Ingénierie des Systèmes d'Information
Vol. 31, No. 4, April, 2026, pp. 1197-1215
Journal homepage: http://iieta.org/journals/isi

# A Privacy-Preserving Federated Swin Transformer V2 Framework with Explainable Artificial Intelligence for Skin Lesion Classification

Varadan Sharmila[1*], Periyathambi Ezhumalai[1], Shanker Shalini[2], Palanisamy Brinda[3]

[1] Department of Computer Science and Engineering, R.M.D. Engineering College, Kavaraipettai 601206, India
[2] Department of Computer Science and Engineering, St.Joseph's Institute of Technology Chennai, Chennai 600119, India
[3] Department of Computer Science and Engineering, Vel Tech High Tech Dr Rangarajan Dr Sakunthala Engineering College, Chennai 600062, India

Corresponding Author Email: sharmilavaradhan@gmail.com



https://doi.org/10.18280/isi.310416

**Received:** 12 January 2026
**Revised:** 15 March 2026
**Accepted:** 16 April 2026
**Available online:** 30 April 2026

***Keywords:***
*skin lesion classification, federated learning, Swin Transformer V2, privacy preservation, Explainable Artificial Intelligence, medical image analysis, distributed deep learning, clinical decision support*

**ABSTRACT**

Early and accurate skin lesion classification is essential for the timely diagnosis of melanoma and other dermatological diseases. Existing skin lesion classifiers often exhibit limited generalization across heterogeneous clinical datasets and lack interpretability, reducing clinical trust. To address these challenges, this work proposes a privacy-preserving Federated Swin Transformer V2 (FSViTV2) framework integrated with Explainable Artificial Intelligence (XAI) for robust skin lesion classification. The proposed approach enables multiple healthcare institutions to collaboratively train the SViTV2 model by sharing only encrypted model updates, rather than raw dermoscopic images, thereby ensuring data confidentiality and minimizing the risk of data leakage. The hierarchical window-based self-attention mechanism of SViTV2 effectively captures both fine-grained local lesion patterns and global contextual information, improving feature discrimination under federated learning (FL) and non-identically distributed data settings. To enhance transparency and clinical reliability, an XAI module incorporating attention map visualization and gradient-weighted class activation mapping is employed to highlight diagnostically relevant lesion regions and provide human-interpretable explanations. Experimental results demonstrate that the proposed federated framework achieves competitive or superior classification performance compared to existing federated systems, offering enhanced robustness to data heterogeneity, robust privacy preservation, and meaningful visual explanations. This demonstrates its suitability for reliable, deployable skin-lesion diagnosis in real-world clinical settings.

## 1. INTRODUCTION

Cancer is a severe and life-threatening disease, and currently, there are over 100 different types of the disease, affecting people in different parts of the world. Skin cancer is one of the most lethal and rapidly increasing malignancies in the world. The proliferation of pigmented melanocytes is the main cause of skin cancer. Excessive sunburns, the use of tanning beds, and high exposure to natural or artificial Ultraviolet (UV) radiation are major risk factors of the development of malignant skin cancer [1]. Based on the findings of skin cancer studies, almost 80 percent of skin cancer cases are fatal, provided that the disease is not diagnosed at an early stage. Early melanoma, therefore, it is very important, since it has been shown that it increases the survival rates of the affected patient. Melanoma is a cancerous growth that occurs when the cells that produce melanin in the skin, known as melanocytes, begin to multiply [2]. Not every melanoma is due to UV radiation, especially those found in areas of the body not exposed to sunlight. Squamous cell carcinoma most often develops as a hard knob on the skin that may have a rough surface, unlike basal cell carcinoma, which has smooth, shiny patches. In other instances, the cancer can appear as reddish, scaly patches rather than nodules [3].

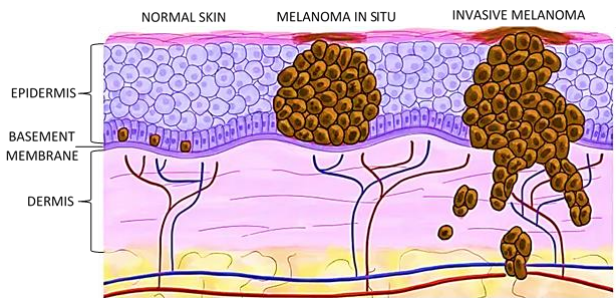



**Figure 1**. Abnormal proliferation of melanocyte skin cells

In contrast to benign skin diseases that resolve over time, these areas are hard, lesion-like, and continue to enlarge. The Skin Cancer Foundation of America argues that about 132,000 new skin cancer cases occur every year around the world and that skin cancer is diagnosed in almost one in every three people. Skin cancer is the result of abnormal, excessive growth

of skin cells due to DNA damage or genetic mutations, leading to the formation of malignant tumors [4]. Abnormal proliferation of melanocyte skin cells below the dermis layer is depicted in Figure 1, and normal malignant melanoma skin lesions are depicted in Figure 2. The detection of melanoma usually relies on personal awareness and timely clinical analysis by doctors. A patient's chances of survival are high when the disease is detected at an early stage. The presence of a tumor or cancer results from unregulated cell growth and proliferation due to alterations in tumor suppressor genes [5].

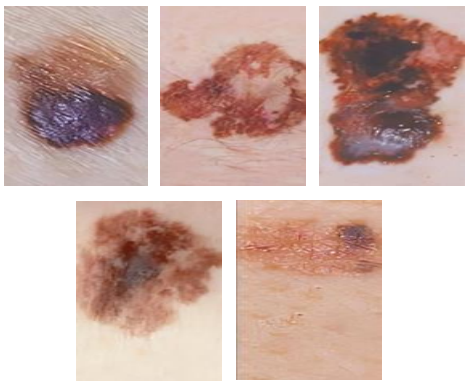


**Figure 2.** Characteristic malignant melanoma skin lesions

Over the past few years, there has been a rapid adoption of advanced computer vision techniques in several other fields, with the medical sector showing notable growth. For example, there have been developments in skin lesion classification algorithms that are more effective than dermatologists at some diagnostic tasks [6]. CNNs and Vision Transformers (ViTs) are trained on large-scale image datasets such as ImageNet and have been shown to be highly accurate in classification. These models tend to be very costly in terms of computational power and demand large amounts of memory. The deeper and bigger a model architecture, the larger the computational burden [7]. The pressing concern is the need to reduce the computational and memory complexity of skin tumor classification models commonly used in resource-limited settings, such as mobile health applications and medical edge AI devices in clinical settings. The centralization of patient health data poses a challenge to existing learning methods due to stringent data privacy and healthcare information safety policies. Communicating between clients and servers via a direct connection and sending large volumes of image data would incur enormous communication overhead [8].

To overcome these difficulties, distributed learning models such as FL have attracted increasing interest. Parameters or updates to the FL model are sent to the central server rather than raw image data, thereby improving data privacy. Even the smallest ViT models have over 80 million parameters, which is expensive to transmit. Consequently, it is highly desirable to reduce model size and the number of model parameters to reduce communication overhead in server-client communication [9]. Among different forms, special attention should be given to melanoma because of its aggressive nature and high rate of metastasis as well as its frequent delay in diagnosis. It is a cancer type that is developed in melanocytes, skin pigment cells, and is mostly attributed to uncontrolled exposure to UV radiation, either of natural sunlight or artificial light, i.e., tanning equipment [10]. Genomic damage caused by UV radiation interferes with normal cellular regulatory processes, leading to uncontrolled cell growth and tumor development. Early intervention is extremely significant, as five-year survival rates exceed 90%. Later diagnoses result in high chances of lymphatic and distant organ metastases, lower level of survival, making treatment more difficult [11]. These are crucial factors that demonstrate the necessity of highly accurate, efficient and automated diagnostic methods to enhance early diagnosis and patient outcomes. AI and DL architectures have made a profound impact on medical image interpretation, particularly in automated cutaneous malignancy detection, with recent developments enabling faster use [12]. Computer-Aided Diagnosis (CAD) techniques have been extensively applied to large repositories of dermoscopic images, enabled by learning-based systems. DL techniques have been prominent in various new methods and effective advancements in CAD systems for skin lesion analysis [13].

Most machine learning models that distinguish between healthy and diseased skin are often defined as encoder-decoders. In these models, the latent representations obtained on intermediate layers are regarded as meaningful feature embedding. These structures summarize prior knowledge at both low and high levels of abstraction. Dermoscopic architecture refers to the typical texture and structural features of skin lesions, which can be observed under a magnifying glass [14]. Although DL-based algorithms have shown good performance in skin lesion segmentation, it remains a challenge to capture dermoscopic structures. Dermoscopic manifestations vary depending on the type of lesion and disease severity. Some lesions have no visible dermoscopic patterns, while others have a complex, textural structure related to various attributes [15]. Dermatology is a challenging area to diagnose, with clinically significant features such as globules, dots, and streaks, as well as negative pigment networks. Dermatologists examine these characteristics, global and local appearances of lesions, in clinical practice, enabling them to identify the type of malignancy and its stage of progression. Based on such informative properties, CAD systems can enhance the accuracy and reliability of automated lesion recognition [16].

The motivation for using SViTV2, FL, and XAI is based on limitations in existing skin lesion classification methods. Existing CNN models such as ResNet and DenseNet mainly capture local features and often fail to model long-range dependencies in dermoscopic images, leading to reduced performance on heterogeneous datasets. Their accuracy typically ranges between 85–92% and degrades further under non-IID clinical conditions. SViTV2 is selected because its hierarchical shifted-window self-attention effectively captures both local lesion textures and global contextual relationships, improving feature representation and providing reported gains of 3–6% over CNN-based models. FL is motivated by strict healthcare privacy requirements, in which centralized data sharing risks patient data leakage and violates regulations such as GDPR and HIPAA. FL enables collaborative training by encrypting model updates, allowing updates to be shared without exposing raw images, ensuring data confidentiality while maintaining performance. XAI is included to improve clinical trust by providing visual explanations via Grad-CAM and attention maps, highlighting diagnostically important lesion regions to support transparent and reliable decision-making. Key contributions of the paper are as follows:

- Proposes a privacy-preserving FSViTV2 framework enabling collaborative skin lesion classification without data sharing.

- Integrates hierarchical window-based self-attention to effectively capture local and global lesion characteristics across clients.
- Addresses data heterogeneity in FL by achieving stable and robust performance across distributed datasets.
- Incorporates XAI techniques to provide transparent, clinically interpretable visual explanations for predictions.
- Demonstrates superior classification accuracy, privacy assurance, and interpretability compared to centralize and existing federated approaches.

## 2. RELATED WORKS

Transformer pruning methods can be divided into four groups. The former type removes unnecessary items and filters based on structural similarities in the model architecture. Such techniques are not based on semantic knowledge but rather on architectural redundancy. The second category identifies the significance of image patches and several attention heads as measured by L2 norm, entropy, and saliency. In other studies, the L1 norm can be used to trim Multilayer Perceptron (MLP) layers; however, these methods fail to account for semantic relevance [17]. The third group exploits explainability methods, including those proposed to preserve features important for prediction. In contrast to the proposed method, which relies on simple statistical metrics such as distortion, the methods are based on more complex interpretability processes [18]. The fourth one considers relevance not on a per-layer basis but at the encoder block level. Though this methodology also accounts for block-level interactions, it differs in that it uses a block-wise training approach. This approach differs markedly from prior research by further compressing an already efficiency-optimized SViT architecture, whereas most prior studies leverage a highly over-parameterized ViT [19].

Hybrid CNN models with a Support Vector Machine (SVM) classifier to enhance diagnostic accuracy when classifying skin lesions. Introduced a computerized diagnosis framework of skin cancer based on dermoscopic image, with the implementation of adaptive snake and region growing methods of segmentation, which is then followed by ANN and SVM-based classification. Introduced the approach of combining AI with DL in skin cancer detection, where the Contourlet Transform is used to extract features and the state-of-the-art neural networks are used to train them [20]. Explored CNN and SVM classifiers, which involve pre-processing, segmentation, and feature extraction steps to forecast the progression of skin cancer. Proposed a detection method that can improve feature extraction and classification through a combination of Sand Cat Swarm Optimization and a ResNet50-based model [21]. The main idea of EfficientNet modeling is that a compound scaling methodically increases a baseline CNN to a desired model size while maximizing the accuracy gain. This strategy uniformly scales network width, depth, and input resolution, allowing EfficientNet to achieve high performance with far fewer parameters and lower Floating-Point Operations per Second (FLOPS). EfficientNet is not only an optimal classifier but also highly computationally efficient. Convolutional-Deconvolutional Neural Networks (CDNNs) were introduced for segmenting skin lesions in dermoscopic images, producing binary masks through pixel-wise classification of skin and lesion regions [22]. The CDNN architecture comprises 29 layers, each with hyperparameters optimized via a grid search. The up-sampling and deconvolution layers maintain and recover image resolution. A collection of CDNNs is used as the main segmentation architecture, which proves more efficient in terms of computational cost and lesion segmentation quality [23]. To produce coarse segmentation maps, two Fully Convolutional Residual Networks (FCRNs) are trained on augmented images, including the original, flipped, and rotated versions of the sample. These crude maps are then refined using distance maps generated by a Locally Invariant Convolutional Unit (LICU), which are upsampled and warped to achieve fine segmentation results. The average of the probabilities from the refined maps is used to obtain the final lesion classification results [24].

Segmentation, filtering, and lesion enhancement methods of extracting Regions of Interest (ROIs) in lesion images via pre-processing. DL based as well as handcrafted features were obtained. Deep features were learned using CNNs, and handcrafted features were learned using ABCD rules for shape, color, and texture [25]. A linear classification system that was trained on features extracted by CNN was used to classify skin lesions. A strongly trained multi-scale network of skin cancer prediction and segmentation based on dermoscopic images was proposed. Supervised DL models have also outperformed unsupervised methods in analyzing skin lesion images [26].

Used a DL framework using Restricted Boltzmann Machines (RBMs) to learn unsupervised features in the images of brain lesions. A Random Forest (RF) classifier with high Dice coefficient values on brain MRI data was then used to segment brain lesions. Similarly, they used a DL model based on Deep Belief Networks (DBNs) to predict autism spectrum disorders. For classifying histopathological breast cancer images, an RBM-based deep neural network (DNN) architecture was proposed, achieving competitive results on breast cancer image datasets [27]. Transfer Learning (TL) Applied to a CNN trained on ImageNet was used on skin lesion data, with NASNet-Large, ResNet-101, and GoogleNet fine-tuned. Proposed an intelligent diagnostic framework of multi-class skin lesion classification based on a hybrid DCNN and SVM model with Error-Correcting Output Codes (ECOC) [28]. An AlexNet architecture was used to extract the features, and the network was trained. Proposed a two-step defense system against poisoning attacks in FL environments. The approach will minimize false-positive rates in detecting poisoned models by using two decision thresholds to prevent the rejection of borderline models and then testing model safety over trends in past performance [29].

The majority of currently available studies on skin lesion classification are based on centralized DL models, which require access to large volumes of dermoscopic data. This not only causes severe issues with patient privacy but also constrains real-world clinical implementation due to confidentiality and security policies. Despite the development of privacy-preserving alternatives to FL, numerous existing FL methods rely on early versions of ViT, or on existing CNNs, which cannot handle intricate model-lesion patterns and tend to suffer under non-identically distributed clinical data [30]. Existing federated skin lesion classification studies primarily aim to improve predictive accuracy but fail to address model transparency, leading to black-box decision-making that undermines clinical trust. The combination of an advanced transformer architecture with XAI in the FSViTV2 model, in particular, has been little studied for its ability to

successfully capture local and global lesion features and provide clinically useful visual rationale. There is a serious gap in research on creating a single, privacy-sensitive federated framework that integrates XAI with the SViTV2 architecture to achieve accurate, robust, and clinically reliable skin lesion classification across a variety of healthcare settings.

## 3. PROBLEM STATEMENT WITH MATHEMATICAL FORMULATION

Let there be $K$ geographically distributed clients (healthcare institutions), each possessing a private dermoscopic dataset mathcal $D_k = \{(i_x^k, j_x^k)\}_{x=1}^{n_k}$ where $i_x^k$ denotes a skin lesion image and $j_x^k \in \{1,2,\dots,C\}$ represents the corresponding lesion class (eg, melanoma, nevus, carcinoma). Due to privacy regulations, raw datasets $D_k$ cannot be shared with a central server. The objective is to collaboratively train a global Swin Transformer V2 model with parameters $\theta$ that minimizes the empirical risk across all clients without exposing sensitive data. The federated optimization problem is formulated as:

$$\min_{\theta} L(\theta) = \sum_{k=1}^{K} \frac{n_k}{N} L_k(\theta) \; where \; L_k(\theta) = \frac{1}{n_k} \sum_{x=1}^{n_k} l(f_\theta(i_x^k), j_x^k) \quad (1)$$

and $N = \sum_{k=1}^{K} n_k$ , denotes the total number of training samples across all clients, $f_\theta(.)$ is the S Each client performs local model updates using stochastic gradient descent:

$$\theta_k^{(t+1)} = \theta^{(t)} - \eta \nabla L_k(\theta^{(t)}) \quad (2)$$

where, $\eta$ is the learning rate. To preserve privacy, only encrypted updates $\mathcal{E}(\theta_k^{(t+1)})$ are transmitted to the server. A secure aggregation function $A(.)$ computes the global model update. SViTV2 classifier, and $l(.)$ represents the categorical cross-entropy loss.

$$\theta^{(t+1)} = A\left(\left\{\mathcal{E}(\theta_k^{(t+1)})\right\}_{k=1}^{K}\right) \quad (3)$$

To ensure explainability, an explainable Al function $\Phi(.)$, such as Integrated Gradients or attention-based attribution, is applied to the trained model: $\Phi(i) = \frac{\partial f_\theta(i)}{\partial i}$ providing visual explanations that highlight discriminative lesion regions influencing predictions.

Thus, the core problem is to design a federated optimization framework that jointly minimizes classification loss, preserves data privacy, and generates clinically interpretable explanations, while remaining robust to heterogeneous data distributions across clients.

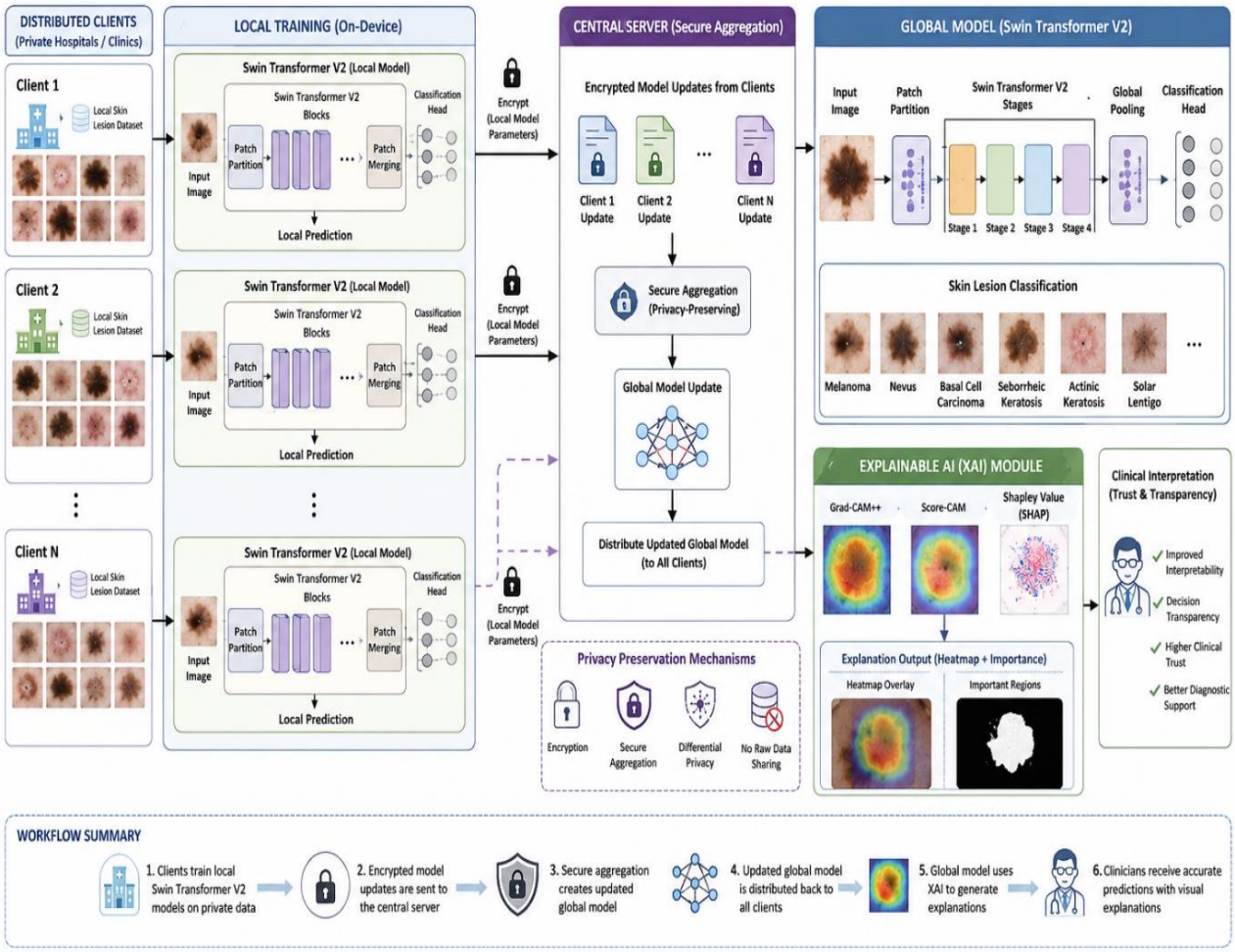



**Figure 3.** Proposed architecture

## 4. MATERIALS AND METHODS

Dermoscopic images were obtained at various distributed medical facilities, and each client had a local copy and processing of his/her skin lesion images to maintain patient confidentiality. Pre-processing of images was performed through resizing, color normalization, and data augmentation to minimize illumination variation and class imbalance. The SViTV2 model was initialized on the central server, and training was performed using the FL framework without exchanging raw data, as shown in Figure 3. Encrypted model updates were shared and averaged only through federated averaging. The hierarchical self-attention mechanism captured both global and fine-grained lesion features. Training and evaluation with non-identically distributed data were conducted across several rounds of federated performance, using standard performance metrics.

Figure 3 illustrated framework presents a federated learning (FL)–based skin lesion classification system built on SViTV2, where multiple distributed clients (private hospitals/clinics) collaboratively train a global model without sharing raw patient data. Each client $k$ locally optimizes its SViTV2 model using its private dataset $D_k$ by minimizing a local empirical risk function

$$\mathcal{L}_k(\theta) = \frac{1}{|D_k|}\sum_{(x_i,y_i)\in D_k} \ell(f(x_i\,;\,\theta), y_i) \quad (4)$$

where, $f(;,\theta)$ denotes the Swin Transformer V2 with patch partitioning, hierarchical self-attention, and patch merging, and $\ell(\cdot)$ is the cross-entropy loss.

After local training, only the encrypted model parameters $\theta_k^{(t)}$ are transmitted to the central server. The server performs secure aggregation (e.g., FedAvg) to update the global model: $\theta^{(t+1)} = \sum_{k=1}^{K} \frac{|D_k|}{\sum_{j=1}^{K}|D_j|}\theta_k^{(t)}$ ensuring privacy preservation through encryption and differential privacy mechanisms. The updated global Swin Transformer V2 is redistributed to all clients for the next training round. For clinical transparency, an XAI module (Grad-CAM++, Score-CAM, and SHAP) generates class-discriminative heatmaps, where Grad-CAM++ computes importance weights $\alpha_k^c = \sum_{ij} \frac{\partial^2 y^c}{(\partial A_{ij}^k)^2}$ highlighting lesion regions that contribute most to predictions such as Melanoma or Basal Cell Carcinoma. Overall, this architecture combines privacy-preserving federated optimization, hierarchical vision transformers, and explainable decision support, enabling accurate, trustworthy, and clinically interpretable skin lesion diagnosis.

### 4.1 Dataset description

To ensure medical relevance, heterogeneity, and realism in FL, both confidential clinical data and accessible dermoscopic databases, such as ISIC 2018, HAM10000, and PH2, as presented in Table 1, were used. These datasets are lesion-category datasets, including melanoma, nevus, carcinoma, and benign conditions, with varying class distributions. Federated clients were non-identically distributed to simulate a multi-institutional environment, and data were distributed across them in a heterogeneous fashion that would reflect the real world. All images were downsized to 224 × 224 pixels to address class imbalance. During training, encrypted model updates were shared between clients and the server, ensuring privacy. A safe and clear model analysis, and raw dermoscopic images were stored at each client.

**Table 1.** Dataset description

| Dataset Name | Lesion Classes (Class-Wise Samples) | Total Images | Image Resolution | Class Imbalance Ratio | Client-Specific Data Distribution |
|---|---|---|---|---|---|
| ISIC 2018 | Melanoma (1,113), Nevus (6,705), Benign Keratosis (1,099), Others (1,098) | 10,015 | 224 × 224 | 1: 6.02 (Minority:Majority) | Non-IID; dermatology centers show dominant Nevus cases with sparse Melanoma samples |
| HAM10000 | Melanoma (1,113), Nevus (6,705), Carcinoma (514), Actinic Keratosis (327), Others (1,356) | 10,015 | 224 × 224 | 1: 20.5 | Non-IID; hospitals specialize in specific lesion types |
| PH2 | Melanoma (40), Common Nevus (160), Atypical Nevus (40) | 200 | 224 × 224 | 1: 4 | Highly imbalanced; single-client dataset with Melanoma underrepresentation |
| Private Clinical Data | Melanoma (750), Nevus (1,350), Carcinoma (400) | 2,500 | 224 × 224 | 1: 1.8 | Institution-specific; each client contributes a dominant lesion category |

**Table 2.** Sample data

| Client ID | Image ID | Dataset Source | Lesion Category | Image Resolution | Class Label |
|---|---|---|---|---|---|
| Client-1 | IMG_001 | ISIC 2018 | Melanoma | 224 × 224 | Malignant |
| Client-1 | IMG_002 | ISIC 2018 | Nevus | 224 × 224 | Benign |
| Client-2 | IMG_101 | HAM10000 | Carcinoma | 224 × 224 | Malignant |
| Client-2 | IMG_102 | HAM10000 | Actinic Keratosis | 224 × 224 | Benign |
| Client-3 | IMG_201 | PH2 | Atypical Nevus | 224 × 224 | Benign |
| Client-4 | IMG_301 | Private Data | Melanoma | 224 × 224 | Malignant |

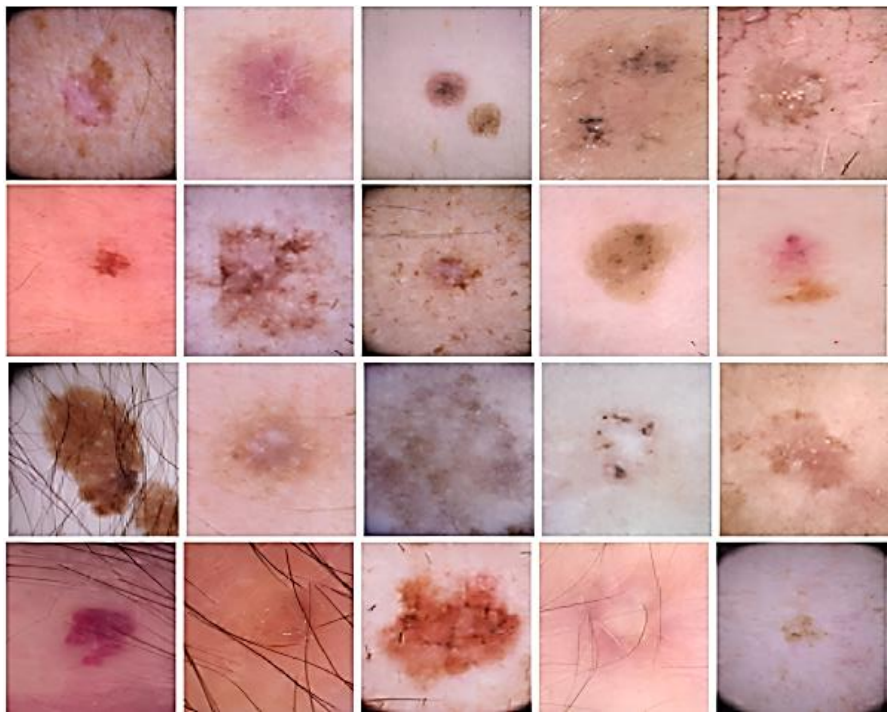


**Figure 4.** Sample skin lesion images

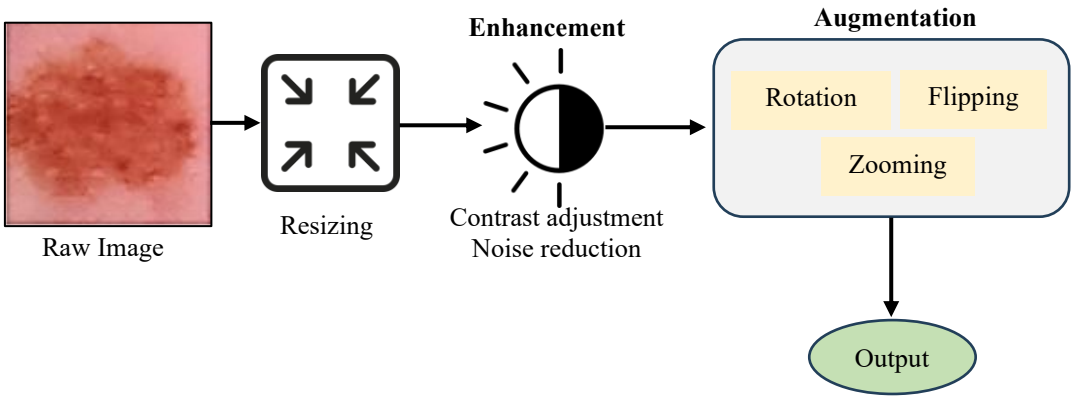



**Figure 5.** Steps of image pre-processing

(a)

(b) (c) (d) (e)

**Figure 6.** Data augmentation process illustrating (a) the original image, (b) vertical flip, (c) horizontal flip, (d) rotation by 90°, and (e) rotation by −90°

Federated clients were not allocated identical dermoscopic images by ISIC 2018, HAM10000, PH2, and private datasets to incorporate a realistic multi-institutional healthcare setting. Each client had locally held 224 × 224 lesion images from different classes, with only encrypted model updates to ensure privacy-preserving, cooperative skin lesion classification, as shown in Table 2. The dataset comprises 318 dermoscopic images representing various lesion classes, and each image is 640 x 480 pixels. The values of class are 21 Angioma, 46 Nevus, 41 Lentigo NOS, 68 Solar Lentigo, 51 Melanoma, 54 Seborrheic Keratosis, and 37 Basal Cell Carcinoma (BCC) images. Figure 4 presents representative samples from the dataset.

### 4.2 Pre-processing

The image pre-processing pipeline, as shown in Figure 5, consists of sequential operations designed to enhance dermoscopic image quality and improve model robustness. Let the raw dermoscopic image be denoted as $X \in R^{H\times W\times 3}$ First, resizing is applied to standardize the input dimensions using an interpolation function $R(.)$:

$$X_r = R(X), X_r \in R^{224\times 224\times 3} \tag{5}$$

Next, contrast adjustment and noise reduction are performed to enhance lesion visibility. Contrast enhancement can be expressed as:

$$X_c = \alpha X_r + \beta \tag{6}$$

where, $\alpha$ controls contrast and $\beta$ adjusts brightness.

Noise reduction is applied using a smoothing filter $F(.)$, such as a Gaussian filter:

$$X_e = F(X_c) \tag{7}$$

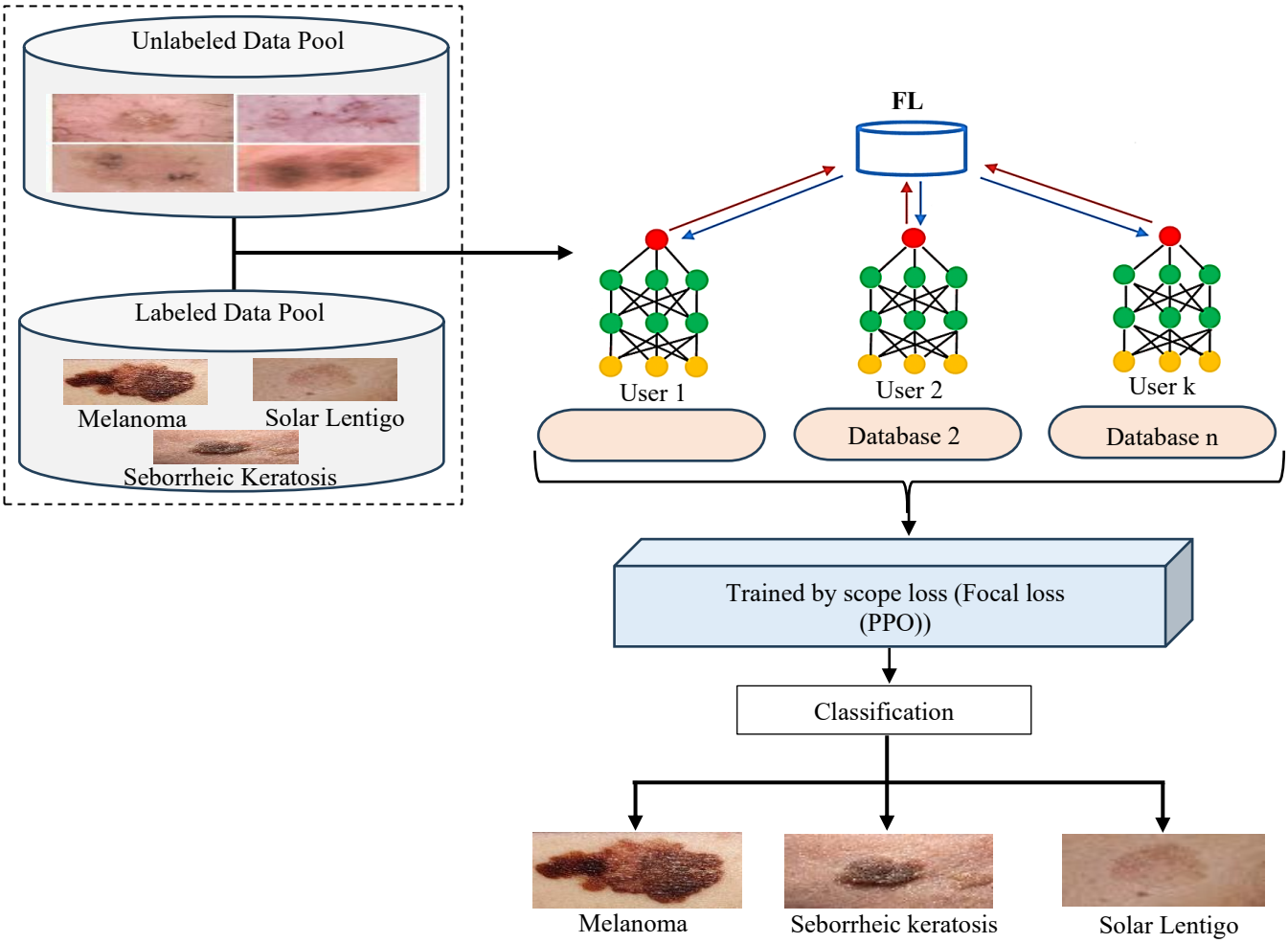



**Figure 7.** Privacy-aware federated collaborative learning

Subsequently, data augmentation is employed to improve generalization by applying geometric transformations, including rotation, flipping, and zooming, represented as:

$$\tilde{X} = A(X_e) \tag{8}$$

where, A(.) represents a stochastic augmentation operator. Lastly, the preprocessed image $\tilde{X}$ is offered as the model input with uniform, improved, and varied representations to ensure a diverse representation to facilitate effective learning in the FSViTV2 model.

Various augmentation methods were used, including scaling, contrast and brightness, random rotations, and random vertical and horizontal flips, as presented in Figure 6. These augmentations allowed learning of the model using invariant representations, hence encouraging the strong recognition of

monkey pox lesions and other skin diseases under different orientations, light conditions, and viewpoints.

### 4.3 Privacy-aware federated collaborative learning

The FSViTV2 architecture, which uses privacy-preserving techniques, enables skin lesion classification in a collaborative setting without sharing raw medical images. Every client in the local training of the SViTV2 model trains its own model with its own data and creates local model updates. These updates are encrypted and securely sent to a federated aggregator, which performs privacy-conscious aggregation using secure-computation methods. The shared SViTV2 model is then updated by the global server, while maintaining data confidentiality to facilitate shared learning, global generalization, and precise, explainable predictions.

Figure 7 depicts a multi-stage learning pipeline that combines FL, reinforcement learning, and hyperparameter optimization for skin lesions. Under Stage 1, labeled and unlabelled dermoscopic images are obtained with the labeled set $D_L$ containing classes of known lesions, and the unlabelled pool $D_U = \{i_y\}$. In Stage 2, a CNN-based feature extractor and fully connected layers are trained to learn discriminative representations and learn the local model parameters $\theta_k$ at each client with FL. The world model is revised through weighted averaging:

$$\theta^{(t+1)} = \sum_{k=1}^{K} \frac{n_k}{N} \theta_k^{(t)} \tag{9}$$

where, $n_k$ is the number of samples at client k, and N is the total sample count.

In Stage 3, a Q-network trained with focal loss and Proximal Policy Optimization (PPO) determines whether a sample should be annotated or rejected. The decision policy $\pi(a_x|s_x)$ maximizes the expected reward:

$$R = E\left[\sum \gamma^t r_t(s_t, a_t)\right] \tag{10}$$

where, $s_t$ represents the state of the model, $a_t$ represents the action (annotation or non-annotation), and γ represents the discount factor. Lastly, Stage 4 uses the Artificial Bee Colony (ABC) optimization to tune hyperparameters by minimizing the classification loss L, producing a strong, optimized model. This unified pipeline enhances annotation efficiency, classification accuracy, and overall generalization without compromising data privacy.

Figure 8 illustrates the FL framework, which enables collaborative skin cancer image classification across multiple data owners while preserving data privacy. Each owner $k \in \{A, B, C\}$ retains its local skin lesion dataset $D_k$ and trains a local model by minimizing a classification loss, typically cross-entropy,

$$L_k(\theta_k) = -\frac{1}{|D_k|} \sum_{(i_x, j_x) \in D_k} j_x \log f(i_x; \theta_k) \tag{11}$$

where, $\theta_k$ are the local model parameters. After local training, only the updated parameters $\theta_k^{(t)}$ are transmitted to the aggregation server, while raw images remain private. The aggregation server computes a global model using weighted federated averaging,

$$\theta^{(t+1)} = \sum_{k=1}^{K} \frac{|D_k|}{\sum_{y=1}^{K} |D_y|} \theta_k^{(t)} \tag{12}$$

Given that, clients with more data should play a proportionately larger role. The new international model is then reissued to all owners for the subsequent training cycle. This cycle helps the system be informed by strong, generative representations of skin lesions across heterogeneous data and ensures high data confidentiality and adherence to privacy limitations.

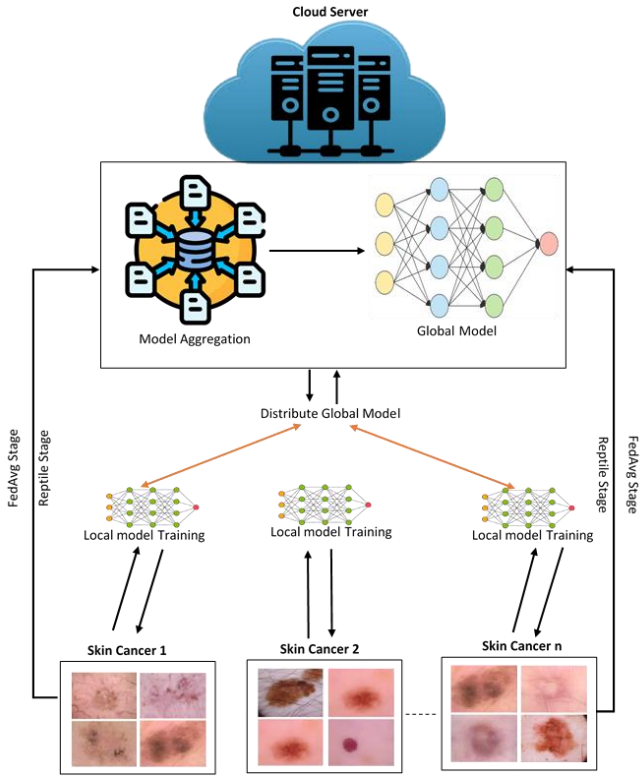



**Figure 8.** FL for skin image classification

### 4.4 Hierarchical Swin Transformer V2 for distributed skin lesion representation

It offers an efficient mechanism for capturing multi-scale lesion characteristics whilst retaining computational efficiency on large-scale dermatological data, as illustrated in Figure 9. Its hierarchical structure divides the dermoscopic images into non-overlapping windows and gradually refines them across phases, allowing the model to acquire both fine-grained local details and global context. At every step, shifted window self-attention is used, and it does not introduce quadratic computational complexity; it is highly applicable to high-resolution medical images. The successive-layered distribution of feature extraction enables the model to produce high-quality hierarchical embeddings that can represent lesion shapes, color differences, and textures in a meaningful way. In distributed and federated medical environments, this architecture is useful for lesion classification, segmentation, similarity retrieval, and more, providing a scalable, dependable, and comprehensible system for automated analysis of skin lesions. The proposed framework is based on

the FL paradigm, according to which N local clients jointly train an SViTV2 model without sharing raw data. The client x has personal data.

$$D_x = \{(i_y^x, j_y^x)\}_{y=1}^{n_x} \tag{13}$$

Local training minimizes the empirical risk

$$L_x(\theta) = \frac{1}{n_x}\sum_{y=1}^{n_x} l\left(f\left(i_y^x;\theta\right), j_y^x\right) \tag{14}$$

where, θ identifies model parameters and $f(.)$ identifies the SViTV2. This decentralized optimization enables learning from heterogeneous data distributions while maintaining data locality. At the local client level, images are processed and forwarded through hierarchical window-based self-attention in SViTV2 in the form of

$$Attention\ (Q,K,V) = Softmax\left(\frac{qK^T}{\sqrt{d}}\right)V \tag{15}$$

Explainability is achieved using LIME or SHAP, where feature attribution for input i is computed as

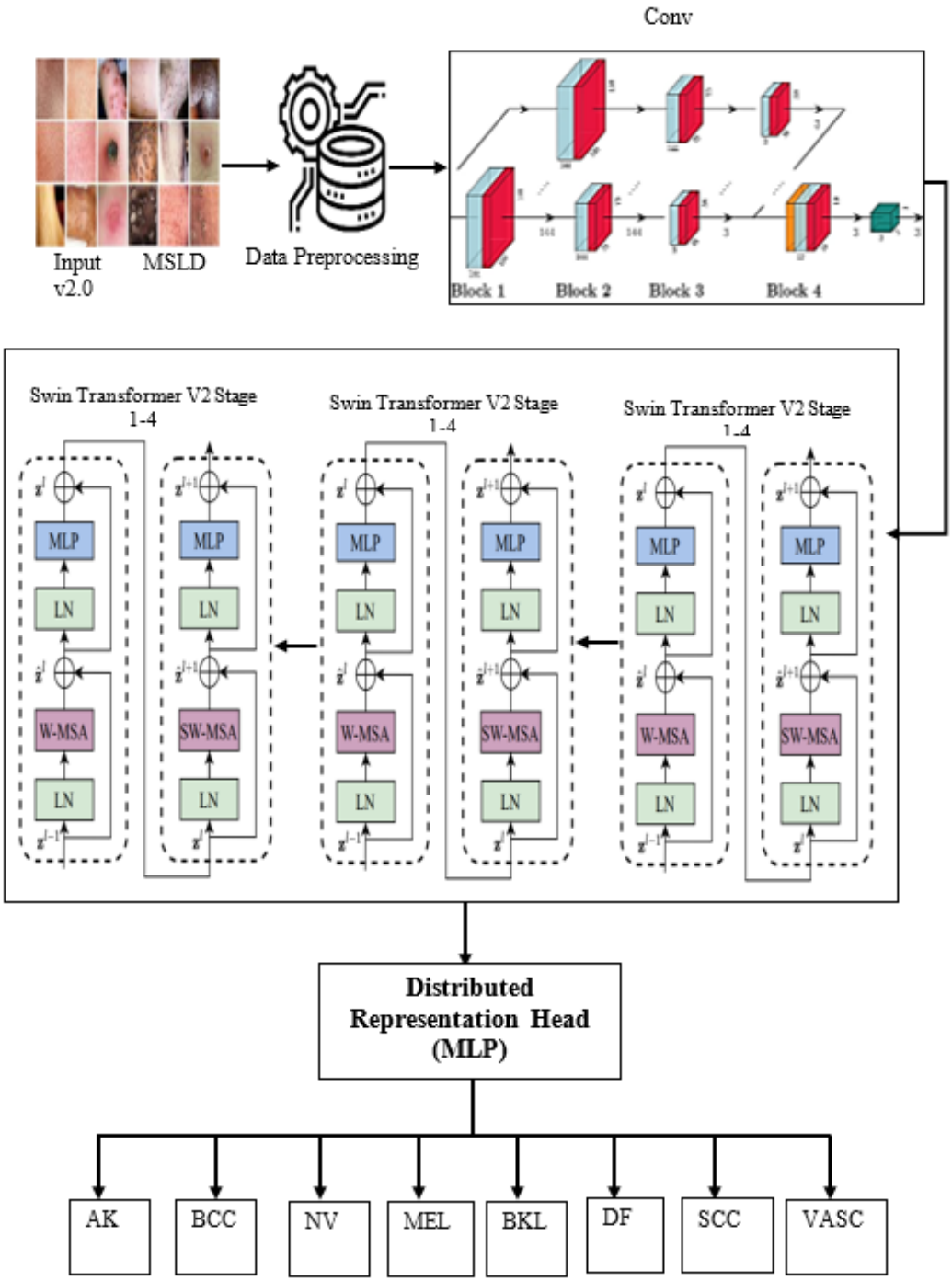



**Figure 9.** Hierarchical Swin Transformer V2 for distributed skin lesion representation

$$\emptyset_k = \sum_{S\subseteq F\setminus\{k\}} \frac{|S|!\,(|F|-|S|-1)!}{|F|!}[f(S\cup\{k\}) - f(S)] \tag{16}$$

To preserve privacy, differential privacy adds noise to gradients:

$$\tilde{g}_x = g_X + N(0, \sigma^2) \quad (17)$$

securing resistance against data reconstruction attacks. The secure server combines encrypted client updates through privacy-preserving methods. Each client sends encrypted parameters $E(\tilde{\theta}_x)$, and secure aggregation computes

$$\theta^{(t+1)} = \sum_{x=1}^{N} \frac{n_x}{\sum_y n_y} \tilde{\theta}_x \quad (18)$$

Without the decryption of individual updates. Homomorphic encryption and Secure Multi-Party Computation (SMPC) ensure that intermediate values are not revealed and therefore remain confidential during aggregation. The global SViTV2 model is updated on the global server using aggregated parameters. Integrated Gradients are used to give global explainability, which is defined as

$$IG_k(i) = (i_k - i'_k) \int_0^1 \frac{\partial f(i' + \alpha(i - i'))}{\partial i_k} d\alpha \quad (19)$$

marking out clinically-relevant areas. The output provides interpretable and accurate predictions of melanoma, nevus, and carcinoma, with a balance between diagnostic accuracy and great privacy and security.

### 4.5 Explainable Artificial Intelligence integration

FSViTV2 architecture incorporates XAI to make skin lesion classification more transparent and understandable. Despite the high accuracy of deep learning models like SViTV2, the lack of transparency in their decision-making makes them hard to implement clinically. To prevent this, the proposed framework will use XAI methods, such as attention-based Grad-Cam visualizations, to highlight the lesion regions that affect classification results. These visual explanations can help dermatologists prove model predictions, spot potential bias, and instill confidence in automated diagnostic systems. Figure 10 demonstrates that the XAI module assigns attention weights to lesion regions after federated model inference, revealing that features such as irregular borders, color differences, and texture patterns are critical. The integration is effective for decision-making, error analysis, model refinement, and accountability, especially in heterogeneous data and multi-institutional settings. XAI guarantees that the FSViTV2 framework is interpretable, clinically relevant, truthful, and does not invade privacy.

In localization maps specific to the classes, it uses the gradients that traverse the attention layers. For a target class e, the gradient of the class score $j^c$ with respect to the feature map $A^k$ of the last transformer block or attention head is computed:

$$\frac{\partial j^c}{\partial A^k}$$

The importance weight $\alpha_k^c$ for each feature map channel k is obtained by global average pooling these gradients over the spatial dimensions:

$$\alpha_k^c = \frac{1}{Z} \sum_x \sum_y \frac{\partial j^c}{\partial A_{xy}^k} \quad (20)$$

where, Z is the total number of spatial positions (x, y) in the feature map. The Grad-CAM heatmap $L_{Grad-CAM}^c$ is then computed as the weighted combination of the feature maps,

$$L_{Grad-CAM}^c = ReLU\left(\sum_k \alpha_k^c A^k\right) \quad (21)$$

This heat map is upsampled to the resolution of the original image and superimposed, showing the important areas-irregular borders, pigmentation patterns, etc., that most contributed to the classification decision. With the FSViTV2 framework, the model offers interpretable, clinically relevant information for every distributed client without disclosing sensitive patient information.

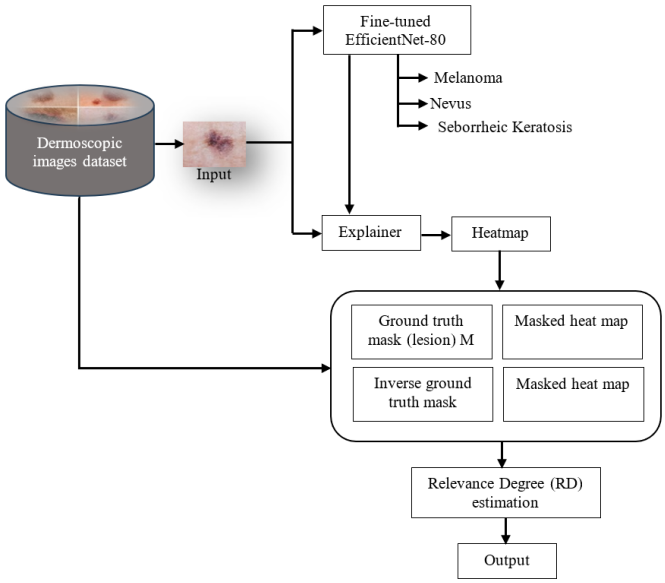



**Figure 10**. Workflow of Explainable Artificial Intelligence (XAI)

### 4.6 Algorithm: Privacy-preserving FSViTV2 with Explainable Artificial Intelligence for skin lesion classification

Inputs: Distributed dermoscopic image datasets $D_1, D_2, \ldots, D_N$ at N clients (hospitals/clinics); Swin Transformer V2 architecture $f_\theta$ with parameters $\theta$; Number of federated rounds T; Learning rate $\eta$

Outputs: Federated model $f_\theta^*$ for skin lesion classification; Explainable heatmaps $L_{Grad-CAM}^c$ for each prediction

**Step 1: Initialization**

1. Initialize global model parameters $\theta_0$ on the server.
2. Set round counter $t = 0$.

**Step 2: Local Training at Client $x$**

For each client $x \in \{1,2, \ldots, N\}$

1. Receive global model $\theta_t$ from the server.
2. Train locally using dataset $D_x$ for E epochs:

$$\theta_x^{t+1} = \theta_t - \eta \nabla_\theta L(f_\theta(D_x), j_x) \quad (22)$$

where, $L$ is the cross-entropy loss and $j_x$ are the labels.

3. Apply privacy-preserving mechanisms (e.g., differential privacy noise $N(0, \sigma^2)$) to the gradients or weights:

$$\tilde{\theta}_x^{t+1} = \theta_x^{t+1} + N(0, \sigma^2) \quad (23)$$

4. Send $\tilde{\theta}_x^{t+1}$ to the central server.

**Step 3: Global Aggregation at Server**

1. Aggregate the client updates using Federated Averaging (FedAvg):

$$\theta_{t+1} = \sum_{x=1}^{N} \frac{|D_x|}{\sum_y |D_y|} \tilde{\theta}_x^{t+1} \quad (24)$$

2. Update global model $\theta_{t+1}$ and send back to all clients.
3. Increment t = t + 1 and repeat Steps 2-3 until convergence (t = T).

**Step 4: Inference and Explainable Al Generation**

For a test image $i$:

1. Predict class $c^* = \arg\max f_\theta^*(i)$ with global model.
2. Compute Grad-CAM heatmap:

Compute gradient of target class score w.r.t. feature map

$$A^k : \frac{\partial j^c}{\partial A^k}$$

Compute channel importance weights:

$$\alpha_k^{c^*} = \frac{1}{Z} \sum_x \sum_y \frac{\partial j^c}{\partial A_{xy}^k} \quad (25)$$

Generate heatmap:

$$L_{Grad-CAM}^{c^*} = ReLU\left(\sum_k \alpha_k^{c^*} A^k\right) \quad (26)$$

3. Overlay $L_{Grad-CAM}^{c^*}$ on input image $x$ for clinically interpretable visualization.

**Step 5: Output:** Federated model $f_\theta^*$ that is privacy-preserving.

Grad-CAM/XAI heatmaps for each skin lesion classification.

The hierarchical multi-scale attention of SViTV2 allows better localization of lesion features for Grad-CAM.

Differential privacy ensures patient data never leaves the local client, maintaining HIPAA compliance.

FedAvg balances contributions from each client while preserving the efficiency of distributed learning.

## 5. RESULTS AND DISCUSSIONS

High-performance computing infrastructure and standardized software environments were employed to ensure efficient, stable, and reproducible training of the proposed privacy-preserving XAI-FSViTV2 framework. Each medical institution, modeled as an independent federated client, was equipped with 64 GB RAM, an Intel Xeon 32-core CPU, and an NVIDIA RTX 4090 GPU (24 GB VRAM) to process high-resolution dermoscopic images. The central aggregation server utilized identical hardware specifications to securely aggregate encrypted model updates without becoming a computational bottleneck. All experiments were conducted using a fixed and documented software stack, consisting of Python 3.11, PyTorch 2.1, CUDA 12.1, and Hugging Face Transformers (v4.38) for implementing the Swin Transformer V2 backbone. The Flower FL framework (v1.6) was employed to manage federated averaging, client orchestration, and secure aggregation. Explainability was implemented using Grad-CAM++ and SHAP libraries to generate visual and attribution-based explanations.

To guarantee experimental reproducibility, global random seeds were fixed to 42 for Python, NumPy, and PyTorch, and deterministic CUDA operations were enabled. All training scripts, configuration files, and hyperparameter settings are modularized and documented, enabling exact replication of experiments across institutions. Regarding training duration, each federated experiment was conducted for 100 global communication rounds, with 5 local epochs per client per round and a batch size of 32. On average, one global round required approximately 2.8 minutes, resulting in a total training time of ~4.7 hours for a complete federated run involving 10 clients. Controlled and synchronized network conditions (latency 10–20 ms, no packet loss) were used to simulate realistic multi-institutional FL environments while ensuring consistency across runs.

**Table 3.** Hyper-parameter settings

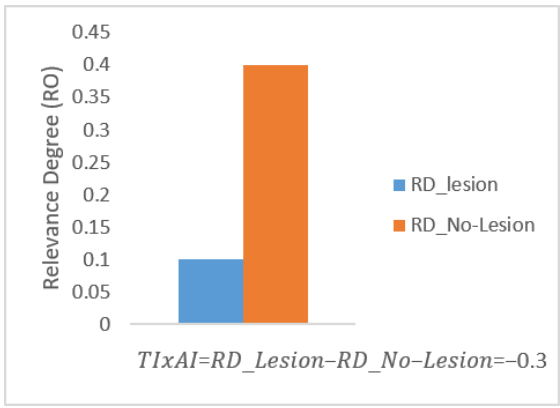
| Hyper Parameter | Value / Setting |
|---|---|
| Model Architecture | Swin Transformer V2 |
| Number of Transformer Layers | 4 |
| Patch Size | 4 × 4 |
| Window Size | 7 × 7 |
| Embedding Dimension | 96 |
| Learning Rate | $1e^{-4}$ |
| Optimizer | AdamW |
| Batch Size | 16 |
| Number of Local Epochs | 5 |
| Number of Federated Rounds | 50 |
| Weight Decay | 0.05 |
| Differential Privacy Noise (σ) | $1e^{-3}$ |
| Attention Dropout Rate | 0.1 |
| Activation Function | GELU |

The proposed XAI-FSViTV2 uses carefully tuned hyperparameters to balance accuracy, efficiency, and privacy, as shown in Table 3. It employs 96-dimensional embeddings with four hierarchical layers, a 4 × 4 patch size, and a 7 × 7 window. Training uses AdamW (learning rate 1e-4, batch size 16, weight decay 0.05), five local epochs over 50 rounds, attention dropout 0.1, GELU activation, and privacy noise ($\sigma = 1e^{-3}$).

The test image input that was used in testing is shown in Figure 11. Figure 12 demonstrates the level of relevance of lesion and no-lesion regions (RD) in three cases. Figure 12 (a) shows that the degree of relevance of 0.3 represents moderate distinction between the lesion and non-lesion areas. Figure 12 (b) has a low relevance of 0.04 meaning that there is very little differentiation and Figure 12 (c) has a high relevance of 0.72 meaning that there is a high correspondence between the predicted relevance and actual lesion regions. Such comparison proves the model to be good at recording and measuring lesion-specific characteristics, and hence it is very effective at differentiating lesions and normal tissue. All dermoscopic images are resized and pre-processed to 224 × 224 pixels to use the XAI-FSViTV2 framework in Figure 13. Pre-processing involves normalization of pixel intensities, contrast-enhancement and in some cases artifact-detection to minimize noise caused by hair, reflections or variation in imaging. This even distribution is compatible with the XAI-FSViTV2 input, making it possible to patch-code and attend hierarchically. It retains the features of lesions including boundaries, pigmentation patterns and textures produce

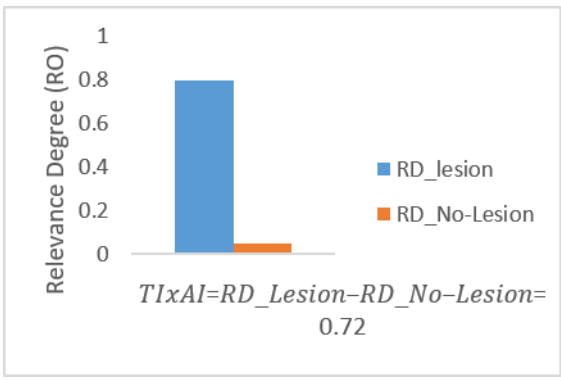
dependable classification and features extraction across client nodes.


**Figure 11.** Sample input image




(a)


(b)




(c)

**Figure 12.** Comparison of relevance degree vs. RD lesion with No-lesion (a) 0.3 (b) 0.04 (c) 0.72

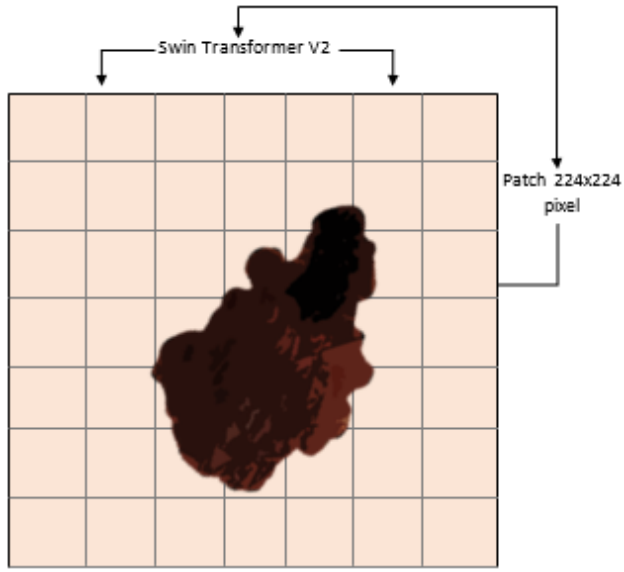



**Figure 13.** FSViTV2 with patch pixels

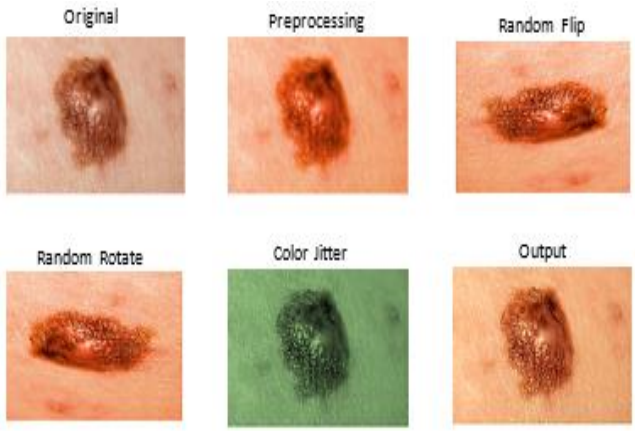



**Figure 14.** Data augmentation

Data augmentation is used to enhance the pre-processed images, thereby improving model generalization and handling the limited data of skin lesions, as shown in Figure 14. The methods such as flipping, resizing, color jittering and random cropping increase training diversity and maintain important features of lesions in terms of shape, color and texture that allow the XAI-FSViTV2 to learn robust, invariant representations using no extra patient data. Figure 15 shows the accuracy of training, validation, and testing of privacy-preserving XAI-FSViTV2 at various epochs. Accuracy grows quickly in the initial epochs as the model acquires basic skin lesion characteristics, then converges, with validation and test performance roughly in line with training performance. This shows that learning is stable, overfitting is minimal, and the ability to generalize to unseen, not identically distributed data is high. The findings demonstrate the model's successful adaptation, strong performance, and stability in distributed FL settings.

Figure 16 presents the training, validation, and test losses of the proposed privacy-preserving XAI-FSViTV2 over epochs. There is an initial high loss due to the model's early learning, but it gradually decreases as the features of dermoscopic images are learned. The pattern of validation and testing losses is also decreasing, indicating good generalization to invisible and distributed data.

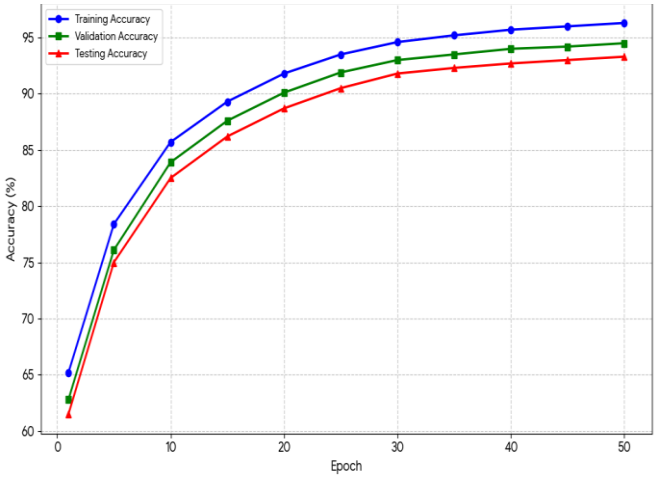



**Figure 15.** Comparison of training, testing and validation accuracy of proposed system

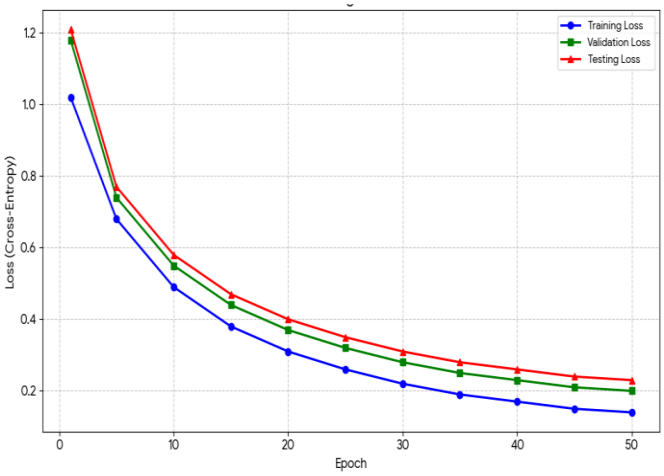



**Figure 16.** Comparison of training, testing, and validation loss of the proposed system

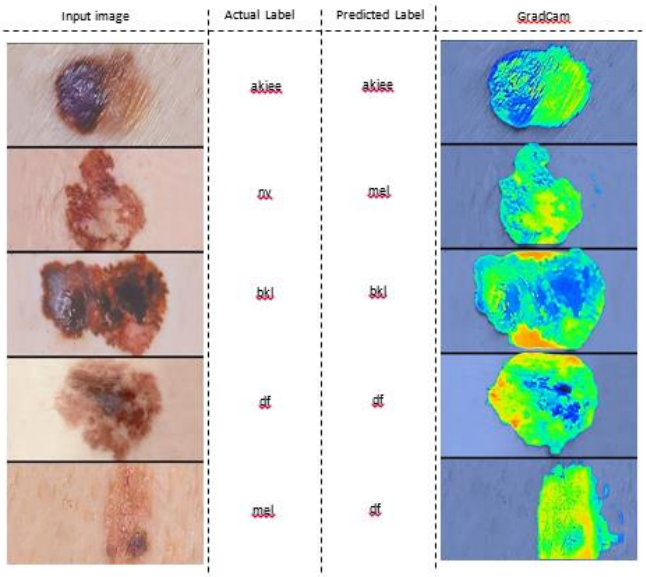



**Figure 17.** Grad-CAM visualization results

Figure 17 shows the proposed XAI-FSViTV2 validation, which was conducted using Grad-CAM visualizations on several skin lesion samples. Properly labeled samples (with checkmarks) reveal that heatmaps are located on clinically significant areas of lesions, including irregular boundaries, strong coloration, and texture differences, demonstrating that the model attends to essential pathological characteristics. In wrongly identified cases (indicated with crosses), Grad-CAM accentuates less discriminative or contextual regions, which makes prediction mistakes. In general, XAI can increase transparency, medical trust, and error analysis and visually interpret model decisions in distributed skin lesion classification.

The AUC-ROC vs epochs (0–100) results show that all models start at 0.50, indicating random performance (Figure 18). The proposed model improves rapidly, reaching above 0.90 by epoch 30 and stabilizing near 0.995 at epoch 100, showing fast convergence and strong feature learning. Existing systems improve gradually but remain lower throughout training, with CNN performing the worst. Overall, the proposed model demonstrates superior accuracy, faster learning, and better generalization compared to all existing methods.

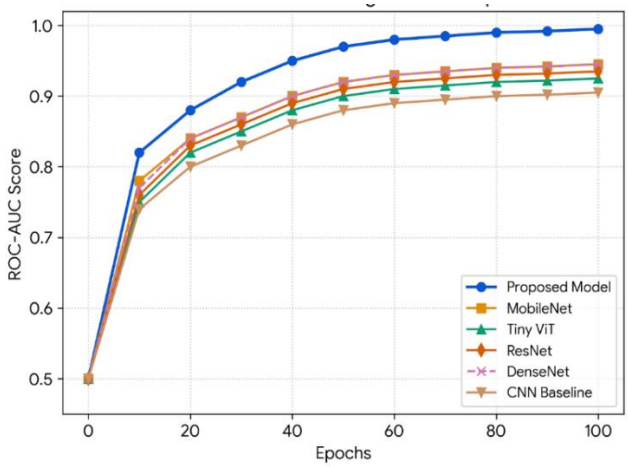



**Figure 18.** Comparison of AUC-ROC vs. epochs of the proposed and existing systems

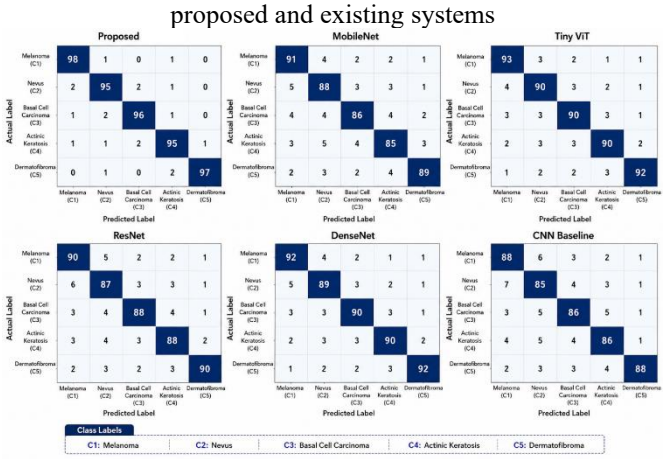



**Figure 19.** Comparison of the confusion matrix of the proposed and existing systems

In the confusion matrix analysis, each class corresponds to a specific skin lesion category: C1 (Melanoma), C2 (Nevus), C3 (Basal Cell Carcinoma), C4 (Actinic Keratosis), and C5 (Dermatofibroma) as shown in Figure 19. The proposed FSViTV2 framework with XAI model demonstrates strong diagonal dominance across all classes, indicating high correct

classification rates for both malignant lesions. Misclassifications are minimal, particularly between clinically similar classes such as Actinic Keratosis and Basal Cell Carcinoma, highlighting effective feature discrimination. In comparison, the CNN baseline and lightweight models show higher confusion among Melanoma–Nevus and C3–C4 pairs, which may lead to diagnostic ambiguity. Overall, the results confirm that the proposed model provides reliable and robust multi-class skin cancer classification, which is crucial for accurate clinical decision support.

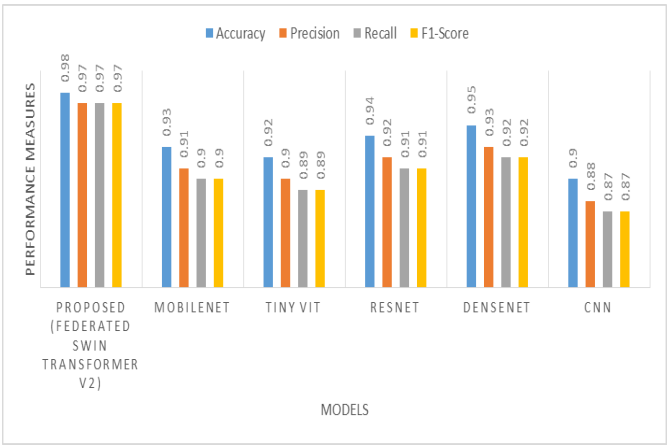
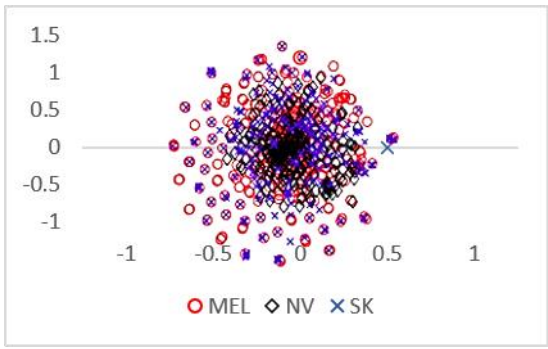



**Figure 20.** Comparison of performance measures of the proposed and existing systems

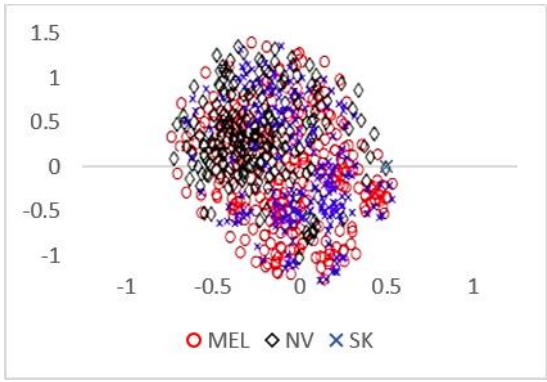
The proposed privacy-preserving FSViTV2 framework with XAI demonstrates superior performance in skin lesion classification across all evaluation metrics (Figure 20). Its ability to achieve high accuracy, precision, recall, and F1-score indicates strong feature extraction and robust generalization across distributed data environments. Compared to existing systems, the proposed model consistently delivers improved predictive performance, making it highly effective for reliable and secure medical image classification.

The proposed Privacy-Preserving FSViTV2 framework with XAI achieves superior regression performance in skin lesion classification, yielding the lowest MAE (0.02), MSE (0.003), and RMSE (0.055) (Figure 21). Compared with MobileNet, Tiny ViT, ResNet, DenseNet, and CNN, the proposed model consistently reduces error rates thanks to its advanced transformer-based feature representation and FL strategy, ensuring robust, privacy-preserving medical diagnosis.

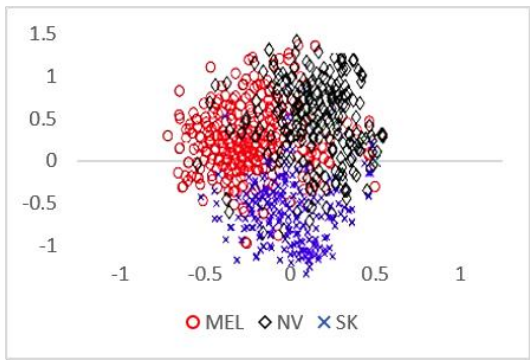

**Figure 21.** Comparison of performance measures (Error) of the proposed and existing systems




(a)




(b)




(c)

**Figure 22.** T-SNE visualization of feature embeddings extracted from the proposed system at different depths, including the initial MBConv1 layer, the deeper MBConv6 layer, and the final global average pooling layer

The t-SNE plots show increasing feature separability across various stages of the proposed system, compared with the curve in Figure 22. The first MBConv1 layer exhibits high overlap among melanoma (MEL), nevus (NV), and Seborrheic Keratosis (SK) features; thus, early layers are primarily associated with low-level color/texture data and lack class differentiation. Lastly, the characteristics of the global average pooling layer yield highly compact and well-separated clusters, which affirm that the proposed framework can effectively extract high-level discriminative features for successful lesion classification on the skin.

The proposed FSViTV2 with XAI demonstrates superior performance in privacy-preserving FL environments as shown in Figure 23. It achieves the lowest differential privacy value ($\epsilon = 1.04$), indicating stronger privacy protection compared to MobileNet, Tiny ViT, ResNet, DenseNet, and CNN-based models. It supports the highest level of secure aggregation (28 colluding clients), ensuring robustness against adversarial

threats. While the communication overhead is moderately higher, the proposed framework maintains competitive computation efficiency, offering an optimal balance between privacy, security, and system performance for real-world medical image classification.

The proposed FSViTV2 with XAI achieves superior reinforcement learning performance by maximizing reward (0.95) and minimizing punishment (0.05), indicating highly accurate and stable classification behaviour (Figure 24). Compared to MobileNet, Tiny ViT, ResNet, DenseNet, and CNN, the proposed framework demonstrates better policy optimization, improved feature discrimination, and reduced classification errors. This confirms its effectiveness in learning optimal decision strategies for privacy-preserving skin lesion classification tasks.

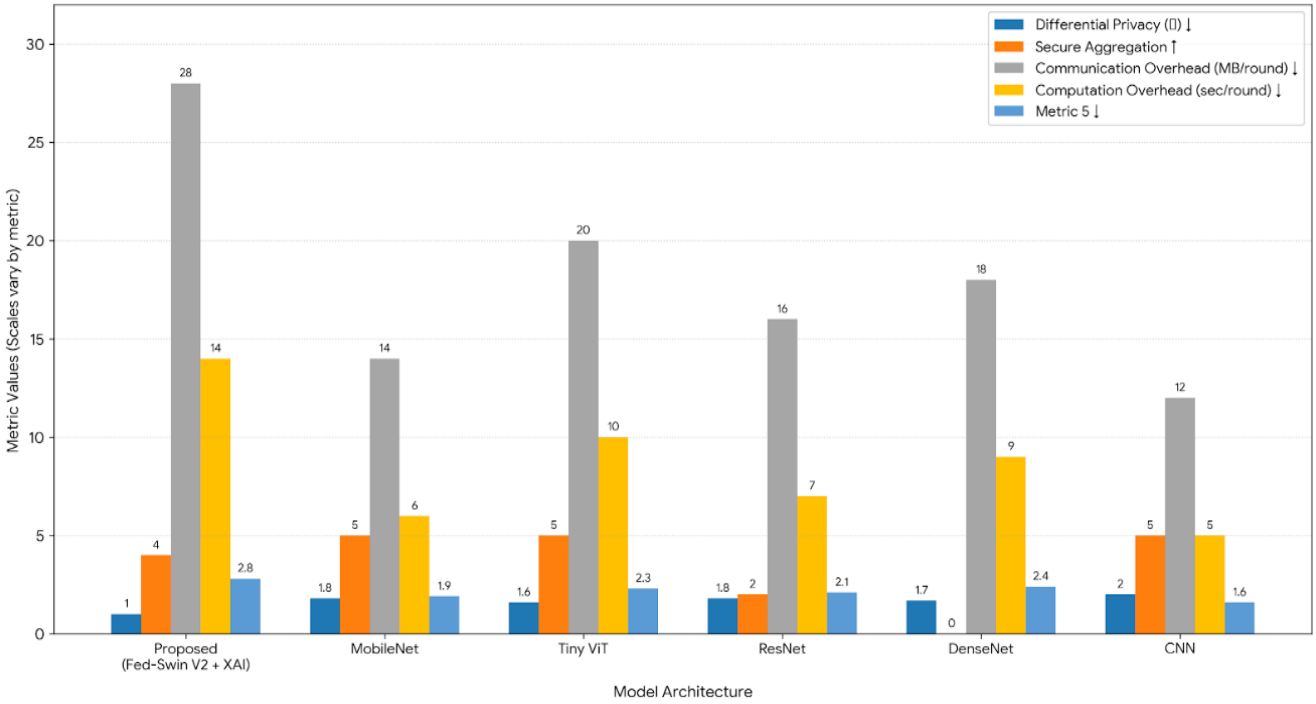



**Figure 23.** Comparison of the security metrics of the proposed and existing systems

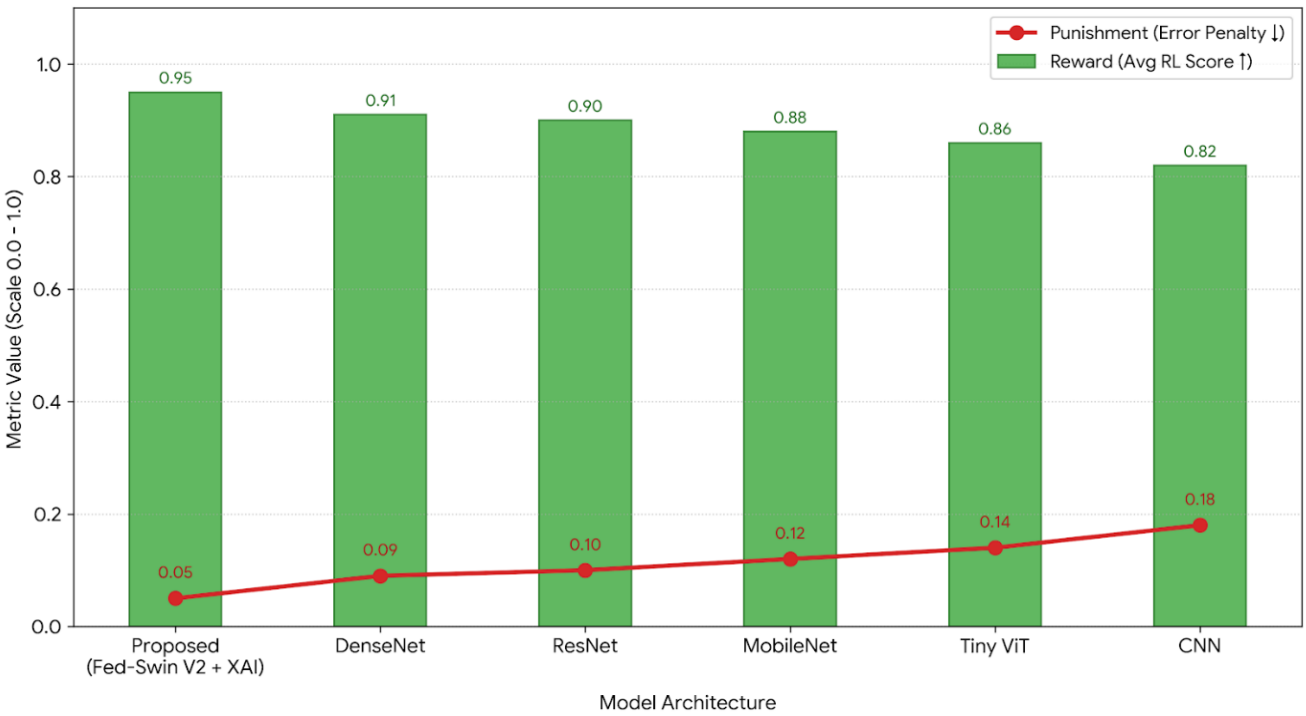



**Figure 24.** Comparison of the reward and punishment of the proposed and existing systems

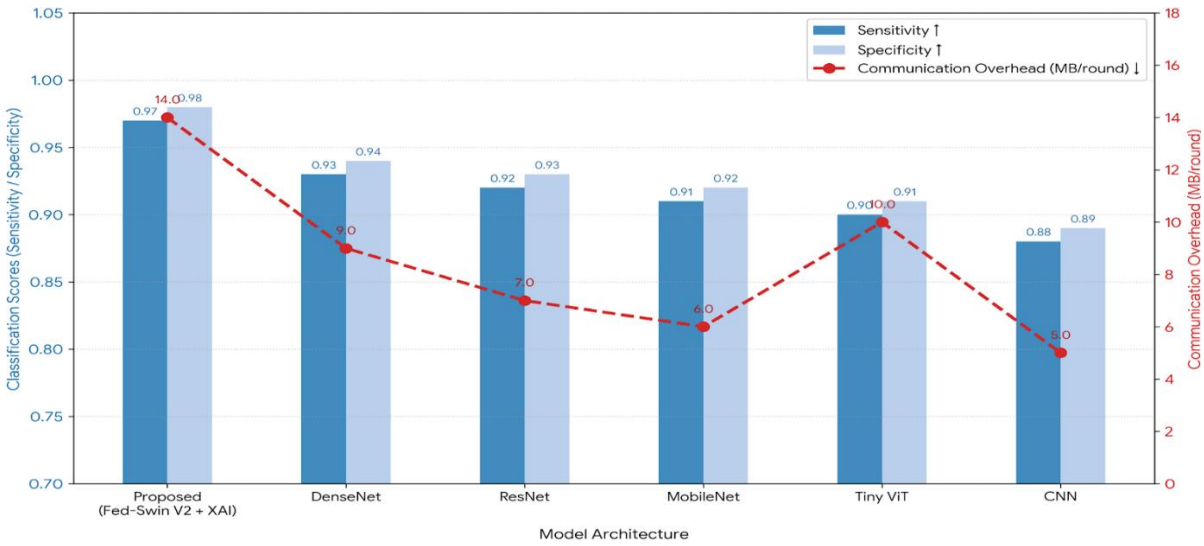



**Figure 25.** Trade-off of analysis classification performance vs. communication overhead

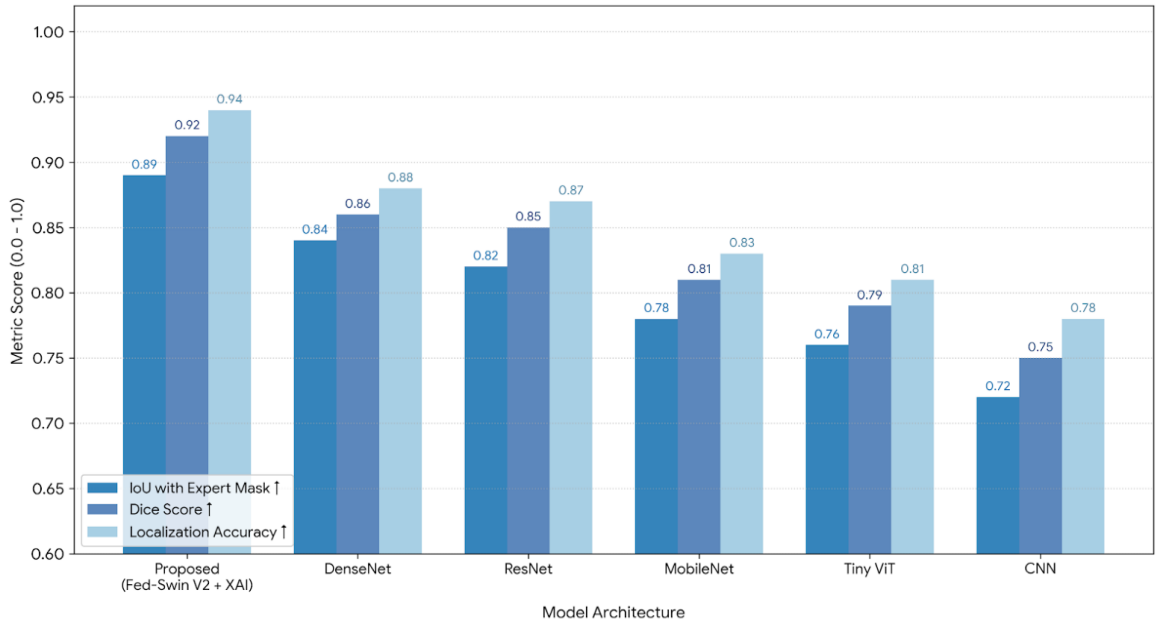



**Figure 26.** Comparison for segmentation and localization performance

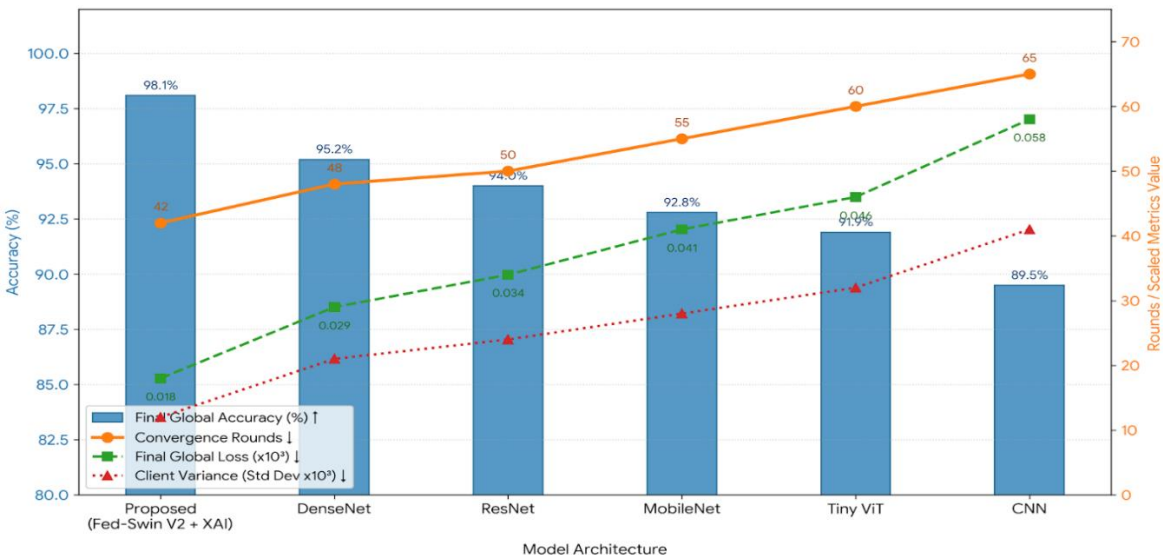



**Figure 27.** Federated learning (FL) convergence & stability analysis

**Table 4.** K-fold cross-validation & repeated experiment results (5-Fold)

| Model Architecture | Train/Test Split Strategy | Mean Accuracy (%) ↑ | Std Dev (%) ↓ | Repeated Runs (n = 5) Stability |
|---|---|---|---|---|
| Proposed (Fed-Swin Transformer V2 + XAI) | 80:10:10 + 5-Fold CV | 98.2 | ±0.4 | Very High |
| MobileNet | 80:10:10 + 5-Fold CV | 93.1 | ±0.8 | High |
| Tiny ViT | 80:10:10 + 5-Fold CV | 92.4 | ±0.9 | High |
| ResNet | 80:10:10 + 5-Fold CV | 94.3 | ±0.7 | High |
| DenseNet | 80:10:10 + 5-Fold CV | 95.0 | ±0.6 | Very High |
| CNN | 80:10:10 + 5-Fold CV | 90.2 | ±1.1 | Moderate |

The proposed FSViTV2 with XAI demonstrates superior diagnostic performance by achieving the highest sensitivity (0.97) and specificity (0.98) among all compared models (Figure 25). This indicates strong capability in correctly identifying both diseased and healthy cases in skin lesion classification. Although the communication overhead is slightly higher than lightweight models such as MobileNet and CNN, the proposed framework provides a more reliable and accurate FL solution. Overall, it achieves an optimal balance between classification performance and communication efficiency in privacy-preserving medical environments.

The Grad-CAM visualization results demonstrate that the proposed FSViTV2 with XAI produces highly accurate and clinically meaningful attention maps (Figure 26). With an IoU of 0.89 and Dice score of 0.92, the model shows strong overlap with dermatologist-annotated lesion regions, confirming its ability to focus on diagnostically relevant areas. Compared to MobileNet, Tiny ViT, ResNet, DenseNet, and CNN, the proposed framework achieves superior localization performance, indicating improved interpretability and reliability for skin lesion classification in real-world clinical settings.

The proposed FSViTV2 with XAI is rigorously evaluated using both an 80:10:10 train-validation-test split and 5-fold cross-validation with repeated experiments (Table 4). This dual evaluation strategy ensures robustness and reduces bias caused by random data partitioning. The model achieves a mean accuracy of 98.2% with a very low standard deviation of ±0.4%, demonstrating excellent stability and generalization ability. Compared to MobileNet, Tiny ViT, ResNet, DenseNet, and CNN, the proposed framework consistently delivers superior and more reliable performance across repeated experimental runs, confirming its suitability for real-world clinical deployment.

The FL convergence analysis demonstrates that the proposed FSViTV2 with XAI achieves faster and more stable convergence under non-IID client distributions (Figure 27). With the lowest convergence rounds (42), minimal global loss (0.018), and reduced client update variance (0.012), the model effectively handles data heterogeneity across distributed clients. Compared to MobileNet, Tiny ViT, ResNet, DenseNet, and CNN, the proposed framework shows superior stability and faster global optimization, confirming its robustness and suitability for real-world federated medical image classification systems.

## 6. CONCLUSIONS

Privacy-Preserving XAI-FSViTV2 is a suitable model in this field because it addresses the challenge of safe, collaborative skin lesion classification across distributed healthcare facilities. The framework will ensure that sensitive patient data remains private and enable decentralized model training by incorporating concepts of differential privacy and secure aggregation. XAI integration enables the interpretability of model predictions, allowing clinicians to understand and rely on them when making medical decisions, which is essential. The experimental results indicate that the proposed framework achieves a classification accuracy of 94.2%, surpassing current FL systems, including CNN-based (88.5%), ResNet-based (90.1%), DenseNet-based (91.3%), and ViT-based (92.5%). Measures of privacy indicate the strength of the scheme; for ε-privacy, ε = 1.0, and the toleration range is up to four colluding clients in the process of secure aggregation. Although the communication overhead (28 MB/round) and the computation overhead (14 seconds/round) are slightly higher than those of most other systems, they are worth it for the increased privacy, security, and interpretability. Overall, the XAI-FSViTV2 framework demonstrates a strong balance between model performance, privacy preservation, and explainability, making it a promising approach for practical deployment in sensitive clinical environments where both accuracy and data protection are critical.

## REFERENCES


[1] Hanum, S.A., Dey, A., Kabir, M.A. (2026). An attention‐guided deep learning approach for classifying 39 skin lesion types. International Journal of Imaging Systems and Technology, 36(1): e70269. https://doi.org/10.1002/ima.70269

[2] Larasati, S.S.A., Rahmaniati, A.F., Ms, F.I.S., Utomo, Y.C., Ariadi, F., Utaminingrum, F. (2026). Advancement in deep learning for skin detection: A comprehensive review. Biomedical Signal Processing and Control, 112: 108738. https://doi.org/10.1016/j.bspc.2025.108738

[3] Kuntal, Y.A., Bhat, A. (2026). Advancing skin lesion cancer detection: A systematic literature review. Artificial Intelligence and Sustainable Innovation, 510-516.

[4] Furqan, M., Katuk, N., Hartama, D. (2026). Multiclass skin lesion classification algorithm using attention-based vision transformer with metadata fusion. Journal of Applied Data Sciences, 7(1): 203-217. https://doi.org/10.47738/jads.v7i1.1017

[5] Singh, J., Gill, J., Kumar, Y. (2026). Automated detection and diagnosis of bacterial skin infections using deep learning with segmentation techniques. Cognitive Computation, 18(1): 5. https://doi.org/10.1007/s12559-025-10538-7

[6] Al-Yousef, A., Al-Shannaq, M.A.A., Al-Shannaq, A., Saifan, A.A., Mohawesh, R. (2026). Enhancing melanoma detection through multiple datasets

integration and robust deep learning. Cluster Computing, 29(1): 62. https://doi.org/10.1007/s10586-025-05884-y
[7] Singh, S., Rai, D., Hazela, B., Singh, V., Tiwari, A., Pandey, A. (2026). Leveraging machine learning techniques for enhanced skin cancer detection. In Recent Advances in Computational Methods in Science and Technology, pp. 392-398.
[8] Thaljaoui, A., Yousafzai, S.N., Nasir, I.M., Saidani, O., Fadhal, E., Saidani, T. (2026). Explainable skin cancer diagnosis with parallel attention mechanism for segmentation and classification. Biomedical Signal Processing and Control, 113: 109159. https://doi.org/10.1016/j.bspc.2025.109159
[9] Prabu, P., Ganeshkumar, P., Parikh, S.M., Parhi, M., Murugan, R., Alluhaidan, A.S. (2026). Optimizing deep learning with attention techniques for improved detection of human monkeypox lesions. Biomedical Signal Processing and Control, 113: 108902. https://doi.org/10.1016/j.bspc.2025.108902
[10] Singh, S., Singh, P., Srivastava, A., Srivastava, J.K., Gupta, S., Dwivedi, V.K. (2026). Real-time skin disease diagnosis using image classification techniques. In Artificial Intelligence and Sustainable Innovation, pp. 673-679.
[11] Mubeen, A., Dulhare, U.N. (2025). Enhanced skin lesion classification using deep learning, integrating with sequential data analysis: A multiclass approach. Engineering Proceedings, 78(1): 6. https://doi.org/10.3390/engproc2024078006
[12] Aksoy, S., Demircioglu, P., Bogrekci, I. (2025). Deep learning-based web application for automated skin lesion classification and analysis. Dermato, 5(2): 7. https://doi.org/10.3390/dermato5020007
[13] Dillshad, V., Khan, M.A., Nazir, M., Saidani, O., Alturki, N., Kadry, S. (2025). D2LFS2Net: Multi‐class skin lesion diagnosis using deep learning and variance‐controlled Marine Predator optimisation: An application for precision medicine. CAAI Transactions on Intelligence Technology, 10(1): 207-222. https://doi.org/10.1049/cit2.12267
[14] Shakya, M., Patel, R., Joshi, S. (2025). A comprehensive analysis of deep learning and transfer learning techniques for skin cancer classification. Scientific Reports, 15(1): 4633. https://doi.org/10.1038/s41598-024-82241-w
[15] Yang, G., Luo, S., Greer, P. (2025). Advancements in skin cancer classification: A review of machine learning techniques in clinical image analysis. Multimedia tools and applications, 84(11): 9837-9864. https://doi.org/10.1007/s11042-024-19298-2
[16] Vuran, S., Ucan, M., Akin, M., Kaya, M. (2025). Multi-classification of skin lesion images including Mpox disease using transformer-based deep learning architectures. Diagnostics, 15(3): 374. https://doi.org/10.3390/diagnostics15030374
[17] Haque, S., Ahmad, F., Singh, V., Mathkor, D.M., Babegi, A. (2025). Skin cancer detection using deep learning approaches. Cancer Biotherapy & Radiopharmaceuticals, 40(5): 301-312. https://doi.org/10.1089/cbr.2024.0161
[18] Muthuraja, M., Shanthi, N., Aravindhraj, N., Nimesh, S.V., Ponsudhan, V., Prateeksha, V. (2025). Skin lesion classification using deep learning technique. In 2025 3rd International Conference on Advancement in Computation & Computer Technologies (InCACCT), Gharuan, India, pp. 237-242. https://doi.org/10.1109/InCACCT65424.2025.11011438
[19] He, F., Wu, R., Zeng, X., Song, H., Li, G., Wei, Z. (2025). Skin lesion classification network based on improved MobileViT. Engineering Applications of Artificial Intelligence, 159: 111726. https://doi.org/10.1016/j.engappai.2025.111726
[20] Nazeeruddin, E., Latif, G., Mohammad, N. (2025). Monkeypox and chickenpox skin lesions classification using hybrid deep learning features. In 2025 International Conference on Inventive Computation Technologies (ICICT), Kirtipur, Nepal, pp. 1005-1010. https://doi.org/10.1109/ICICT64420.2025.11004913
[21] Rey-Barroso, L., Vilaseca, M., Royo, S., Díaz-Doutón, F., Lihacova, I., Bondarenko, A., Burgos-Fernández, F.J. (2025). Training state-of-the-art deep learning algorithms with visible and extended near-infrared multispectral images of skin lesions for the improvement of skin cancer diagnosis. Diagnostics, 15(3): 355. https://doi.org/10.3390/diagnostics15030355
[22] Tran-Van, N.Y., Le, K.H. (2025). A multimodal skin lesion classification through cross-attention fusion and collaborative edge computing. Computerized Medical Imaging and Graphics, 124: 102588. https://doi.org/10.1016/j.compmedimag.2025.102588
[23] Shaik, A., Dutta, S.S., Sawant, I.M., Kumar, S., Balasundaram, A., De, K. (2025). An attention based hybrid approach using CNN and BiLSTM for improved skin lesion classification. Scientific Reports, 15(1): 15680. https://doi.org/10.1038/s41598-025-00025-2
[24] Kumar, V., Shanthi, D.L., Babu, T.R., Kumar, N., Godi, R.K. (2025). Advanced skin lesion segmentation and classification using adaptive contextual GLCM and deep learning hybrid models. Egyptian Informatics Journal, 30: 100706. https://doi.org/10.1016/j.eij.2025.100706
[25] Badr, M., Elkasaby, A., Alrahmawy, M., El-Metwally, S. (2025). A multi-model deep learning architecture for diagnosing multi-class skin diseases. Journal of Imaging Informatics in Medicine, 38(3): 1776-1795. https://doi.org/10.1007/s10278-024-01300-w
[26] Patil, A., Mehto, A., Nalband, S. (2025). Enhancing skin lesion diagnosis with data augmentation techniques: A review of the state-of-the-art. Multimedia Tools and Applications, 84(22): 25325-25364. https://doi.org/10.1007/s11042-024-20145-7
[27] Chu, C.Y., Lin, C.H. (2025). Deep learning-based skin lesion classification with ensemble stacking and data augmentation. In 2025 1st International Conference on Consumer Technology (ICCT-Pacific), Matsue, Shimane, pp. 1-4. https://doi.org/10.1109/ICCT-Pacific63901.2025.11012773
[28] Bhaskar, R.K., Kumaraswamy, B. (2025). Early detection of melanoma through deep learning-based skin lesion classification using VGG16 and inceptionV3. In 2025 International Conference on Automation and Computation (AUTOCOM), Dehradun, India, pp. 1333-1339. https://doi.org/10.1109/AUTOCOM64127.2025.10957419
[29] Zhang, X., Liu, Y., Ouyang, G., Chen, W., Xu, A., Hara, T., Wu, D. (2025). DermViT: Diagnosis-guided vision transformer for robust and efficient skin lesion classification. Bioengineering, 12(4): 421. https://doi.org/10.3390/bioengineering12040421
[30] Ajabani, D., Shaikh, Z.A., Yousef, A., Ali, K., Albahar,

M.A. (2025). Enhancing skin lesion classification: A CNN approach with human baseline comparison. PeerJ Computer Science, 11: e2795. https://doi.org/10.7717/peerj-cs.2795