Benchmarking DenseNet121, ViT-Base, and Swin-Tiny for Multiclass Mpox Skin Lesion Classification

Mukhriddin Arabboev [1], Shohruh Begmatov [1], Sukhrob Bobojanov [2], Nor Erne Nazira Bazin [3] and Akhram Nishanov [1]

[1] *Tashkent University of Information Technologies named after Muhammad al-Khwarizmi, Amir Temur 108, 100084 Tashkent, Uzbekistan*
[2] *Urgench Ranch University of Technology, Khanka 26, 220100 Urgench, Uzbekistan*
[3] *Universiti Teknologi Malaysia, Skudai, 81310 Johor Bahru, Malaysia*

Abstract

Mpox-related skin lesion recognition is challenging due to strong visual similarity among infectious skin diseases, limited labelled data, and class imbalance. This paper presents a comparative benchmark of three deep learning architectures on the official five-fold protocol of the Mpox Skin Lesion Dataset version 2.0 (MSLD v2.0): DenseNet121, ViT-Base, and Swin-Tiny. The proposed evaluation pipeline combines model-specific ImageNet preprocessing, imbalance-aware learning, validation macro-F1 checkpoint selection, and test-time augmentation. Experiments were conducted on the official Train/Valid/Test splits provided by MSLD v2.0, which contains six classes: Chickenpox, Cowpox, HFMD, Healthy, Measles, and Monkeypox. The results show a clear advantage of transformer-based models over the CNN baseline. Swin-Tiny achieved the best overall performance with 81.43 ± 4.19% accuracy, 79.70 ± 4.50% macro precision, 81.85 ± 4.24% macro recall, 80.03 ± 4.54% macro-F1, and 96.73 ± 1.52% macro ROC-AUC. These findings indicate that hierarchical vision transformers are particularly effective for balanced multiclass skin lesion classification under class imbalance.

Keywords

Mpox, skin lesion classification, Swin Transformer, Vision Transformer, DenseNet121, class imbalance, medical image analysis, deep learning

1. Introduction

Mpox is a zoonotic viral disease caused by an orthopoxvirus. It has remained an important public health concern because of its potential for human-to-human transmission and its visually confusing cutaneous manifestations [1, 2]. The 2022 multi-country outbreak renewed interest in rapid and accessible screening tools, especially in settings where laboratory confirmation or specialist dermatological assessment is not immediately available [2]. In image-based triage scenarios, diagnosis is difficult because mpox lesions may resemble other skin conditions such as chickenpox, cowpox, measles, or hand-foot-mouth disease, and because available datasets are often limited and imbalanced [3].

Deep learning has shown strong promise in dermatological image analysis, including skin cancer recognition and lesion-based decision support [4]-[8]. CNNs such as DenseNet remain strong feature extractors due to efficient gradient flow and feature reuse [9]. More recently, transformer-based models have demonstrated strong performance in computer vision and medical imaging owing to their ability to model long-range dependencies and broader contextual relationships [10]-[13]. In particular, Vision Transformer (ViT) provides global patch-level attention, whereas Swin Transformer introduces hierarchical shifted-window attention that improves locality, multi-scale representation learning, and scalability [10, 11].

IVUS 2026: Information Society and University Studies 2026, May 12, Kaunas, Lithuania
mukhriddin.arabboev@tuit.uz (M. Arabboev); sh.begmatov@tuit.uz (S. Begmatov); w.suxrob.w@gmail.com (S. Bobojanov); erne@utm.my (N.E.N. Bazin); a.nishanov@tuit.uz (A. Nishanov)
0000-0001-5733-5889 (M. Arabboev); 0000-0002-2441-916X (S. Begmatov); 0009-0001-4860-6971 (S. Bobojonov); 0000-0001-5080-8209 (N.E.N. Bazin); 0000-0002-5652-8977 (A. Nishanov)
Copyright © 2026 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-06

CEUR-WS.org/Vol-4317/paper3.pdf

For mpox-related lesion analysis, recent literature has shifted from simple binary screening toward more clinically relevant multiclass benchmarks. A notable example is the Mpox Skin Lesion Dataset version 2.0 (MSLD v2.0), which supports six-class differential diagnosis and offers an official fold-based Train/Valid/Test structure for more reproducible evaluation [3]. Recent studies on this benchmark suggest that transformer-based architectures may outperform conventional CNNs, but fair protocol-aware comparisons remain limited [3], [14].

Motivated by this gap, this paper presents a controlled benchmark of DenseNet121, ViT-Base, and Swin-Tiny on the official five-fold MSLD v2.0 protocol using an imbalance-aware training pipeline. The main contributions are as follows.

- A reproducible benchmark of three representative architectures on the official MSLD v2.0 fold protocol.
- An imbalance-aware training strategy combining class weighting, focal loss for transformer models, weighted cross-entropy for the CNN baseline, and weighted sampling.
- A multi-metric evaluation using accuracy, macro precision, macro recall, macro-F1, and macro ROC-AUC.
- A structured comparison with existing studies, including both all-dataset comparison and same-dataset comparison.

2. Related Work

Large-scale deep learning studies established the feasibility of dermatologist-level image-based lesion classification and created the foundation for modern skin image analysis research [4]. Public resources such as HAM10000 and ISIC further accelerated progress by enabling reproducible training and evaluation across lesion recognition tasks [5], [7, 8]. DenseNet-based architectures became especially popular because dense connectivity improves gradient propagation and parameter efficiency, making them suitable baselines for medical imaging [9].

Transformer architectures further expanded the field. ViT demonstrated that pure transformer models can achieve competitive image recognition performance through patch-based tokenisation and self-attention [10]. Swin Transformer improved practical vision performance by introducing hierarchical shifted-window attention, which preserves both local detail and broader context [11]. Their growing impact in medical imaging has been documented in dedicated surveys [12, 13].

For mpox-specific imaging, Ali et al. [3] introduced a web-based screening system using state-of-the-art deep learning models, emphasising racial diversity and distinguishing mpox from clinically similar diseases. Campana et al. [15] proposed a transfer learning and explainable framework for mpox detection from smartphone images, highlighting the relevance of mobile and practical deployment scenarios. Dahiya et al. [16] reported strong results using hyperparameter-tuned deep learning for human monkeypox detection, but their setting did not focus on the official six-class MSLD v2.0 fold protocol [15]. Alhasson et al. [17] proposed a mobile application for monkeypox detection, again demonstrating feasibility but with a different dataset and task design.

More recent works have moved toward stronger benchmarking and broader frameworks. Vuran et al. [14] compared transformer-based models on multiclass skin lesion images, including mpox, and demonstrated the strength of Swin-based architectures. Hossain et al. [18] proposed a reliable deep learning framework for monkeypox diagnosis using skin images and highlighted the importance of robustness and practical diagnostic framing. Chakroborty [19] introduced a hybrid deep learning framework for early mpox detection using image data. Saha et al. [20] proposed the Mpox-XDE ensemble model with explainable AI. However, this work used a different four-class dataset and is therefore not directly comparable to the six-class MSLD v2.0 benchmark.

Despite this progress, several limitations remain. Many studies use different datasets, custom splits, binary settings, or reduced-class problems, which makes direct comparison difficult. Some report high accuracy, but these results are often obtained under easier conditions than those of six-class differential diagnosis. In addition, class imbalance is not always explicitly addressed, even though it strongly affects macro-level performance. These limitations motivate a fair, same-dataset, protocol-aware benchmark using balanced evaluation metrics.

3. Materials and Methods

Section 3 presents the overall methodological pipeline adopted for six-class mpox-related skin lesion classification using the MSLD v2.0 dataset. The study begins with the preparation of dermoscopic and clinical skin lesion images belonging to six categories, followed by standardised preprocessing, resizing, normalisation, and mild data augmentation to improve generalisation while preserving lesion characteristics. Three pretrained deep learning architectures, namely DenseNet121, ViT-Base, and Swin-Tiny, are then comparatively evaluated under an imbalance-aware training framework that incorporates class weighting, appropriate loss design, two-stage fine-tuning, validation-based checkpoint selection, and test-time augmentation. Finally, model performance is assessed using the official five-fold protocol and multiple classification metrics, including accuracy, macro precision, macro recall, macro-F1, and macro ROC-AUC. Figure 1 illustrates the overall workflow of the study.

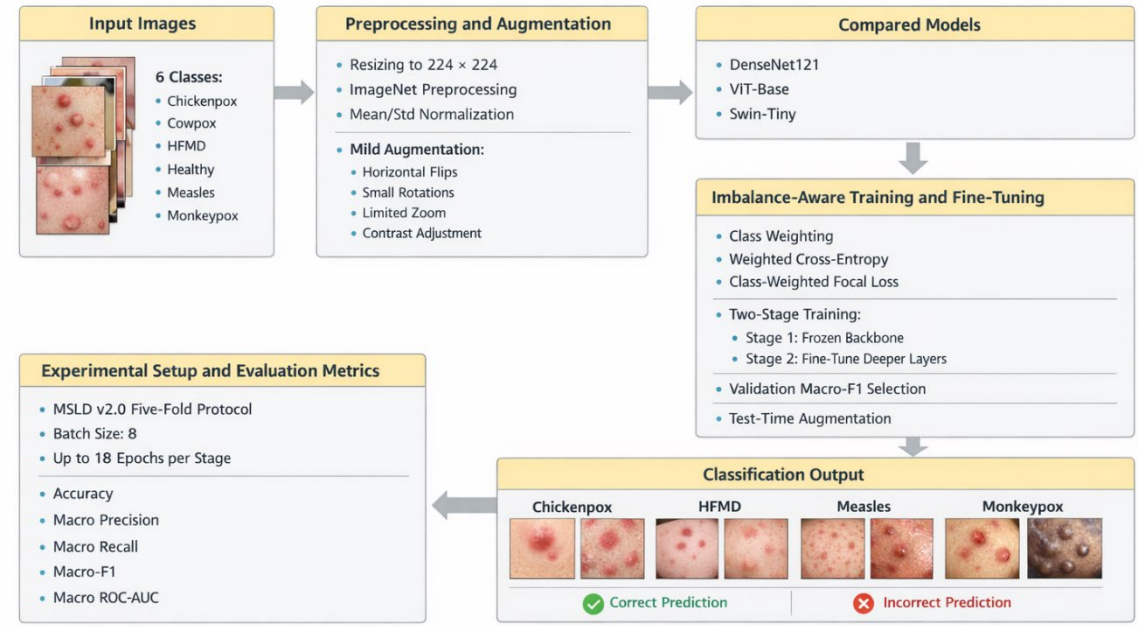


Figure 1: Overview of the proposed study design for six-class mpox-related skin lesion classification.

Figure 1 presents the overall workflow of the proposed study for the six-class classification of mpox-related skin lesions. The pipeline starts with preparing MSLD v2.0 images, followed by preprocessing and mild augmentation, and then proceeds to the comparative evaluation of DenseNet121, ViT-Base, and Swin-Tiny. To address class imbalance and improve generalisation, the training stage incorporates class weighting, loss adaptation, two-stage fine-tuning, validation macro-F1-based checkpoint selection, and test-time augmentation. The final performance is assessed using the official five-fold protocol with accuracy, macro precision, macro recall, macro-F1, and macro ROC-AUC.

3.1. Dataset

Experiments were conducted on MSLD v2.0, a six-class skin-lesion dataset containing images of Chickenpox, Cowpox, HFMD, Healthy, Measles, and Monkeypox [3]. Unlike ad hoc random splitting, MSLD v2.0 provides an official five-fold Train/Valid/Test structure, which was followed directly in this work to ensure reproducibility and comparability [3]. Table 1 presents an overview of the dataset.

Table 1
Class-wise distribution of the MSLD v2.0 dataset used in this study

	Chickenpox	Cowpox	HFMD	Healthy	Measles	Monkeypox	Total
Number of images	75	66	161	114	55	284	755

The MSLD v2.0 dataset used in this study contains 755 images across six classes, with Monkeypox as the largest and Measles as the smallest.



Figure 2: Representative sample images from the six classes of the MSLD v2.0 dataset used in this study.

Figure 2 shows representative examples from the six classes of the MSLD v2.0 dataset used in this study. The samples illustrate the visual diversity of lesion appearance across Chickenpox, Cowpox, HFMD, Healthy, Measles, and Monkeypox categories.

3.2. Preprocessing and Augmentation

All images were resized to 224 × 224 pixels. Model-specific preprocessing was applied to match pretrained backbones. DenseNet121 used DenseNet ImageNet preprocessing, whereas ViT-Base and Swin-Tiny used ImageNet mean and standard deviation normalisation. To reduce overfitting while preserving lesion characteristics, mild augmentation was applied during training, including horizontal flips, small rotations, limited zoom, and moderate contrast variation.

3.3. Compared Architectures

DenseNet121 was used as the CNN baseline because it is a widely accepted and computationally efficient architecture for medical image classification. ViT-Base models an image as a sequence of patches and uses global self-attention to capture long-range relationships. Swin-Tiny uses hierarchical shifted-window attention and is particularly suitable for fine-grained vision tasks because it balances local detail modelling with global contextual reasoning.

3.4. Imbalance-Aware Training

Class imbalance is common in medical datasets and can bias performance if only overall accuracy is optimised. To improve balanced learning, class weights were computed from each training fold. DenseNet121 was trained with weighted cross-entropy, whereas ViT-Base and Swin-Tiny used class-weighted focal loss [21]. Transformer models additionally used weighted random sampling during training.

3.5. Two-Stage Fine-Tuning

All models were trained in two stages. In Stage 1, only the task-specific head was trained, with the backbone frozen. In Stage 2, deeper layers were unfrozen and fine-tuned with a smaller learning rate. Checkpoint selection was based on validation macro-F1 rather than validation loss, since macro-F1 better reflects balanced class-wise performance in imbalanced multiclass tasks.

3.6. Test-Time Augmentation and Evaluation Metrics

At inference time, predictions from the original image and a horizontally flipped version were averaged. Test-time augmentation can improve robustness and reduce prediction instability in image classification [22]. Performance was evaluated using accuracy, macro-precision, macro-recall, macro-F1, and macro ROC-AUC.

3.7. Experimental Setup

All experiments were performed using the official MSLD v2.0 fold protocol [3]. The batch size was set to 8. Training was conducted for up to 18 epochs in each stage, with early stopping based on validation macro-F1. The initial learning rate was 1×10^{-4}, and the fine-tuning learning rate was reduced to 5×10^{-6}. DenseNet121 was implemented in TensorFlow/Keras, whereas ViT-Base and Swin-Tiny were implemented in PyTorch using pretrained backbones. Fixed seeds, GPU acceleration, and mixed precision were used to improve reproducibility and efficiency.

4. Results and Discussion

4.1. Quantitative and Qualitative Results

Table 2 summarises the mean and standard deviation values across the five official folds. The ranking is clear: Swin-Tiny outperformed ViT-Base and DenseNet121 on all reported metrics.

Table 2
Performance comparison on MSLD v2.0 (mean ± std across five official folds)

Model	Accuracy (%)	Precision_macro (%)	Recall_macro (%)	F1_macro (%)	ROC-AUC_macro (%)
DenseNet121	57.88 ± 10.02	56.60 ± 7.74	58.68 ± 4.51	53.87 ± 7.61	88.44 ± 2.07
ViT-Base	75.74 ± 5.02	74.89 ± 4.82	76.73 ± 2.24	74.63 ± 3.25	95.81 ± 0.97
Swin-Tiny	81.43 ± 4.19	79.70 ± 4.50	81.85 ± 4.24	80.03 ± 4.54	96.73 ± 1.52

The quantitative results show a clear advantage of transformer-based models over the CNN baseline. Swin-Tiny achieved the highest accuracy, macro precision, macro recall, macro-F1, and macro ROC-AUC, indicating more balanced recognition across all six classes. ViT-Base also performed strongly, while DenseNet121 remained a useful but clearly weaker baseline.

The performance gap suggests that hierarchical transformer representations are particularly effective for this dataset. Swin-Tiny likely benefits from combining local pattern modelling with broader contextual attention, which is important when lesions from different diseases share overlapping visual traits. Compared with ViT-Base, Swin-Tiny achieved better macro recall and macro-F1, indicating more balanced recognition across classes rather than improvement limited to dominant categories.

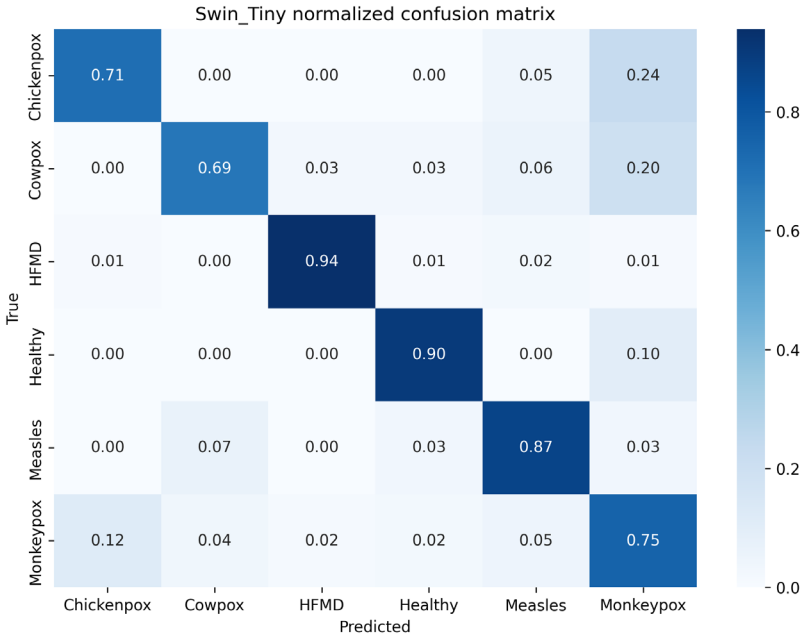


Figure 3: Normalised confusion matrix of Swin-Tiny on the combined test folds.

Figure 3 helps reveal the principal inter-class confusions. Such errors are clinically plausible because several lesion categories share related texture, colour, and distribution patterns. Even when overall macro-F1 is high, class-wise plots remain important because minority classes can still be more difficult than dominant categories in imbalanced medical datasets.

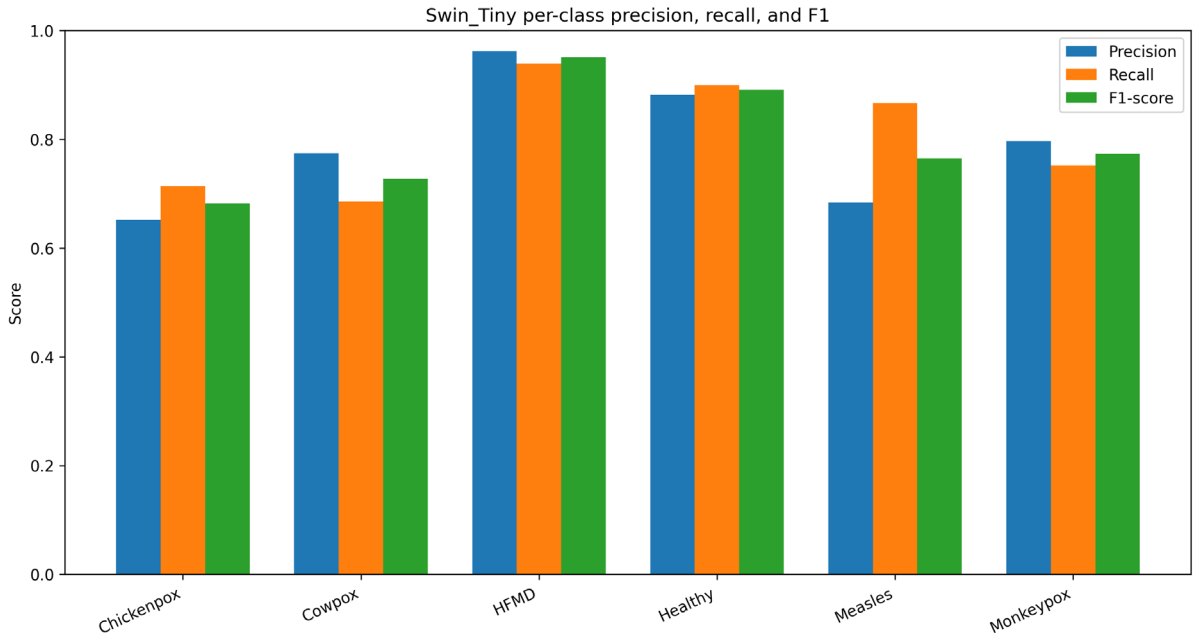


Figure 4: Per-class precision, recall, and F1-score of Swin-Tiny on the combined test folds.

Figure 4 shows class-specific precision, recall, and F1 scores. This provides a more transparent view of whether classes such as Cowpox or Measles remain weaker than larger classes. In an imbalanced setting, such plots complement macro-F1 by highlighting which categories still need further improvement.

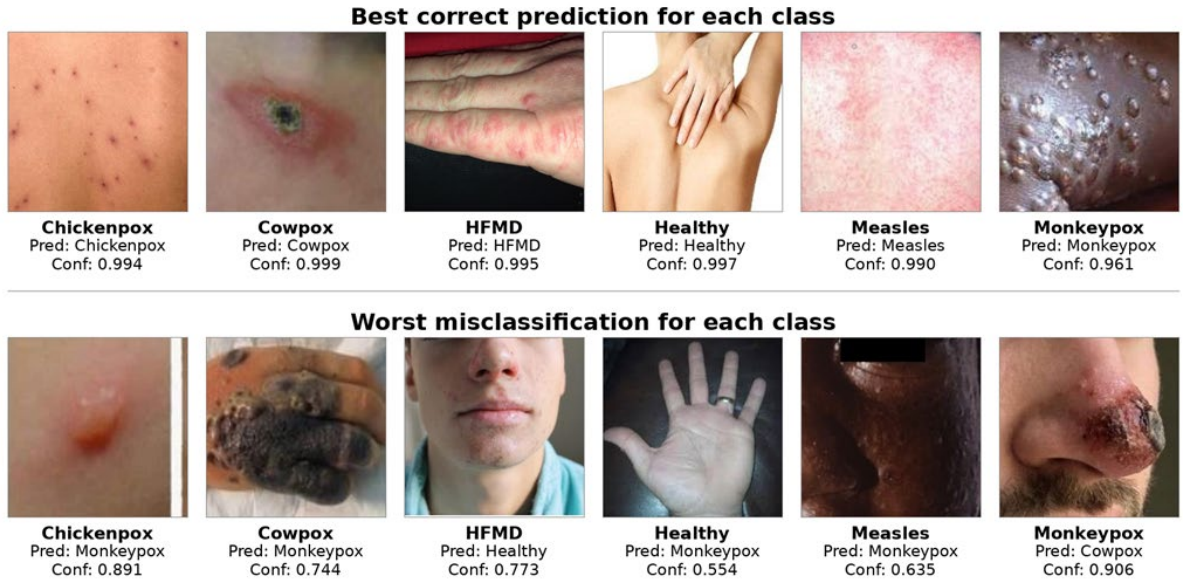


Figure 5: Best and worst per-class prediction examples of the Swin-Tiny model across all test folds.

Figure 5 presents per-class qualitative examples of the Swin-Tiny model across all test folds, showing the best correct prediction for each class in the top row and the worst misclassification for each class in the bottom row. Each sample is annotated with its predicted label and confidence score.

4.2. Comparison with Existing Studies

Table 3 compares representative related studies across different datasets and task settings. The comparison shows that many prior works reported strong results, but most are not directly comparable because they use different datasets, class definitions, or evaluation protocols.

Table 3
Comparison of related studies across different datasets

Study	Dataset	Task	Method	Main finding	Relation to our work
[3]	MSLD v2.0	6-class classification	CNN TL	Acc ≈ 83.59%	Same benchmark; CNN
[15]	Smartphone images	Binary/multiclass	TL + XAI	F1 = 87.9% (multiclass)	Different data/protocol
[16]	Human mpox images	Mpox detection	Tuned DL	Acc = 98.18%	Easier than a 6-class task
[17]	Mpox image dataset	Detection	Mobile deep learning app	Acc = 99%, F1 = 98%	Different data/task
[18]	Public mpox datasets	Diagnosis	Reliable DL framework	Acc = 95%	Not the same benchmark
[19]	Mpox image dataset	Early detection	Hybrid DL	Acc = 91.49%	Different data
[20]	MSID	4-class classification	Ensemble + XAI	Acc = 98.70%	Different data and class structure
[14]	MSLD v2.0	6-class classification	Transformer models	Acc =93.71%	Closest architecture study

Proposed work	MSLD v2.0	6-class official folds	DenseNet121, ViT-Base, Swin-Tiny	Acc = 81.43%, 80.03% macro-F1, 96.73% AUC	Controlled benchmark

Note: Direct numerical comparison across studies should be interpreted with caution, as the included works employ different datasets, class structures, preprocessing strategies, and evaluation protocols.

4.3. Same-Dataset Comparison

Table 4 presents the strictest comparison, limited to studies that used MSLD v2.0. This view is more scientifically meaningful because it distinguishes between same-dataset works that are only partially comparable and the present protocol-aware benchmark.

Table 4
Direct comparison with studies using the same dataset (MSLD v2.0)

Study	Protocol	Method	Reported result	Same protocol?	Remark
[3]	6-class benchmark	CNN TL	Acc ≈ 83.59%	Partial	Foundational same-data study
[14]	6-class transformer evaluation	ViT, MAE, DINO, Swin Transformer	Acc =93.71%	Partial	Closest same-data comparison
This work	Official 5-fold split	DenseNet121, ViT-Base, Swin-Tiny + imbalance-aware training + TTA	Swin-Tiny: 81.43% acc., 80.03% macro-F1, 96.73% AUC	Yes	Protocol-aware benchmark

The same-dataset comparison supports the same conclusion as the aggregate benchmark: transformer architectures, especially Swin-based variants, are better suited to the fine-grained six-class MSLD v2.0 task. The present paper strengthens that conclusion by emphasising macro-level metrics and a strict official-fold evaluation design.

4.4. Overall Interpretation

Taken together, the results indicate that transformer-based models are more suitable than the tested CNN baseline for this imbalanced six-class mpox-related lesion classification task. Swin-Tiny offers the best trade-off between overall recognition and class-wise robustness, while ViT-Base remains a strong alternative. DenseNet121 remains a valuable baseline because it shows that the performance improvement is not simply due to transfer learning, but rather to the representational strength of transformer architectures on this benchmark.

5. Conclusion

This paper presented a comparative benchmark of DenseNet121, ViT-Base, and Swin-Tiny on the official MSLD v2.0 five-fold protocol for imbalanced multiclass mpox skin lesion classification. The proposed pipeline incorporated class-weighted learning, focal loss for transformer models, weighted sampling, validation macro-F1 checkpointing, and test-time augmentation. The experiments demonstrated a clear advantage of transformer-based models over the CNN baseline.

Among the evaluated methods, Swin-Tiny achieved the best overall performance, reaching 81.43 ± 4.19% accuracy, 80.03 ± 4.54% macro-F1, and 96.73 ± 1.52% macro ROC-AUC. These findings suggest

that hierarchical vision transformers are particularly effective for balanced recognition of mpox and visually similar skin lesion categories. Future work should include external validation, calibration analysis, explainability, and evaluation on broader dermatological benchmarks.

Declaration on Generative AI

During the preparation of this work, the authors used Grammarly solely for grammar, spelling, and punctuation checking. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

References

[1] "Mpox(Fact sheet)," *World Health Organization*, no. 26 August 2024. 2024. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/mpox

[2] J. P. Thornhill *et al.*, "Monkeypox Virus Infection in Humans across 16 Countries — April–June 2022," *N. Engl. J. Med.*, vol. 387, no. 8, pp. 679–691, 2022, doi: 10.1056/nejmoa2207323.

[3] S. N. Ali *et al.*, "A web-based mpox skin lesion detection system using state-of-the-art deep learning models considering racial diversity," *Biomed. Signal Process. Control*, vol. 98, no. January, p. 106742, 2024, doi: 10.1016/j.bspc.2024.106742.

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

[5] P. Tschandl, C. Rosendahl, and H. Kittler, "Data descriptor: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Sci. Data*, vol. 5, pp. 1–9, 2018, doi: 10.1038/sdata.2018.161.

[6] M. Arabboev, S. Begmatov, M. Rikhsivoev, S. Bobojanov, G. Tangriberganov, and J. Uraimov, "Deep learning-based skin cancer classification: a comparative study of pre-trained CNN models," in *30th International Conference Information Society and University Studies - IVUS 2025*, 2025.

[7] D. Gutman *et al.*, "Skin Lesion Analysis toward Melanoma Detection: A Challenge at the International Symposium on Biomedical Imaging (ISBI) 2016, hosted by the International Skin Imaging Collaboration (ISIC)," *arXiv*, pp. 1–5, 2016, [Online]. Available: http://arxiv.org/abs/1605.01397

[8] N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schlaefer, "Skin lesion classification using ensembles of multi-resolution EfficientNets with meta data," *MethodsX*, vol. 7, p. 100864, 2020, doi: 10.1016/j.mex.2020.100864.

[9] G. Huang, L. Van Der Maaten, and K. Q. Weinberger, "Densely Connected Convolutional Networks," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2017)*, 2017, pp. 2261–2269.

[10] A. Dosovitskiy *et al.*, "An Image Is Worth 16X16 Words: Transformers for Image Recognition At Scale," *ICLR 2021 - 9th Int. Conf. Learn. Represent.*, 2021.

[11] Z. Liu *et al.*, "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows," in *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, Oct. 2021, pp. 9992–10002. doi: 10.1109/ICCV48922.2021.00986.

[12] F. Shamshad *et al.*, "Transformers in medical imaging: A survey," *Med. Image Anal.*, vol. 88, pp. 1–41, 2023, doi: 10.1016/j.media.2023.102802.

[13] S. Khan, H. Ali, and Z. Shah, "Identifying the role of vision transformer for skin cancer—A scoping review," *Front. Artif. Intell.*, vol. 6, 2023, doi: 10.3389/frai.2023.1202990.

[14] S. Vuran, M. Ucan, M. Akin, and M. Kaya, "Multi-Classification of Skin Lesion Images Including Mpox Disease Using Transformer-Based Deep Learning Architectures," *Diagnostics*, vol. 15, no. 3, 2025, doi: 10.3390/diagnostics15030374.

[15] M. G. Campana, M. Colussi, F. Delmastro, S. Mascetti, and E. Pagani, "A Transfer Learning and Explainable Solution to Detect mpox from Smartphones images," *Pervasive Mob. Comput.*, vol. 98, 2024, doi: 10.1016/j.pmcj.2023.101874.

[16] N. Dahiya *et al.*, "Hyper-parameter tuned deep learning approach for effective human

monkeypox disease detection," *Sci. Rep.*, vol. 13, no. 1, pp. 1–18, 2023, doi: 10.1038/s41598-023-43236-1.
[17] H. F. Alhasson, E. Almozainy, M. Alharbi, N. Almansour, S. S. Alharbi, and R. U. Khan, "A Deep Learning-Based Mobile Application for Monkeypox Detection," *Appl. Sci.*, vol. 13, no. 23, 2023, doi: 10.3390/app132312589.
[18] M. S. Hossain, M. Ahmed, and M. S. Rahman, "From survey to solution: A deep learning framework for reliable monkeypox diagnosis using skin images," *Array*, vol. 28, no. July, p. 100554, 2025, doi: 10.1016/j.array.2025.100554.
[19] S. Chakroborty, "A hybrid deep learning framework for early detection of Mpox using image data," *Healthc. Anal.*, vol. 7, no. January, p. 100396, 2025, doi: 10.1016/j.health.2025.100396.
[20] D. Kumar Saha *et al.*, "Mpox-XDE: an ensemble model utilising deep CNN and explainable AI for monkeypox detection and classification," *BMC Infect. Dis.*, vol. 25, no. 1, 2025, doi: 10.1186/s12879-025-10811-y.
[21] T. Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollar, "Focal Loss for Dense Object Detection," *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2017-Octob, pp. 2999–3007, 2017, doi: 10.1109/ICCV.2017.324.
[22] L. Cino, C. Distante, A. Martella, and P. L. Mazzeo, "Skin Lesion Classification Through Test Time Augmentation and Explainable Artificial Intelligence," *J. Imaging*, vol. 11, no. 1, pp. 1–28, 2025, doi: 10.3390/jimaging11010015.