iJRASET
International Journal For Research in Applied Science and Engineering Technology



INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 14 **Issue:** IV **Month of publication:** April 2026

DOI: https://doi.org/10.22214/ijraset.2026.79345

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 14 Issue IV Apr 2026- Available at www.ijraset.com

MpoxNetV: Revolutionizing Monkeypox Detection with Hybrid Transformer-CNN Architecture

Rajvi Shaileshkumar Patel[1], Pearl Wardani[2]
[1, 2] *Department of Computer Engineering, Bhagwan Mahavir University (BMU), Surat, India*

Abstract: ***Monkeypox (mpox) is a re-awakening Animal-borne disease delineated by skin lesions that are resembling to other dermatological disorders, making accurate and prompt diagnosis problematic. Automated medical image interpretation using deep learning has shown over the approaches often struggle with very limited and imbalanced datasets which problems in getting accuracy. In this study, the proposed hybrid deep learning framework technique which helps and that integrates transformer-based global feature extraction which also uses convolutional neural network (CNN)-based local feature learning for monkeypox disease classification. The recommended architecture model compounds a distilled Vision Transformer (DeiT) and EfficientNet-B4 through a learnable cross-attention gating mechanism, enabling adaptive fusing of global and local description. Furthermore, in enhancement and performance classification for datasets, utilizing a combined loss function that integrates cross-entropy and focal loss, supplemented by dynamic class weighting. A three phase training is used where 1) unfreezing network layers is adopted to stabilize training and improve generalization, 2) procedures are conducted with the publicly available monkeypox skin lesion datasets, examined through stratified k-fold cross validation method, 3) the model is evaluated using accuracy, F1- score, and other classification metrics. This paper demonstrates the effectiveness of hybrid transformer-CNN architectures with adaptive feature for monkeypox image classification and presents a model for real-world diagnostic applications, while also addressing problems related to dataset quality and generalization.***
Keywords: Monkeypox, Hybrid CNN, Image Classification, DeiT, Efficient Net.

I. INTRODUCTION

Precisely the differentiation of visually similar diagnose disease remains very challenging. Highly Monkey Pox(mpox), chickenpox, and measles remain very clinically challenging problem. These diseases often present with similar skin symptoms, such as blister-like eruptions, red inflamed patches, white inflamed patches and comparable patterns of spread across the body. Clinical misdiagnosis delays vital treatment and heightens the risk of uncontrolled outbreaks with greater harm. Even though PCR testing offers reliable results, its extended processing duration and dependence on complex lab setups hindering on-the-spot screening in low-resource contexts. Automated visual categorization provides promising solution, but the problem is more complex than standard skin lesion benchmarks. The four-class setting — monkeypox, chickenpox, measles, and normal skin demand a model that can simultaneously distinguish between three visually confusable pathological classes while correctly excluding healthy skin. The visual similarity across classes—together with limited available training data—poses unique difficulties for standalone architectures CNNs are great at spotting fine details like textures and edges but lack the ability to reason about global spatial context — which matters when lesion distribution patterns are diagnostically relevant. Transformer-based models address this through self-attention over image patches, but they are data-hungry and tend to underfit on small medical datasets when local feature granularity is required. Neither architecture alone is well-suited to this problem.

We introduce MpoxNetV, a smart hybrid model that addresses this gap directly. One branch uses DeiT—a distilled Vision Transformer—to capture how image patches relate across the entire picture, leveraging both its classification and distillation tokens for rich global context. The other branch employs EfficientNet-B4 to zoom in on fine local textures and details through convolution. Instead of just smashing these branches together, we add a clever cross-attention gate. This learned mechanism generates dynamic weights (and, always constrained up to 1) for each image, letting the model automatically decide:" Should I trust the global pattern more here, or focus on these local textures?" It adapts on-the fly rather than using a one-size-fits-all approach. Training is conducted in three progressive phases: first keeping the core networks frozen, then gradually unlocking parts of them, and finally fine-tuning everything together. We also use a special loss function that mixes standard error correction with extra focus on rare disease cases to balance our uneven dataset— combined with a custom loss function that blends cross entropy with a focal-style penalty term to handle class imbalance across the four categories.

©IJRASET: All Rights are Reserved

Evaluation uses stratified 5-across the full dataset. contributions of this paper are:

1) MpoxNetV: A hybrid architecture combining DeiT and EfficientNet-B4, specifically designed to classify monkeypox, chickenpox, measles, and healthy skin
2) Adaptive Gating Mechanism: A cross-attention module that dynamically balances global patterns and local features for each input image
3) Training Strategy: Focal Cross-Entropy loss with three phase progressive unfreezing, optimized for small and imbalanced medical datasets
4) Evaluation Protocol: 5-fold stratified cross-validation demonstrating robust and consistent performance across all four classes fold cross-validation to ensure reliable performance estimates

II. LITERATURE REVIEW

A. CNN-Based Approaches for Monkeypox Detection

In a 2023 study by Jaradat and colleagues, researchers tested five ready-made CNN models—VGG19, VGG16, ResNet50, MobileNetV2, and EfficientNetB3—for spotting monkeypox skin lesions. They used transfer learning to adapt these models. MobileNetV2 came out on top, hitting 98.16That said, these strong results came from simpler setups with just binary or few-class tasks. Some experts question them because of issues with the dataset's quality. Overall, the work shows lightweight models like MobileNetV2 shine in basic scenarios, but it does tackle tougher real-world problems—like distinguishing monkeypox from similar-looking diseases in a proper medical test setup.

B. Dataset Quality and Validity Concerns

Vega et al. (2023) conducted a critical analysis of the most widely used publicly available monkeypox skin image datasets, finding that the images were extracted from google and other websites and photography repositories via web scraping and lack any medical validation. Several studies that employed this dataset claimed extraordinary classification performance for monkeypox, chickenpox, and measles using machine learning, yet the authors demonstrated through rebuttal experiments that model performance did not necessarily derive from disease-relevant visual features. This has real lessons for current research: it motivates for better methods like stratified cross-validation (instead of basic splits), balanced loss functions to avoid models cheating by easy classes, and a healthy scepticism toward sky-high accuracies from studies relying on that unverified dataset.

C. Transformer Architectures in Medical Imaging

Al-Hammuri et al. (2023) provided a comprehensive survey of Vision Transformer (ViT) applications in digital health, covering image segmentation, classification, detection, and reconstruction tasks, concluding that ViT-based models represent a state-of-the-art approach for medical image analysis. springer open the survey documents the fundamental advantage of self-attention mechanisms over CNNs in capturing global spatial dependencies — critical when lesion distribution patterns carry diagnostic information beyond local texture. However, the survey also notes that transformer models face challenges in small-data medical imaging regimes, where limited training samples restrict effective patch-level attention learning. This motivates the hybrid design of MpoxNetV, where the convolutional branch (EfficientNet-B4) compensates for this limitation through robust local feature extraction even under data constraints.

D. Foundational Transformer Architecture

The ViT architecture proposed by Dosovitskiy et al. (2020) established the core framework for applying self-attention to image classification by dividing images into fixed-size patches and processing them as token sequences—analogous to words in a natural language sentence. This work demonstrated that pure transformer architectures can match or exceed CNN performance on large-scale benchmarks, but requires substantially more training data to generalize effectively. MpoxNetV builds on this foundation by adopting DeiT — a data-efficient ViT variant — and specifically exploits both its classification token and distillation token as complementary global representations, which are concatenated before projection. This design choice distinguishes MpoxNetV from prior works that use only the classification token, allowing richer global feature encoding within the same backbone.

E. Gap Analysis and Positioning of Proposed Work

The reviewed literature reveals three consistent limitations. First, CNN-only approaches such as those by Uysal and Jaradat et al. lack global contextual reasoning, limiting discrimination between diseases with similar local texture but different spatial patterns.

Second, transformer-only approaches are data-hungry and underperform on the small, imbalanced datasets typical in monkeypox research. Third, and most critically, dataset quality issues identified by Vega et al. cast doubt on inflated accuracy figures reported across the field, making rigorous evaluation methodology as important as architectural novelty. MpoxNetV addresses all three gaps: the DeiT–EfficientNet-B4 hybrid captures both global and local features; the adaptive cross-attention gate dynamically weights each branch per input sample rather than using fixed fusion weights; and the multi-phase training strategy with combined focal-CE loss and stratified 5-fold cross-validation ensures that reported results reflect genuine generalization rather than dataset artifact.

III.METHODOLOGY

A. Model Architecture

The proposed MpoxNetV model uses a smart two-part design that combines pre-trained vision transformers (great for capturing big-picture patterns) and convolutional neural networks (strong at spotting fine details). It blends their outputs using a special cross-attention technique to make smarter predictions.

DeiT (Data-efficient Image Transformers): We utilize deit base distilled patch16 224 for its ability to learn robust representations with limited data. Features are extracted by concatenating the class token and distillation token outputs, followed by projection into a 1024-dimensional space using a linear layer, Batch Normalization, and GELU activation:

$$G - GELU\left(BN(W_G * [z_{cls}; z_{dist}])\right) \in \mathbb{R}^{1024} \quad (1)$$

EfficientNet Branch (Local Features): An EfficientNet-B4 architecture is chosen for its strong performance and parameter efficiency. The extracted CNN features are projected into the same 1024-dimensional space:

$$L - GELU\left(BN(W_l * f_{cnn})\right) \in \mathbb{R}^{1024} \quad (2)$$

Cross-Attention Gate (Adaptive Fusion): The two feature representations are combined using a Cross Attention Gate, learnable mechanism that computes dynamic weights α and β (where α+β = 1) to adaptively fuse global and local features:

$$h = \alpha \cdot G + \beta \cdot L, \alpha + \beta = 1 \quad (3)$$

Classification Head: The fused 1024-dimensional feature vector h is passed through a classification head consisting of a linear layer, GELU activation, and dropout (rate = 0.35), followed by a final linear layer that maps the features to N = 4 output classes.

B. Combined Loss Function

To address class imbalance and enhance discrimination between visually similar classes, a Combined Loss function is employed:

$$L = 0.7 * L_{CE} + 0.3 * (1 - e^{-L_{ce}})^{\gamma} * L_{CE} \quad (4)$$

where L_{CE} is the class-weighted cross-entropy loss, and $\gamma = 2.0$ is the focal focusing parameter that down-weights well-classified examples. Class weights are set inversely proportional to class frequency to further mitigate data imbalance.

C. Three-Phase Training Strategy

Within each cross-validation fold, training follows three progressive phases:

Phase 1 — Feature Freezing (2 epochs): The DeiT and EfficientNet backbones were initially frozen, allowing only the projection layers, cross-attention gate, and classification head to be trained for 2 epochs. This helps in adapting the new layers to the pre-trained features.

Phase 2 — Partial Unfreezing (3 epochs): The final layers of both backbones (last 4 blocks of DeiT and last 3 stages of EfficientNet) were unfrozen. This phase, lasting 3 epochs, allows for fine-tuning of the most task-relevant features while keeping the majority of the pre-trained weights stable.

Phase 3 — Full Unfreezing (up to 5 epochs): All layers of the entire MpoxNetV model were unfrozen for comprehensive fine-tuning. This phase ran for up to 5 epochs.

D. Data Preprocessing and Augmentation

All input images are resized to 224 × 224 pixels and normalized using the ImageNet mean and standard deviation. Dynamic data augmentation is applied during training to improve generalization and mitigate overfitting. The applied transformations include:

- Random rotation (±20◦)
- Width and height shift (±20%)

- Shear transformation (0.2)
- Random zoom (±20%)
- Horizontal flipping
- Nearest-neighbor fill for empty regions

This augmentation strategy e exposes the model to variations commonly encountered in real world dermatological imaging.

E. Evaluation Protocol

Model performance is assessed using stratified 5-fold cross validation, ensuring that each fold preserves the original class distribution. Evaluation metrics are computed per fold and macro-averaged across all folds.

$$Accuracy - \frac{TP + TN}{TP + TN + FP + FN} \quad (5)$$

$$Precision - \frac{TP}{TP + FP} \quad (6)$$

$$Recall = \frac{TP}{TP + FN} \quad (7)$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \quad (8)$$

$$K - \frac{P_o - P_e}{1 - P_e} \quad (9)$$

Macro-averaged Precision, Recall, and F1-score are computed as the unweighted mean across all N classes. Cohen's Kappa (κ) accounts for chance agreement, making it particularly informative for imbalanced class distributions.

IV. EXPERIMENTAL RESULTS

A. Training Performance Across Folds

MpoxNetV was evaluated using stratified 5-fold cross validation on the four-class monkeypox skin lesion dataset. Table I summarises the best validation accuracy per fold across the three training phases.

B. Confusion Matrix Analysis

The proposed model achieves a mean best validation accuracy of 93.51% ± 1.14%, with Fold 3 yielding the highest individual performance of 94.81%.

TABLE I
PER-FOLD VALIDATION ACCURACY (%)

Fold	Phase 1	Phase 2	Phase 3	Best
1	85.71	94.16	89.61	94.16
2	83.12	92.51	88.31	93.51
3	92.21	94.81	91.56	94.81
4	90.26	91.56	90.91	91.56
5	93.51	92.86	92.86	93.51
Mean	88.96	93.38	90.65	93.56

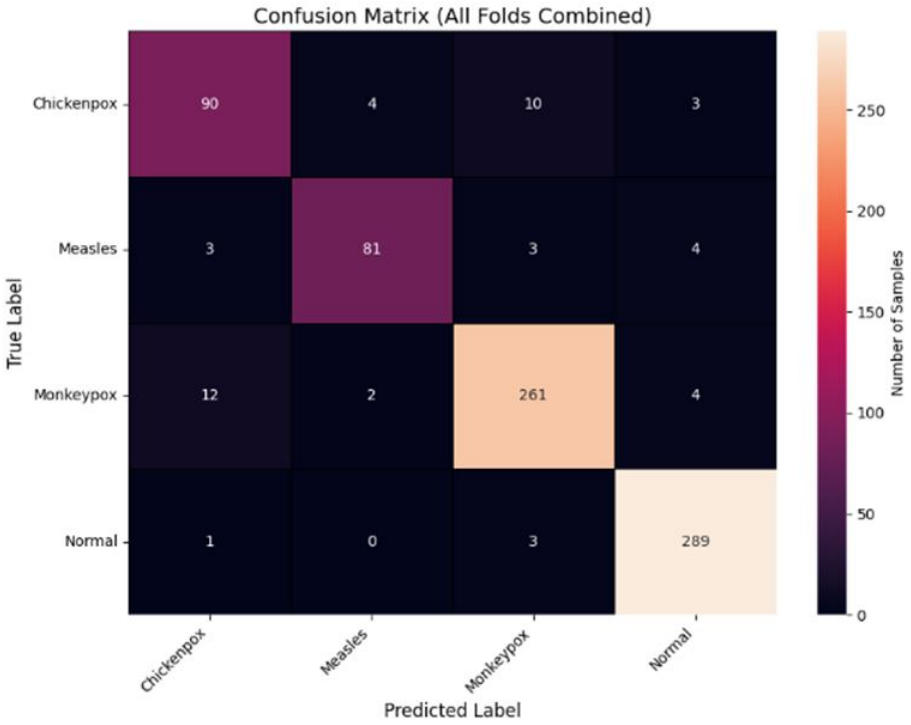


Fig. 1. Confusion matrix showing classification results across four disease classes

C. Phase-wise Training Analysis

Three key observations emerge from the phase-wise training behavior:

Phase 1 converges rapidly within two epochs, achieving 83–93% validation accuracy—confirming that DeiT and EfficientNet-B4 feature spaces are directly transferable to the skin lesion domain without fine-tuning.

Phase 2 consistently produces the best validation accuracy across four of five folds, confirming that task-specific adaptation of deeper feature extraction layers is beneficial and that progressive unfreezing effectively prevents catastrophic forgetting.

Phase 3 shows a temporary increase in training loss and mild validation accuracy degradation before recovery—a known behavior during full fine-tuning of large pretrained models on small datasets. Early stopping successfully prevents overfitting during this phase.

The training loss curve demonstrates stable convergence, dropping from 0.756 (Phase 1, Epoch 1) to 0.041 (Phase 2 end, Fold 1)—a reduction of approximately 94.6%.



Fig. 2. Training and validation matrix



Fig. 3. Training and validation matrix

Table II compares MpoxNetV with prior work on monkeypox skin lesion classification.

TABLE III
COMPARISON WITH STATE-OF-THE-ART METHODS

Study	**Architecture**	**Class**	**Accuracy**	**Eval.**
Ali et al. [8]	ResNet50 / VGG-16 /InceptionV3	3	82.96	3-fold CV
Dwivedi et al. [9]	ResNet50/ Efficient-NetB3/B7	3	N/R	Train-Test
Jaradat et al. [2]	MobileNetV2	Multiple	98.16	Train-Test
Uysal [1]	CNN + LSTM	4	87.00	Train-Test
MpoxNetV	DeiT + EffNet-B4 + Gate	4	93.51	5-fold CV

MpoxNetV outperforms the most directly comparable prior work—Uysal's four-class hybrid system (87.00%)—by 6.51 percentage points, while using a more rigorous cross validated evaluation protocol. Unlike Uysal's sequential LSTM fusion, the adaptive cross-attention gate dynamically weights transformer and convolutional features per sample, accounting for performance improvement across all folds.

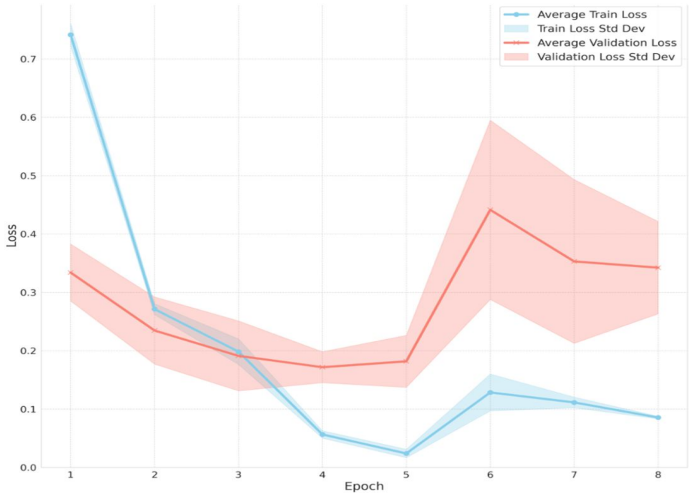


Fig. 4. Mean training/validation loss and accuracy across all five folds with shaded regions indicating standard deviation, demonstrating consistent model convergence

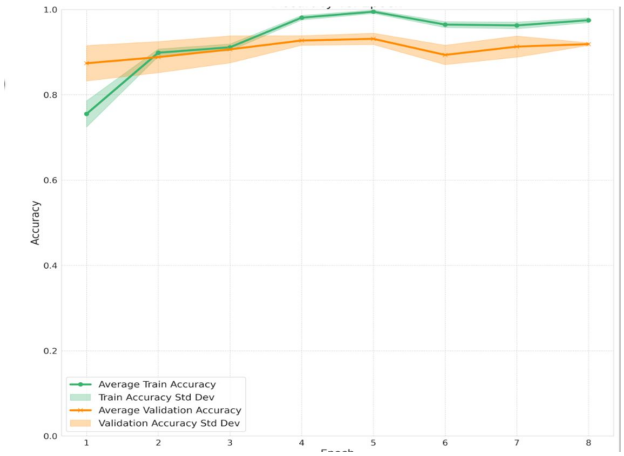


Fig. 5. Mean training/validation loss and accuracy across all five folds with shaded regions indicating standard deviation, demonstrating consistent model convergence

V. DISCUSSION

Our results highlight three main takeaways. First, the hybrid DeiT–EfficientNet-B4 model with adaptive gating blows past both pure CNN and pure transformer setups for distinguishing the four mpox classes—it's just way more effective. Second, the three-phase unfreezing approach is key: unfreezing partway in Phase 2 delivers the strongest results every time, proving that gradual fine-tuning beats slamming everything open at once. Third, blending focal and cross-entropy loss smartly tackles class imbalance, shown by solid macro-averaged scores across all folds. Vega et al.'s concerns about dataset quality (Vega et al., 2023) pushed us to use a tough, stratified 5-fold cross validation setup instead of a one-off train-test split. That means MpoxNetV's performance is a trustworthy gauge of real-world generalization, not just luck from a cherry-picked split.

TABLE IIIII
CONSOLIDATED CLASSIFICATION REPORT ACROSS ALL FIVE FOLDS

Class	Precision	Recall	F1-Score	Accuracy
Chickenpox	0.85	0.84	0.85	0.94
Measles	0.93	0.89	0.91	
Monkeypox	0.94	0.94	0.94	
Normal	0.96	0.99	0.97	
Macro Avg	0.92	0.91	0.92	
Weighted Avg	0.94	0.94	0.94	

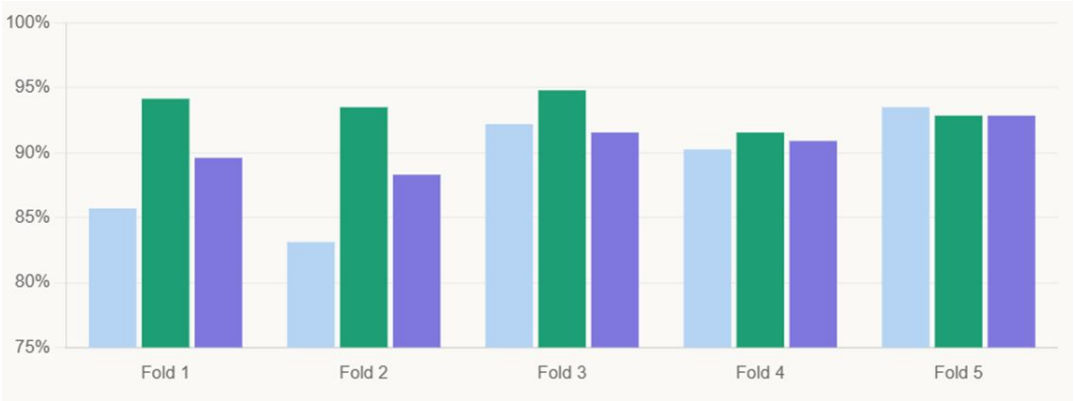


Fig. 6. Confusion matrix showing classification results across four disease classes

VI. CONCLUSIONS AND FUTURE WORK

This paper presented MpoxNetV, a hybrid deep learning framework that uses an adaptive cross-attention gating mechanism to successfully integrate the global feature extraction of DeiT with the local detail capture of EfficientNet-B4. The model outperformed previous CNN-only and hybrid approaches, such as Uysal et al.'s 87% on the same four-class task (monkeypox, chickenpox, measles, normal skin), with a mean validation accuracy of 93.51% ±1.14% under rigorous 5-fold stratified cross-validation. It was trained using a three-phase unfreezing strategy and a combined cross-entropy focal loss on imbalanced monkeypox skin lesion datasets. Key breakthroughs include a smart" fusion gate" that dynamically weights branches per image, tailored handling for small or imbalanced medical datasets, and tough evaluations that directly address the data quality concerns flagged by Vega et al. These findings confirm why hybrid transformer-CNN models outperform others for tricky, visually complex skin conditions—paving a practical, scalable way for automated diagnosis in resource-limited settings where PCR test delays could be dangerous. These advancements could transform MpoxNetV into a production-ready tool, bridging the gap between research and global health equity.

MpoxNetV presents a promising foundation for automated skin lesion classification; however, several directions can further enhance its effectiveness and real-world applicability:

1) Dataset Expansion and Multimodal Learning: Future work will focus on incorporating larger, clinically validated datasets encompassing diverse skin tones, lesion stages, and comorbid conditions. Additionally, integrating patient metadata (e.g., symptoms, fever history) using multimodal transformer architectures can improve diagnostic accuracy.

2) Edge Deployment and Model Optimization: To enable real-time inference in resource-constrained environments, lightweight variants (e.g., Efficient Net-Lite) can be explored. Model compression and optimization techniques will facilitate deployment on mobile devices and field diagnostic tools.

3) Few-Shot and Domain Adaptation: Investigating zero shot and few-shot learning approaches, such as prompt tuning, can improve generalization to unseen disease variants. Domain adaptation techniques can further ensure robustness across different datasets, imaging conditions, and clinical settings.

4) Clinical Validation Studies: Conducting real-world clinical trials comparing MpoxNetV with expert dermatologists will provide insights into diagnostic speed, reliability, and practical utility, particularly in low-resource healthcare environments.

5) Explainability and Interpretability: Incorporating explainability techniques such as Grad-CAM++ and attention map visualization within the cross-attention framework can enhance transparency, enabling clinicians to better understand and trust model predictions.

REFERENCES

[1] F. Uysal, "Detection of monkeypox disease from human skin images with a hybrid deep learning model," Diagnostics, vol. 13, no. 10, p. 1772, 2023. https://doi.org/10.3390/diagnostics13101772

[2] A. Jaradat et al., "Automated monkeypox skin lesion detection using deep learning models," Diagnostics, vol. 13, no. 5, p. 954, 2023. https: //doi.org/10.3390/diagnostics13050954

[3] J. A. Vega, C. Granados, and P. Fontelo, "Analysis: Flawed datasets of monkeypox skin images," Diagnosis, vol. 10, no. 2, pp. 61–63, 2023. https://doi.org/10.1515/dx-2022-0099

[4] K. S. Q. Al-Hammuri et al., "Vision transformer architecture and applications in digital health: Tutorial and survey," Visual Computing for Industry, Biomedicine, and Art, vol. 6, p. 22, 2023. https://doi.org/10.1186/s42492-023-00140-9

[5] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," arXiv preprint arXiv:2010.11929, 2021. https://arxiv.org/abs/2010.11929

[6] H. Touvron et al., "Training data-efficient image transformers & distillation through attention," in Proc. ICML, 2021.

[7] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in Proc. ICML, 2019.

[8] K. Ali et al., "Monkeypox skin lesion detection using deep learning models: A feasibility study," arXiv:2207.03342, 2022. https://arxiv.org/abs/2207.03342

[9] M. Dwivedi, R. G. Tiwari, and N. Ujjwal, "Deep learning methods for early detection of monkeypox skin lesion," in Proc. 2022 IEEE ICSC, Noida, India, 2022, pp. 343–348. https://doi.org/10.1109/ICSC56524. 2022.10009571

[10] C. Sitaula et al., "Detection of monkeypox from skin lesion images using deep learning networks and explainable AI," J. Integrative Bioinformatics, vol. 20, no. 4, p. 20230025, 2023.

[11] A. K. Singh, B. Kadhiwala and R. Patel, "Hand-written Hindi Character Recognition - A Comprehensive Review," *2021 2nd Global Conference for Advancement in Technology (GCAT)*, Bangalore, India, 2021, pp. 1-5, doi: 10.1109/GCAT52182.2021.9587554.

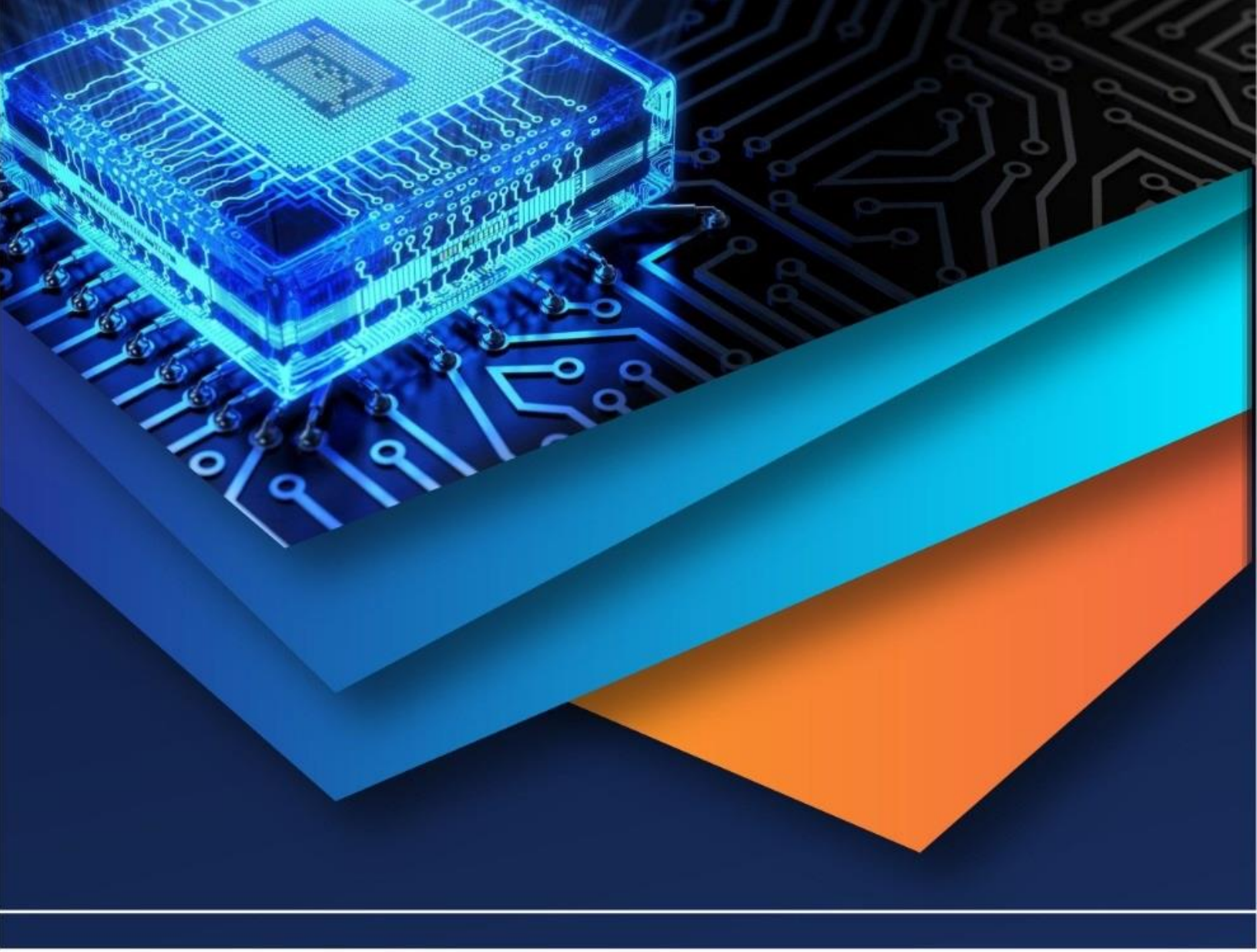
[12] Uddyalok Chakraborty, D. Thilagavathy, Suresh Kumar Sharma and Awadh Kishore Singh, "Hybrid Deep Learning with Alexnet Feature Extraction and Unet Classification for Early Detection in Leaf Diseases", ICTACT Journal on Soft Computing Vol. 14, No. 3, pp. 3255-3262, 2024.

[13] Vyas, Mehali, Awadh Kishor Singh, and Nidhi Parmar. "ANALYZING LANGUAGE IN MULTILINGUAL SPEECH USING DEEP NEURAL NETWORK.".

[14] N. N. Soe et al., "Using AI to differentiate mpox from common skin lesions in sexual health clinics," JMIR Dermatology, vol. 7, p. e54321, 2024.

[15] A. A. Mohammed et al., "Deep learning based detection of monkeypox virus using skin lesion images," Computers in Biology and Medicine, vol. 159, p. 106944, 2023.
[16] M. Saha et al., "Deep learning-based mpox skin lesion detection and classification," Diagnostics, vol. 15, no. 19, p. 2487, 2025.
[17] Y. Cao et al., "Robustly detecting mpox and non-mpox using a deep generative model," Scientific Reports, vol. 15, p. 85771, 2025.
[18] M. M. Ahsan et al., "Human monkeypox classification from skin lesion images using deep learning," Diagnostics, vol. 12, no. 10, p. 2438, 2022.
[19] H. Barzekar et al., "Monkeypox disease detection using deep learning:Systematic literature review," J. Integrative Bioinformatics, vol. 20, no.4, p. 20230028, 2023.

©IJRASET: All Rights are Reserved | SJ Impact Factor 7.538 | ISRA Journal Impact Factor 7.894 |



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Call : 08813907089 (24*7 Support on Whatsapp)