

PAPER • OPEN ACCESS

Exploiting data diversity in multi-domain federated learning

To cite this article: Hussain Ahmad Madni *et al* 2024 *Mach. Learn.: Sci. Technol.* **5** 025041

View the [article online](#) for updates and enhancements.

You may also like

- [Federated learning for intelligent fault diagnosis based on similarity collaboration](#)
Yonghong Zhang, Xingan Xue, Xiaoping Zhao et al.
- [Open-set federated adversarial domain adaptation based cross-domain fault diagnosis](#)
Shu Xu, Jian Ma and Dengwei Song
- [Privacy-preserving quantum federated learning via gradient hiding](#)
Changhao Li, Niraj Kumar, Zhixin Song et al.



PAPER


Exploiting data diversity in multi-domain federated learning

OPEN ACCESS

RECEIVED
19 November 2023REVISED
30 March 2024ACCEPTED FOR PUBLICATION
3 May 2024PUBLISHED
16 May 2024

Original Content from
this work may be used
under the terms of the
[Creative Commons
Attribution 4.0 licence](#).

Any further distribution
of this work must
maintain attribution to
the author(s) and the title
of the work, journal
citation and DOI.

Hussain Ahmad Madni^{1,*} , Rao Muhammad Umer² and Gian Luca Foresti¹¹ Department of Mathematics, Computer Science and Physics, University of Udine, Via delle Scienze 206, Udine 33100, Italy² Institute of AI for Health, Helmholtz Zentrum Munchen—German Research Center for Environmental Health, Neuherberg 85764, Germany

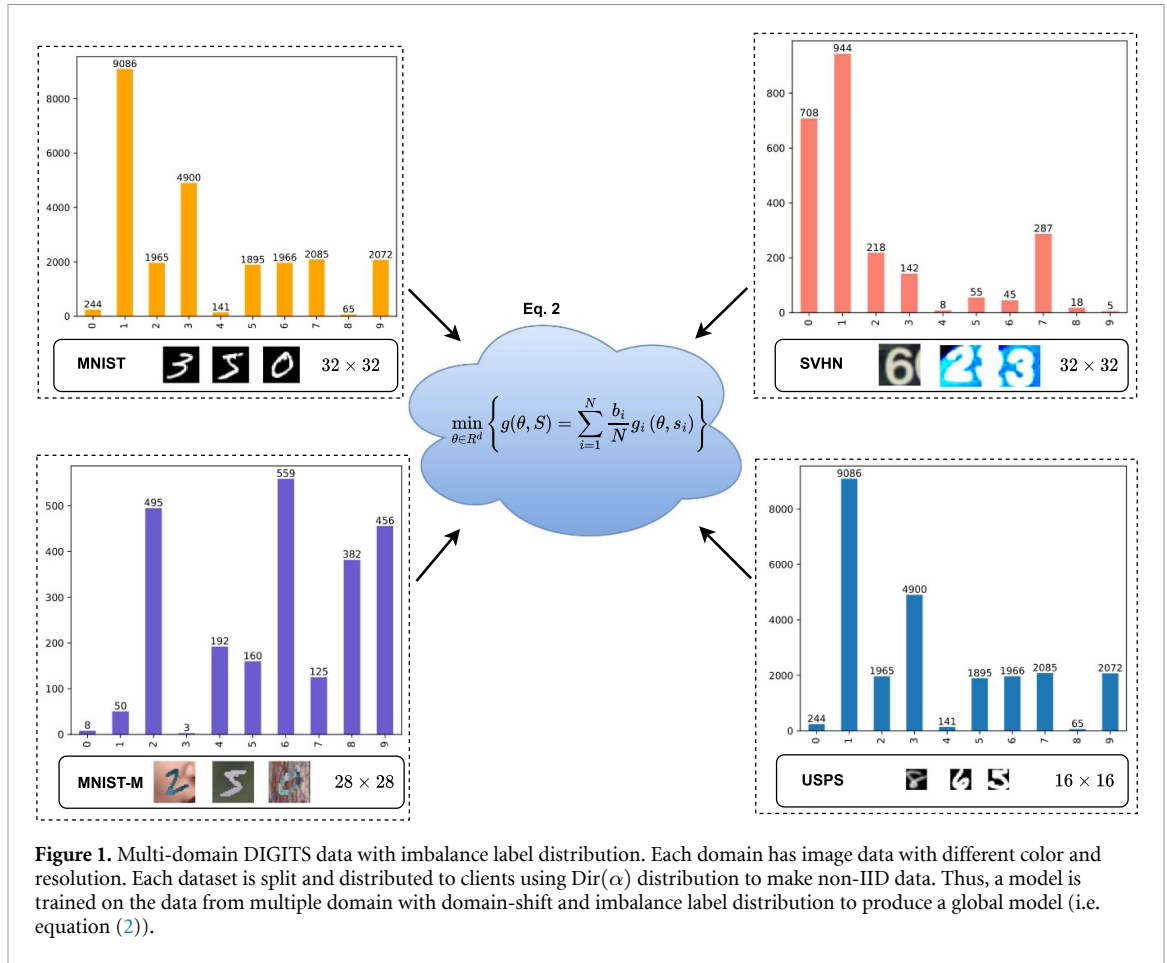
* Author to whom any correspondence should be addressed.

E-mail: hamadnig@gmail.com**Keywords:** class-imbalance, data heterogeneity, domain-shift, federated learning, multi-domain data**Abstract**

Federated learning (FL) is an evolving machine learning technique that allows collaborative model training without sharing the original data among participants. In real-world scenarios, data residing at multiple clients are often heterogeneous in terms of different resolutions, magnifications, scanners, or imaging protocols, and thus challenging for global FL model convergence in collaborative training. Most of the existing FL methods consider data heterogeneity within one domain by assuming same data variation in each client site. In this paper, we consider data heterogeneity in FL with different domains of heterogeneous data by raising the problems of domain-shift, class-imbalance, and missing data. We propose a method, multi-domain FL as a solution to heterogeneous training data from multiple domains by training robust vision transformer model. We use two loss functions, one for correctly predicting class labels and other for encouraging similarity and dissimilarity over latent features, to optimize the global FL model. We perform various experiments using different convolution-based networks and non-convolutional Transformer architectures on multi-domain datasets. We evaluate the proposed approach on benchmark datasets and compare with the existing FL methods. Our results show the superiority of the proposed approach which performs better in term of robust FL global model than the exiting methods.

1. Introduction

Federated learning (FL) is a distributed and collaborative machine learning technique in which multiple clients train a global model while keeping the data at local locations [1]. In centralized training, a model is trained and updated on a cumulative train and test data at a single location (i.e. either client or server). However, in FL, the training dataset is decentralized and located on multiple participating clients where each client only shares its model parameters (i.e. gradients) with the central server rather than sharing the original raw data for training. Moreover, each client trains its local model using its local training data for a given number of local communication rounds, and sends model gradients to the server during the global communication round. After the server update, the revised global model is disseminated to all participants for the subsequent global communication round, and this process iterates for a given number of global rounds. In the recent past, FL was considered as a privacy-preserving machine learning approach to train a global model without sharing clients' private data with the server [1]. However, many recent studies have introduced the vulnerabilities known as gradient leakage in FL, caused by the adversarial networks [2, 3]. To address the problem of gradient leakage, different methods such as gradient clipping [4], representation learning [3] and gradient-free optimization [2] have been introduced in the literature. Thus, FL with additional security layers, has evolved by enabling decentralized training of multiple participants without sharing of confidential data. This characteristic of FL makes it useful and beneficial in different areas such as communication networks [5], health care [6–10], organizations [11, 12], and smart cities [13] where privacy and confidentiality of sensitive data are crucial for the data owners. Although, FL provides rich opportunities in many fields, there are many research challenges in the implementation of FL for real-world problems.



One of the most common problem in FL is the data heterogeneity or different data distribution among multiple domains. Moreover, imbalanced class problem or label distribution in multi-domain environment is common for the real-world data in which non-uniform labels are distributed across the classes in a dataset. In real-world data, some classes contain only a few samples, but many others have a large number of samples of such classes [14–16]. Many methods have been proposed to solve the problem of non-uniform label distribution in collaborative training [15–20]. However, most of the existing methods face challenges of non-independent and identical data distribution (non-IID), causing uncertainty in fast and complete model convergence [21–23]. Such existing methods have achieved good results by solving the problem of data heterogeneity in FL, but most of the methods focus on single domain scenario in which data splits across participants are taken from the same domain for all participating clients. These existing methods train a model on imbalanced data from different splits of the same domain, and try to generalize on a balanced test dataset.

Instead of focusing on the improvement of FL optimization as most of the existing methods [24, 25], our work is based on multi-domain data heterogeneity i.e. class-imbalance and domain-shift in the scenario of collaborative training with different participants from different domains. In the proposed multi-domain FL (MDFL) scenario, we use data from different domains each having the same general categories (i.e. class labels) but different distributions and underlying patterns. Thus, severe domain-shifts and diverse distributions are major challenges for the convergence of global model in FL. As shown in figure 1, the key challenges in MDFL are as follows:

- Domain-shift due to data heterogeneity within and across the domains. Data from different domains have different variations such as color variation and image resolution.
- Different distribution of labels (i.e. class-imbalance) within and across the domains. Class-imbalance becomes more challenging for local training of a participant and eventually global model, when dataset of a single domain is split into multiple clients.
- Minority and missing classes within and across the domains. Thus, some of the participants have minority and missing classes in their local training data.

Transformer models [26, 27] have been used to solve the problem of data heterogeneity in classification tasks [28–31] by demonstrating the resilience against heterogeneous data [32, 33]. The robustness exhibited by Transformers makes them well-suited for self-supervised learning [34, 35], especially for data heterogeneity based on domain and distribution shifts in training data. Thus, in our method, we exploit Transformer architectures for the training on heterogeneous data with domain-shifts and diverse distributions across multiple participating domains in FL. Existing studies [32, 36] have indicated the superior performance of Transformer models in comparison to commonly used ResNet [37] and other convolution-based networks [38, 39]. The reason for the success lies in the Transformer architecture comprising attention heads that help contextual awareness in image interpretation [36]. As the attention mechanism helps to capture contextual dependencies more effectively, we posit that this property contributes to superior performance in heterogeneous FL. The convergence of Transformers is fast and their global model is suitable for most devices. We compare our results with the existing FL methods, and conclude that vision transformers (ViTs) perform better without additional hyperparameters and training. Therefore, they are appropriate for the future research in FL problems.

Moreover, to minimize the inconsistency of domain-shift and diversity in multi-domain data distribution, we use two loss functions to optimize and improve the performance of global model. During training, one loss is computed on the latent feature vectors and the other one is calculated by class-logits of the model to optimize the global model in MDL. We evaluate and compare our model with the existing methods using non-IID and heterogeneous data splits with domain-shift from multiple source domains. Experimental results suggest that performance of the proposed method is better compared to the similar existing methods.

Our main contributions are as follows:

- (i) We formulate the problem of data heterogeneity based on class-imbalance and domain-shift within and across the domains in FL.
- (ii) We propose a method MDL, by training robust Transformer model to improve the performance of global model trained on heterogeneous data with diverse distribution, class-imbalance and domain-shift from multiple domains in FL.
- (iii) We use two loss functions; a loss \mathcal{L}_C calculated by the class-logits of the model to correctly predict the class labels, as given in equation (7), and a loss \mathcal{L}_B computed on the latent feature vectors to align the classes across domains, as given in equation (8), to optimize and improve the global model trained on diverse data splits from multiple domains.
- (iv) We evaluate our method on benchmark datasets by training our model on multi-domain data with diverse distribution and domain-shift in MDL. Experimental results indicate the better performance of the proposed method.

2. Related work

2.1. Class-imbalance and label distribution

Much research has been done on the class-imbalance problem [15–20, 40], and different solutions have been proposed to solve this problem including under-sampling and over-sampling [41, 42], reconciliation of loss function [15, 17, 43, 44], and learning paradigms such as self-supervised learning [16, 45], transfer learning [18], ensemble learning [46, 47], metalearning [48], and metric learning [49]. All these methods have been used in the scenario of a single domain and use the data splits for all participants from the same domain, while we extend the data heterogeneity problem to multi-domain and imbalance classes in FL environment.

2.2. Multi-domain learning

In multi-domain learning, a model must be adaptive and robust to data from multiple domains containing different label distributions [50] which is similar to transfer learning [51]. The objective of domain adaptation is to learn a model for a single target domain [28, 51], while multi-domain learning is focused on the average performance of all source domains and their distributions [52]. The existing methods are based on a single-domain data [50, 53], which exploit domain-invariant features [52, 54–56] and multi-task learning [57]. We are focused on the class-imbalance within and across domains in FL environment. Our problem is similar to domain generalization, in which a model is trained on multiple domains and generalized for an unseen domain [58]. Most existing methods are based on data augmentation [59, 60], domain-invariant features [54, 55, 61], meta-learning [62, 63], and casual relationships [64, 65]. These methods are based on a single domain and have not explored the class-imbalance problem within and across

domains in the scenario of domain-shift, especially in FL environment. In this paper, we investigate the effect of data heterogeneity and class-imbalance in MDL environment.

2.3. FL

FL provides distributed and collaborative training on private data from multiple sources [1]. There are two main categories of effective distributed training [66] that have been evolved: (1) serial FL methods allow training of multiple clients in a cyclic and serial manner such as split training [67] and cyclic weight transfer (CWT) [7], whereas (2) in parallel FL methods, training of each participant is parallel, such as FedAvg [1]. FL presents the challenge of domain-shift and class-imbalance across participants in FL training. Such data heterogeneity in FL causes non-guaranteed model convergence and forgetting problem for cyclic FL methods [7, 68, 69], and divergence in model weights for parallel FL methods [22, 70–72].

FedAvg algorithm [1] has been widely used in different variations such as FedAVGM [73] to use the server momentum to mitigate the problem of class-imbalance and distribution-shift for each client. It has been used with some optimization methods, such as matching feature layers [74, 75], collaborative replay [76], model distillation [77], and unsupervised contrastive learning [72] to address the problem of heterogeneity in data. It has also been implemented as FedAvg-Share [72] by sharing small chunks of data among participating users with an additional proximal term (FedProx) to the local objective, which reduces the potential weight divergence [22].

At the same time, several methods have been presented to solve the problem of catastrophic forgetting in cyclic FL methods. Such methods restrict the weight updates that are required and important for historical tasks, known as elastic weight consolidation [78]. These methods implement cyclic weighted objectives to reduce loss due to the skewness of the label distribution [28], and deep generative replay to mimic data from historical tasks or the client [76, 79]. However, most existing methods focus on optimization techniques without inspecting the model architecture for domain and distribution shift of data, to increase the robustness and performance of the model. In our work, the experimental results are consistent with the hypothesis that an architectural change in the model makes a huge difference and should be explored for optimization methods, which is the main focus of our work.

2.4. Transformers

The Transformer architecture firstly proposed by [26], has been implemented in sequence-to-sequence machine translation, and then in self-supervised natural language processing tasks [34]. Transformers have been widely used in image and video tasks in recent years. For example, self-attention has been applied to the local neighborhoods of the image in [80]. Similarly, global self-attention has been applied to full-size images using ViT [81] for the ImageNet classification task, and state-of-the-art performance is achieved. Therefore, Transformers have shown a prominent performance increase compared to classical vision networks such as CNN [37, 82], language models (i.e. LSTMs [83]), and are attracted to understanding the causes of their effectiveness. Many existing methods have proven the effectiveness and robustness of ViTs to severe domain-shifts, perturbations, and occlusions [5, 32]. Furthermore, recent methods have demonstrated the effectiveness and suitability of Transformers for multi-modal and heterogeneous data [33, 35, 84]. Inspired by the above studies, our hypothesis is that ViTs will perform better by adapting the domain-shift, class-imbalance, and overall heterogeneity of the data in FL. We conducted a considerable number of experiments and gave a detailed empirical analysis to validate the hypothesis.

3. Methodology

In this section, we formulate the problem of data diversity in MDL using transferability, and demonstrate our approach to minimize the heterogeneity effect of multi-domain data. Moreover, we demonstrate our proposed approach using end-to-end model training with Transformer architectures and proposed losses in MDL environment.

3.1. Transferability

To explain the problem of multi-domain data heterogeneity, we visualize the data distribution of different domains with respect to their variations across the domains. In a classification task, a domain space $D = 1, 2, 3, \dots, D$, each domain having a label space $C = 1, 2, 3, \dots, C$ can be represented as a training set $T = \{(d_i, x_i, c_i)\}_{i=1}^N$, where $x_i \in \mathbb{R}$ is input, $c_i \in C$ denotes the class label, and $d_i \in D$ represents a domain in the dataset. To represent the variation between domains and classes, we denote domain samples d and class samples c as domain-class pairs (d, c) as part of the training data, represented as $T_{d,c} \subseteq T$. When we have

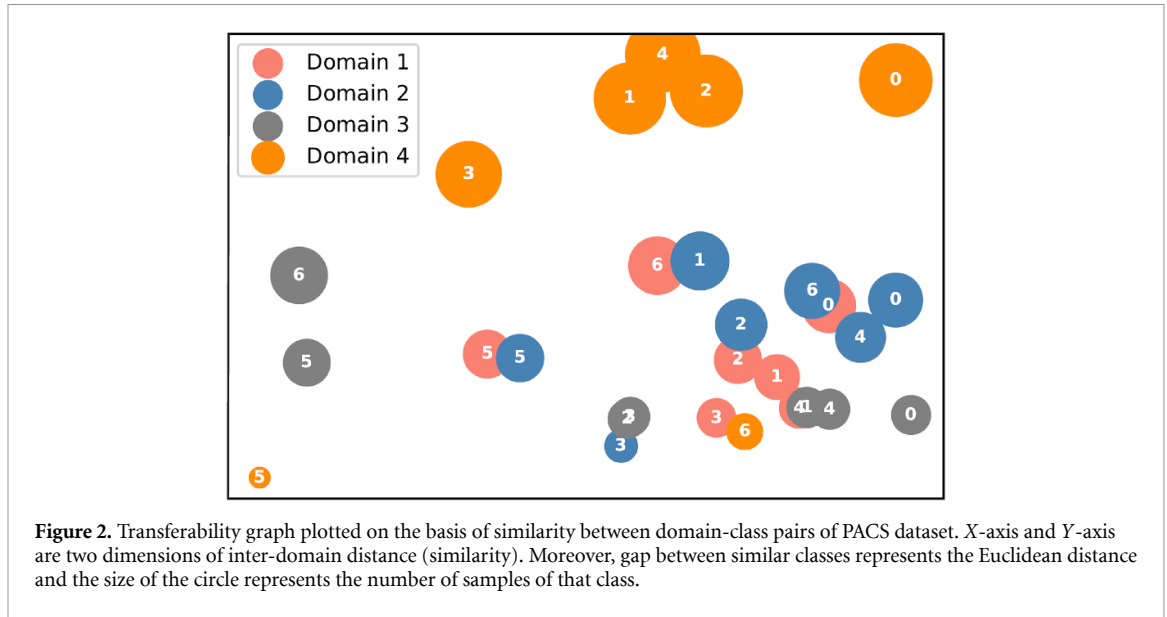


Figure 2. Transferability graph plotted on the basis of similarity between domain-class pairs of PACS dataset. X-axis and Y-axis are two dimensions of inter-domain distance (similarity). Moreover, gap between similar classes represents the Euclidean distance and the size of the circle represents the number of samples of that class.

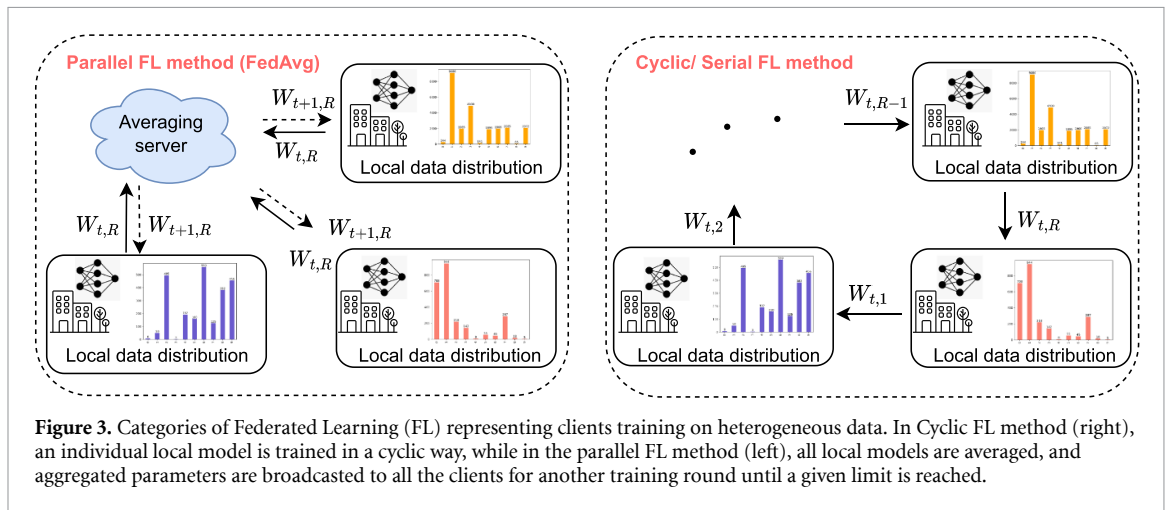


Figure 3. Categories of Federated Learning (FL) representing clients training on heterogeneous data. In Cyclic FL method (right), an individual local model is trained in a cyclic way, while in the parallel FL method (left), all local models are averaged, and aggregated parameters are broadcasted to all the clients for another training round until a given limit is reached.

domain-class pairs, our objective is to compute the transferability between two domain-class pairs, which is the average distance between their individual feature representations. Transferability characterizes and measures the closeness of domain-class pairs in the feature space. Thus, in a feature space, the transferability of a domain-class pair is the transformation from the original domain-class pair (d, c) to the transformed domain-class pair (d', c') , which is represented as follows.

$$Trf\{(d, c), (d', c')\} \approx [E_d(z, \mu_{d', c'})], \tag{1}$$

where E_d is the Euclidean distance measured by first-order statistics μ (i.e. mean) from representation space $z = g(x, \theta)$, and $g : X \rightarrow Z$ maps input $x \in \mathbb{R}$ to representation space $z \in \mathbb{R}$.

Transferability can be visualized as a graph in a 2D Cartesian space by taking the average of $Trf\{(d, c), (d', c')\}$ as a similarity measure. As a representative case, we plot the transferability graph for PACS [85] data. For each domain-class pair (i.e. (d, c)), the mean $\mu_{d, c}$ is estimated from the learned representations and the distance matrix is calculated for the transferability graph as shown in figure 2, where each color represents domain samples with the circle size as the number of samples in that domain. Moreover, the distance among similar classes indicates the Euclidean distance between them. Our objective is to minimize the distance between domain-class pairs in all domains of a dataset.

3.2. MDFL

As shown in figure 3, there are two main categories of FL based on the gradient merging mechanism for all participating models. (1) CWT in which an individual local model is trained to become an ultimate global

model in a serial and a cyclic way for every round of communication. Subsequently, the global model is transferred to the next client for the same process [7]. In this way, each client participates in the training for a given number of local epochs. This process continues until the global model converges and reaches a specific number of given rounds for communication cycles. (2) Federated Averaging (FedAvg) in which a local model is trained on local data of a participating client performing a stochastic gradient descent. Afterwards, all local models are averaged, and these averaged parameters are broadcasted to every participant. The process is repeated until the global model converges for the given number of rounds [1]. In our experiments, we apply the most common parallel FL method, FedAvg [1] as used by most existing methods [11, 86, 87]. The overall objective of FL is to minimize the loss to achieve the best global model as given below.

$$\min_{\theta \in \mathbb{R}^d} \left\{ g(\theta, S) = \sum_{i=1}^N \frac{b_i}{N} g_i(\theta, s_i) \right\}, \quad (2)$$

where b_i represents the batch size for the participant $i \in N$ having model parameters θ , and $s_i \in S$ is the local data of a participant.

Furthermore, we employ both convolution-based networks and non-convolutional Transformer architectures for training and evaluating the model across multi-domain data. We implement a variety of commonly used convolution-based classification models, including our custom model (referred to as CustomNet), LeNet-1 [39], CNN-2 [38], ResNet18 [37], and EfficientNet [88]. These models incorporate convolution and pooling operations within their architectures. CustomNet is similar to CNN-2 [38] model, except that it contains an additional maxpool layer before each activation layer in the architecture. We also use an adaptive-average pooling layer in each convolution-based network to handle the variation in image size from different domains. We exploit ViTs as non-convolutional models, specifically ViT (T) and ViT (S), in our implementation for a fair comparison. In contrast to traditional convolutional models, these architectures do not utilize convolutional layers in their designs.

Transformers use patches of the image known as tokens to learn the features. The robustness of their architecture is due to self-attention, which aggregates the image information. Their non-convolutional architecture is based on layers, which investigate the average distance among learned weights. The attention distance and its variability from higher to lower layers is compared, which is almost uniform throughout the network going deeper. This ability of Transformers is significant for contextual relationship to interpret an image which is different from convolution-based architectures. If a domain i contains data samples s_i , then a domain-adaptive Transformer attention Att with Q_i , K_i , and V_i can be represented as follows.

$$\text{Att}(Q_i, K_i, V_i) = \text{softmax} \left(\frac{Q_i K_i^T}{\sqrt{d_k}} \right) V_i, \quad (3)$$

where d_k represents dimensionality of the key vectors.

For multi-domain data, aggregated attention is formulated as:

$$A_{\text{agg}} = \sum_{i=1}^N \gamma_i \cdot \text{Att}(Q_i, K_i, V_i), \quad (4)$$

where γ_i is weight of a domain i , and $\sum_{i=1}^N \gamma_i = 1$.

Moreover, for a domain i , if the input data x_i is transformed to embedding $E_i = \text{Emb}(x_i)$, and loss \mathcal{L}_i for that domain is $\mathcal{L}_i = (E_i, \theta_i)$, then the loss function \mathcal{L}_C (i.e. cross-entropy loss) for MDFL can be mathematically represented as follows.

$$\mathcal{L}_C = \sum_{i=1}^N \beta_i \cdot \mathcal{L}_i(E_i, \theta_i), \quad (5)$$

where hyperparameter β controls the domain importance, E_i is embedding, and θ_i are model parameters for a domain i . The loss \mathcal{L}_C is minimized to optimize and generalize the model trained on multi-domain data with heterogeneous domain and label distribution.

3.3. Training loss

For the training of a global model, we use two loss functions to formulate an overall loss to minimize for the optimization, as given below.

$$\mathcal{L} = \mathcal{L}_C + \lambda \mathcal{L}_B, \quad (6)$$

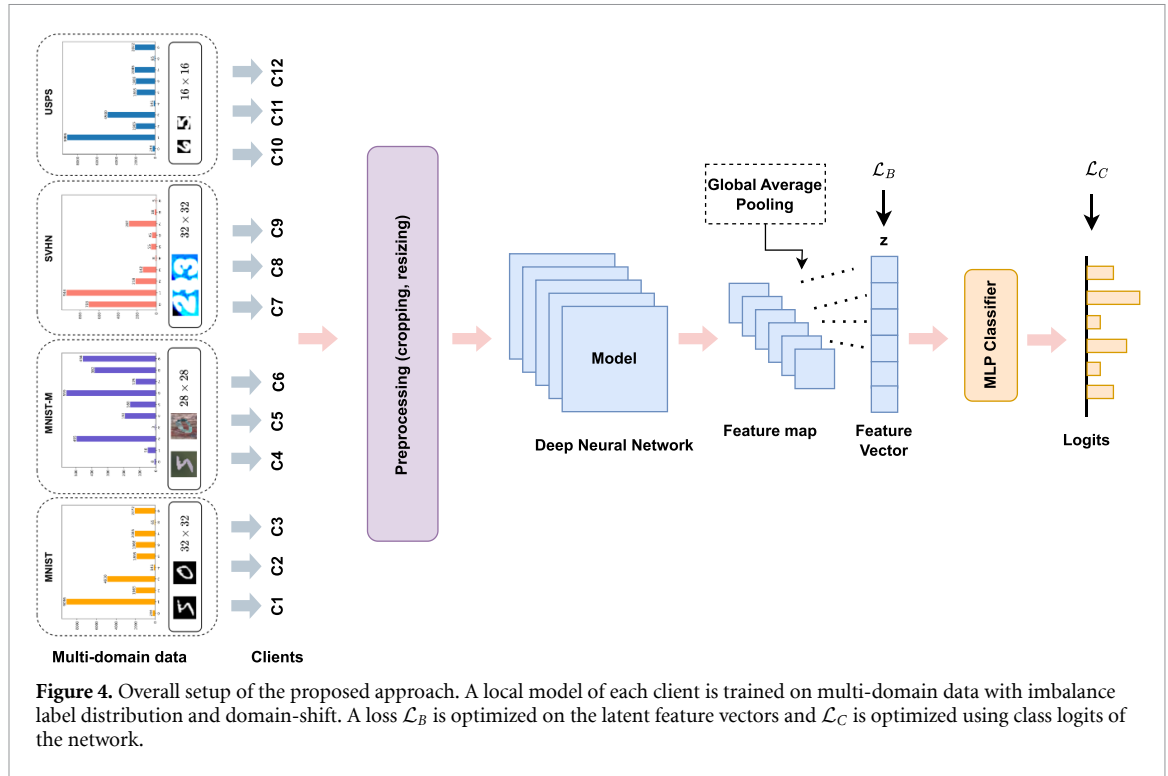


Figure 4. Overall setup of the proposed approach. A local model of each client is trained on multi-domain data with imbalance label distribution and domain-shift. A loss \mathcal{L}_B is optimized on the latent feature vectors and \mathcal{L}_C is optimized using class logits of the network.

where λ is a hyperparameter representing trade-off between two losses. \mathcal{L}_C denotes the standard cross-entropy loss computed on the final output layer (i.e. class logits) and can be defined as follows.

$$\mathcal{L}_C(y, y') = -\frac{1}{b} \sum_{n=1}^N y_n \log y'_n + (1 - y_n) \log (1 - y'_n), \quad (7)$$

where y'_n is the predicted output for input image n and b is the batch size. In (6), \mathcal{L}_B is used to represent the balanced domain-class distribution alignment (BoDA) loss as proposed in [89]. \mathcal{L}_B is a special loss introduced to reduce the heterogeneity effect in a dataset. Here, we use \mathcal{L}_B to align the classes across domains, and to minimize the negative impact on FL training produced by data diversity from multiple domains. It can be mathematically formulated as follows.

$$\mathcal{L}_B(z, \mu) = \sum_{z_i \in \mathcal{Z}} \frac{-1}{|\mathcal{D}| - 1} \sum_{d \in \mathcal{D} \setminus \{d_i\}} \log \frac{\exp\left(-p_{d_i, c_i}^{d, c_i} E_d(z_i, \mu_{d, c_i})\right)}{\sum_{(d', c') \in \mathcal{M} \setminus \{(d_i, c_i)\}} \exp\left(-p_{d_i, c_i}^{d', c'} E_d(z_i, \mu_{d', c'})\right)} \quad (8)$$

where numerator represents the positive distance (i.e. E_d) of domain-class pairs (d, c) which has to be minimized to attract the same classes in the training, while the denominator is a negative distance of domain-class pairs that should be maximized to isolate different classes during training. Therefore, this loss function aims to minimize the distance between domain-class pairs that are similar, while increasing the separation between pairs that are dissimilar. Here, μ is the first-order statistics estimated from feature representations of domain-class pairs, and the calibration parameter $p_{d_i, c_i}^{d', c'}$ is used to control the transferability from (d, c) to (d', c') depending on their sample size. Moreover, Euclidean distance $E_d(z, \mu) = \sqrt{(z - \mu_{d, c})(z - \mu_{d, c})^T}$ takes over the first-order statistics (i.e. mean). The overall training setup of the proposed approach is shown in figure 4, where multi-domain data are distributed to clients (i.e. C1 to C12), and a local model generates feature map which is transformed to feature vectors through global average pooling. The loss \mathcal{L}_B is calculated by the feature vectors, and the loss \mathcal{L}_C is measured by the class logits which is the final output after multi layer perceptron classifier.

4. Experimental results

In our approach, we train the Transformer models and support our hypothesis that they produce a superior global FL model compared to traditional convolution-based architectures. Employing Transformer models

enhances the optimization process of the FL model. Moreover, these models are robust for heterogeneous data within and across domains (i.e. domain-shift), and must be adaptive for new and unseen-domain data.

We use two loss functions to optimize and further improve the performance of the model. Additional loss to model optimization is useful for data heterogeneity, especially domain-shift which is reduced as a measure of transferability. The performance of the FL model, which incorporates transformers and uses two loss functions (i.e. \mathcal{L}_C and \mathcal{L}_B), is assessed and contrasted with the performance of the same model only using \mathcal{L}_C .

4.1. Datasets

In the experiments, DIGITS [54, 90–92] data and PACS [85] data are used by different split-categories. In the DIGITS dataset, domains are divided into multiple subdomains (i.e. clients) using Dirichlet Distribution (i.e. $\text{Dir}(\alpha)$). As used in [87, 93], dataset is split using smaller value of α (i.e. $\alpha = 0.5$) to distribute data as non-IID data among participants. However, PACS dataset is used as multidomain dataset with one domain as a single client without further split distribution. Additional information regarding both datasets is provided below.

4.1.1. DIGITS dataset

We evaluate the proposed method for the digit classification task. In the experiments, we use: (1) MNIST [90] handwritten digits having 32×32 grayscale images with 60 000 samples as train set and 10 000 as test set, (2) MNIST-M dataset [54] having 28×28 color images with about 59 000 training examples and 9 000 testing examples, (3) the street view house numbers (SVHN) [91] dataset with 32×32 color images having 73 257 train samples and 26 032 test samples, and (4) US Postal Service (USPS) dataset [92] having 16×16 grayscale images with 7291 training instances and 2007 testing samples. DIGITS dataset is visualized and described in appendix A.1.

We split each domain data into three clients using $\text{Dir}(\alpha = 0.5)$ to make non-IID as used by other existing methods such as [87, 93]. We use the minimum value of α (i.e. 0.5) to make the data more heterogeneous with imbalanced label distribution across the clients as higher the value of α leads to higher homogeneous data distribution and vice versa. Moreover, we use 3 different domains for training out of 4 domains, each having above mentioned datasets. We distribute the data across clients with different label distribution within and across the domains. We use test set of each domain for its local model evaluation. We perform training to evaluate the model by leave-one-domain-out cross-validation on an unseen-domain data. Thus, for each domain, there are 3 clients and a total of 9 clients for 3 different domains which participate to produce a global model that is evaluated on the 4th unseen domain. We use the test set of 4th domain as a global test set to evaluate the global model. For the implementation of Transformers, we transform the training and testing data to a larger size of 224×224 . In each experiment, a local model is optimized using loss from (6). Moreover, 5 local epochs and 100 communication rounds are used for each data split scenario.

4.1.2. PACS dataset

In the experiments, PACS [85] dataset is used to evaluate the robustness of the proposed method. PACS dataset contains the images of natural photos, art paintings, cartoons, and sketches. In each experiment, a domain is assigned to a participant without further split into subdomains and clients. Thus, 4 domains (i.e. 4 clients) participate in the scenario of PACS dataset. Other hyperparameter settings are same as used by the DIGITS dataset. For further visualization and description of PACS data, please see appendix A.2.

4.2. Technical details

We implement the proposed method with PyTorch configured on the Linux Operating System (i.e. Ubuntu 22.04 LTS) with the installation of NVIDIA GPU (GeForce RTX 3090) having a memory of 24 GB. Moreover, a CPU (i7-8700) with a memory of 50 GB has been used in the experiments. We follow [66] for the hyperparameter settings used in our experiments as summarized in table 1.

4.3. Evaluation metric

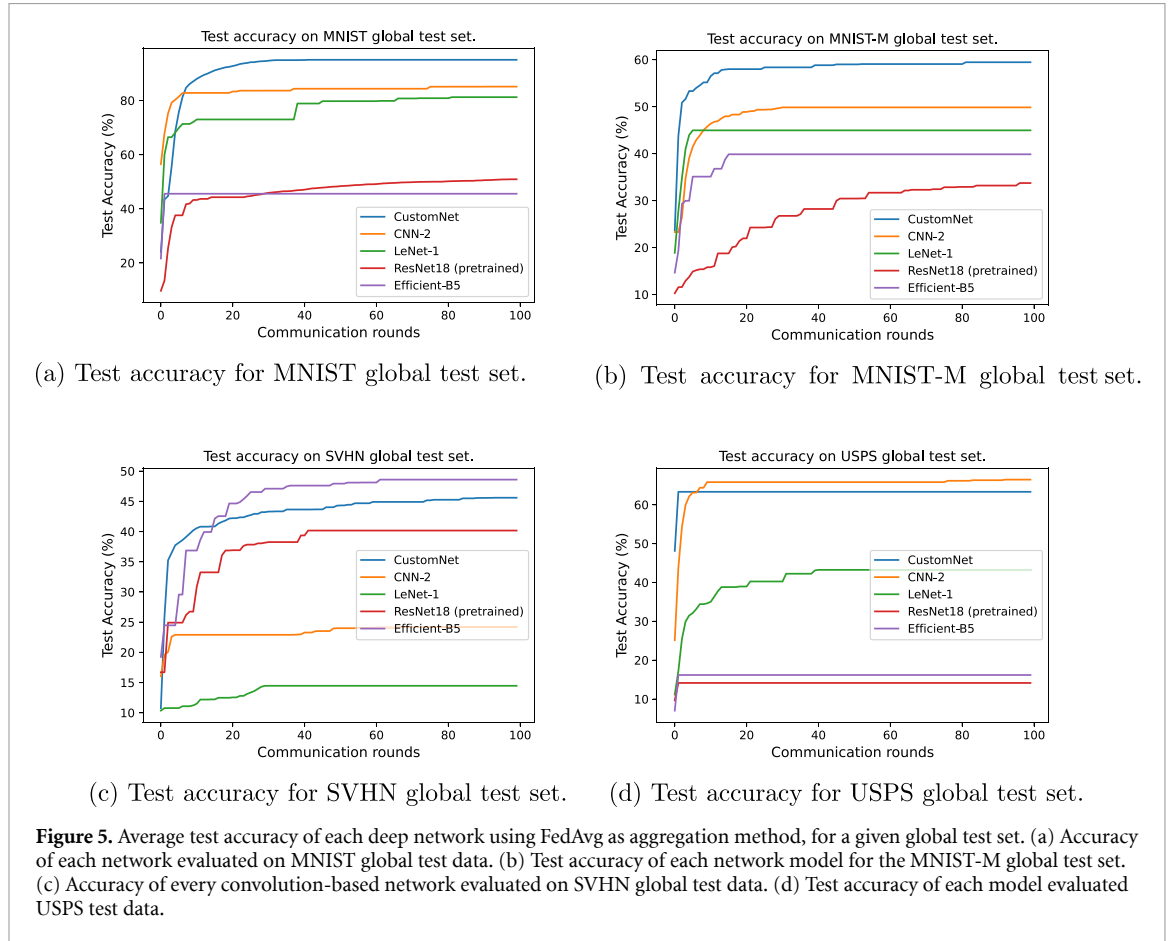
We use a common evaluation metric accuracy in our experiments because of the balanced evaluation data as used in other similar methods [22, 78, 95]. We compare the proposed method based on accuracy with other existing methods.

4.4. Comparison with existing methods

We measure the test accuracy of each model trained as Leave-One-Domain-Out validation on heterogeneous data splits from three different domains of DIGITS dataset. The global model is evaluated on global test set obtained from 4th domain other than included domains in training (i.e. leave-one-domain-out). Test

Table 1. Hyperparameter settings and values used in the experiments.

Hyperparameter	Value (s)
N (domains)	3 & 9
b_i (batch-size)	64
λ	1
α	0.5
local epoch	5
global epoch	100
Optimizer	Adam [94]
Initial learning rate	0.001



accuracy of all convolution-based networks against communication rounds is shown in figure 5. It is observed that in some cases, models are converged, but pretrained ResNet18 and Efficient(B5) are not converged and adaptive when evaluated on USPS test data. Moreover, CNN-2 and LeNet-1 do not perform well in case of SVHN data.

We also measure the test accuracy of each model evaluated on global test data from each individual domain and compare with existing methods as shown in table 2. It is clear from table 2 that the performance of the convolutional models is comparatively lower, while non-convolution-based ViTs performs better which are more robust against domain-shift and imbalanced distribution in the scenario of heterogeneous data from multiple domains. Moreover, it is also noteworthy that ViT(S) performs better as compared to ViT(T). The performance variance observed between ViT(T) and ViT(S) could stem from their distinct tokenization strategies, with ViT(T) employing token mixing tokenization and ViT(S) utilizing spatial tokenization. Moreover, the effectiveness of tokenization strategies also depends on the dataset characterization such as ViT(T) performs better in case of complex spatial relationship that cannot be adequately captured by non-overlapping patches, while ViT(S) performs better if the dataset is comprised of spatially correlated features because it preserves spatial information by dividing the image into

Table 2. Average accuracy with non-homogeneous data ($\alpha = 0.5$) and unseen domain. All results are reported according to leave-one-domain-out settings. These results are also a comparison of the proposed method and other recent FL methods based on the accuracy measured on homogeneous data from multiple domains. The bold text shows the improved performance of the proposed method using transformer models.

Model	MNIST	MNIST-M	SVHN	USPS	Average (%)
CustomNet	95.04	59.29	45.49	63.33	65.79
LeNet-1	77.14	45.76	13.69	43.00	59.90
CNN-2	84.18	50.11	22.61	66.67	55.89
ResNet18 (pretrained)	52.20	28.30	28.93	22.09	32.88
EfficientNet(B5)	53.88	31.73	73.05	15.68	43.58
R50-FedAVG [78]	70.52	75.66	73.58	75.12	73.72
FedProx [22]	81.85	77.45	74.22	69.28	75.70
FedAVG-Share [95]	84.25	79.88	75.96	78.43	79.63
ViT(T)	98.71	84.69	83.58	77.08	86.01
ViT(S)	98.89	88.98	88.15	81.56	89.40

Table 3. Individual and average test accuracy of each domain from PACS dataset including Art painting, cartoons, natural photos, and sketch images.

Model	Art painting	Cartoon	Photo	Sketch	Average (%)
ResNet18	94.01	96.03	98.02	96.60	96.16
EfficientNet(B5)	93.64	94.48	94.85	95.00	94.49
R50-FedAVG [78]	95.00	96.06	98.00	96.90	96.49
FedProx [22]	95.50	95.72	95.26	96.08	95.64
FedAVG-Share [95]	96.18	95.98	96.30	95.86	96.08
ViT(T)	97.58	97.89	98.16	97.56	97.78
ViT(S)	97.45	97.98	98.32	98.11	97.96

non-overlapping patches. However, for both DIGITS and PACS data, ViT(T) and ViT(S) perform better than other convolution-based networks in all cases.

To evaluate the performance of the model for a different dataset, we also use PACS [85] dataset in the experiments. In this dataset, each domain is assigned to a single participant and hence 4 participants perform training on their individual domain data on the basis of Leave-One-Domain-Out settings. The experimental results for the PACS dataset training with selected convolutional and non-convolutional models based on previous results computed on DIGITS dataset, are given in table 3 in which accuracy of an individual domain is measured, and average accuracy is also given in the last column. Table 3 also demonstrates the supremacy of ViTs over other convolutional networks trained on the PACS dataset.

We further exploit additional loss functions given in (6) to optimize the model trained on heterogeneous data of DIGITS from multiple domains. The model's performance experiences a notable improvement when incorporating these additional loss functions for both convolution-based networks and non-convolutional Transformers, as illustrated in figure 6. This is because of the transferability and distance minimization between domain-class pairs which eventually improves the overall performance of trained FL model.

Finally, we measure and compare the average accuracy of each method for both DIGITS and PACS datasets. Accuracy of the global model with and without additional loss has been given in table 4 which shows that ViTs are robust, when used in collaboration with additional loss functions, to heterogeneous data with imbalanced label distribution and domain-shift. Moreover, additional loss increases the performance of the learned model by decreasing the transferability and Euclidean distance between domain-class pairs of multiple domains.

We compare the proposed method with existing FL methods such as ResNet50-FedAVG (i.e. R50-FedAVG) [78], FedProx [22], and FedAVG-Share [95] as given in table 4. Most of the existing methods use convolution-based networks, and the performance of these methods varies from domain to domain. The comparative results clearly demonstrate that ViTs exhibit greater robustness when employed with FedAvg aggregation and the inclusion of additional (i.e. BoDA) loss. They also demonstrate an ability to adapt effectively to new and unseen domains. Moreover, ViTs perform better than existing methods [22, 78, 95] in case of data heterogeneity and different distribution on the basis of domain-shift when data from multiple domains are used for training. Thus, ViTs in addition to losses (equation (6)) perform better compared to existing state-of-the-art methods, and are suitable for MDLFL by solving the problem of data heterogeneity, catastrophic forgetting, and domain-shift.

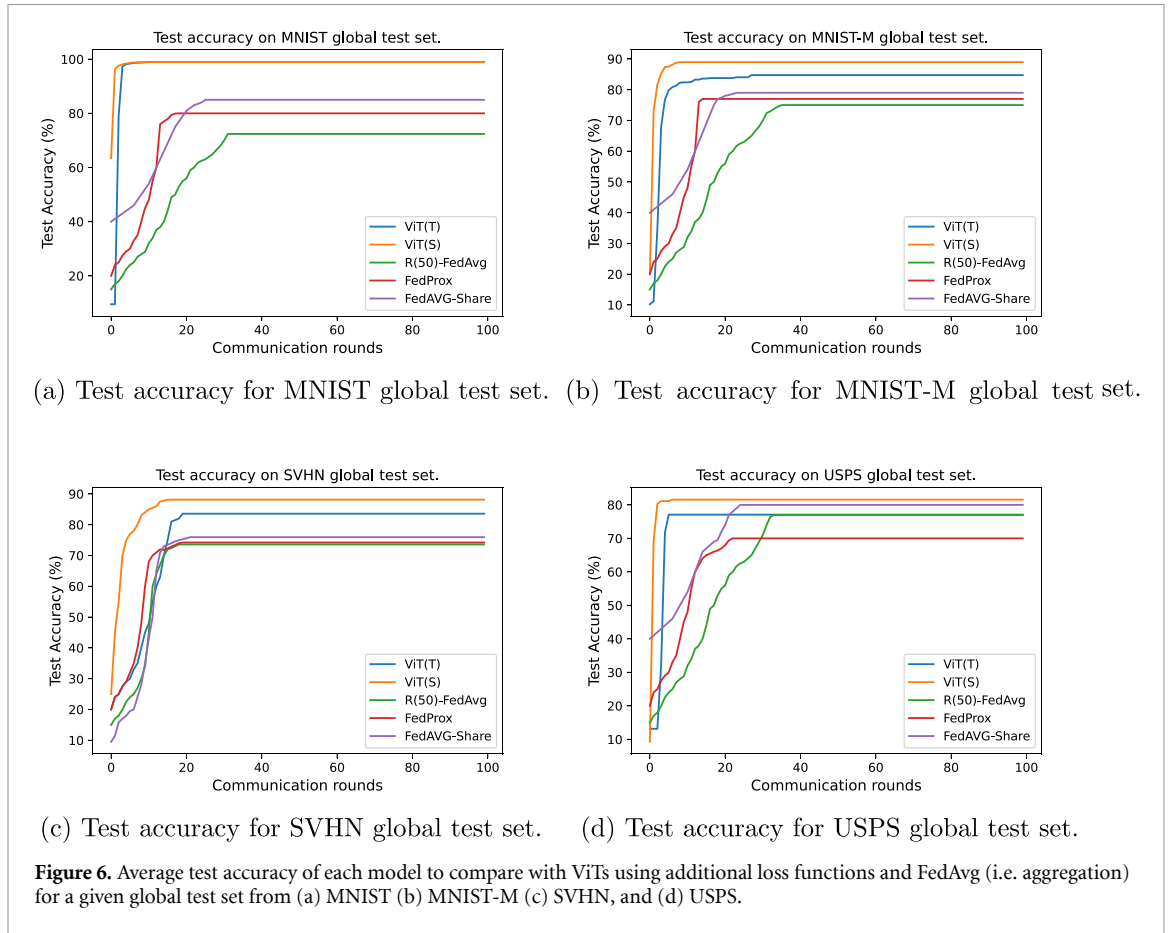


Table 4. Overall average performance accuracy measured for different existing methods and the proposed method using both DIGITS and PACS datasets. Average accuracy is measured for the model with and without addition loss, given in equation (8). Here, + represents the results when \mathcal{L}_B is used for the model training on a given dataset.

Model	DIGITS (%)	DIGITS+ \mathcal{L}_B (%)	PACS (%)	PACS+ \mathcal{L}_B (%)
ResNet18	32.88	95.80	95.70	96.16
EfficientNet (B5)	43.58	96.10	93.74	94.49
R50-FedAVG [78]	73.72	96.04	94.97	96.49
FedProx [22]	75.70	95.96	95.07	95.64
FedAVG-Share [95]	79.63	96.75	94.84	96.08
ViT(T)	86.01	97.20	95.30	97.78
ViT(S)	89.40	97.63	96.50	97.96

4.5. Analysis

As given in section 2, most existing methods perform training on data splits from the same domain to improve the optimization of FL methods. However, in real-world data, different domains possess a divergent data distribution. For example, in our experiments, we use 4 different domains which contain different datasets with variation in colors, image resolution, domain-shift, and heterogeneity in data based-on label distribution. When training the FL model in a real-world scenario, a model should adapt the diversity across domains. Thus, we use data splits from a single domain as well as multiple domains using the leave-one-domain-out training method, so that the test domain does not overlap with the training data. Furthermore, we evaluate the proposed model using unseen domain data that were not included in training or validation of the model. Our results show the effectiveness of ViTs and the additional loss used in such real-world scenarios. We solve the problem of data heterogeneity in the scenario of imbalanced class distribution and domain-shift within and across domains.

To evaluate the robustness of the proposed method, we have used two different data, DIGITS and PACS data sets that contain multiple domains having different datasets. DIGITS dataset is subdivided into multiple splits as non-IID for the participants, while the PACS dataset is distributed to participants as one domain

allocation for one client. Moreover, DIGITS dataset contains digit images with different color, shape and resolution. Likewise, PACS dataset is also heterogeneous on the basis of different label distribution and features containing different images of natural photos, sketches, art paintings and cartoons. Thus, both datasets represent the real-world scenario with multiple domains having heterogeneous data with domain-shift and imbalanced label distribution.

To tackle with the data heterogeneity in MDLFL, we have exploited ViTs that have been used as robust classifiers in different fields. However, we implement ViTs to address the problem of data diversity due to imbalanced label distribution and domain-shift when used in FL model training. Moreover, we incorporate a generalization loss into our MDLFL approach to mitigate the impact of heterogeneous distribution and domain-shift in the training data. The experimental results clearly demonstrate that, within our MDLFL settings, ViTs outperform other convolution-based architectures when combined with the BoDA loss (i.e. equation (8)).

5. Conclusion

We train the robust Transformer model for MDLFL using data from multiple domains characterized by imbalanced class distribution and domain-shift, to solve the problem of data diversity. Transformer architectures are able to solve the problem of device forgetting and learn features efficiently compared to convolution-based deep networks. However, ViTs typically have a large number of parameters, making communication expensive, especially in FL where models are trained across multiple decentralized devices or servers. Moreover, these models require substantial computational resources and significant memory resources during training due to computational complexity and large number of parameters. As we are not concerned with the computational resources in this work, we take advantage of their robustness in case of training with heterogeneous data. We train the global FL model by optimizing two loss functions in latent feature space and class logits of the model. We evaluated the proposed method and compared with the existing methods based on accuracy of the global model. We achieve excellent results in term of accuracy which show the supremacy of the proposed method. In addition to addressing optimization challenges, we can address the prevalent issue of data leakage in FL by incorporating the proposed model with state-of-the-art defense methods in the future.

Data availability statement

All data that support the findings of this study are included within the article (and any supplementary files).

Acknowledgments

This work was partially supported by the Department Strategic Project of the University of Udine within the Inter Department Project on Artificial Intelligence (2020-25).

Appendix. Datasets detail

Here, details of both datasets (i.e. DIGITS and PACS) used in our experiments, are provided.

A.1. DIGITS dataset

DIGITS dataset is comprised of multiple domains including MNIST (i.e. grayscale handwritten digits), MNIST-M (i.e. handwritten digits with color images), SVHN (i.e. colored images of house numbers), and USPS (i.e. handwritten digits with grayscale images) datasets. Each domain has different image data with different properties such as color, size, and resolution as shown in figure A1. To make the data non-IID, $\text{Dir}(\alpha = 0.5)$ is used to split dataset into multiple clients. In our scenario, each domain is divided into 3 subdomains (i.e. clients).

A.2. PACS dataset

PACS [85] dataset contains 4 different domains with images of natural photos, art paintings, cartoons, and sketch as shown in figure A2. The dataset contains images with different attributes (i.e. color, size, and resolution) for each domain. PACS dataset has 4 domains with their individual heterogeneous data on the basis of domain-shift and label distribution, where each domain is assigned to a participating client, so 4 participants train a global model using this dataset in FL environment. To illustrate the domain shift in PACS

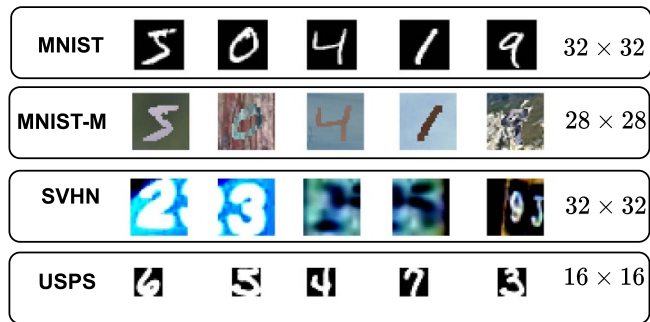


Figure A1. DIGITS dataset: Heterogeneous datasets from different domains with different colors, resolution, and data distribution (i.e. domain shift).

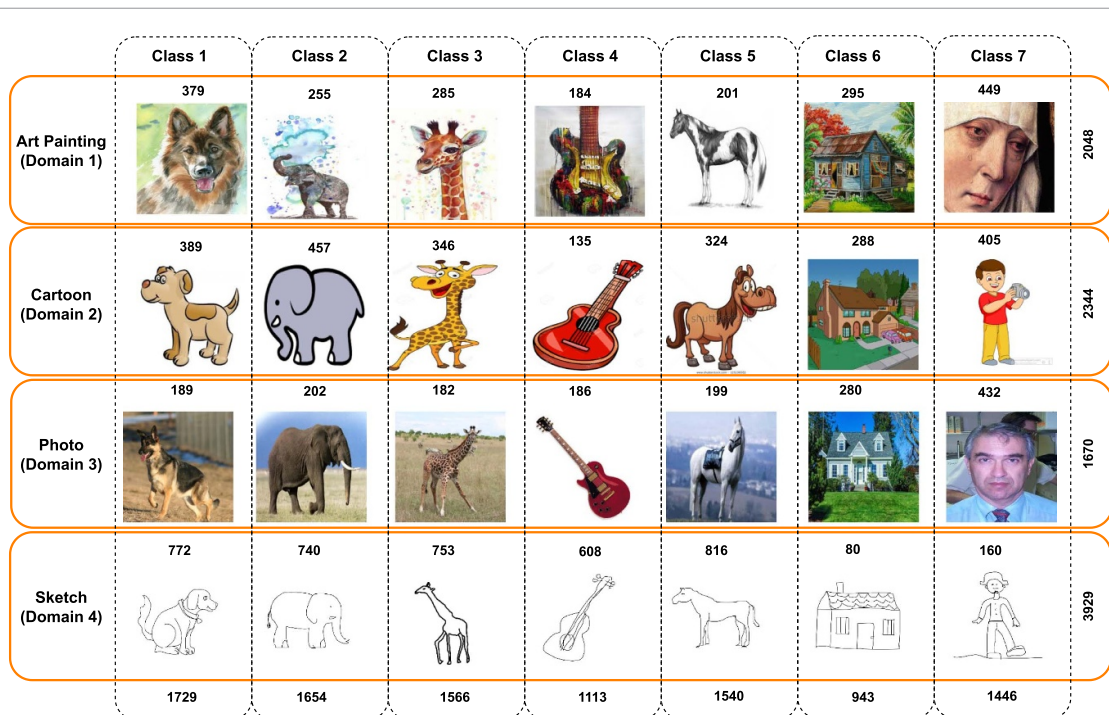


Figure A2. Heterogeneous PACS data from different domains with different colors, size, and data distribution (i.e. domain-shift).

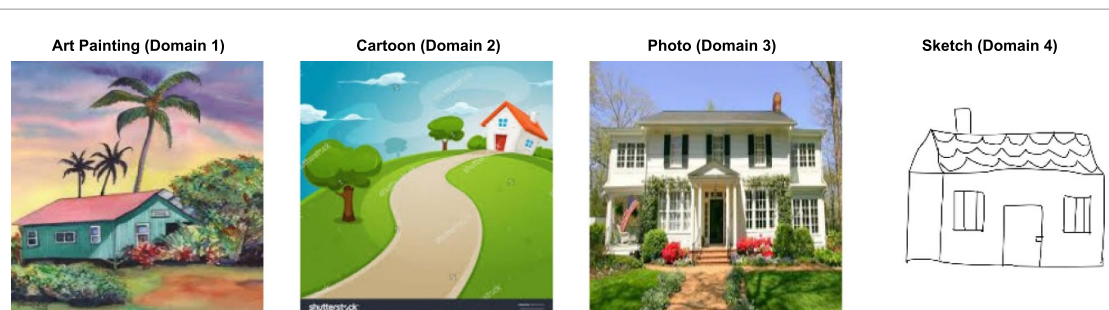


Figure A3. PACS dataset: Heterogeneous dataset containing different domains with different features, colors, and domain-shift. The dataset contains painting, cartoons, natural photos, and sketch images. A global model is trained on this divergent data.

dataset, a diagram is presented in figure A3, and the class distribution is depicted in figure A4. While the imbalanced class distribution may not be notably conspicuous, there is a substantial domain-shift in the dataset, as visualized in figures A3 and A4, respectively.

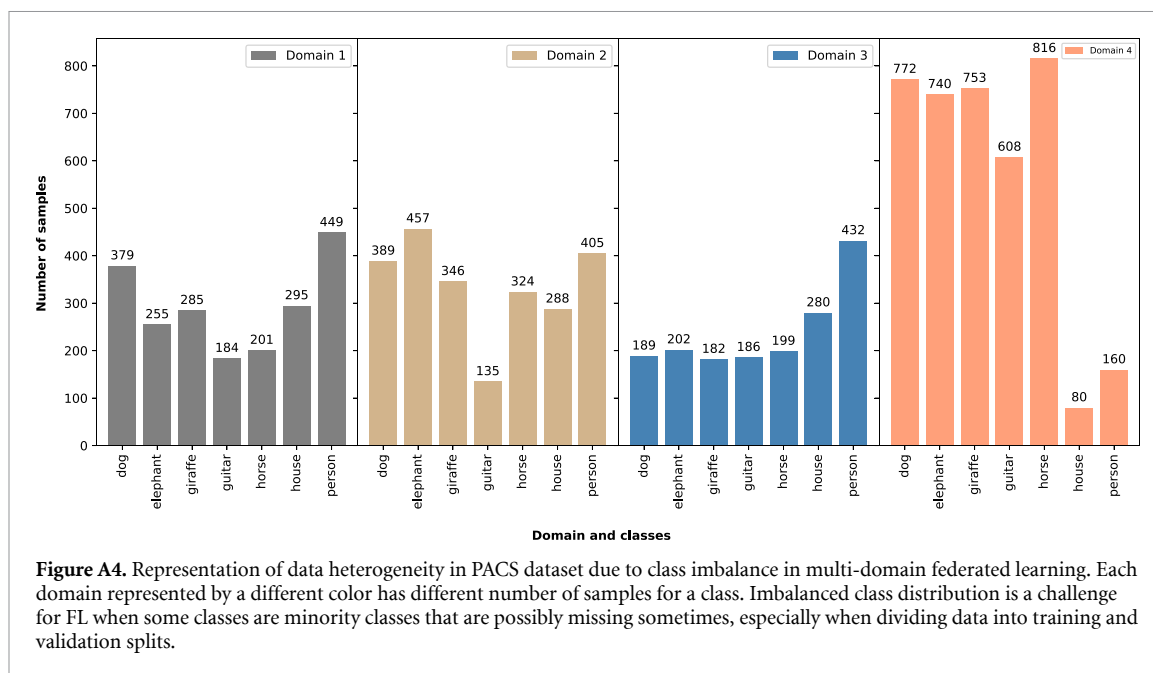


Figure A4. Representation of data heterogeneity in PACS dataset due to class imbalance in multi-domain federated learning. Each domain represented by a different color has different number of samples for a class. Imbalanced class distribution is a challenge for FL when some classes are minority classes that are possibly missing sometimes, especially when dividing data into training and validation splits.

ORCID iD

Hussain Ahmad Madni  <https://orcid.org/0000-0003-1227-524X>

References

- [1] McMahan B, Moore E, Ramage D, Hampson S and Arcas B A Y 2017 Communication-efficient learning of deep networks from decentralized data *Artificial Intelligence and Statistics* (PMLR) pp 1273–82
- [2] Li Z, Zhang J, Liu L and Liu J 2022 Auditing privacy defenses in federated learning via generative gradient leakage *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 10132–42
- [3] Sun J, Li A, Wang B, Yang H, Li H and Chen Y 2021 Soteria: provable defense against privacy leakage in federated learning from representation perspective *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 9311–9
- [4] Zhang X, Chen X, Hong M, Wu Z S and Yi J 2022 Understanding clipping for federated learning: Convergence and client-level differential privacy *Int. Conf. on Machine Learning (ICML 2022)*
- [5] Naseer M M, Ranasinghe K, Khan S H, Hayat M, Shahbaz Khan F and Yang M-H 2021 Intriguing properties of vision transformers *Advances in Neural Information Processing Systems* vol 34 pp 23296–308
- [6] Brisimi T S, Chen R, Mela T, Olshevsky A, Paschalidis I C and Shi W 2018 Federated learning of predictive models from federated electronic health records *Int. J. Med. Inform.* **112** 59–67
- [7] Chang K, Balachandar N, Lam C, Yi D, Brown J, Beers A, Rosen B, Rubin D L and Kalpathy-Cramer J 2018 Distributed deep learning networks among institutions for medical imaging *J. Am. Med. Inform. Assoc.* **25** 945–54
- [8] Guo P, Wang P, Zhou J, Jiang S and Patel V M 2021 Multi-institutional collaborations for improving deep learning-based magnetic resonance image reconstruction using federated learning *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 2423–32
- [9] Liu Q, Chen C, Qin J, Dou Q and Heng P-A 2021 Feddgm: Federated domain generalization on medical image segmentation via episodic learning in continuous frequency space *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 1013–23
- [10] Madni H A, Umer R M and Foresti G L 2023 Federated learning for data and model heterogeneity in medical imaging *Int. Conf. on Image Analysis and Processing* (Springer) pp 167–78
- [11] Madni H A, Umer R M and Foresti G L 2023 Blockchain-based swarm learning for the mitigation of gradient leakage in federated learning *IEEE Access* **11** 16549–56
- [12] Madni H A, Umer R M and Foresti G L 2023 Swarm-fhe: fully homomorphic encryption-based swarm learning for malicious clients *Int. J. Neural Syst.* **33** 2350033
- [13] Jiang J C, Kantarci B, Oktug S and Soyata T 2020 Federated learning in smart city sensing: Challenges and opportunities *Sensors* **20** 6230
- [14] Buda M, Maki A and Mazurowski M A 2018 A systematic study of the class imbalance problem in convolutional neural networks *Neural Netw.* **106** 249–59
- [15] Cao K, Wei C, Gaidon A, Arechiga N and Ma T 2019 Learning imbalanced datasets with label-distribution-aware margin loss *Advances in Neural Information Processing Systems* vol 32
- [16] Yang Y and Xu Z 2020 Rethinking the value of labels for improving class-imbalanced learning *Advances in Neural Information Processing Systems* vol 33 pp 19290–301
- [17] Cui Y, Jia M, Lin T-Y, Song Y and Belongie S 2019 Class-balanced loss based on effective number of samples *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 9268–77
- [18] Liu Z, Miao Z, Zhan X, Wang J, Gong B and Yu S X 2019 Large-scale long-tailed recognition in an open world *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 2537–46
- [19] Ren J et al 2020 Balanced meta-softmax for long-tailed visual recognition *Advances in Neural Information Processing Systems* vol 33 pp 4175–86

- [20] Yang Y, Zha K, Chen Y, Wang H and Katabi D 2021 Delving into deep imbalanced regression *Int. Conf. on Machine Learning (PMLR)* pp 11842–51
- [21] Li Q, He B and Song D 2021 Model-contrastive federated learning *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 10713–22
- [22] Li T, Sahu A K, Zaheer M, Sanjabi M, Talwalkar A and Smith V 2020 Federated optimization in heterogeneous networks *Proc. Machine Learning and Systems* vol 2 pp 429–50
- [23] Gao L, Fu H, Li L, Chen Y, Xu M and Xu C-Z 2022 Feddc: Federated learning with non-iid data via local drift decoupling and correction *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 10112–21
- [24] Huang W, Ye M, Shi Z, Li H and Du B 2023 Rethinking federated learning with domain shift: a prototype view *2023 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)* (IEEE) pp 16312–22
- [25] Zhang P, Chen N, Li S, Choo K-K R, Jiang C and Wu S 2023 Multi-domain virtual network embedding algorithm based on horizontal federated learning *IEEE Trans. Inform. Forensics Secur.* **18** 3363–75
- [26] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser Ł and Polosukhin I 2017 Attention is all you need *Advances in Neural Information Processing Systems* vol 30
- [27] Dosovitskiy A et al 2021 An image is worth 16×16 words: transformers for image recognition at scale *Int. Conf. on Learning Representations* pp 1–21
- [28] Balachandrar N, Chang K, Kalpathy-Cramer J and Rubin D L 2020 Accounting for data variability in multi-institutional distributed deep learning for medical imaging *J. Am. Med. Inform. Assoc.* **27** 700–8
- [29] Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, Lin S and Guo B 2021 Swin transformer: hierarchical vision transformer using shifted windows *Proc. IEEE/CVF Int. Conf. on Computer Vision* pp 10012–22
- [30] Krizhevsky A et al 2009 *Learning multiple layers of features from tiny images* (University of Toronto)
- [31] Liu Z, Luo P, Wang X and Tang X 2015 Deep learning face attributes in the wild *Proc. IEEE Int. Conf. on Computer Vision* pp 3730–8
- [32] Bhojanapalli S, Chakrabarti A, Glasner D Li D, Unterthiner T and Veit A 2021 Understanding robustness of transformers for image classification *Proc. IEEE/CVF Int. Conf. on Computer Vision* pp 10231–41
- [33] Tsai Y-H H, Bai S, Liang P P, Kolter J Z, Morency L-P and Salakhutdinov R 2019 Multimodal transformer for unaligned multimodal language sequences *Proc. Conf. Association for Computational Linguistics. Meeting* vol 2019 (NIH Public Access) p 6558
- [34] Devlin J, Chang M-W, Lee K and Toutanova K 2018 Bert: pre-training of deep bidirectional transformers for language understanding (arXiv:1810.04805)
- [35] Hu R and Singh A 2021 Unit: multimodal multitask learning with a unified transformer *Proc. IEEE/CVF Int. Conf. on Computer Vision* pp 1439–49
- [36] Paul S and Chen P-Y 2022 Vision transformers are robust learners *Proc. AAAI Conf. on Artificial Intelligence* vol 36 pp 2071–81
- [37] He K, Zhang X, Ren S and Sun J 2016 Deep residual learning for image recognition *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* pp 770–8
- [38] Reddi S, Charles Z, Zaheer M, Garrett Z, Rush K, Konecny J, Kumar S and McMahan H B 2020 Adaptive federated optimization (arXiv:2003.00295)
- [39] Verdhan V and Verdhan V 2021 Image classification using lenet *Computer Vision Using Deep Learning: Neural Network Architectures With Python and Keras* pp 67–101
- [40] Zhang Y, Kang B, Hooi B, Yan S and Feng J 2021 Deep long-tailed learning: a survey (arXiv:2110.04596)
- [41] Chawla N V, Bowyer K W, Hall L O and Kegelmeyer W P 2002 Smote: synthetic minority over-sampling technique *J. Artif. Intell. Res.* **16** 321–57
- [42] He H, Bai Y, Garcia E A and Li S 2008 Adasyn: adaptive synthetic sampling approach for imbalanced learning *2008 IEEE Int. Joint Conf. on Neural Networks (IEEE World Congress on Computational Intelligence)* (IEEE) pp 1322–8
- [43] Dong Q, Gong S and Zhu X 2018 Imbalanced deep learning by minority class incremental rectification *IEEE Trans. Pattern Anal. Mach. Intell.* **41** 1367–81
- [44] Huang C, Li Y, Loy C C and Tang X 2019 Deep imbalanced learning for face recognition and attribute prediction *IEEE Trans. Pattern Anal. Mach. Intell.* **42** 2781–94
- [45] Li T, Cao P, Yuan Y, Fan L, Yang Y, Feris R S, Indyk P and Katabi D 2022 Targeted supervised contrastive learning for long-tailed recognition *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 6918–28
- [46] Wang X, Lian L, Miao Z, Liu Z and Yu S X 2020 Long-tailed recognition by routing diverse distribution-aware experts (arXiv:2010.01809)
- [47] Zhang Y, Hooi B, Hong L and Feng J 2021 Self-supervised aggregation of diverse experts for test-agnostic long-tailed recognition (arXiv:2107.09249)
- [48] Shu J, Xie Q, Yi L, Zhao Q, Zhou S, Xu Z and Meng D 2019 Meta-weight-net: learning an explicit mapping for sample weighting *Advances in Neural Information Processing Systems* vol 32
- [49] Zhang X, Fang Z, Wen Y, Li Z and Qiao Y 2017 Range loss for deep face recognition with long-tailed training data *Proc. IEEE Int. Conf. on Computer Vision* pp 5409–18
- [50] Dredze M, Kulesza A and Crammer K 2010 Multi-domain learning by confidence-weighted parameter combination *Mach. Learn.* **79** 123–49
- [51] Pan S J and Yang Q 2010 A survey on transfer learning *IEEE Trans. Knowl. Data Eng.* **22** 1345–59
- [52] Schoenauer-Sebag A, Heinrich L, Schoenauer M, Sebag M, Wu L F and Altschuler S J 2019 Multi-domain adversarial learning (arXiv:1903.09239)
- [53] Xiao T, Li H, Ouyang W and Wang X 2016 Learning deep feature representations with domain guided dropout for person re-identification *Proc. IEEE Conf. on Computer Vision and Pattern Recognition* pp 1249–58
- [54] Ganin Y, Ustinova E, Ajakan H, Germain P, Larochelle H, Laviolette F, Marchand M and Lempitsky V 2016 Domain-adversarial training of neural networks *J. Mach. Learn. Res.* **17** 2096–2030
- [55] Li Y, Gong M, Tian X, Liu T and Tao D 2018 Domain generalization via conditional invariant representations *Proc. AAAI Conf. on Artificial Intelligence* vol 32 pp 1–9
- [56] Sun B and Saenko K 2016 Deep coral: correlation alignment for deep domain adaptation *Computer Vision–ECCV 2016 Workshops (Proc., Part III) (Amsterdam, The Netherlands, 8–10 and 15–16 October 2016)* vol 14 (Springer) pp 443–50
- [57] Yang Y and Hospedales T M 2014 A unified perspective on multi-domain and multi-task learning (arXiv:1412.7489)
- [58] Zhou B, Cui Q, Wei X-S and Chen Z-M 2020 Bbn: bilateral-branch network with cumulative learning for long-tailed visual recognition *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 9719–28

- [59] Carlucci F M, D’Innocente A, Bucci S, Caputo B and Tommasi T 2019 Domain generalization by solving jigsaw puzzles *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 2229–38
- [60] Zhou K, Yang Y, Qiao Y and Xiang T 2021 Domain generalization with mixstyle (arXiv:2104.02008)
- [61] Muandet K, Baldazzi D and Schölkopf B 2013 Domain generalization via invariant feature representation *Int. Conf. on Machine Learning* (PMLR) pp 10–18
- [62] Li D, Yang Y, Song Y-Z and Hospedales T 2018 Learning to generalize: Meta-learning for domain generalization *Proc. AAAI Conf. on Artificial Intelligence* vol 32 pp 1–8
- [63] Zhang M, Marklund H, Dhawan N, Gupta A, Levine S and Finn C 2021 Adaptive risk minimization: learning to adapt to domain shift *Advances in Neural Information Processing Systems* vol 34 pp 23664–78
- [64] Arjovsky M, Bottou L, Gulrajani I and Lopez-Paz D 2019 Invariant risk minimization (arXiv:1907.02893)
- [65] Krueger D, Caballero E, Jacobsen J-H, Zhang A, Binas J, Zhang D, Le Priol R and Courville A 2021 Out-of-distribution generalization via risk extrapolation (rex) *Int. Conf. on Machine Learning* (PMLR) pp 5815–26
- [66] Qu L, Zhou Y, Liang P P, Xia Y, Wang F, Adeli E, Fei-Fei L and Rubin D 2022 Rethinking architecture design for tackling data heterogeneity in federated learning *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 10061–71
- [67] Vepakomma P, Gupta O, Swedish T and Raskar R 2018 Split learning for health: distributed deep learning without sharing raw patient data (arXiv:1812.00564)
- [68] Gupta S et al 2021 Addressing catastrophic forgetting for medical domain expansion (arXiv:2103.13511)
- [69] Sheller M J et al 2020 Federated learning in medicine: facilitating multi-institutional collaborations without sharing patient data *Sci. Rep.* **10** 1–12
- [70] Hsieh K, Phanishayee A, Mutlu O and Gibbons P 2020 The non-iid data quagmire of decentralized machine learning *Int. Conf. on Machine Learning* (PMLR) pp 4387–98
- [71] Zhang M, Qu L, Singh P, Kalpathy-Cramer J and Rubin D L 2022 Splitavg: a heterogeneity-aware federated deep learning method for medical imaging *IEEE J. Biomed. Health Inform.* **26** 4635–44
- [72] Zhuang W, Gan X, Wen Y, Zhang S and Yi S 2021 Collaborative unsupervised visual representation learning from decentralized data *Proc. IEEE/CVF Int. Conf. on Computer Vision* pp 4912–21
- [73] Hsu T-M H, Qi H and Brown M 2019 Measuring the effects of non-identical data distribution for federated visual classification (arXiv:1909.06335)
- [74] Wang H, Yurochkin M, Sun Y, Papailiopoulos D and Khazaeni Y 2020 Federated learning with matched averaging (arXiv:2002.06440)
- [75] Zhang L, Luo Y, Bai Y, Du B and Duan L-Y 2021 Federated learning for non-iid data via unified feature learning and optimization objective alignment *Proc. IEEE/CVF Int. Conf. on Computer Vision* pp 4420–8
- [76] Qu L, Balachandar N, Zhang M and Rubin D 2022 Handling data heterogeneity with generative replay in collaborative learning for medical imaging *Med. Image Anal.* **78** 102424
- [77] Gong X, Sharma A, Karanam S, Wu Z, Chen T, Doermann D and Innanjan A 2021 Ensemble attention distillation for privacy-preserving federated learning *Proc. IEEE/CVF Int. Conf. on Computer Vision* pp 15076–86
- [78] Kirkpatrick J et al 2017 Overcoming catastrophic forgetting in neural networks *Proc. Natl Acad. Sci.* **114** 3521–6
- [79] Shin H, Lee J K, Kim J and Kim J 2017 Continual learning with deep generative replay *Advances in Neural Information Processing Systems* vol 30
- [80] Parmar N, Vaswani A, Uszkoreit J, Kaiser L, Shazeer N, Ku A and Tran D 2018 Image transformer *Int. Conf. on Machine Learning* (PMLR) pp 4055–64
- [81] Dosovitskiy A et al 2020 An image is worth 16x16 words: transformers for image recognition at scale (arXiv:2010.11929)
- [82] LeCun Y, Bottou L, Bengio Y and Haffner P 1998 Gradient-based learning applied to document recognition *Proc. IEEE* **86** 2278–324
- [83] Hochreiter S and Schmidhuber J 1997 Long short-term memory *Neural Comput.* **9** 1735–80
- [84] Lu K, Grover A, Abbeel P and Mordatch I 2021 Pretrained transformers as universal computation engines (arXiv:2103.05247)
- [85] Li D, Yang Y, Song Y-Z and Hospedales T M 2017 Deeper, broader and artier domain generalization *Proc. IEEE Int. Conf. on Computer Vision* pp 5542–50
- [86] Arivazhagan M G, Aggarwal V, Singh A K and Choudhary S 2019 Federated learning with personalization layers (arXiv:1912.00818)
- [87] Cheng A, Wang P, Zhang X S and Cheng J 2022 Differentially private federated learning with local regularization and sparsification *Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition* pp 10122–31
- [88] Tan M and Le Q 2019 Efficientnet: rethinking model scaling for convolutional neural networks *Int. Conf. on Machine Learning* (PMLR) pp 6105–14
- [89] Yang Y, Wang H and Katabi D 2022 On multi-domain long-tailed recognition, imbalanced domain generalization and beyond *Computer Vision—ECCV 2022: 17th European Conf. (Proc., Part XX) (Tel Aviv, Israel, October 23–27 2022)* (Springer) pp 57–75
- [90] LeCun Y, Cortes C and Burges C J 2009 The mnist database of handwritten digits (2010) (available at: <http://yann.lecun.com/exdb/mnist>)
- [91] Netzer Y, Wang T, Coates A, Bissacco A, Wu B and Ng A 2011 Reading digits in natural images with unsupervised feature learning *Neural Information Processing Systems*
- [92] Hull J 1994 A database for handwritten text recognition research *IEEE Trans. Pattern Anal. Mach. Intell.* **16** 550–4
- [93] Guo P et al 2022 Auto-fedrl: federated hyperparameter optimization for multi-institutional medical image segmentation *Computer Vision—ECCV 2022: 17th European Conf. (Proc., Part XXI) (Tel Aviv, Israel, 23–27 October 2022)* (Springer) pp 437–55
- [94] Kingma D P and Ba J 2014 Adam: a method for stochastic optimization (arXiv:1412.6980)
- [95] Zhao Y, Li M, Lai L, Suda N, Civin D and Chandra V 2018 Federated learning with non-iid data (arXiv:1806.00582)