IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 27 August 2025, accepted 16 September 2025, date of publication 22 September 2025,
date of current version 26 September 2025.

Digital Object Identifier 10.1109/ACCESS.2025.3612477

RESEARCH ARTICLE

Federated ViT: A Distributed Deep Learning Framework for Skin Cancer Classification

RIAZ H. JUNEJO[1], QAISAR ABBAS[2], (Member, IEEE), MUHAMMAD AWAIS[3], (Senior Member, IEEE), TALLHA AKRAM[4], AND MUTLAQ B. ALDAJANI[2]

[1]Department of Computer Engineering, COMSATS University Islamabad, Wah Campus, Wah 47040, Pakistan
[2]College of Computer and Information Sciences, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 11432, Saudi Arabia
[3]Department of Computer Science, College of Computer, Qassim University, Buraydah, Qassim 52571, Saudi Arabia
[4]Department of Information Systems, College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, Al-Kharj 11942, Saudi Arabia

Corresponding author: Muhammad Awais (mu.aslam@qu.edu.sa)

This work was supported and funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) (grant number IMSIU-DDRSP2501).

ABSTRACT Skin cancer remains one of the most prevalent and life-threatening diseases globally, where early and accurate diagnosis is crucial for effective treatment and improved patient outcomes. Although deep learning and digital pathology have advanced lesion classification performance, significant challenges remain due to complex histopathological patterns, privacy concerns over centralized data storage, and inconsistencies across datasets in terms of annotation, quality, and structure. These disparities hinder model generalization and complicate performance benchmarking. To overcome these issues, this study proposes a novel skin lesion classification framework that integrates Federated Learning (FL) with Vision Transformers (ViTs). FL enables decentralized model training across multiple clinical sites without sharing sensitive data, inherently accommodating data heterogeneity while preserving privacy. ViTs, with their self-attention mechanisms, effectively capture global contextual features, making them robust to structural and visual variations in non-standardized datasets. The framework incorporates key components such as client-side preprocessing, data augmentation, federated transfer learning, model aggregation, and feature optimization. Comprehensive experiments were conducted on two publicly available skin lesion datasets—HAM10000 and ISIC2019. The proposed model achieved a global test accuracy of 90%, with a sensitivity of 88.2% and specificity of 91.4% on the HAM10000 dataset. On the ISIC2019 dataset, the model attained a test accuracy of 87.6%, with Area Under the Curve (AUC) values reaching 0.96 for the best-performing class, demonstrating strong discriminative ability. These results confirm the framework's robustness in real-world scenarios with non indentical data and its potential to support privacy-preserving, scalable clinical applications in dermatology. This work contributes to advancing FL-based computer vision solutions for skin cancer detection and lays the groundwork for future exploration of ViTs in decentralized healthcare environments.

INDEX TERMS Deep learning, vision transformers, federated learning, computer vision, image classification.

I. INTRODUCTION

Skin cancer, a potentially fatal disease resulting from the uncontrolled proliferation of abnormal skin cells, is predominantly induced by prolonged exposure to ultraviolet (UV) radiation from sunlight or artificial sources. This exposure causes irreversible DNA mutations, leading to tumor formation [1]. According to the World Cancer Research Fund (WCRF), skin melanoma ranked as the 17th most common cancer globally in 2022, with an estimated 331,722 new cases, ranking 13th and 15th in incidence among men and women, respectively [2].

The associate editor coordinating the review of this manuscript and approving it for publication was Vincenzo Conti.

Skin cancer encompasses a spectrum of malignancies such as melanoma, actinic keratosis, basal and squamous cell carcinomas, Merkel cell carcinoma, and atypical moles [3], [4], [5], [6]. While benign lesions like moles display regulated growth, malignant counterparts exhibit uncontrolled proliferation and structural abnormalities [7]. Their visual similarity often complicates early diagnosis, necessitating a combination of visual inspection, dermoscopy, histopathology, and advanced imaging for accurate classification and staging [8], [9], [10], [11].

Recent advances in computer vision, particularly deep learning, have led to the development of automated frameworks for skin lesion classification using dermoscopic images [12]. However, training high-performance models requires large-scale datasets that are often distributed across institutions with varying data characteristics. Traditional machine learning typically relies on the assumption that data is independent and identically distributed (IID), meaning that each sample is statistically independent and drawn from the same underlying distribution. However, this assumption rarely holds true in real-world medical imaging scenarios. In federated learning (FL), the data is often non-IID—exhibiting variations across institutions in terms of class frequencies, imaging protocols, and even feature spaces such as image resolutions or staining methods. Additionally, samples from certain clients may be statistically dependent, as is common in specialized clinics where patient cases are inherently correlated. This non-IID nature introduces several challenges for FL: local models may experience client drift by overfitting to their institution-specific distributions; biased aggregation can occur when standard federated averaging disproportionately reflects dominant data sources; and slow convergence arises due to heterogeneous gradients across clients, often requiring more communication rounds to achieve global model alignment [13], [14]. Convolutional neural networks (CNNs) have shown remarkable efficacy in extracting hierarchical features from raw images [15], [16], while more recently, Vision Transformers (ViTs) have demonstrated superior performance by capturing global contextual relationships through self-attention mechanisms [17]. Unlike CNNs, which rely on local receptive fields, ViTs can model long-range dependencies and emphasize diagnostically relevant regions, enhancing interpretability and robustness to non-IID data variations.

The distributed nature of medical data across institutions, combined with strict privacy regulations, presents significant challenges for centralized model training. Federated Learning (FL) addresses these challenges by enabling decentralized model training across multiple sites without sharing raw data. When integrated with ViTs, FL not only preserves patient confidentiality but also leverages inter-institutional data variability, fostering the development of secure, generalizable diagnostic models. This fusion presents a promising direction for advancing automated, privacy-aware skin lesion classification in real-world clinical environments where data heterogeneity is the norm rather than the exception.

1) FEDERATED LEARNING

To cope with the challenges unaddressed by the traditional deep learning methods, specially the issue of data privacy, google developed the framework of Federated Learning in 2016 [18], [19].

Image classification using federated learning has revolutionized the traditional machine learning by enabling collaborative model training over decentralized devices, keeping data privacy intact. In this particular approach, a global model is initialized and distributed to participating clients, which then individually train the model over their local datasets. Only the model's updated weights, not the datasets, are exchanged with a central server for aggregation [20]. Through iterative training cycles, the global model learns from the collective knowledge acquired by all devices, while making sure that sensitive user data remains localized. Federated learning offers advantages such as data privacy, low communication costs, and scalability to large and distributed datasets. However, challenges including ensuring model convergence, addressing communication delays and failures, and managing security and privacy concerns must be carefully addressed. Nevertheless, with ongoing advancements, federated learning holds promise for collaborative and privacy-preserving image classification and other machine learning applications.

We can categorize FL on the basis of six factors: machine learning model, data partitioning, communication architecture, privacy method, size & motive of the federation

Figure 1 illustrates the primary classifications of federated learning (FL). One categorization is based on participant types: cross-device FL involves numerous devices with small local datasets, while cross-silo FL includes fewer entities (e.g., hospitals) with larger data volumes. Another classification is based on data partitioning: horizontal FL uses datasets with similar feature spaces but different samples (e.g., skin lesion images from different hospitals), whereas vertical FL shares sample IDs but varies in feature types (e.g., patient records from different departments). Federated transfer learning supports heterogeneous feature and sample spaces, offering greater flexibility without requiring strict alignment. FL can also be categorized by model type, including linear models, decision trees, and neural networks. Communication architectures are either centralized, where clients interact with a central server for aggregation, or decentralized, enabling peer-to-peer updates without a central coordinator. Regarding privacy, FL employs methods such as differential privacy, which adds controlled noise to updates, and cryptographic techniques like homomorphic encryption to secure model aggregation.

II. LITERATURE REVIEW

This section surveys recent advances in FL for medical image classification, focusing on architectures, optimization strategies, and privacy-preserving mechanisms. Adjei-Mensah et al. [21] proposed a Multi-Efficient Channel Attention Network (MECAN) for COVID-19 diagnosis

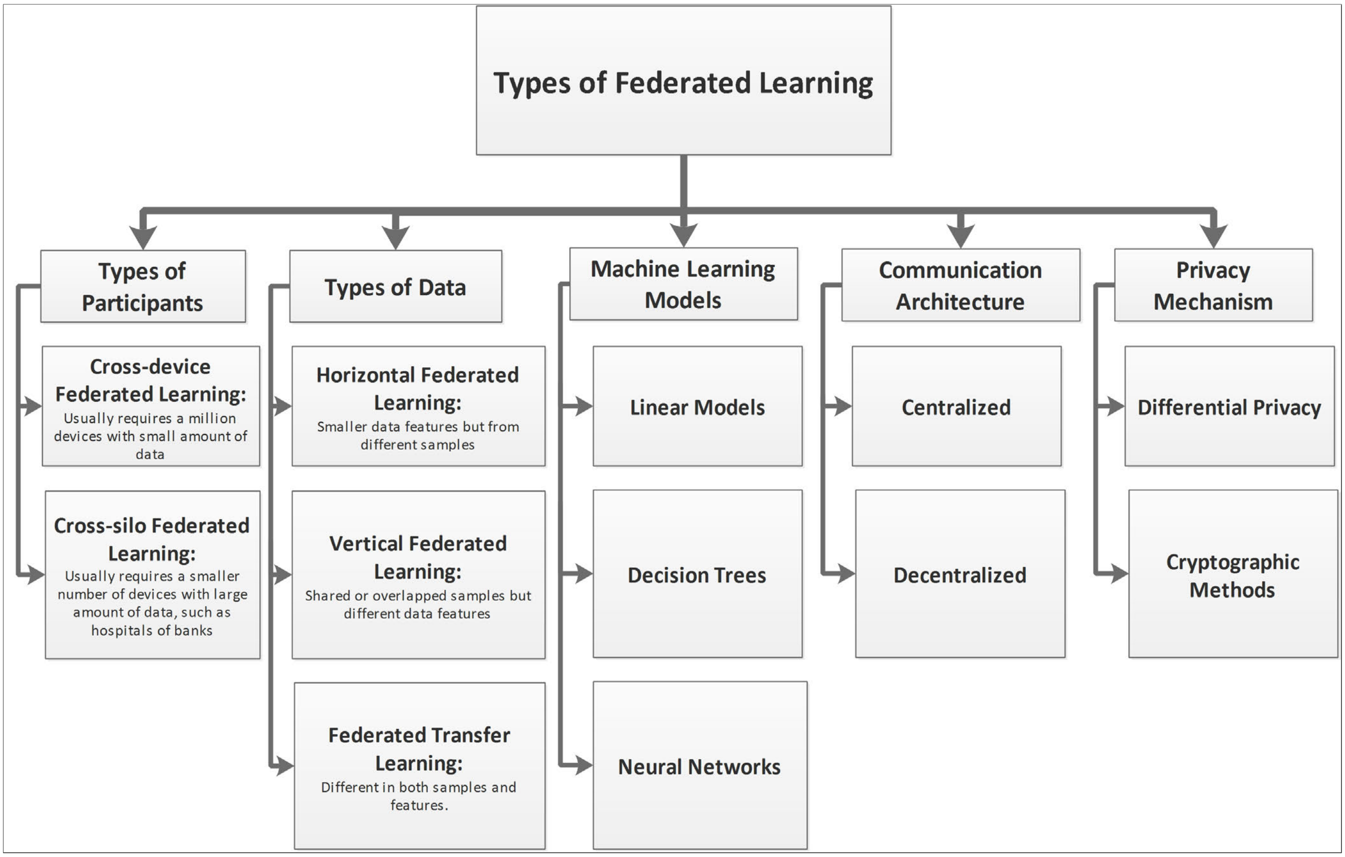

Types of Federated Learning
Types of Participants
Cross-device Federated Learning: Usually requires a million devices with small amount of data
Cross-silo Federated Learning: Usually requires a smaller number of devices with large amount of data, such as hospitals of banks
Types of Data
Horizontal Federated Learning: Smaller data features but from different samples
Vertical Federated Learning: Shared or overlapped samples but different data features
Federated Transfer Learning: Different in both samples and features.
Machine Learning Models
Linear Models
Decision Trees
Neural Networks
Communication Architecture
Centralized
Decentralized
Privacy Mechanism
Differential Privacy
Cryptographic Methods

FIGURE 1. Types of Federated Learning.

using X-ray images, achieving 87.7% accuracy and a ROC-AUC of 92.4%. Zhang et al. [22] introduced a doubly contrastive learning approach (PerFCL) that improves FL performance by jointly optimizing global and personalized representations.

Casella et al. [20] tested a federated model on COVID-19 prognosis and Alzheimer's stratification, reporting better generalization and performance compared to centralized and single-input models. Nilsson et al. [13] evaluated multiple FL algorithms on MNIST under both IID and non-IID settings, showing that FedAvg performs comparably to centralized models under IID conditions but falls short in non-IID cases. Deng et al. [23] proposed APFL, which adaptively balances local and global contributions using optimal mixing parameters, demonstrating improved personalization and generalization. Sharma and Guleria [24] conducted a comprehensive survey on FL-based disease detection, highlighting unresolved challenges in model convergence and personalization.

Lu et al. [25] introduced FedMedVLP, a federated vision-and-language pretraining framework, which achieved state-of-the-art performance across five tasks, including clinical image-text retrieval and visual query resolution. Similarly, Ma et al. [26] proposed FedAR, an adaptive reweighting method that balances client contributions by data volume and model performance, outperforming baseline FL methods. Kareem et al. [27] benchmarked popular CNN models like ResNet-50 under FL for real-time medical classification tasks, reporting high accuracy (93%) using AUC as the evaluation metric. Lusnig et al. [28] applied FL to histological liver biopsy data, achieving 90% accuracy.

Yang and Wang [29] improved tuberculosis CT-scan classification by 11.54% over FedAvg using a customized FL method. Woubie et al. [30] emphasized privacy-preserving federated training without secure aggregation on the CelebA dataset, showing minimal performance degradation. Park and Ko [31] introduced FedHM, an aggregation scheme supporting heterogeneous model architectures (e.g., ResNet variants) across FL frameworks, with strong performance on CIFAR-10 and FEMNIST. Beltrán et al. [32] presented FedStellar, a dynamic FL framework allowing real-time federation formation, which demonstrated robust F1 scores (up to 98%) under various training topologies and device settings.

Yan et al. [33] proposed FedEYE, a benchmarking framework evaluating CNN and transformer models (ResNet-45, ViT, Swin-Transformer) on both private and public datasets, but without ViT-specific optimization for FL.

In [34], the authors propose a fuzzy ensemble-based federated learning framework for EEG-based emotion recognition in the Internet of Medical Things. It integrates TCN, LSTM, and GRU models using a Gompertz function-based fuzzy rank approach and introduces an asynchronous dropout algorithm to enhance global model aggregation. Experiments on GAMEEMO, SEED, and DEAP datasets demonstrate that the proposed method consistently outperforms individual models and existing ensemble/FL baselines in terms of accuracy and F1 score. However, the framework has several shortcomings: it relies only on federated learning for privacy without employing stronger methods like differential privacy or encryption, assumes clients behave honestly (leaving it vulnerable to adversarial updates), and is sensitive to hyperparameter tuning. Tricomi et al. [35] propose HERALD, a hybrid IL-FL framework that effectively mitigates catastrophic forgetting in COVID-19 chest X-ray classification while preserving data privacy across institutions. Their work is noteworthy for its pragmatic integration of DILoCC with federated averaging, demonstrating strong empirical results on heterogeneous datasets. However, the framework's reliance on synchronous FL iterations and centralized aggregation may limit scalability in real-world settings with highly unbalanced or non-IID data distributions. Furthermore, while privacy is maintained by design, the study does not address potential model inversion or membership inference attacks—a significant concern in medical FL. Nonetheless, HERALD provides a valuable benchmark for privacy-preserving continual learning in clinical imaging and underscores the importance of adaptive aggregation in federated environments.

Table 1 summarizes the key aspects of published studies discussed above. The literature review indicates that most FL-based medical image classification methods predominantly utilize CNN architectures like ResNet and MobileNet. However, these CNN-driven approaches often perform poorly on non-IID medical data, with accuracy drops of up to 15%. This limitation stems from CNNs' reliance on local receptive fields, which are insufficient for capturing the diverse feature distributions found across institutions—an issue especially critical in skin lesion analysis, where global characteristics such as border irregularity and texture variation are key diagnostic cues.

In contrast, our work integrates ViT-B/16 into a federated framework, adapting its patch embedding and self-attention mechanisms to address non-IID data and communication efficiency. While FedEYE [33] touches on ViTs, it lacks FL-specific optimizations. Additionally, unlike previous studies [21], [22] that apply transfer learning to CNNs, we introduce a ViT-specific federated transfer learning pipeline that replaces the final classification layer to align with lesion classes while preserving ImageNet-pretrained knowledge—mitigating cold-start issues and improving convergence. Furthermore, existing aggregation strategies like FedAvg [33] and FedProx [14] often fail under data heterogeneity, and although APFL [23] and FedAR [26]introduce client-weighting, they remain CNN-focused. We enhance FedAvg with client-level weighting based on both data volume and stochastic participation (Eq. 8), achieving better global generalization across diverse datasets (HAM10000, ISIC2019). Unlike prior works, that typically report overall accuracy only, our evaluation includes detailed metrics—accuracy, sensitivity, specificity, F1-score, and per-class AUC—showing strong discriminative performance. Finally, while frameworks like FedMedVLP [25] and FedStellar [32] explore multi-modal or dynamic FL scenarios, they neither target ViTs nor dermatological use cases. Our proposed framework is tailored for decentralized dermatology, incorporating standardized preprocessing and consistent augmentation to support clinically viable, privacy-preserving deployment.

III. CONTRIBUTIONS

The primary contributions of this work are as follows:

1) A federated learning pipeline is proposed that trains ViTs on decentralized skin lesion datasets without transferring raw data, ensuring compliance with privacy regulations like HIPAA and GDPR.
2) An adaptation of ViTs is performed for FL settings. The core processing steps and parameters of ViT are meticolously tailed for distributed training.
3) This work proposes a robust aggregation strategy based on Federated Averaging (FedAvg) with client-level weighting to handle non-IID data distributions and improve global model convergence.
4) Rigorous testing on benchmark datasets (HAM10000 and ISIC2019), achieving competitive accuracy (up to 99% on client data and 90% globally) while highlighting challenges like class imbalance and overfitting.

IV. PROPOSED FRAMEWORK

As discussed earlier, medical image classification faces key challenges, particularly in data acquisition due to privacy restrictions, as datasets are typically patient- or institution-controlled, complicating centralized machine learning. To overcome these limitations, this work proposes an FL based framework incorporating ViTs, offering a decentralized and privacy-preserving approach to skin cancer classification. Figure 2 shows the pipeline of the proposed distributed learning framework. The setup consists of N decentralized clients $(C_1, C_2, \ldots, C_N)$ and a central aggregation server that coordinates the training process. Each client holds its own private dataset, which remains stored locally to maintain privacy and reduce the risk of data leakage.

In this framework, we adopt the ViT-B/16 as the base model architecture for image classification tasks. ViT processes input images by dividing them into patches and applying transformer-based self-attention mechanisms to capture global contextual relationships—making it highly suitable for complex visual tasks like medical image analysis.

The training unfolds in multiple communication rounds, following these steps:

TABLE 1. Summary of Some Significant Works in the Domain of Federated Learning.

FL Studies	Dataset	Data Partitioning	Model Implementation	Privacy Mechanism	Communication Architecture	Remarks
Cov-Fed [21]	CSM SARS MERS	Horizontal	MobileNetV2 ResNet-18 ShuffleNetV2 and VGG-11	Harmonic Encryption	Decentralized	Poor Performance with increase in number of connected clients
FedProx [14]	MNIST FEMNIST Shakespear Sent140	Horizontal	NN	Differential privacy	Secure multiparty communication	
VIRTUAL [36]	FEMNIST MNIST PMNIST VSN HAR NLP	Vertical	NN	Differential privacy	Wireless networks	
PerfCL [22]	CIFAR10 CIFAR100 Tiny-ImageNet	Horizontal	CNN ResNet	Differential		Doesn't consider the different number of client categories
MERGE [20]	CoViD-CXR ADNI	Horizontal	CNN	Differential		Lacks an intermediate validated federation setting
[37]	CIFAR-10 GoogleBigQuery	Horizontal	CNN		Structured and sketched updates,	
APFL [23]	MNIST1 CIFAR10 EMNIST Synthetic dataset	Horizontal	CNN MLP	Differential privacy		
FedAR [26]	ISIC2018 ChestXRay14	Horizontal	ResNet50 Res2Net50		Decentralized	Doesn't guarantee inferring of private information
[27]	ChestXRay-14	Horizontal	Resnet-50 AlexNet Densenet Inception VGG19			
FedVS [38]	FashionMNIST	Vertical	MLP CNN PN	Differential Privacy	Lagrange Coded Computing (LLC)	
FedCR [39]	MNIST FMNIST, CIFAR10 CIFAR100	Horizontal	CNN			No privacy preserving technique applied
[40]	BraTS	Horizontal	U-Net	Differential Privacy	Decentralized	Does not scale to many devices with modest data volumes.
FedEYE [33]	EDDL	Horizontal	ViT, Swin-T, ResNet	Differential Privacy	Decentralized	No data preprocessing is employed
PyVertical [41]	MNIST	Vertical	SplitNN	Differential Privacy	Decentralized	
Fedstellar [32]	MNIST, CIFAR-10	Vertical	SplitNN	Differential Privacy	Decentralized	
[34]	GAMEEMA, SEED, DEAP	Horizontal	Fuzzy ensemble of TCN, LSTM and GRU		Decentralized	No privacy mechanism
Herald [35]	Pnemonia Covid-Xray, Covid-19 Image and Radiography dataset	Horizontal	custom CNN		Decentralized	No privacy mechanism
Proposed Method	**HAM10000**	**Horizontal**	**ViT-B/16**	**Differential Privacy**	**Decentralized**	**Preprocessing is applied at randomized clients**

1) **Global Model Initialization:** The central server initializes a global ViT model w_0 and sends it to all participating clients.
2) **Local Training at Clients:** At each round t, client C_i receives the global ViT model w_t and trains it locally using its private dataset $\mathcal{D}_i$. Each client performs a few local training epochs using optimization methods such as stochastic gradient descent (SGD) or Adam, resulting in an updated local model w_t^i.
3) **Model Update Transmission:** Once local training is complete, clients send only their model updates w_t^i back to the central server. No raw data is ever transmitted, ensuring privacy.
4) **Global Aggregation:** The server aggregates the collected local ViT model updates using *Federated Averaging (FedAvg)*:
$$w_{t+1} = \frac{1}{n}\sum_{i=1}^{N} n_i \cdot w_t^i \tag{1}$$
where w_t^i is the updated ViT model from client C_i, n_i is the number of data samples on client C_i, and $n = \sum_{i=1}^{N} n_i$ is the total number of data samples across all clients.
5) **Model Distribution:** The newly aggregated ViT model w_{t+1} is redistributed to all clients to initiate the next training round.

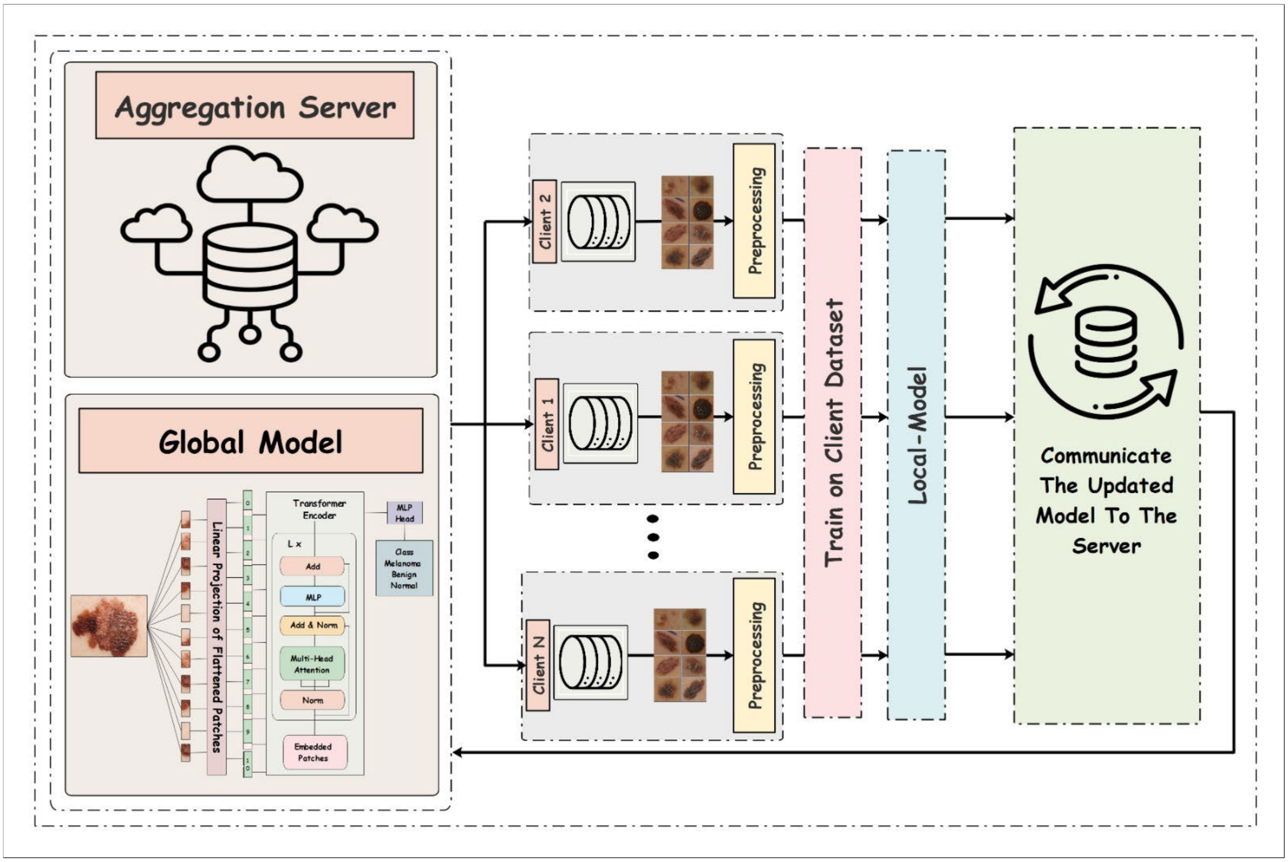


FIGURE 2. Proposed Methodology.

This iterative training continues until the model converges or a stopping criterion is reached. Integrating ViT within this federated learning framework allows for powerful, privacy-preserving, and globally aware image classification—especially valuable in decentralized medical and clinical settings where data heterogeneity and privacy concerns are significant.

A. BASE MODEL: VIT-B/16

The flow diagram of the ViT-B/16 is shown in Figure 3. Here, we summarize the key steps involved in the ViT architecture: The first step in the ViT architecture is Image Patching, where the input image is divided into fixed-size patches, which are treated as tokens similar to the ones used in Natural Language Processing (NLP) [42]. The ViT-B/16 performs the splitting of images in patches of size 16 × 16. In the second step, called Linear Embedding, each image patch is embedded into a lower-dimensional vector space, enabling the model to learn spatial relations and patterns between the patches [42].

Next, Positional Encoding is applied to the patches to provide the model with spatial information. Since transformers do not inherently understand the sequence of inputs, positional encoding is used to inject this spatial knowledge. The core of the ViT model consists of several Transformer Encoder Blocks, each containing feedforward neural networks and self-attention mechanisms. These encoder blocks allow the model to capture long-range dependencies and relationships between patches.

Within each transformer block, Multi-Head Self-Attention is used to enhance feature representation by allowing the model to focus on different aspects of the input. Attention is calculated as:

$$\text{Attention}(C, \kappa, \nu) = \text{Softmax}\left(\frac{C\kappa^T}{\sqrt{d}}\right)\nu \tag{2}$$

where C, κ, ν are computed as linear projections from the input $X \in \mathbb{R}^{N\times D}$. The computations are:

$$C = XW_C \tag{3}$$

$$\kappa = XW_\kappa \tag{4}$$

$$\nu = XW_\nu \tag{5}$$

Following this, Normalization and Residual Connections are applied after each subblock to stabilize the training process and improve the flow of information through the network [43]. Finally, the output from the transformer blocks passes through a classification head, typically consisting of a fully connected layer that makes the final predictions based on the aggregated representations of the image patches [42].

Despite its powerful architecture, the ViT-B/16 model shares some common limitations with CNNs, including the

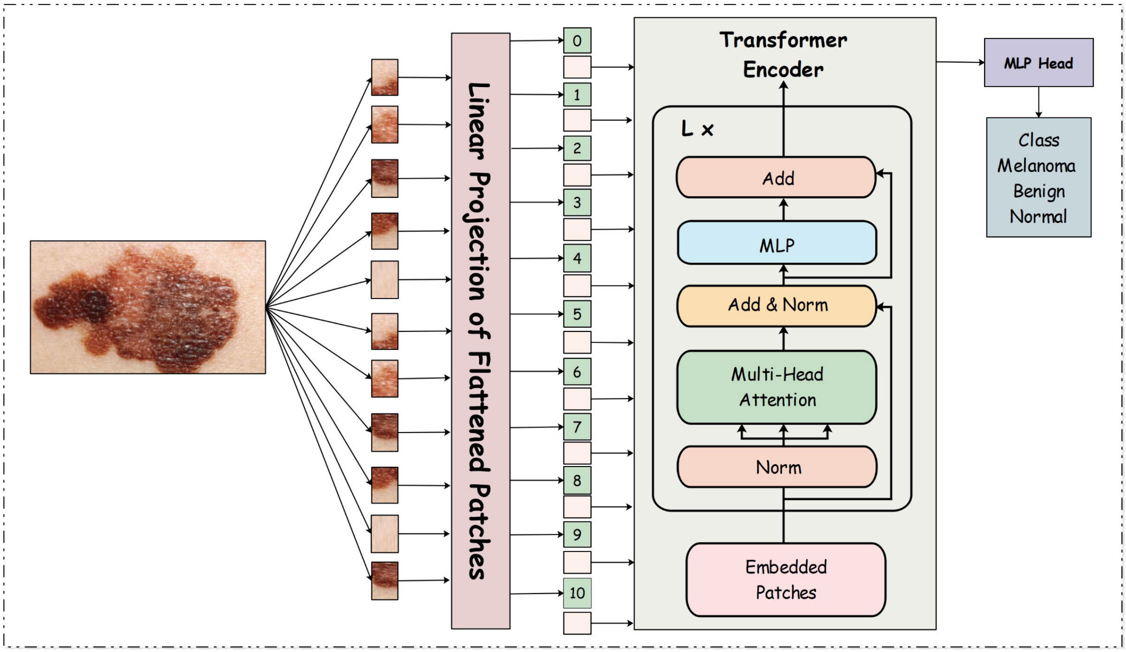


FIGURE 3. A Generic Architecture of ViT-B/16.

lack of inherent support for data privacy, high computational cost, and the requirement for large amounts of labeled data.

1) FEDERATED TRANSFER LEARNING OF VIT-B/16

The original proposed ViT-B/16 model was pretrained on the ImageNet dataset to leverage its ability to learn rich, generalized visual features. In the context of the proposed federated learning-based framework of skin lesion classification, the final fully connected (FC) layer, originally configured for 1000 ImageNet classes, is removed. A new fully connected layer is appended, sized according to the number of classes in the given skin lesion dataset (e.g., melanoma, nevus, basal cell carcinoma, etc.). This adapted ViT model is distributed to all participating clients.

Prior to transfer learning, each client performs standardized preprocessing of its private dataset to comply with the ViT-B/16 architecture specifications. The pipeline consists of the following steps:

1) **Data Augmentation**: To improve generalization and mitigate overfitting, we apply stochastic transformations including:
   - Random horizontal/vertical flipping ($p = 0.5$)
   - Rotation within ±15°
   - Scaling with a factor range of [0.9, 1.1]
2) **Spatial Normalization**: All images are resized to 224×224 pixels using bicubic interpolation, matching the input dimensions required by ViT's patch embedding layer [44].
3) **Intensity Normalization**: Pixel values are scaled to the interval [0, 1] through min-max normalization:

$$I_{\text{norm}} = \frac{I - I_{\min}}{I_{\max} - I_{\min}} \tag{6}$$

where I denotes the input image tensor.

This preprocessing ensures compatibility with the transformer's expectation of fixed-size inputs while preserving discriminative features for lesion classification.

B. FEDERATED TRAINING PROCESS

Each client k trains a local ViT-B/16 model using the cross-entropy loss over its private dataset $\mathcal{D}_k$:

$$\mathcal{L}_k(w) = \frac{1}{|\mathcal{D}_k|} \sum_{(x_i,y_i)\in\mathcal{D}_k} \text{CE}\big(f_w(x_i), y_i\big) \tag{7}$$

where:

- $f_w(x_i)$ denotes the ViT-B/16 logits for input x_i,
- CE$(\cdot)$ is the cross-entropy loss,
- $|\mathcal{D}_k| = n_k$ is the number of samples for client k.

To address the non-IID problems and enhance scalability, client participation is stochastic. Let p_k be the probability of selecting client k in a given round. The global aggregation at round τ becomes:

$$w^{\tau+1} = \sum_{k\in S_\tau} \frac{p_k n_k}{\sum_{j\in S_\tau} p_j n_j} w_k^{\tau+1} \tag{8}$$

where:

- $S_\tau \subseteq \{1, \ldots, K\}$ is the subset of participating clients at round τ,
- The weighting term $\frac{p_k n_k}{\sum_j p_j n_j}$ ensures fair contribution accounting for both selection probability and data quantity.

Next, the each client updates its weights using SGD/Adam:

$$\omega_k^{\tau+1} = \omega_k^\tau - \eta \nabla F_k(\omega_k^\tau) \tag{9}$$

Each client k evaluates its model on the test segment of its local dataset i.e. $\mathcal{D}_k^{test}$, computing accuracy using the

$$Acc_k = \frac{1}{|\mathfrak{D}_k^{test}|} \sum_{i \in D_k^t est} 1(\arg\max \hat{y}_i = y_i) \tag{10}$$

where $\hat{y}_i = \text{Softmax}(f_\omega(x_i))$ is the predicted label of the model, y_i is the actual class label, and $1(.)$ is the indicator function, which returns 1 if the predicted class label matches the actual class, otherwise it is 0.

After training, clients communicate their weights to the aggregating server, which computes the new weights by aggregating the client weights using 1. This process repeats for T iterations, after which the final model is used for evaluation.

TABLE 2. Class Distribution of HAM10000 Dataset.

Label	Lesion Type	No. of Images
AKIEC	Actinic keratoses and intraepithelial carcinoma	327
BCC	Basal cell carcinoma	514
BKL	Benign keratosis-like lesions	1,099
DF	Dermatofibroma	115
MEL	Melanoma	1,113
NV	Melanocytic nevi	6,705
VASC	Vascular lesions	142
Total		**10,015**

TABLE 3. Class Distribution of ISIC 2019 Dataset.

Label	Lesion Type	No. of Images
AKIEC	Actinic keratoses and intraepithelial carcinoma	867
BCC	Basal cell carcinoma	3,323
BKL	Benign keratosis-like lesions	2,624
DF	Dermatofibroma	239
MEL	Melanoma	4,522
NV	Melanocytic nevi	12,875
VASC	Vascular lesions	253
SCC	Squamous cell carcinoma	628
Total		**25,331**

C. FINAL FORMULATED MODEL

The federated learning objective combines local client objectives weighted by their data distribution:

$$\min_w \sum_{k=1}^{K} \frac{n_k}{N} \mathcal{L}_k(w) \tag{11}$$

The final Federated Learning objective function is given as;

$$\min_w \sum_{k=1}^{K} \frac{n_k}{N} \left(\frac{1}{n_k} \sum_{i \in \mathcal{D}_k} - \sum_{c=1}^{C} y_{i,c} \log \hat{y}_{i,c} \right) \tag{12}$$

subject to:

- Preprocessing: $x_i' = \frac{x_i - \mu}{\sigma}$
- ViT forward pass: z_L = BatchNorm(Dropout(MLP (LN(MSA(LN($z_{\ell-1}$))))))
- Dropout and BatchNorm applied in ViT layers.
- Global model aggregation via FedAvg.

Each client computes the train accuracy using the 13 [45]

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{13}$$

where TP denotes the true positives i.e., correctly predicted positive samples, TN denotes correctly predicted negative dataset samples, FP denotes incorrectly predicted positive dataset samples and TN denotes incorrectly predicted negative dataset samples

The cross-entropy loss for a batch of N samples is:

$$\mathcal{L} = -\sum_{i=1}^{N} \sum_{c=1}^{C} y_{i,c} \log(\hat{y}_{i,c}) \tag{14}$$

where:

- $y_{i,c} \in \{0, 1\}$ = Ground truth label (one-hot encoded)
- $\hat{y}_{i,c} \in [0, 1]$ = Predicted probability for class c
- C = Total number of classes

V. RESULTS AND DISCUSSION

A. EXPERIMENTAL SETUP

The proposed FL framework consists of a central aggregation server and three client nodes. Each client locally trains a vision transformer, ViT-B/16, on its respective skin lesion dataset while retaining raw data on-device to preserve privacy. Upon completing local training in each communication round, clients upload their model updates to a secure cloud storage. The central server retrieves these updates, aggregates them using Federated Averaging, and distributes the refined global model back to all clients for subsequent rounds.

All experiments were conducted within a private network environment. Each client was equipped with an Intel Core i7-12700K CPU (12 cores, 3.6 GHz), an NVIDIA RTX 3080 GPU (10 GB VRAM), 32 GB DDR4 RAM, and a 1 TB NVMe SSD, running Ubuntu 22.04 LTS. The central server maintained similar hardware with enhanced networking and storage capabilities to optimize synchronization. The Flower framework facilitated FL orchestration, including model synchronization and round coordination.

B. DATASET AND TRAINING CONFIGURATION

This study utilizes two publicly available dermatological image datasets: HAM10000 and ISIC2019. The HAM10000

TABLE 4. Model training parameters for federated learning Using vision transformers.

Parameter	Value	Parameter	Value
ViT Model	Pretrained ViT-B/16	Framework	Python (Pytorch)
Training function	SGD	Learning parameter	1×10^{-4}
Loss function	Cross Entropy	Federation Algorithm	FedAvg
Max. Epochs per client	3	Mini Batch Size	32
Global aggregation cycles	50	Stride	1

dataset comprises 10,015 images spanning 7 lesion categories, while ISIC2019 includes 25,331 images across 8 lesion types. The class distributions are detailed in Tables 2 and 3, with representative samples shown in Figure 5. To emulate real-world data heterogeneity, the datasets were partitioned across clients using a Dirichlet distribution based allocation (Dir(α)), a widely used approach for FL benchmarks [46], where α controls the degree of label skewness. Each class k's samples were split into shards following Dir($\alpha = 0.5$), and shards were assigned to clients such that no client held all classes. This ensured realistic imbalances in lesion type distributions. Each client's local dataset was further divided into training and testing subsets using an 80:20 hold-out split.

The ViT-B/16 model was fine-tuned using transfer learning, with hyperparameters detailed in Table 4. Hyperparameters were selected via grid search on client-held validation sets, balancing convergence and computational costs. For instance, a batch size of 32 and 3 local epochs prevent overfitting while minimizing communication overhead [47]. The learning rate (1e-4) follows ViT-FL benchmarks [17], and stride = 1 aligns with ViT-B/16's non-overlapping patch embedding. We conducted 50 communication rounds (global aggregations), with each client training for 3 epochs per round. Federated training spanned 50 communication rounds, with clients executing 3 local epochs per round using stochastic gradient descent (SGD) with momentum. After each round, model updates were transmitted to the server for FedAvg-based aggregation, and the global model was redistributed for subsequent training.

C. RESULTS OF HAM10000 DATASET

Figure 4 illustrates the training accuracy and loss trajectories of three clients over federated rounds on the HAM10000 dataset. Accuracy steadily improves, surpassing 99% after 30 rounds, while training loss declines sharply to near zero—demonstrating the ViT model's effectiveness in learning from decentralized, non-IID data. These trends highlight the robustness of the FedAvg-based aggregation and suggest strong model convergence. Figure 6 presents the corresponding confusion matrices, reflecting the ViT-B/16 model's strong classification performance across all clients after 50 communication rounds. Each client evaluates the final global model on its respective HAM10000 test partition, confirming consistent generalization across distributed environments.

Table 5 summarizes the key performance metrics obtained by the three clients on their respective test sets using the updated global ViT-B/16 model. The key performance metrics include accuracy, precision, sensitivity, specificity, and F1-score. For a class i, these metrics are computed as;

$$\text{Precision}_i = \frac{TP_i}{TP_i + FP_i} \tag{15}$$

$$\text{Sensitivity}_i = \frac{TP_i}{TP_i + FN_i} \tag{16}$$

$$\text{Specificity}_i = \frac{TN_i}{TN_i + FP_i} \tag{17}$$

$$\text{F1-score}_i = \frac{2 \cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i} \tag{18}$$

$$\text{Accuracy}_i = \frac{TP_i + TN_i}{TP_i + TN_i + FP_i + FN_i} \tag{19}$$

The values reported in Table 5 are macro-averaged, i.e., obtained by computing each metric per class and then averaging across all classes. All clients achieved high accuracy, with Client 2 leading at 98.90%, followed closely by Client 1 and Client 3 at 98.61% and 98.52%, respectively. Precision and specificity remained consistently high across clients, with values above 91% and 98%, respectively. Client 3 achieved the highest sensitivity (93.34%), while Client 2 showed the best F1-score (92.5%). These results reflect strong model generalization, effective handling of non-IID data, and the robustness of the FedAvg aggregation in preserving diagnostic accuracy across decentralized nodes.

TABLE 5. Test set performance metrics of three clients on HAM10000 dataset.

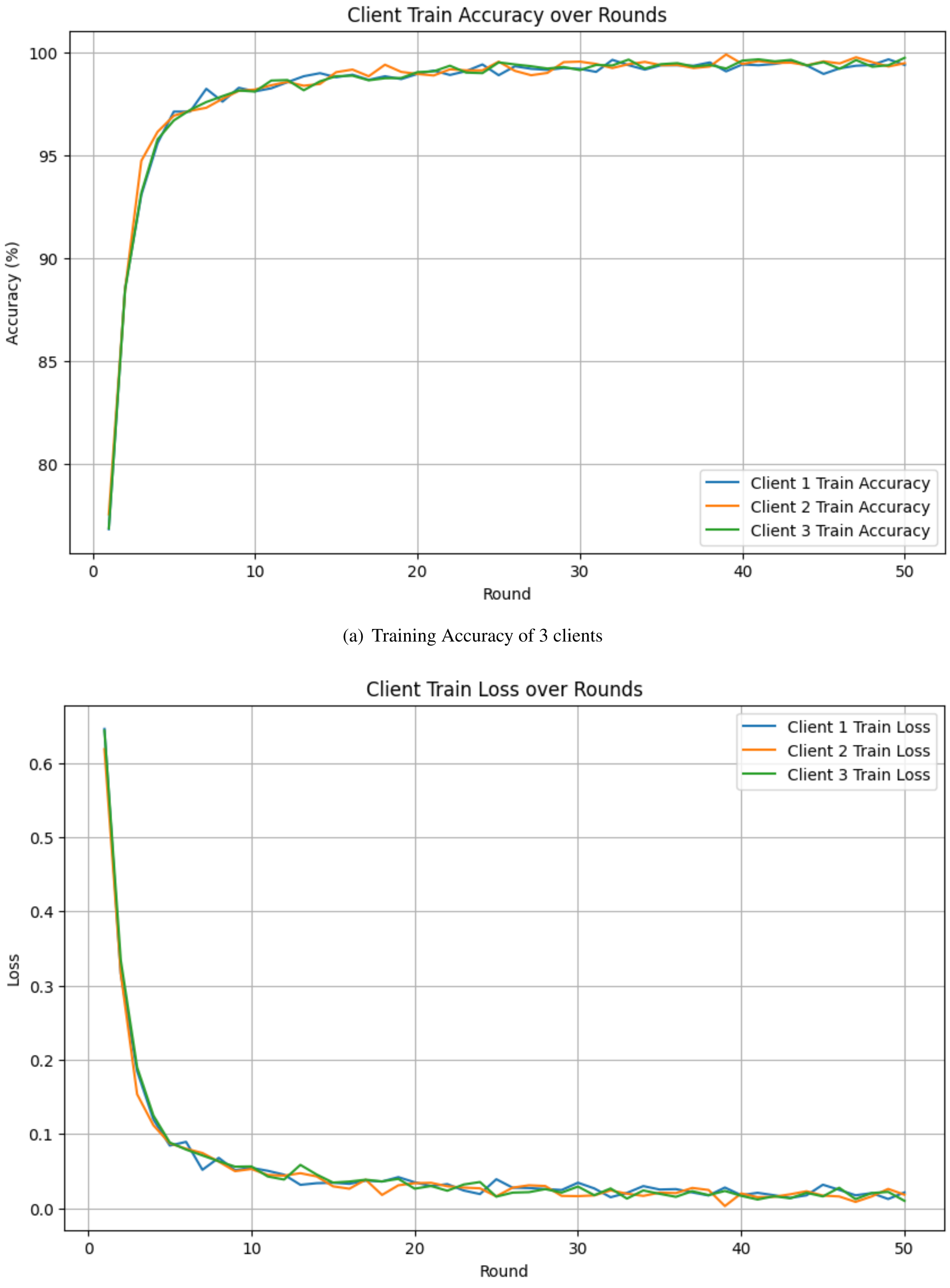
Client	Accuracy	Precision	Sensitivity	Specificity	F1-Score
Client 1	98.61%	94.98%	91.92%	98.99%	0.922
Client 2	98.9%	95.6%	91.3%	99.1 %	0.925
Client 3	98.52%	91.03%	93.34%	98.82%	0.9174

TABLE 6. Test set performance metrics of three clients on ISIC2019 Dataset.

Client	Accuracy	Precision	Sensitivity	Specificity	F1-Score
Client 1	97.18%	93.21%	98.80%	99.51%	0.959
Client 2	99.4%	94.23%	98.96%	99.56 %	0.9623
Client 3	97.50%	93.89%	99.0%	99.56%	0.9606

D. RESULTS OF ISIC2019 DATASET

The same FL pipeline of Figure 2 was applied to the ISIC2019 dataset. The data was partitioned non-IID across three clients,



(a) Training Accuracy of 3 clients

(b) Training Loss of 3 clients

FIGURE 4. Training accuracy and loss plots of three clients over FL iterations (rounds) for HAM10000 Dataset.

with each client's share further split into training and testing subsets for ViT-B/16 model training. Figure 7 presents the training accuracy and loss trajectories of the three federated clients over 50 communication rounds using the ISIC2019 dataset. All clients exhibit rapid convergence during the initial rounds. Accuracy improves significantly from around 55–60% to over 90% within the first 10 rounds, reflecting effective early learning and fast adaptation of the ViT-B/16

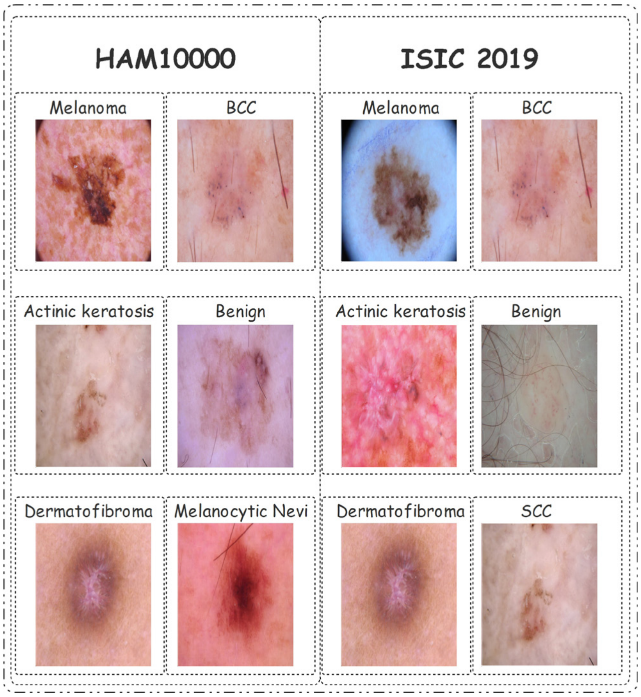


FIGURE 5. Representative samples of various skin cancer types from HAM10000 and ISIC 2019 datasets.

model to local data distributions. Beyond round 30, accuracy stabilizes near or above 96%, indicating strong generalization and sustained performance.

Concurrently, the training loss curves demonstrate a steep decline, falling from over 0.5 to below 0.05 within the first few rounds. This sharp reduction illustrates the model's ability to minimize classification error efficiently. Loss values converge to near-zero levels by round 30 and remain consistently low thereafter, with minimal variance across clients—highlighting the robustness and stability of the training process despite non-IID data partitions. The close alignment among all clients in both accuracy and loss trends underscores the effectiveness of the federated averaging strategy (FedAvg) in synchronizing local model updates. These results collectively validate the scalability and performance of the proposed federated ViT framework in handling complex, decentralized medical imaging tasks.

Figure 8 illustrates the confusion matrix for three clients using the ViT-B/16 model on the ISIC2019 dataset. The model demonstrates high classification performance across all eight lesion types, with strong diagonal dominance. Notably, melanocytic nevi (NV) and melanoma (MEL) show high true positive rates, though some confusion exists between them due to visual similarity, reflecting typical diagnostic challenges in distinguishing pigmented lesions. Classes such as BCC, BKL, AK, DF, SCC, and VASC are classified with near-perfect accuracy and minimal misclassification. Overall, the results confirm the model's robust generalization in a federated setting, with only minor confusion in closely related classes.

Table 6 summarizes the classification performance of the proposed federated ViT-B/16 model across three clients using the ISIC2019 dataset. All clients achieved high accuracy, with Client 2 outperforming others at 99.4%, followed by Client 3 (97.50%) and Client 1 (97.18%). The F1-scores across all clients remained consistently strong (above 0.95), indicating a balanced trade-off between precision and recall. Sensitivity and specificity values exceeding 98.9% and 99.5%, respectively, across all clients highlight the model's effectiveness in correctly identifying both positive and negative cases. Client 2 also achieved the highest F1-score (0.9623), reinforcing the model's ability to generalize well across decentralized and non-IID data distributions. These results demonstrate the robustness and scalability of the federated learning framework for complex skin lesion classification tasks in real-world medical imaging scenarios.

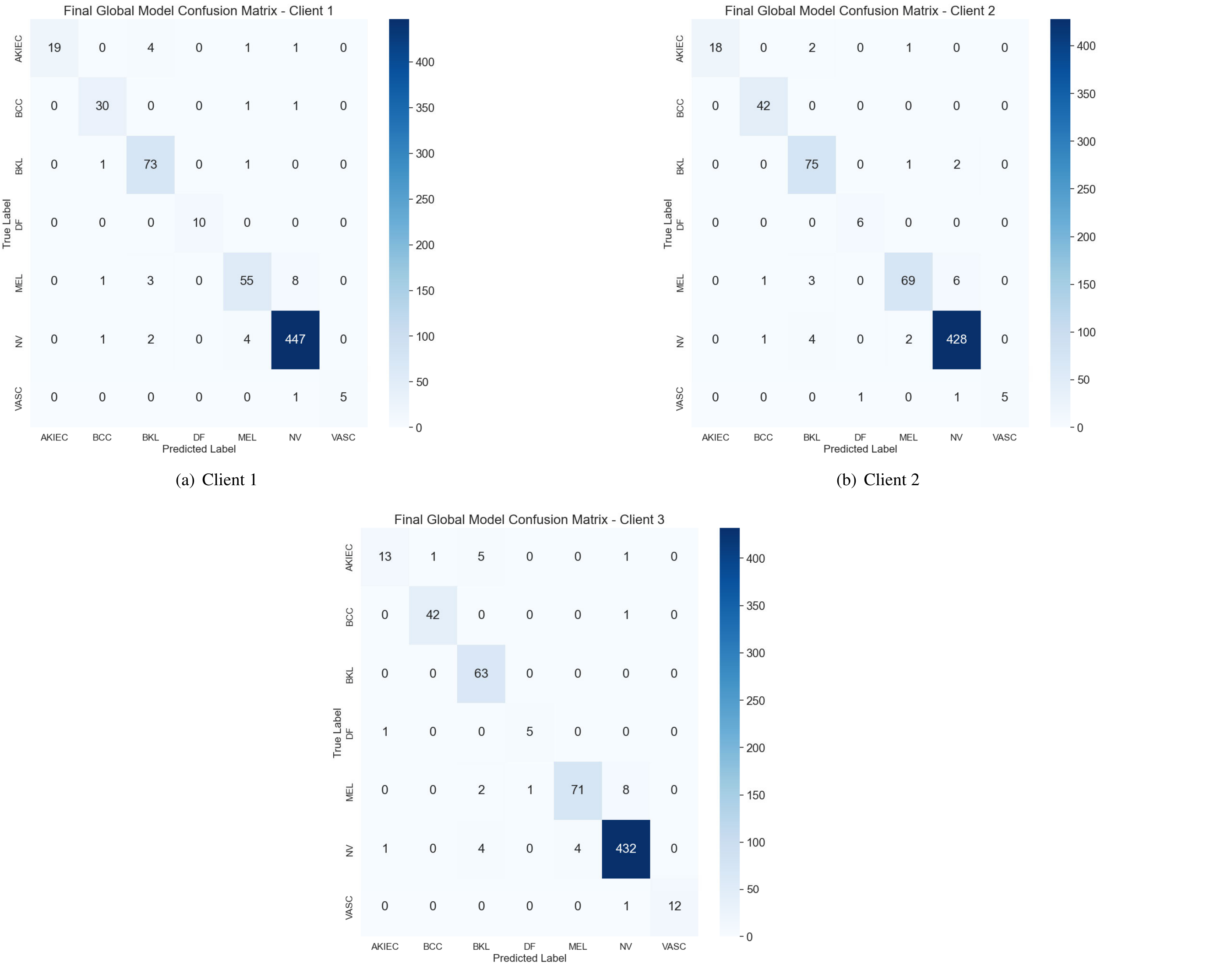
E. DISCRIMINATIVE PERFORMANCE EVALUATION VIA AUC

To evaluate the discriminative ability of the proposed federated learning framework, the Area Under the Receiver Operating Characteristic Curve (AUC-ROC) was computed for both the HAM10000 and ISIC2019 datasets. AUC is a threshold-independent metric that measures the model's ability to distinguish between classes by capturing the trade-off between the true positive rate (sensitivity) and the false positive rate. It is particularly useful in medical imaging tasks where class imbalance is common and classification confidence is critical. Given the multiclass nature of both datasets, AUC was computed using the one-vs-rest (OvR) strategy, where each class is evaluated independently against all others.

Figure 9 displays the AUC scores of the global ViT-B/16 model for each lesion class at the final communication round (Round 50) using the HAM10000 dataset. An AUC value closer to 1.0 indicates superior discriminative ability, while values closer to 0.5 suggest the model performs no better than random guessing. As illustrated, the model achieved consistently high AUC values across all seven classes, indicating excellent discriminative capability. The highest AUC was observed for the vascular lesions (VASC) class, reaching nearly 1.00, which reflects the model's ability to perfectly differentiate these lesions from all others. Melanocytic nevi (NV) and dermatofibroma (DF) followed closely, each attaining AUC values above 0.98, further confirming strong class-specific learning.

Other lesion types, such as melanoma (MEL), benign keratosis-like lesions (BKL), and basal cell carcinoma (BCC), also recorded high AUCs ranging from approximately 0.96 to 0.98, suggesting stable performance with minimal overlap among predictions. Actinic keratoses and intraepithelial carcinoma (AKIEC), while slightly lower than others, still achieved an AUC above 0.96, indicating reliable detection even for less frequent or more challenging classes.

Figure 10 illustrates the per-class AUC scores for ISIC2019 dataset. The model also performs strongly, though with slightly more variance across the eight lesion classes. The



(a) Client 1

(b) Client 2

(c) Client 3

FIGURE 6. Test Confusion Matrices of three clients using updated global model on Ham10000 dataset.

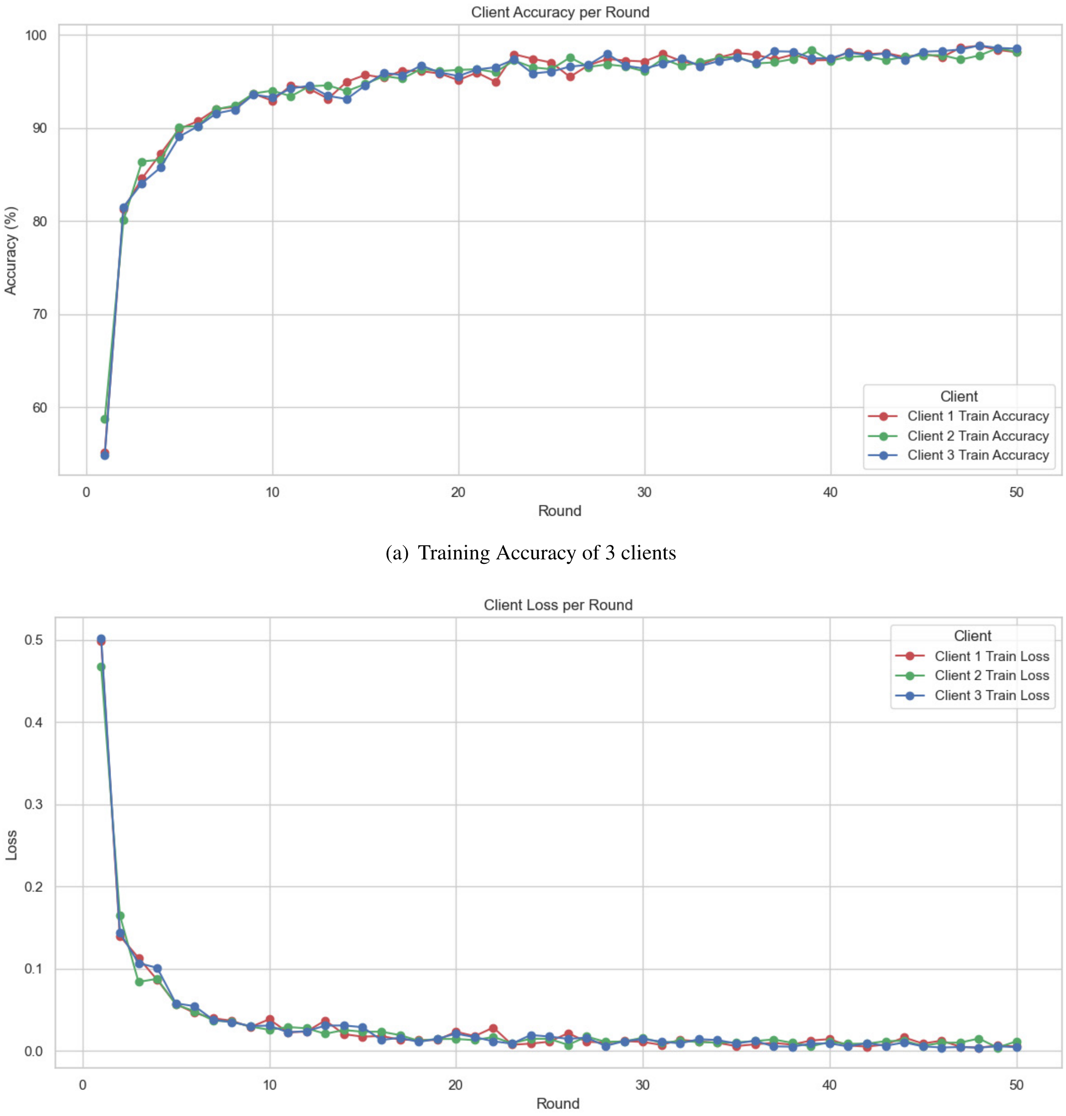
TABLE 7. Performance Comparison with Some Existing Works on FL based Skin Lession Classification.

Method	FL/ViT?	Key Strength	Accuracy (HAM10000)	AUC	Non-IID Robustness
FedEYE [33]	ViT+FL	First ViT-FL benchmark	∼85%	0.88	Low
FedAR [26]	FL (CNN)	Adaptive client weighting	86% (ISIC2018)	0.92	Moderate
[43]	ViT	Centralized multi-modal	89% (ISIC 2019)	0.94	N/A
Ours	**ViT+FL**	**Privacy + global features**	**90%**	**0.96**	**High**

highest AUCs were observed for VASC, SCC, NV, and DF, all approaching or exceeding 0.90, signifying high confidence in class separation. Classes such as MEL, BCC, and BKL showed slightly lower but still reliable AUC scores in the range of 0.85 to 0.90. Actinic keratoses (AK) recorded the lowest, though still acceptable, AUC values slightly above 0.80, reflecting higher intra-class ambiguity and overlap with similar lesion types.

The overall high and stable AUC scores across both datasets confirm the effectiveness of the ViT-based federated learning framework in achieving strong class-level discrimination under non-IID conditions. The results validate the model's capability to generalize well across decentralized medical image data, ensuring reliable skin lesion classification performance in real-world scenarios.

In Table 7, a performance comparison is presented between our proposed Federated ViT framework and state-of-the-art methods in skin lesion classification. Our approach demonstrates clear advantages over existing solutions, particularly in addressing the key challenges of privacy preservation and



(a) Training Accuracy of 3 clients

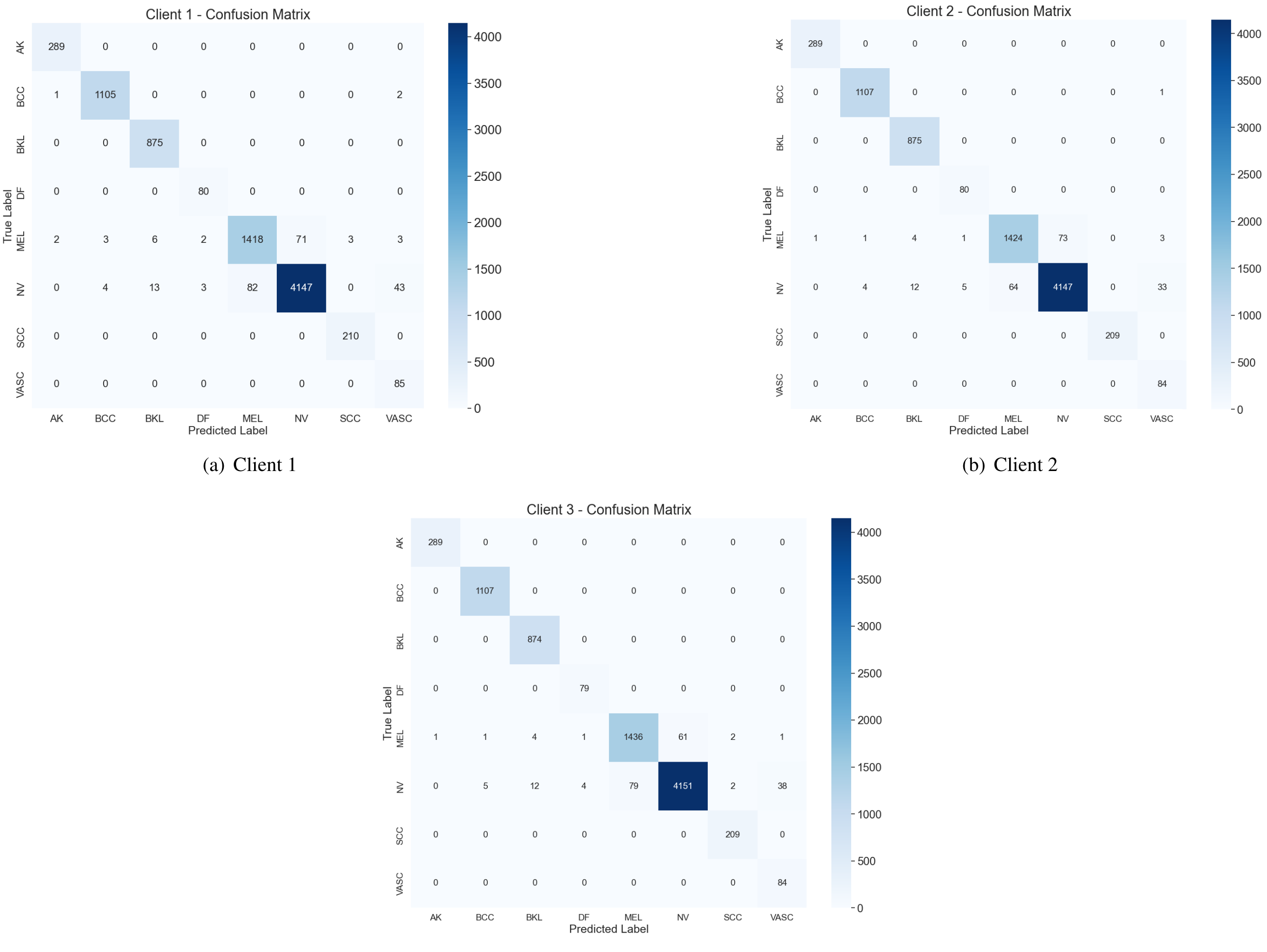
(b) Training Loss of 3 clients

FIGURE 7. Training accuracy and loss plots of three clients over FL iterations (rounds) for ISIC2019 Dataset.

non-IID data distributions. Compared to FedEYE [33]–the only other FL study employing ViTs–our framework achieves 5% higher accuracy on HAM10000 (90% vs. 85%) through novel client-level weighting and ViT-specific optimizations for federated settings. While CNN-based FL methods like FedAR [26] incorporate adaptive client weighting, they remain fundamentally limited by CNNs' local receptive fields, whereas our ViT-based solution better models global lesion characteristics, as evidenced by superior AUC scores (0.96 vs. 0.92). Remarkably, our federated approach even matches the performance of centralized ViT methods like [43] (90% vs. 89% accuracy) while providing the critical benefit of data privacy. This comparison underscores our framework's unique capability to combine ViTs' diagnostic strengths with FL's privacy guarantees while maintaining robustness to heterogeneous data distributions–a combination unmatched by any existing method in the literature.

F. DISCUSSION

Our experimental results demonstrate that the proposed federated ViT framework achieves strong performance in skin lesion classification, with global test accuracies of 90% (HAM10000) and 87.6% (ISIC2019). The high AUC scores (up to 0.96) further confirm the model's discriminative ability across all lesion classes, including rare categories. These outcomes validate that ViTs can effectively operate in federated settings while preserving data privacy—a critical requirement for medical applications. The framework's decentralized architecture ensures all training data remains



(a) Client 1

(b) Client 2

(c) Client 3

FIGURE 8. Test Confusion Matrices of three clients using updated global model on ISIC2019 dataset.

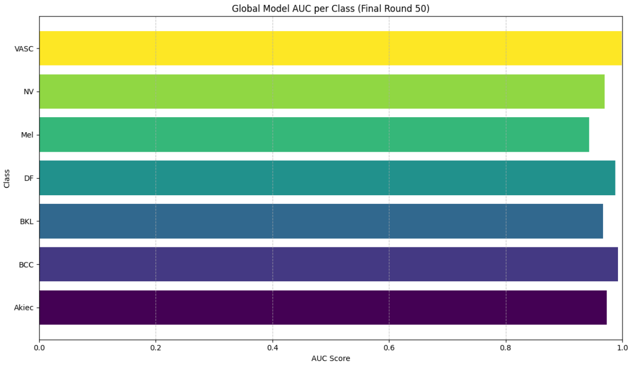


FIGURE 9. AUC per class For Global Model using aggregated HAM10000 dataset.

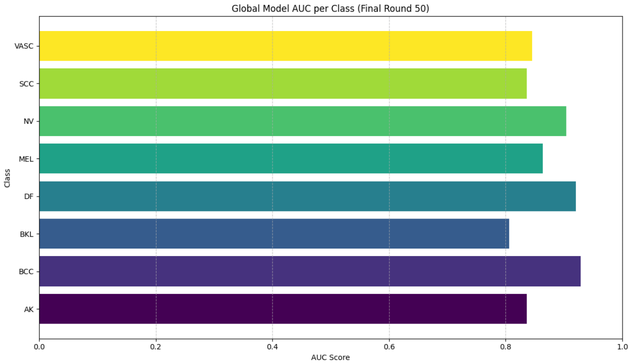


FIGURE 10. AUC per class For Global Model using aggregated ISIC2019 dataset.

localized on client devices (e.g., hospitals or diagnostic centers), with only model updates—never raw images—transmitted to the aggregation server. This design inherently complies with stringent medical data regulations, including HIPAA and GDPR, by eliminating centralized data storage vulnerabilities. Stochastic client selection further enhances participant anonymity.

While our experiments focus on skin cancer classification using HAM10000 and ISIC2019, the framework is designed for broad applicability across dermatological and other healthcare domains. The modular architecture allows seamless integration of new datasets with varying resolutions, class distributions, and imaging modalities (e.g., dermoscopy, clinical photos, or histopathology). Since ViTs inherently capture global contextual features rather than local inductive biases (unlike CNNs), they generalize well to diverse lesion morphologies, including rare or underrepresented conditions in federated settings.

The framework's scalability is demonstrated by its efficient handling of non-IID data, a common challenge in real-world healthcare collaborations. By weighting client contributions based on dataset size (Eq. 8) and employing federated averaging (FedAvg), the system adapts to imbalances without requiring centralized data pooling. This flexibility suggests applicability to other dermatological conditions (e.g. psoriasis, eczema etc.), multi-model health care data and global health initiatives involving low-resource clinics with small, heterogeneous datasets.

However, several limitations must be acknowledged that impact clinical applicability. First, while ViTs excel at capturing long-range dependencies, their computational complexity may hinder deployment on resource-constrained edge devices common in healthcare. Second, the communication overhead of federated ViT training, though manageable in our experiments, could become prohibitive in bandwidth-limited real-world scenarios. Third, our current implementation uses a modified FedAvg approach; comparisons with state-of-the-art FL methods (e.g., FedProx, MOON) tailored for non-IID data would strengthen the robustness claims. Finally, the framework assumes high-quality annotations across clients, whereas real-world settings often face label noise and inter-institutional variability.

Beyond these technical limitations, real-world clinical deployment faces additional challenges of data heterogeneity that extend beyond non-IID label distribution. Differences in dermoscope vendors, lighting conditions, patient skin tones, and imaging protocols across hospitals represent a significant "domain shift" that can substantially impact model performance when deploying the global model in new clinical environments with different data characteristics. Furthermore, system heterogeneity—variations in computational resources, network connectivity, and hardware capabilities across participating institutions—could lead to stragglers that slow down the federated averaging process or even cause client dropouts, challenging the stability and efficiency of the training loop.

Integration of interpretability is a critical step for the clinical acceptance of federated AI systems. In this regard, Gradient-weighted Class Activation Mapping (Grad-CAM) and SHapley Additive exPlanations (SHAP) can be incorporated into our proposed Federated ViT framework to enhance transparency and trust. Grad-CAM generates class-specific visual explanations by backpropagating gradients to the final attention layers of the ViT, thereby producing heatmaps that highlight the most relevant regions of an input image influencing the prediction. In the case of skin cancer classification, such heatmaps can validate whether the model is attending to clinically meaningful lesion characteristics, including irregular borders, pigment distribution, or vascular patterns. While originally designed for CNNs, Grad-CAM has been successfully extended to ViTs by leveraging patch embeddings and attention weights from the last transformer encoder block, allowing class-discriminative localization maps to be projected back onto dermoscopic images. SHAP, on the other hand, complements Grad-CAM by providing feature-level interpretability through game-theoretic attribution values that quantify the contribution of each input pixel or patch to the final prediction. This dual perspective—spatial localization from Grad-CAM and feature contribution analysis from SHAP—offers a comprehensive explanation of model behavior. Within our federated pipeline, both techniques can be applied locally at each client, ensuring interpretability without compromising privacy: clients can generate explanations on their private test images, the aggregated global model can be validated through local interpretability checks, and clinicians can be directly involved by reviewing highlighted lesion regions and feature attributions in a clinician-in-the-loop feedback process. This integration offers multiple benefits in sensitive medical applications—providing transparency into decision-making, supporting error analysis on misclassified cases, improving clinical trust by aligning predictions with dermatological features, and preserving privacy since raw images remain local. As a future extension of this work, we plan to systematically evaluate the quality of Grad-CAM heatmaps and SHAP attributions against expert annotations, compare their effectiveness across lesion types, and explore federated explainability aggregation where clients can contribute anonymized interpretability statistics without exposing raw data.

To address the identified limitations and enhance clinical applicability, future work will pursue several key directions. First, we will develop lightweight ViT variants such as MobileViT to reduce computational and communication costs for deployment on resource-constrained devices. Second, we will integrate advanced FL optimization techniques including FedBN (to mitigate feature shift from different imaging devices through local batch normalization), FedProx (to prevent client drift and improve convergence under system heterogeneity), and SCAFFOLD for improved handling of data heterogeneity. Third, we will extend validation to multi-modal medical data (combining dermoscopy with clinical metadata) and noisy-label scenarios through robust learning techniques. Additionally, we will explore federated segmentation capabilities to provide crucial spatial information beyond classification and investigate fairness constraints to ensure equitable performance across diverse demographic groups and skin tones. These improvements aim to bridge the gap between research prototypes and clinically viable tools while maintaining strict privacy guarantees, ultimately enabling the framework to integrate with Hospital Information Systems (HIS) and provide structured reports that fit seamlessly into dermatologists' workflows.

VI. CONCLUSION

This study presents a robust and privacy-preserving federated learning framework for automated skin lesion classification using the Vision Transformer (ViT-B/16) architecture. By leveraging federated learning, the proposed system enables collaborative model training across multiple decentralized clients without exposing sensitive patient data. The model was trained and evaluated using two benchmark dermatological image datasets—HAM10000 and ISIC2019—partitioned in a non-IID manner to simulate real-world clinical settings. The experimental results demonstrate that the ViT-based global model achieves high classification performance across all clients, with consistently strong metrics including accuracy, precision, sensitivity, specificity, F1-score, and AUC-ROC. Both per-class and macro-averaged AUC scores confirm the model's ability to distinguish between visually similar skin lesions, despite data heterogeneity and class imbalance. Moreover, the federated setup exhibited excellent scalability and communication efficiency, with convergence achieved in a limited number of rounds. The high fidelity of training and validation performance across multiple rounds and datasets indicates that ViTs, when integrated with federated learning, serve as a powerful solution for real-world, privacy-sensitive medical imaging tasks. In future work, the framework can be extended to include heterogeneous model architectures, additional dermatological datasets, and real-time clinical deployment scenarios, further enhancing its utility in tele-dermatology and collaborative AI-assisted diagnostics

DATASET AND CODE AVAILABILITY

The HAM10000 dataset is made available by the authors of [48] at the URL: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T. The ISIC2019 dataset is available at the URL: https://challenge.isic-archive.com/data/#2019. The complete code of the proposed VIT assisted federated learning pipeline is available at the URL: https://github.com/RiazJunejo/ViT-FederatedLearning.

REFERENCES

[1] R. Manne, S. Kantheti, and S. Kantheti, "Classification of skin cancer using deep learning, convolutionalneural networks-opportunities and vulnerabilities—A systematic review," *Int. J. Mod. Trends Sci. Technol.*, vol. 6, pp. 2455–3778, Jan. 2020.

[2] (2022). *Skin Cancer Statistics Melanoma Skin Cancer Rates*. [Online]. Available: https://www.wcrf.org/cancer-trends/skin-cancer-statistics/

[3] H. Ghosh, I. S. Rahat, S. N. Mohanty, J. Ravindra, and A. Sobur, "A study on the application of machine learning and deep learning techniques for skin cancer detection," *Int. J. Comput. Syst. Eng.*, vol. 18, no. 1, pp. 51–59, 2024.

[4] M. A. Khan, M. Sharif, T. Akram, S. A. C. Bukhari, and R. S. Nayak, "Developed Newton-Raphson based deep features selection framework for skin lesion recognition," *Pattern Recognit. Lett.*, vol. 129, pp. 293–303, Jan. 2020.

[5] N. H. Khan, M. Mir, L. Qian, M. I. Baloch, M. F. A. Khan, A. U. Rehman, E. E. Ngowi, D. Wu, and X. Ji, "Skin cancer biology and barriers to treatment: Recent applications of polymeric micro/nanostructures," *J. Adv. Res.*, vol. 36, pp. 223–247, Jan. 2021.

[6] J. K. Cullen, J. L. Simmons, P. G. Parsons, and G. M. Boyle, "Topical treatments for skin cancer," *Adv. Drug Del. Rev.*, vol. 153, pp. 54–64, Jan. 2020.

[7] P. Hermosilla, R. Soto, E. Vega, C. Suazo, and J. Ponce, "Skin cancer detection and classification using neural network algorithms: A systematic review," *Diagnostics*, vol. 14, no. 4, p. 454, Feb. 2024.

[8] M. Goyal, T. Knackstedt, S. Yan, and S. Hassanpour, "Artificial intelligence-based image classification methods for diagnosis of skin cancer: Challenges and opportunities," *Comput. Biol. Med.*, vol. 127, Dec. 2020, Art. no. 104065.

[9] A. F. Jerant, J. T. Johnson, C. D. Sheridan, and T. J. Caffrey, "Early detection and treatment of skin cancer," *Amer. Family Physician*, vol. 62, no. 2, pp. 357–368, 2000.

[10] L. Rey-Barroso, S. Peña-Gutiérrez, C. Yáñez, F. J. Burgos-Fernández, M. Vilaseca, and S. Royo, "Optical technologies for the improvement of skin cancer diagnosis: A review," *Sensors*, vol. 21, no. 1, p. 252, Jan. 2021.

[11] C. Reggiani, M. Manfredini, V. D. Mandel, F. Farnetani, S. Ciardo, S. Bassoli, A. Casari, S. Guida, G. Argenziano, A. Lallas, M. Ulrich, G. Pellacani, and C. Longo, "Update on non-invasive imaging techniques in early diagnosis of non-melanoma skin cancer," *G Ital Dermatol Venereol*, vol. 150, no. 4, pp. 393–405, 2015.

[12] M. Atikur Rahman, E. Bazgir, S. M. Saokat Hossain, and M. Maniruzzaman, "Skin cancer classification using NASNet," *Int. J. Sci. Res. Arch.*, vol. 11, no. 1, pp. 775–785, Jan. 2024.

[13] A. Nilsson, S. Smith, G. Ulm, E. Gustavsson, and M. Jirstrand, "A performance evaluation of federated learning algorithms," in *Proc. 2nd Workshop Distrib. Infrastruct. Deep Learn.*, Dec. 2018, pp. 1–8.

[14] T. Li, A. K. Sahu, A. Talwalkar, and V. Smith, "Federated learning: Challenges, methods, and future directions," *IEEE Signal Process. Mag.*, vol. 37, no. 3, pp. 50–60, May 2020.

[15] M. K. Hasan, L. Dahal, P. N. Samarakoon, F. I. Tushar, and R. Martí, "DSNet: Automatic dermoscopic skin lesion segmentation," *Comput. Biol. Med.*, vol. 120, May 2020, Art. no. 103738.

[16] M. A. Al-Masni, D.-H. Kim, and T.-S. Kim, "Multiple skin lesions diagnostics via integrated deep convolutional networks for segmentation and classification," *Comput. Methods Programs Biomed.*, vol. 190, Jul. 2020, Art. no. 105351.

[17] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[18] T. Gafni, N. Shlezinger, K. Cohen, Y. C. Eldar, and H. V. Poor, "Federated learning: A signal processing perspective," *IEEE Signal Process. Mag.*, vol. 39, no. 3, pp. 14–41, May 2022.

[19] P. Mary Mammen, "Federated learning: Opportunities and challenges," 2021, *arXiv:2101.05428*.

[20] B. Casella, W. Riviera, M. Aldinucci, and G. Menegaz, "MERGE: A model for multi-input biomedical federated learning," *Patterns*, vol. 4, no. 11, Nov. 2023, Art. no. 100856.

[21] I. Adjei-Mensah, X. Zhang, I. O. Agyemang, S. B. Yussif, A. A. Baffour, B. M. Cobbinah, C. Sey, L. D. Fiasam, I. A. Chikwendu, and J. R. Arhin, "Cov-Fed: Federated learning-based framework for COVID-19 diagnosis using chest X-ray scans," *Eng. Appl. Artif. Intell.*, vol. 128, Feb. 2024, Art. no. 107448.

[22] Y. Zhang, Y. Xu, S. Wei, Y. Wang, Y. Li, and X. Shang, "Doubly contrastive representation learning for federated image recognition," *Pattern Recognit.*, vol. 139, Jul. 2023, Art. no. 109507.

[23] Y. Deng, M. Mahdi Kamani, and M. Mahdavi, "Adaptive personalized federated learning," 2020, *arXiv:2003.13461*.

[24] S. Sharma and K. Guleria, "A comprehensive review on federated learning based models for healthcare applications," *Artif. Intell. Med.*, vol. 146, Dec. 2023, Art. no. 102691.

[25] S. Lu, Z. Liu, T. Liu, and W. Zhou, "Scaling-up medical vision-and-language representation learning with federated learning," *Eng. Appl. Artif. Intell.*, vol. 126, Nov. 2023, Art. no. 107037.

[26] B. Ma, Y. Feng, G. Chen, C. Li, and Y. Xia, "Federated adaptive reweighting for medical image classification," *Pattern Recognit.*, vol. 144, Dec. 2023, Art. no. 109880.

[27] A. Kareem, H. Liu, and V. Velisavljevic, "A federated learning framework for pneumonia image detection using distributed data," *Healthcare Anal.*, vol. 4, Dec. 2023, Art. no. 100204.

[28] L. Lusnig, A. Sagingalieva, M. Surmach, T. Protasevich, O. Michiu, J. McLoughlin, C. Mansell, G. de' Petris, D. Bonazza, F. Zanconati, A. Melnikov, and F. Cavalli, "Hybrid quantum image classification and federated learning for hepatic steatosis diagnosis," *Diagnostics*, vol. 14, no. 5, p. 558, Mar. 2024.

[29] Z. Yang and X. Wang, "A study on tuberculosis CT image classification based on federated learning methods," *Int. J. Comput. Sci. Inf. Technol.*, vol. 2, no. 2, pp. 369–379, Apr. 2024.

[30] A. Woubie, E. Solomon, and J. Attieh, "Maintaining privacy in face recognition using federated learning method," *IEEE Access*, vol. 12, pp. 39603–39613, 2024.

[31] J. Park and J. Ko, "FedHM: Practical federated learning for heterogeneous model deployments," *ICT Exp.*, vol. 10, no. 2, pp. 387–392, Apr. 2024.

[32] E. T. M. Beltrán, Á. L. P. Gómez, C. Feng, P. M. S. Sánchez, S. L. Bernal, G. Bovet, M. Gil Pérez, G. M. Pérez, and A. H. Celdrán, "Fedstellar: A platform for decentralized federated learning," *Expert Syst. Appl.*, vol. 242, May 2024, Art. no. 122861.

[33] B. Yan, D. Cao, X. Jiang, Y. Chen, W. Dai, F. Dong, W. Huang, T. Zhang, C. Gao, Q. Chen, Z. Yan, and Z. Wang, "FedEYE: A scalable and flexible end-to-end federated learning platform for ophthalmology," *Patterns*, vol. 5, no. 2, 2024, Art. no. 100928.

[34] J. Weiwei, Z. Yang, H. Haoyu, L. Xiaozhu, G. Jeonghwan, G. Weixi, S. Achyut, and M. Carsten, "Fuzzy ensemble-based federated learning for EEG-based emotion recognition in Internet of Medical Things," *J. Ind. Inf. Integr.*, vol. 44, Mar. 2025, Art. no. 100789. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2452414X25000135

[35] G. Tricomi, G. Cicceri, I. Ficili, S. Vitabile, G. Merlino, and A. Puliafito, "HERALD: A hybrid distributEd learning incrementAL & federated solution for knowledge distillation in COVID-19 classification," *Future Gener. Comput. Syst.*, vol. 174, Jan. 2026, Art. no. 107991.

[36] L. Corinzia, A. Beuret, and J. M. Buhmann, "Variational federated multi-task learning," 2019, *arXiv:1906.06268*.

[37] J. Konecný, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," 2016, *arXiv:1610.05492*.

[38] S. Li, D. Yao, and J. Liu, "FedVS: Straggler-resilient and privacy-preserving vertical federated learning for split models," 2023, *arXiv:2304.13407*.

[39] H. Zhang, C. Li, W. Dai, J. Zou, and H. Xiong, "Fedcr: Personalized federated learning based on across-client common representation with conditional mutual information regularization," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 41314–41330.

[40] M. Sheller, G. A. Reina, B. Edwards, J. Martin, and S. Bakas, "Multi-institutional deep learning modeling without sharing patient data: A feasibility study on brain tumor segmentation," in *Proc. Int. MICCAI Brainlesion Workshop*, Sep. 2018, pp. 92–104.

[41] D. Romanini, A. J. Hall, P. Papadopoulos, T. Titcombe, A. Ismail, T. Cebere, R. Sandmann, R. Roehm, and M. A. Hoeh, "PyVertical: A vertical federated learning framework for multi-headed SplitNN," 2021, *arXiv:2104.00489*.

[42] Y. Dai, Y. Gao, and F. Liu, "TransMed: Transformers advance multi-modal medical image classification," *Diagnostics*, vol. 11, no. 8, p. 1384, Jul. 2021.

[43] C. Matsoukas, J. F. Haslum, M. Söderberg, and K. Smith, "Is it time to replace CNNs with transformers for medical images?" 2021, *arXiv:2108.09038*.

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.

[45] M. Khalifa and M. Albadawy, "AI in diagnostic imaging: Revolutionising accuracy and efficiency," *Comput. Methods Programs Biomed. Update*, vol. 5, Jan. 2024, Art. no. 100146.

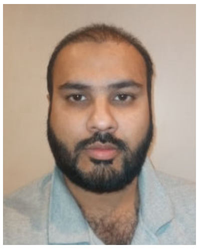
[46] Y. Zhao, M. Li, L. Lai, N. Suda, D. Civin, and V. Chandra, "Federated learning with non-IID data," 2018, *arXiv:1806.00582*.

[47] H. B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y. Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1273–1282.

[48] P. Tschandl, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, no. 1, pp. 1–9, 2018.

RIAZ H. JUNEJO is currently pursuing the Ph.D. degree with the Department of Computer Engineering, COMSATS University Islamabad, Wah Campus. His research interests include computer vision and deep learning methods for biomedical image classification.

QAISAR ABBAS (Member, IEEE) is currently a Distinguished Researcher and a Faculty Member at Imam Mohammad Ibn Saud Islamic University (IMSIU) and an Expert in artificial intelligence (AI). He has contributed to early disease detection, sustainable practices, industrial efficiency, and economic modeling through innovative AI techniques, including deep learning (DL) and federated learning (FL). His work bridges cutting-edge research with real-world applications, fostering impactful advancements for societal and industrial benefit.



MUHAMMAD AWAIS (Senior Member, IEEE) received the Ph.D. degree from the VLSI Laboratory, Department of Electronics and Telecommunications, Politecnico di Torino, Italy, in 2014. He is currently an Associate Professor at the Department of Computer Science, College of Computer, Qassim University, Saudi Arabia. His research interests include advanced digital design and algorithms from various domains, including wireless communication, machine learning, and signal processing. He is doing research in the domains of deep learning, computer vision, and applied optimization of feature selection algorithms for the biomedical domain.

TALLHA AKRAM received the B.Sc. degree in computer engineering from COMSATS University Islamabad, Abbottabad Campus, Pakistan, in 2006, the M.Sc. degree in embedded systems and control engineering from Leicester University, U.K., in 2008, and the Ph.D. degree in computer vision and pattern recognition from Chongqing University, in 2014. He is currently an Associate Professor with the Department of Information Systems, College of Computer Engineering and Sciences, Prince Sattam bin Abdulaziz University, Saudi Arabia. He is the author of several peer-reviewed journals and conferences. His research interests include computer vision, pattern recognition, machine learning, artificial intelligence, and applied optimization.

MUTLAQ B. ALDAJANI is currently a Professor of information systems at Imam Mohammad Ibn Saud Islamic University (IMSIU). He held multiple leadership positions throughout his career, including the Chair of the IS Department, in 2010, and the Vice Dean for academic affairs of the College of Computer and Information Sciences, in 2012. His primary research interests include interactive systems, with a particular focus on multimodal interaction and adaptive user interfaces. Additionally, he explores consumer behavior within the field of information systems research, emphasizing the application of user acceptance of technology. He has published a diverse range of publications covering various topics, including e-business and e-commerce, customer knowledge management, cloud computing, mobile computing, and usability heuristics.