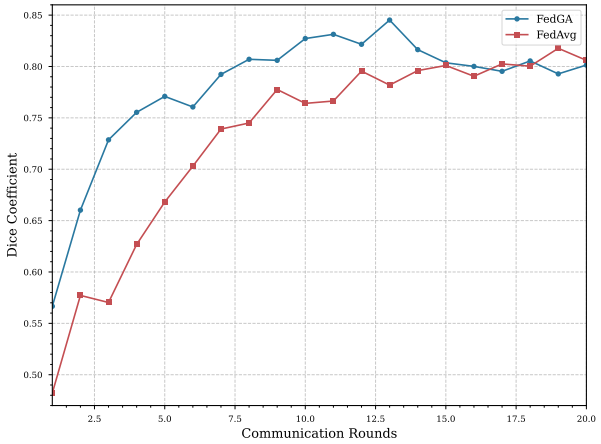


Fig. 8: DSCs of global models for FedGA and FedAvg on unseen data of *Site F* over 20 communication rounds of FL for the breast lesion segmentation task.

TABLE XI: Optimal runtime (calculated as the average runtime per round multiplied by the number of rounds required for convergence) for each method (FedAvg and FedGA), along with the cumulative time savings (%) achieved by FedGA.

FedAvg (s)	FedGA (s)	Saving (%)
17179.42	13522.34	-21.29%

G. Ablation Study of GA Parameters

Finally, we conduct an ablation study for the two most critical parameters of our GA module: P_c and P_m. The parameter space considered for P_c is $\{0.4, 0.5, 0.6\}$, while for P_m it is $\{0.05, 0.1, 0.2\}$. The same experimental setup described in Section VII-C is used here for breast lesion segmentation. The DSC and HD95 values for each P_c-P_m pair are shown as heatmaps in Figure 9. A P_c of 0.5 and a P_m of 0.1 –as we have used so far– yielded the highest scores, balancing exploration and exploitation. High crossover and mutation rates make the GA too random, lowering performance. In contrast, high crossover with low mutation rates leads to premature convergence due to inadequate diversity.

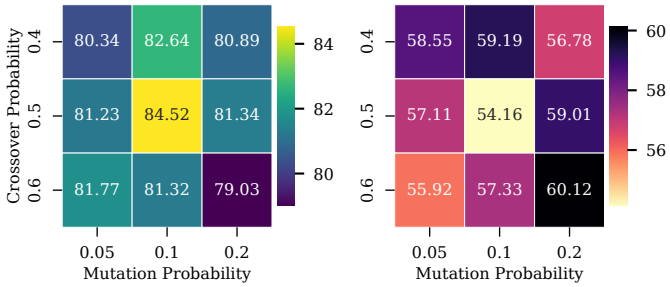


Fig. 9: Heatmaps from the ablation study of GA parameters, showing the DSC (left) and HD95 (right) across different combinations of crossover and mutation probabilities.

VIII. Potential Clinical Impact

The potential clinical impact of FedGA lies in its ability to facilitate the joint development of a robust and generalized clinical decision support tool for medical image segmentation across diverse healthcare institutions (despite variations in patient demographics, imaging protocols, and imaging systems), while ensuring patient privacy and compliance with relevant regulations. Moreover, by allowing access to diverse datasets from multiple institutions, FedGA could facilitate more reliable diagnosis and treatment planning, particularly for underrepresented populations and hospitals with limited data, computational resources or clinical expertise. This could eventually improve patient outcomes and support broader and more equitable access to AI-assisted healthcare.

It is worth mentioning is that the variability in annotation quality and the scale of participating institutions can affect the generalizability of the model. Further validation on larger, multi-institutional datasets and across diverse imaging modalities is essential to confirm its future clinical use.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Journal of Biomedical and Health Informatics. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JBHI.2026.3676321

IX. Conclusions and Future Work

We have presented FedGA, a method that integrates a genetic algorithm into the server-side global model optimization process to address the fundamental challenge of feature distribution shift across clients in FL for (medical image) segmentation. A comprehensive evaluation of two segmentation tasks across two different medical imaging modalities shows substantial improvements versus related works.

Future work will explore alternative metaheuristic approaches, such as *differential evolution* [34] and *particle swarm optimization* [35], as potential replacements for genetic algorithms to reduce the server-side computational overhead of model optimization. We also plan to investigate FedGA's performance and generalizability in federated multimodal datasets and multitask learning [36] settings for simultaneous medical image segmentation and classification tasks.

References

[1] Q. Li, Y. Diao, Q. Chen, and B. He, "Federated learning on non-IID data silos: An experimental study," in *2022 IEEE 38th international conference on data engineering (ICDE)*. IEEE, 2022, pp. 965–978.

[2] H. Guan, P.-T. Yap, A. Bozoki, and M. Liu, "Federated learning for medical image analysis: A survey," *Pattern Recognition*, p. 110424, 2024.

[3] A. E. Durmus, Z. Yue, M. Ramon, M. Matthew, W. Paul, and S. Venkatesh, "Federated learning based on dynamic regularization," in *International conference on learning representations*, 2021.

[4] T. Li, A. K. Sahu, M. Zaheer, M. Sanjabi, A. Talwalkar, and V. Smith, "Federated optimization in heterogeneous networks," *Proceedings of Machine learning and systems*, vol. 2, pp. 429–450, 2020.

[5] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[6] M. Tan and Q. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," in *International conference on machine learning*. PMLR, 2019, pp. 6105–6114.

[7] 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: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*. Springer, 2015, pp. 234–241.

[8] V. Badrinarayanan, A. Kendall, and R. Cipolla, "Segnet: A deep convolutional encoder-decoder architecture for image segmentation," *IEEE transactions on pattern analysis and machine intelligence*, vol. 39, no. 12, pp. 2481–2495, 2017.

[9] J. Chen, J. Mei, X. Li, Y. Lu, Q. Yu, Q. Wei, X. Luo, Y. Xie, E. Adeli, Y. Wang *et al.*, "Transunet: Rethinking the U-Net architecture design for medical image segmentation through the lens of transformers," *Medical Image Analysis*, p. 103280, 2024.

[10] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 801–818.

[11] J. Wang, Q. Liu, H. Liang, G. Joshi, and H. V. Poor, "Tackling the objective inconsistency problem in heterogeneous federated optimization," *Advances in neural information processing systems*, vol. 33, pp. 7611–7623, 2020.

[12] J. Hu, J. Du, Z. Wang, X. Pang, Y. Zhou, P. Sun, and K. Ren, "Does differential privacy really protect federated learning from gradient leakage attacks?" *IEEE Transactions on Mobile Computing*, 2024.

[13] J. Wang, Y. Mao, Z. Wang, J. Wu, and J. Wu, "Joint focal loss and dominant gradient correction for gradient conflict in federated learning," in *2023 IEEE 20th International Conference on Mobile Ad Hoc and Smart Systems (MASS)*. IEEE, 2023, pp. 243–251.

[14] S. P. Karimireddy, S. Kale, M. Mohri, S. Reddi, S. Stich, and A. T. Suresh, "Scaffold: Stochastic controlled averaging for federated learning," in *International conference on machine learning*. PMLR, 2020, pp. 5132–5143.

[15] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. Agüera y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[16] Y. Shi, J. Liang, W. Zhang, V. Y. Tan, and S. Bai, "Towards understanding and mitigating dimensional collapse in heterogeneous federated learning," *arXiv:2210.00226*, 2022.

[17] X. Li, M. Jiang, X. Zhang, M. Kamp, and Q. Dou, "Fedbn: Federated learning on non-IID features via local batch normalization," *arXiv:2102.07623*, 2021.

[18] S. Ioffe, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," *arXiv:1502.03167*, 2015.

[19] Q. Li, B. He, and D. Song, "Model-contrastive federated learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 10 713–10 722.

[20] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PmLR, 2020, pp. 1597–1607.

[21] X. Xu, H. H. Deng, T. Chen, T. Kuang, J. C. Barber, D. Kim, J. Gateno, J. J. Xia, and P. Yan, "Federated cross learning for medical image segmentation," in *Medical Imaging with Deep Learning*. PMLR, 2024, pp. 1441–1452.

[22] J. Wang, Y. Jin, D. Stoyanov, and L. Wang, "Feddp: Dual personalization in federated medical image segmentation," *IEEE Transactions on Medical Imaging*, vol. 43, no. 1, pp. 297–308, 2023.

[23] M. Jiang, H. R. Roth, W. Li, D. Yang, C. Zhao, V. Nath, D. Xu, Q. Dou, and Z. Xu, "Fair federated medical image segmentation via client contribution estimation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 16 302–16 311.

[24] V. Siomos, J. Passerat-Palmbach, and G. Tarroni, " FedCLAM: Client Adaptive Momentum with Foreground Intensity Matching for Federated Medical Image Segmentation ," in *proceedings of Medical Image Computing and Computer Assisted Intervention – MICCAI 2025*, vol. LNCS 15965. Springer Nature Switzerland, September 2025.

[25] C. Cavallaro, V. Cutello, M. Pavone, and F. Zito, "Machine learning and genetic algorithms: A case study on image reconstruction," *Knowledge-Based Systems*, vol. 284, p. 111194, 2024.

[26] C. W. Ahn and R. S. Ramakrishna, "Elitism-based compact genetic algorithms," *IEEE Transactions on Evolutionary Computation*, vol. 7, no. 4, pp. 367–385, 2003.

[27] N. Bloch, A. Madabhushi, H. Huisman, J. Freymann, J. Kirby, M. Grauer, A. Enquobahrie, C. Jaffe, L. Clarke, and K. Farahani, "NCI-ISBI 2013 Challenge: Automated Segmentation of Prostate Structures," *The Cancer Imaging Archive*, vol. 370, pp. 6, 12, 2015.

[28] G. Lemaître, R. Martí, J. Freixenet, J. C. Vilanova, P. M. Walker, and F. Meriaudeau, "Computer-aided detection and diagnosis for prostate cancer based on mono and multi-parametric MRI: a review," *Computers in biology and medicine*, vol. 60, pp. 8–31, 2015.

[29] G. Litjens, R. Toth, W. Van De Ven, C. Hoeks, S. Kerkstra, B. Van Ginneken, G. Vincent, G. Guillard, N. Birbeck, J. Zhang *et al.*, "Evaluation of prostate segmentation algorithms for mri: the promise12 challenge," *Medical image analysis*, vol. 18, no. 2, pp. 359–373, 2014.

[30] Q. Liu, C. Chen, J. Qin, Q. Dou, and P.-A. Heng, "Feddg: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 1013–1023.

[31] W. Gómez-Flores, M. J. Gregorio-Calas, and W. Coelho de Albuquerque Pereira, "Bus-bra: a breast ultrasound dataset for assessing computer-aided diagnosis systems," *Medical Physics*, vol. 51, no. 4, pp. 3110–3123, 2024.

[32] W. Al-Dhabyani, M. Gomaa, H. Khaled, and A. Fahmy, "Dataset of breast ultrasound images," *Data in brief*, vol. 28, p. 104863, 2020.

[33] A. A. Ardakani, A. Mohammadi, M. Mirza-Aghazadeh-Attari, and U. R. Acharya, "An open-access breast lesion ultrasound image database: Applicable in artificial intelligence studies," *Computers in Biology and Medicine*, vol. 152, p. 106438, 2023.

[34] A. K. Qin, V. L. Huang, and P. N. Suganthan, "Differential evolution algorithm with strategy adaptation for global numerical optimization," *IEEE transactions on Evolutionary Computation*, vol. 13, no. 2, pp. 398–417, 2008.

[35] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of ICNN'95-international conference on neural networks*, vol. 4. ieee, 1995, pp. 1942–1948.

[36] Y. Zhou, H. Chen, Y. Li, Q. Liu, X. Xu, S. Wang, P.-T. Yap, and D. Shen, "Multi-task learning for segmentation and classification of tumors in 3d automated breast ultrasound images," *Medical image analysis*, vol. 70, p. 101918, 2021.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/