IJSRA

International Journal of Science and Research Archive

eISSN: 2582-8185

Cross Ref DOI: 10.30574/ijsra

Journal homepage: https://ijsra.net/

Int. J. Sci. Res. Arch.
IJSRA
International Journal of Science and Research Archive
Research Journal Archive, INDIA

(Research Article)

Efficient and interpretable monkeypox detection using vision transformers with explainable visualizations

Sanjida Akter [1], Mohammad Rasel Mahmud [2], Md Ariful Islam [3], Md Ismail Hossain Siddiqui [4, *] and Anamul Haque Sakib [3]

[1] *Department of Mathematics and Natural Sciences, BRAC University, Dhaka, Bangladesh.*
[2] *Department of Management Information System, International American University, CA 90010, USA.*
[3] *Department of Business Administration, International American University, CA 90010, USA.*
[4] *Department of Engineering/Industrial Management, Westcliff University, Irvine, CA 92614, USA.*

International Journal of Science and Research Archive, 2025, 15(01), 1811-1822

Publication history: Received on 13 March 2025; revised on 22 April 2025; accepted on 24 April 2025

Article DOI: https://doi.org/10.30574/ijsra.2025.15.1.1162

Abstract

Monkeypox is a zoonotic disease that poses diagnostic challenges due to its resemblance to other pox-type skin lesions like measles and chickenpox. Traditional deep learning (DL) methods, especially convolutional neural networks (CNNs), often struggle with generalization when trained on small, imbalanced datasets. These methods also tend to lack interpretability and computational efficiency, limiting their use in real-time, resource-constrained settings. This study introduces a lightweight, explainable DL framework based on EfficientFormerV2, which merges the advantages of convolutional inductive biases with efficient token-mixing strategies. We used the publicly available Monkeypox Skin Image Dataset (MSID), which contains 770 images across four categories: Monkeypox, Chickenpox, Measles, and Normal. Through advanced preprocessing and augmentation, we expanded the dataset to 4,000 images, improving class representation and reducing overfitting. Also, we evaluated five models—EfficientFormerV2, T2T-ViT, DeiT, Xception, and MobileNetV4—using metrics like F1-score, specificity, PR AUC, and Matthews Correlation Coefficient (MCC) with 10-fold stratified cross-validation. EfficientFormerV2 performed the best, achieving an F1-score of 98.73%, specificity of 99.63%, PR AUC of 99.86%, and MCC of 94.15%. We used Grad-CAM visualizations to create class-specific heatmaps for better interpretability. This framework combines an efficient architecture, data-centric augmentation, and explainable AI (XAI), offering high accuracy and low-latency predictions, making it suitable for real-time monkeypox screening, especially in low-resource settings.

Keywords: Skin Lesion; Vision Transformer; Hybrid Deep Learning; Explainable AI(XAI); Monkeypox

1. Introduction

Monkeypox is a viral zoonosis that is raising global health concerns due to its rapid spread [1]. As of late 2023, the World Health Organization (WHO) reported over 91,000 confirmed cases and more than 160 deaths across 114 countries, making this the largest outbreak since the virus was first identified in 1970 [2] [3]. Symptoms include pustular skin lesions, fever, lymphadenopathy, and fatigue, which can be similar to chickenpox and measles [4]. This similarity complicates diagnosis, particularly in non-endemic areas with limited access to laboratory testing. Early identification is crucial for containing transmission, enabling timely treatment, and reducing illness.

In response to the need for better diagnostic methods has led to the use of dermatological imaging and artificial intelligence (AI) for non-invasive, rapid screening and diagnosis. DL has shown promise in medical image classification within dermatology [5]. However, most current approaches are limited by small, imbalanced datasets, leading to

[*] Corresponding author: Md Ismail Hossain Siddiqui

overfitting and poor generalization. CNNs focus on localized features, which makes it hard to capture important long-range spatial relationships needed to differentiate between similar skin conditions. While transformer-based models can grasp global context better, they are computationally heavy and require large, labeled datasets—often unavailable for rare diseases like monkeypox. There is a substantial gap in efficient, explainable, and lightweight models for detecting monkeypox lesions, despite increased research on skin disease recognition. Few studies have explored hybrid transformer architectures that combine accuracy with computational efficiency, and even fewer have precisely tested these models on well-augmented and balanced datasets. Also, explainability methods like Grad-CAM are underused, which hampers transparency and clinical adoption.

This study aims to develop an effective and explainable DL framework for classifying monkeypox and pox-type skin lesions using a lightweight hybrid transformer architecture. The objectives are to: (1) use advanced data augmentation to address small and imbalanced datasets; (2) evaluate both convolutional and transformer models under the same conditions; and (3) incorporate XAI to enhance clinical interpretability; and (4) achieve low-latency, high-accuracy performance for real-time diagnostics in resource-limited environments.

To achieve our objectives, we introduce EfficientFormerV2, a hybrid vision transformer architecture that combines convolutional inductive biases with efficient token-mixing strategies to capture fine-grained and global features (Figure 1). We enhanced our dataset with a robust augmentation pipeline, increasing the number of images while ensuring balanced class representation. In our comparative evaluation, we included four benchmark models, trained using 10-fold stratified cross-validation and assessed with various metrics. Grad-CAM visualizations were employed for heatmap-based interpretability, enhancing clinical transparency. The following are the key contributions of our study:

- A hybrid diagnostic framework that combines convolutional inductive bias with transformer token-mixing strategies to achieve high accuracy and computational efficiency, ensuring stable, low-latency predictions for real-time web deployment in resource-limited environments.
- Utilized advanced preprocessing and data augmentation methods were used to expand the dataset which addressed class imbalance and enhanced training diversity.
- Grad-CAM integration provides interpretable and transparent predictions by visualizing specific attention regions related to lesions that enhance clinical reliability and model trustworthiness.
- Achieved state-of-the-art performance by outperforming previous models in both classification accuracy and prediction stability across multiple evaluation metrics.

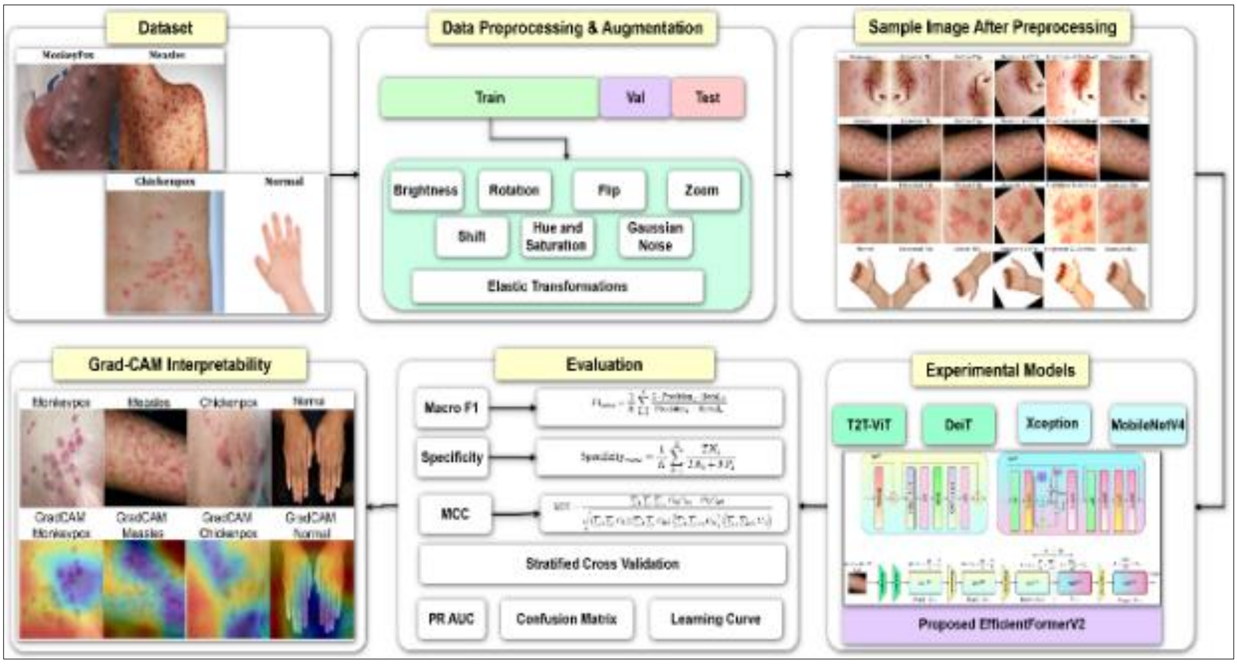

Figure 1 Proposed methodology

The remainder of this paper is structured as follows: Section 2 reviews related work. Section 3 describes the datasets, preprocessing techniques, and architecture. Section 4 presents the results along with a comparative analysis. Section 5 provides a critical discussion, and Section 6 concludes the study.

2. Related Works

Optimization-based CNN models have shown early success in monkeypox classification. Eliwa et al. [6] proposed a CNN model enhanced by the Grey Wolf Optimizer (GWO) for classifying monkeypox skin lesions. They utilized a Kaggle dataset containing 25,000 clinical samples across 11 features and achieved an accuracy of 95.31%. Despite its effectiveness, this approach faced limitations related to data imbalance, the risk of overfitting, and limited generalizability.

Attention mechanisms and evolutionary algorithms have been used to improve CNN-based models. Almars et al. [7] introduced DeepGenMon, a lightweight CNN integrated with attention mechanisms and a Genetic Algorithm (GA) for optimization. Evaluated on two datasets of 847 and 659 images across six and four classes respectively, it achieved accuracy of 98.5% and 98.2%. However, it lacked explainability and was not validated on large-scale datasets.

XAI techniques have been incorporated into classical CNN models. Abbas et al. [8] developed a VGG16-based classifier enhanced by Layer-wise Relevance Propagation (LRP) to detect monkeypox, measles, chickenpox, and normal skin lesions. Trained on 2,310 augmented images, the model attained an accuracy of 93.29%. Its incorporation of XAI improved interpretability, although the diversity of the dataset was limited.

Ensemble models combining multiple CNN architectures have also been explored. Muñoz-Saavedra et al. [9] utilized an ensemble of ResNet50, EfficientNet-B0, and MobileNetV2 along with Grad-CAM visualizations to classify monkeypox from a custom dataset of 300 images (across three classes). They reported an accuracy of 98.33% but noted that high computational costs and the limited dataset size were drawbacks.

Hybrid CNN-Transformer ensembles have pushed the boundaries of classification and interpretability. Saha et al. [10] proposed Mpox-XDE, which combined Xception, DenseNet201, and EfficientNetB7 models with SwinViT and Grad-CAM for enhanced detection. Trained on 770 images across four classes, the model achieved an accuracy of 98.70%. This study excelled in combining ensemble learning with explainability but faced challenges due to limited data and model complexity.

ViT-CNN fusion models have shown potential in multi-class monkeypox classification. Oztel et al. [11] presented an ensemble model that fused Vision Transformer and DenseNet201 architectures to classify monkeypox and six other skin conditions. Evaluated on the PAD-UFES-20 and MSLD datasets, which comprised 2,400 images, it achieved an accuracy of 81.91%. While the approach was innovative in combining fine-tuned ViT and bagging, it was hindered by the dataset size.

Transformer-only ensembles have demonstrated superior performance over CNNs in some settings. Vuran et al. [12] proposed a transformer-based ensemble model incorporating ViT, MAE, DINO, and SwinTransformer for multi-class skin lesion classification, including monkeypox. Using the MSLD v2.0 dataset of 755 images, the SwinTransformer achieved an accuracy of 93.71%. Although the superiority of transformers over CNNs was demonstrated, high computational costs and small data volumes remained a concern.

Multi-scale transformer-based models have introduced new architectural innovations. Huan et al. [13] introduced MSMP-Net, a ConvNeXt-based multi-scale model for end-to-end classification of monkeypox. Trained on the MSLD v2.0 dataset (755 images across six classes), the model reached an accuracy of 87.03%. Its innovation lies in multi-scale feature fusion; however, the absence of XAI and a small dataset limits its interpretability and scalability.

3. Materials and Methods

3.1. Data Description

This study utilized the MSID, a publicly available dataset curated by Bala and Hossain [14] and hosted on the Mendeley Data platform. The dataset contains a total of 770 high-resolution dermatological images categorized into four clinically relevant classes: Monkeypox (279 images), Chickenpox (107 images), Measles (91 images), and Normal (293 images). Each image showcases distinct skin lesion characteristics necessary for classifying visually similar pox-type conditions. Sample image form each class is shown in Figure 2. To ensure robust training and effective performance evaluation, the dataset was divided into training, validation, and test sets using a stratified splitting strategy in the ratio of 80:10:10. This approach maintained consistent class distribution across all subsets, minimizing bias during model evaluation. The final class-wise distribution after the split is summarized in Table 1.

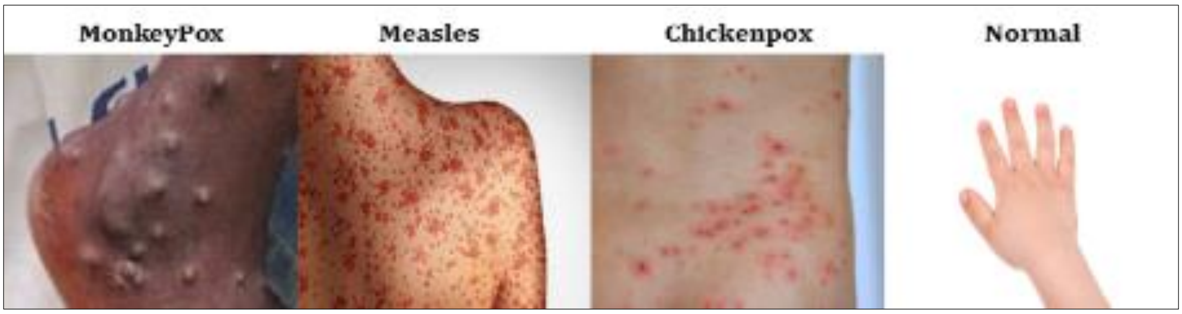

Figure 2 Sample image from each class of MSID dataset

Table 1 Class distribution after splitting the dataset

Class	Train	Validation	Test
Monkeypox	223	27	29
Chickenpox	85	10	12
Measles	72	9	10
Normal	234	29	30

3.2. Data Preprocessing and Augmentation

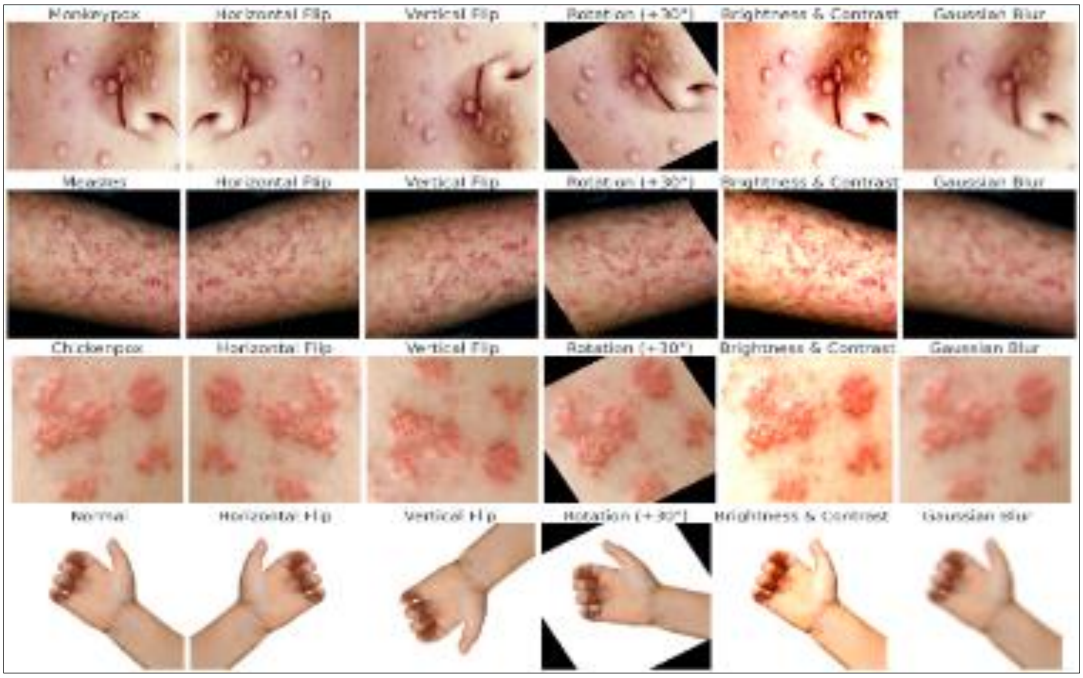

Figure 3 Sample augmented image from each class

All images in the dataset were resized to 224×224 pixels and normalized to a pixel intensity range of [0, 1]. These preprocessing steps were applied consistently across all subsets to ensure standardized input dimensions and facilitate efficient model training [15]. To address the limited sample size and class imbalance, we implemented a data augmentation strategy for the training set. This approach increased the number of samples in each class to 1,000, resulting in a balanced dataset of 4,000 images across four classes. This expansion helped reduce overfitting, especially in underrepresented classes like Measles and Chickenpox, and provided the DL models with diverse samples to learn from [16]. A balanced training set also minimizes class bias during optimization and improves the classifier's generalization across all categories. The augmentation pipeline used both geometric and photometric transformations. Geometric operations included random horizontal and vertical flipping, rotations of ±20 degrees, zooming by 10%, and shifting dimensions by up to 10%. These changes helped simulate different lesion orientations and sizes. For photometric adjustments, brightness and contrast were altered between 0.8 and 1.2, hue and saturation were adjusted, and Gaussian noise with a standard deviation of 0.01 to 0.05 was added. These modifications aim to mimic various lighting conditions and camera artifacts [17], [18]. Sample augmented image is illustrated in Figure 3.

Advanced techniques like random cutouts and elastic transformations were employed to help models learn more robust representations. This approach enables the models to focus on multiple lesion regions rather than just fixed areas. Elastic transformations simulate realistic non-linear distortions by applying a displacement field $\mathcal{D}(\boldsymbol{x})$ to each pixel location $(\boldsymbol{x})$. The displacement field is computed using Equation 1, where 34 is the deformation coefficient, a Gaussian kernel $(G_{4.5})$ with a standard deviation of 4.5, and a randomly initialized displacement field $\mathcal{N}(\boldsymbol{x})$. These transformations increased training diversity, reduced overfitting, and promoted the learning of generalizable features across different lesion patterns.

$$\mathcal{D}(\boldsymbol{x}) = \boldsymbol{x} + 34 \cdot G_{4.5} * \mathcal{N}(\boldsymbol{x}) \tag{1}$$

3.3. Experimental Models

To address the issue of limited labeled medical image data, this study employed advanced transfer learning models, including convolutional and transformer-based architectures. Each model was selected for its effectiveness in visual recognition, suitability for medical imaging, and efficiency in fine-tuning on moderate datasets.

3.3.1. Transfer Learning Models

The Tokens-to-Token Vision Transformer (T2T-ViT) was selected for its capacity to preserve local structural information through progressive tokenization, as well as its ability to capture global context [19]. In contrast to standard ViTs, which flatten image patches at an early stage, T2T-ViT utilizes multiple stages of soft splitting and attention mechanisms [20], [21]. This effectively models both low-level textures and high-level semantic patterns. DeiT employs knowledge distillation to learn rich feature embeddings without requiring extremely large-scale pretraining data [22]. Its efficient design and superior generalization capabilities make it ideal for tasks with limited annotated samples.

Xception is an extension of the Inception architecture that utilizes depthwise separable convolutions, allowing it to learn fine-grained spatial hierarchies while maintaining computational efficiency [23], [24]. This model has shown strong performance in medical imaging, particularly in dermatology, due to its ability to capture subtle variations in texture and shape. MobileNetV4 is chosen for its speed and efficiency in real-time applications. It's optimized for edge devices and mobile use, offering low-latency performance with minimal computational cost [25]. While it has a lower representational capacity than transformer-based models, its ease of deployment and compatibility with XAI tools make it a strong candidate for web-based diagnostic systems.

3.3.2. Proposed EfficientFormerV2

EfficientFormerV2 is a lightweight vision transformer designed for image classification. It balances accuracy and latency by combining convolutional inductive biases with efficient token-mixing strategies. The architecture features hierarchical feature extraction using MB^{4}D (Mobile Block 4D) and MB^{3}D (Mobile Block 3D) stages, optimizing performance and computational efficiency while capturing fine-grained local textures and global semantic context (Figure 4). The model processes an input image of size B×3×H×W, where B is the batch size. It passes through two convolutional stem layers that reduce spatial dimensions while improving feature representation by learning low-level elements like edges and textures.

The architecture starts with a convolutional stem, followed by MB^{4}D blocks in the first three stages. Each MB^{4}D block consists of a depthwise separable convolution unit with two 1×1 pointwise convolutions, batch normalization (BN), and a GeLU activation. Additionally, a residual pooling path captures extra spatial context. The blocks operate at decreasing spatial resolutions: H/4×W/4, H/8×W/8, and H/16×W/16, enabling the model to learn more abstract features. As the model advances, spatial features are flattened into tokens and processed through MB^{3}D blocks starting at Stage 4. Here, the features are reshaped into B×(HW/16) ^{2}×C tokens. The MB^{3}D block performs multi-head self-attention by calculating Query (Q), Key (K), and Value (V) vectors, followed by dot-product attention with softmax activation. The output is refined through linear projections, layer normalization (LN), GeLU activations, and feed-forward layers, allowing for global contextual learning across the image.

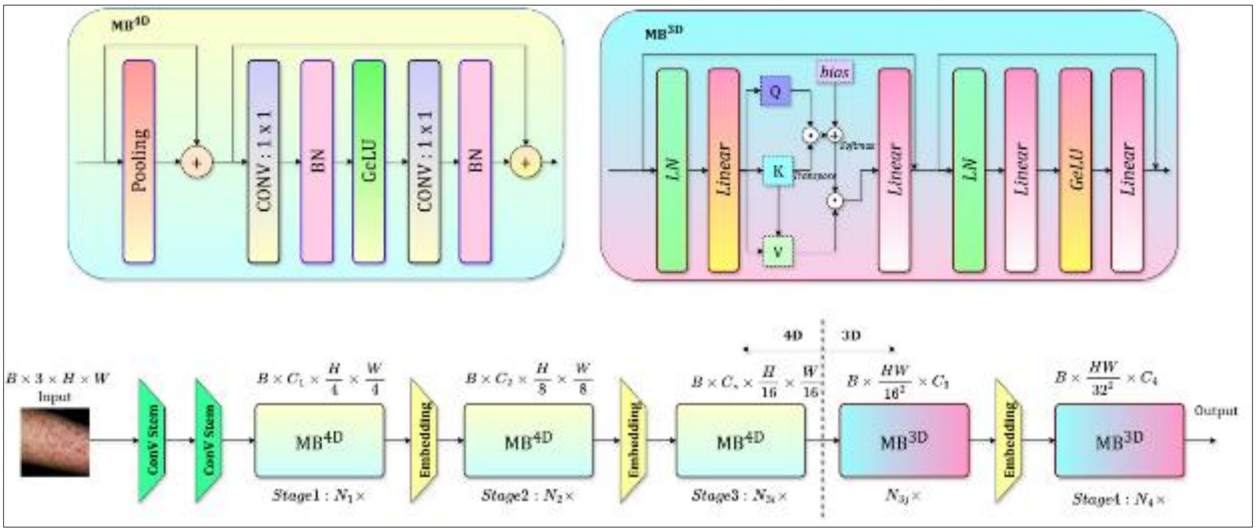
MB^{4D}
Pooling
CONV : 1 x 1
BN
GeLU
CONV : 1 x 1
BN
MB^{3D}
LN
Linear
Q
K
V
bias
Softmax
Linear
LN
Linear
GeLU
Linear
$B \times 3 \times H \times W$
Input
ConV Stem
ConV Stem
$B \times C_1 \times \frac{H}{4} \times \frac{W}{4}$
MB^{4D}
Stage1 : $N_1 \times$
Embedding
$B \times C_2 \times \frac{H}{8} \times \frac{W}{8}$
MB^{4D}
Stage2 : $N_2 \times$
Embedding
4D
3D
$B \times C_3 \times \frac{H}{16} \times \frac{W}{16}$
MB^{4D}
Stage3 : $N_{3_1} \times$
$B \times \frac{HW}{16^2} \times C_3$
MB^{3D}
$N_{3_2} \times$
Embedding
$B \times \frac{HW}{32^2} \times C_4$
MB^{3D}
Stage4 : $N_4 \times$
Output

Figure 4 Proposed EfficientFormerV2 architecture

Embedding layers are incorporated between stages to align feature dimensions and improve communication between blocks. Finally, the output tokens are aggregated and processed through a classification head to generate predictions for target classes. EfficientFormerV2 is designed for high classification accuracy and low inference latency, making it ideal for real-time web-based diagnostic systems in resource-limited settings [26], [27]. Its lightweight architecture ensures efficient deployment, and it works well with explainability methods like Grad-CAM for integration into clinical decision-making processes [28].

3.4. Evaluation and Training Parameters

Table 2 Hyperparameter ranges and selected values

Hyperparameter	Ranges	Selected Value
Learning Rate	{1e-5, 5e-5, 1e-4, 5e-4, 1e-3}	1e-4
Batch Size	{16, 32, 64, 128}	64
Dropout Rate	{0.1, 0.3, 0.5, 0.6}	0.3
Optimizer	{SGD, Adam, AdamW}	AdamW
Weight Decay	{1e-6, 1e-5, 1e-4, 5e-4}	1e-5
Learning Rate Schedule	{Constant, Step, Cosine, OneCycle}	Cosine Annealing
Warm-up Steps	{0, 100, 300, 500}	300
Max Epochs	{30, 50, 75, 100}	30
Early Stopping Patience	{3, 5, 7, 10}	7

The evaluation was conducted using key metrics: F1 Score, specificity, PR AUC, MCC. The F1 Score balanced precision and recall addressing class imbalances, while specificity evaluated the model's ability to correctly identify negative instances and minimize false positives. PR AUC provided insights into the trade-off between precision and recall at different thresholds. MCC combined true and false positives and negatives into a single score for multi-class evaluations. The confusion matrix helped visualize misclassification patterns among similar disease types. Learning curves for accuracy and loss were plotted over 30 epochs to track model performance, convergence, and potential issues like underfitting or overfitting. Furthermore, early stopping, a learning rate scheduler, and model checkpointing were applied to enhance generalization and reduce overfitting in our training process. We applied 10-fold stratified cross-validation to ensure reliable performance estimation, maintaining class distribution across folds during training and validation. We also conducted a hyperparameter search detailed in Table 2. We set the learning rate to 5e-4 to achieve fast and stable convergence while avoiding divergence. After testing batch sizes of 16, 32, 64, and 128, we chose 64 for optimal gradient stability and GPU efficiency. We implemented a dropout rate of 0.3 for regularization, effectively reducing overfitting without hindering learning. Among the optimizers tested, AdamW was the most effective due to its decoupled weight decay mechanism. We found that a weight decay of 1e-5 provided the best balance between generalization and learning dynamics. Cosine annealing was used for the learning rate schedule, resulting in smoother convergence than constant or step-based methods. We included 300 warm-up steps to stabilize initial gradients. The model training lasted a maximum of 30 epochs, with early stopping applied if validation performance did not improve for 7 consecutive epochs.

4. Results and Discussion

The evaluation of five advanced models was performed with and without data augmentation. Table 3 shows that data augmentation improved all evaluation metrics, enhancing generalization and reducing overfitting. EfficientFormerV2 outperformed the others, achieving an F1 score of 97.34% and a MCC of 93.92% before augmentation, and improving these to 98.73% and 94.15% after augmentation, along with a specificity of 99.63%. The low standard deviations indicate the model's high stability and consistency. These findings demonstrate EfficientFormerV2's effectiveness in distinguishing visually similar pox-type lesions.

T2T-ViT demonstrated a notable increase in performance after augmentation, with its PR AUC rising from 97.18% to 98.90%. It maintained a high MCC, demonstrating strong predictive abilities and a good balance between sensitivity and specificity. DeiT improved its F1-Score from 95.57% to 97.41%, though it experienced a slight decrease in MCC, indicating some instability in class handling. Xception's MCC increased from 90.89% to 91.53%, but it still fell short compared to transformer-based models, highlighting their superior ability to model intricate visual patterns. MobileNetV4, the lightest model, also improved with augmentation, but its lower MCC and higher variability suggests it is less suitable for high-stakes diagnostics, even though it's beneficial for use in resource-constrained environments.

Table 3 Performance comparison of all models before and after augmentation

Augmentation Status	Model	Specificity	F1-Score	PR AUC	MCC
Before	EfficientFormerV2	98.19 ± 0.31	97.34 ± 0.29	98.42 ± 0.28	93.92 ± 0.22
	T2T-ViT	97.04 ± 0.47	96.59 ± 0.49	97.18 ± 0.27	93.30 ± 0.33
	DeiT	95.96 ± 0.65	95.57 ± 0.17	95.82 ± 0.46	92.76 ± 0.26
	Xception	94.61 ± 0.98	94.11 ± 1.21	95.19 ± 0.83	90.89 ± 1.07
	MobileNetV4	93.31 ± 1.13	93.58 ± 1.32	93.92 ± 0.98	89.29 ± 1.30
After	EfficientFormerV2	99.63 ± 0.24	98.73 ± 0.13	99.86 ± 0.13	94.15 ± 0.13
	T2T-ViT	98.64 ± 0.46	98.22 ± 0.44	98.90 ± 0.01	93.61 ± 0.27
	DeiT	97.76 ± 0.42	97.41 ± -0.01	97.66 ± 0.21	92.58 ± 0.15
	Xception	96.59 ± 0.13	96.25 ± 0.54	97.11 ± 0.25	91.53 ± 0.46
	MobileNetV4	95.25 ± 1.02	95.60 ± 0.78	96.02 ± 0.94	90.62 ± 1.10

The standard deviation is a key indicator of model performance stability in 10-fold cross-validation. EfficientFormerV2 not only achieved the highest performance scores but also had the lowest standard deviation across all metrics after augmentation, with a ±0.13 standard deviation in the F1-Score, indicating very stable predictions. T2T-ViT and DeiT showed low variability as well, suggesting good generalization. In contrast, MobileNetV4 and Xception had higher standard deviations, particularly in the MCC and F1-Score, indicating greater sensitivity to specific training subsets and less reliability in complex classification tasks.

The learning curves of the EfficientFormerV2 model, shown in Figure 5, highlight its training dynamics, convergence behavior, and generalization performance over 30 epochs, both before and after applying data augmentation. The loss curves indicate that the model trained on the non-augmented dataset shows a consistent decrease in training loss. However, the validation loss varies widely, especially between epochs 5 and 20, suggesting difficulty in generalization due to the original dataset's limitations and imbalances. The growing gap between training and validation loss in later epochs points to mild overfitting, where the model is too tailored to the training data, leading to poor performance on new samples. In contrast, after implementing a strong data augmentation pipeline, the loss curves show improved training stability. Both training and validation loss decrease steadily with less fluctuation, and the gap between the two curves narrows. This indicates that the augmented dataset has improved the model's ability to generalize across different skin lesion patterns. The increase in diversity and balance in the training data helps the model avoid overfitting and learn more representative features across all classes.

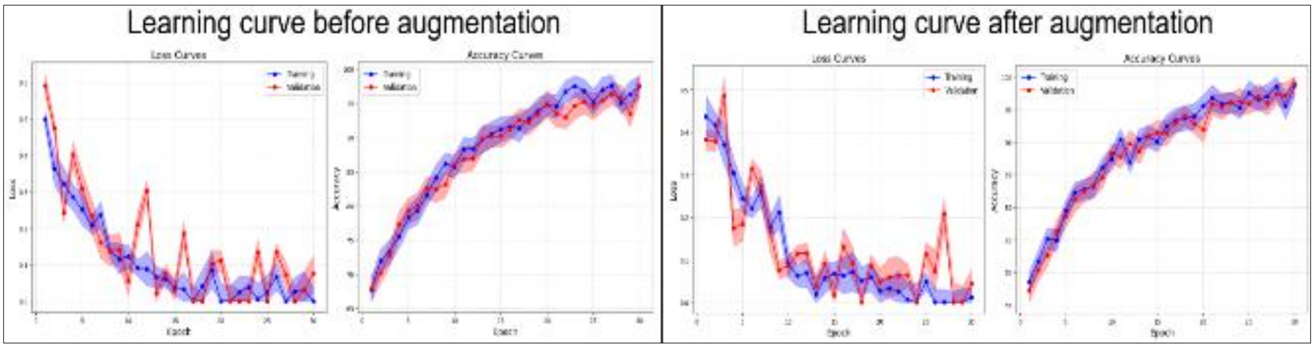

Figure 5 Learning curve of the proposed EfficientFormerV2 before and after augmentation

Before augmentation, the training and validation accuracy increased steadily, but the validation accuracy lagged and showed notable variability, particularly between epochs 15 and 25, indicating inconsistency. After augmentation, both accuracy curves progressed nearly in parallel, surpassing 98% by the end of training. The close alignment and minimal variance between the curves demonstrate strong generalization, confirming that the augmented data improved learning efficiency and enhanced classification reliability.

The confusion matrices before and after data augmentation provide a clear overview of the class-wise prediction performance of the EfficientFormerV2 model across all the categories (Figure 6). Before augmentation, the model had several misclassifications. In the Monkeypox category, it correctly identified 192 out of 194 samples, misclassifying one as Chickenpox and another as Normal. The Chickenpox category had 76 correct predictions out of 80, with two misclassified as Measles and one each as Monkeypox and Normal. For Measles, the model accurately classified 64 out of 67 samples, misclassifying one as Chickenpox and two as Normal. The Normal category performed well, with only one error, where a sample was misclassified as Monkeypox.

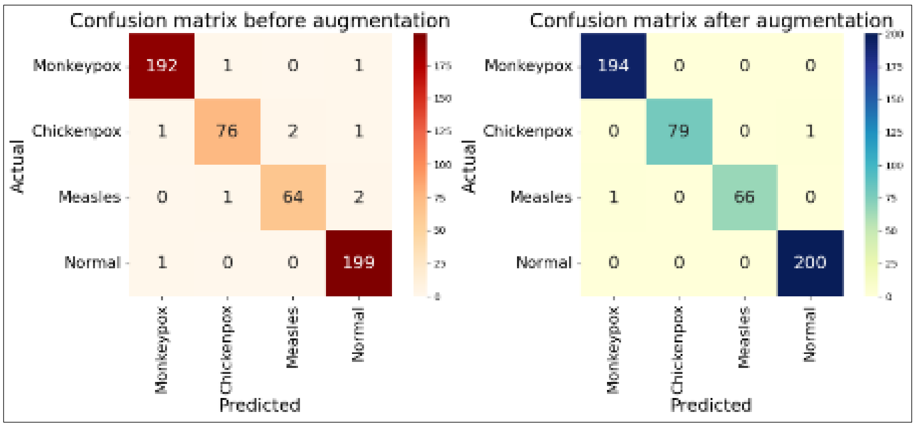

Figure 6 Confusion matrix of the proposed EfficientFormerV2 before and after augmentation

After implementing data augmentation, the model's performance improved significantly. Both Monkeypox and Normal categories achieved 100% accuracy, eliminating all previous misclassifications. Chickenpox had one error, misclassifying one sample as Normal, while Measles had a single misclassification as Monkeypox. These results demonstrate the positive impact of data augmentation on the model's ability to distinguish between classes, particularly in diseases with subtle visual differences. The reduced confusion between Chickenpox and Measles indicates that the model has developed better class-specific features from the varied training data.

The Grad-CAM visualization demonstrates (Figure 7) how the EfficientFormerV2 model interprets four classes of skin lesions: Monkeypox, Measles, Chickenpox, and Normal. The top row displays the original images, while the bottom row presents heatmaps that highlight the region's most influential to the model's predictions. For Monkeypox, the model shows strong activation over clustered pustular lesions, effectively capturing key features such as raised vesicles. The heatmap for Measles reveals a broader attention pattern, consistent with the diffuse nature of its rash, showcasing the model's ability to identify subtle visual cues. In the case of Chickenpox, the focus is on grouped vesicles, but with less intensity than Monkeypox, which may contribute to occasional misclassifications. For the Normal class, the heatmap shows minimal activation, confirming that the model correctly identifies healthy skin without being influenced by non-lesion areas.

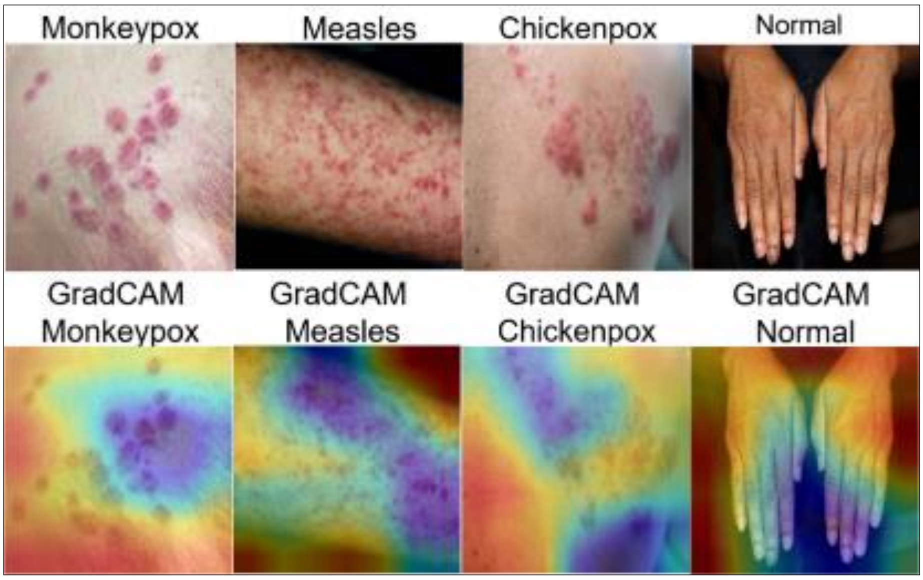

Figure 7 Grad-CAM visualizations highlighting lesion-specific model attention

Table 4 compares various models for classifying monkeypox and pox-type skin lesions. Among the models listed, the proposed EfficientFormerV2 model stands out with a performance score of 99.86% on a balanced dataset of 4,000 images, outperforming all other models in accuracy and dataset size. Moreover, it incorporates XAI features like Grad-CAM, which helps in interpreting AI decisions in clinical settings. The use of diverse augmentation techniques likely enhanced the model's generalization and robustness.

Table 4 Performance comparison of existing models with proposed model

Model	Data	Result	Augmentation	XAI
ViT + DenseNet201 Ensemble [11]	2,400	81.91	Yes	No
Mpox-XDE [10]	770	98.7	Yes	Yes
Ensemble CNN [9]	300	98.33	Yes	Yes
Ensemble [12]	755	93.71	Yes	No
VGG16 + LRP [8]	2,310	93.29	Yes	Yes
DeepGenMon [7]	1,506	98.5	Yes	No
MSMP-Net [13]	755	87.03	Yes	No
Our proposed EfficientFormerV2	4,000	99.86	Yes	Yes

In comparison, models like Mpox-XDE and DeepGenMon achieved high performance scores of 98.7% and 98.5%, respectively, but they were trained on smaller datasets. DeepGenMon's lack of explainability limits its clinical transparency. The Ensemble CNN model performed well with a score of 98.33%, but with only 300 images, its scalability and generalization are questionable. Other models, such as VGG16 + LRP, Ensemble, and MSMP-Net, scored between 87.03% and 93.71%, despite using data augmentation and some explainability tools. The ViT + DenseNet201 Ensemble scored 81.91%, possibly due to inadequate training or dataset issues.

Our proposed framework outperforms traditional architectures with its hybrid design, merging convolutional inductive biases and efficient token mixing through Mobile Block 4D (MB^{4}D) and Mobile Block 3D (MB^{3}D). This structure enables fine-grained local feature extraction early on and models long-range dependencies with multi-head self-attention in deeper layers. Unlike standard Vision Transformers, EfficientFormerV2 uses spatially aware tokenization after convolutional encoding to preserve essential morphological details in skin lesions. The architecture features depthwise separable convolutions, residual pooling paths, GeLU activations, and LayerNorm, resulting in improved representational efficiency, low latency, and stable convergence, validated by stratified cross-validation.

Our data preprocessing and augmentation pipeline enhanced generalization and reduced overfitting. We normalized inputs and applied geometric (rotation, shifting, zooming) and photometric (brightness, hue, contrast) augmentations to mimic real-world variability. Additional transformations like Gaussian noise and elastic distortion increased intra-class diversity, particularly for similar categories like Chickenpox and Measles. Techniques such as cutout and color jittering improved global lesion context learning and minimized over-reliance on localized features. This balanced training set resulted in significant improvements across all evaluation metrics, demonstrating the pipeline's effectiveness in developing robust decision boundaries with limited data.

The proposed framework combines data-driven learning with explainability and deployment efficiency. EfficientFormerV2 features a low parameter count and token-efficient attention, making it suitable for real-time diagnostics on edge devices. Its integration with Grad-CAM provides saliency maps that highlight lesion-relevant regions, enhancing diagnostic validation and regulatory compliance while fostering user trust in AI-assisted healthcare tools. With a low computational footprint and minimal memory requirements, EfficientFormerV2 is ideal for mobile and point-of-care systems, especially in resource-limited settings. Its quick inference capabilities and clear outputs support dermatologist-AI collaboration for fast monkeypox screening, reducing the need for expert interpretation or lab testing. The system's real-time interpretable predictions make it a strong candidate for scalable public health surveillance.

Despite these advantages, several technical limitations persist. It was trained and validated on a single dataset (MSID), limiting its exposure to variations between clinics and diverse demographics. While Grad-CAM provides explanations, it doesn't offer causal interpretability and is sensitive to noise and input changes. Additionally, EfficientFormerV2's fixed receptive fields and static patch tokenization may perform poorly on lesions with irregular or varying shapes. It also doesn't address robustness against domain shifts, adversarial attacks, or changes in real-world data distribution, which are essential for reliable deployment. Lastly, the assumption of independent and identically distributed (i.i.d.) data may not hold true in multi-site clinical settings.

Future work will focus on improving domain generalization and adaptation through techniques like CORAL, MMD-based alignment, and adversarial feature disentanglement to handle domain shifts. We will enhance spatial modeling using deformable attention modules and scale-invariant convolutions for better detection of irregular lesion structures. We also plan to explore interpretable architectures such as ProtoPNet and Concept Bottleneck Models (CBMs) for clearer decision-making. To reduce reliance on large, labeled datasets, we will implement self-supervised learning methods, including contrastive frameworks (MoCo, SimCLR) and masked image modeling. Lastly, we will evaluate certified adversarial robustness using interval-bound propagation and randomized smoothing to ensure safe deployment in high-risk clinical settings.

5. Conclusion

This study presents a new hybrid DL framework for monkeypox and pox-type lesions recognition using the efficient EfficientFormerV2 architecture. Our approach combines convolutional inductive biases with token-efficient attention, achieving high accuracy and low inference latency, making it viable for real-time clinical use. We implemented a strong augmentation pipeline and stratified training to tackle issues of limited and imbalanced datasets, enhancing the model's generalization and stability. Grad-CAM integration provides transparency by highlighting key areas crucial for diagnosis. Our model outperforms current methods in identification and supports XAI in dermatology. This framework shows promise for trustworthy AI-assisted diagnosis in resource-limited settings. Future work will focus on domain adaptation and interpretable architectures to improve reliability and clinical application.

Compliance with ethical standards

Disclosure of conflict of interest

There is not conflict of interests.

References

[1] D. Kmiec and F. Kirchhoff, "Monkeypox: A New Threat?," International Journal of Molecular Sciences 2022, Vol. 23, Page 7866, vol. 23, no. 14, p. 7866, Jul. 2022, doi: 10.3390/IJMS23147866.

[2] B. Moss, "Understanding the biology of monkeypox virus to prevent future outbreaks," Nature Microbiology 2024 9:6, vol. 9, no. 6, pp. 1408–1416, May 2024, doi: 10.1038/s41564-024-01690-1.

[3] M. O. Meo, M. Z. Meo, I. M. Khan, M. A. Butt, A. M. Usmani, and S. A. Meo, "Rising epidemiological trends in prevalence and mortality of mpox: Global insights and analysis," Saudi Med J, vol. 45, no. 12, p. 1334, Dec. 2024, doi: 10.15537/SMJ.2024.45.12.20240720.

[4] A. Al Noman et al., "Monkeypox Lesion Classification: A Transfer Learning Approach for Early Diagnosis and Intervention," Proceedings of International Conference on Contemporary Computing and Informatics, IC3I 2024, pp. 247–254, 2024, doi: 10.1109/IC3I61595.2024.10828678.

[5] J. Hasan et al., "Transforming Leukemia Classification: A Comprehensive Study on Deep Learning Models for Enhanced Diagnostic Accuracy," PEEIACON 2024 - International Conference on Power, Electrical, Electronics and Industrial Applications, pp. 266–271, 2024, doi: 10.1109/PEEIACON63629.2024.10800693.

[6] E. H. I. Eliwa, A. M. El Koshiry, T. Abd El-Hafeez, and H. M. Farghaly, "Utilizing convolutional neural networks to classify monkeypox skin lesions," Scientific Reports 2023 13:1, vol. 13, no. 1, pp. 1–20, Sep. 2023, doi: 10.1038/s41598-023-41545-z.

[7] A. M. Almars, "DeepGenMon: A Novel Framework for Monkeypox Classification Integrating Lightweight Attention-Based Deep Learning and a Genetic Algorithm," Diagnostics 2025, Vol. 15, Page 130, vol. 15, no. 2, p. 130, Jan. 2025, doi: 10.3390/DIAGNOSTICS15020130.

[8] S. Abbas, F. Ahmed, W. A. Khan, M. Ahmad, M. A. Khan, and T. M. Ghazal, "Intelligent skin disease prediction system using transfer learning and explainable artificial intelligence," Scientific Reports 2025 15:1, vol. 15, no. 1, pp. 1–13, Jan. 2025, doi: 10.1038/s41598-024-83966-4.

[9] 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 2023, Vol. 23, Page 7134, vol. 23, no. 16, p. 7134, Aug. 2023, doi: 10.3390/S23167134.

[10] D. Kumar Saha et al., "Mpox-XDE: an ensemble model utilizing deep CNN and explainable AI for monkeypox detection and classification," BMC Infect Dis, vol. 25, no. 1, pp. 1–19, Dec. 2025, doi: 10.1186/S12879-025-10811-Y/TABLES/9.

[11] G. Yolcu Oztel, "Vision transformer and CNN-based skin lesion analysis: classification of monkeypox," Multimed Tools Appl, vol. 83, no. 28, pp. 71909–71923, Aug. 2024, doi: 10.1007/S11042-024-19757-W/TABLES/3.

[12] 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 2025, Vol. 15, Page 374, vol. 15, no. 3, p. 374, Feb. 2025, doi: 10.3390/DIAGNOSTICS15030374.

[13] E. Huan and H. Dun, "MSMP-Net: A Multi-Scale Neural Network for End-to-End Monkeypox Virus Skin Lesion Classification," Applied Sciences 2024, Vol. 14, Page 9390, vol. 14, no. 20, p. 9390, Oct. 2024, doi: 10.3390/APP14209390.

[14] D. Bala and M. S. Hossain, "Monkeypox Skin Images Dataset (MSID)," vol. 6, 2023, doi: 10.17632/R9BFPNVYXR.6.

[15] R. Haque et al., "Advancing Early Leukemia Diagnostics: A Comprehensive Study Incorporating Image Processing and Transfer Learning," BioMedInformatics 2024, Vol. 4, Pages 966-991, vol. 4, no. 2, pp. 966–991, Apr. 2024, doi: 10.3390/BIOMEDINFORMATICS4020054.

[16] A. Al Noman et al., "Monkeypox Lesion Classification: A Transfer Learning Approach for Early Diagnosis and Intervention," 2024 7th International Conference on Contemporary Computing and Informatics (IC3I), pp. 247–254, Sep. 2024, doi: 10.1109/IC3I61595.2024.10828678.

[17] M. D. Hosen et al., "Parasitology Unveiled: Revolutionizing Microorganism Classification Through Deep Learning," 2024 6th International Conference on Electrical Engineering and Information &amp; Communication Technology (ICEEICT), pp. 1163–1168, May 2024, doi: 10.1109/ICEEICT62016.2024.10534322.

[18] A. Al-Sakib, F. Islam, R. Haque, M. B. Islam, A. Siddiqua, and M. M. Rahman, “Classroom Activity Classification with Deep Learning,” 2nd International Conference on Integrated Circuits and Communication Systems, ICICACS 2024, 2024, doi: 10.1109/ICICACS60521.2024.10498187.

[19] Y. Wang, Y. Deng, Y. Zheng, P. Chattopadhyay, and L. Wang, “Vision Transformers for Image Classification: A Comparative Survey,” Technologies 2025, Vol. 13, Page 32, vol. 13, no. 1, p. 32, Jan. 2025, doi: 10.3390/TECHNOLOGIES13010032.

[20] S. Akça, Z. Garip, E. Ekinci, and F. Atban, “Automated classification of choroidal neovascularization, diabetic macular edema, and drusen from retinal OCT images using vision transformers: a comparative study,” Lasers Med Sci, vol. 39, no. 1, pp. 1–8, Dec. 2024, doi: 10.1007/S10103-024-04089-W/TABLES/4.

[21] S. B. Vijayakumar, K. T. Chitty-Venkata, K. Arya, and A. K. Somani, “ConVision Benchmark: A Contemporary Framework to Benchmark CNN and ViT Models,” AI 2024, Vol. 5, Pages 1132-1171, vol. 5, no. 3, pp. 1132–1171, Jul. 2024, doi: 10.3390/AI5030056.

[22] Z. Hu et al., “Weakly Supervised Classification for Nasopharyngeal Carcinoma with Transformer in Whole Slide Images,” IEEE J Biomed Health Inform, 2024, doi: 10.1109/JBHI.2024.3422874.

[23] Md. R. Ahmed et al., “Towards Automated Detection of Tomato Leaf Diseases,” 2024 6th International Conference on Electrical Engineering and Information &amp; Communication Technology (ICEEICT), pp. 387–392, May 2024, doi: 10.1109/ICEEICT62016.2024.10534559.

[24] M. S. Rahman et al., “Effective Disease Recognition in Cucumbers: A Web-Based Application Using Transfer Learning Models,” 2024 IEEE 3rd International Conference on Robotics, Automation, Artificial-Intelligence and Internet-of-Things, RAAICON 2024 - Proceedings, pp. 59–64, 2024, doi: 10.1109/RAAICON64172.2024.10928353.

[25] R. Haque et al., “Advancements in Jute Leaf Disease Detection: A Comprehensive Study Utilizing Machine Learning and Deep Learning Techniques,” 2024 IEEE International Conference on Power, Electrical, Electronics and Industrial Applications (PEEIACON), pp. 248–253, Sep. 2024, doi: 10.1109/PEEIACON63629.2024.10800378.

[26] Y. R. Chen, C. C. Chen, C. F. Kuo, and C. H. Lin, “An efficient deep neural network for automatic classification of acute intracranial hemorrhages in brain CT scans,” Comput Biol Med, vol. 176, p. 108587, Jun. 2024, doi: 10.1016/J.COMPBIOMED.2024.108587.

[27] E. B. Hamdi and Hidayaturrahman, “Ensemble of pre-trained vision transformer models in plant disease classification, an efficient approach,” Procedia Comput Sci, vol. 245, no. C, pp. 565–573, Jan. 2024, doi: 10.1016/J.PROCS.2024.10.283.

[28] D. Varam, L. Khalil, and T. Shanableh, “On-Edge Deployment of Vision Transformers for Medical Diagnostics Using the Kvasir-Capsule Dataset,” Applied Sciences 2024, Vol. 14, Page 8115, vol. 14, no. 18, p. 8115, Sep. 2024, doi: 10.3390/APP14188115.