Medical Image Analysis 105 (2025) 103714

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Efficient one-shot federated learning on medical data using knowledge distillation with image synthesis and client model adaptation

Myeongkyun Kang [a,b], Philip Chikontwe [c], Soopil Kim [a,b], Kyong Hwan Jin [d], Ehsan Adeli [b], Kilian M. Pohl [b], Sang Hyun Park [a,*]

[a] Department of Robotics and Mechatronics Engineering, Daegu Gyeongbuk Institute of Science and Technology (DGIST), Daegu, Republic of Korea
[b] Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA 94305, USA
[c] Department of Biomedical Informatics, Harvard Medical School, Boston, MA 02115, USA
[d] School of Electrical Engineering, Korea University, Seoul, Republic of Korea

ARTICLE INFO

Dataset link: https://github.com/myeongkyunkang/E-FedISCA

Keywords:
Client model adaptation
Image synthesis
Knowledge distillation
Noise
One-shot federated learning

ABSTRACT

One-shot federated learning (FL) has emerged as a promising solution in scenarios where multiple communication rounds are not practical. Though previous methods using knowledge distillation (KD) with synthetic images have shown promising results in transferring clients' knowledge to the global model on one-shot FL, overfitting and extensive computations still persist. To tackle these issues, we propose a novel one-shot FL framework that generates pseudo intermediate samples using mixup, which incorporates synthesized images with diverse types of structure noise. This approach (i) enhances the diversity of training samples, preventing overfitting and providing informative visual clues for effective training and (ii) allows for the reuse of synthesized images, reducing computational resources and improving overall training efficiency. To mitigate domain disparity introduced by noise, we design noise-adapted client models by updating batch normalization statistics on noise to enhance KD. With these in place, the training process involves iteratively updating the global model through KD with both the original and noise-adapted client models using pseudo-generated images. Extensive evaluations on five small-sized and three regular-sized medical image classification datasets demonstrate the superiority of our approach over previous methods.

1. Introduction

Though the accuracy of machine learning is influenced by the size of the dataset, the difficulty in centralizing data poses challenges in robust model training (Kang et al., 2023c). Federated learning (FL) has emerged as a solution to facilitate collaborative model training while maintaining data privacy compliance (Li et al., 2020b; Yang et al., 2021). However, FL methods often require multiple communication rounds during training, limiting their applicability in extreme scenarios where (a) communication is prohibitively expensive, (b) model transfer is infeasible, and (c) minimizing the risk of attack is crucial (Kairouz et al., 2021; Zhang et al., 2022). Consider a scenario where (a) patient data is stored on an isolated local server secured with firewalls and network segmentation (Eichelberg et al., 2020), and accessible only within a closed offline network. In such cases, collaborators need to copy the trained client model to another server with external network access so that the model can be transferred to the central server. Alternatively, researchers may need to visit hospitals to train the client models onsite and then transfer them to the central server in person. Frequent model transfers significantly increase both communication costs and training time, and ultimately limiting the practicality of FL. Furthermore, (b) when access to a dataset is restricted to a limited period due to privacy regulations or storage constraints, client models can no longer be trained, making FL infeasible. In addition, (c) frequent communication in poor network infrastructure increases the risk of interception by attackers, making it necessary to reduce communication rounds to minimize security risks (Park et al., 2021; Chen et al., 2024). To overcome these limitations, one-shot FL serves as a viable solution. One-shot FL (Zhang et al., 2022) enables collaborative model training through a single communication round, thereby addressing the aforementioned challenges by: (a) reducing communication costs, (b) enabling legacy model integration, and (c) minimizing the risk of attacks.

While FL can handle client heterogeneity through multiple communication rounds (Kang et al., 2024), it remains challenging in one-shot scenarios, leading to low-accuracy global models. To address this, previous one-shot FL methods use generated images and knowledge

* Corresponding author.
E-mail address: shpark13135@dgist.ac.kr (S.H. Park).

https://doi.org/10.1016/j.media.2025.103714
Received 27 September 2023; Received in revised form 25 March 2025; Accepted 30 June 2025
Available online 14 July 2025
1361-8415/© 2025 Elsevier B.V. All rights are reserved, including those for text and data mining, AI training, and similar technologies.

distillation (KD) (Hinton et al., 2015) to transfer knowledge from multiple client models for global model training (Zhang et al., 2022). However, the limited diversity of these generated images often results in overfitting, a significant drawback for one-shot FL in practice. To tackle this issue, Zhang et al. (2022) and Micaelli and Storkey (2019) propose an approach to enhance the transferability of client models by generating diverse natural images near the decision boundary. However, the decision boundaries in medical data often tend to be more complex compared to natural images, limiting the applicability of existing one-shot FL methods in medical data. Particularly, our preliminary work (Kang et al., 2023a) demonstrated that the feature distribution in medical data is more complex than in natural images, a finding further supported by Konz et al. (2022) and Konz and Mazurowski (2024). Konz et al. (2022) conducted extensive experiments demonstrating that learning the intrinsic dimension (Pope et al., 2021) of medical images is more challenging than for natural images and requires larger datasets for training, even though medical data being represented in a lower-dimensional space. Therefore, in the medical domain, it is crucial to capture the features of samples with ambiguous distributions, as even small changes in data distribution can result in label shifts, also referred to as higher label sharpness in a dataset (Konz and Mazurowski, 2024). In this context, though ambiguous and unrealistic generated images may perform sufficiently in one-shot FL with natural images, they could fail in medical image analysis.

In this paper, we propose a one-shot FL method to avoid global model overfitting by utilizing synthesized images ranging from random noise to realistic images. Unlike the previous one-shot FL methods that select the best image as a training source for KD, we stored all intermediate samples generated during image synthesis and used them for training. Additionally, we designed noise-adapted client models employing adaptive batch normalization (AdaBN) (Li et al., 2018), which enhances the KD signal for random noise. With these methodologies, we trained a global model through KD using both the original and noise-adapted client models with all synthetic images as a training source. The preliminary work was introduced at a conference (Kang et al., 2023a), but it faced a significant computational challenge for image synthesis (see Fig. 5), limiting the practical applicability of one-shot FL.

To address this, we synthesize images only once and then reuse them as a training source for KD rather than performing image synthesis multiple times. Since reusing a limited number of synthesized images may not be sufficient to solve the overfitting problem, we employ both structure noise (Baradad Jurjo et al., 2021) and *mixup* (Zhang et al., 2018) to enhance the diversity of the synthesized images and ensure that images possess properties of the target medical task. Note that reusing synthesized images by leveraging structured noise and *mixup* distinguishes this approach from our preliminary work (Kang et al., 2023a), significantly reducing computational resources. Furthermore, combining structure noise and *mixup* leads to better visual representations, ultimately enhancing model accuracy. Specifically, structural noise, which has demonstrated effectiveness in visual representation learning (Baradad Jurjo et al., 2021; Kataoka et al., 2020), can serve as an informative source for training, thereby reducing computational costs. However, due to style and semantic differences between natural and medical images (Morra et al., 2021; Konz et al., 2022), structural noise is limited in its applicability for datasets with significant differences in style, e.g., X-ray. Empirical experiments will demonstrate that incorporating even a small number of synthesized images through *mixup* significantly improves accuracy (see Table 2). This implies that synthesized images that possess properties of the target medical task play a crucial role in medical one-shot FL. In summary, this approach not only reduces computational resources but also contributes to diverse and valuable image synthesis, leading to a model with higher accuracy.

In the training process, we first gather client models on the central server, where each client model is trained on its own dataset. Next, we synthesize images from random noise once and store intermediate samples in memory that exhibit a loss lower than a certain threshold. We then generate pseudo images by adjusting the noise level of *mixup* using synthesized images and structure noise. Additionally, we design noise-adapted client models using AdaBN, enhancing the KD signal for structure noise. Finally, we train a global model through KD with both the original and noise-adapted client models using pseudo-generated images, repeating this process until the global model converges. In summary, the contributions of our work are as follows:

- We introduce a novel one-shot FL approach that utilizes synthetic images and structure noise with client model adaptation. This enables effective knowledge transfer from client models to the global model, preventing overfitting and providing a wide range of synthetic images containing informative visual clues.
- We propose an efficient one-shot FL by leveraging *mixup* and structure noise. This optimization significantly reduces computational resources while enhancing the diversity of the synthesized images, leading to improved model accuracy.
- We evaluate our method in multi-shot and natural image settings to highlight its effectiveness in diverse settings and demonstrate its efficiency by comparing the computation speed to other methods.

2. Related work

2.1. One-shot FL

Federated Learning (FL) is a distributed learning paradigm that trains a global model without sharing private data (Lu et al., 2022; Kim et al., 2024). To address privacy concerns and reduce communication costs, one-shot FL has emerged as a promising solution, enabling global model training with a single communication round. Due to the challenges of one-shot FL, prior methods have used public data. Guha et al. (2019) propose a one-shot FL method that trains support vector machines (SVMs) using unlabeled proxy data, while Li et al. (2020a) extend PATE (Private Aggregation of Teacher Ensembles) (Papernot et al., 2017) for one-shot FL by utilizing an unlabeled dataset for training. Apart from that, Zhou et al. (2020) uses dataset distillation, and Dennis et al. (2021) shares additional information for training. However, these assumptions may not hold in many real-world scenarios, limiting the applicability of these methods. Recently, Zhang et al. (2022) proposed DENSE, an approach for one-shot FL that transfers knowledge from an ensemble of client models using KD and generated images. DENSE aims to enhance the transferability of client models by generating diverse natural images near the decision boundary. However, DENSE is not appropriate for one-shot FL in the medical domain, due to the complexity of decision boundaries in medical data (Konz et al., 2022; Chikontwe et al., 2021). In contrast, we utilize synthesized images ranging from noise to realistic images rather than using a generator to diversify image generation. This prevents overfitting while enabling effective KD in one-shot FL settings.

2.2. Data-free KD

Knowledge distillation (KD) (Hinton et al., 2015) transfers knowledge from a teacher model to a student model by minimizing the Kullback–Leibler divergence between the student and teacher predictions (Yang et al., 2023b; Wang et al., 2023; Xie et al., 2023). However, this approach may not be applicable when images are not available during training. To address this limitation, data-free KD methods have been proposed, which involve generating images and using them as a training source for KD. DAFL (Chen et al., 2019) employs a generator with a teacher model as a discriminator to build a training dataset. ZSKT (Micaelli and Storkey, 2019) synthesizes images that exhibit a mismatch between the student and teacher models, enhancing the

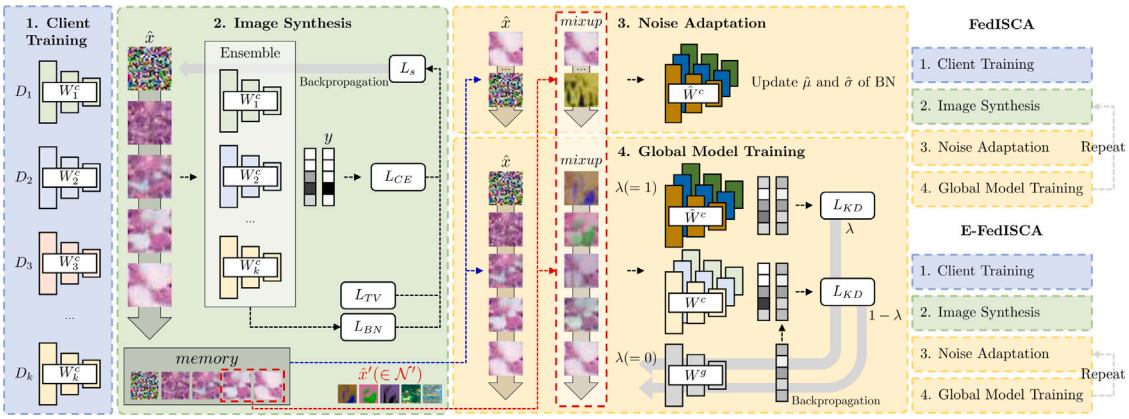

Fig. 1. Illustration of our proposed method. W_k^c denotes a client model with respect to data D_k and W^g denotes a global model. W^c denotes original client models and $\hat{W}^c$ denotes noise-adapted client models. $\hat{x}$ denotes random noise and λ indicates noise level. $\mathcal{N}'$ denotes diverse types of structure noise (Baradad Jurjo et al., 2021) and *mixup* generates a pseudo input with the corresponding noise level λ. For FedISCA, $\hat{x}$ is optimized to exhibit properties similar to all D_k datasets using losses L_{CE}, L_{BN}, and L_{TV}. Subsequently, all synthesized images serve as the training source for KD during the global model training (depicted by the blue dashed arrow). For E-FedISCA, we select the synthesized images that exhibit a lower loss (L_s) than a specified threshold (L_{th}). We then generate pseudo intermediate samples by applying *mixup* with structure noise $\mathcal{N}'$ and synthesized images. These pseudo intermediate samples are utilized for global model training (depicted by the red dashed arrow).

robustness of the trained student model. An alternative approach to image synthesis involves optimizing RGB pixels to construct a training dataset. DeepDream (Mordvintsev et al., 2015) synthesizes images corresponding to specific labels using cross-entropy and regularization losses. To improve image synthesis quality, DeepInversion (Yin et al., 2020) additionally minimizes feature statistics in batch normalization (BN) layers, enforcing feature similarity from low to high levels. Although these data-free methods have been applied to various settings, they have not been explicitly proposed for one-shot FL scenarios. Furthermore, our approach differs from these methods which choose the best image as the training source for KD. Instead, we utilize all intermediate synthesized images to prevent overfitting. While Raikwar and Mishra (2022) introduced a method for KD using random noise as a training source, it requires real images during training and needs multiple iterative updates to BN layer statistics. In contrast, our one-shot FL method does not necessitate real images during training and is more suitable for real-world settings.

3. Method

We introduce a novel one-shot FL, which leverages KD and synthesized images. We first offer the backgrounds of the key components of our method, including federated learning, knowledge distillation, image synthesis, and model adaptation. Subsequently, we describe our proposed method: (i) a method that uses synthesized images ranging from noise to realistic images with client model adaptation (FedISCA) and (ii) an efficient method that uses mixup and structure noise for improving training efficiency (E-FedISCA).

3.1. Backgrounds

3.1.1. Federated learning

Federated learning (FL) aims to train a global model W^g that represents all K datasets $D = \{D_1, \dots, D_k\}$ without sharing each client data D_k. Formally,

$$W^{g*} = arg\min_{W^g} L(W^g) = \frac{1}{K}\sum_{k=1}^{K} L_k(W^g), \tag{1}$$

where $L(\cdot)$ represents the loss for the global model and $L_k(\cdot)$ represents the loss on the kth client model. However, one-shot FL is restricted to training a global model with a single communication round, indicating that K client models $W^c = \{W_1^c, \dots, W_k^c\}$ with respect to data D_k are only available during the global model training. Since the accuracy of previous FL methods, such as federated averaging (McMahan et al., 2017), heavily relies on multiple communication rounds, a single transmission of parameters significantly degrades the accuracy of the global model ($= \frac{1}{K}\sum_k W_k^c$). To address this issue, recent methods (Zhang et al., 2022) employ KD (Hinton et al., 2015) to facilitate the transfer of knowledge from client models W^c to the global model W^g. In the following section, details regarding KD are described.

3.1.2. Knowledge distillation

Knowledge distillation (KD) is a method that transfers knowledge from a teacher model (i.e., a model trained with more parameters and data) to a student model. As stated in Hinton et al. (2015), soft targets, which are the smoothed class probability distributions produced by the teacher model, provide richer information compared to hard labels, enabling more effective KD. Soft targets corresponding to the ith class can be estimated using the softmax function as follows:

$$p(z_i, T) = \frac{exp(z_i/T)}{\sum_j exp(z_j/T)}, \tag{2}$$

where z_i is the logit for the ith class of a model. T denotes the temperature hyper-parameter, which controls the smoothness of the probability distribution; increasing T results in a smoother probability distribution.

In FL, our goal is to train a global model representing K datasets; therefore, averaging the logits of client models W^c (i.e., ensembling) is set as the teacher model, while the global model W^g is set as the student model. Formally, the logits of the teacher model z^c are defined as:

$$z^c = \frac{1}{K}\sum_{k=1}^{K} z_k^c, \tag{3}$$

where z_k^c denotes the logits of the kth client model W_k^c, i.e., $z_k^c = W_k^c(\cdot)$, and the notation for the ith class is omitted to avoid confusion. Given the client models (i.e., teacher model) W^c and the corresponding ensemble of logits z^c, along with the global model (i.e., student model) W^g and the corresponding logits z^g, KD is defined as follows:

$$L_{KD}(z^c, z^g) = KL(p(z^c, T), p(z^g, T)), \tag{4}$$

where $KL(\cdot)$ represents the Kullback–Leibler divergence between the two probability distributions. Through KD, knowledge is transferred from the client models W^c to the global model W^g, enabling W^g to represent all K datasets. However, the restricted access to the datasets D poses a challenge in applying KD to one-shot FL. Therefore, prior works (Zhang et al., 2022; Chen et al., 2019; Yin et al., 2020) employ synthesized images $\hat{x}$ as a training source for KD instead of using real images from D. Details regarding the image synthesis process are described in the following section.

3.1.3. Image synthesis

Given a $\mathcal{N}(0,1)$ initialized image $\hat{x} \in \mathbb{R}^{H\times W\times C}$, where H, W, and C represent the height, width, and number of channels, we optimize the RGB pixels of $\hat{x}$ with a specific label y using client models W^c by minimizing the following:

$$L_s(\hat{x}, y; W^c) = L_{CE}(\hat{x}, y; W^c) + \lambda_{BN} L_{BN}(\hat{x}; W^c) + \lambda_{TV} L_{TV}(\hat{x}; W^c), \tag{5}$$

where L_{CE} is the cross-entropy loss, L_{BN} is the batch normalization (BN) loss, and L_{TV} is the total variation loss (Mahendran and Vedaldi, 2015). To balance the losses, hyper-parameters λ_{BN} and λ_{TV} are used. The cross-entropy loss ensures that the synthesized image $\hat{x}$ corresponds to the label y by employing the client models, i.e., $\frac{1}{K}\sum_k CE(W_k^c(\hat{x}), y)$, while the total variation loss encourages consistency in the image synthesis process. The BN loss enforces feature similarity from low to high levels by minimizing the distance between the statistics of feature maps. This ensures that $\hat{x}$ and the real images in dataset D are similar at all levels, resulting in high-quality image synthesis. The BN loss is formulated as follows:

$$L_{BN}(\hat{x}) = \frac{1}{K}\sum_{k=1}^{K}\sum_{l}(\|\mu_{k,l}(\hat{x}) - \mu_{k,l}\| + \|\sigma_{k,l}^2(\hat{x}) - \sigma_{k,l}^2\|), \tag{6}$$

where $\mu_{k,l}(\hat{x})$ and $\sigma_{k,l}^2(\hat{x})$ denote the batch-wise mean and variance corresponding to the lth BN layer of the kth client model W_k^c for $\hat{x}$, while $\mu_{k,l}$ and $\sigma_{k,l}^2$ are the stored statistics of the lth BN layer of the trained client model W_k^c, respectively. Note that the synthesized image $\hat{x}$ can be used as an alternative input to the dataset D for KD, facilitating the transfer of knowledge from the client models W^c to the global model W^g. The following section describes model adaptation, which mitigates the detrimental impact of visual differences in samples (Kang et al., 2023b), enhancing the effectiveness of our approach.

3.1.4. Model adaptation

Model adaptation mitigates prediction degradation caused by domain discrepancies. Employing adaptive batch normalization (AdaBN) can resolve domain discrepancies by developing an adapted model (Li et al., 2018). AdaBN facilitates the model's adaptation to the target domain by updating the stored mean $\mu_{k,l}$ and variance $\sigma_{k,l}^2$ statistics of the lth BN layer of the kth client model W_k^c, as follows:

$$\mu_{k,l} = \alpha\mu_{k,l} + (1-\alpha)\mu_{k,l}(\hat{x}),\ \sigma_{k,l}^2 = \alpha\sigma_{k,l}^2 + (1-\alpha)\sigma_{k,l}^2(\hat{x}), \tag{7}$$

where α represents the momentum, and $\mu_{k,l}(\hat{x})$ and $\sigma_{k,l}^2(\hat{x})$ denote the batch-wise mean and variance for the lth BN layer of the kth client model W_k^c for a target domain image $\hat{x}$. The assumption underlying model adaptation is that domain-specific knowledge is stored in the BN statistics, while label-related knowledge is stored in the model weights. By adjusting the BN statistics, the adapted model enhances its predictions for samples from different domains. This contributes to an overall improvement in KD and ensures that the client model becomes better equipped to handle samples from various domains.

3.2. FedISCA: Federated learning using image synthesis and client model adaptation

Despite the restricted access to dataset D, KD with the ensemble of logits z^c from the client models W^c and synthesized images $\hat{x}$ enables the training of a global model W^g that represents all K datasets. However, overfitting remains a major challenge in one-shot FL. To address this issue, we adopt a diverse range of samples, including random noise $\mathcal{N}(0,1)$ and D's characteristic synthetic images, for global model training. Though this contributes to preventing overfitting, visual differences between random noise $\mathcal{N}(0,1)$ and samples from D can result in poor model predictions, negatively impacting the training process. To mitigate this problem, we introduce noise-adapted client models, which compensate for KD signals for random noise through model adaptation.

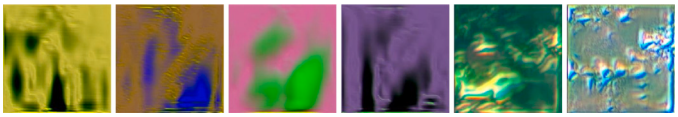

Fig. 2. Samples of structure noise (Baradad Jurjo et al., 2021) that used to generate a pseudo intermediate samples.

The training processes are illustrated in Fig. 1 and Algorithm 1. Initially, we gather trained client models W^c on the central server. Next, we synthesize $\mathcal{N}(0,1)$ initialized images $\hat{x}$ to possess the characteristics of D by optimizing Eq. (5) for S steps. During image synthesis, we store all intermediate samples sequentially in *memory*. Following, the stored samples ranging from $\mathcal{N}(0,1)$ to D are used to update the BN statistics of W^c using Eq. (7) for S steps. The updated BN statistics $\hat{\mu}$ and $\hat{\sigma}^2$ are applied to prepare noise-adapted client models denoted as $\hat{W}^c$. Lastly, we train W^g using both W^c and $\hat{W}^c$ with KD for S steps. The robust training of the global model while avoiding negative impacts from random noise $\mathcal{N}(0,1)$ is defined as:

$$L_d(\hat{z}^c, z^c, z^g, \lambda) = \lambda L_{KD}(\hat{z}^c, z^g) + (1-\lambda)L_{KD}(z^c, z^g), \tag{8}$$

where λ denotes the noise level, defined as $1 - s/S$ for the sth step, and z^g indicates the logits of the global model W^g for $\hat{x}$, i.e., $W^g(\hat{x})$. Additionally, $\hat{z}^c$ and z^c indicate the ensemble logits of the noise-adapted client models $\hat{W}^c$ and the client models W^c for $\hat{x}$, respectively. To clarify, random noise helps avoid overfitting, while noise-adapted client models improve the KD signal for random noise, thereby enhancing the robustness of global model training. Once *(1) Client Training* is completed, the processes *(2) Image Synthesis*, *(3) Noise Adaptation*, and *(4) Global Model Training* are repeated until the global model W^g converges.

3.3. E-FedISCA: Efficient federated learning using image synthesis and client model adaptation

FedISCA has demonstrated the ability to train a robust global model, but its computational demands during training limit its widespread use in real-world applications. Among the FedISCA training processes, image synthesis requires the most computational resources (see Fig. 5) thus reusing the synthesized images significantly reduces overall computations and enhances training efficiency. To accomplish this, we propose a new learning framework that leverages pseudo input, incorporating diverse types of structure noise $\hat{x}' \in \mathcal{N}'$ (Baradad Jurjo et al., 2021) (see Fig. 2) and *mixup* (Zhang et al., 2018). Formally, *mixup* with $\hat{x}$ and $\hat{x}'$ is defined as:

$$mixup(\hat{x}, \hat{x}', \lambda) = \lambda\hat{x}' + (1-\lambda)\hat{x}, \tag{9}$$

where λ denotes a noise level. This allows us to generate pseudo intermediate samples that range from $\mathcal{N}'$ to D by adjusting λ using *mixup*. In contrast to FedISCA, which synthesizes all training samples using Eq. (7), our approach first synthesizes a set of $\hat{x}$ that exhibits the characteristics of D. We then generate pseudo inputs by applying *mixup* to $\hat{x}$ and $\hat{x}'$ with the corresponding noise level λ, and use these generated pseudo inputs for global model training.

The overall training processes are illustrated in Fig. 1 and Algorithm 2. First, we gather trained client models on the central server. Next, we synthesize $\hat{x}$ to possess the characteristics of D by optimizing Eq. (5) for S steps. During image synthesis, n intermediate samples that exhibit lower loss (L_s) than a threshold (L_{th}) are stored in *memory*. Subsequently, we employ *mixup* to generate pseudo inputs ranging from random noise $\mathcal{N}'$ to D, and adjust the BN statistics μ and σ^2 for the noise-adapted client models $\hat{W}^c$. Finally, we train W^g using Eq. (8) for S steps by feeding the generated pseudo images as a training source for KD. Unlike FedISCA, this approach completes the *(1) Client Training* and *(2) Image Synthesis* stages first, and then iteratively repeats

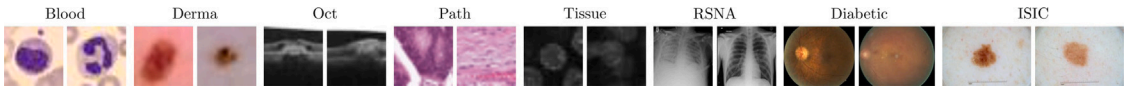

Fig. 3. Sample images of the dataset used in this work. Blood, Derma, Oct, Path, and Tissue are small-sized (28 × 28) datasets and RSNA, Diabetic, and ISIC are regular-sized (224 × 224) datasets.

Algorithm 1 Training process of FedISCA.

Input: K client models $W^c = \{W_1^c, \dots, W_k^c\}$ with corresponding μ_k and σ_k^2 (for simplicity, we omit the index of the BN layers l), a global model W^g, a step S, a learning rate of image synthesis η_s, a learning rate of KD η_d, a momentum α.
$\hat{W}^c \leftarrow W^c, \hat{\mu} \leftarrow \mu, \hat{\sigma}^2 \leftarrow \sigma^2$
while until convergence **do**
  Initialize a batch of random noise $\hat{x}$ and arbitrary labels y.
  $memory \leftarrow [\ ]$
  for $s = 1, \cdots, S$ **do**
    $\hat{x} \leftarrow \hat{x} - \eta_s \nabla L_s(\hat{x}, y; W^c)$
    $memory.append(\hat{x})$
  end for
  for $s = 1, \cdots, S$ **do**
    $\hat{x} \leftarrow memory[S - s]$
    for $k = 1, \cdots, K$ **do**
      $\hat{\mu}_k \leftarrow \alpha\hat{\mu}_k + (1-\alpha)\mu_k(\hat{x}),\ \hat{\sigma}_k^2 \leftarrow \alpha\hat{\sigma}_k^2 + (1-\alpha)\sigma_k^2(\hat{x})$
    end for
  end for
  for $s = 1, \cdots, S$ **do**
    $\hat{x}, \lambda \leftarrow memory[s], 1 - s/S$
    $\hat{z}^c \leftarrow \frac{1}{K}\sum_{k=1}^{K} \hat{W}_k^c(\hat{x}),\ z^c \leftarrow \frac{1}{K}\sum_{k=1}^{K} W_k^c(\hat{x}),\ z^g \leftarrow W^g(\hat{x})$
    $W^g \leftarrow W^g - \eta_d \nabla L_d(\hat{z}^c, z^c, z^g, \lambda)$
  end for
end while
Output: Trained global model W^g.

Algorithm 2 Training process of E-FedISCA.

Input: K client models $W^c = \{W_1^c, \dots, W_k^c\}$ with corresponding μ_k and σ_k^2, a global model W^g, a step S, a learning rate of image synthesis η_s, a learning rate of KD η_d, a momentum α, structure noise $\mathcal{N}'$, a threshold L_{th}.
$\hat{W}^c \leftarrow W^c, \hat{\mu} \leftarrow \mu, \hat{\sigma}^2 \leftarrow \sigma^2$
$memory \leftarrow [\ ]$
Initialize a batch of random noise $\hat{x}$ and arbitrary labels y.
for $s = 1, \cdots, S$ **do**
  $\hat{x} \leftarrow \hat{x} - \eta_s \nabla L_s(\hat{x}, y; W^c)$
  if $L_s < L_{th}$ **then**
    $memory.append(\hat{x})$
  end if
end for
while until convergence **do**
  for $s = 1, \cdots, S$ **do**
    $\hat{x} \leftarrow$ `random.choice`$(memory)$
    $\hat{x}' \leftarrow$ `random.choice`$(\mathcal{N}')$
    $\lambda \leftarrow s/S$
    $\hat{x} \leftarrow mixup(\hat{x}, \hat{x}', \lambda)$
    for $k = 1, \cdots, K$ **do**
      $\hat{\mu}_k \leftarrow \alpha\hat{\mu}_k + (1-\alpha)\mu_k(\hat{x}),\ \hat{\sigma}_k^2 \leftarrow \alpha\hat{\sigma}_k^2 + (1-\alpha)\sigma_k^2(\hat{x})$
    end for
  end for
  for $s = 1, \cdots, S$ **do**
    $\hat{x} \leftarrow$ `random.choice`$(memory)$
    $\hat{x}' \leftarrow$ `random.choice`$(\mathcal{N}')$
    $\lambda \leftarrow 1 - s/S$
    $\hat{x} \leftarrow mixup(\hat{x}, \hat{x}', \lambda)$
    $\hat{z}^c \leftarrow \frac{1}{K}\sum_{k=1}^{K} \hat{W}_k^c(\hat{x}),\ z^c \leftarrow \frac{1}{K}\sum_{k=1}^{K} W_k^c(\hat{x}),\ z^g \leftarrow W^g(\hat{x})$
    $W^g \leftarrow W^g - \eta_d \nabla L_d(\hat{z}^c, z^c, z^g, \lambda)$
  end for
end while
Output: Trained global model W^g.

the *(3) Noise Adaptation* and *(4) Global Model Training* steps until the global model W^g converges. This approach significantly improves computational efficiency compared to the original FedISCA, as it only requires a single image synthesis process for global model training. Additionally, the use of $\mathcal{N}'$ allows for generating a wider range of samples compared to those optimized from $\mathcal{N}(0, 1)$, leading to a more robust global model training. Also, utilizing $\mathcal{N}'$ results in using various samples containing informative clues to learn visual representation, ultimately attaining a high-accuracy model.

4. Experiments

In this section, we provide details regarding the datasets and the experimental scenarios used in our experiments. Additionally, we describe the implementation details and the methods used for comparisons.

4.1. Datasets

We conducted experiments on eight medical image classification datasets, as shown in Fig. 3. The datasets consist of five small-sized (28 × 28) datasets from MedMNIST (Yang et al., 2023a) (i.e., Blood, Derma, Oct, Path, and Tissue) and three regular-sized (224 × 224) datasets, including RSNA (RSNA, 2018), Diabetic (EyePACS, 2015), and ISIC (Codella et al., 2018; Tschandl et al., 2018; Combalia et al., 2019). We refer to the 224 × 224 resolution image as regular-sized since it is commonly used for general image analysis and distinguish it from small-sized due to differences in models, training settings, and resulting outcomes. Here are the details of each dataset:

- **Blood:** A dataset with 17,092 blood cell microscope images for classifying eight different types of cells (e.g., neutrophils, eosinophils). The dataset is split into 13,671 samples for the train and validation sets, and 3421 samples for the test set. We used pre-processed samples from MedMNIST (Yang et al., 2023a), where the central square region of the image was cropped and resized to 28 × 28.
- **Derma:** A dataset with 10,015 dermatoscopic images for classifying seven different diseases. The dataset is split into 8010 samples for the train and validation sets, and 2005 samples for the test set. We used pre-processed samples from MedMNIST (Yang et al., 2023a), where the original images were resized to 28 × 28.
- **Oct:** A dataset with 109,309 optical coherence tomography (OCT) images for classifying four different diseases. The dataset is split into 108,309 samples for the train and validation sets, and 1000 samples for the test set. We used pre-processed samples from MedMNIST (Yang et al., 2023a), where the original images were cropped to the length of the short edge from the center and resized to 28 × 28.
- **Path:** A dataset with 107,180 images for classifying nine types of tissues. The dataset is split into 100,000 samples for the train and validation sets, and 7180 samples for the test set. We used pre-processed samples from MedMNIST (Yang et al., 2023a), where the non-overlapping patches of histology slides were resized to 28 × 28.
- **Tissue:** A dataset with 236,386 kidney cortex microscope images for classifying eight tissue specimens. The dataset is split into 189,106 samples for the train and validation sets, and 47,280

samples for the test set. We used pre-processed samples from MedMNIST (Yang et al., 2023a), where $32 \times 32 \times 7$ images were projected as gray-scale images by taking the maximum pixel value of the last dimension (i.e., 7), and resized to 28×28.

- **RSNA:** A dataset with 27,340 chest X-ray images for binary classification (normal and abnormal) (RSNA, 2018). The dataset is split into 24,348 samples for the train and validation sets, and 2992 samples for the test set. We resized the images to 224×224 for pre-processing.
- **Diabetic:** A dataset with 88,696 retina images for classifying five disease classifications (EyePACS, 2015). The dataset is split into 35,120 samples for the train and validation sets, and 53,576 samples for the test set. We resized the images to 224×224 for pre-processing.
- **ISIC:** A dataset with 23,247 dermoscopy images for classifying eight different melanoma. The dataset is split into 18,597 samples for the train and validation sets, and 4650 samples for the test set. The images were resized to 224×224 for pre-processing. Note that this dataset is based on ISIC2019 (Codella et al., 2018), HAM (Tschandl et al., 2018), and BCN20000 (Combalia et al., 2019), and it provides image acquisition information for each sample.

In the FL scenario, the datasets are partitioned into multiple clients, with each client using a non-overlapping subset of the data for training. The partitioning can be done as follows: Independent and Identically Distributed (IID) clients, where each client has a uniform label distribution, or Dirichlet (Yurochkin et al., 2019) distributed clients, where each client has a different label distribution, causing difficulties in FL training. Adjusting the α parameter in Dirichlet allows for updating the severity of the label distribution, with a lower α resulting in more diverse label distributions among clients. The non-overlapping subsets for each client are split into 90% for the train set and 10% for the validation set. To prevent data leakage, we use the official test data splits provided by the source dataset (Yang et al., 2023a; EyePACS, 2015; Codella et al., 2018; Tschandl et al., 2018; Combalia et al., 2019).

For the ISIC dataset, image acquisition information is available for each sample. To create more realistic FL scenarios, we divide each client's samples using the predefined setting (Ogier du Terrail et al., 2022). We denote this modified dataset as **ISIC′**, where each client has a different image acquisition system. Note that although ISIC′ has no label distribution shift among clients, the difference in imaging acquisition results in style differences between clients, posing a challenge in FL.

4.2. Experimental scenarios

We conducted experiments to evaluate various aspects of our proposed method. The following scenarios were explored:

- **Comparison against previous one-shot FL methods:** We used the Blood, Derma, Oct, Path, Tissue, RSNA, Diabetic, and ISIC datasets with IID clients. Additionally, we conducted experiments on the ISIC′ dataset, which represents an FL scenario with different image acquisition systems between clients. For comparison, we employed three one-shot FL methods: FedAvg (McMahan et al., 2017) with a single communication round, DAFL (Chen et al., 2019), and DENSE (Zhang et al., 2022). We also reported the upper bound accuracy of FedAvg with 100 communications.
- **Ablation studies:** We used the Blood, Derma, Oct, Path, Tissue, RSNA, Diabetic, ISIC, and ISIC′ datasets with IID clients. We conducted ablation studies to evaluate the contributions of different components to the final accuracy. Specifically, we evaluated the accuracy with (a) image synthesis (w/ IS), (b) noise (w/ $\mathcal{N}$), (c) image synthesis and noise (w/ IS&$\mathcal{N}$), (d) noise and noise adaptation (w/ $\mathcal{N}$&Ada), and compared them against (e) our method (w/ IS&$\mathcal{N}$&Ada). Note that noise images are required for noise adaptation; therefore, we excluded "image synthesis and noise adaptation" from the ablation studies. Notably, although we used the same components for the ablation studies, there exist differences between FedISCA and E-FedISCA. Specifically, in FedISCA, the use of (a) image synthesis is identical to DeepInversion (Yin et al., 2020) in the one-shot FL scenario, where we leverage the best synthesized images (samples with the lowest L_s) for training. On the other hand, in E-FedISCA, (a) image synthesis differs from FedISCA as it employs n synthesized samples from a single image synthesis process. Furthermore, the ablation on (b) noise, the difference between FedISCA and E-FedISCA lies in their use of different noise for training. Specifically, FedISCA employs intermediate samples for training, while $\mathcal{N}'$ (Baradad Jurjo et al., 2021) is used for E-FedISCA.
- **Experiments on non-IID data heterogeneity:** We used the Blood, Derma, Oct, Path, and Tissue datasets with Dirichlet distributed clients on $\alpha = 0.6$ and $\alpha = 0.3$.
- **Experiments on model heterogeneity:** We used the Blood, Derma, Oct, Path, and Tissue datasets with IID, Dirichlet ($\alpha = 0.6$), and Dirichlet ($\alpha = 0.3$) distributed clients. For model heterogeneity, different client models, including ResNet18 (He et al., 2016), ResNet34 (He et al., 2016), WRN-16-2 (Zagoruyko and Komodakis, 2016), VGG16 (with BN) (Simonyan and Zisserman, 2014), and VGG8 (with BN) (Simonyan and Zisserman, 2014), were used. For comparison methods, we used DAFL (Chen et al., 2019) and DENSE (Zhang et al., 2022) as the methods are capable of training a global model in scenarios with model heterogeneity.
- **Quantitative analysis on computation speed:** To analyze the training efficiency, we measured the computation speed of image synthesis, noise adaptation, and global model training. To calculate the time required for each training process, we recorded the time before and after each operation within each mini-batch step, reporting the average speed over 100 runs. These measurements were performed using a small-sized dataset, with evaluations conducted on an RTX A5000 GPU. For a fair comparison with FedISCA, the time spent on the Image Synthesis in E-FedISCA was divided by the number of Image Synthesis performed in FedISCA. Since the training was conducted over 100 epochs in the small-sized dataset experiment, the final computation speed was calculated by dividing the total result by 100.
- **Experiments on multi-shot:** We used the Blood, Derma, Oct, Path, Tissue, RSNA, Diabetic, ISIC, and ISIC′ datasets with IID clients. The number of shots was increased from 1 to 30, and we evaluated the highest accuracy of our method and the highest accuracy of the upper bound (i.e., FedAvg with 100 communications). In the multi-shot experiment, we again transmit the trained global model to the client and we retrain each client model using their respective datasets. Afterward, we train the global model using the updated client models and report the accuracy.
- **Experiments on larger clients:** We used the Blood, Oct, Path, and Tissue datasets with IID clients. The number of clients was increased to 20, and we verified whether our method achieves similar accuracy compared to results on smaller(=5) clients.
- **Impact on natural dataset:** We used the Cifar10 dataset with Dirichlet distributed clients on $\alpha = 0.5$, $\alpha = 0.3$, and $\alpha = 0.1$. We compared our method against three one-shot FL methods: FedAvg (McMahan et al., 2017) with a single communication round, DAFL (Chen et al., 2019), and DENSE (Zhang et al., 2022), to evaluate whether our method performs well on natural data.

4.3. Implementation details

We adopted ResNet18 (He et al., 2016) as the base model for our experiments and set the number of clients to five by default. On small-sized datasets, we reported the accuracy of the global model on the test

Table 1
Classification accuracy on eight datasets. The first and second sub-rows show the accuracy of upper bounds and the previous one-shot FL methods. For regular-sized datasets, balanced accuracy was used as the metric. Bold indicates the best accuracy among one-shot FL methods.

	Blood	Derma	Oct	Path	Tissue	RSNA	Diabetic	ISIC	ISIC′
FedAvg (McMahan et al., 2017)	93.51	74.61	75.60	84.54	63.64	88.16	49.04	62.88	57.15
FedAvg(1)	13.74	66.88	25.00	5.86	32.07	78.65	35.60	38.05	18.08
DAFL (Chen et al., 2019)	7.13	66.43	25.00	7.63	11.55	50.55	22.63	14.51	18.37
DENSE (Zhang et al., 2022)	39.37	66.93	33.80	21.89	21.35	55.06	23.51	13.69	16.46
FedISCA	**87.99**	70.12	70.20	**84.18**	**61.90**	85.34	40.08	48.39	22.47
E-FedISCA	87.31	**71.47**	**71.30**	79.48	57.96	**85.46**	**41.32**	**51.17**	**27.01**

Table 2
Classification accuracy of ablation studies on eight datasets. w/ indicates "with" and IS indicates image synthesis. $\mathcal{N}$ indicates noise ($\mathcal{N}(0,1)$ for FedISCA and $\mathcal{N}'$ for E-FedISCA) and Ada indicates noise adaptation.

		IS	$\mathcal{N}$	Ada	Blood	Derma	Oct	Path	Tissue	RSNA	Diabetic	ISIC	ISIC′
(a)	FedISCA w/ (Yin et al., 2020)	✓			87.02	68.73	60.20	77.90	57.86	50.61	28.30	25.61	14.80
(b)	FedISCA w/		✓		7.13	11.12	35.80	4.72	7.13	50.00	20.02	12.50	12.52
(c)	FedISCA w/	✓	✓		81.61	68.33	**70.30**	82.08	59.34	81.56	**40.91**	47.21	21.72
(d)	FedISCA w/		✓	✓	9.09	66.88	25.20	24.69	23.70	50.00	20.07	12.50	11.29
(e)	FedISCA	✓	✓	✓	**87.99**	**70.12**	70.20	**84.18**	**61.90**	**85.34**	40.08	**48.39**	**22.47**
(a)	E-FedISCA w/	✓			85.56	68.28	57.70	77.45	55.90	62.88	29.74	32.83	18.73
(b)	E-FedISCA w/		✓		78.40	69.63	63.60	59.67	46.03	50.00	39.84	35.88	26.74
(c)	E-FedISCA w/	✓	✓		83.66	70.72	68.20	77.26	55.54	85.37	39.38	50.66	26.72
(d)	E-FedISCA w/		✓	✓	26.22	66.88	25.00	17.17	32.07	79.87	22.56	16.18	13.12
(e)	E-FedISCA	✓	✓	✓	**87.31**	**71.47**	**71.30**	**79.48**	**57.96**	**85.46**	**41.32**	**51.17**	**27.01**

data, while on regular-sized datasets, we used the balanced accuracy of the global model on the test data, following the approach used in Ogier du Terrail et al. (2022). The client models were trained for 100 epochs using the SGD optimizer with an initial learning rate of 1e−3 and a batch size of 128. For FedISCA image synthesis, we used the Adam optimizer with a learning rate of 5e−2 for 100 epochs. We used 500 and 1000 synthesis steps (S) for small- and regular-sized datasets, respectively, with a batch size of 256 for small-sized datasets and 50 for regular-sized datasets. Following Yin et al. (2020), the hyperparameters $\lambda_{TV} = 0.000025$ and $\lambda_{BN} = 10$ were used for image synthesis, with a KD temperature of $T = 20$ and momentum $\alpha = 0.9$. For E-FedISCA image synthesis, we followed the settings of FedISCA but with some differences. We performed image synthesis before global model training with 1000 and 2000 synthesis steps (S) for small- and regular-sized datasets, respectively. The batch size used was 256 for small-sized datasets and 96 for regular-sized datasets. During image synthesis, we stored n intermediate samples in $memory$ that showed lower loss (L_s) than a threshold value $L_{th} = 50$. We empirically set a threshold of 50 that converged and yielded high-quality image synthesis. However, since we store the latest synthesized sample in case we fail to collect $memory$ with a size n, the lower threshold value by itself does not significantly affect the final results (for simplicity, this part is omitted from Algorithm 2). To generate pseudo intermediate samples, we employed StyleGAN-Oriented structure noise (Baradad Jurjo et al., 2021), which obtained the best accuracy in the ImageNet-100 experiment. However, since structure noise is proposed for regular-sized image training thus we resized the images to 28×28 as a pre-processing for small-sized experiments.

For fair comparison, we followed the original implementations of each comparison method and matched all training/parameter settings. All one-shot FL approaches used client models with same trained parameters for global model aggregation. For DAFL, an ensemble of client models was used as the teacher model with KD used for global model training, following the approach in Zhang et al. (2022), Lin et al. (2020) and Zhang et al. (2020). For client model training on regular-sized datasets, we initialize the model to an ImageNet pre-trained model.

5. Results

5.1. Comparison against previous one-shot FL methods

Table 1 shows the accuracy on five small- and four regular-sized datasets. FedISCA achieves improved accuracy than the compared methods on both small- and regular-sized datasets. However, DAFL and DENSE perform poorly on medical data, showing significant accuracy gaps compared to the upper bound, i.e., FedAvg. Additionally, though FedAvg reports higher accuracy for multiple communication rounds, it shows significantly lower accuracy for single communication (FedAvg(1)). The accuracy decrease varies across datasets; for example, on the Path dataset, accuracy decreases significantly, whereas the Derma dataset showed a minor decrease in accuracy. To explain the observed differences, we analyzed the variance in client model parameters. The comparison shows that models with higher variance (e.g., Path IID = 36.10) tend to have lower accuracy, while models with lower variance (e.g., Derma IID = 0.01) achieve higher accuracy. Furthermore, in regular-sized dataset experiments where ImageNet pre-trained parameters are used for initialization, FedAvg(1) shows a relatively smaller accuracy drop due to its potential for facilitating knowledge sharing among clients. These findings suggest a correlation between client model variance and the accuracy of FedAvg(1).

In Fig. 4, we show the synthesized images of E-FedISCA, DAFL (Chen et al., 2019), and DENSE (Zhang et al., 2022) for eight datasets. FedISCA generates more realistic images compared to the competitors. Note that both DAFL and DENSE aim to generate diverse images distributed near the decision boundary, resulting in the synthesis of images that are visually distinct from real data. Qualitatively, the synthesized images in their papers are similar to those shown in Fig. 4, indicating that we synthesized the images as intended by those methods. Although this strategy effectively generates informative samples and improves accuracy on natural images, our experiments show that it even underperforms FedAvg(1), suggesting the assumption does not hold in medical images.

Furthermore, E-FedISCA demonstrates competitive accuracy compared to FedICSA on Blood, Path, and Tissue datasets and outperforms FedISCA on Derma, Oct, RSNA, Diabetic, ISIC, and ISIC′ datasets. This highlights the viability of E-FedISCA in one-shot FL with reduced computational overhead. Notably, E-FedISCA achieves the best accuracy on all regular-sized datasets, indicating that noise input capturing certain structural properties of real data (as verified on ImageNet) encourages robust global model training, leading to high accuracy. However, E-FedISCA obtains lower accuracy compared to FedISCA on small-sized datasets. This could be attributed to the loss of structure noise (Baradad Jurjo et al., 2021) when resizing the image to 28×28 during pre-processing.

Table 3
Classification accuracy on five datasets with different heterogeneity levels.

	Dirichlet ($\alpha = 0.6$)					Dirichlet ($\alpha = 0.3$)				
	Blood	Derma	Oct	Path	Tissue	Blood	Derma	Oct	Path	Tissue
FedAvg (McMahan et al., 2017)	93.60	72.72	76.50	81.48	55.61	87.49	69.88	73.50	77.52	53.26
FedAvg(1)	18.24	66.88	25.00	5.86	32.07	16.92	10.97	25.00	5.86	32.07
DAFL (Chen et al., 2019)	7.13	66.88	34.40	14.97	39.15	7.13	13.62	25.00	18.64	45.00
DENSE (Zhang et al., 2022)	34.52	67.78	39.40	30.31	9.47	30.78	12.77	25.80	19.87	9.33
FedISCA	82.90	**69.83**	**68.60**	**82.92**	**53.04**	**46.59**	15.91	60.50	**79.25**	**51.00**
E-FedISCA	**83.10**	69.68	66.20	78.51	52.87	45.86	**16.96**	**61.60**	73.40	48.45
FedISCA w/ IS (Yin et al., 2020)	80.62	69.58	60.30	75.54	49.06	45.69	13.87	49.20	70.53	46.73

Table 4
Classification accuracy on five datasets with model heterogeneity.

	IID					Dirichlet ($\alpha = 0.6$)					Dirichlet ($\alpha = 0.3$)				
	Blood	Derma	Oct	Path	Tissue	Blood	Derma	Oct	Path	Tissue	Blood	Derma	Oct	Path	Tissue
DAFL (Chen et al., 2019)	7.13	65.69	25.00	15.72	35.66	7.13	67.23	37.10	28.15	39.54	7.13	13.47	45.30	29.68	19.54
DENSE (Zhang et al., 2022)	46.86	66.88	44.00	33.08	38.28	23.47	67.93	40.70	28.68	36.70	34.67	13.42	44.00	39.37	38.37
FedISCA	87.96	71.17	70.00	**83.02**	**61.74**	**73.43**	**69.23**	**64.80**	**82.73**	51.95	44.20	**16.61**	**62.00**	**72.26**	**43.80**
E-FedISCA	**88.31**	**71.72**	**71.00**	80.04	58.96	72.76	**69.23**	64.60	80.84	**52.04**	**47.18**	16.01	58.90	72.17	43.19
FedISCA w/ IS (Yin et al., 2020)	87.55	69.93	51.20	74.05	57.90	68.78	68.93	61.00	72.79	46.91	43.85	15.61	51.50	64.65	39.89

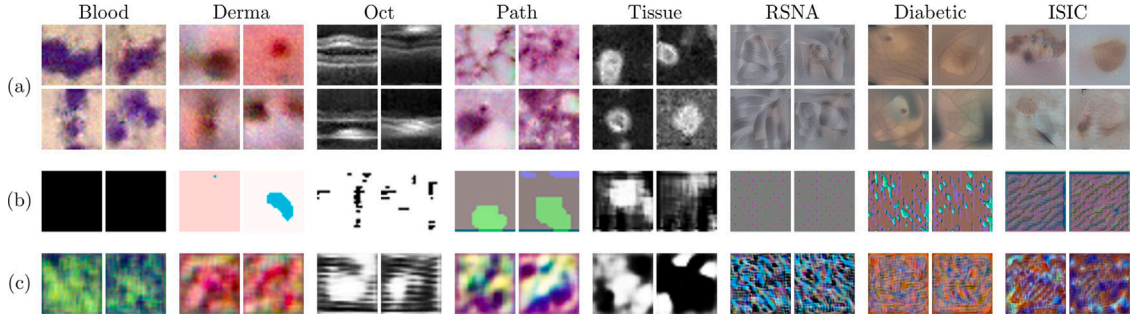

Fig. 4. The synthesized images of (a) E-FedISCA, (b) DAFL (Chen et al., 2019), and (c) DENSE (Zhang et al., 2022) on eight datasets. We do not show the synthesized image of FedISCA as it is identical to E-FedISCA. Overall, E-FedISCA synthesizes more realistic images.

5.2. Ablation studies

We present the ablation results for FedISCA in Table 2. In the medical field, generating realistic images is crucial for one-shot FL, as the accuracy of w/ $\mathcal{N}$ and w/ $\mathcal{N}$&Ada is significantly lower than w/ IS and w/ IS&$\mathcal{N}$. This validates the necessity of image synthesis during one-shot FL. Additionally, using Ada along with $\mathcal{N}$ (i.e., w/ $\mathcal{N}$&Ada) yields lower accuracy, as visual gaps between $\mathcal{N}(0,1)$ and D cannot be addressed via noise-adapted client models, suggesting that image synthesis is essential for successful KD. Except for Blood and Derma, w/ IS&$\mathcal{N}$ achieves higher accuracy compared to w/ IS. This indicates that intermediate samples encourage robust global model training and lead to higher accuracy than training with the best images alone. On the contrary, the accuracy of FedISCA (w/ IS&$\mathcal{N}$&Ada) suggests that noise-adapted client models compensate for the negative effects of random noise, resulting in the best accuracy across all datasets. Overall, the experimental results support the idea that these components play an essential role in medical one-shot FL.

For E-FedISCA, the ablation results in Table 2 similarly highlight the importance of generating realistic images. The accuracy of w/ $\mathcal{N}$ and $\mathcal{N}$&Ada is lower than that of w/ IS and IS&$\mathcal{N}$ on small-sized datasets and the RSNA dataset. However, since w/ IS in E-FedISCA only uses $n(= 500)$ samples from a single image synthesis, this approach may lead to overfitting, resulting in lower accuracy on Diabetic, ISIC, and ISIC′ datasets (see w/ $\mathcal{N}$). Notably, w/ $\mathcal{N}$ in E-FedISCA performs better than w/ $\mathcal{N}$ in FedISCA for all cases. This indicates that the structure noise $\mathcal{N}'$ is more beneficial than random noise $\mathcal{N}$ for one-shot FL. Moreover, on regular-sized datasets including Diabetic, ISIC, and ISIC′, w/ $\mathcal{N}$ outperforms w/ IS, suggesting that $\mathcal{N}'$ provides more information than using n samples from a single image synthesis.

Regarding mixup with structure noise and synthesized images, using both results in higher (or similar) accuracy compared to using them separately, except on the Blood dataset. Particularly, the failure of w/ $\mathcal{N}'$ on RSNA is addressed when synthesized images are used, indicating that an image possessing the properties of D is crucial for successful one-shot FL. In contrast, the ablations in w/ $\mathcal{N}'$&Ada on RSNA achieve higher accuracy compared to w/ $\mathcal{N}'$, while other datasets show a degraded accuracy compared to w/ $\mathcal{N}'$ due to the possibility of label-related knowledge of client models being forgotten. This observation suggests that noise-adapted clients not only mitigate the negative impact of noise but also resolve domain gaps between D and $\mathcal{N}'$, resulting in higher accuracy. Finally, the accuracy of E-FedISCA (w/ IS&$\mathcal{N}$&Ada) demonstrates that the proposed three components are vital for efficient and robust global model training. Even with a single image synthesis and using it with structure noise for KD, E-FedISCA achieves higher accuracy than FedISCA on regular-sized datasets. Overall, the ablation studies support that our design choices are suitable for efficient one-shot FL and provide insights into the role of each component in contributing to success.

5.3. Experiments on non-IID data heterogeneity

Table 3 presents the accuracy of five datasets with different levels of data heterogeneity. FedISCA and E-FedISCA outperform all previous one-shot FL methods across all datasets, regardless of the level of heterogeneity. Notably, due to the challenges presented by non-IID

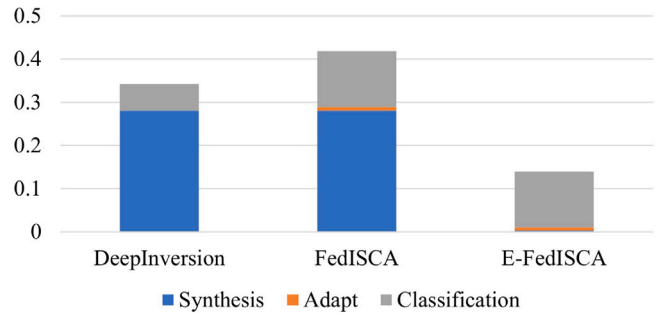

Fig. 5. Computation speed of DeepInversion (FedISCA w/ Yin et al., 2020), FedISCA, and E-FedISCA. E-FedISCA requires the lowest computational resources than the competitors.

scenarios (e.g., Dirichlet with $\alpha = 0.3$), the upper bound reports reduced accuracy compared to the IID setting. Similar to IID scenarios, DAFL, DENSE, and FedAvg with single communication (FedAvg(1)) yield poor results, indicating that their assumptions also fail in non-IID data heterogeneity scenarios. However, when realistic synthesized images are used for KD (see FedISCA w/ IS Yin et al., 2020), the accuracy improves significantly, resulting in small differences in accuracy between the upper bound. The gaps in accuracy further decrease when employing our proposed modules, resulting in the best accuracy across all non-IID scenarios. Moreover, E-FedISCA shows competitive or even higher accuracy in non-IID data heterogeneity settings. These experimental results suggest that the effectiveness of our method has been verified in non-IID data heterogeneity problems.

5.4. Experiments on model heterogeneity

Table 4 presents the accuracy of a global model trained on client models with model heterogeneity. FedISCA achieves the best accuracy among all previous one-shot FL methods, demonstrating the effectiveness of our method in one-shot FL with diverse types of model architectures. On the other hand, previous approaches consistently yield low accuracy across all comparisons, highlighting their inadequacy in handling the challenges posed by model heterogeneity. Moreover, the experimental results suggest that our approach performs well in challenging scenarios where both model and data heterogeneity are exhibited. Additionally, E-FedISCA obtains similar or even outperforms the accuracy of FedISCA on model heterogeneity, demonstrating the validity and robustness of our approach to diverse heterogeneity scenarios.

5.5. Quantitative analysis on computation speed

In Fig. 5, we present the computation speed for each component of our method. E-FedISCA demonstrates the fastest speed compared to FedISCA and DeepInversion (Yin et al., 2020). Clearly, this is due to the reuse of synthesized images, which significantly reduces the overall computational overhead. Additionally, though FedISCA and E-FedISCA necessitate additional computation resources for noise-adapted client models than DeepInversion, E-FedISCA still remains the most efficient approach. In summary, E-FedISCA enables training with minimal computation resources, enhancing the applicability of one-shot FL.

5.6. Experiments on multi-shot

Fig. 6 shows the accuracy of our method on eight datasets in the multi-shot FL scenario. Overall, our method exhibits improved accuracy across all datasets as the number of shots increases. Notably, in 5-shot scenarios, our approach outperforms the upper bound (FedAvg (McMahan et al., 2017), as represented by the blue dashed line), except on the Diabetic and ISIC′ datasets. Furthermore, increasing the number of shots to 12 demonstrates that our method obtains the upper bound on the ISIC′ dataset. This suggests that increases in the number of shots result in achieving accuracy comparable to the upper bound for challenging datasets. The orange dashed line represents the highest accuracy achieved by E-FedISCA when the number of shots is increased to 30. Our method converges faster, as the increased number of shots confirms that the highest accuracy is achieved in the early training stages, highlighting its superior training efficiency. However, though we increase the number of shots to 30 on Diabetic dataset, our method achieves competitive accuracy with a marginal gap compared to the accuracy of FedAvg. The development of a more efficient multi-shot FL method is reserved for future research.

Table 5
Classification accuracy of 20 clients on five datasets with IID clients.

	Blood	Derma	Oct	Path	Tissue
FedISCA (5 clients)	87.99	70.12	70.20	84.18	61.90
FedISCA (20 clients)	78.31	69.23	67.20	84.57	56.93
E-FedISCA (5 clients)	87.31	71.47	71.30	79.48	57.96
E-FedISCA (20 clients)	79.07	69.18	64.50	81.45	54.62
E-FedISCA (20 clients) 2-shot	86.23	69.48	69.80	83.19	58.57

Table 6
Classification accuracy on Cifar10 datasets with different Dirichlet α.

	$\alpha = 0.5$	$\alpha = 0.3$	$\alpha = 0.1$
FedAvg(1)	43.67	27.72	23.93
DAFL (Chen et al., 2019)	58.59	53.89	47.34
DENSE (Zhang et al., 2022)	62.19	59.76	50.26
FedISCA	**71.39**	**65.30**	**54.55**
E-FedISCA	67.94	60.94	50.90

5.7. Experiments on larger clients

Table 5 presents the accuracy on five datasets with 20 clients. Due to the increased number of clients, both FedISCA and E-FedISCA achieved lower accuracy compared to the experiments with five clients. However, the challenges posed by a larger number of clients can be alleviated in the 2-shot scenario. The results of the 2-shot setting demonstrate competitive accuracy compared to the one-shot settings with five clients. These experimental results confirm that our method is capable of training a robust global model regardless of the number of clients and maintains its effectiveness in larger client scenarios.

5.8. Impact on natural dataset

Table 6 shows the accuracy of the Cifar10 dataset with different levels of heterogeneity. Compared to the state-of-the-art one-shot FL method (DENSE Zhang et al., 2022) on natural images, FedISCA exhibits a significant improvement in accuracy, achieving the highest accuracy among all settings. Furthermore, although E-FedISCA shows a slightly decreased accuracy compared to FedISCA, it still outperforms DENSE and ranks second in terms of accuracy. Notably, the accuracy gap between our method and previous methods is more significant in the medical dataset than in the natural dataset (see Table 1). As stated in Konz et al. (2022), medical data is crucial for capturing a few, difficult-to-learn features, indicating that synthesized images should meet these criteria for high accuracy. However, as shown in Fig. 4, previous methods generated images that are visually distinct from real data (e.g., ambiguous or unrealistic), resulting in the failure of one-shot FL on medical data. The experimental results suggest that our method not only performs exceptionally well in the medical domain but also exhibits strong accuracy in the natural image domain, highlighting the effectiveness of our approach across diverse domains.

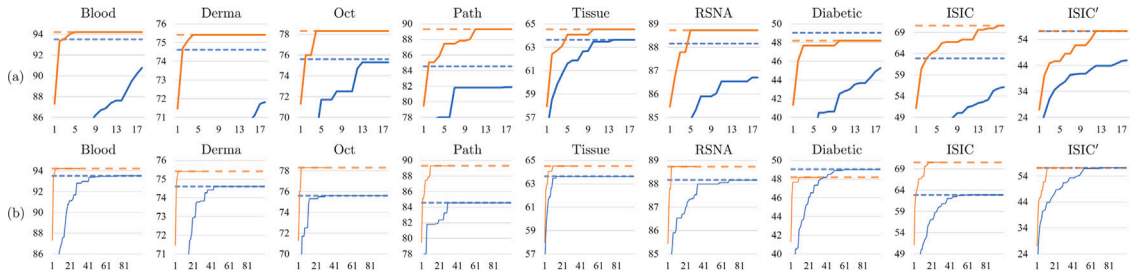

Fig. 6. Classification accuracy in multi-shot FL scenarios on eight datasets. (a) shows the accuracy for the zoomed-in early phase, while (b) shows the accuracy for the entire phase across 100 shots. Orange solid line indicates the accuracy of E-FedISCA and orange dashed line indicates the highest accuracy of E-FedISCA across 30 shots. Blue solid line indicates the accuracy of FedAvg and blue dashed line indicates the highest accuracy of FedAvg across 100 shots. The X-axis represents the number of shots in FL, while the Y-axis represents the highest accuracy achieved with that number of shots.

5.9. Limitations

Through experiments, we confirmed that our method achieves higher accuracy compared to the baseline methods. Furthermore, as presented in Section 5.6, E-FedISCA surpasses the upper bound of FedAvg with fewer communication rounds, overcoming its limitations and demonstrating its superiority. However, some challenges remain unresolved, and further research is required to address them in the future. In Section 5.3, although our method demonstrates higher accuracy compared to the baselines, its accuracy decreases when heterogeneity becomes more severe, i.e., from Dirichlet ($\alpha = 0.6$) to Dirichlet ($\alpha = 0.3$). Since the accuracy degradation in FedAvg is relatively small under conditions of severe heterogeneity, developing an algorithm that robustly trains models under severe heterogeneity will enhance the applicability of one-shot FL. As discussed in Sections 5.6 and 5.7, the development of an efficient multi-shot FL method and training with a larger number of clients remain for future research. Since accuracy degradation arises from parameter drift between client models, leveraging foundation models with fewer client parameter updates (e.g., by using adapters) and applying novel fine-tuning approaches that consider future global model aggregation is expected to address these challenges.

Lastly, as state in Bujotzek et al. (2024), once the FL infrastructure is established, communication costs will no longer pose a bottleneck in FL, which limits the contribution of the proposed method. However, scenarios necessitating a one-shot FL are often encountered. First, setting up a secure FL infrastructure may not be feasible for small- and mid-scale hospitals due to limited budgets. In such cases, employing one-shot FL can reduce infrastructure costs and provide a viable solution for allowing more hospitals to participate in FL. Secondly, even large-scale hospitals often face challenges in establishing an FL infrastructure due to stringent cybersecurity requirements (Eichelberg et al., 2020). In environments where privacy is critical and any form of data sharing over external networks is highly restricted, such as when data is stored on an isolated local server, significant investment is required to establish the hardware infrastructure necessary for secure external network connections. Reducing communication frequency can effectively support model training in these constrained network environments. Additionally, reducing model transmission and improving efficiency (Park et al., 2021; Chen et al., 2024) are critical challenges in FL, as noted by Kairouz et al. (2021). Our method addresses this by achieving converged accuracy with a low number of communication rounds, surpassing the upper bound, thus demonstrating its efficiency. Consequently, one-shot FL still serves as a viable solution in such cases.

6. Conclusion

We present a novel one-shot FL framework that uses synthetic- and noise-images with KD. First, we demonstrate that random noise significantly reduces the risk of overfitting, resulting in robust global model training. Next, by employing mixup and structure noise, we generate diverse intermediate samples, effectively reducing computational resources and enhancing training efficiency. Additionally, noise-adapted client models improve the KD signal for noise, resulting in higher accuracy in the trained models. Finally, through extensive experiments on eight medical datasets, we validate that our method outperforms state-of-the-art one-shot FL methods, demonstrating its effectiveness and superiority.

CRediT authorship contribution statement

Myeongkyun Kang: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Philip Chikontwe:** Writing – review & editing, Writing – original draft, Validation, Methodology, Conceptualization. **Soopil Kim:** Writing – review & editing, Writing – original draft, Methodology, Conceptualization. **Kyong Hwan Jin:** Writing – review & editing, Validation, Supervision, Conceptualization. **Ehsan Adeli:** Writing – review & editing, Validation, Supervision, Conceptualization. **Kilian M. Pohl:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Funding acquisition, Data curation, Conceptualization. **Sang Hyun Park:** Writing – review & editing, Writing – original draft, Validation, Supervision, Resources, Project administration, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported by the DGIST R&D program of the Ministry of Science and ICT of KOREA (22-KUJoint-02) and the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00516124) and the Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. RS-2024-00439264, No. RS-2025-02219277).

Data availability

We use publicly available data. Code is available at https://github.com/myeongkyunkang/E-FedISCA.

References

Baradad Jurjo, M., Wulff, J., Wang, T., Isola, P., Torralba, A., 2021. Learning to see by looking at noise. Adv. Neural Inf. Process. Syst. 34, 2556–2569.

Bujotzek, M.R., Akünal, Ü., Denner, S., Neher, P., Zenk, M., Frodl, E., Jaiswal, A., Kim, M., Krekiehn, N.R., Nickel, M., et al., 2024. Real-world federated learning in radiology: Hurdles to overcome and benefits to gain. J. Am. Med. Inform. Assoc. ocae259.

Chen, M., Jiang, M., Zhang, X., QI, D., Wang, Z., Li, X., 2024. Local superior soups: A catalyst for model merging in cross-silo federated learning. Adv. Neural Inf. Process. Syst. 37, 20858–20886.

Chen, H., Wang, Y., Xu, C., Yang, Z., Liu, C., Shi, B., Xu, C., Xu, C., Tian, Q., 2019. Data-free learning of student networks. In: International Conference on Computer Vision. pp. 3514–3522.

Chikontwe, P., Luna, M., Kang, M., Hong, K.S., Ahn, J.H., Park, S.H., 2021. Dual attention multiple instance learning with unsupervised complementary loss for COVID-19 screening. Med. Image Anal. 72, 102105.

Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al., 2018. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In: 2018 IEEE 15th International Symposium on Biomedical Imaging. ISBI 2018, IEEE, pp. 168–172.

Combalia, M., Codella, N.C., Rotemberg, V., Helba, B., Vilaplana, V., Reiter, O., Carrera, C., Barreiro, A., Halpern, A.C., Puig, S., et al., 2019. Bcn20000: Dermoscopic lesions in the wild. arXiv preprint arXiv:1908.02288.

Dennis, D.K., Li, T., Smith, V., 2021. Heterogeneity for the win: One-shot federated clustering. In: International Conference on Machine Learning. PMLR, pp. 2611–2620.

Eichelberg, M., Kleber, K., Kämmerer, M., 2020. Cybersecurity in PACS and medical imaging: An overview. J. Digit. Imaging 33 (6), 1527–1542.

EyePACS, 2015. Diabetic retinopathy detection.

Guha, N., Talwalkar, A., Smith, V., 2019. One-shot federated learning. arXiv preprint arXiv:1902.11175.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: Computer Vision and Pattern Recognition. pp. 770–778.

Hinton, G., Vinyals, O., Dean, J., 2015. Distilling the knowledge in a neural network. arXiv preprint arXiv:1503.02531.

Kairouz, P., McMahan, H.B., Avent, B., Bellet, A., Bennis, M., Bhagoji, A.N., Bonawitz, K., Charles, Z., Cormode, G., Cummings, R., et al., 2021. Advances and open problems in federated learning. Found. Trends® Mach. Learn. 14 (1–2), 1–210.

Kang, M., Chikontwe, P., Kim, S., Jin, K.H., Adeli, E., Pohl, K.M., Park, S.H., 2023a. One-shot federated learning on medical data using knowledge distillation with image synthesis and client model adaptation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer.

Kang, M., Chikontwe, P., Won, D., Luna, M., Park, S.H., 2023b. Structure-preserving image translation for multi-source medical image domain adaptation. Pattern Recognit. 144, 109840.

Kang, M., Kim, S., Jin, K.H., Adeli, E., Pohl, K.M., Park, S.H., 2024. FedNN: Federated learning on concept drift data using weight and adaptive group normalizations. Pattern Recognit. 149, 110230.

Kang, M., Won, D., Luna, M., Chikontwe, P., Hong, K.S., Ahn, J.H., Park, S.H., 2023c. Content preserving image translation with texture co-occurrence and spatial self-similarity for texture debiasing and domain adaptation. Neural Netw. 166, 722–737.

Kataoka, H., Okayasu, K., Matsumoto, A., Yamagata, E., Yamada, R., Inoue, N., Nakamura, A., Satoh, Y., 2020. Pre-training without natural images. In: Proceedings of the Asian Conference on Computer Vision.

Kim, S., Park, H., Kang, M., Jin, K.H., Adeli, E., Pohl, K.M., Park, S.H., 2024. Federated learning with knowledge distillation for multi-organ segmentation with partially labeled datasets. Med. Image Anal. 95, 103156.

Konz, N., Gu, H., Dong, H., Mazurowski, M., 2022. The intrinsic manifolds of radiological images and their role in deep learning. In: Medical Image Computing and Computer Assisted Intervention. Springer, pp. 684–694.

Konz, N., Mazurowski, M.A., 2024. The effect of intrinsic dataset properties on generalization: Unraveling learning differences between natural and medical images. In: The Twelfth International Conference on Learning Representations.

Li, X., Gu, Y., Dvornek, N., Staib, L.H., Ventola, P., Duncan, J.S., 2020b. Multi-site fMRI analysis using privacy-preserving federated learning and domain adaptation: ABIDE results. Med. Image Anal. 65, 101765.

Li, Q., He, B., Song, D., 2020a. Practical one-shot federated learning for cross-silo setting. In: International Joint Conference on Artificial Intelligence.

Li, Y., Wang, N., Shi, J., Liu, J., Hou, X., 2018. Adaptive batch normalization for practical domain adaptation. Pattern Recognit. 80, 109–117.

Lin, T., Kong, L., Stich, S.U., Jaggi, M., 2020. Ensemble distillation for robust model fusion in federated learning. Adv. Neural Inf. Process. Syst. 33, 2351–2363.

Lu, M.Y., Chen, R.J., Kong, D., Lipkova, J., Singh, R., Williamson, D.F., Chen, T.Y., Mahmood, F., 2022. Federated learning for computational pathology on gigapixel whole slide images. Med. Image Anal. 76, 102298.

Mahendran, A., Vedaldi, A., 2015. Understanding deep image representations by inverting them. In: Computer Vision and Pattern Recognition. pp. 5188–5196.

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. PMLR, pp. 1273–1282.

Micaelli, P., Storkey, A.J., 2019. Zero-shot knowledge transfer via adversarial belief matching. Adv. Neural Inf. Process. Syst. 32.

Mordvintsev, A., Olah, C., Tyka, M., 2015. Inceptionism: Going deeper into neural networks. URL https://research.googleblog.com/2015/06/inceptionism-going-deeper-into-neural.html.

Morra, L., Piano, L., Lamberti, F., Tommasi, T., 2021. Bridging the gap between natural and medical images through deep colorization. In: 2020 25th International Conference on Pattern Recognition. ICPR, IEEE, pp. 835–842.

Ogier du Terrail, J., Ayed, S.-S., Cyffers, E., Grimberg, F., He, C., Loeb, R., Mangold, P., Marchand, T., Marfoq, O., Mushtaq, E., Muzellec, B., Philippenko, C., Silva, S., Teleńczuk, M., Albarqouni, S., Avestimehr, S., Bellet, A., Dieuleveut, A., Jaggi, M., Karimireddy, S.P., Lorenzi, M., Neglia, G., Tommasi, M., Andreux, M., 2022. FLamby: Datasets and benchmarks for cross-silo federated learning in realistic healthcare settings. In: Advances in Neural Information Processing Systems.

Papernot, N., Abadi, M., Erlingsson, U., Goodfellow, I., Talwar, K., 2017. Semi-supervised knowledge transfer for deep learning from private training data. Int. Conf. Learn. Represent..

Park, Y., Han, D.-J., Kim, D.-Y., Seo, J., Moon, J., 2021. Few-round learning for federated learning. Adv. Neural Inf. Process. Syst. 34, 28612–28622.

Pope, P., Zhu, C., Abdelkader, A., Goldblum, M., Goldstein, T., 2021. The intrinsic dimension of images and its impact on learning. Int. Conf. Learn. Represent..

Raikwar, P., Mishra, D., 2022. Discovering and overcoming limitations of noise-engineered data-free knowledge distillation. In: Advances in Neural Information Processing Systems.

RSNA, R.S.o.N.A., 2018. RSNA pneumonia detection challenge.

Simonyan, K., Zisserman, A., 2014. Very deep convolutional networks for large-scale image recognition. Int. Conf. Learn. Represent..

Tschandl, P., Rosendahl, C., Kittler, H., 2018. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data 5 (1), 1–9.

Wang, Y., Wang, Y., Cai, J., Lee, T.K., Miao, C., Wang, Z.J., 2023. Ssd-kd: A self-supervised diverse knowledge distillation method for lightweight skin lesion classification using dermoscopic images. Med. Image Anal. 84, 102693.

Xie, H., Liu, Y., Lei, H., Song, T., Yue, G., Du, Y., Wang, T., Zhang, G., Lei, B., 2023. Adversarial learning-based multi-level dense-transmission knowledge distillation for AP-ROP detection. Med. Image Anal. 84, 102725.

Yang, Y., Guo, X., Ye, C., Xiang, Y., Ma, T., 2023b. CReg-KD: Model refinement via confidence regularized knowledge distillation for brain imaging. Med. Image Anal. 89, 102916.

Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B., 2023a. MedMNIST v2: A large-scale lightweight benchmark for 2D and 3D biomedical image classification. Sci. Data 10 (1), 41.

Yang, D., Xu, Z., Li, W., Myronenko, A., Roth, H.R., Harmon, S., Xu, S., Turkbey, B., Turkbey, E., Wang, X., et al., 2021. Federated semi-supervised learning for COVID region segmentation in chest CT using multi-national data from China, Italy, Japan. Med. Image Anal. 70, 101992.

Yin, H., Molchanov, P., Alvarez, J.M., Li, Z., Mallya, A., Hoiem, D., Jha, N.K., Kautz, J., 2020. Dreaming to distill: Data-free knowledge transfer via deepinversion. In: Computer Vision and Pattern Recognition. pp. 8715–8724.

Yurochkin, M., Agarwal, M., Ghosh, S., Greenewald, K., Hoang, N., Khazaeni, Y., 2019. Bayesian nonparametric federated learning of neural networks. In: International Conference on Machine Learning. PMLR, pp. 7252–7261.

Zagoruyko, S., Komodakis, N., 2016. Wide residual networks. In: British Machine Vision Conference. BMVC.

Zhang, J., Chen, C., Li, B., Lyu, L., Wu, S., Ding, S., Shen, C., Wu, C., 2022. DENSE: Data-free one-shot federated learning. In: Advances in Neural Information Processing Systems.

Zhang, H., Cisse, M., Dauphin, Y.N., Lopez-Paz, D., 2018. Mixup: Beyond empirical risk minimization. In: International Conference on Learning Representations.

Zhang, S., Liu, M., Yan, J., 2020. The diversified ensemble neural network. Adv. Neural Inf. Process. Syst. 33, 16001–16011.

Zhou, Y., Pu, G., Ma, X., Li, X., Wu, D., 2020. Distilled one-shot federated learning. arXiv preprint arXiv:2009.07999.