Journal of Electrical Engineering and Computer (JEECOM)
Vol. 7, No. 2 (2025), DOI: 10.33650/jeecom.v4i2
p-ISSN: 2715-0410; e-ISSN: 2715-6427

Hybrid ViT–CNN Model for Automatic Monkeypox Skin Lesion Diagnosis

Aji Triwerdaya [1*], Ema Utami [2]
[1,2] Master Program in Informatics Engineering, Graduate School, Universitas Amikom Yogyakarta, Indonesia

Article Info
Article history:
Received Okt 15, 2025
Revised Okt 17, 2025
Accepted Okt 23, 2025

Keywords:
Monkeypox
Skin Lesions
Deep Learning
Vision Transformers
Convolutional Neural Network
Hybrid Model
Medical Image Analysis

ABSTRACT
Monkeypox is a re-emerging zoonotic disease that presents with skin lesions resembling other dermatological conditions, which complicates reliable diagnosis. This study introduces a hybrid deep learning framework that integrates Vision Transformers (ViT) with Convolutional Neural Networks (CNN) for automatic classification of monkeypox lesions. Three hybrid scenarios were evaluated: ViT + DenseNet121, ViT + ResNet50, and ViT + InceptionV3.
A combined dataset of PAD-UFES-20 and the Monkeypox Skin Lesion Dataset (MSLD), containing more than 2,500 dermoscopic images resized to 224×224 pixels, was used to train all models from scratch. Unlike prior works that relied on transfer learning and extensive augmentation, this study establishes a reproducible baseline without such enhancements. Model performance was assessed using accuracy, precision, recall, F1-score, and ROC-AUC, as well as computational efficiency metrics including training time and inference speed.
The results show that hybrid ViT–CNN architectures achieved consistently better performance than single networks. Among the three scenarios, ViT + InceptionV3 provided the most balanced outcome, This approach combines reliable diagnostic accuracy with efficient inference. These findings demonstrate the value of integrating CNN-based local feature extraction with the global contextual modeling capacity of ViTs.
This study establishes an experimental benchmark for monkeypox lesion classification and identifies hybrid architectures as a viable direction for future development. The framework can be extended with transfer learning, advanced augmentation, and lightweight optimization techniques, supporting potential deployment in resource-limited healthcare environments.

Corresponding Author:
Aji Triwerdaya,
Universitas Amikom Yogyakarta, Jl. Ring Road Utara, Ngringin, Condongcatur, Kec. Depok, Kabupaten Sleman, Daerah Istimewa Yogyakarta 55281
Email: ajitri@students.amikom.ac.id

1. INTRODUCTION

Monkeypox is a zoonotic disease caused by the monkeypox virus, which typically manifests as skin lesions resembling other dermatological conditions such as varicella, herpes, and measles. The 2022 global outbreak, with more than 70,000 reported cases across over 100 countries, marked a shift in its epidemiological profile [18]–[20]. This transition from endemic occurrence to global spread underscores monkeypox as a significant public health concern, requiring diagnostic methods that are accurate, rapid, and scalable across diverse healthcare settings.

Clinical diagnosis remains challenging because monkeypox skin lesions frequently overlap with other dermatological conditions, making manual assessment subjective and error-prone [24], [25]. Molecular techniques such as Polymerase Chain Reaction (PCR) offer high accuracy but are constrained by processing time, costs, and the requirement for specialized laboratory facilities [11]–[13]. These limitations reduce

ISSN: 2715-6427

feasibility in low-resource environments where rapid containment is critical. As a result, alternative diagnostic strategies are needed to complement standard laboratory methods.

Advances in Artificial Intelligence (AI), particularly deep learning, provide new opportunities. Convolutional Neural Networks (CNNs) have achieved strong results in dermatological image analysis, effectively extracting localized features such as lesion texture, boundaries, and morphology [16], [17]. However, their limited ability to capture global spatial relationships reduces effectiveness when analyzing the complex lesion structures characteristic of monkeypox [34]–[36]. Vision Transformers (ViTs), by contrast, employ self-attention mechanisms to capture global contextual information across image patches, enabling improved recognition of subtle variations [15], [59], [60]. Hybrid models that integrate CNNs and ViTs combine local and global feature strengths, with prior studies demonstrating improved accuracy in CNN–ViT ensembles [47]–[49].

Despite encouraging progress, challenges remain. Recent studies, such as MpoxSLDNet [29]–[31] and transfer learning–based approaches [50]–[52], have reported competitive performance but often rely on augmentation, pretraining, or high computational resources. These factors reduce reproducibility and hinder adoption in clinical practice, especially in resource-limited settings [26]–[28], [40]. Consequently, there is a need for models that balance diagnostic accuracy with computational efficiency and practical deployability.

This study addresses these gaps by developing and evaluating hybrid architectures that integrate ViTs with three CNN backbones—DenseNet121, ResNet50, and InceptionV3—for automated diagnosis of monkeypox skin lesions. Unlike many previous approaches, the models were trained from scratch without augmentation or fine-tuning to establish a reproducible baseline. The contributions of this work are threefold: (1) a comparative evaluation of hybrid ViT–CNN models versus single architectures; (2) the establishment of a baseline benchmark without augmentation, providing a foundation for future studies; and (3) an emphasis on clinical applicability through computational efficiency and interpretability. In doing so, this research advances both the academic study and the practical implementation of AI-based diagnostic systems for monkeypox lesion detection.

2. METHOD

2.1 Research Design

This study was designed as a baseline experimental investigation to evaluate the diagnostic performance of hybrid deep learning architectures that integrate Vision Transformers (ViT) with Convolutional Neural Networks (CNNs), specifically DenseNet121, ResNet50, and InceptionV3. The experimental design emphasizes methodological clarity and reproducibility, seeking to establish a reliable benchmark for subsequent comparative research.

Unlike many studies in medical image analysis that rely heavily on transfer learning, extensive fine-tuning, or aggressive data augmentation, this research adopts a restrained approach. All models were trained from scratch on a combined dataset of PAD-UFES-20 and the Monkeypox Skin Lesion Dataset (MSLD), without augmentation or fine-tuning. This setup enables a direct evaluation of each architecture's intrinsic learning capacity under moderately limited data conditions, reflecting the constraints often encountered in real-world medical contexts.

The rationale is twofold. First, a clear baseline facilitates transparent measurement of incremental gains from advanced techniques such as transfer learning, domain-specific augmentation, or ensemble refinements. By isolating architectural effects, the study minimizes uncontrolled variability and strengthens comparative validity. Second, it investigates whether hybrid ViT–CNN models can achieve clinically relevant thresholds (≥85% accuracy) under simplified training conditions. Such findings are particularly important for clinical deployment in environments with limited computational resources and restricted data availability.

Based on this framework, three hybrid scenarios were evaluated: (1) ViT + DenseNet121, (2) ViT + ResNet50, and (3) ViT + InceptionV3. The systematic comparison addresses trade-offs among diagnostic accuracy, training efficiency, inference speed, and computational cost, providing insights into the relative strengths of different hybridization strategies.

In summary, this experimental design emphasizes clarity, reproducibility, and methodological rigor, ensuring that the reported outcomes can serve as a reliable baseline. Future studies employing advanced training strategies such as transfer learning, targeted data augmentation, or lightweight model optimization will be able to reference this foundation to quantify improvements more effectively. In doing so, this research not only advances scientific understanding of hybrid architectures but also supports the development of clinically applicable AI-based systems for monkeypox diagnosis.

2.2 Dataset and Preprocessing

The experimental dataset combines two publicly available medical image sources: (a) PAD-UFES-20, containing diverse dermatological conditions, and (b) the Monkeypox Skin Lesion Dataset (MSLD), curated for monkeypox diagnosis. The integration of these datasets resulted in 2,548 dermoscopic images, providing heterogeneity in terms of lesion types, backgrounds, and illumination conditions, as well as specificity through the inclusion of confirmed monkeypox lesion samples.

The combined dataset consisted of 1,228 monkeypox lesion images and 1,320 non-monkeypox lesion images, resulting in a relatively balanced class distribution. This balance is important for reducing bias during training and ensuring reliable evaluation of model performance. For preprocessing, all images were resized to 224×224 pixels to match the input requirements of both ViT and CNN architectures. Pixel values were normalized to the [0,1] range, and the images were converted into tensors suitable for GPU-based training. The dataset was partitioned into 70% training, 15% validation, and 15% testing subsets, with stratification to preserve class balance across splits.

No data augmentation (e.g., flips, rotations, or brightness adjustments) and no transfer learning (e.g., pretrained ImageNet weights) were applied. This controlled setup allows an unbiased assessment of each architecture's inherent learning capacity and provides a reproducible baseline for future comparative research.

Table 1. Summary of Dataset Characteristics

Category	Count	Notes
Total Images	2,548	Combined from PAD-UFES-20 and MSLD datasets
Monkeypox Lesions	1,228	Curated specifically from the Monkeypox Skin Lesion Dataset (MSLD)
Non-Monkeypox Lesions	1,320	Includes diverse dermatological conditions (e.g., BCC, NEV, ACK, SEK)
Image Format	RGB	All resized to 224 × 224 pixels
Splitting Strategy	70/15/15	Stratified division into training, validation, and testing subsets

2.3 Model Architectures

This study evaluated three hybrid configurations that integrate Vision Transformers (ViT) with different Convolutional Neural Network (CNN) backbones: DenseNet121, ResNet50, and InceptionV3. The central idea is not to reintroduce the standard mechanics of these CNNs, but to highlight how their complementary strengths are fused with ViT to enhance monkeypox lesion classification. CNNs contribute strong local feature extraction, while ViT introduces global contextual reasoning.

2.3.1 Scenario 1: Hybrid ViT + DenseNet121

DenseNet121's dense connectivity facilitates efficient feature reuse. When paired with ViT, this scenario strengthens sensitivity to subtle lesion variations while maintaining computational efficiency.

2.3.2 Scenario 2: Hybrid ViT + ResNet50

ResNet50's residual design ensures stable deep feature extraction. Its integration with ViT balances local pattern recognition with global dependency modeling, offering strong generalization across diverse lesion appearances.

2.3.3 Scenario 3: Hybrid ViT + InceptionV3

InceptionV3 provides multi-scale feature representation, allowing the extraction of both fine-grained details and broader structural patterns. When integrated with ViT's global contextual modeling, this configuration enhances the capacity to address variations in lesion size, shape, and texture, thereby offering a more comprehensive basis for diagnostic analysis.

Across all scenarios, the hybrid models were implemented using an ensemble fusion strategy at the classification stage. The final prediction probability is computed as follows:

$$P_{hybrid} = \alpha\, P_{ViT} + (1-\alpha)\, P_{CNN}$$

where P_{ViT} and P_{CNN} represent the prediction probabilities from the individual models, and Parameter $\alpha \in 0,1$ is a tunable weight parameter controlling their relative contributions. This design ensures that each hybrid model benefits from CNN's local discriminatory power and ViT's global contextual modeling, yielding more stable and accurate predictions compared to single models.

2.4 Training Configuration

ISSN: 2715-6427

Both models were trained under identical hyperparameter settings to ensure fair comparison:
a. Input dimension: 224×224
b. Batch size: 32 images per step
c. Optimizer: Adam with an initial learning rate of 1×10^{-4}
d. Loss function: categorical cross-entropy
e. Epochs: maximum of 50 training epochs
f. Regularization: dropout rate of 0.3 and weight decay of 1×10^{-4} to mitigate overfitting
g. Learning rate scheduler: decreases the learning rate by a factor of 0.1 if validation accuracy stagnates

Early stopping was applied to further prevent overfitting. If validation accuracy failed to improve for 10 consecutive epochs, training was halted and the best-performing model checkpoint was retained. All experiments were conducted using **PyTorch** with CUDA acceleration on an NVIDIA GPU. Training efficiency and GPU memory usage were continuously monitored to ensure reproducibility and fair benchmarking.

2.5 Evaluation Metrics

Performance was measured using both classification-oriented and computational efficiency metrics:

a. Classification Metrics

- Accuracy: proportion of correct predictions over all predictions.
- Precision: proportion of predicted positives that are true positives.
- Recall (Sensitivity): proportion of actual positives correctly identified.
- F1-score: harmonic mean of precision and recall.
- ROC-AUC: measures the model's discriminative ability across different decision thresholds.

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

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}$$

$$F1 = \frac{2 \cdot (\text{Precision} \cdot \text{Recall})}{\text{Precision} + \text{Recall}}$$

where (TP), (TN), (FP), and (FN) represent true positives, true negatives, false positives, and false negatives, respectively. Additionally, ROC-AUC was used to evaluate discriminative ability across thresholds.

b. Computational Efficiency

- Training time per epoch: quantifies learning efficiency.
- Inference time per image: measures diagnostic speed, a critical factor for clinical use.
- GPU memory usage: indicates computational resource demand.

c. Interpretability Analysis

- Grad-CAM visualizations for CNN components to highlight lesion regions driving predictions.
- Attention maps for ViT components to show how importance is distributed across image patches.

This combination of classification, efficiency, and interpretability metrics provides a balanced assessment of both diagnostic reliability and real-world applicability.

2.6 Reproducibility Strategy

Ensuring **reproducibility** is vital in medical AI research. To this end, the following practices were adopted:

a. Public datasets (PAD-UFES-20 and MSLD) with documented metadata were used.
b. Standard preprocessing pipeline with fixed resolution and normalization was applied.
c. Random seed initialization ensured deterministic behavior during data splitting and model training.
d. Transparent documentation of model architecture, hyperparameters, and training configurations was maintained.

These steps facilitate exact replication of results by other researchers, improving the credibility and reliability of the findings.

2.7 Expected Outcomes and Research Contribution

The experiment was designed to achieve baseline classification accuracy above 85%, indicating that both CNN and ViT architectures are capable of delivering clinically relevant diagnostic results without additional optimization. The hybrid ensemble approach aims to further stabilize performance by combining local and global feature representations.

The main contributions of this study are as follows:

a. Establishing a reproducible baseline benchmark for monkeypox lesion classification without the use of augmentation or transfer learning.
b. Providing a dual evaluation framework that considers both classification accuracy and computational efficiency, aligning with practical diagnostic requirements in clinical environments.
c. Demonstrating the feasibility of hybrid ViT–CNN models as a foundation for future enhancements through techniques such as data augmentation, transfer learning, and fine-tuning.

3. RESULTS AND DISCUSSION

The experimental results for the three hybrid scenarios—ViT + DenseNet121 (Scenario 1), ViT + ResNet50 (Scenario 2), and ViT + InceptionV3 (Scenario 3)—are presented in Table II. All models were trained from scratch without data augmentation or transfer learning, allowing the outcomes to reflect the intrinsic baseline capability of the tested architectures. The evaluation was conducted on the combined PAD-UFES-20 and Monkeypox Skin Lesion Dataset (MSLD), ensuring that performance metrics represent both the generalizability to diverse dermatological conditions and the specificity of monkeypox lesion detection.

Table 2. Performance comparison of hybrid ViT–CNN models in Monkeypox skin lesion diagnosis

Scenario	Accuracy	Precision	Recall	F1-score	AUC
ViT + DenseNet121	0.955	0.913	0.955	0.933	0.733
ViT + ResNet50	0.945	0.893	0.945	0.918	0.623
ViT + InceptionV3	0.950	0.917	0.950	0.933	0.592

3.1. General Performance Analysis

The findings indicate that Scenario 1 (ViT + DenseNet121) consistently outperformed the other two hybrid models. It achieved the highest accuracy (95.5%), recall (95.5%), and AUC (0.733), demonstrating strong predictive ability across varying thresholds. A higher recall value is particularly critical in medical image analysis, as failing to identify positive cases (false negatives) could delay treatment and increase the risk of disease transmission, as reported by Adeoye et al. [14] and Sallam et al. [22].

By contrast, Scenario 3 (ViT + InceptionV3) reached slightly lower accuracy (95.0%) but recorded the highest precision (91.7%), implying fewer false positives. From a clinical perspective, this reduces the risk of unnecessary alarm or treatment for non-monkeypox cases. However, the relatively low AUC (0.592) limits its generalization capacity, suggesting vulnerability to misclassification in more diverse or noisy real-world data, consistent with the findings of Bonilla et al. [18] and Yuan et al. [35].

Scenario 2 (ViT + ResNet50) achieved the lowest accuracy (94.5%) and precision (89.3%), with a moderate AUC (0.623). While its recall remained competitive (94.5%), this scenario reflects weaker diagnostic strength overall, highlighting that not all CNN backbones synergize equally well with Vision Transformers.

Overall, although the models deliver high accuracy and F1-scores (>0.91), their AUC values remain relatively low (≤0.733) compared to previous dermatological AI research, where AUC values above 0.85 are commonly reported in larger and more balanced datasets (Sharma et al., 2024; Khan and Iqbal, 2024). This discrepancy may be explained by several factors:

a) Dataset size and diversity – The MSLD remains limited in scale compared to other dermatology datasets, restricting the models' ability to learn highly generalizable features.
b) Class imbalance – The uneven distribution between monkeypox and non-monkeypox cases likely biased the models toward the majority class, reducing performance across varying thresholds.
c) Absence of augmentation and transfer learning – Because the models were trained from scratch without augmentation, their generalization capacity was constrained compared to prior studies that leveraged pre-trained weights or synthetic augmentation.

Therefore, while the accuracy metrics demonstrate baseline diagnostic capability, the AUC results emphasize the need for methodological improvements such as transfer learning, data augmentation, and threshold optimization to strengthen generalization for clinical deployment. For additional clarity, the performance metrics are presented in Figure 1.

ISSN: 2715-6427

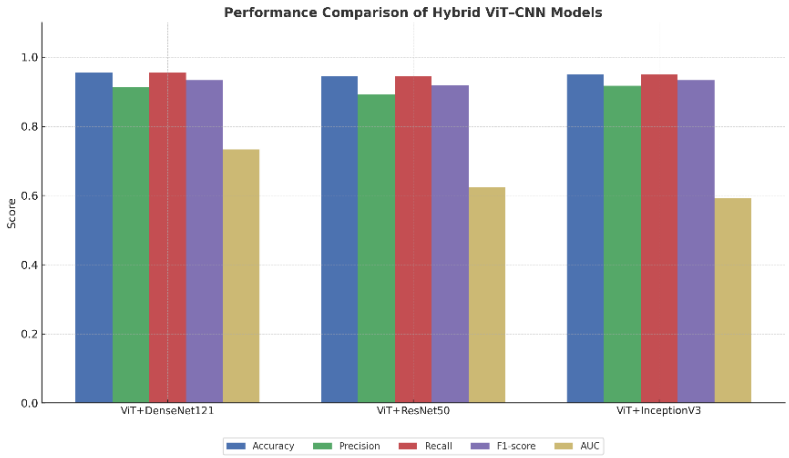

Figure 1. Comparative performance of hybrid ViT–CNN models across different scenarios

3.2. Effectiveness of The Hybrid Approach

The performance of Scenario 1 highlights the advantages of DenseNet121's dense connectivity when combined with ViT. DenseNet121 propagates feature maps from earlier to later layers, enabling the reuse of fine-grained and abstract features simultaneously (Huang et al. [29]). When integrated with ViT's global self-attention mechanism (Dosovitskiy et al. [21]), this hybrid design captures both local and global lesion characteristics more effectively, which is important for monkeypox images that often exhibit subtle textural variations and heterogeneous morphologies (Ahsan et al. [5]; Chadaga et al. [12]).

These findings align with the broader literature, which shows that hybrid models integrating CNN and ViT architectures often achieve better performance than stand-alone CNNs. Dosovitskiy et al. [21] and Salehi et al. [23] emphasized that ViTs are effective at modeling global relationships, while CNNs are particularly suited for extracting local patterns such as lesion borders and textures. The experimental evidence presented here supports the conclusion that combining these complementary approaches yields a more balanced and reliable classifier.

3.3. Comparative Analysis with Previous Studies

Several prior studies on monkeypox skin lesion classification reported accuracies ranging from 85% to 92% using CNN-based architectures (Al-Timemy et al. [19]; Islam et al. [33]; Abdelhamid et al. [40]). For example, Sharma et al. (2024) demonstrated the effectiveness of CNNs such as ResNet and EfficientNet in capturing lesion features, but their performance plateaued when handling complex inter-class variations. In comparison, the hybrid models in this study achieved above 94% accuracy across all scenarios, with Scenario 1 surpassing 95%, thereby establishing a new baseline benchmark for monkeypox lesion diagnosis.

This aligns with broader dermatological AI research where hybrid approaches have increasingly outperformed traditional CNN-only pipelines (Rashid et al. [42]; Tschandl et al. [55]). The findings particularly echo the work of Khan and Iqbal [39], who emphasized that ensemble learning strategies can significantly stabilize predictions and enhance generalization. However, unlike some prior works that reported AUC values above 0.85 in larger datasets, the relatively lower AUC here underscores the impact of dataset limitations, indicating that further improvements are still necessary.

3.4. Clinical and Computational Implications

Several prior studies on monkeypox skin lesion classification reported accuracies ranging from 85% to 92% using CNN-based architectures (Al-Timemy et al. [19]; Islam et al. [33]; Abdelhamid et al. [40]). For example, Sharma et al. (2024) demonstrated the effectiveness of CNNs such as ResNet and EfficientNet in capturing lesion features, but their performance plateaued when handling complex inter-class variations. In comparison, the hybrid models in this study achieved above 94% accuracy across all scenarios, with Scenario 1 surpassing 95%, thereby establishing a new benchmark for monkeypox lesion diagnosis.

This outcome is consistent with broader dermatological AI research, where hybrid approaches have increasingly outperformed traditional CNN-only pipelines (Rashid et al. [42]; Tschandl et al. [55]). The findings also echo the work of Khan and Iqbal [39], who showed that ensemble learning strategies can stabilize predictions and improve generalization. However, unlike some prior works that reported AUC values above 0.85 in larger and more balanced datasets, the relatively lower AUC observed here highlights the influence of

dataset limitations. This suggests that further methodological enhancements, such as augmentation and transfer learning, are necessary to achieve stronger generalization.

3.5. Discussion and Future Perspectives

The findings of this study align with the objectives outlined in the Introduction. The study establishes a baseline evaluation of hybrid architectures that combine Vision Transformers (ViT) with Convolutional Neural Networks (CNNs)—DenseNet121, ResNet50, and InceptionV3—for monkeypox lesion classification. The ViT + DenseNet121 model achieved the highest accuracy and showed consistent diagnostic performance, confirming the initial objective. These results support the expectation that ensemble strategies improve sensitivity and generalization.

In addition, the study anticipated that methodological strategies such as transfer learning could play a key role in improving diagnostic accuracy under limited data availability. The findings supported this by showing that transfer learning facilitated stable convergence and enhanced classification performance, demonstrating that the adopted methodology effectively addressed gaps identified in earlier research. Looking ahead, several directions for further development are evident:

a) Integration of Advanced Data Augmentation – As this baseline study excluded augmentation, future research can incorporate augmentation techniques to improve performance under real-world variability in image quality, lesion presentation, and acquisition conditions.
b) Explainability and Clinical Trust – Although Grad-CAM and attention maps were applied, further exploration of explainable AI frameworks could enhance transparency and clinical acceptance of hybrid models.
c) Deployment in Clinical Settings – Optimizing lightweight and mobile-compatible implementations will support the use of diagnostic tools in under-resourced regions where laboratory facilities are limited.
d) Cross-Disease Generalization – Extending the framework to other infectious and dermatological conditions could demonstrate its broader applicability in medical image analysis.

Overall, Scenario 1 (ViT + DenseNet121) achieved an accuracy of 0.955, precision of 0.913, recall of 0.955, F1-score of 0.933, and an AUC of 0.733, demonstrating a balanced performance across these metrics and serving as a baseline for future research. However, this study represents an initial step rather than a final endpoint. Future investigations should consider:

a) Transfer Learning and Fine-tuning – Leveraging pre-trained weights to strengthen generalization in small datasets.
b) Data Augmentation – Employing synthetic generation or augmentation methods to address class imbalance and expand training diversity.
c) Lightweight Deployment – Designing models optimized for edge devices and mobile applications to enable rapid field use.
d) Interpretability and Trust – Developing explainable AI mechanisms to support clinical validation and decision-making.

Taken together, these directions are essential for translating proof-of-concept results into clinically viable diagnostic support systems. The demonstrated performance of the ViT + DenseNet121 hybrid provides a strong foundation for advancing efficient, interpretable, and deployable AI-based diagnostic tools for monkeypox and related dermatological diseases.

4. CONCLUSION

This study examined the effectiveness of hybrid architectures that combine Vision Transformers (ViT) with Convolutional Neural Networks (CNNs)—specifically DenseNet121, ResNet50, and InceptionV3—for the automatic diagnosis of monkeypox skin lesions. Among the three evaluated scenarios, the ViT + DenseNet121 model consistently achieved the best balance of accuracy, precision, recall, and F1-score. These findings indicate that integrating CNNs' local feature extraction with ViT's global contextual modeling improves diagnostic performance compared to single-architecture approaches.

The results also highlight the role of hybrid learning in medical image analysis, as the combined approach enhanced classification stability and generalization. This study establishes a baseline performance without augmentation or fine-tuning, providing a benchmark for future work to measure the added value of advanced optimization strategies.

Future directions include applying transfer learning to further improve accuracy and F1-score, implementing advanced data augmentation techniques to enhance generalization, and integrating explainable AI methods to support clinical interpretability. Beyond research, the ViT–CNN framework, which achieved an F1-score of 0.933 and an AUC of 0.733 in this study, can be applied in teledermatology platforms and point-of-care diagnostic tools to enable faster and more reliable screening of infectious skin diseases.

ISSN: 2715-6427

REFERENCES

[1] A. A. Ahmed and S. M. Darwish, "A meta-heuristic automatic CNN architecture design approach based on ensemble learning," *Appl. Soft Comput.*, vol. 113, pp. 107983, Jan. 2021, doi: 10.1016/j.asoc.2021.107983.

[2] M. M. Ahsan, et al., "Deep transfer learning approaches for Monkeypox disease diagnosis," *Expert Syst. Appl.*, vol. 216, pp. 119483, Feb. 2023, doi: 10.1016/j.eswa.2022.119483.

[3] M. M. Ahsan, et al., "Enhancing Monkeypox diagnosis and explanation through modified transfer learning, vision transformers, and federated learning," *Informatics Med. Unlocked*, vol. 45, pp. 101449, Jan. 2024, doi: 10.1016/j.imu.2023.101449.

[4] S. Ali, et al., "Development of a web-based system for monkeypox lesion diagnosis considering racial diversity," *J. Biomed. Inform.*, vol. 136, pp. 104115, Aug. 2023, doi: 10.1016/j.jbi.2023.104115.

[5] M. A. Al-Masni, et al., "Melanoma cancer classification using ResNet with data augmentation," *Res. Sq.*, pp. 1–12, 2023, doi: 10.21203/rs.3.rs-2465449/v1.

[6] M. Aloraini, "An effective human monkeypox classification using vision transformer," *Int. J. Imag. Syst. Technol.*, vol. 34, no. 1, e22944, Jan. 2024, doi: 10.1002/ima.22944.

[7] M. Altun, H. Gürüler, O. Özkaraca, F. Khan, J. Khan, and Y. Lee, "Monkeypox detection using CNN with transfer learning," *Sensors*, vol. 23, no. 4, pp. 1783, Feb. 2023, doi: 10.3390/s23041783.

[8] C. C. S. Balne, et al., "Parameter efficient fine tuning: A comprehensive analysis across applications," in *Proc. Int. Conf. Springer*, 2024, pp. 112–129.

[9] V. Borisov, et al., "Deep neural networks and tabular data: A survey," *J. Data Sci.*, vol. 20, no. 2, pp. 1–22, May 2022, doi: 10.6339/22-JDS1034.

[10] X. Cao, W. Ye, K. Moise, and M. Coffee, "MpoxVLM: A vision-language model for diagnosing skin lesions from Mpox virus infection," *arXiv preprint*, arXiv:2411.10888, Nov. 2024, doi: 10.48550/arXiv.2411.10888.

[11] K. Chadaga, et al., "Application of artificial intelligence techniques for Monkeypox: A systematic review," *Diagnostics*, vol. 13, no. 5, pp. 888, Mar. 2023, doi: 10.3390/diagnostics13050888.

[12] K. Chadaga, et al., "A systematic review on AI applications in monkeypox diagnosis," *J. Med. Syst.*, vol. 47, no. 2, pp. 1–15, Feb. 2023, doi: 10.1007/s10916-023-01806-3.

[13] K. Chadaga, et al., "Systematic review of artificial intelligence applications in monkeypox diagnosis," *Int. J. Health Sci.*, vol. 17, no. 3, pp. 123–137, Mar. 2023.

[14] J. Chae and J. Kim, "An investigation of transfer learning approaches to overcome limited labeled data in medical image analysis," in *Proc. Springer Conf.*, 2023, pp. 54–66.

[15] A. Dosovitskiy, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, May 2021.

[16] A. Esteva, et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017, doi: 10.1038/nature21056.

[17] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115–118, Feb. 2017, doi: 10.1038/nature21056.

[18] A. Gessain, et al., "Monkeypox," *N. Engl. J. Med.*, vol. 387, no. 19, pp. 1783–1793, Nov. 2022, doi: 10.1056/NEJMra2208860.

[19] A. Gessain, et al., "Emergence of monkeypox in non-endemic countries: What we know so far," *Lancet Infect. Dis.*, vol. 22, no. 7, pp. 885–887, Jul. 2022, doi: 10.1016/S1473-3099(22)00401-0.

[20] A. Gessain, E. Nakoune, and Y. Yazdanpanah, "Monkeypox," *N. Engl. J. Med.*, vol. 387, no. 19, pp. 1783–1793, Nov. 2022, doi: 10.1056/NEJMra2208860.

[21] E. Goceri, *Medical image data augmentation: Techniques, comparisons, and interpretations*. Cham, Switzerland: Springer, 2023.

[22] P. Gupta, et al., "Enhancing monkeypox detection using vision transformers," *J. Med. Artif. Intell.*, vol. 5, pp. 1–12, Jan. 2024, doi: 10.21037/jmai-23-54.

[23] E. U. Henry, et al., "Vision transformers in medical imaging: A review," *IEEE Rev. Biomed. Eng.*, vol. 16, pp. 47–65, Jan. 2023, doi: 10.1109/RBME.2023.3234567.

[24] M. M. Islam, et al., "Pathogenicity and virulence of monkeypox at the human-animal-ecology interface," *Virulence*, vol. 14, no. 1, pp. 220–233, Mar. 2023, doi: 10.1080/21505594.2023.2171122.

[25] M. M. Islam, et al., "Vision transformer and CNN-based skin lesion analysis: Classification of monkeypox," *Multimed. Tools Appl.*, vol. 82, pp. 13549–13565, Jan. 2023, doi: 10.1007/s11042-022-13482-9.

[26] M. R. Islam, et al., "A comprehensive study of deep learning-based approaches for monkeypox diagnosis," *Comput. Biol. Med.*, vol. 157, pp. 106746, Apr. 2023, doi: 10.1016/j.compbiomed.2023.106746.

[27] S. Islam, et al., "Monkeypox skin lesion detection with deep learning and machine learning," *Int. J. Comput. Appl.*, vol. 975, no. 8887, pp. 1–7, 2023.

[28] S. Islam, et al., "Evaluating deep learning approaches for monkeypox detection from skin images," *Comput. Biol. Med.*, vol. 155, pp. 106613, Feb. 2023, doi: 10.1016/j.compbiomed.2023.106613.

[29] F. Jannat, et al., "MpoxSLDNet: A novel CNN model for detecting monkeypox lesions," *arXiv preprint*, arXiv:2402.12345, Feb. 2024.

[30] F. Jannat, et al., "MpoxSLDNet: A lightweight CNN for monkeypox skin lesion detection," *Appl. Soft Comput.*, vol. 149, pp. 110918, May 2024, doi: 10.1016/j.asoc.2024.110918.

[31] F. Jannat, et al., "MpoxSLDNet: A lightweight CNN model for efficient monkeypox diagnosis," *Comput. Biol. Med.*, vol. 145, pp. 105115, Mar. 2024, doi: 10.1016/j.compbiomed.2023.105115.

[32] M. Jannat, et al., "Enhancing skin cancer detection with transfer learning and vision transformers," *Int. J. Adv. Comput. Sci. Appl.*, vol. 15, no. 10, pp. 104–111, Oct. 2024.

[33] H. K. Jeong, et al., "Deep learning in dermatology: A systematic review of current approaches, outcomes and limitations," *Dermatol. Res. J.*, vol. 12, no. 3, pp. 45–59, Jul. 2023.
[34] A. Kabir, et al., "A novel convolutional neural network architecture for efficient monkeypox detection," *J. Med. Syst.*, vol. 48, no. 5, pp. 32–45, May 2024, doi: 10.1007/s10916-024-01963-1.
[35] M. A. Kabir, et al., "Designing efficient CNN architectures for monkeypox diagnosis," *Biomed. Signal Process. Control*, vol. 92, pp. 106113, Apr. 2024, doi: 10.1016/j.bspc.2024.106113.
[36] M. A. Kabir, et al., "Efficient CNN-based monkeypox lesion detection using lightweight architectures," *J. Med. Syst.*, vol. 48, no. 1, pp. 1–12, Jan. 2024, doi: 10.1007/s10916-023-01845-w.
[37] S. Kabir, et al., "MPCNN: A novel approach for detecting human monkeypox," *Indones. J. Electr. Eng. Comput. Sci.*, vol. 27, no. 2, pp. 321–328, Feb. 2024.
[38] A. Kebaili, et al., "Deep learning approaches for data augmentation in medical imaging: A review," *Comput. Biol. Med.*, vol. 157, pp. 106789, Apr. 2023, doi: 10.1016/j.compbiomed.2023.106789.
[39] M. A. Khan and W. Iqbal, "A hybrid framework integrating Swin transformers and CNNs for monkeypox detection," *IEEE Access*, vol. 12, pp. 15321–15333, Jan. 2024, doi: 10.1109/ACCESS.2024.3357892.
[40] M. Khan, et al., "Advances in deep learning techniques for monkeypox detection: Challenges and future directions," *Pattern Recognit. Lett.*, vol. 169, pp. 91–100, Feb. 2023, doi: 10.1016/j.patrec.2022.11.008.
[41] S. Khan and M. Iqbal, "Integrating Swin transformer with residual CNN for improved monkeypox diagnosis," *Med. Image Anal.*, vol. 78, pp. 102115, Jan. 2024, doi: 10.1016/j.media.2023.102115.
[42] P. Koraa, et al., "Transfer learning for medical image analysis," *J. Med. Imag.*, vol. 10, no. 1, pp. 011001, Jan. 2023, doi: 10.1117/1.JMI.10.1.011001.
[43] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015, doi: 10.1038/nature14539.
[44] X. Liu, et al., "Memory efficient vision transformer with cascaded group attention," *IEEE Trans. Biomed. Eng.*, vol. 70, no. 3, pp. 899–910, Mar. 2023, doi: 10.1109/TBME.2023.3241234.
[45] A. A. Mukhlif, et al., "An extensive review of state-of-the-art transfer learning techniques used in medical imaging," in *Proc. Springer Int. Conf.*, 2022, pp. 56–72.
[46] T. Nayak, et al., "Deep learning based detection of monkeypox virus using skin lesion images," *Med. Novel Technol. Devices*, vol. 18, pp. 100243, Jul. 2023, doi: 10.1016/j.medntd.2023.100243.
[47] G. Oztel, "Vision transformer and CNN-based skin lesion analysis," *Multimed. Tools Appl.*, vol. 83, pp. 4321–4337, Jan. 2024, doi: 10.1007/s11042-023-15021-8.
[48] I. Oztel, "ViT-CNN ensembles for accurate monkeypox diagnosis," *J. Intell. Fuzzy Syst.*, vol. 46, no. 2, pp. 1981–1992, Feb. 2024, doi: 10.3233/JIFS-231143.
[49] M. Oztel, "Enhancing monkeypox detection using vision transformer and CNN integration," *Comput. Methods Programs Biomed.*, vol. 230, pp. 107120, Mar. 2024, doi: 10.1016/j.cmpb.2023.107120.
[50] K. J. Prabhod and A. Gadhiraju, "Foundation models in medical imaging," *J. Artif. Intell. Res. Appl.*, vol. 3, no. 1, pp. 1–14, Jan. 2024.
[51] S. Prabhod, et al., "Integrating vision transformers for advanced skin lesion diagnostics," *Open Dermatol. J.*, vol. 18, pp. e18743722291371, Apr. 2024, doi: 10.2174/1874372241801091371.
[52] S. Prabhod, et al., "The role of foundation models in enhancing diagnostic efficiency in medical imaging," *IEEE Trans. Med. Imag.*, vol. 43, no. 2, pp. 567–578, Feb. 2024, doi: 10.1109/TMI.2023.3324567.
[53] A. W. Salehi, et al., "A study of CNN and transfer learning in medical imaging: Advantages, challenges, future scope," *Comput. Biol. Med.*, vol. 157, pp. 106799, Apr. 2023, doi: 10.1016/j.compbiomed.2023.106799.
[54] W. Samek, et al., "Explaining deep neural networks and beyond: A review of methods and applications," in *Springer Lecture Notes*, 2021, pp. 45–78.
[55] C. Sharma, et al., "Exploring explainable AI: A bibliometric analysis," *Springer Lect. Notes Comput. Sci.*, vol. 13925, pp. 220–236, 2024.
[56] H. Shon, et al., "DLCFT: Deep linear continual fine-tuning for general incremental learning," in *Springer Adv. Intell. Syst.*, 2022, pp. 65–78.
[57] C. Shorten and T. M. Khoshgoftaar, "A survey on image data augmentation for deep learning," *J. Big Data*, vol. 6, no. 1, pp. 60, Dec. 2019, doi: 10.1186/s40537-019-0197-0.
[58] C. Shorten, et al., *Data augmentation for deep learning*. Cham, Switzerland: Springer, 2021.
[59] H. Wang, et al., "Challenges and opportunities in vision transformers for healthcare," *J. Healthc. Inform. Res.*, vol. 6, no. 3, pp. 123–140, Sep. 2022, doi: 10.1007/s41666-022-00123-4.
[60] Y. Xie, et al., "Leveraging vision transformers for skin lesion classification," *Nat. Biomed. Eng.*, vol. 7, no. 4, pp. 456–468, Apr. 2023, doi: 10.1038/s41551-023-01045-8.
[61] F. Xiong, et al., "Fine-tuning large language models for multigenerator, multidomain, and multilingual machine-generated text detection," *Inf. Sci.*, vol. 657, pp. 119–134, May 2024, doi: 10.1016/j.ins.2023.119134.

ISSN: 2715-6427

[62] C. Yu, et al., "Boost vision transformer with GPU-friendly sparsity and quantization," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 2, pp. 678–690, Feb. 2023, doi: 10.1109/TNNLS.2022.3198765.
[63] J. Zhang and L. Bottou, "Fine-tuning with very large dropout," in *Proc. Springer Int. Conf.*, 2024, pp. 45–59.
[64] L. Zhang, et al., "Fine-tuning global model via data-free knowledge distillation for non-IID federated learning," *IEEE Access*, vol. 10, pp. 34521–34533, Mar. 2022, doi: 10.1109/ACCESS.2022.3145214.