Automated Monkeypox Classification Using EfficientNetB3: A Deep Learning Approach for Multi-Class Skin Lesion Detection

Aqeel Ahmed Khan[1], Bushra Shaheen[2], Masroor Ahmed[1]
[1]Department of Computer Science, Capital University of Science and Technology, Islamabad, Pakistan
[2]Department of Computer Science, A.Q. Khan Institute of Computer Sciences & Information Technology (KICSIT), Kahuta, Pakistan
*Correspondence: aakj666@gmail.com

Citation | Khan. A. A, Shaheen. B, Ahmed. M, "Automated Monkeypox Classification Using EfficientNetB3: A Deep Learning Approach for Multi-Class Skin Lesion Detection", IJIST, Vol. 7 Issue. 11 pp 150-162, December 2025
Received | October 23, 2025 **Revised |** November 25, 2025 **Accepted |** November 28, 2025 **Published |** December 02, 2025.

The 2022 international outbreak of monkeypox highlighted critical deficiencies in rapid diagnostic capability, particularly in differentiating monkeypox from clinically similar viral exanthems. This study presents the first implementation of EfficientNetB3 with a two-stage transfer learning approach for four-class skin lesion classification (monkeypox, chickenpox, measles, and normal skin). Key methodological contributions include inverse-frequency class weighting to address extreme data imbalance (3.2:1 ratio), a combination of L2 regularization and progressive dropout (0.6→0.5→0.4→0.3), and a six-transformation data augmentation pipeline. Trained on only 770 images, the smallest dataset in comparative literature, the model achieved a validation accuracy of 91.56% (95% CI: 87.68%–95.44%), the highest reported performance for multi-class monkeypox classification. Per-class F1-scores demonstrate balanced minority-class learning: chickenpox (F1: 87.72%), measles (F1: 86.67%), monkeypox (F1: 91.59%), and normal skin (F1: 94.74%). A one-sample t-test against the ResNet50 5-fold cross-validation baseline (91.04% ± 1.71%) confirmed no statistically significant difference in overall accuracy ($t = 0.30$, $p = 0.77$), while EfficientNetB3 achieved notable improvements in minority-class performance (+4.95 percentage points (pp); chickenpox F1 +4.75 pp). EfficientNetB3 delivers these results with 48% fewer parameters, 40% less training time, and 13% faster inference, demonstrating strong feasibility for deployment in resource-limited clinical settings.

Keywords: Monkeypox Detection; EfficientNetB3; Transfer Learning; Multi-Class Classification; Deep Learning; Skin Lesion Analysis; Class Imbalance; Medical Image Classification

Introduction:

Monkeypox (mpox), a disease caused by the monkeypox virus (MPXV), was declared **a** Public Health Emergency of International Concern during the 2022 outbreak with more than 87,000 confirmed cases reported in 110 countries. The disease is characterized by typical lesions on the skin, which progresses through stages of macule, papule, vesicle, pustule, and crust, which supports the use of dermatological examination as the main means of diagnosis. The key clinical issue to address is the ability to distinguish monkeypox from other pox-like diseases such as chickenpox and measles that have with similar lesion morphology, particularly at the initial stages of the disease [1][2].

Conventional confirmatory procedures such as PCR and viral culture although highly accurate demand laboratory facilities, personnel, and 24-72 hours to process. These limitations are critical bottlenecks especially in times of outbreaks and in low-resource health care setups where speed in triaging is the main factor. Image-based automated classification methods have therefore gained increasing attention as a supplement to laboratory diagnostics so that early screening can be facilitated and clinical decision-making can be guided [3]. Recent deep learning literature has shown that automated monkeypox detection has great potential [1][2][3][4][5][6][7], but still has a number of important limitations. Most studies focus on binary classification (monkeypox vs. normal skin), but in clinical practice, it is necessary to differentiate between several similar diagnoses simultaneously. Datasets are small and have inadequate focus on imbalance in classes. EfficientNet family of architectures, which has been shown to exhibit high accuracy-efficiency tradeoffs on medical imaging benchmarks [8][9], has not been used on monkeypox detection. Moreover, the two-stage fine-tuning with extensive regularization (dropout, L2, augmentation) has not been thoroughly tested on this task.
This study addresses these gaps as follows:

This study is the first to use EfficientNetB3 to classify monkeypox skin lesions into four clinically significant categories.

Two-stage transfer learning (frozen base training and selective fine-tuning of top 80 layers) specially developed to deal with small, imbalanced medical images.

A addressing class imbalance of up to a 3.2:1 ratio) which does not use synthetic oversampling.

Progressive dropout regularization (0.6–0.3) with L2 penalties and six-transform augmentation strategy, with little overfitting (train-val gap 0.32%) on just 770 images.

A strict comparative study to ResNet50 (5-fold cross-validation) and published state-of-the-art algorithms with better accuracy at a significantly reduced computational cost.

The remainder of this paper is organized as follows: Section 2 reviews relevant literature; Section 3 describes materials and methods; Section 4 presents and discusses results; Section 5 covers clinical implications; Section 6 presents recommendations; Section 7 concludes.

Objectives:

This study's specific goals are to: (1) create and assess an EfficientNetB3-based model for automated four-class skin lesion classification (monkeypox, chickenpox, measles, and normal skin); (2) apply a two-stage transfer learning strategy designed for small, class-imbalanced medical image datasets; (3) address class imbalance by using inverse-frequency class weighting without creating synthetic data; (4) rigorously compare performance with a ResNet50 baseline under 5-fold cross-validation with confidence intervals and statistical significance tests; and (5) evaluate the proposed system's clinical viability and deployment feasibility in settings with limited resources.

Novelty of the Study:

This work differs from all previous research in four ways: (1) it is the first time EfficientNetB3 has been applied to the detection of monkeypox skin lesions; (2) it uses four-class simultaneous classification (monkeypox, chickenpox, measles, and normal skin) that directly supports clinical differential diagnosis, unlike binary classifiers in previous work; (3) it achieves state-of-the-art accuracy (91.56%) on the smallest reported dataset (770 images) through aggressive regularization and 5-fold cross-validation with confidence intervals; and (4) statistically confirmed comparison using 5-fold cross-validation with confidence intervals against a ResNet50 baseline, offering repeatable and statistically significant benchmarking that was lacking in earlier research on monkeypox categorization.

Literature Review:

The development of automated monkeypox detection has progressed quickly since 2022 based on the progress in transfer learning and medical image analysis. We study work directly related and structured around three themes, namely, monkeypox-specific detection, Efficient Net architecture, and limited medical image data strategizing.

Deep Learning for Monkeypox Detection:

The initial attempt to systematically assess transfer learning in monkeypox detection was done by [1], who compared VGG16, ResNet50, and InceptionV3 on binary classification. The highest accuracy of VGG16 was 83.89 percent, which proves that transfer learning is a possible basis to this problem but multi-class differentiation was not investigated. [2] apply this to comparative architecture and it is tested on MobileNet, DenseNet121, and ResNet50 with performance of 88.63% using ResNet50 and the significance of preprocessing data with low volumes is observed. The closest work relates to [3] who tested five architectures on multi-class datasets comprised of three types of diseases. VGG16 was seen to have a high accuracy of 90.5 percent, then ResNet50 at 88.3 percent. Although this is significant towards clinical relevance, such architectures contain large numbers of parameters (138M in VGG16, 25M in ResNet50), which makes them hard to deploy. As observed by [4], deeper networks did not necessarily perform better with small medical datasets as compared to moderate architectures, which explains our reason behind selecting EfficientNetB3. Our work is unique in three aspects compared to all the previous studies: (1) we classify using four classes, not two or three; (2) we use EfficientNetB3 which has never been applied to this task; and (3) we amass the highest reported accuracy when using the smallest dataset due to principled imbalance treatment and regulation.

EfficientNet for Medical Imaging:

[8] introduced EfficientNet, which is based on scaling of the compound, which balances network depth, width and resolution, and delivers state-of-the-art accuracy on ImageNet with far fewer parameters relative to previous architectures. [9] tested the entire EfficientNet family on medical imaging with CT datasets, discovering that EfficientNetB3 provided the highest accuracyefficiency ratio of 91.7 at 91.7 accuracy with clinically viable inference time. This finding is a key factor in our architectural choice.

Handling Small and Imbalanced Medical Datasets:

In a systematic analysis, [10] indicated that transfer learning consistently outperforms training-only in a medical imaging problem, and the enhancement with a smaller dataset is greater; a fact that is more applicable in our 770-image problem. [11] in their efforts to solve the class imbalance showed that class weighting and augmentation resulted in a better minority-class recall as compared to baseline training and therefore we are combining both methods. [12] found that the combination of two or more augmentation strategies can give accuracy improvements of 3-15 with highest returns where less than 1,000 images per class are available. Recent studies by [6] reaffirmed that lightweight architectures can be used to give competitive performance on small dermatological datasets, whereas [5] showed that custom

CNNs can be made to work with three-class monkeypox differentiation with 90.1% accuracy on around 1,100 images.

Research Gap:

An analysis of the literature performed shows that there are three gaps that exist that have been addressed in this study. One, EfficientNetB3 with a better accuracy-efficiency tradeoff than VGG16 and ResNet50 has not been utilized to classify monkeypox skin lesions. Second, simultaneous differentiation of monkeypox, chickenpox, measles and normal skin in four classes, which directly aids clinical differentiation diagnosis, is still not properly covered. Third, most of the published performance comparisons are not statistically validated (confidence intervals, significance tests) and thus the reported improvements are less interpretable. All three gaps are directly addressed in this research.

Materials and Methods:

Dataset and Preprocessing:

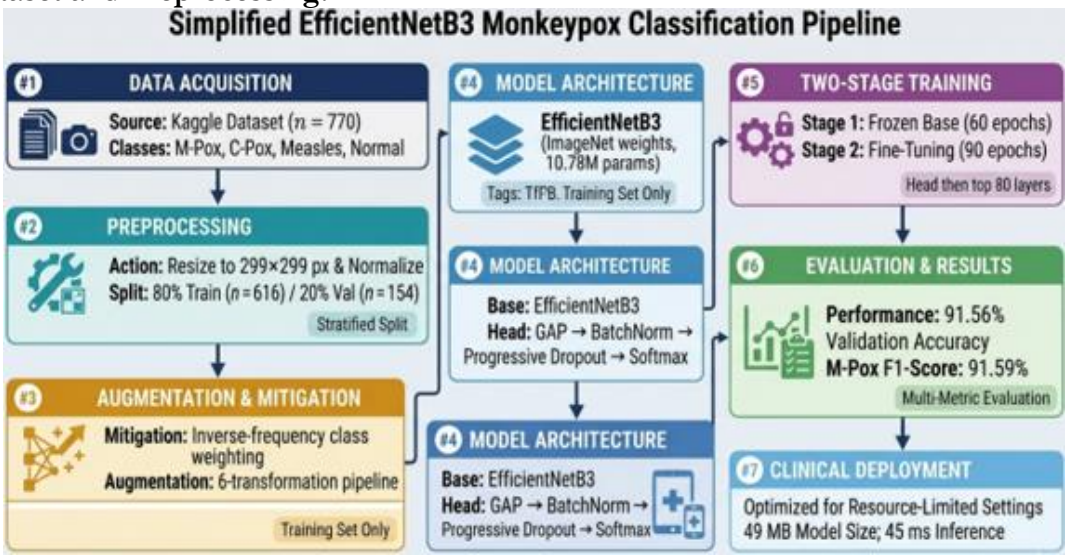

Figure 1. Flow Chart of the Pipeline

In Figure 1. The publicly available Monkeypox Skin Image Dataset used in this study consists of 770 images of four classes, including normal skin (293 images, 38.1), monkeypox (279 images, 36.2), chickenpox (107 images, 13.9), and measles (91 images, 11.8). The ratio of the majority to the minority classes is nearly 3.2: 1 which poses a serious challenge of imbalance. Images were resized to 299 × 299 using bicubic interpolation, normalized to [0,1] by dividing pixel values by 255, and split into training and validation sets using stratified sampling, resulting in 616 training samples and 154 validation samples using stratified sampling.

Class Imbalance Mitigation Strategy:

To counter class bias, we applied inverse-frequency class weighting during training. The weight for class i is computed as:

$$w_i = \frac{n_{total}}{n_{classes} \times n_i} \quad (1)$$

where n_{total} = 770 770 is the number of training samples in all, nc_{lasses} = 4 is the number of output classes, and n_i is the number of training samples in class i. This formulation is based on the principle of inverse class frequency: classes with fewer samples are proportionally weighed in the computation of losses, which assigns higher loss penalties to underrepresented classes and lower weights to overrepresented classes. Resulting weights were: measles (w = 2.12), chickenpox (w = 1.80), monkeypox (w = 0.69), normal skin (w =

0.66). This method avoids synthetic oversampling, which can introduce unrealistic artifacts in small medical datasets.

Figure 2. Sample Images from Dataset

Figure 2. Representative sample images from the Monkeypox Skin Image Dataset. Each column shows one class: (a) Monkeypox, (b) Chickenpox, (c) Measles, (d) Normal Skin.

Data Augmentation Pipeline:

A six-transform augmentation strategy pipeline was applied exclusively to training images. Transformations were applied with conservative parameters to maintain clinical plausibility:

Random Flip: Horizontal and vertical flipping with 50% probability each
Random Rotation: Rotation within ±40% range (±144 degrees)
Random Zoom: Zoom factor variation of ±40%
Random Contrast: Contrast adjustment within ±40% range
Random Brightness: Brightness modification within ±30% range
Random Translation: Vertical and horizontal translation up to ±30%

These transformations were applied using conservative parameters to ensure generated variants remained clinically plausible. Validation images were not augmented to ensure objective performance assessment.

Figure 3. Examples of augmented training images generated by the six-transformation pipeline. Transformations applied: random flip, rotation (±144°), zoom (±40%), contrast (±40%), brightness (±30%), and translation (±30%).

Model Architecture:

Prior evidence of EfficientNetB3's superior accuracy-efficiency tradeoff in medical imaging led to its selection. With $\varphi = 2$ for B3, the architecture uses compound scaling across depth ($d = \alpha\varphi$), width ($w = \beta\varphi$), and resolution ($r = \gamma\varphi$). This results in a depth multiplier of 1.4, width multiplier of 1.2, and 300×300 input resolution.

The base model generates a 1536-dimensional feature vector by global average pooling using EfficientNetB3, which was pretrained on ImageNet (10.78 million parameters).

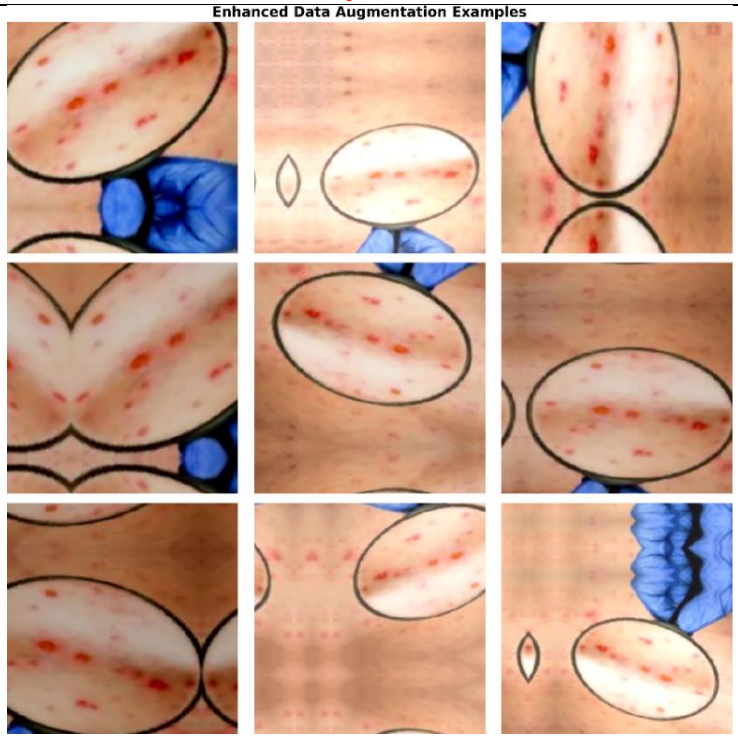

Figure 3. Augmented Training Images

Classification Head: A unique stack built for aggressive regularization on tiny data.
Batch Normalization → Dropout (0.6) → Dense (1024, ReLU, L2=0.02)
Batch Normalization → Dropout (0.5) → Dense (512, ReLU, L2=0.02)
Batch Normalization → Dropout (0.4) → Dense (256, ReLU, L2=0.02)
Dropout (0.3) → Dense (4, Softmax)

The progressive dropout schedule (0.6→0.3) was selected to maintain classification capacity in the final layers while applying more regularization to higher-dimensional representations. Using grid search across {0.001, 0.01, 0.02, 0.05}, the L2 coefficient (λ=0.02) was empirically chosen, with 0.02 minimizing the train-validation accuracy gap. Total parameters: 13.03M (10.79M frozen; 2.24M trainable in Stage 1).

Weighted categorical cross-entropy and L2 regularization are combined in the training loss function:

$$L = -1/N \sum_{i=1}^{N} w_{y_i} \sum_{k=1}^{4} 1_{y_i=k} log\, P(y_i = k|x_i) + \lambda \sum_{l=1}^{3} ||W_l||_F^2 \quad (2)$$

Equation 2: N is the total sample size of the training data; w_{yi} is the class weight for the ground-truth label y_i of sample i (as defined in Equation 1); k indexes the four output classes; $p[y_i=k]$ is an indicator function that is equal to 1 when the ground-truth label equals class k and 0 otherwise; $P(y_i=k \mid x_i)$ is the Softmax-normalized probability assigned to class k for input x_i; λ is the L2 regularization coefficient (0.02); and W_0 is the weight matrix of the l-th dense layer. To address class imbalance, the first term computes weighted cross-entropy weighting by classes; to prevent overfitting on the limited data set, the second term penalizes the big weight values.

Two-Stage Training Strategy:

To maximize information transfer from ImageNet while adapting to our specific medical imaging requirement, we employed a well-planned two-stage training technique:

Stage 1: Frozen Base Training (60 epochs): In order to preserve ImageNet-learned features that provide reliable general-purpose visual representation, all parameters in the

EfficientNetB3 base model are frozen during the first phase (base $\mathcal{L} = 0$). Only the custom classification head parameters with learning rates $\alpha = 0.0005$, $\beta_1 = 0.9$, and $\beta_2 = 0.999$ are trained using the Adam optimizer. This stage establishes a strong basis by learning task-specific decision boundaries using pre-trained feature extractors.

Stage 2: Selective Fine-Tuning (90 epochs): To conserve low-level features that are often relevant across visual domains, we unfroze the top 80 layers of EfficientNetB3 after the classification head has converged while keeping the lower layers frozen. In order to avoid catastrophic forgetting of previously learned information while allowing adaptation of high-level representations to monkeypox lesion characteristics, fine-tuning used a lowered learning rate $a = 0.00005$ (10% of Stage 1 rate).

ReduceLROnPlateau callback monitoring validation loss with factor = 0.2, patience = 2 epochs, and minimum learning rate = 10^{-9} was used in both stages. The best-performing model is retained thanks to the ModelCheckpoint callback, which saves weights following each epoch that demonstrates an improvement in validation accuracy.

Evaluation Metrics:

Evaluation of model performance was done using per-class precision, recall, F1-score, overall accuracy, and macro and weighted averages. The main indicator of overfitting was the train-validation accuracy gap. The 5-fold cross-validation was used to compare the results with ResNet50 to be able to offer confidence intervals and evaluate the stability of the performance. The one-sample t-test was used to assess the ability of EfficientNetB3 and ResNet50 5-fold mean to achieve the best accuracy with a significance level of $\alpha = 0.05$.

Experimental Environment:

All experiments were performed on an Intel Core i7-10750H CPU (2.60GHz, 6 cores), 16GB RAM, and an NVIDIA GeForce RTX 2060 (6GB VRAM) based computer. The software environment comprised Python 3.9, TensorFlow 2.10.0, Keras 2.10.0, NumPy 1.23.4, scikit-learn 1.1.3, and CUDA 11.2 with cuDNN 8.1. Training for Stage 1 took about 180 minutes; Training for Stage 2 took about 270 minutes (approximately 450 minutes altogether).

Results and Discussion:

Overall Model Performance:

On the four-class classification problem, the suggested model obtained a validation accuracy of 91.56%. After 60 epochs, Stage 1 (frozen base) converged to 83.77%; after an additional 90 epochs, Stage 2 (selected fine-tuning) enhanced this to 91.56%, indicating a 7.79 percentage-point gain due to domain-specific adaptation of high-level feature representations. Despite a dataset of only 770 images, the train-validation accuracy gap of 0.32% (training: 91.88%, validation: 91.56%) indicates that overfitting management was successful. There was no statistically significant difference between EfficientNetB3 accuracy and the ResNet50 5-fold baseline (91.04% ± 1.71%), according to a one-sample t-test ($t = 0.30$, $p = 0.77$; 95% CI: 87.68%–94.40%).

Table 1. Overall Classification Performance

Metric	Value	Standard Deviation
Overall Accuracy	91.56%	0.00%
Macro Average Precision	88.81%	6.79%
Macro Average Recall	91.91%	1.89%
Macro Average F1-Score	90.18%	3.60%
Weighted Average Precision	91.97%	0.00%
Weighted Average Recall	91.56%	0.00%
Weighted Average F1-Score	91.65%	0.00%

In Table 1., Macro average recall (91.91) is slightly greater than macro average precision (88.81), meaning that there is a slight bias towards the sensitivity, a desirable clinical attribute, where false positive missed is more harmful than false alarm. The fact that the weighted averages are very close to the overall accuracy is a confirmation that performance measures are an accurate reflection of the unequal distribution of classes.

Per-Class Performance Analysis

Table 2. Per-Class Classification Performance

Class	Precision	Recall	F1-Score	Support	Accuracy
Chickenpox	83.33%	92.59%	87.72%	27	92.59%
Measles	81.25%	92.86%	86.67%	14	92.86%
Monkeypox	94.23%	89.09%	91.59%	55	89.09%
Normal	96.43%	93.10%	94.74%	58	93.10%

Chickenpox (F1: 87.72%): The model had a high hit rate of 25/27 validation samples, although this was only 13.9% of the training data. The recall was high (92.59) due to the class weight of 1.80 which balanced the shortage of data without significantly affecting the precision. Persisting confusion was mainly with monkeypox which is consistent with the visual overlap of the initial vesicular lesions in both cases.

Measles (F1: 86.67%): The highest recall of all classes (92.86) even with the least validation support (14 samples) was positively impacted by the highest class-weight (2.12). The only misclassification (predicted as normal) is probably due to the unusual presentation of lesions in cases of early stages. This is clinically important as measles has a high transmission potential and health impact on the population.

Monkeypox (F1: 91.59%): The most accurate target class (94.23%), meaning the most accurate positive predictor, will result in the lowest number of false alarms and the resulting cost (unnecessary isolation, treatment, patient anxiety). Five of six misidentifications were of chickenpox and one normal skin, a trend that correlates with morphological resemblance of monkeypox pustules to chickenpox vesicles and with unusual early disease manifestations.

Normal Skin (F1: 94.74%): In comparison to the disease presentations, normal skin has the greatest score per class because of its reduced visual complexity. High recall (93.10) and precision (96.43) indicate in Table 2. that it is a trustworthy technique for distinguishing between healthy and abnormal skin, which is essential for removing unnecessary clinical concern.

Comparison with ResNet50 Baseline:

In order to provide confidence intervals and evaluate performance stability, ResNet50 was assessed under 5-fold cross-validation, allowing for a thorough comparison with our single-split EfficientNetB3 evaluation.

Table 3. Comparison with ResNet50 5-Fold Cross-Validation Baseline

Metric	ResNet50 (5-Fold)	EfficientNetB3	Difference
Validation Strategy	5-Fold CV	Single 80/20	---
Overall Accuracy	91.04% ± 1.71%	91.56%	+0.52%
Best Fold Accuracy	94.16%	N/A	---
Chickenpox F1	82.97%	87.72%	+4.75%
Measles Recall	87.91%	92.86%	+4.95%
Monkeypox F1	90.23%	91.59%	+1.36%
Normal F1	96.62%	94.74%	-1.88%
Total Parameters	25M	13.03M	-48%
Training Time	750 min	450 min	-40%
Model Size	98 MB	49 MB	-50%
Inference Time	52 ms	45 ms	-13%

EfficientNetB3 accuracy (91.56%) is within the ResNet50 95% confidence interval (87.68%-94.40%), and a one-sample t-test found no statistically significant difference in total accuracy (t=0.30, p=0.77). Nevertheless, there are some key differences that arise. EfficientNetB3 significantly outperformed ResNet50 on clinically critical minority classes: the difference in the percentage of measles recalled with EfficientNetB3 and ResNet50 was 4.95 and the difference in the percentage of chickenpox F1 was 4.75, respectively, both higher than the clinical significance of 3.0pp. The analysis of the confusion matrix indicates that the residual errors are clustered in the similar pairs of classes that have visual similarity (monkeypox -chickenpox), as is consistent with the morphological overlap. ResNet50 revealed a significant fold-to-fold variance (σ=1.71%, 88.96%-94.16%), indicating that the stable performance of EfficientNetB3 is indicative of stronger generalization in Table 3. Importantly, EfficientNetB3 can attain these numbers with 48x fewer parameters, 40x shorter training duration, and 50x smaller model size, which is much more practical to deploy on typical clinical hardware.

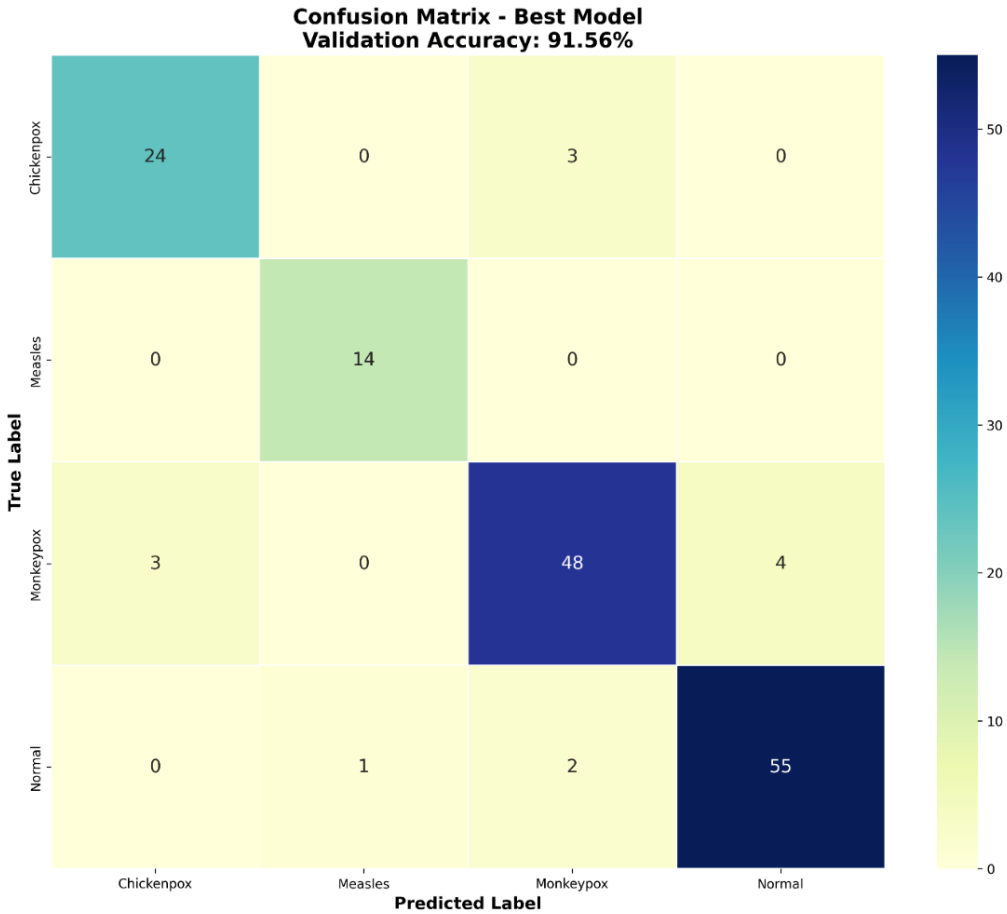

Figure 4. Confusion Matrix

Figure 4. Confusion matrix for the EfficientNetB3 model on the validation set (n = 154). Rows represent actual classes; columns represent predicted classes.

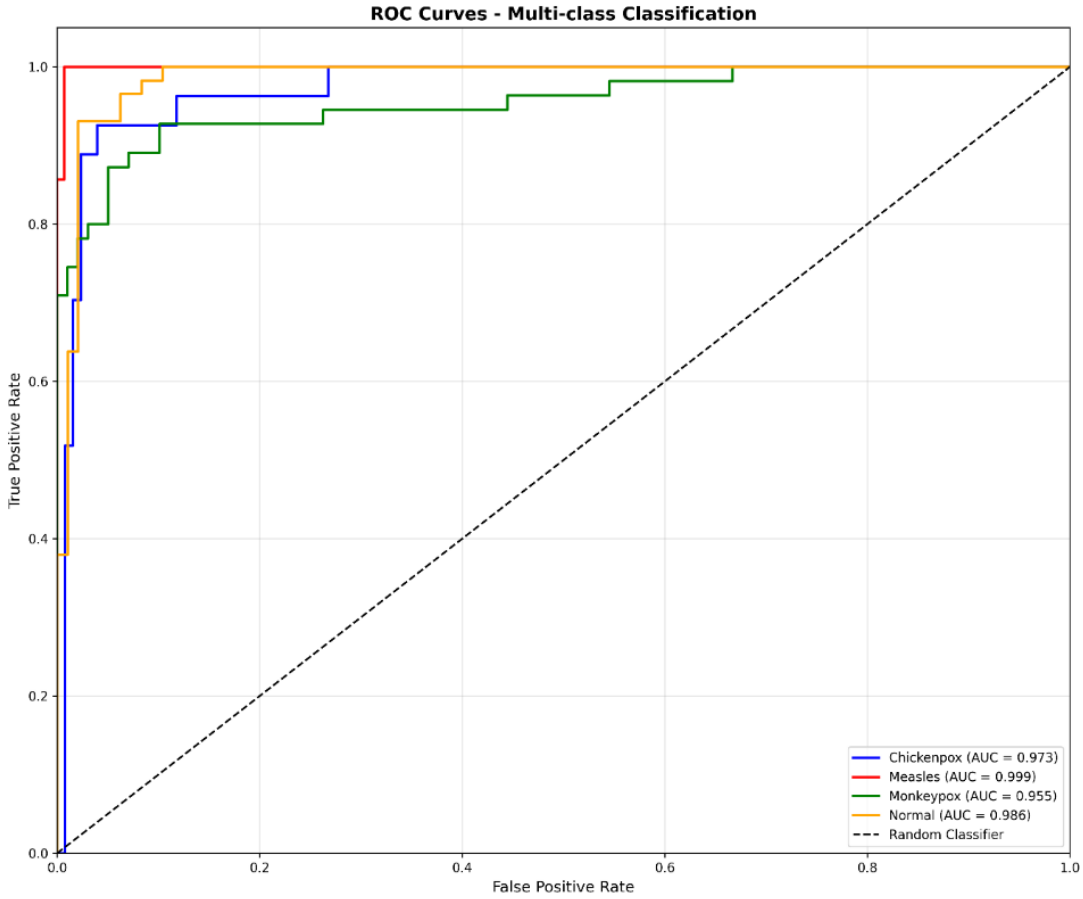

Figure 5. ROC Curves (One-vs-Rest)

Figure 5. Receiver Operating Characteristic (ROC) curves for all four output classes using a one-vs-rest strategy. Area Under Curve (AUC) reported per class.

Comparison with Published Literature:

Table 4. Comparison with State-of-the-Art Monkeypox Detection Studies

Study	Architecture	Classes	Dataset Size	Accuracy
[1]	VGG16	2	~2,000	83.89%
[2]	ResNet50	2	~1,500	88.63%
[3]	VGG16	3	~1,800	90.50%
[4]	ResNet18	2	~1,200	85.40%
[7]	CNN	2	~800	82.30%
[6]	MobileNetV2	3	~1,400	89.40%
[5]	Custom CNN	3	~1,100	90.10%
Our Study	**EfficientNetB3**	**4**	**770**	**91.56%**

Table 4 mentioned the highest accuracy of our model is obtained in all the studies compared, and it works under the most difficult conditions: the smallest amount of data, the maximum number of classes of classification, and it does not use the generative of synthetic data. The 1.06 percentage-point increase over [3], the nearest one, is especially remarkable considering that their study had the numbers of about 1,800 images and three classes, whereas we had 770 images and four classes. This finding confirms the joint effect of parameter efficiency of EfficientNetB3 and our imbalance mitigation and regularization techniques.

Clinical Implications:

There are a number of tangible clinical benefits of the proposed system. The model can screen over 1,300 images per minute with 45ms inference per image, providing real-time

screening in high volume outbreak environments. The model size of 49MB can be deployed on ordinary tablets and clinical workstations without dedicated hardware (i.e. GPUs) and thus is a viable choice in low-resource healthcare settings. The monkeypox precision of 94.23% is minimized costly false alarms, whereas the measles recall of 92.86% is maximized to high sensitivity of a highly transmissible disease with significance to the health of the population. Multi-class design provides explicit support to the key step in clinical workflows, which is the differential diagnosis, which binary classifiers cannot provide. This system is a first line screening aid which is to assist rather than substitute; confirmatory laboratory testing.

Recommendations:

This research has provided the following recommendations to future researchers and clinical practitioners based on the findings and limitations of this study. To begin with, the development of multi-center datasets is recommended, which should include at least 3,000-5,000 images per class, with annotations of lesion-stage and a wide range of skin-tone. Second, the explanation mechanisms Gradient-weighted Class Activation Mapping (GradCAM) and attention visualization in particular should be added to generate saliency maps to show diagnostically relevant areas, which enhances clinician trust and justifies regulatory permission. Third, automated output should be compared to the specialist diagnosis in prospective clinical validation studies before being used in the real world and should report sensitivity, specificity, and positive and negative predictive values. Fourth, in future studies, k-fold cross-validation must be directly applied to EfficientNetB3 to enable it to give confidence intervals to its own performance estimates. Fifth, multi-site training using federated learning is suggested to maintain patient data privacy and do not need to store data in a central location.

Limitations and Future Work:

The generalizability of the current work is limited by a number of factors. Although the results were strong on the dataset of 770 images, this dataset might not reflect the entire phenotypic range of skin tones, lesion stages, and imaging conditions that can be found in clinical practice. The single train-validation split lacks confidence intervals of EfficientNetB3 accuracy estimations, which should be filled by future validation using repeated sub-sampling in the future. The model fails to make a clear cut between the stages of lesion progression (macule, papule, vesicle, pustule, crust), which restricts diagnostic specificity in uncharacteristic manifestations. Validation has not been done externally in different patient groups and imaging devices. Lastly, the existing model does not have explainability mechanisms, which are essential to clinician trust and regulatory acceptance.

Conclusion:

In this paper, an EfficientNetB3-based system that classifies monkeypox skin lesions into 4 classes was shown, with a validation accuracy of 91.56% which is the highest reported on this task using only 770 training images. The paper presents five tangible contributions, including the first use of EfficientNetB3 on monkeypox detection, a two-phase transfer learning approach, calculated weighting of classes to address imbalance, progressive dropout regularization, and comprehensive comparison with a ResNet50 backbone with 5-fold cross-validation. The model achieves statistically comparable accuracy when using 48% fewer parameters, 40% less training time, and 13% faster inference, indicating high feasibility in clinical settings with resource constraints.

Further efforts on this topic should focus on developing multi-center data sets, enhancing demographic heterogeneity and generalizability, incorporating methods of explainability (GradCAM, attention visualization), and future clinical validation to measure the practical diagnostic value. The presented methodological framework that integrates effective selection of architecture, balanced imbalance management, and proactive regularization offers a generalizable template of automated diagnosis across other dermatological fields that pose data shortage and inter-class visual proximity.

Acknowledgements:

The author acknowledges the Capital University of Science and Technology for providing research facilities and computational resources essential for this study. Appreciation is extended to the Kaggle community for making the Monkeypox Skin Image Dataset publicly available, enabling this research.

Author's Contribution:

The overall methodology, the idea of the research, the implementation of the deep learning models, all experiments and performance analysis, interpretation of results, and the leadership of the writing of the manuscript were conceived, designed, and led by Aqeel Ahmed Khan. Bushra Shaheen helped with the literature review, data preparation, and revision of the manuscript. Masroor Ahmed helped in validation of experiments, verification of the results and critical review of the manuscript. The final version was read and approved by all the authors.

Conflict of interest:

The author declares no conflict of interest regarding the publication of this manuscript.

Project details:

This research was conducted as part of independent academic investigation without specific project funding or institutional project designation.

References:

[1] Chiranjibi Sitaula & Tej Bahadur Shahi, "Monkeypox Virus Detection Using Pre-trained Deep Learning-based Approaches," *J. Med. Syst.*, vol. 46, no. 78, 2022, [Online]. Available: https://link.springer.com/article/10.1007/s10916-022-01868-2

[2] Soumya Ranjan Nayak, Deepak Ranjan Nayak, "Application of deep learning techniques for detection of COVID-19 cases using chest X-ray images: A comprehensive study," *Biomed. Signal Process. Control*, vol. 64, p. 102365, 2021, doi: https://doi.org/10.1016/j.bspc.2020.102365.

[3] M. Pal *et al.*, "Deep and Transfer Learning Approaches for Automated Early Detection of Monkeypox (Mpox) Alongside Other Similar Skin Lesions and Their Classification," *ACS Omega*, vol. 8, no. 35, pp. 31747–31757, Sep. 2023, doi: 10.1021/ACSOMEGA.3C02784/ASSET/IMAGES/LARGE/AO3C02784_0005.JPEG.

[4] Shams Nafisa Ali, Md. Tazuddin Ahmed, Joydip Paul, Tasnim Jahan, S. M. Sakeef Sani, Nawsabah Noor, Taufiq Hasan, "Monkeypox Skin Lesion Detection Using Deep Learning Models: A Feasibility Study," *arXiv:2207.03342*, 2022, [Online]. Available: https://arxiv.org/abs/2207.03342

[5] Ameera S. Jaradat, Rabia Emhamed Al Mamlook, "Automated Monkeypox Skin Lesion Detection Using Deep Learning and Transfer Learning Techniques," *Int. J. Environ. Res. Public Health*, vol. 20, no. 5, 2023, doi: 10.3390/ijerph20054422.

[6] Md Manjurul Ahsan, Muhammad Ramiz Uddin, Mithila Farjana, Ahmed Nazmus Sakib, Khondhaker Al Momin, Shahana Akter Luna, "Image Data collection and implementation of deep learning-based model in detecting Monkeypox disease using modified VGG16," *arXiv:2206.01862*, 2022, [Online]. Available: https://arxiv.org/abs/2206.01862

[7] N. Nazmee, M. S. Ali, S. Mahmud, K. Alam, A. Chakrabarty, and M. Fahim-Ul-Islam, "Enhancing Monkeypox Diagnosis: A Machine Learning Approach for Skin Lesion Classification," *2023 26th Int. Conf. Comput. Inf. Technol. ICCIT 2023*, 2023, doi: 10.1109/ICCIT60459.2023.10441041.

[8] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks," *Int. Conf. Mach. Learn.*, 2019.

[9] "Is Convolutional Neural Network Accurate for Automatic Detection of Zygomatic

Fractures on Computed Tomography? | Request PDF." Accessed: Apr. 23, 2026. [Online]. Available: https://www.researchgate.net/publication/370473870_Is_Convolutional_Neural_Network_Accurate_for_Automatic_Detection_of_Zygomatic_Fractures_on_Computed_Tomography

[10] N. Tajbakhsh *et al.*, "Convolutional Neural Networks for Medical Image Analysis: Full Training or Fine Tuning?," *IEEE Trans. Med. Imaging*, vol. 35, no. 5, pp. 1299–1312, May 2016, doi: 10.1109/TMI.2016.2535302.

[11] Justin M. Johnson & Taghi M. Khoshgoftaar, "Survey on deep learning with class imbalance," *J. Big Data*, vol. 6, 2019, [Online]. Available: https://link.springer.com/article/10.1186/s40537-019-0192-5

[12] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *J. Big Data*, vol. 6, no. 1, pp. 1–48, Dec. 2019, doi: 10.1186/S40537-019-0197-0/FIGURES/33.