FedCIAL: Federated Color-Invariant Adversarial Learning for Enhancing Fairness and Performance in Skin Lesion Classification

Rahmat Izwan Heroza John. Q. Gan Haider Raza
School of Computer Science and Electronic Engineering, University of Essex
Colchester CO4 3SQ, United Kingdom
{rh22708,jqgan,h.raza}@essex.ac.uk

Abstract

Federated learning (FL) enables healthcare institutions to collaboratively train a global model without sharing patient data. However, this approach can still introduce bias into the learning process. In the context of skin lesion classification, bias is a significant challenge due to the diversity in skin tone representation. Models trained on imbalanced datasets that underrepresent darker skin tones may exhibit reduced accuracy and reliability for those population. In this work, we propose Federated Color-Invariant Adversarial Learning (FedCIAL), a novel approach that leverages known color distribution shifts to generate target samples. This allows us to train a color-invariant feature extractor using domain adaptation techniques without requiring any shared data. Experimental results on the Fitzpatrick17k dataset show that FedCIAL outperforms the state-of-the-art model FeSViBS, achieving an average accuracy of 0.7754, compared to 0.7666 for the baseline, with a statistically significant improvement (p = 0.044). Additionally, FedCIAL improves model fairness, reducing the standard deviation across clients to 0.044, compared to 0.053 for the baseline. These findings demonstrate that FedCIAL enhances performance and offers a promising solution for fairer federated learning models in medical image analysis.

1. Introduction

Federated learning (FL) is a decentralized machine learning paradigm, where multiple clients (e.g., edge devices, hospitals, mobile phones) collaboratively train a shared model without centralizing their data [15, 17, 21]. Each client trains the local model locally on its own data and only the model updates or gradients are communicated to a central server that aggregates the local models to form a global model. This framework not only preserves data privacy, but also leverages heterogeneous distributed data sources, making it particularly useful in scenarios where data collection is sensitive or restricted by regulatory policies.

FL has emerged as a transformative approach in healthcare, enabling collaborative machine learning across institutions without sharing sensitive patient data [26, 27, 29]. By decentralizing the training process, FL preserves privacy while allowing models to learn from diverse datasets, addressing critical challenges in medical AI development. This atribute of FL also helps comply with strict regulations like GDPR and HIPAA [22, 25].

In FL, data on each client is often non-independent and identically distributed (non-IID). This means that each client's data can have different underlying distributions due to variations in user demographics, sensor characteristics, or environmental conditions [12, 16, 17]. A significant challenge arises from a distribution shift, where the statistical properties of the data vary across clients. For example, when considering bias variables such as Fitzpatrick skin type, certain clients might predominantly have data from specific skin tones, which can lead to models that perform unevenly across different groups [8]. Addressing such distribution shifts is critical to ensuring fairness and robust model performance across all client populations.

An invariant feature extractor is designed to learn representations that are robust to variations or biases in the data [7, 20, 23], which is introduced by non-IID distributions. The goal is to capture features that are relevant for the main prediction task while ignoring domain-specific characteristics that could lead to bias. In the context of this work, the feature extractor is trained to be invariant to differences in Fitzpatrick skin type. This is achieved by incorporating a domain branch along with a gradient reversal layer [6]. The gradient reversal layer enables adversarial training. While the main network minimizes the loss for the primary task, the domain branch is simultaneously trained to predict the skin type. By reversing the gradient during backpropagation, the feature extractor is forced to discard skin type specific features, resulting in representations that are more domain-invariant and unbiased.

In producing bias-invariant representation, the Domain-

Adversarial Neural Networks (DANN) use gradient reversal layers to achieve domain invariance [6]. However, DANN was originally designed for centralized learning, where all data is available on a single server, and does not inherently account for the decentralized and privacy-preserving nature of federated learning. CIRCLe [24] leverages color-invariant representation learning to extract features from skin lesion images that are robust against color variations, thus reducing bias from heterogeneous imaging conditions. However, it assumes centralized access to skin-type labels across all training data to train the skin color transformer, which makes it less suitable for federated settings, where data is decentralized, heterogeneous, and subject to strict privacy constraints.

In this study, we focus on the task of skin lesion classification, as skin cancer is among the most prevalent forms of cancer and requires precise early detection to significantly enhance patient survival rates [1, 3, 9]. We leverage the known color distribution shifts in Fitzpatrick skin types to inform the training of a color-invariant representation, rather than relying on a skin color transformer such as CIRCLe. This method directly uses explicit domain knowledge of color variations to guide a domain-adversarial training process, ensuring that the learned features remain invariant to biases introduced by color differences across decentralized data. By integrating this strategy within a federated learning framework, we address both privacy concerns and non-IID challenges inherent to decentralized environments. The contribution lies in mitigating bias in skin lesion classification by aligning feature representations using known color distribution shifts, thereby fostering fairer and more robust models in real-world, heterogeneous settings.

2. Related Work

2.1. Feature Skew in Federated Learning

One significant form of non-IIDness in federated learning is feature skew, where the distribution of input features varies between clients. Feature skew is a case of the covariate shift [18], which occurs when the input feature distribution changes but the conditional distribution of the output given the features remains the same [30, 31]. This can lead to higher generalization error, reducing the performance of models across different environments [10].

Liu *et al*. [19] carried out comprehensive experiments to develop a prostate segmentation model using heterogeneous MRI data collected from multiple sites. The variability in MRI acquisition protocols across these sites resulted in significant inter-site discrepancies. Similarly, Li *et al*. [18] trained a classification model on datasets from four medical institutions participating in ABIDE I, where functional brain imaging data were obtained using diverse imaging devices and scanning protocols. Another example in medical imaging arises in skin lesion classification task. Xu *et al*. [28] conducted experiments aimed at classifying skin lesion images sourced from clients with varying skin types. This skew can lead to models that perform well on locally seen data but generalize poorly to unseen global distributions [8]. This is particularly problematic because the feature space, such as skin tone, is closely related to the target output, leading to potential bias and poor transferability to other skin types. Addressing feature skew is crucial for equitable and robust learning in FL for healthcare, particularly in scenarios demanding fairness and generalization across diverse demographics.

2.2. Bias Mitigation in Machine Learning

Feature skew can lead to bias in the model. The mitigation of bias in machine learning, particularly color bias, is critical to ensuring fair and equitable outcomes among diverse populations. In skin lesion classification, color bias arises when machine learning models disproportionately rely on skin tone as a feature for prediction, leading to reduced generalization across individuals with varying skin types. This problem is exacerbated by the underrepresentation of darker skin tones in datasets, as highlighted in previous studies [4, 8], which show that biased datasets can result in significant performance disparities.

DANNs utilize gradient reversal layers to foster the learning of domain-invariant features, effectively minimizing the discrepancy between data distributions across domains [6]. By incorporating a domain classifier that predicts the domain labels from the learned features and then reversing the gradients during backpropagation, DANN encourages the feature extractor to produce representations that are agnostic to domain-specific information. This adversarial training mechanism has proven highly effective in centralized settings, where all data is available on a single server, allowing the model to align feature distributions across varied domains. However, this original design does not naturally address the challenges of decentralized, privacy-preserving environments inherent in federated learning, where data heterogeneity and restricted data sharing are primary concerns.

CIRCLe leverages color-invariant representation learning to extract features from skin lesion images that remain robust against variations in color, thereby reducing bias stemming from inconsistent imaging conditions [24]. This method relies on the availability of centralized skin-type labels to train a skin color transformer that normalizes the influence of color discrepancies, which can otherwise skew diagnostic accuracy. While this approach is effective in settings where a uniform and comprehensive color distribution is accessible, its reliance on centralized label availability makes it less adaptable to federated learning environments. In federated settings, data is inherently decentralized and subject to strict privacy constraints, which challenges the collection and utilization of consistent skin-type labels across all clients, thereby limiting the direct applicability of CIRCLe's methodology.

FeSViBS [2] integrates Federated Learning and Split Learning to enable collaborative learning while addressing privacy concerns and computational constraints. A key innovation in FeSViBS is its block sampling module, which optimally utilizes intermediate features extracted by a Vision Transformer (ViT) [5] at the server. Instead of relying solely on final-layer representations, FeSViBS strategically samples patch tokens from an intermediate transformer block. These selected features are then distilled into a pseudo class token, a compact representation that captures essential feature information, before being sent back to the client. By incorporating these pseudo class tokens, FeSViBS enhances feature augmentation and improves model generalizability across different client distributions. This approach mitigates the loss of information that may occur in federated settings due to data heterogeneity and limited client-side computations. Furthermore, the method ensures that critical learned representations are efficiently shared without exposing raw data, maintaining privacy while still benefiting from global knowledge aggregation. In this study, We use FeSViBS as the baseline model and enhance it with our proposed method for federated color-invariant adversarial learning.

3. Methodology

3.1. Preliminaries

3.1.1. Federated Learning

In federated learning, there are N clients $C = \{C_1, C_2, \dots, C_N\}$ and a central server. Each client C_k (where $k \in [1, N]$) trains a local model w_k on its private dataset D_k with a local objective function $\mathcal{L}_k$, defined as:

$$\mathcal{L}_k(w) = \frac{1}{|D_k|} \sum_{i \in D_k} \ell(w; x_i, y_i) \tag{1}$$

where $\ell(w; x_i, y_i)$ is the loss function (e.g., cross-entropy loss) evaluated on a data sample (x_i, y_i).

The server aggregates the local models to form a global model w by minimizing the weighted average of the client losses:

$$\min_w \mathcal{L}(w) = \sum_{k=1}^{N} \alpha_k \mathcal{L}_k(w) \tag{2}$$

where

$$\alpha_k = \frac{|D_k|}{\sum_{j=1}^{N} |D_j|} \tag{3}$$

This formulation enables collaborative training across clients while preserving data privacy, as raw data remains on local devices.

3.1.2. Feature Skew in Federated Learning

One significant challenge in FL is the presence of data heterogeneity across clients, particularly feature skew. Feature skew occurs when the distribution of input features varies across clients. This often occurs due to differences in data acquisition methods, imaging devices, or population demographics.

Formally, given two clients C_i and C_j, feature skew implies that their data distributions satisfy $P_i(x) \neq P_j(x)$, even if $P_i(y|x) \approx P_j(y|x)$. This mismatch in feature distributions can degrade the performance of the aggregated global model, as it may not generalize well across domains with differing feature characteristics.

3.2. Federated Color-Invariant Adversarial Learning

Federated Color-Invariant Adversarial Learning (FedCIAL) proposed in this paper is a novel method for invariant feature extraction in a federated learning environment by leveraging known color distribution shifts associated with Fitzpatrick skin types. Instead of employing a skin color transformer as used in previous approaches such as CIRCLe [24], our approach integrates explicit domain knowledge of color variations directly into the training process. We incorporate a gradient reversal layer, a core component of DANN [6], to construct an adversarial training framework. This layer inverts gradients from a domain classifier that attempts to predict Fitzpatrick skin type, thereby encouraging the feature extractor to learn representations that are robust against color-induced biases. This integration not only mitigates bias in skin lesion classification but also aligns feature distributions across diverse data sources without requiring centralized access to consistent skin type labels. The overview of our proposed method is shown in Figure 1 and the pseudocode for the main training loop and the color-invariant adversarial learning procedure are presented in Algorithm 1 and Algorithm 2, respectively.

Algorithm 1 Main FL Training Loop

```
Require: Number of clients N, communication rounds T
Initialize global model weights w_0
for each round t = 1, 2, ..., T do
    Server selects a subset of clients S_t ⊆ {1, ..., N}
    for each client k ∈ S_t in parallel do
        Client receives global model w_t
        Client performs Color-Invariant Adversarial Learning local updates with w_t
        Send updated model w^k_{t+1} to the server
    end for
    Server aggregates models:
```

$$w_{t+1} = \sum_{k \in \mathcal{S}_t} \frac{n_k}{\sum_{j \in \mathcal{S}_t} n_j} w^k_{t+1}$$

```
end for
```

Algorithm 2 Color-Invariant Adversarial Learning

```
Require: Model weights w_t, local epochs E, learning rate η, gradient reversal coefficient λ
for each local epoch e = 1, ..., E do
    for each mini-batch {x_i, y_i, d_i} from local dataset D_k do
        Sample a random target domain label d'_i ≠ d_i
        Generate transformed input x'_i = Transform(x_i, d'_i)
        Extract features: f_i = F(x_i), f'_i = F(x'_i)
        Predict class: ŷ_i = C(f_i)
        Predict domains: d̂_i = D(f_i), d̂'_i = D(f'_i)
        Compute losses:
```

$$\mathcal{L}_{\text{cls}} = \ell(\hat{y}_i, y_i)$$
$$\mathcal{L}_{\text{dom}} = \ell(\hat{d}_i, d_i) + \ell(\hat{d}'_i, d'_i)$$

```
        Compute total loss:
```

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{cls}} + \mathcal{L}_{\text{dom}}$$

```
        Apply standard backpropagation for L_cls from classifier C to feature extractor F
        Apply backpropagation for L_dom from domain classifier D to F via gradient reversal layer (GRL), where GRL multiplies the gradient by −λ
    end for
end for
```

3.2.1. Color-Invariant Representation Learning

Color-invariant representation learning aims to extract features that remain consistent regardless of variations in color information inherent in imaging data. In the context of skin lesion analysis, these variations often arise from differences in imaging conditions, device characteristics, and intrinsic properties of the skin, such as Fitzpatrick skin type.

The central idea of our approach is to leverage known color distribution shifts, specifically those associated with Fitzpatrick skin types, to guide the training process. Unlike many abstract biases, color variations are measurable and can be quantified, providing a concrete basis for correcting bias. By explicitly incorporating these known shifts into the training process, the model is forced to learn features that are robust to these systematic variations. This not only minimizes the risk of misclassification due to color differences but also aligns the feature representations across data from different sources, enhancing the overall fairness and accuracy of the model in heterogeneous settings.

Methods such as CIRCLe employ skin color transformers to normalize color discrepancies by using centralized access to skin-type labels for effective adjustment of the color space [24]. While effective in centralized environments, this approach is less suitable for federated learning where data is decentralized and strict privacy constraints limit the availability of uniform skin-type labels across clients. In contrast, our methodology directly integrates the knowledge of color distribution shifts into the feature extraction process using an adversarial training framework, thereby enabling decentralized learning of invariant features without relying on centralized label aggregation. This not only circumvents privacy issues but also ensures a more robust adaptation to the diverse color distributions encountered in real-world federated datasets.

The input RGB image is first converted to the Lab color space to separate luminance and chromatic components. The Lab color space is better suited for representing how colors are perceived by the human eye, as it more accurately aligns with human visual perception compared to standard RGB models [11]. A transformation function is then applied to shift the image towards the target Fitzpatrick Lab values. Predefined Lab color values for each Fitzpatrick skin type is presented in Table 1. The Lab color values assigned to each Fitzpatrick skin type are empirically derived based on their expected ITA values [8], ensuring a consistent mapping between skin pigmentation levels and color space representation. This approach follows the established relationship between ITA and Lab components, where lighter skin types exhibit higher L values and darker skin types show lower ITA with increased b values.

3.2.2. Integration of Domain-Adversarial Techniques

DANNs provide a robust framework for learning representations that are invariant to domain-specific variations, making them especially useful in decentralized environments. A key element in DANN is the gradient reversal layer, which is situated between the feature extractor and the domain classifier. During forward propagation, this layer func-

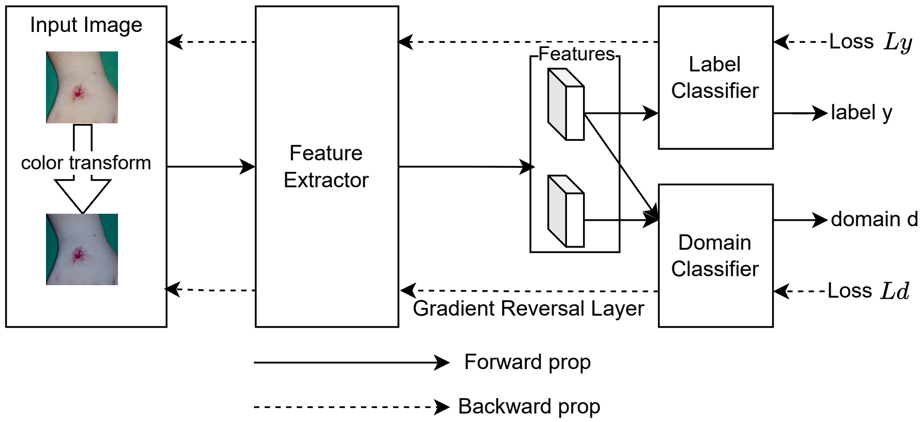

Figure 1. Method Overview

Table 1. Predefined Lab color values for each Fitzpatrick skin type (FST).

FST	**L**	**a**	**b**
1	85	4	12
2	78	6	18
3	69	9	25
4	62	14	33
5	55	19	42
6	48	23	51

tions as an identity operation, transmitting the extracted features without modification. However, during backpropagation, it multiplies the gradients coming from the domain classifier by a negative constant, effectively inverting the gradient signal. This inversion changes the optimization objective: while the domain classifier is trained to accurately predict domain labels, the feature extractor is simultaneously driven to produce representations that obscure domain-specific cues. The result is a model that learns robust and invariant features, a property that is crucial for ensuring that a global model generalizes well across the heterogeneous and decentralized datasets, typical in federated learning.

In addition to the adversarial adaptation achieved via the gradient reversal layer, our approach further enhances invariance by incorporating transformed images as target samples during training. We leverage known color distribution shifts associated with Fitzpatrick skin types to generate transformed versions of the input images, aligning them with a predefined target color profile. Because in the original DANN framework shared target samples are used, we instead utilize these transformed images as the target samples. These transformed images serve as explicit targets that guide the feature extractor to not only become invariant to domain-specific color variations but also to normalize representations against a consistent color distribution.

Forward propagation follows the conventional approach, where both the original and transformed images are passed through a feature extractor to generate feature representations. The label classifier minimizes the label prediction loss using features derived from the original images, treated as source samples, while the domain classifier aims to minimize the domain prediction loss based on features from both original and transformed images. During backpropagation, a gradient reversal layer connected to the feature extractor multiplies the gradients from the domain classification task by a negative constant. This encourages the feature extractor to learn domain-invariant features by effectively reversing the domain-related gradient signals.

To control the strength of domain adversarial learning over the course of training, we used a sigmoid-based scheduling strategy for the domain adaptation coefficient λ, similar to the approach in DANN. The coefficient is updated using the following function:

$$\lambda = \frac{2}{1 + \exp\left(-10 \times \frac{r}{\text{epochs}}\right)} - 1$$

where r represents the current training round and epochs is the total number of training rounds. This scheduling strategy allows λ to start near 0, minimizing the effect of domain adversarial loss in the early stages of training. As training progresses, λ gradually increases toward 1, ensuring stronger domain adaptation and encouraging the feature extractor to learn domain-invariant representations.

To create a binary domain classification similar to traditional domain adaptation, we grouped Fitzpatrick skin types into two broad categories based on their melanin levels and visual appearance. Specifically, Fitzpatrick skin type 1, 2, and 3, which correspond to lighter skin tones, were mapped to domain 0, while Fitzpatrick skin type 4, 5, and 6, representing darker skin tones, were mapped to domain 1. This binary domain division allows the model to learn invariant features by distinguishing between two broad groups rather than handling each Fitzpatrick type separately. By structuring the domain in this manner, we ensure that the adversarial domain adaptation process can effectively reduce the influence of skin color on feature extraction while maintaining strong classification performance across different skin tones.

4. Experiments

4.1. Experimental Setup

4.1.1. Dataset

The Fitzpatrick17k [8] dataset consists of 16,500 clinical images annotated with six Fitzpatrick skin types, making it a valuable resource for studying the impact of skin tone on dermatological diagnosis. In addition to skin type annotations, the dataset encompasses a wide range of lesion characteristics, ensuring diversity in both skin conditions and imaging variations. At the highest level, skin conditions in the dataset are categorized into benign, malignant, and non-neoplastic classes.

For this experiment, we follow Xu *et al.* [28] that partition the dataset into six clients, where each client exclusively contains images corresponding to a single Fitzpatrick skin type. This partitioning strategy simulates the non-IID (non-independent and identically distributed) nature of real-world federated learning scenarios, where medical institutions may serve populations with differing skin type distributions.

Prior to model training, we apply several preprocessing steps to standardize image inputs and improve model robustness. These include geometric transformations such as rotation, flipping, and shearing to enhance data diversity, resizing images to a consistent input resolution, and normalization to standardize pixel intensity distributions across images using ImageNet statistics [14].

4.1.2. Baseline

We utilize the Federated Split Learning of Vision Transformer with Block Sampling (FeSViBS) [2] as the model architecture. FeSViBS is designed to leverage the advantages of federated split learning and Vision Transformers (ViTs) in handling large-scale, distributed image data.

We used FeSViBS as the baseline model for our experiments. To evaluate the effectiveness of our approach, we integrated FedCIAL into FeSViBS by training the feature extractor to produce color-invariant representations using transformed images as target samples. Additionally, we introduced two comparative methods: (1) enhancing the baseline by training the feature extractor with a regularization loss applied to color-augmented samples, and (2) integrating the baseline with adversarial learning but without utilizing target samples.

Table 2. Average balance accuracy and standard deviation.

Method	Bal. Acc.
Baseline	0.7666 +- 0.0054
Color-invariant rep. learning	0.7682 +- 0.0122
Adversarial learning wo. target samples	0.7686 +- 0.0056
FedCIAL	**0.7754 +- 0.0098**

4.1.3. Implementation Details

All methods are implemented using PyTorch. We employ the cross-entropy loss function for both the lesion classification and domain classification objectives. We set the batch size to 32. The Adam [13] optimizer is used for optimization, with a learning rate of 0.0001. Training is conducted over 100 communication rounds, where each round consists of one epoch of local training per client. To ensure valid and robust results, we conducted each experiment using 5-fold cross-validation, splitting the dataset into five folds for comprehensive evaluation.

4.2. Results

We report the average performance across 5-fold cross-validation in Table 2. Specifically, FedCIAL achieved an average performance of 0.7754, compared to the baseline score of 0.7666, demonstrating a clear performance gain. To assess the statistical significance of the improvement between FedCIAL and the baseline, we conducted a paired t-test. The Shapiro-Wilk test yielded a p-value of 0.522, indicating that the data follows a normal distribution, thus validating the use of a paired t-test. The resulting t-test p-value of 0.044 confirms that our proposed method significantly improves upon the baseline.

Another benchmark for this dataset is the work of Groh *et al.* [8], which achieved an accuracy of 62.4% in a centralized setting using a randomly held-out test set. In contrast, FedCIAL, designed for a federated learning environment, surpasses this baseline while operating under decentralized and privacy-preserving constraints, highlighting its effectiveness in improving classification performance despite the challenges of distributed data.

FedCIAL also enhances model fairness, as evidenced by a reduction in the standard deviation across clients. The average standard deviation for FedCIAL is 0.044, compared to 0.053 for the baseline, indicating more consistent performance across diverse clients.

To demonstrate the effectiveness of our method, we mea-

Figure 2. Original image, transformed and anchor image

Table 3. Distance between original image, transformed and anchor image.

Metric	a	b	c
Baseline			
Manhattan (L1)	24.7	27.7	11.8
Euclidean (L2)	1.1	1.3	0.5
Cosine Distance	0.11	0.14	0.03
FedCIAL			
Manhattan (L1)	22.3	24.6	9.8
Euclidean (L2)	1.0	1.1	0.4
Cosine Distance	0.09	0.11	0.01

sure the distance between an original image, a transformed image, and an anchor image for comparison as in 2. The distances between these images are presented in Table 3. The results demonstrate that FedCIAL effectively reduces the distance between the original image and the color-transformed image across multiple distance metrics: Manhattan (L1), Euclidean (L2), and Cosine distance. In each case, the distance between the original image and the color-transformed image (denoted as c) decreases when applying FedCIAL, suggesting that the method successfully brings the feature representations of the transformed images closer to those of the original images. The reduction in c is highlighting FedCIAL's ability to address color-related discrepancies and improve feature alignment between the original and transformed images. This improvement across different distance metrics shows that FedCIAL enhances the model's ability to maintain consistent feature representations, which is essential for ensuring fairness and robustness in skin lesion classification, especially across diverse skin types.

5. Conclusion

This paper proposes FedCIAL to address bias in skin lesion classification within federated learning. By leveraging color-invariant representation learning based on known color distribution shifts, FedCIAL improves both model accuracy and fairness across clients. Experiments on the Fitzpatrick17k dataset demonstrates that FedCIAL significantly outperforms the state-of-the-art method FesViBS in terms of improved average accuracy and reduced standard deviation between clients. These results underscore the potential of FedCIAL to reduce skin tone bias and ensure more equitable and robust models in federated medical imaging applications, providing a promising direction for future research in fair AI in healthcare.

Acknowledgement

We thank the Brain-Computer Interfaces and Neural Engineering Lab at the University of Essex for providing the GPU server and professional technical support.

References

[1] Abdulmateen Adebiyi, Nader Abdalnabi, Eduardo J Simoes, Mirna Becevic Phd, Emily Hoffman, Smith Md, and Praveen Rao. Transformers in skin lesion classification and diagnosis: A systematic review. *medRxiv*, page 2024.09.19.24314004, 2024. 2

[2] Faris Almalik, Naif Alkhunaizi, Ibrahim Almakky, and Karthik Nandakumar. Fesvibs: Federated split learning of vision transformer with block sampling. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 14221 LNCS:350–360, 2023. 3, 6

[3] Muhammad Azeem, Kaveh Kiani, Taha Mansouri, and Nathan Topping. Skinlesnet: Classification of skin lesions and detection of melanoma cancer using a novel multi-layer deep convolutional neural network. *Cancers*, 16:108, 2023. 2

[4] Marin Benčević, Marija Habijan, Irena Galić, Danilo Babin, and Aleksandra Pižurica. Understanding skin color bias in deep learning-based skin lesion segmentation. *Computer Methods and Programs in Biomedicine*, 245:108044, 2024. 2

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 3

[6] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, Victor Lempitsky, Urun Dogan, Marius Kloft, Francesco Orabona, Tatiana Tommasi, and al Ganin.

Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17:1–35, 2016. 1, 2, 3

[7] Chenzhong Gao and Wei Li. An invariant feature extraction for multi-modal images matching. 2023. 1

[8] Matthew Groh, Caleb Harris, Luis Soenksen, Felix Lau, Rachel Han, Aerin Kim, Arash Koochek, and Omar Badri. Evaluating deep neural networks trained on clinical images in dermatology with the fitzpatrick 17k dataset. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 1820–1828, 2021. 1, 2, 4, 6

[9] Rahmat Izwan Heroza, John Q. Gan, and Haider Raza. Enhancing skin lesion classification: A self-attention fusion approach with vision transformer. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 14860 LNCS: 309–322, 2024. 2

[10] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *32nd International Conference on Machine Learning, ICML 2015*, 1:448–456, 2015. 2

[11] Geunho Jung, Semin Kim, and Sangwook Yoo. Skin tone analysis through skin tone map generation with optical approach and deep learning. *Skin Research and Technology*, 30:e70088, 2024. 4

[12] Sai Praneeth Karimireddy, Satyen Kale, Mehryar Mohri, Sashank J. Reddi, Sebastian U. Stich, and Ananda Theertha Suresh. Scaffold: Stochastic controlled averaging for federated learning. *37th International Conference on Machine Learning, ICML 2020*, PartF168147-7:5088–5099, 2019. 1

[13] Diederik P. Kingma and Jimmy Lei Ba. Adam: A method for stochastic optimization. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 2014. 6

[14] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. *Communications of the ACM*, 60:84–90, 2017. 6

[15] Qinbin Li, Zeyi Wen, Zhaomin Wu, Sixu Hu, Naibo Wang, Yuan Li, Xu Liu, and Bingsheng He. A survey on federated learning systems: Vision, hype and reality for data privacy and protection. *IEEE Transactions on Knowledge and Data Engineering*, 35:3347–3366, 2019. 1

[16] Tian Li, Anit Kumar Sahu, Manzil Zaheer, Maziar Sanjabi, Ameet Talwalkar, and Virginia Smith. Federated optimization in heterogeneous networks. 2018. 1

[17] Tian Li, Anit Kumar Sahu, Ameet Talwalkar, and Virginia Smith. Federated learning: Challenges, methods, and future directions. *IEEE Signal Processing Magazine*, 37:50–60, 2019. 1

[18] Xiaoxiao Li, Meirui Jiang, Xiaofei Zhang, Michael Kamp, and Qi Dou. Fedbn: Federated learning on non-iid features via local batch normalization. *ICLR 2021 - 9th International Conference on Learning Representations*, 2021. 2

[19] Quande Liu, Qi Dou, Lequan Yu, and Pheng Ann Heng. Ms-net: Multi-site network for improving prostate segmentation with heterogeneous mri data. *IEEE Transactions on Medical Imaging*, 39:2713–2724, 2020. 2

[20] Wang Lu, Jindong Wang, Haoliang Li, Yiqiang Chen, and Xing Xie. Domain-invariant feature exploration for domain generalization. 2022. 1

[21] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Aguera y Arcas. Communication-efficient learning of deep networks from decentralized data, 2017. 1

[22] Karthik Meduri, Geeta Sandeep Nadella, Akhila Reddy Yadulla, Vinay Kumar Kasula, Mohan Harish Maturi, Steven Brown, Snehal Satish, and Hari Gonaygunta. Leveraging federated learning for privacy-preserving analysis of multi-institutional electronic health records in rare disease research. *Journal of Economy and Technology*, 3:177–189, 2025. 1

[23] Shaohui Mei, Ruoqiao Jiang, Jingyu Ji, Jun Sun, Yang Peng, and Yifan Zhang. Invariant feature extraction for image classification via multi-channel convolutional neural network. *2017 International Symposium on Intelligent Signal Processing and Communication Systems, ISPACS 2017 - Proceedings*, 2018-January:491–495, 2017. 1

[24] Arezou Pakzad, Kumar Abhishek, and Ghassan Hamarneh. Circle: Color invariant representation learning for unbiased classification of skin lesions. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 13804 LNCS: 203–219, 2023. 2, 3, 4

[25] Sarthak Pati, Sourav Kumar, Amokh Varma, Brandon Edwards, Charles Lu, Liangqiong Qu, Justin J. Wang, Anantharaman Lakshminarayanan, Shih han Wang, Micah J. Sheller, Ken Chang, Praveer Singh, Daniel L. Rubin, Jayashree Kalpathy-Cramer, and Spyridon Bakas. Privacy preservation for federated learning in health care. *Patterns*, 5:100974, 2024. 1

[26] Nicola Rieke, Jonny Hancox, Wenqi Li, Fausto Milletarì, Holger R. Roth, Shadi Albarqouni, Spyridon Bakas, Mathieu N. Galtier, Bennett A. Landman, Klaus Maier-Hein, Sébastien Ourselin, Micah Sheller, Ronald M. Summers, Andrew Trask, Daguang Xu, Maximilian Baust, and M. Jorge Cardoso. The future of digital health with federated learning. *npj Digital Medicine 2020 3:1*, 3:1–7, 2020. 1

[27] Zhen Ling Teo, Liyuan Jin, Siqi Li, Di Miao, Xiaoman Zhang, Wei Yan Ng, Ting Fang Tan, Deborah Meixuan Lee, Kai Jie Chua, John Heng, Yong Liu, Rick Siow Mong Goh, and Daniel Shu Wei Ting. Federated machine learning in healthcare: A systematic review on clinical applications and technical architecture. *Cell Reports Medicine*, 5:101419, 2024. 1

[28] Gelei Xu, Yawen Wu, Jingtong Hu, and Yiyu Shi. Achieving fairness in dermatological disease diagnosis through automatic weight adjusting federated learning and personalization. 2022. 2, 6

[29] Jie Xu, Benjamin S. Glicksberg, Chang Su, Peter Walker, Jiang Bian, and Fei Wang. Federated learning for healthcare informatics. *Journal of Healthcare Informatics Research*, 5: 1–19, 2021. 1

[30] Yunlu Yan, Huazhu Fu, Yuexiang Li, Jinheng Xie, Jun Ma, Guang Yang, and Lei Zhu. A simple data augmentation for feature distribution skewed federated learning. 2023. 2

[31] Tianfei Zhou and Ender Konukoglu. Fedfa: Federated feature augmentation. *11th International Conference on Learning Representations, ICLR 2023*, 2023. 2