Monkeypox Disease Classification Based on Skin Images Using Hierarchical Swin Transformer-Based Convolutional Neural Network Approach

Putu Desiana Wulaning Ayu[1,*], Sasiwimol Sukket[2], Sutikno[3], Putu Manik Prihatini[4], Gede Angga Pradipta[5], Dandy Pramana Hostiadi[6]

[1,4]*Information Technology Department, Politeknik Negeri Bali, Indonesia*

[2]*Faculty of Industrial Technology Valaya Alongkorn Rajabhat University Pathum Thani, Thailand*

[3]*Department of Informatics, Faculty of Science and Mathematics, Diponegoro University, Indonesia*

[5,6]*Department of Magister Information System Institut Teknologi dan Bisnis STIKOM Bali, Indonesia*

(Received: November 15, 2025; Revised: January 10, 2026; Accepted: March 15, 2026; Available online: April 26, 2026)

Abstract

Monkeypox diagnosis can initially be conducted through expert physical examination based on characteristic lesions. However, laboratory confirmation using PCR is still essential, these tests are often hampered by limitations such as high costs, lengthy processing times, and a general lack of detailed symptom knowledge among patients. In light of these issues, image-based diagnostic methods offer a more efficient solution, given that monkeypox manifests as visible lesions on the skin that can be accurately detected using a deep learning. This study employs Transformer network-based deep learning for classifying skin diseases. To improve model robustness and mitigate the limitations of the relatively small dataset, we designed a comprehensive data augmentation pipeline that incorporates both positional and color transformations, including rotation, horizontal and vertical flipping, zooming, shearing, and brightness, contrast, hue, and saturation adjustments. Furthermore, a k-fold cross-validation strategy was employed, where the entire dataset was partitioned into k equal-sized folds to ensure a reliable and unbiased evaluation of the model performance. The Swin Transformer leverages advanced transformer network to analyze images, emphasizing hierarchical relationships within images. Swin Transformer enhances the convolutional Transformer architecture by substituting the standard multi-head-self-attention (MSA) mechanism with a shifted window-based MSA module It enhances efficiency over traditional transformer models by incorporating a shifted window mechanism, which reduces computational demands. The average global accuracy achieved was 0.99 (99%), which is further supported by the AUC values obtained for each disease category. The model achieved an AUC of 1.00 for chickenpox, cowpox, and hand-foot-mouth disease (HFMD), indicating excellent discriminative capability for these classes. Meanwhile, the remaining classes, including healthy skin, measles, and monkeypox, achieved AUC values of 0.99 and 0.98, respectively. These results demonstrate that the proposed Hierarchical Swin Transformer model provides highly reliable classification performance across all skin disease categories included in the dataset.

Keywords: Classification, Monkeypox, Skin Image, Hierarchical Swin Transformer Network, Deep Learning

1. Introduction

In recent years, the global health community has been challenged by the emergence and spread of monkeypox, a zoonotic disease that can be transmitted from animals to humans and poses a potential risk of widespread outbreaks. The first cluster of cases reported outside endemic regions was identified in the United Kingdom on May 7, 2022, involving a patient who had recently traveled from Nigeria. Following this initial case, additional infections were reported in several non-endemic countries across different WHO regions, including Europe, the Americas, and the Western Pacific [1]. As the outbreak progressed, the World Health Organization (WHO) continued to monitor and report the global spread of the disease, highlighting the increasing public health concern associated with monkeypox. Indonesia also reported its first confirmed monkeypox case in 2022 involving an international traveler returning from

***Corresponding author: Putu Desiana Wulaning Ayu (wulaning_ayu@pnb.ac.id)**
DOI: https://doi.org/10.47738/jads.v7i2.1313

ISSN 2723-6471

abroad [2] . These developments underscore the importance of early detection and reliable diagnostic tools to support rapid identification and prevent further transmission. This has raised concerns among both the public and governmental authorities, particularly due to the potential for disease transmission in densely populated regions. The rash develops into lesions with the stages of macules, papules, vesicles, and pustules. Diagnosis of monkeypox can be made through a physical examination by an expert who sees the typical lesions. However, laboratory diagnosis using PCR tests is also necessary to confirm infection [3]. Several studies have shown the use of image-based artificial intelligence technology and expert systems as a solution for the diagnosis of monkeypox. However, the use of expert systems is less effective because sufferers generally do not know in detail the symptoms that arise[4]. Therefore, another solution using images is more appropriate, because monkeypox causes lesions on the skin that can be recognized through images. The deep learning convolutional neural network (CNN) technique has shown good performance in image-based research, including in the diagnosis of monkeypox.

Convolutional neural networks (CNNs) have demonstrated strong performance in many medical image classification tasks. Several previous studies have applied CNN architectures such as VGG-19 and ResNet50 for monkeypox detection, achieving classification accuracies of 93.33% and 82.96%, respectively [5], [6]. However, deeper CNN architectures often contain a large number of parameters, which may increase computational complexity and the risk of overfitting when trained on relatively small datasets. For instance, the widely used ResNet50 architecture contains approximately 25 million trainable parameters, which can require substantial computational resources and careful regularization to avoid overfitting in limited medical datasets. These limitations have motivated researchers to explore alternative architectures such as Transformer based models that can capture long range dependencies and hierarchical representations more efficiently [7]. In particular, the Swin Transformer introduces a shifted window self-attention mechanism that enables efficient feature learning while maintaining linear computational complexity with respect to image size. Several studies have successfully applied these AI models to medical images such as computed tomography (CT) scans and X-ray images, claiming performance improvements in their detection processes with Transformer models using TransUnet and SwinUnet algorithms [8]. Other research has also claimed increased accuracy with Transformer models for MRI images, specifically for osteosarcoma [9]. In the field of remote sensing, the use of Swin Transformer enhanced with a Dynamic High-Pass Preservation module showed significant improvements in sharpening techniques, producing images with finer details and more complete spectral information [10]. Research Peng et al [11] introduced modifications to the Swin Transformer that focus on spatial feature extraction for spectral analysis. In recent years, Transformer based architecture has gained increasing attention in computer vision tasks due to their ability to model long range dependencies and capture global contextual information. Models such as Vision Transformer (ViT) and Swin Transformer have demonstrated strong performance not only in medical imaging applications but also in other complex visual recognition tasks. These advances suggest that Transformer-based architectures have significant potential for improving the classification of skin lesion images, including monkeypox detection.

However, laboratory diagnosis via PCR testing is necessary to confirm the infection. Despite this, expert systems have limitations, as patients often do not fully understand or identify their symptoms, leading to challenges such as inaccurate test results, high costs, and extended wait times. As an alternative, using image based methods offers a more viable solution, as monkeypox manifests in skin lesions that can be effectively recognized through the application of deep learning models. This study focuses on developing a reliable framework for detecting patients with monkeypox using skin images obtained with Transformer model. The novelty and contributions of this study are stated as follows:

First, we design a dedicated data augmentation pipeline that combines extensive positional and color transformation including rotation, horizontal and vertical flipping, zoom, shear, height and width shifting, brightness, contrast, hue, and saturation jittering, as well as appropriate fill modes to address the strong variability of web-collected images and to enlarge the limited dataset. Second, to mitigate the risk of overfitting on a small dataset and to rigorously assess robustness, we adopt a k-fold cross validation strategy in which entire dataset is split into k-equal size fold, and finally we compare Swin Transformer, Vision Transformer, VGG16 and Modified Resnet50, and demonstrate that the proposed Transformer-based model achieve competitive accuracy with lower and smaller standard deviations across folds, indicating more stable and reliable performance. And the third, we develop a reliable deep learning framework for detecting monkeypox patients from skin images using a hierarchical Swin Transformer-based architecture

ISSN 2723-6471

specifically tailored for skin-image classification. Unlike conventional CNN-based models and plain Vision Transformers used in previous monkeypox studies, our framework exploits two key design characteristics. First, it utilizes a multistage representation with progressive patch merging that constructs a hierarchical feature representation across multiple resolution levels (56×56 → 28×28 → 14×14 → 7×7). This hierarchical design enables the model to capture both low level texture information and high-level semantic patterns in skin lesion images. Second, the architecture incorporates shifted window multi-head self-attention mechanisms (W-MSA and SW-MSA), which allow efficient modeling of long-range spatial dependencies while maintaining linear computational complexity with respect to image size.

The remainder of the research paper is structured as follow: Section 2 presents the literature review; section 3 is proposed methodology. Section 4 covers the experiment result, while Section 5 provides the conclusion along with future work.

2. Literature Review

Deep learning has been employed by various researchers to support the diagnosis of the Monkeypox and skin disease[12]. Ahsan et al [13] used VGG-16 to diagnose Monkeypox from skin lesion images. They also created a data set called 'Monkeypox2022'. The accuracy, sensitivity, recall and F1-score obtained by the model were 97%, 97%, 97% and 97%, respectively. To interpret the results the local interpretable model explainer (LIME) was utilized. Abdelhamid et al. [14] proposed a monkeypox detection framework that utilizes the AI-Biruni Earth Radius Optimization algorithm combined with GoogLeNet for feature extraction. The study reported a maximum classification accuracy of 98.8%, along with additional evaluation metrics including sensitivity and F1-score. These results demonstrate the potential of deep learning approaches for monkeypox detection from skin lesion images.

A mobile application was developed to diagnose Monkeypox from skin lesion images [15]. Java and android were used to develop the application. A maximum accuracy of 91.11% was obtained. The sensitivity, recall and F1-score obtained were 85%,94% and 89%, respectively. Islam et al. [16] used deep learning to diagnose the monkeypox virus. The data set contained images of monkeypox, chickenpox, smallpox, cowpox and measles. Seven different classifiers were utilized in this research. The accuracy, sensitivity, recall and F1-score were 83%, 85%, 94% and 89%. Sitaula et al. used pretrained DL classifiers to diagnose the monkeypox. Eight different classifiers were used to distinguish the four classes. A maximum accuracy, sensitivity, recall, and F1-score of 87.13%, 85%, 85% and 85% were obtained. Bala D et all [17] proposed Monkey-Net Architecture for monkeypox detection and classification and reached an accuracy of 93.19%.

Based on previous research, this study proposed a deep learning model using Transformer-based approach, specifically the Swin Transformer. Furthermore, the performance of this model is compared with that Convolutional Neural Networks (CNNs), including the hyperparameter tuning. Yolcu Oztel et al [18] design a quantitative and objective classification tool that merges two skin-lesson datasets (PAD-UFES-and MLSD) to form a seven-class problem including monkeypox. The study compared a state-of-the-art Vision Transformer with several popular CNN architecture under a transfer-learning regime. Result showed that Vision transformer can match or surpass CNN performance even with limited data when carefully fine-tuned, higliht the promise of Transformer based models for complex multi class dermatology task. Fei Ma Jian et al [19] proposed a leverage a multiscale inflated convolutional feature fusion and attentional Swin-Unet approach. In this method, a multiscale extended convolution module is employed in the coding stage of the Swin-Unet network to enhance complementary features while preserving different features at different scales. Chakroborty [20] developed a framework allows us to construct hybrid deep learning models combining deep learning architectures as a feature extraction tool with machine learning classifiers and perform a comprehensive analysis of Mpox detection from image data, and best performing model consists of MobileNetV2 with LightGBM classifier achives an accuracy of 91.49%. Vuran et al [21] investigated several Transformer-based architectures for multi-class classification of skin-lesion images, explicitly including a class for mpox. Using models such as ViT and Swin Transformer they compared performance across multiple metrics and reported that Transformer models can achieve strong accuracy 93.10% and stable standard deviation values across folds.

ISSN 2723-6471

3. Methodology

The proposed method is illustrated in figure 1, where develop a reliable deep learning framework based on a Swin Transformer architecture to detect monkeypox patient from skin images.

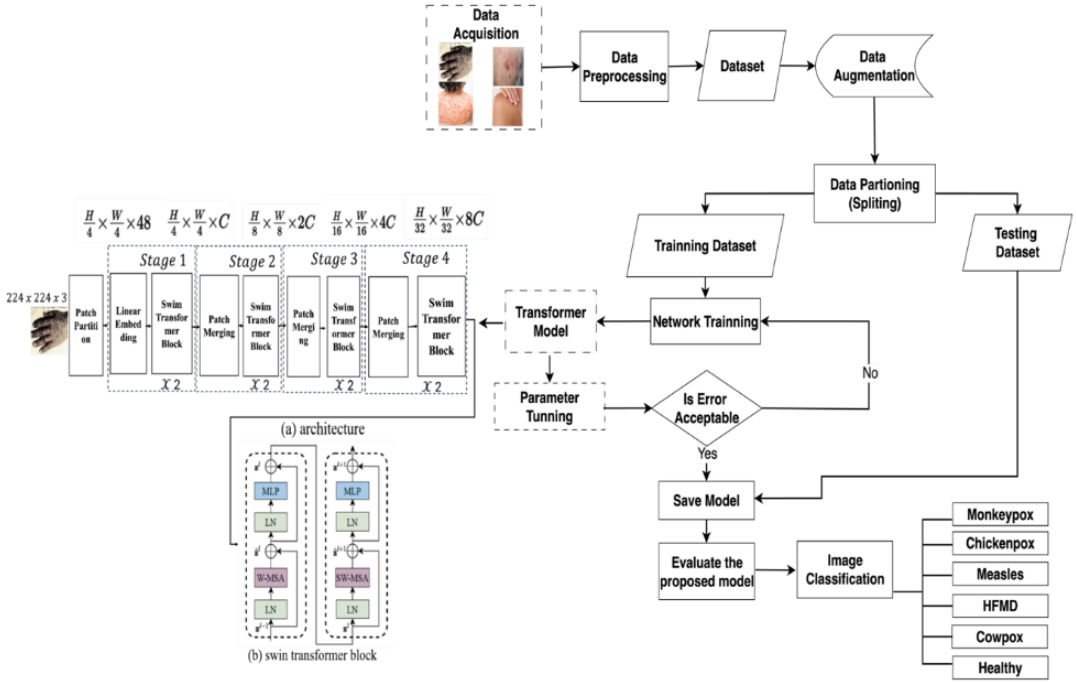

Figure 1. Research Proposed Method

3.1. Dataset collection

This study uses a skin image dataset obtained from research conducted by Ali S.N et al. [22] and online dataset available in "Kaggle dataset" [22]. Total images from six distinct classes, namely Mpox (284 images), Chickenpox (75 images), Measles (55 images), Cowpox (66 images), Hand-foot-mouth disease or HFMD (161 images), and Healthy (114 images). The dataset includes 755 original skin lesion images sourced from 541 distinct patients, ensuring a representative sample. This was followed by an image preprocessing step, where the images were cropped to standardize the dimensions to 224 x 224 pixels. This resizing aims to create uniformity in the image dimensions, ensuring a consistent format that aligns with the parameters tested in transformer model research. Example our dataset has shown in figure 2.

Figure 2. Dataset Samples

3.2. Data Augmentations

Data augmentation techniques were applied to increase the diversity of the training dataset and to improve model generalization. Two main categories of augmentation were used, namely positional transformations and color-based transformations. Positional augmentations were applied to simulate variations in object orientation and spatial positioning, while color augmentations were applied to address illumination variability since the images were collected from online sources with diverse lighting conditions. The positional augmentation operations include random rotation (within ±15°), horizontal and vertical flipping, zoom scaling (range 0.9–1.1), shear transformation (range ±10°), height shifting (up to 10% of the image height), and width shifting (up to 10% of the image width). In addition, several color-based transformations were applied, including brightness adjustment (range 0.8–1.2), contrast jitter (factor 0.2), hue jitter (range ±0.02), and saturation jitter (factor 0.2). A reflection based fill mode was used to handle newly created pixels during geometric transformations. Through the application of these augmentation techniques, the dataset size increased from the original 755 images to 8689 images. The augmented dataset consists of 1777 chickenpox images, 1502 measles images, 2642 monkeypox images, and 2768 normal skin images.

3.3. Data Portioning

All preprocessed data were divided into training and testing sets within a patient wise 5-fold cross-validation framework. The training set was used to train the Swin Transformer and other baseline models, while the testing set was used to evaluate the model performance in classifying monkeypox and other skin disease categories. We create a separate validation, training, and test set, instead, we utilized the testing set as an evaluation set to validate the model's performance iteratively. This approach ensures that the models' accuracy is measured on data that they have not seen during training, providing a robust evaluation of their performance. The patch tokens are processed using multiple Transformer block that incorporate a modified self-attention mechanism, known as Swin Transformer block. These blocks preserve the total number of tokens, given by $\frac{H}{4} x \frac{W}{4}$ along with the linear embedding.

3.4. Network training (Swin Transformer Model)

Figure 3 provides an overview of the Swin Transformer architecture, specifically the tiny version (Swin-T). The Swin Transformer architecture is adopted in this study as the main backbone for skin lesion classification. The Swin Transformer introduces a hierarchical vision transformer that computes self-attention within local windows while allowing cross-window interaction through a shifted window mechanism. This design enables efficient feature learning with linear computational complexity relative to image size. The RGB image with a resolution of 224×224 is first divided into non-overlapping patches using a patch partition module. In this study, patch size of 4×4 pixels are used. Each patch is flattened and projected into a feature embedding vector through a linear embedding layer. These embedded tokens are then processed by multiple Swin Transformer blocks organized in hierarchical stages. Each stage consists of a sequence of transformer blocks followed by a patch merging layer that reduces spatial resolution while increasing feature dimensionality. Through progressive patch merging, the network constructs a hierarchical feature representation across multiple spatial resolutions (e.g., $56 \times 56 \rightarrow 28 \times 28 \rightarrow 14 \times 14 \rightarrow 7 \times 7$). This hierarchical structure enables the model to capture both low-level texture patterns and high-level semantic features of skin lesion images. Each Swin Transformer block consists of two main components: a window-based multi-head self-attention module (W-MSA) or shifted window multi-head self-attention module (SW-MSA), followed by a multi-layer perceptron (MLP). Layer normalization (LN) and residual connections are applied to stabilize training. The output of the transformer block at layer $l$can be formulated as follows:

$$\hat{z}^l = W\text{-MSA}(LN(z^{l-1})) + z^{l-1} \quad (1)$$

$$z^l = MLP(LN(\hat{z}^l)) + \hat{z}^l \quad (2)$$

z^lrepresents the output feature of the l-th transformer block and $LN(\cdot)$denotes the layer normalization operation. In order to enable cross-window information exchange, the Swin Transformer introduces a shifted window multi-head self-attention mechanism. In this mechanism, the window partitioning is shifted between consecutive layers, allowing interactions between neighboring windows. The shifted attention computation is defined as:

$$\hat{z}^{l+1} = SW\text{-MSA}(LN(z^l)) + z^l \quad (3)$$

 ISSN 2723-6471

$$z^{l+1} = MLP(LN(\hat{z}^{l+1})) + \hat{z}^{l+1} \tag{4}$$

The self-attention operation itself is computed using the standard scaled dot-product attention formulation:

$$Attention(Q, K, V) = Softmax\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{5}$$

Q, K, and Vrepresent the query, key, and value matrices derived from the input features, and $d$denotes the dimensionality of the key vectors. In the Swin Transformer, the self-attention computation is performed locally within each window to reduce computational complexity. The window-based multi-head self-attention can be expressed as:

$$W\text{-MSA}(X) = Attention(Q, K, V) + B \tag{6}$$

B denotes the relative position bias used to incorporate spatial positional information into the attention computation. By combining hierarchical feature representation, window-based attention, and shifted window interactions, the Swin Transformer architecture efficiently captures both local texture patterns and global contextual relationships within skin lesion images. This capability is particularly beneficial for distinguishing visually similar dermatological conditions such as monkeypox, chickenpox, measles, and other skin diseases.

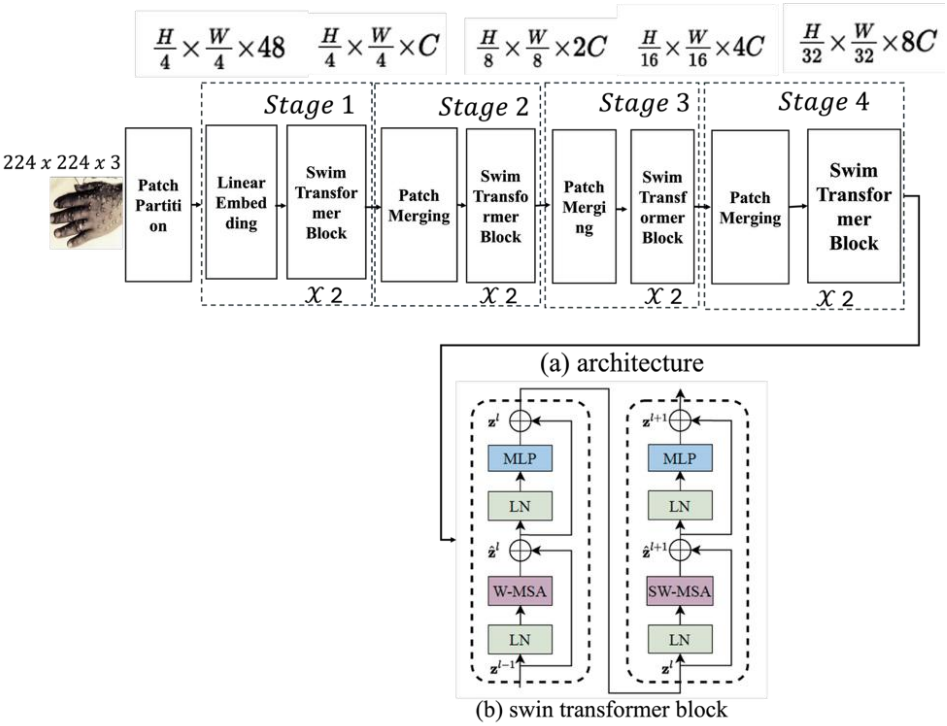

Figure 3. (a) The architecture of a Swin Transformer (Swin-T), (b) Swin Transformers Block

3.5. Hyperparameter Tuning

Developed model is crucial to improving classification performance. Hyperparameter tuning plays a critical role in machine learning and deep learning, as the chosen hyperparameters greatly influence the model's effectiveness. In this study, hyperparameter tuning methods are employed to determine the parameters of the optimizer used. An optimizer is an algorithm or method in artificial intelligence that plays a crucial role in adjusting parameters such as weights and biases, aiming to reduce the loss function or enhance production efficiency. This facilitates changes in weight values and adjusts the learning rate in neural networks so that losses can be minimized [25]. Table 1 and 2 lists the hyperparameters for pretraining phase and fine-tuning phase the current and proposed models. During the pretraining stage, the Mean Squared Error (MSE) loss was used to stabilize the learning process and facilitate the extraction of general feature representations from the input images. After this initial phase, the model was fine-tuned using the categorical cross-entropy loss function, which is widely used for multi-class classification tasks. This two-stage

 ISSN 2723-6471

optimization strategy allows the network to first learn robust visual representations and then optimize its discriminative capability for the final classification task.

Table 1. Parameter setting for Swin Transformer

Parameters	Pretraining Phase	Fine-tuning Phase
Batch Size	64	32
Optimizer	Adam, AdamW	AdaMax
Optimizer Parameters	β_1=0.9, β_2=0.999, ε=1e-8	$_1$=0.9, β_2=0.999, ε=1e-8
Weight Decay	0.05	0.01
Initial Learning Rate	$1e^{-3}$	$1e^{-4}$
Learning Rate Schedule	Cosine annealing with warmup	Fixed
Total Epochs	50	25
Loss Function	Mean Squared Error	Cross-Entropy
Early Stopping Patience	20 epochs	10 epochs
Dropout Rate	0.1 (in all transformer blocks)	0.1 (in all transformer blocks)
Data Augmentation	rotation, horizontal flip vertical flip, zoom range, shear range, height shift range, width shift range, brightness range, contrast jitter, hue jitter, saturation jitter, and fill mode	rotation, horizontal flip vertical flip, zoom range, shear range, height shift range, width shift range, brightness range, contrast jitter, hue jitter, saturation jitter, and fill mode

Table 2. Parameter setting for deep learning model (VGG16, RESNET50 modified model)

Parameters	Range
Learning rate	0.0001, 0.001, 0.01
Epoch	20
Batch size	32
Optimizer	Adam, AdamW, AdaMax
Momentum	0.098
Weight decay	0.0001

3.6. Evaluate Performance

To evaluate the Classification model used a confusion matrix, where performance parameters such as Accuracy, Recall, Precision, and F1-score are determined. These parameters are calculated using a confusion matrix created for each model. Accuracy is calculated to determine the percentage of correct predictions. Precision is calculated to determine the probability of positive classification. Specificity determines the percentage of negative classifications correctly predicted from all parameters. Unlike specificity, recall determines the percentage of positive classes correctly predicted. The F1-score is used to determine the balance between specificity and recall. The parameters are expressed in the following Eq. (7-10):

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{7}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{8}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{9}$$

$$F1 - Score = 2\left(\frac{Precision\ x\ Recall}{Precision + Recall}\right) \tag{10}$$

ISSN 2723-6471

4. Results and Discussion

Our dataset consists of 755 original skin lesion images collected from 541 distinct patients; because some patients contribute more than one image, the patient is treated as the independent unit for evaluating clinical generalization. To prevent patient-identity leakage and artificially inflated metrics, we revised the evaluation protocol to patient-wise 5-fold cross-validation by using patient_id as the grouping key, ensuring that all images from the same patient remain within a single fold and never appear in both the training and test sets within any fold. In each iteration, the model is trained on patients from four folds (≈80%) and evaluated on the held-out fold (≈20%), and split safety is explicitly verified by confirming that the patient intersection between training and test is zero, i.e., $| P_{train} \cap P_{test} |= 0$ for all folds. After fold construction, we apply augmentation only to the training data (training-only augmentation), so that no augmented version or near-duplicate of an image can enter the test fold; consequently, each test fold contains only original images for unbiased evaluation. The total dataset size after augmentation is 8689 images, which corresponds to the global augmentation ratio relative to the original data ($8689/755 \approx 11.51$); under this leakage-safe scheme, the ratio is applied per fold only to the training portion, so the augmented training size is computed as $N^{(k)}_{\text{train,aug}} \approx r \cdot N^{(k)}_{\text{train,orig}}$, while $N^{(k)}_{\text{test}}$ remains original-only. In addition to the patient-overlap check, we conduct a cross-fold duplicate/near-duplicate screening step (e.g., using pHash or embedding similarity) to ensure that no identical or near-identical images cross the train–test boundary after filtering. The complete fold composition (patient and original-image counts per train/test), training-only augmentation, and leakage verification outcomes (patient overlap and cross-fold duplicates) are transparently summarized in table 3, making the split logic auditable and the experimental evidence sufficient to support clinical discrimination claims without bias from data leakage.

Table 3. Leakage verification and fold composition (patient-wise)

Fold	#Patients (Train)	#Patients (Test)	#Images (Train, original)	#Images (Test, original)	Augmentation applied to Train only?	#Augmented images (Train, added)	#Images (Train, after aug)	|Ptrain ∩ Ptest|
1	433	108	604	151	Yes	6347	6951	0
2	433	108	604	151	Yes	6347	6951	0
3	433	108	604	151	Yes	6347	6951	0
4	433	108	604	151	Yes	6347	6951	0
5	432	109	604	151	Yes	6347	6951	0
Mean ± SD	432.8 ± 0.4	108.2 ± 0.4	604.0 ± 0.0	151.0 ± 0.0	-	6347 ± 0	6951 ± 0	0

To ensure a fair and controlled comparison, all baseline models were trained and evaluated under an identical experimental pipeline. Specifically, we used the same patient-wise 5-fold splits for every method (table 3), with fold construction performed prior to augmentation. All experiments were conducted under a leakage-safe protocol using patient-wise 5-fold cross-validation, where fold assignment is performed prior to augmentation and no patient appears in both training and test splits within any fold. Each input image was converted to RGB and resized to 224×224 using bilinear interpolation (aspect ratio preserved via resizing followed by center-crop to 224×224). Pixel intensities were normalized using ImageNet statistics (mean = [0.485, 0.456, 0.406], std = [0.229, 0.224, 0.225]). Data augmentation was applied only to the training split and implemented online during training: random horizontal flip (p=0.5), random rotation (±15°), random brightness/contrast/saturation adjustment (ColorJitter with magnitude 0.2), random zoom/scale (0.9–1.1), and random erasing (p=0.25, scale 0.02–0.2). No augmentation was applied to the held-out test fold. We used Swin Transformer Tiny (Swin-T) as the backbone (patch size 4, window size 7, input resolution 224), initialized with ImageNet-1K pretrained weights. The classification head was replaced with a randomly initialized linear layer for 5 classes, with dropout 0.2 before the final layer. Training used AdamW (learning rate 5×10^{-5}, betas (0.9, 0.999), weight decay 0.05) for 50 epochs with batch size 32 (gradient accumulation 1; effective batch size 32). We employed a cosine learning-rate schedule with 5-epoch warmup and a minimum learning rate of 1×10^{-6}. Mixed precision training (AMP) was enabled. Early stopping was applied based on validation macro-F1 with patience 10 epochs, and the best checkpoint per fold was selected by the highest validation macro-F1. To support replication, we fixed random seeds

(seed=42) and enabled deterministic settings where possible; we report the full experiment manifest in table 4, including hyperparameters.

Table 4. Experiment manifest for replication (leakage-safe patient-wise CV).

Category	Setting (exact)
Data split	Patient-wise 5-fold CV; folds built before augmentation; overlap check: |Ptrain ∩ Ptest|=0 (Table X)
Input	RGB; resize 224×224 (bilinear); preserve aspect ratio → resize shortest side then center-crop 224×224
Normalization	ImageNet mean/std: mean (0.485, 0.456, 0.406); std (0.229, 0.224, 0.225)
Augmentation (train only)	Horizontal flip p=0.5; rotation ±15°; scale/zoom 0.9–1.1; ColorJitter 0.2; random erasing p=0.25, area 0.02–0.2; online; none on test
Model	Swin-Tiny (Swin-T); patch=4, window=7, input=224; dropout=0.2; classes=5
Pretraining	ImageNet-1K pretrained backbone; head randomly initialized
Freezing	No layer freezing (full fine-tuning from epoch 1)
Loss	Cross-entropy; label smoothing 0.1
Optimizer	AdamW; lr=5e−5; betas=(0.9, 0.999); weight decay=0.05
Scheduler	Cosine annealing; warmup=5 epochs; min lr=1e−6
Batch/epochs	batch=32; grad accumulation=1; epochs=50
Early stopping	metric=val macro-F1; patience=10; min_delta=1e−4
AMP	Enabled (mixed precision)
Seeds	global seed=42; fold seeds fixed; deterministic flags enabled where feasible
Checkpointing	best model per fold selected by max val macro-F1
Evaluation	Accuracy, Precision, Recall, macro-F1, report mean±std across folds

According to these experimental tasks, our proposed model with Swin Transformer, Vision Transformer, VGG16 [27] and modified Resnet50 [26] performed with a loss, accuracy and standard deviation on the augmented dataset, respectively, as well as lower standard deviation values for each metric on both datasets. The experimental result, presented in tables 5-8, show that Swin Transformer and Vision Transformer (ViT) model outperform the VGG16 and modified ResNet50 models. Both Transformer-based model achieved a training accuracy (*Train_acc*) and validation accuracy (*Val_acc*) of 1.00 by the fifth epoch. Although the validation accuracy of both Swin Transformer and Vision Transformer approaches reaches near perfect values in later training epochs, several precautions were taken to prevent overfitting and data leakage. Specifically, the experiments were conducted using a patient wise 5-fold cross-validation strategy, ensuring that images from the same patient do not appear in both training and validation sets. Additionally, all data augmentation operations were applied only to the training data within each fold. These precautions help ensure that the reported performance reflects the model's ability to generalize to unseen samples rather than memorizing training data.

Table 5. 5-Fold cross validation results for augmented dataset RESNET50 model

Fold	Evaluation Metrics					
	Train_Loss	**Train_acc**	**Train_ std**	**Val_ loss**	**Val_ acc**	**Val_std**
Fold-1	0.013 ± 0.00365	0.99 ± 0.001	1.724 ± 0.00416	0.881 ± 0.3735	0.784 ± 0.09223	1.741 ± 0.03444
Fold-2	0.004 ± 0.00415	0.99 ± 0.003	1.727 ± 0.00713	0.053 ± 0.4432	0.981 ± 0.03231	1.756 ± 0.05453
Fold-3	0.008± 0.00212	0.99 ± 0.011	1.741 ± 0.00316	0.280 ± 0.3212	0.941 ± 0.02344	1.634 ± 0.02332
Fold-4	0.005± 0.00311	0.99 ± 0.012	1.726 ± 0.00531	0.006 ± 0.2314	1.00 ± 0.03232	1.716 ± 0.01233
Fold-5	0.005 ± 0.00322	0.99 ± 0.011	1.735 ± 0.00234	0.003 ± 0.1232	**1.00 ± 0.09112**	1.733 ± 0.05454

Table 6. 5-Fold cross validation results for augmented dataset for swin transformer model

Fold	Evaluation Metrics					
	Train_Loss	**Train_acc**	**Train_ std**	**Val_ loss**	**Val_ acc**	**Val_std**
Fold-1	0.0005± 0.001	1.00 ± 0.022	1.671 ± 0.011	0.636± 0.010	0.833± 0.001	1.572
Fold-2	0.0002± 0.006	1.00 ± 0.012	1.727 ± 0.010	0.116± 0.009	0.969± 0.001	1.721
Fold-3	0.0001± 0.002	1.00 ± 0.001	1.744 ± 0.011	0.459± 0.008	0.941± 0.002	1.618

Fold-4	0.0130± 0.001	0.99 ± 0.009	1.726 ± 0.009	0.126± 0.007	0.946± 0.001	1.631
Fold-5	0.0002± 0.002	1.00 ± 0.021	1.736 ± 0.011	0.0134± 0.005	**1.00**± 0.001	1.733

Table 7. 5-Fold cross validation results for augmented dataset vision transformer model

Fold	Evaluation Metrics					
	Train_Loss	**Train_acc**	**Train_ std**	**Val_ loss**	**Val_ acc**	**Val_std**
Fold-1	0.0013± 0.001	1.00± 0.002	1.671± 0.011	0.563± 0.001	0.84± 0.002	1.611
Fold-2	0.0007± 0.001	1.00± 0.011	1.728± 0.002	0.145± 0.012	0.957± 0.002	1.720
Fold-3	0.0008± 0.001	1.00± 0.012	1.744± 0.002	0.189± 0.011	0.954± 0.003	1.671
Fold-4	0.0004± 0.001	1.00± 0.022	1.736± 0.003	0.0004± 0.001	1.00± 0.002	1.733
Fold-5	0.0002± 0.001	1.00± 0.011	1.736± 0.001	0.0134± 0.001	**1.00**± 0.001	1.733

Table 8. 5-Fold cross validation results for augmented dataset VGG-16 model

Fold	Evaluation Metrics					
	Train_Loss	**Train_acc**	**Train_ std**	**Val_ loss**	**Val_ acc**	**Val_std**
Fold-1	0.0669± 0.001	0.97± 0.012	1.723± 0.011	2.349± 0.012	0.68± 0.002	1.669
Fold-2	0.0248± 0.002	0.99± 0.022	1.726± 0.012	0.362± 0.011	0.902± 0.002	1.723
Fold-3	0.0042± 0.002	0.98± 0.011	1.741± 0.022	0.506± 0.010	0.856± 0.001	1.711
Fold-4	0.0043± 0.002	1.00± 0.022	1.727± 0.012	0.0004± 0.009	0.986± 0.003	1.693
Fold-5	0.0111± 0.002	0.99± 0.011	1.735± 0.012	0.078± 0.008	**0.967**± 0.003	1.729

Figures 4 and 5 display the loss and accuracy curves of the Swin Transformer model. Figures 4 Show the validation accuracy is high and close to the training accuracy it starts 0.922 (92.2%), climbs sharply until about epoch 3-4, then plateaus around 0.993 (99.3%). Train accuracy shows the model performing on the training data during each epoch, and accuracy starts at 0.98 (98%) and quickly improves to 1.0 (100%) by around the 5th epoch, remaining constant afterward. The second experimental scenario evaluates the performance of different optimizers and learning rate. The optimizers used are Adam and AdaMax, each combined with three learning rates: 0.01, 0.001, and 0.0001. Table 9-12 present the result of the experiments under the first scenario across the model architecture.

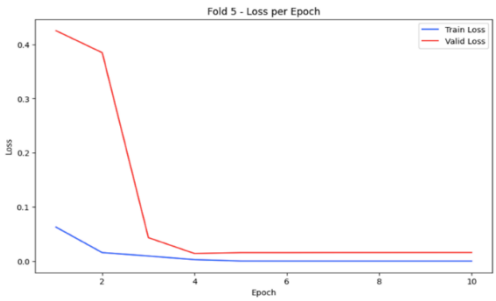

Figure 4. Loss vs epoch for hierarchical swin transformer

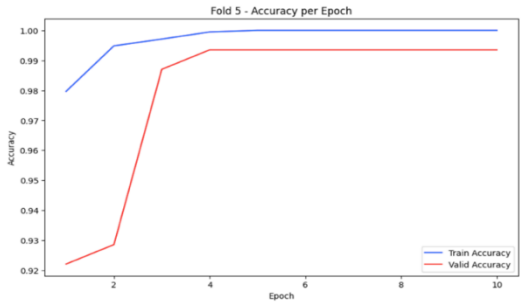

Figure 5. Accuracy vs epoch for hierarchical swin transformer model

Table 9-12 present the performance results of the three optimizers used in this study. Among the tested models, the Swin Transformer architecture combine with the AdamW optimizer and a learning rate 0.0001 achieved the best performance.

Table 9. Optimizer and learning rate result (accuracy) for VGG 16 model

Fold	ADAM			ADAMAX			ADAMW		
	0.0001	**0.001**	**0.01**	**0.0001**	**0.001**	**0.01**	**0.0001**	**0.001**	**0.01**
Fold-1	0.74	0.40	0.34	0.59	0.81	0.50	0.75	0.49	0.45

ISSN 2723-6471

Fold-2	0.82	0.41	0.36	0.69	0.82	0.51	0.84	0.51	0.36
Fold-3	0.82	0.55	0.38	0.71	0.80	0.41	0.82	0.56	0.38
Fold-4	0.78	0.62	0.34	0.78	0.80	0.42	0.83	0.48	0.34
Fold-5	0.81	0.43	0.32	0.69	0.84	0.49	0.79	0.43	0.33

Table 10. Optimizer and learning rate result (accuracy) for RESNET 50 model

Fold	ADAM			ADAMAX			ADAMW		
	0.0001	**0.001**	**0.01**	**0.0001**	**0.001**	**0.01**	**0.0001**	**0.001**	**0.01**
Fold-1	0.78	0.78	0.51	0.81	0.82	0.55	0.81	0.74	0.59
Fold-2	0.79	0.78	0.57	0.76	0.84	0.41	0.80	0.73	0.29
Fold-3	0.81	0.78	0.48	0.83	0.84	0.69	0.84	0.81	0.47
Fold-4	0.84	0.78	0.32	0.83	0.80	0.47	0.77	0.81	0.40
Fold-5	0.87	0.89	0.59	0.89	0.87	0.60	0.90	0.83	0.43

Table 11. Optimizer and learning rate result (accuracy) for swin transformer model

Fold	ADAM			ADAMAX			ADAMW		
	0.0001	**0.001**	**0.01**	**0.0001**	**0.001**	**0.01**	**0.0001**	**0.001**	**0.01**
Fold-1	0.72	0.73	0.36	0.74	0.73	0.7	0.72	0.76	0.36
Fold-2	0.70	0.63	0.36	0.71	0.68	0.66	0.66	0.65	0.39
Fold-3	**0.83**	0.78	0.38	**0.83**	0.81	0.78	**0.80**	0.80	0.44
Fold-4	0.67	0.68	0.35	0.68	0.74	0.69	0.75	0.77	0.36
Fold-5	0.73	0.75	0.29	0.73	0.79	0.75	0.79	0.76	0.36

Table 12. Optimizer and learning rate result (accuracy) for vision transformer model

Fold	ADAM			ADAMAX			ADAMW		
	0.0001	**0.001**	**0.01**	**0.0001**	**0.001**	**0.01**	**0.0001**	**0.001**	**0.01**
Fold-1	0.73	0.36	0.36	0.86	0.77	0.36	0.81	0.36	0.36
Fold-2	0.83	0.38	0.38	0.79	0.76	0.38	0.80	0.38	0.38
Fold-3	0.76	0.34	0.34	0.88	0.73	0.34	0.81	0.34	0.34
Fold-4	0.74	0.33	0.32	0.80	0.75	0.33	0.80	0.33	0.33
Fold-5	0.86	0.29	0.29	0.89	0.83	0.29	**0.90**	0.29	0.29

As shown in table 12, this configuration attained the highest global accuracy of 0.90, outperforming the other model. In the third scenario, the results from Scenario 2 are compared with the proposed Swin Transformer architecture and presented in table 13, where only the performance for the best K-fold (2), is shown for each class. Testing results of the proposed Swin Transformer model achieved a global accuracy of 0.99. This indicates that the proposed architecture significantly improves performance, both at the class level and in terms of overall accuracy. Specifically, for the *monkeypox* class, the model achieved an accuracy of 0.99 (99%), a precision of 0.97 (97%), a recall of 1.00 (100%), and an F1-score of 0.98 (98%).

Table 13. Evaluation matrix performance

Class	Proposed Hierarchical SwinTransformer				Swin Transformer+ADAMW Optimizer			
	Accuracy	**Precision**	**Recall**	**F1-Score**	**Accuracy**	**Precision**	**Recall**	**F1-Score**
Chickenpox	0.9	1.0	0.91	0.95	0.71	1.0	0.71	0.83
Cowpox	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
HFMD	1.0	1.0	1.0	1.0	1.0	0.88	1.0	0.94
Healthy	1.0	1.0	1.0	1.0	0.9	0.82	0.9	0.86
Measles	1.0	1.0	1.0	1.0	0.66	0.8	0.67	0.73

Monkeypox	0.99	0.97	1.0	0.98	0.94	0.94	0.94	0.94

Figure 6 shows the confusion matrix for each class in the test dataset using the hierarchical Swin Transformer architecture. Based on the experimental results, we conclude that the hierarchical Swin Transformer backbone is a suitable and effective architecture for skin-image classification, particularly for monkeypox detection. By partitioning input images into patches and progressively merging then into a multi-scale feature hierarchy from $\frac{H}{4} x \frac{W}{4}$ to $\frac{H}{32} x \frac{W}{32}$, the model can capture both fine-grained texture details and high-level semantic pattern of skin lesions. The combination of windows-based and shifted-window multi head self-attention provides rich and global context while maintaining linear computational complexity, which result in stable training and robust generalization on our dataset. Overall, the Swin Transformer-based framework delivers discriminative feature representations that enable accurate separation between monkeypox and non-monkeypox class, confirming our design choice to adopt these architectures as the core classifier in this study. Additionally, figure 7 presents the ROC curves for all six classes. The Area Under the Curve (AUC) values for classes 1 to 3 (*chickenpox*, *cowpox*, and *HFMD*) reached 1.0, while classes 4 to 6 (*healthy*, *measles*, and *monkeypox*) achieved AUC values of 0.99 and 0.98. These results reflect a high true positive rate (TPR) and a low false positive rate (FPR), further confirming that the Swin Transformer-based method demonstrates excellent performance and outperforms several Vision Transformer variants as well as conventional CNN architectures.

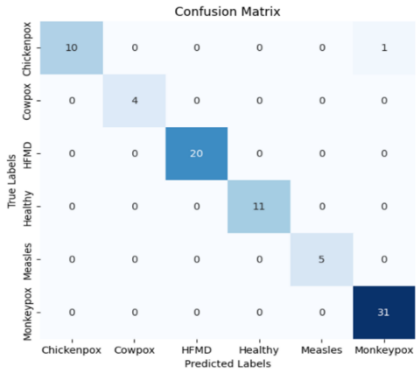

Figure 6. Confusion Matrix with Swin Transformer Architecture

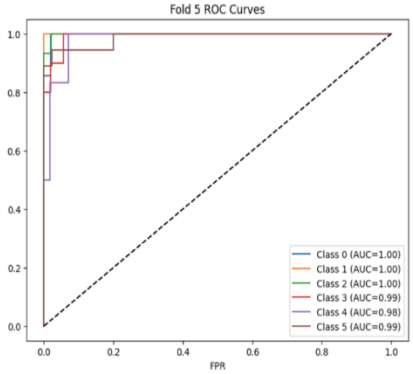

Figure 7. ROC Curves for 6 class

To ensure a fair and auditable comparison with prior studies, we divide the literature into two categories: comparable benchmarks and non-comparable settings. This separation is necessary because reported performance in literature can vary substantially due to differences in dataset version and inclusion criteria, class taxonomy and label mapping, sample composition, and evaluation protocol (e.g., holdout vs k-fold, patient-wise vs image-wise splitting, and augmentation policies). A study is treated as comparable only if it satisfies all of the following conditions: (i) the same dataset version/source (or an explicitly stated identical dataset) with consistent inclusion criteria; (ii) the same class taxonomy and label mapping (i.e., identical set of classes and definitions), (iii) validation scheme, and (iv) an augmentation policy that does not introduce leakage (augmentation applied after split on training only, or an explicitly leakage-safe procedure). If any of these elements are not satisfied or not reported, the study is categorized as non-comparable and is used only as literature context. Our primary experimental evidence is derived from our leakage-safe evaluation protocol (patient-wise fold construction prior to augmentation and explicit leakage checks), summarized in table 3. In Table 14(a) our proposed method with Swin Transformer Architecture reached 99% accuracy and showed an increase of almost 16% than previous research [18] with the same dataset.

Table 14(a). Reports comparable (apple-to-apple) benchmarks

Study reference	Method	Key result	Dataset Images	Name of Classes
Ali et al. [18]	DenseNet121	82.26% accuracy	755 images	Chickenpox, Measles, HFMD, Healthy Skin, Monkeypox
Proposed Model in this Research	Hierarchical Swin Transformer	99% accuracy	755 images	Chickenpox, Measles, HFMD, Healthy Skin, Monkeypox

Table 14(a) provides a controlled comparison under a consistent and leakage-safe protocol; therefore, any performance improvement claimed in this manuscript is based only on the comparable benchmarks in table 14(a). In this table, Ali

et al. [18] is listed because it reports a DenseNet121 model evaluated on a dataset with the same nominal scale (755 images) and the same six-class taxonomy used in our work (Chickenpox, Measles, HFMD, Healthy Skin, Monkeypox). Under that setting, Ali et al. reports a key result of 82.26% accuracy, which provides an informative baseline reference for how a conventional CNN backbone performs on a similarly defined multi-class problem. Our proposed method, the Hierarchical Swin Transformer, is reported in the second row with 99% accuracy on the same five-class formulation and the same dataset size (755 original images), reflecting the performance obtained under our experimental pipeline. Interpreted at a high level, table 14(a) suggests that, when the problem is framed with an identical class list and a dataset of comparable size, a transformer-based hierarchical architecture can yield substantially higher accuracy than a standard DenseNet121 backbone, indicating improved feature representation and discrimination capability for fine-grained lesion patterns. In contrast, table 14(b) is intentionally presented as non-comparable literature context because prior studies differ in task formulation (binary vs multi-class), class definitions, dataset composition, and/or evaluation procedures, or they do not report sufficient details to verify equivalence. By separating comparable and non-comparable settings, we prevent misleading conclusions while preserving a complete overview of the research landscape. Finally, our generalization claims are supported primarily by the leakage-safe patient-wise validation procedure and verification reported in table 14(b), which reduces the risk of inflated performance due to patient overlap or augmentation leakage.

Table 14(b). Summary of Related Studies (Non-Comparable Settings)

Study reference	Method	Key result	Dataset Images	Name of Classes
Ahsan et al. [27]	Modified VGG16	97% accuracy, 97.2% AUC	90 for study one and 1754 for study two	Chickenpox vs Monkeypox and Monkeypox vs others for study one and two
Ahsan et al. [13]	GRA-TLA on CNN models	77-88% accuracy	76 images	Monkeypox, Normal
Bala et al. [17]	13 pre-trained DL models	85.44% precision, 87.13 accuracy	770 images	Chickenpox, Measles, Healthy Skin, Monkeypox
Dahiya et al. [28]	Yolov5 model-based	98.18% accuracy	971 images	Chickenpox and Monkeypox
Pramanik et al.[29]	CNN aided with Beta function-based normalization	93.39%	228 images	Monkeypox. others
Monuz et al. [30]	Ensemble of CNNs	98% accuracy	300 images	Monkeypox, Healthy, Others
Gasshani et al. [31]	MobileNetV2	95.5% accuracy	770 images	Chickenpox, Measles, Monkeypox, and Normal)
Proposed Model in this Research	Hierarchical Swin Transformer	99% accuracy	755 images	Chickenpox, Measles, HFMD, Healthy Skin, Monkeypox

5. Conclusion

Contribution in this study, the Swin Transformer method demonstrated the best performance compared to several CNN architectures and even Vision Transformer methods based on its global accuracy as well as per-class accuracy, precision, recall, and F1-score. The average global accuracy achieved was 0.99 (99%), which is further supported by the AUC values. For classes 1 to 3 (*chickenpox*, *cowpox*, and *HFMD*), the AUC values were 1.0, while for classes 4 to 6 (*healthy*, *measles*, and *monkeypox*), the AUC values were 0.99 and 0.98, respectively. These results indicate that the proposed Hierarchical Swin Transformer model exhibits exceptionally high classification performance across all classes, with AUC values consistently above 0.98. Architecture Hierarchical Swin Transformer Model showed achieved the best accuracy compared with previous research, as shown in table 14(a). In table 14(a), previous study with Ali et all [22], where our dataset is same (apple to apple), showed our proposed method using hierarchical Swin Transformer achieved the higher accuracy. Although the proposed Swin Transformer based framework demonstrates strong classification performance on the evaluated dataset, several limitations should be considered. First, the dataset used in this study was collected from publicly available online sources, which may introduce potential dataset bias related to image quality, acquisition conditions, and class distribution. Second, the model was evaluated on a limited dataset size, and its generalization ability across different clinical environments or imaging devices remains to be further validated. Third, while the proposed approach shows promising performance for automated skin disease classification, real-world clinical deployment would require additional validation on large-scale clinical datasets, as well as integration with dermatological diagnostic workflows. Future work will focus on evaluating the proposed

framework on larger and more diverse datasets and exploring its potential integration into clinical decision support systems.

6. Declarations

6.1. Author Contributions

Conceptualization: P.D.W.A. and S.S.; Methodology: P.D.W.A. and G.A.P.; Software: P.D.W.A. and S.; Validation: P.D.W.A. and D.P.H.; Formal Analysis: P.D.W.A. and P.M.P.; Data Curation: G.A.P. and P.M.P.; Writing Original Draft Preparation: P.D.W.A. and S.; Writing Review and Editing: P.D.W.A. and S.S.; Visualization: D.P.H.; Supervision: P.D.W.A.; Funding Acquisition: P.D.W.A.

6.2. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

6.3. Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

6.4. Institutional Review Board Statement

Not applicable.

6.5. Informed Consent Statement

Not applicable.

6.6. Declaration of Competing Interest

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

References

[1] S. T. Al Awaidy, F. Khamis, M. Sallam, R. M. Ghazy, and H. Zaraket, "Monkeypox (mpox) outbreak more queries posed as cases soar globally," *Sultan Qaboos Univ. Med. J.*, vol. 23, no. 1, pp. 1–4, 2023, doi: 10.18295/squmj.8.2022.046.

[2] R. K. Mohapatra, "Monkeypox lineages amid the ongoing COVID-19 pandemic: a global public health concern – correspondence," *Int. J. Surg.,* vol. 107, no. Sep., pp. 9–11, 2022, doi: 10.1016/j.ijsu.2022.106968.

[3] H. Hatami, "Demographic, epidemiologic, and clinical characteristics of human monkeypox disease pre- and post-2022 outbreaks: a systematic review and meta-analysis," *Biomedicines,* vol. 11, no. 3, pp. 1–12, 2023, doi: 10.3390/biomedicines11030957.

[4] T. Nayak, "Detection of monkeypox from skin lesion images using deep learning networks and explainable artificial intelligence," *Appl. Math. Sci. Eng.,* vol. 31, no. 1, pp. 1–12, 2023, doi: 10.1080/27690911.2023.2225698.

[5] A. A. Aouragh and M. Bahaj, "Comparison results of hybrid CNN-machine learning algorithms architectures for monkeypox images classification," *Int. J. Appl. Sci. Eng. Technol.*, vol. 2023, no. Jan., pp. 1–6, 2023, doi: 10.1109/IRASET57153.2023.10153062.

[6] M. Lakshmi and R. Das, "Classification of monkeypox images using LIME-enabled investigation of deep convolutional neural network," *Diagnostics,* vol. 13, no. 9, pp. 1–12, 2023, doi: 10.3390/diagnostics13091639.

[7] W. Yao, J. Bai, W. Liao, Y. Chen, M. Liu, and Y. Xie, "From CNN to transformer: a review of medical image segmentation models," *J. Imaging Inform. Med.*, vol. 37, no. 4, pp. 1529–1547, 2024, doi: 10.1007/s10278-024-00981-7.

[8] D. Nam and A. Pak, "Overview of transformer-based models for medical image segmentation," *Sci. J. Astana IT Univ.*, vol. 2023, no. Jan., pp. 64–75, 2023, doi: 10.37943/13bkbf2003.

[9] K. He, F. Gou, and J. Wu, "Image segmentation technology based on transformer in medical decision-making system," *IET Image Process.,* vol. 17, no. 10, pp. 3040–3054, 2023, doi: 10.1049/ipr2.12854.

[10] W. Li, "A swin transformer with dynamic high-pass preservation for remote sensing image pansharpening," *arXiv,* vol. 2023, no. Jan., pp. 1–10, 2023.

[11] Y. Peng, J. Ren, J. Wang, and M. Shi, "Spectral-Swin transformer with spatial feature extraction enhancement for hyperspectral image classification," *arXiv,* vol. 2023, no. Jan., pp. 1–19, 2023.

[12] M. Tounsi, E. Aram, A. T. Azar, A. Al-Khayyat, and I. K. Ibraheem, "A comprehensive review on biomedical image classification using deep learning models," Eng. Technol. Appl. Sci. Res., vol. 15, no. 1, pp. 19538–19545, 2025, doi: 10.48084/etasr.8728.

[13] M. M. Ahsan, "Deep transfer learning approaches for monkeypox disease diagnosis," *Expert Syst. Appl.*, vol. 216, no. Jan., pp. 1–12, 2023, doi: 10.1016/j.eswa.2022.119483.

[14] A. A. Abdelhamid, "Classification of monkeypox images based on transfer learning and the Al-Biruni earth radius optimization algorithm," *Mathematics,* vol. 10, no. 19, pp. 1–12, 2022, doi: 10.3390/math10193614.

[15] V. H. Sahin, I. Oztel, and G. Y. Oztel, "Human monkeypox classification from skin lesion images with deep pre-trained network using mobile application," *J. Med. Syst.*, vol. 46, no. 11, pp. 1–12, 2022, doi: 10.1007/s10916-022-01863-7.

[16] M. E. Haque, M. R. Ahmed, R. S. Nila, and S. Islam, "Classification of human monkeypox disease using deep learning models and attention mechanisms," *arXiv,* vol. 2022, no. Nov., pp. 1–12, 2022.

[17] D. Bala, "MonkeyNet: a robust deep convolutional neural network for monkeypox disease detection and classification," *Neural Netw.*, vol. 161, no. Jan., pp. 757–775, 2023, doi: 10.1016/j.neunet.2023.02.022.

[18] G. Y. Oztel, "Vision transformer and CNN-based skin lesion analysis: classification of monkeypox," *Multimed. Tools Appl*., vol. 83, no. 28, pp. 71909–71923, 2024, doi: 10.1007/s11042-024-19757-w.

[19] J. F. Ma, P. F. He, C. L. Li, and R. Nie, "Mpox virus image segmentation based on multiscale expansion convolution," *IEEE Access,* vol. 12, no. Sep., pp. 117608–117616, 2024, doi: 10.1109/ACCESS.2024.3448364.

[20] S. Chakroborty, "A hybrid deep learning framework for early detection of mpox using image data," *Healthcare Anal.*, vol. 7, no. Jan., pp. 1–12, 2025, doi: 10.1016/j.health.2025.100396.

[21] S. Vuran, M. Ucan, M. Akin, and M. Kaya, "Multi-classification of skin lesion images including mpox disease using transformer-based deep learning architectures," *Diagnostics,* vol. 15, no. 3, pp. 1–12, 2025, doi: 10.3390/diagnostics15030374.

[22] S. N. Ali, "A web-based mpox skin lesion detection system using state-of-the-art deep learning models considering racial diversity," *Biomed. Signal Process. Control*, vol. 98, no. Jan., pp. 1–9, 2024, doi: 10.1016/j.bspc.2024.106742.

[23] Z. Liu, "Swin transformer: hierarchical vision transformer using shifted windows," *IEEE Access,* vol. 2021, no. Oct., pp. 9992–10002, 2021, doi: 10.1109/ICCV48922.2021.00986.

[24] W. Shi, J. Xu, and P. Gao, "SSformer: a lightweight transformer for semantic segmentation," *IEEE Access*, vol. 2022, no. Sep., pp. 1–5, 2022, doi: 10.1109/MMSP55362.2022.9949177.

[25] P. D. W. Ayu and G. A. Pradipta, "U-Net tuning hyperparameter for segmentation in amniotic fluid ultrasonography image," *IEEE Access,* vol. 2022, no. Jun., pp. 1–6, 2022, doi: 10.1109/ICORIS56080.2022.10031294.

[26] P. D. W. Ayu, G. A. Pradipta, I. M. D. Susila, D. P. Hostiadi, and M. Liandana, "Deep learning based detection and classification of amniotic fluid echogenicity type for enhanced prenatal diagnosis," *Int. J. Intell. Eng. Syst.*, vol. 18, no. 1, pp. 246–267, 2025, doi: 10.22266/ijies2025.0229.18.

[27] M. M. Ahsan, M. R. Uddin, M. Farjana, A. N. Sakib, K. Al Momin, and S. A. Luna, "Image data collection and implementation of deep learning-based model in detecting monkeypox disease using modified VGG16," *arXiv,* vol. 2022, no. Jun., pp. 1–12, 2022.

[28] N. Dahiya, "Hyper-parameter tuned deep learning approach for effective human monkeypox disease detection," *Sci. Rep*., vol. 13, no. 1, pp. 1–19, 2023, doi: 10.1038/s41598-023-43236-1.

[29] R. Pramanik, B. Banerjee, G. Efimenko, D. Kaplun, and R. Sarkar, "Monkeypox detection from skin lesion images using an amalgamation of CNN models aided with beta function-based normalization scheme," *PLoS One,* vol. 18, no. 4, pp. 1–21, 2023, doi: 10.1371/journal.pone.0281815.

[30] L. Muñoz-Saavedra, E. Escobar-Linero, J. Civit-Masot, F. Luna-Perejón, A. Civit, and M. Domínguez-Morales, "A robust ensemble of convolutional neural networks for the detection of monkeypox disease from skin images," *Sensors,* vol. 23, no. 16, pp. 1–24, 2023, doi: 10.3390/s23167134.

[31] M. S. A. M. Al-Gaashani, W. Xu, and E. Y. Obsie, "MobileNetV2-based deep learning architecture with progressive transfer learning for accurate monkeypox detection," *Appl. Soft Comput*., vol. 169, no. Oct., pp. 1–12, 2025, doi: 10.1016/j.asoc.2024.112553.