Iraqi Journal for Applied Science (IJAS)
Vol. 3, No. 2, June, 2026, pp. 35-42
E-ISSN: 3006-5828 DOI: 10.69923/qfnn3n63

Research Article

Robust Patient-Level Evaluation of Post-Training Quantization for Edge-Deployable MPox Skin Lesion Classification

Haitham Qutaiba Ghadhban[1], **Dina Fitria Murad**[2]
[1] Department of Computer Engineering, College of Engineering, Diyala University, Diyala, IRAQ.
[2] Information Systems Department, Bina Nusantara University, Jakarta, Indonesia.

Article Information

Article history:

Received: 03, 01, 2026
Revised: 30, 03, 2026
Accepted: 01, 05, 2026
Published: 30, 06, 2026

Keywords:

Artificial intelligence,
MPox classification,
Patient-level split,
INT8 inference,
Post-training quantization,
Medical image,

Abstract

Portable computational imaging has emerged as a promising approach for infectious disease screening. However, many medical image classification studies rely on image-level random data splits and evaluate model performance using a single training run, often leading to overly optimistic performance estimates. In this study, we implemented Post-Training Quantization (PTQ) for Mpox skin lesion classification using a patient-level evaluation protocol based on a stringent pre-registered experimental design. The publicly available MSLD v2.0 dataset was preprocessed at the pixel level, resulting in 755 unique original images from six diagnostic classes representing 524 patients. A stratified patient-level split (80% training, 10% validation, and 10% testing) was applied to prevent data leakage. Each experiment was repeated across three independent data splits and three training seeds to assess performance variability and robustness. A two-stage transfer learning framework based on MobileNetV2 was trained and evaluated using FP32, Float16 PTQ, and full INT8 quantization implemented through TensorFlow Lite (TFLite). The FP32 baseline achieved an accuracy of 69%, a macro F1-score of 69%, and a macro-AUC of 94.7%, with a model size of 4.27 MB. Float16 quantization maintained comparable predictive performance and model size, whereas full INT8 quantization reduced the model size to 2.59 MB, representing a 39% reduction compared with the FP32 baseline. Top-2 accuracy remained stable at approximately 89% across all precision regimes, supporting the feasibility of deployment in triage-oriented clinical settings. Variability analyses revealed that performance fluctuations were driven primarily by patient partitioning rather than reduced numerical precision, demonstrating that INT8 quantization preserves clinically relevant performance while substantially reducing computational and storage requirements for deployment in resource-constrained edge environments.

Corresponding Author: Haitham Qutaiba Ghadhban, Department of Computer Engineering, College of Engineering, University of Diyala, Iraq. ***Email***: haithamqutaiba@uodiyala.edu.iq

1. INTRODUCTION

The worldwide distribution of Mpox, formerly called monkeypox, has become a major public health issue because of its high transmissibility and inability to differentiate from visually similar skin lesions at an early stage using clinical manifestations only. Mpox still represents a health burden and may require rapid point-of-care tests to aid clinical management as well as for surveillance purposes, according to the World Health Organization. Deep learning has shown promising results in medical image analysis and skin lesion classification, providing automatic recognition of dermatologic disorders, which present a similar pattern to that of Mpox-related infections [1]. Several recent studies have employed deep CNNs for Mpox classification in skin lesion images. For instance, in this work [2], they developed and evaluated a series of these modern CNN architectures, such as MobileNetV2 and EfficientNet, on the Mpox Skin Lesion Dataset (MSLD v2. 0) and have shown that deep models can differentiate Mpox from other infectious skin diseases. In other studies, transfer learning is also used with pre-trained models, such as VGG, ResNet and mobileNet variants, that can reach high classification accuracy between Mpox vs. non-Mpox cases from the appointed dataset [3].

However, a majority of these works use traditional image-level random split and single-run evaluation that may cause artificial performance boost through data leakage between training and test sets and do not provide insight into the variability under different training conditions. Meanwhile, model quantization has become a popular approach to facilitate CNN inference in resource-constrained edge and fog environments by decreasing the model precision to low-bit ones like INT8 or FP16. Quantization reduced memory usage and model complexity, which is required for AI models running on mobile and edge devices [4]. In the context of PTQ, pretrained full-precision models are transformed into lower precision models after training without retraining process, which benefits practical deployment. However, existing research in quantization applied to medical imaging focused on segmentation or general classification benchmarks and did not assess conditional performance robustness under strong patient-level separation protocols [5].

In addition, previously published work on Mpox imaging models rarely accounted for quantized inference or an extensive robustness study among multiple data set partitions and training seeds. To counter these limitations, this work explores the robustness of Mpox skin lesion classification for strong patient-level separation and low-precision inference. A defined test routine is used to analyze model stability in various data splits and training setups. Furthermore, posttraining quantization approaches are investigated to evaluate their influence on the predictive capacity and size of models in an edge setting. In health care distributed architecture, it is possible to perform inference on the edge and IoT layer to minimize latency and protect data privacy. A quantized model with lightweight is more suitable for such edge-fog deployment, which has limited computational capacity and storage. This work reports on the following contributions:

1. Implements a rigorous stratified patient-level splitting strategy on 755 distinct Mpox skin lesion images to prevent intrapatient data contamination.
2. Proposes a two-stage transfer learning model with focal loss using MobileNetV2 for effective phase classification.
3. Assesses model robustness over three patient levels splits and three training seeds per split, with mean SD reported for all performance scores.
4. Evaluates the effect of posttraining quantization (Float16 PTQ and INT8 PTQ) on classification accuracy and model size for the edge scenario.
5. Deep assessment of performance through accuracy, macro-F1, macro-AUC and MKP recall.

2. METHOD

Through structured searches of established academic databases, including Scopus, Web of Science, IEEE Xplore, SpringerLink, ScienceDirect, and PubMed, using keywords such as “Mpox” or “Monkeypox”, ”skin lesion classification”, “deep learning”, “transfer learning” and “vision transformer”, relevant literature was identified in turn. This review focused on image-based Mpox skin lesion classification studies in publicly indexed and peer-reviewed literature, principally those employing publicly available datasets such as MSLD or MSLD v2.0 with sufficient methodological details, including dataset composition, model architecture, split protocol, patient-level separation strategy, multi-seed or multi-run testing, quantization or deployment precision analysis, and evaluation protocol. Studies were excluded if synthetic images provided the basis of those investigations, or if nonimage diagnostic modalities were addressed. Studies that did not provide sufficient transparency to allow meaningful protocol-level comparison were also excluded. This structured screening procedure means that the review emphasizes only those Mpox classification systems, which are reproducible and image-based, with praticular attention to split rigor, patient-level data separation, robustness across multiple seeds or partitions, post-training quantization, and external validation reporting.

One of the main catalysts for this recent progress was readily available data on which multiclass classification could be performed reliably by any human(Table 1). This research [2] presents the MSLD v2.0 (Mpox-Skin Lesion Dataset), which is a multi-label dataset including Mpox and five confounder classes providing benchmarks for numerous transfer-learning CNN backbones. Numerous techniques are employed to counteract bias in notion or augmentation, although color space is a particularly sharp straw in this respect. Implanted with the logic of screening for the 21st century, this work was later included in a journal version that maintained its focus on datasets and consideration in racially diversified settings, alongside an initial web application. These dataset-based contributions made MSLD v2.0 a well-recognized target and incentivized follow-on studies, which systematically compared deep learning architectures under similar formulations. Since then, CNN baselines have been broadened to examine transformer variants and hybrid architectures. This work [6] was the first to benchmark various contemporary architectures, including CNNs and transformers, for Mpox-related dermatological classification, and the authors described the use of cross-validation and high performance scores for select models. Meanwhile, systems-level papers have integrated classification pipelines into larger monitoring or mobile health systems. For instance, a recent diagnostics study proposed a mobile Al-based monitoring system for Mpox and compared the performances of pretrained CNN and transformer models on binary, and multiclass Mpox classification and described mobile deployment components [7].

Other lines of research investigate different modeling approaches and mixed data set generation. A springer conference chapter proposed a hybrid ensemble model with reported high accuracy, where their focus was on architecture fusion and not deployment precision [8]. Another study examined a ViT-CNN ensemble on a concatenated data set consisting of several sources, which is informative for architecture design since such mixed-dataset pipelines make it difficult to directly compare with MSLD v2. 0 only evaluation [9].

Other research emphasizes the understandability and comparable evaluation of prevailing pretrained CNNs for Mpox detection classification, normally generalizing standard performance metrics and visualization tools such as Grad-CAM [10]. Mpox classification studies are gradually introduced to the mobile or web deployment, but in this area, reduced-precision inference and PTQ as experimental variables have been mentioned somewhat peripherally until now. In general, medical image quantization has been recognized as an application to reduce model size and improve inference efficiency, but a systematic evidence review documents that research across different domains of medical imaging tasks is highly technical and methodologically disparate [4]. Practical PTQ studies also stress that simulated quantization can give a false picture of real deployment behavior, so evaluation on actual inference engines is motivated [5].

Overall, prior Mpox lesion classification research has demonstrated feasibility across CNNs, transformers, and hybrid designs, but reported results often remain difficult to compare due to differences in dataset construction, evaluation design, and limited reported stability across multiple random seeds or independent partitions. Quantization and deployment precision regimes are rarely evaluated as first-class variables in Mpox-focused studies. This leaves an actionable gap in robustness-oriented, deployment-aware evaluation under strict data separation and reported-run protocols.

Table 1. Illustrates the recent studies for Mpox classification

Study/Year	Dataset	Techniques	Evaluation	Limitations
[2] 2024	MSLD v2.0	CNN benchmark and prototype web screening, augmentation focus.	System framing	Emphasis on dataset and generalization, no PTQ/INT8 analysis.
[6]2025	Mpox-related dermatological dataset	CNNs with transformer.	Cross-validation described, high reported performance	Primary focus is architecture benchmarking, quantization not studied
[7] 2025	Public dataset	Transfer learning with CNNs and transformers.	Binary and multiclass evaluation reported	Quantization not treated as primary variable, protocol differs from robustness-focused reported runs.
[8] 2025	Mpox lesion dataset	Hybrid ensemble	Single reported results emphasizing accuracy.	Focus on hybrid architecture, no reduced-precision inference evaluation.
[9] 2024	Combined dataset from two sources	ViT fine-tuning and CNNs ensemble.	Standard split pipeline described	Uses mixed datasets, not directly comparable to MSLD v2.0-only studied, quantization not evaluated.
[10] 2025	MSLD and MSLD v2.0	Pretrained CNN	Standard metrics reported	No PTQ/INT8 evaluation.
[11] 2024	New released Mpox dataset	CLIP encoder with ViT classifier and LLM	VLM-style diagnostic framing	Not comparable to pure image multiclass MSLD.
[4] 2026	Broad medical imaging	Meta-analysis of quantization in medical imaging	Synthesis across studies	Not Mpox-specific, highlights reporting heterogeneity rather than providing a Mpox benchmark.
[5] 2025	3D medical segmentation	PTQ on real inference engines	Practical PTQ evaluation	Segmentation domain, not Mpox specific.

Despite this impressive performance, direct comparison with these studies is necessary given the difference in the structure, split protocol and evaluation process. Additionally, even though deployment is often cited as a reason for a given approach to be preferable to others, post-training quantization and robustness (FP32, FP16 and INT8) are still relatively rare as an first-class experimental variable in Mpox-focused works. Such gaps call for deployment-aware evaluations that characterize performance robustness in a well-defined operational separation and arranged experimental setting.

3. MATERIALS AND METHODS

3.1. Dataset preparation

The experiments were performed on the publicly available Mpox Skin Lesion Data set v2.0 (MSLD v2. 0) [2] with images of Mpox and look-alike dermatologic conditions, such as chickenpox, measles or cowpox, hand-foot- mouth disease (HFMD), and normal skin. The data are pre-divided into folds, and augmented subsets for benchmarking. To maintain methodological transparency and minimize data leakage, this study used original images only. The packaging of data set has several folds where each identical image repeats in other partition folder. As such, a collection of all original images was manually cleaned to keep only unique original samples. Image uniqueness was assessed by the use of identical filename identifiers for each lesion instance, and duplicates were removed within folds. As a result of this process, 755 original unique images were obtained from the six classes, which correspond to 524 different patients. The final curated data set was used for all further analyses. Artificial data were generated or external images were used. Focusing on the unique original samples and removing fold level repetition. The work guarantees that model assessment is based on real generalization but not performance induced by artifacts.

3.2. Patient-level stratified splitting

Data set splitting was conducted at the patient rather than at the image level to avoid intrapatient data leakage and maintain clinically meaningful assessment. Because the same individual may be matched from multiple lesion images, random splitting could cause test samples to be identical with the training sample and bring about a maximum performance rate. To resolve this problem, all images of one patient were assigned to one partition.

An 80/10/10 stratified split was used on the curated data set to generate training, validation and test sets, maintaining class distribution at patient level. The stratification was conducted over class labels to ensure that class imbalance did not affect different partitions. To investigate the consistency of our results concerning data splitting, we also repeated the patient-level split using three different split seeds. For each split configuration, a different set of patients was assigned to training, validation and test sets with preserved class proportions.

This setup allows to study how the choice of data set splitting affects performance instead of using only one fixed split. Strict non overlapping of patient numbers between the partitions was required throughout all splits. Therefore, the reported results are generalization to new patients and not memorization of patient-specific visual features. The details of the class distribution are shown in Table 2. Note that some classes contain notably small test sets such as Measles, Cowpox, and Chickenpox, which may increase variability in class specific recall estimates for those categories.

Table 2. Patient and image distribution per class across splits.

Class	Train	Validation	Test
Chickenpox	58	10	7
Cowpox	56	4	6
Healthy	92	11	11
HFMD	125	19	17
Monkeypox	230	22	32
Measles	45	5	5
Total	606	71	78

3.3. Model Architecture and Training Strategy

We used MobileNetV2 [12] for transfer learning, which is light enough to be deployed in constrained devices. MobliNetV2 uses an inverted residual structure with a linear bottleneck that is more efficient for the representation information and maintain low computational complexity. The backbone was pretrained with ImageNet weights, and the final classification layer was replaced with a fully connected layer specific to the six classes of interest, followed by activation using softmax. The MobileNetV2 backbone contains around 2.3 M parameters. In first stage, 6K parameters were used to train in the classification head and in second stage, 30 layers unfroze with 500k parameters for fine-tuning. Input images were resized to 224*224 pixels and processed using a batch size of 32. Pixel values were normalized to [0,1] by dividing by 255 before feeding into the model. To improve generalization and mitigate overfitting, online data augmentation was applied during training, including random horizontal flipping, small-angle rotation 8%, zooming 10% and contrast adjustment 10%. No synthetic data generation or external augmentation sources were introduced.

A two-stage training strategy was employed. In stage 1, the backbone network was frozen, and only the classification head was trained using the Adam optimizer with a learning rate of 1x10-3. In stage 2, the last 30 layers of the backbone were unfrozen to allow fine-tuning and the learning rate was reduced to 1x10-4 to stabilize convergence. The categorical cross-entropy loss was used without label smoothing.No explicit class weighting was applied, class imbalance was addressed through focal loss and stratified splitting. Training was conducted for a maximum of 20 epochs with early stopping applied based on validation performance using patience of 3 epochs. Deterministic operations and controlled random seeds were enabled within TensorFlow to ensure reproducibility across runs. No dropout or L2 regularization was applied beyond the implicit regularization provided by early stopping and data augmentation.

3.4. Post-Training Quantization

To evaluate deployment feasibility under reduced-precision inference, post-training quantization (PTQ) was applied to the trained FP32 models using the TensorFlow Lite (TFLite) conversion framework. Three precision regimes were considered: the first precision is the FP32 Baseline, and the original trained model using 32-bit floating point precision was retained as the reference configuration. The second precision is Float16, hybrid float16 quantization was applied, in which model weights were converted to 16-bit floating-point representation while inference computations remained in floating-point precision. This strategy saves memory for storing the model and maintains numerical stability during inference. The final precision is Full INT8, where we performed full integer quantization by transforming both weights and activation into 8-bit integers. A randomly selected subset of the training images is used as a sample calibration data set upon conversion to compute activation scaling parameters for integer inference. The resulting model can perform full inference using INT8 calculation throughout the TFLite runtime.

Only posttraining quantization was used, and we did not perform any quantization-aware training (QAT) or further fine-tuning with lower precision. All quantized models were tested against the TFLite interpreter to get results on realistic deployment conditions instead of simulating the use of quantization. Note that the model size was calculated directly from serialized TFLite files, and compression ratios were relative to FP32 baseline. Quantized models were evaluated on patient-level test partitions identical to those used for the FP32 model to ensure fair comparisons across precision levels. Such a design allows directly investigating the compromise of reduced model size vs. predictive quality under precision reduction tailored for deployment. The general methodology setup and details of patient-level splitting, multiseed training, and training quantization evaluation are presented in Figure 1.

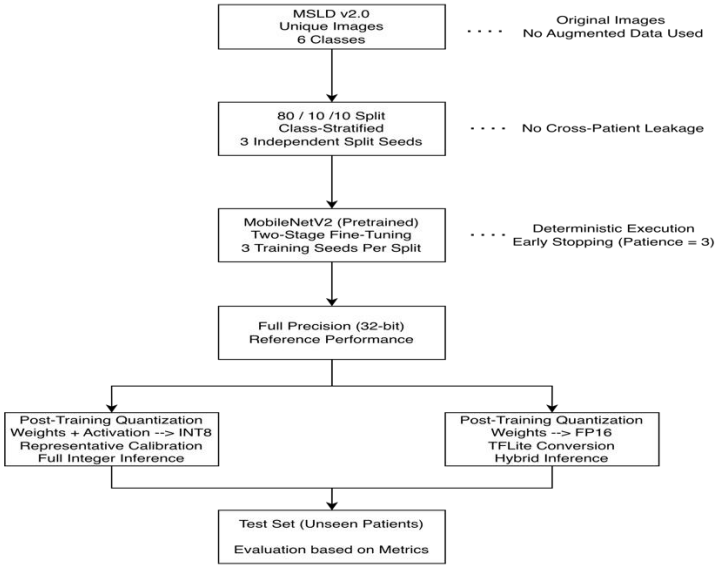

Figure 1. The framework methodological shown patient-level evaluation and post-training quantization.

3.5. Evaluation Protocol

Model performance was evaluated using a structured multi-split, multiseed experimental design to assess robustness with respect to both dataset partitioning and stochastic training variability. For each of the three independent patient-level split seeds described in Section 3.2, models were trained using three distinct training seeds, resulting in nine independent training configurations per precision regime. This design enables the separation of variability arising from data partitioning and random initialization effects.

The validation set was used exclusively for early stopping and model selection, while all final performance metrics were computed strictly on the held-out test set composed of previously unseen patients. The following evaluation metrics were reported, Accuracy, Macro-F1, macro-AUC, class-specific recall for Mpox (MKP Recall), Top-2 accuracy, and model size. Macro-average was adopted to account for class imbalance across the six diagnostic categories. Macroaverage was computed using a one-vs-rest strategy and averaged across classes. Top-2 accuracy was defined as a correct prediction when the ground truth was among the two highest-confidence class probabilities. For each metric, results are reported as mean SD across all split and training seed combinations. All quantized models were evaluated using identical test partitions to ensure direct comparability with the FP32 baseline model size 4.27 MB.

4. RESULTS AND DISCUSSION

4.1. Overall Performance Across Precision Regimes

Table 3 summarizes the aggregated performance across all patient-level splits and training seeds. Results are reported as mean ± SD over nine independent configurations per precision regime.

Table 3. Comparative Performance Across Precision Regimes (Mean ± std)

Precision	Accuracy	Macro-F1	Macro-AUC	MKP Recall	Top-2 ACC	Model Size (MB)	Compression Ratio
FP32	0.693 ± 0.063	0.690 ± 0.090	0.947 ± 0.022	0.506 ± 0.103	0.888 ± 0.049	4.27	1.00x
Float16 (PTQ)	0.693 ± 0.062	0.689 ± 0.087	0.947 ± 0.022	0.506 ± 0.103	0.880 ± 0.052	4.27	~1.6x
Full INT8 (PTQ)	0.677 ± 0.060	0.678 ± 0.082	0.944 ± 0.026	0.470 ± 0.101	0.885 ± 0.072	2.59	~2.7x

With strict patient-level voting analysis, the FP32 baseline yielded a mean accuracy of 0.693 ± 0.063 and macro-F1 = 0.690 ± 0.090, respectively. The macro-AUC was consistently high at 0.947 ± 0.022, with stable interclass separability over independent splits. The slightly higher variance seen for MKP recall 0.506 ± 0.103 than the other aggregate metrics indicates that it is sensitive to class imbalance and partition-specific sample distribution.

4.2. Statistical Assessment of Quantization Impact

Statistical significance testing paired t-test was performed to check the performance difference between FP32 and Full INT8 models for each split seed configuration. The p-values were 0.358 for accuracy, 0.494 for macro-F1 and 0.155 for MKP recall. None of the differences was statistically significant at the P = 0.05 level. The effect sizes (Cohen's d) were small for accuracy (d = 0.325) and macro-F1 (d = 0.239) and moderate for MKP recall (d = 0.523). The results show that the loss of performance induced by full INT8 quantization is not statistically significant using the tested protocol. The difference lies within the fluctuation caused by CNN training and patient division.

4.3. Robustness Across Patient-Level Splits

Note that the interquartile ranges of different precisions, including FP32, Float16 and INT8, intersect with each other(Figure 2, Figure 3), indicating that more variation is due to partitioning rather than precision loss. Remarkably, the gap between different splits is larger than the average accuracy loss due to INT8 quantization. This emphasizes the necessity of multisplit evaluation when evaluating medical image classification systems over patient-level isolation.

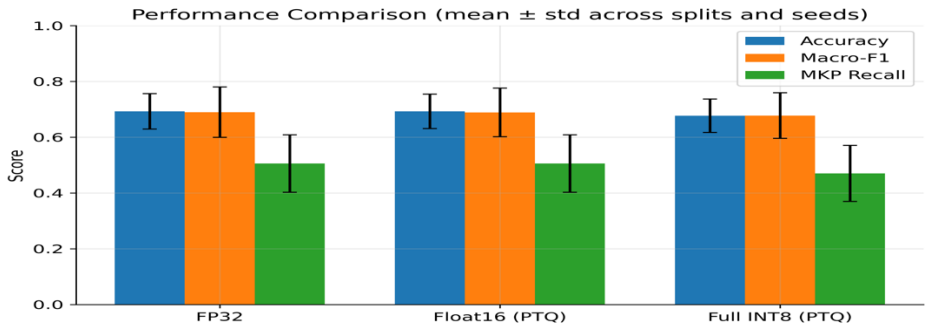

Figure 2. Illustrates aggregated mean ± standard deviation performance across precision regimes.

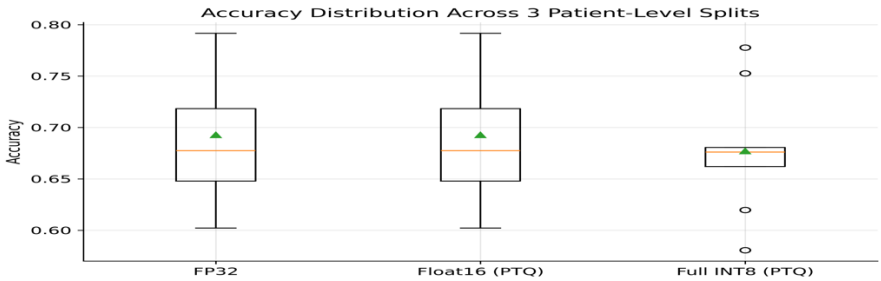

Figure 3. Presents the distribution of accuracy across the three independent patient-level splits.

4.4. Deployment Trade-Off Analysis

After applying float16 quantization, the model size decreased to 4.27 MB, followed by full INT8 compression leading to a final model size of 2.59 MB, which corresponds to a nearly 40% reduction compared to float16 representation and can be appreciated as significant in comparison to a full precision deployment. While the storage requirement has decreased by this reduction in accuracy is only 1.6 percentage points on average, which is even not significant according to our statistical testing. These findings indicate that posttraining integer quantization achieves a balance between computational efficiency and the stability of predictive performance, showing potential for edge deployment. All experiments were conducted on a kaggle through Tesla P100 of 16 GB RAM. TFLite inference was executed on the same hardware using the CPU backend to simulate realistic edge deployment conditions. (Figure 4)

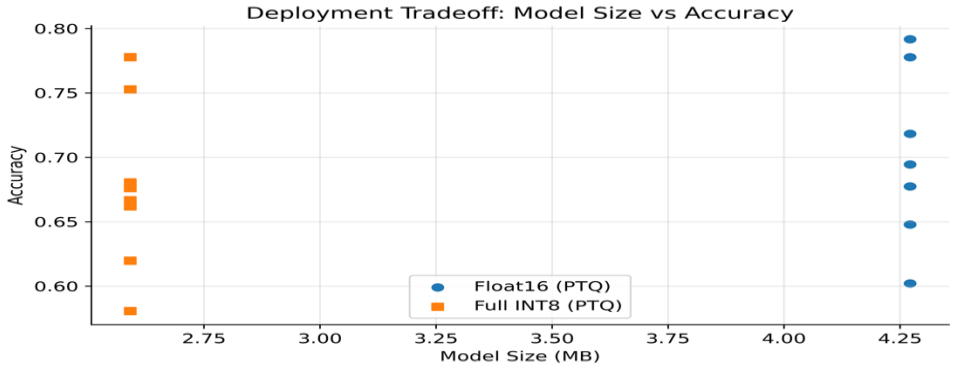

Figure 4. Represents the relationship between model size and accuracy across precision regimes.

4.5. Class-Level Sensitivity and Top-2 Analysis

Whereas the full retrieval measures stay constant under quantization, MKP recall is more unstable across runs, thus showing sensitivity to class distribution on specific splits. The decrease in MKP recall with INT8 quantization was not significant. Top-2 accuracy also remained high for all precision regimes at approximately 0.88, indicating that when the top-1 prediction is wrong, the true disease often still exists among the highest-confidence outputs. This is consistent with potential use as a triage or decision-support system rather than a diagnostic replacement.

4.6. Comparison with Recent Studies

Most Recent MSLD Studies Table 4 presents selected recent findings using the MSLDv2. 0 multiclass data set. Due to essential differences in the split protocol between these studies, and different augmentation strategy or patient-level isolation, evaluation design and direct numerical accuracy ranking across these works are meaningless. The point of comparison is therefore methodological: it shows not which study achieved the largest headline number, but what evaluation choices each one made. Finally, the present work is unique in its patient-level stratification, multi-split × multi-seed robustness design, statistical significance testing, and systematic post-training quantization evaluation, none of which are reported in the studies considered to date.

Table 4. Comparison with Recent MSLD v2.0 Multiclass Classification studies.

Study	Dataset/Task	Split strategy	Best model	Reported performance	Robustness reporting	Deployment/ Quantization
[2]	MSLD v2.0 / 6 classes	5-fold cross-validation, extensive augmentation	DenseNet121	83.59% accuracy	Mean ± std across fold	No quantization analysis
[13]	MSLD v2.0 / 6 classes	Predefined fold protocol	SwinTransformer	93.71% accuracy	Model comparison across folds	No quantization analysis
[7]	MSLD v2.0 / 6 classes	Train-test split + 5-fold cross-validation	Hybrid CNN/ViT	92.16% test accuracy	Cross-validation reported	Deployment system focus, no PTQ study
This work	MSLD v2.0 / 6 classes (755 unique originals)	Strict patient-level stratified splitting, multi-split x multi-seed	MobileNetV2 (fine-tuned)	0.693 ± 0.063 accuracy, 0.690 ± 0.090 macro-F1	Mean ± std across splits x seeds, statistical testing	Float16 and full INT8 PTQ evaluated, compression trade-off analyzed

Across recent MSLD v2.0 studies, the reported multiclass accuracies are often obtained on top of pre-determined folds or cross-validation settings and often also on top of extensive augmentation. In contrast, the current study employs explicit patient-level stratification, multi-split × multi-seed robustness evaluation and deployment-focused precision analysis. While this more conservative protocol does yield lower headline accuracy, it offers a higher burden of proof that generalization stability is maintained under stringent patient isolation and provides the first quantification of post-training quantization effects on classifier performance. Notably, we used statistical testing at P = 0.05 to show that the performance differences between FP32 and INT8 regimes are not statistically significant, thus verifying integer quantization is actually feasible with minimal loss. Thus, instead of presenting the proposed framework as a method that provides state-of-the-art raw performance, the contribution here is in demonstrating the combination of rigorous patient-level evaluation with systematic deployment-oriented compression analysis, which is not commonly reported in previous multiclass Mpox classification studies.

5. CONCLUSION

In this study, we presented a comprehensive patient-level assessment framework for multiclass Mpox skin lesion classification on the MSLD v2.0, designed with the highest standards of rigor data set. The proposed protocol removes intrapatient information leakage by limiting the data set to unique original images and implementing a strict patient-level stratified split that serves as a conservatively realistic approximation of generalization performance. The FP32 baseline was stable across many independent splits and training seeds when utilizing a MobileNetV2 backbone and two-stage fine-tuning. Compared to single-run evaluations, the multi-split × multi-seed design was able to explicitly quantify this performance variability and show that partition-based variability is greater than marginal effects from reduced-precision inference. This evaluation of posttraining quantization was systematic and under realistic TFLite inference conditions. Full INT8 quantization does not lead to a statistically significant drop in accuracy or macro-F1 P = 0.05 while reducing the model size to 2.59 MB, statistical testing conforms. Thus, these results demonstrate that coupled with this restrictive patient isolation, lightweight convolutional architectures enable models with a high degree of reduction in size whilst maintaining diagnostic relevant performance, which is beneficial for low-latency edge and fog deployments in resource constrained healthcare infrastructures.

This work emphasizes the rigor and reproducibility of the evaluation and feasibility of deployment over optimal headline accuracy. Results does autonomous MSLD v2.0 iterates on these to show that simultaneous patient-level generalization and precise-regime stability are possible and to demonstrate this through benchmarks. Several limitations should be acknowledged. First, only one publicly available data set was used to evaluate these approaches, meaning domain shift was not considered. Further, independent cohorts are necessary to assess robustness. Second, while leakage is less of a concern with patient-level splitting, this leads to a further constraint in the overall data set size with reduced class-specific recall variability. Third, only posttraining quantization was studied; future work should explore quantization-aware training and hardware-level latency evaluation to better define deployment effectiveness. Lastly, the framework is a decision-support system and not a stand-alone clinical diagnostic system. Both the implementation and the experimental setup are entirely reproducible, and further extensions toward cross-dataset validation and hardware-aware optimization are valuable avenues for future research.

REFERENCES

[1] A. H. Thieme et al., "A deep-learning algorithm to classify skin lesions from mpox virus infection," Nat. Med., vol. 29, no. 3, pp. 738–747, 2023. DOI: 10.1038/s41591-023-02225-7

[2] 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, p. 106742, 2024. DOI: 10.1016/j.bspc.2024.106742

[3] A. S. Jaradat et al., "Automated monkeypox skin lesion detection using deep learning and transfer learning techniques," Int. J. Environ. Res. Public Health, vol. 20, no. 5, p. 4422, 2023. DOI: 10.3390/ijerph20054422

[4] E. F. Rivera-Guzmán, L. F. Guerrero-Vásquez, and V. E. Robles-Bykbaev, "Quantization of Deep Neural Networks for Medical Image Analysis: A Systematic Review and Meta-Analysis," Technologies, vol. 14, no. 1, p. 76, 2026. DOI: 10.3390/technologies14010076

[5] C. Qu et al., "Post-Training Quantization for 3D Medical Image Segmentation: A Practical Study on Real Inference Engines," arXiv Prepr. arXiv2501.17343, 2025. DOI: 10.48550/arXiv.2501.17343

[6] M. S. Elhadidy, A. T. Elgohr, A. Mousa, A. Safwat, R. I. Abdelfatah, and H. M. Kasem, "Benchmarking pre-trained CNNs and vision transformers for Mpox-related dermatological image classification on MSLD v2. 0," Results Eng., p. 108071, 2025. DOI: 10.1016/j.rineng.2025.108071

[7] H. Alghoraibi et al., "Deep Learning-Based Mpox Skin Lesion Detection and Real-Time Monitoring in a Smart Healthcare System," Diagnostics, vol. 15, no. 19, p. 2505, 2025. DOI: 10.3390/diagnostics15192505

[8] T. Singh and A. Solanki, "ResEffNet: Web-Based Mpox Skin Lesion Detection System," in International Conference on Data Analytics & Management, 2025, pp. 314–325. DOI: 10.1007/978-3-032-03740-4_25

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

[10] M. A. Rahim, M. N. Arefin, M. M. Rahman, M. A. Hossain, and A. Moustafa, "An empirical study for the early detection of Mpox from skin lesion images using pretrained CNN models leveraging XAI technique,". DOI: 10.48550/arXiv.2507.15915

[11] X. Cao, W. Ye, K. Moise, and M. Coffee, "Mpoxvlm: A vision-language model for diagnosing skin lesions from mpox virus infection," arXiv Prepr. arXiv2411.10888, 2024. DOI: 10.48550/arXiv.2411.10888

[12] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 4510–4520. DOI: 10.1109/CVPR.2018.00474

[13] 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, p. 374, 2025. DOI: 10.3390/diagnostics15030374

BIOGRAPHIES OF AUTHORS

Dr. Haitham Qutaiba Ghadhban is a lecturer at the College of Engineering, University of Diyala, Iraq. He Holds a Ph.D. degree in Information Technology with specialization in Artificial Intelligent. His research areas are deep learing, machine learning, computer vision, network technology. He has published several scientific papers in national and international conferences and journals. He can be contacted at email: haithamqutaiba@uodiyala.edu.iq.

Dr. Dina Fitria Murad is a senior lecturer and researcher in the field of Information Systems. She is a faculty member at the Department of Information Systems- BINUS Online Learning, Bina Nusantara University, Jakarta, Indonesia. She actively contributes to academic development, accreditation, digital learning innovation and is an auditor, assessor, and reviewer in several reputable journals. She is also involved in research focusing on the utilization of technology and AI for online learning (e-learning), information system development, learning analytics, artificial intelligence in education, and user experience with h-index 13. Currently, Dr. Dina supervises and examiner for undergraduate and doctoral students (including external examiners of doctoral students outside Indonesia). Guest Lecturer related to digital transformation, e-learning, and information system innovation. Her interest is in community empowerment through technology, Community service for regional MSMEs, particularly by integrating AI, data analytics, and information systems to create smarter education and future-ready digital careers.