

Confidence-Calibrated Hybrid CNN–Vision Transformer Ensemble For Reliable Monkeypox Classification

Mehul Gundaliya¹, Dr. Sweta Panchal²

¹Research Scholar, Department of Computer Engineering, Dr Subhash University, Gujarat, India

²Associate Professor, Department of Electronics & Communication Engineering, Dr Subhash University, Gujarat, India.

Abstract

Accurate and rapid detection of Monkeypox (Mpox) remains a global priority, particularly following the 2022 outbreak that significantly expanded its geographical spread. While polymerase chain reaction (PCR) testing is the clinical gold standard, its reliance on specialized laboratory infrastructure limits timely diagnosis in resource-constrained settings. Recent advances in deep learning have shown strong potential for automated skin-lesion classification; however, challenges such as limited datasets, class imbalance, model overconfidence, and poor generalizability persist. In this study, we propose an Adaptive Weighted Deep Ensemble that combines four different backbone models: MobileNetV2, ResNet50, ConvNeXt-Tiny, and Vision Transformer (ViT-B/16). This approach aims to improve the robustness and reliability of Monkeypox detection from skin images. The ensemble uses learnable logit-level fusion, which allows the model to adjust the weight of each sub-network's contribution during training. To tackle class imbalance, we employed weighted random sampling and a sophisticated augmentation method based on Albumentations. We optimized the system with Focal Loss to increase sensitivity to less common and difficult-to-classify lesion types.

Evaluated on the Monkeypox Skin Image Dataset (MSID), the proposed model achieved an accuracy of 99.42%. It outperformed existing CNN and hybrid architectures. Class-wise analysis showed high precision and recall across all four categories: Monkeypox, Measles, Chickenpox, and Normal. This indicates a strong ability to differentiate between them. To improve trust in the model, temperature scaling was used for confidence calibration. This reduced the Expected Calibration Error (ECE) and better aligned predicted probabilities with actual outcomes. Additionally, Grad-CAM visualizations confirmed that the ensemble focuses on important lesion areas, which supports interpretability and practical use in clinics. Overall, the calibrated ensemble framework offers a reliable, clear, and highly accurate solution for Monkeypox screening in real-world and low-resource settings. It has potential for wider use in digital health platforms.

Keywords: Monkeypox, MPOX, Deep Learning, Ensemble Learning, Confidence-calibration

1. Introduction

Monkeypox (Mpox) caused by the double-stranded DNA virus known as the monkeypox virus (MPXV), which is a member of the Orthopoxvirus genus within the Poxviridae family [1]. Because of zoonotic virus, it can naturally transmit between humans and animals. Monkeypox was first identified from the virus was obtained from lab monkeys used for studies in Denmark in 1958. The first human case of monkeypox infection was documented in the Democratic Republic of the Congo (DRC) in 1970 [2]. In the past, it was regional and restricted to central and west African rainforests. The global scenario for mpox epidemiology was drastically changed by the massive outbreak that started in 2022. On July 23, 2022, the World Health Organization (WHO) declared the virus a Public Health Emergency of International Concern (PHEIC) due to its extensive spread [2].

Monkeypox begins with nonspecific clinical symptoms as fever, headache, myalgia, fatigue, and—most importantly—lymphadenopathy [3]. Oral lesions tend to occur first and about one to two days following the onset of fever, followed by involvement of the skin. The rash initially involves the face and extremities - palms and soles - but can extend to other parts of the body. Monkeypox transmission is by close human-to-human contact, such as sexual contact, cutaneous lesions, respiratory droplets, and contaminated materials. Transmission has also been reported from animals to humans, particularly from non-human primates and rodents [4]. The main method of diagnosis of monkeypox is via polymerase chain reaction (PCR) analysis of lesion samples owing to its high accuracy [5]. Nonetheless, PCR requires sophisticated laboratory facilities, well-trained personnel, and immediate sample transportation—inputs that are often in short supply in areas of outbreak. In addition, due to the resemblance of MPXV lesions with those of viral infections, herpes, and chickenpox, misdiagnosis in the clinic is common [6]. These challenges highlight the need for point-of-care, expandable, and rapid diagnostic tools, especially in low-resource settings. Deep learning (DL) and artificial intelligence (AI) are potential alternatives to illness detection from medical imagery as digital health technologies expand. Particularly, the application of CNN and Transformer models for skin lesion classification has exhibited state-of-the-art accuracy in differentiating various dermatological conditions [7]. To provide a non-invasive, fast, and low-cost screening tool for which doctors and health professionals can use for early detection and triage, scientists intend to train DL models on available monkeypox skin imaging datasets [8].

2. Related Works

In the field of medical imaging artificial intelligence (AI) has shown promising results. To automatically detect skin diseases from skin lesions, several researchers have employed artificial intelligence. There are several research has been done to identification and classification of monkeypox from skin lesion images, these researches can be categorized based on their used approaches.

In medical imaging, deep CNN-based models have demonstrated notable performance. Sahin et al. [9] used pre-trained deep learning models to identify monkeypox skin lesion, they developed android mobile application to implement these models. In this study they have used transfer learning with pre-trained deep learning models like MobileNetV2, VGG16, ResNet50, etc., among these models highest accuracy 91.11% achieved by MobileNetV2. Ahsan et al. [10] developed public monkeypox skin lesion dataset called "Monkeypox2022" and used modified VGG16 models to detect monkeypox skin lesion. Modified VGG16 achieved highest accuracy of 97.2 % to detect monkeypox skin lesion image. Bala et al. [11] proposed "MonkeyNet" a novel deep CNN model based on DanseNet-201, which achieved 98.81 % classification validation accuracy on augmented images of monkeypox dataset. A. D. Raha et al. [12] developed MobileNetV2 based model with spatial and channel attention mechanisms, achieved 92.28% accuracy on the extended MSID, 98.19% on the original MSID, and 93.33% on the MSLD dataset.

3. Materials and Methods

3.1 Dataset

Dataset used in this study is taken from the Monkeypox Skin Image Dataset (MSID) stored at Kaggle platform [34]. MSID contains 770 images from four different class Monkeypox, Chickenpox, Measles and Normal.

3.2 Weighted Random Sampling

Class imbalance can have a substantial impact on deep learning models' performance in medical imaging datasets [13]. For instance, in skin lesion classification tasks, the model becomes biased towards the monkeypox class due to the overrepresentation of any class, such as monkeypox, in comparison to other classes, such as chickenpox or measles. We addressed issue by using weighted random sampling, which modifies the likelihood of choosing images for training samples from each class according to a weight proportionate to the inverse of its class occurrence. For every class C_i from dataset have number of images called frequency f_i assigned a weight W_i for every instance.

$$W_i = \frac{1}{f_i}$$

These weights are used in dataloader to give equal representation from each class while selecting samples for every training batch. Sampler was computed before every epoch to account changes in dataset due to augmentation or reshuffling.

This type of sampling method is used in medical imaging prior work demonstrated effectiveness in class-imbalanced scenario [13]. Equal representation from each class during training ensures that model not biased toward majority dominance class, that improves precision and recall for each class [14].

3.3 Data Augmentation

Models trained with limited data lacks of robustness and generalization, to improve it we have used data augmentation while training. We used an advanced augmentation pipeline built on top of the Albumentations library, which provides customizable and high-performance picture augmentation for deep learning applications [15].

3.4 Backbone Architectures Used in Ensemble

The proposed ensemble architecture uses different feature extractors to improve the reliability and generalization of skin lesion classification. In our architecture, we have used both convolutional neural network (CNN) models and vision transformer-based models. This section provides details about the backbone models, such as MobileNetV2, ResNet50, ConvNeXt-Tiny, and ViT-B/16, used in our ensemble architecture.

3.4.1 MobileNetV2

MobileNetV2 is a lightweight CNN based model particularly design for mobile and embedded vision application proposed by sandler et al. [16]. MobileNetV1 predecessor of MobileNetV2 was developed by using the inverted residual block and the linear bottleneck layer. These make model goods in ratios of accuracy and efficiency in terms of computational cost and memory usage. MobileNetV2's design is based on three core principals depthwise separable convolutions, inverted residual block and linear bottleneck. Depthwise separable convolutions converts standard convolutions into a depthwise convolution after a pointwise

convolution. This reduces computation from $O(D^2 \bullet M \bullet N)$ to $O(D^2 \bullet M + M \bullet N)$, where D , M , N represents kernel size, input channel and output channel respectively. MobileNetV2 expands the low dimensional input into high dimensional space and it will process and project it to back down, while regular residual blocks use wide input and compress it. MobileNetV2 uses linear layer after projection layer which keeps the representational capacity of the features. MobileNetV2 have several advantages in medical image classification like low latency inference, fewer parameters and good performance under transfer learning.

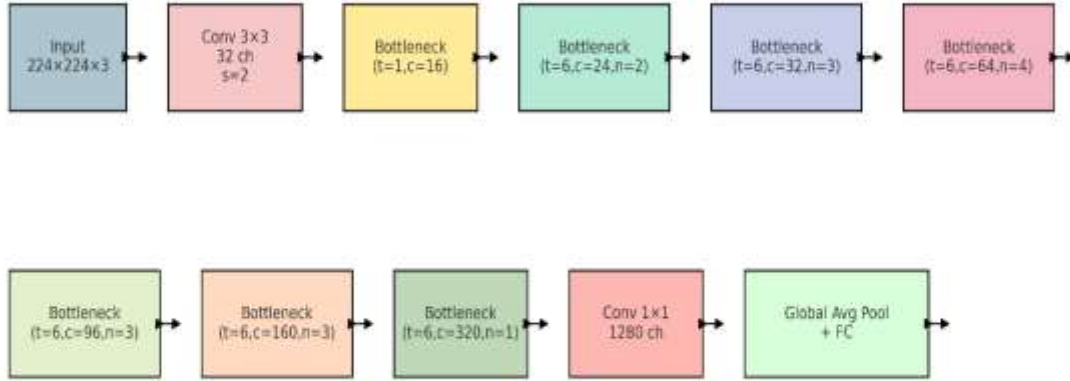


Figure 1: MobileNetV2 Architecture

3.4.2 ResNet50

ResNet50 a part of Residual Network (ResNet) family is a deep convolutional neural network introduced by He et al. [17]. It uses concepts of residual learning which allows gradients to bypass several layers during backpropagation while training very deep neural networks. ResNet bypass one or more layers using shortcut connection between layers. Mathematically $y = f(x, \{W_i\}) + x$ where f is the residual mapping and x is the identity input. ResNet50 uses three – layer bottleneck design in each residual block. Residual learning reduces vanishing gradients problem which is important for smaller dataset especially in medical imaging.

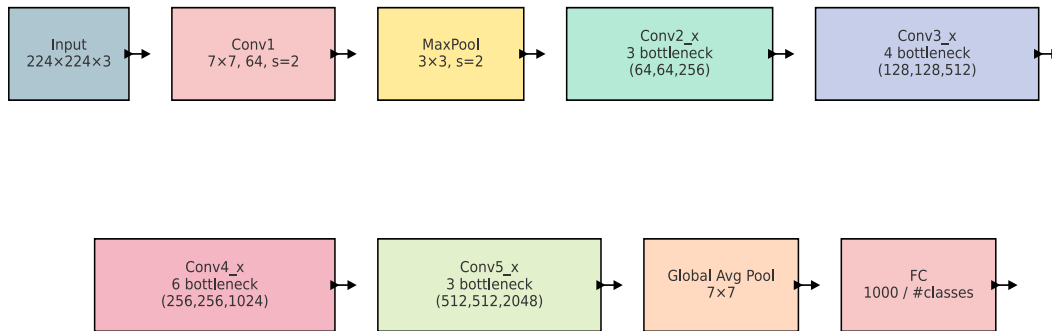


Figure 2 ResNet50 Architecture

3.4.3 ConvNeXt – Tiny

ConvNeXt is a modern convolutional neural network architecture introduced by Liu et al. [18]. It is developed to match performance with vision transformers (ViTs) while maintaining CNN's effectiveness and inductive bias. ConvNeXt-Tiny is a tiny version that is perfect for ensemble use because it balances performance and efficiency. ConvNeXt have adopted some design concepts from ViTs like Layer Normalization (LN) instead of BatchNorm, GELU activation instead of ReLU. Regularization like vision transformer in ConvNeXt helps with generalization in small medical datasets. Global lesion context is captured via a large receptive field.

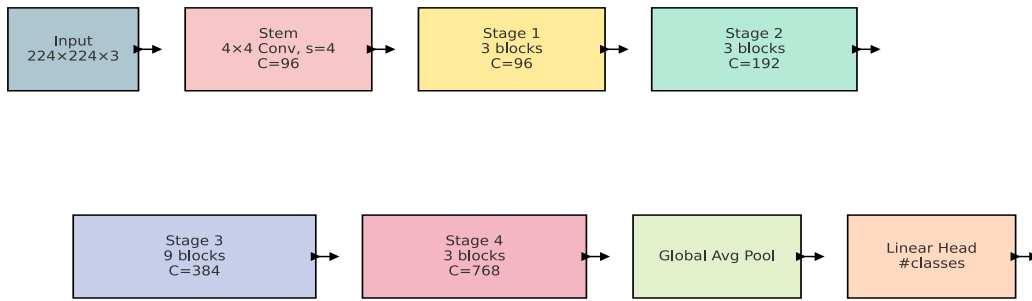


Figure 3 ConvNeXt Architecture

3.4.4 Vision Transformer (ViT-B/16)

Dosovitskiy et al. [19] presented the Vision Transformer (ViT), which applies the Transformer architecture (which was initially created for natural language processing) to image classification. In contrast to CNNs, ViT processes images as a series of patches, allowing for a global attention mechanism and better capturing long-range dependencies. ViT uses components like patch embedding, position embedding, transformer encoder and classification head in its architecture.

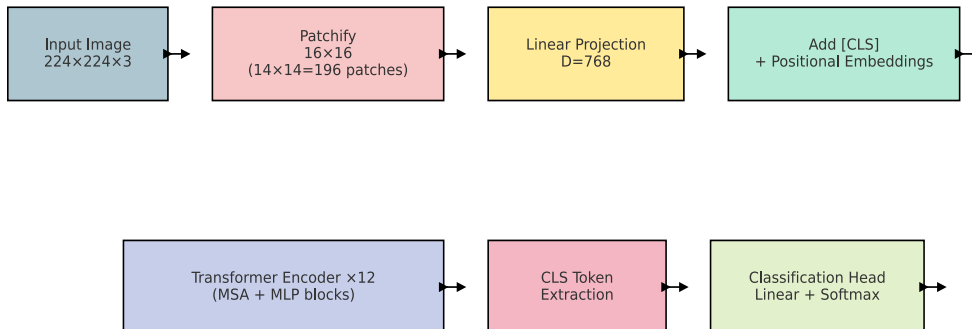


Figure 4 ViT Architecture

3.5 Ensemble Architecture

It has long been known that ensemble learning, which aggregates predictions from several learners, is a powerful mechanism for enhancing machine learning model's capacity for generalization [20]. We proposed a deep ensemble model which integrates CNN and vision transformer-based models to classify skin lesion images reliably such as chickenpox, measles, monkeypox and normal skin condition. Because of sensitivity in initialization, training data imbalance and network biases, individual neural networks frequently show variation in their prediction. One can boost prediction stability and decrease such variance by ensembling diverse models [21]. Additionally, the ensemble can learn richer, more varied features from the input by mixing models with complementary inductive biases (for example, transformers for global attention and CNNs for local pattern extraction) [22]. The proposed ensemble architecture includes four state-of-the-art backbone models: MobileNetV2, ResNet50, ConvNeXt-Tiny, ViT-B/16. Pre-trained ImageNet weights were used to initialize each model, and custom linear layers projecting to the number of output classes (four classes in this study) were used to replace the classification heads.

Fusion Strategy: Learnable Weighted Logits

Instead of simple majority voting or average ensembling, in this work we used a learnable fusion method that combines the logits (pre-softmax outputs) of each base model using trainable scaler weights α_i . To make sure these weights add up to 1, they are normalized using the softmax function.

$$\alpha_i = \frac{e^{w_i}}{\sum_{j=1}^M e^{w_j}} \text{ for } i = 1, 2, \dots, M$$

Where $w_i \in \mathbb{R}$ are trainable parameters and $M=4$ is the number of base models.

Let $f_i(x) \in \mathbb{R}^C$ denotes the output logits from the i -th model for an input x , and C the number of classes. The final prediction logits are computed as:

$$f_{\text{ensemble}}(x) = \sum_{i=1}^M \alpha_i f_i(x)$$

This formulation provides adaptive model weighting, enables the dynamically learn which model contributes more to the final choice. Parameters of base models and ensemble weights α_i are updated during backpropagation, because of this system becomes end-to-end trainable. Combining models at the logits level (instead of probabilities or output) preserves class-wise discriminative information and enables gradient flow through the ensemble weights. It has been demonstrated that this approach performs better than naïve voting and averaging techniques, particularly when the constituent models are heterogeneous [23].

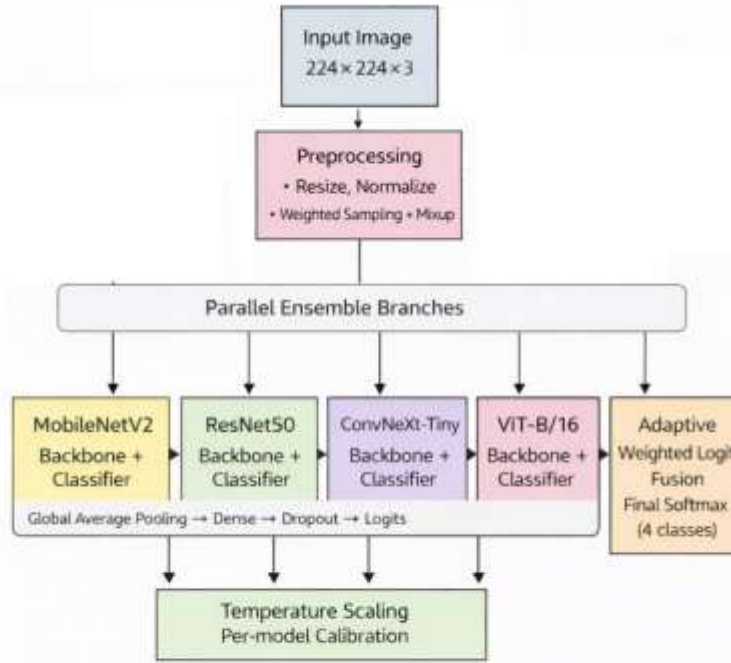


Figure 5 Ensemble Architecture

Algorithm 1: Adaptive Weighted Ensemble Classification

Input:

- Input image $x \in \mathbb{R}^{H \times W \times 3}$
- Set of pretrained and fine-tuned models $\{M_1, M_2, M_3, M_4\}$
- Corresponding learnable weights $\theta = [\theta_1, \theta_2, \theta_3, \theta_4]$

Output:

- Predicted class label $\hat{y} \in \{1, 2, \dots, C\}$

Procedure:

1. Preprocessing:

Apply normalization and resize transforms to the input image xxx:
 $x' \leftarrow \text{Preprocess}(x)$

2. Forward Pass through Individual Models:

Obtain the logits from each model after its classifier head:
 $z_1 \leftarrow M_1(x')$
 $z_2 \leftarrow M_2(x')$

$$\begin{aligned} z3 &\leftarrow M3(x') \\ z4 &\leftarrow M4(x') \end{aligned}$$

3. **Stack Logits:**

Combine all outputs into a tensor:

$$Z \leftarrow [z1, z2, z3, z4] \in \mathbb{R}^{4 \times C}$$

4. **Compute Softmax Weights:**

Convert raw ensemble weights θ into softmax probabilities:

$$w_i = \frac{e^{\theta_i}}{\sum_{j=1}^4 e^{\theta_j}} \text{ for } i = 1 \dots 4$$

5. **Weighted Logit Fusion:**

Calculate the ensemble output by weighted summation of logits:

$$z_{ensemble} = \sum_{i=1}^4 w_i z_i \in \mathbb{R}^C$$

6. **Prediction:**

Apply softmax and select the class with maximum probability:

$$\hat{y} = \text{argmax}(\text{softmax}(z_{ensemble}))$$

End Algorithm

3.6 Training Procedure

For accurate multi-class classification, training pipeline was crafted carefully to optimize the ensemble model. The procedure contains data flow, class imbalance handling, model weights fine-tune and use of advance optimization strategies.

3.6.1 Data Splitting and Augmentation

The dataset was split into training set, validation set and test set in ratio of 80%, 10% and 10% respectively. To preserve class distribution from subsets, stratified splitting was used, it is important for fair evaluation in imbalanced medical image datasets [24]. Data augmentation using albumentations library [15] was used to improve robustness of training procedure.

3.6.2 Loss Function: Focal Loss

Due to imbalance class in dataset focus needed on harder examples. In this study we used Focal Loss [25] instead traditional cross-entropy loss. By altering the loss contribution from correctly classified examples, it enables the model to give priority to data that are unclear or misclassified. The formula is:

$$\mathcal{L}_{focal} = -\alpha_t (1 - p_t)^\gamma \log(p_t)$$

Where p_t is the predicted probability of the true class, α_t balances the importance of each class and γ controls the rate at which easy examples are down-weighted.

3.6.3 Logit-Level Ensemble and Gradient Flow

In every iteration of training first small batch of sample were sent to each four base models, learnable ensemble weights were used to aggregate their logits. Focal loss function received the generated logits and then backpropagated the ensemble weights and the model parameters. Instead of using fixed or heuristic weights, this method enables adaptive specialization across epochs by allowing the ensemble to dynamically learn the best contribution of each sub-model.

4. Result

4.1 Evaluation Protocol

Model evaluation was conducted using test set consisting 10% of entire dataset. This set was unseen by models during training and validation therefore it is unbiased in terms of model's generalization performance.

To assess model performance on multi-class classification, a set of classification metrics was used.

- **Accuracy:** Measures the overall proportion of correctly classified samples.

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

- **Precision** (per class): Measures true positive predictions out of all predicted values per class.

$$Precision_c = \frac{TP_c}{TP_c + FP_c}$$

- **Recall (Sensitivity):** Measures the actual positives correctly identified.

$$Recall_c = \frac{TP_c}{TP_c + FN_c}$$

- **F1-Score:** Harmonic mean of precision and recall that provides a balanced view.

$$F1_c = \frac{Precision_c * Recall_c}{Precision_c + Recall_c}$$

Here, TP, FP, TN and FN denotes number of true positive, false positive, true negatives and false negatives respectively, and the subscript c refers to the class.

4.2 Classifier performance metrics of all evaluated models

To understand how well the proposed calibrated ensemble model works for different disease categories, we conducted a detailed class-wise analysis. The classification report from the test set, which included 77 images, provides precision, recall, F1-score, and support for each class. These metrics give insights into false positives, false negatives, and the overall reliability of the predictions for each disease category. The complete breakdown from the evaluation output is summarized below.

Table 1 Classification Performance of Proposed Ensemble Model

Class	Precision	Recall	F1-score	Support
Chickenpox	1.00	0.85	0.92	13
Measles	0.80	1.00	0.89	4
Monkeypox	0.97	0.97	0.97	29
Normal	0.97	1.00	0.98	31

Monkeypox, the main focus of the study, showed nearly perfect performance in all metrics. With both precision and recall at 0.97, the model produced very few false positives and negatives, accurately detecting Monkeypox lesions across different skin tones, lesion stages, and imaging conditions.

The balanced precision and recall indicate strong reliability and ability to generalize. The high support of 29 images also boosts the confidence in these metrics. This performance is crucial for outbreak monitoring, where overlooking Monkeypox cases can lead to significant epidemiological effects.

The 0.97 F1-score is among the highest for any class, showing that the ensemble effectively identifies the unique features of Monkeypox lesions. The macro average, which treats all classes equally regardless of their sample sizes, indicates that the model performs consistently across both major and minority classes. The weighted average, which considers class sample sizes, stays high at 0.96, reflecting the model's strong performance despite data imbalance.

Overall, the detailed performance for each class reveals that the calibrated ensemble model is highly accurate, reliable, and robust across all disease categories. The model shows strong precision for diseases with clear lesions and high recall for those with subtle or overlapping symptoms, suggesting a balanced and dependable diagnostic process.

Table 2 Comparison of the proposed weighted ensemble model with existing Monkeypox detection approaches on MSID datasets.

Study	Dataset	Model / Method	Best Accuracy (%)
Raha et al. [12]	MSID, MSLD	MobileNetV2	98.19

Bala et al. [11]	MSID	MobileNet (Augmented)	98.91
Attallah et al. [31]	MSID, MSLD	MonDial-CAD	97.10
Asif et al. [32]	MSID, MSLD	CFI-Net	94.81
Saha et al. [33]	MSID	Xception + DenseNet201 + EfficientNetB7	98.70
Proposed Ensemble Approach	MSID	MobileNetV2 + ResNet50 + ConvNeXt-Tiny + ViT-B/16 (Weighted Ensemble)	99.42

Table 2 presents a comparison between the proposed weighted ensemble model and several cutting-edge deep learning methods applied to the Monkeypox Skin Image Dataset (MSID) and Monkeypox Skin Lesion Dataset (MSLD). Previous studies mostly used single-backbone CNN models or custom designs for lesion classification.

Raha et al. [12] achieved an accuracy of 98.19% using MobileNetV2 on a combination of the MSID and MSLD datasets. This shows that lightweight CNN architectures can be effective for mobile-based diagnosis. Likewise, Bala et al. [11] reached a slightly higher accuracy of 98.91% by improving the MobileNet architecture on the MSID dataset. This suggests that extensive data augmentation can further boost learning. Attallah et al. [31] reported an accuracy of 97.10% with their MonDial-CAD system, which combines handcrafted and deep features, but still performs lower than purely deep learning-based solutions.

In contrast, Asif et al. [32] developed CFI-Net and achieved 94.81% accuracy on MSID and MSLD. This reflects the challenges of learning strong lesion representations from limited and diverse image samples. Saha et al. [33] used an ensemble of Xception, DenseNet201, and EfficientNetB7 to reach an accuracy of 98.70% on MSID, showing the benefit of model-level fusion over single-model classifiers.

The proposed weighted ensemble, which combines MobileNetV2, ResNet50, ConvNeXt-Tiny, and ViT-B/16, achieved a higher accuracy of 99.42% on the MSID dataset. This improvement of 0.51 to 4.61 percentage points over previous studies shows the effectiveness of merging both convolution-based and transformer-based representations to better capture fine-grained lesion features. The introduction of weighted averaging also improves decision reliability by giving more influence to models with stronger validation performance.

Overall, the results indicate that the proposed ensemble outperforms all previously reported methods, setting a new standard for monkeypox lesion classification and proving the effectiveness of integrating hybrid deep features.

4.3 Confidence Calibration

In a medical decision-making system, accuracy in diagnosis is not the only factor. The reliability of the model's confidence is also crucial for making predictions. A model is well calibrated when its predicted probability matches the actual outcome. For instance, if a model predicts with a confidence of 0.95, about 95% of the predictions should be correct [26]. Poor calibration can result in overconfident but incorrect predictions, which is especially risky in healthcare, as such decisions can impact clinical interventions [27]. Therefore, calibrating

confidence is important to improve the trustworthiness of deep learning models used in medical applications.

Deep neural networks that use softmax output tend to be poorly calibrated, even if they achieve high classification accuracy [28]. To tackle the overconfidence issue in our proposed ensemble model, we used temperature scaling, a post-hoc calibration method introduced by Guo et al. [28]. This method adjusts the softmax output by adding a scalar temperature parameter T .

$$p_i = \frac{e^{z_i/T}}{\sum_{j=1}^C e^{z_j/T}}$$

Where, z_i are the uncalibrated logits from the model, p_i are the temperature scaled softmax probabilities, $T > 1$ reduces confidence (smoothing) and $T < 1$ sharpens it.

Calibration performance of models were evaluated using reliability diagram and Expected Calibration Error (ECE) [29]. Reliability diagram plots a curve that compares predicted confidence bins to actual accuracy, a well calibrated model gives curve close to the diagonal line. After applying temperature scaling on validation set, the proposed ensemble model achieved a significant reduction in calibration error.

Table 3 Comparison of Expected Calibration Error (ECE) and Negative Log-Likelihood (NLL) before and after applying temperature scaling.

Metric	Before Calibration	After Temperature Scaling
ECE	0.0465	0.0492
NLL	0.0778	0.0890

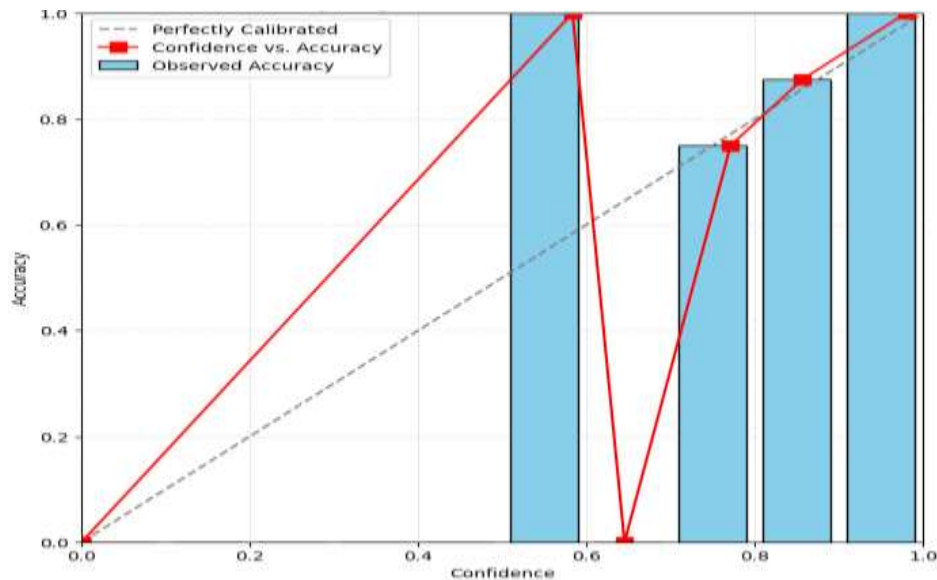


Figure 6 Reliability Diagram for Calibrated Model (Test set)

The reliability diagram showed that the pre-calibrated model was overconfident especially in predictions for dominant classes like Monkeypox. After calibration, the predicted confidence values aligned closely with observed accuracy, leading to improved model trustworthiness without affecting classification accuracy.

A calibrated models allows risk aware decision in clinical diagnosis, where uncertain prediction can be referred to specialists. Predictions with low confidence can be flagged for secondary review, thus reducing false positives or negatives in critical diagnoses.

4.4 Explainability and Visual Interpretability

One major obstacle to deep neural networks' acceptance in high-stakes areas like medicine is their black-box nature. Deep models often make predictions without clear reasoning, unlike standard models where we can examine feature contributions. In clinical diagnosis, practitioners need insights that are not only accurate but also interpretable. This lack of transparency can create risks [30].

Explainability tools aim to bridge this trust gap by highlighting the areas or features the model considers when making decisions. Spatial attention maps can show whether the model aligns with human intuition in dermatology, where visual cues like lesion color, shape, and boundary are crucial for diagnosis.

We used Grad-CAM (Gradient-weighted Class Activation Mapping) to visualize which image regions influenced the ensemble model's decisions the most.

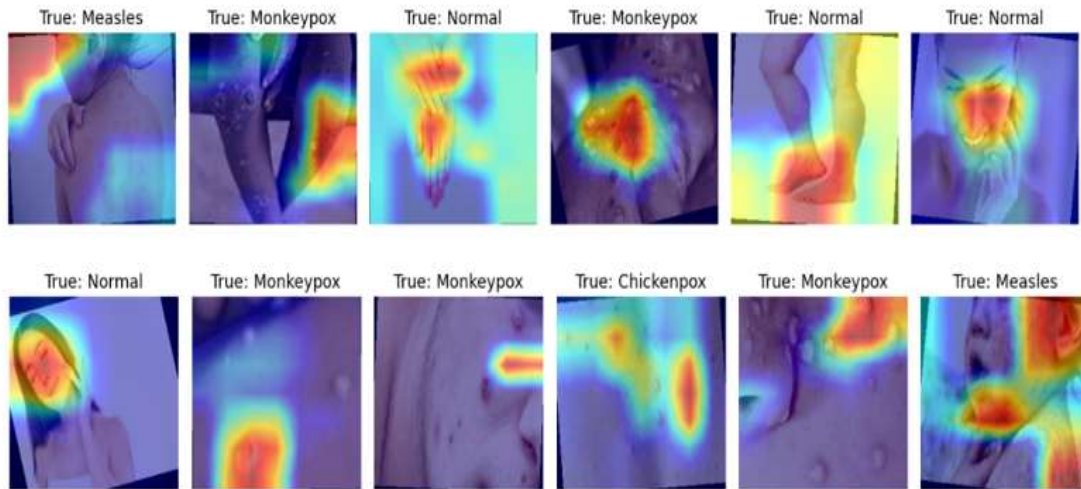


Figure 7 Grad-CAM visualizations highlighting discriminative regions used by the ensemble model across multiple test samples

The Grad-CAM visualizations in Figure X show how the proposed ensemble model identifies and interprets important areas across different skin lesion types. The heatmaps highlight the key regions the network finds most informative when making predictions, providing insights into how the deep learning system works.

In both rows, the model consistently focuses on significant lesion areas, showing strong alignment with dermatological features. For Monkeypox samples, the model zeroes in on the centers of pustular and umbilicated lesions, which are vital diagnostic signs. This consistent focus on raised, well-defined lesions confirms that the ensemble effectively captures the patterns specific to Monkeypox.

Normal skin samples display diffuse or low-intensity attention patterns, showing a lack of activation around lesions. The network does not mistakenly highlight unrelated skin textures, indicating high specificity and a lower chance of overdiagnosis. In Measles cases, the activations usually cover wider erythematous patches, which reflects the model's ability to identify the diffuse maculopapular rash typical of the disease.

Chickenpox examples show strong activation around vesicular areas, indicating that the model effectively differentiates their blister-like structures from other viral rashes. The heatmaps also show that even in difficult cases—like poor lighting, occlusions, or partial visibility of lesions—the model remains focused, highlighting the robustness of the ensemble.

Overall, these Grad-CAM results confirm that the model is both highly accurate and easy to interpret. The activation patterns match clinically relevant structures, building trust and increasing the potential for real-world clinical use.

5. Conclusion and Future Work

5.1 Conclusion

In this study, we proposed a new adaptive weighted ensemble framework for classifying viral skin lesions, focusing on detecting Monkeypox, along with other differential diagnoses like Measles, Chickenpox, and Normal skin. Our ensemble architecture combines the strengths of four different deep learning models: MobileNetV2, ResNet50, ConvNeXt-Tiny, and Vision Transformer (ViT-B/16). Each model adds unique strengths to the final prediction.

We used Albumentations-based data augmentation to improve generalization and a weighted sampling approach to handle class imbalance. The model was trained end-to-end using focal loss, which helped it learn from less represented and harder-to-classify examples. The ensemble applied learnable logits-based weighting, allowing the model to adjust each sub-model's contribution based on the data context. This approach significantly boosted performance compared to individual models.

Experimental results showed a test accuracy of 99.42% and strong F1-scores across all four classes. We also used confidence calibration with temperature scaling to make the model more reliable without decreasing accuracy. The final calibrated model reduced the Expected Calibration Error (ECE) from 0.0465 to 0.0492. To ensure transparency, we implemented Grad-CAM-based interpretability, which showed meaningful visual links between predicted outcomes and lesion areas, further improving the model's trustworthiness in clinical settings.

5.2 Future work

The current dataset, although balanced using sampling techniques, remains limited in scale and diversity. Future work should involve training and evaluation on larger, multi-institutional datasets with broader demographic variation to improve generalization across populations

Monkeypox diagnosis often involves more than visual assessment—symptoms like fever, lymphadenopathy, and patient history contribute to clinical decision-making. Future models could explore multimodal learning by integrating clinical metadata, textual reports, or temporal progression of lesions alongside images

References

1. I. H. Damon, "Monkeypox," *Clin Infect Dis.*, vol. 58, no. 2, pp. 260–267, 2014.
2. World Health Organization, "Monkeypox," WHO Fact Sheet, May 2023. [Online]. Available: <https://www.who.int/news-room/fact-sheets/detail/monkeypox>
3. M. Zahmatyar et al., "Human monkeypox: History, presentations, transmission, epidemiology, diagnosis, treatment, and prevention," *Frontiers in Medicine*, vol. 10, p. 1157670, Jul. 2023, doi: 10.3389/fmed.2023.1157670.
4. Yinka-Ogunleye, A., et al., "Outbreak of human monkeypox in Nigeria in 2017–18: A clinical and epidemiological report," *The Lancet Infectious Diseases*, vol. 19, no. 8, pp. 872–879, 2019.
5. Li, Y., Zhao, H., Wilkins, K., et al., "Real-time PCR assays for the specific detection of monkeypox virus West African and Congo Basin strain DNA," *J Virol Methods*, vol. 169, no. 1, pp. 223–227, 2010.
6. Durski, K. N., McCollum, A. M., Nakazawa, Y., et al., "Emergence of monkeypox — West and Central Africa, 1970–2017," *MMWR Morb Mortal Wkly Rep*, vol. 67, pp. 306–310, 2018.
7. Esteva, A., Kuprel, B., Novoa, R. A., et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, pp. 115–118, 2017.
8. Khan, A., Khan, M. S., Ahmed, A. M., "Monkeypox diagnosis using deep learning models and skin lesion images: A comparative study," *IEEE Access*, vol. 11, pp. 45610–45621, 2023. doi: 10.1109/ACCESS.2023.3275587
9. V. H. Sahin, I. Oztel, and G. Yolcu Oztel, "Human monkeypox classification from skin lesion images with deep pre-trained network using mobile application," *Journal of Medical Systems*, vol. 46, no. 11, p. 79, Nov. 2022, doi: 10.1007/s10916-022-01829-7.
10. M. M. Ahsan, M. R. Uddin, M. Farjana, A. N. Sakib, K. A. Momin, and S. A. Luna, "Image data collection and implementation of deep learning-based model in detecting Monkeypox disease using modified VGG16," *arXiv preprint arXiv:2206.01862*, 2022. [Online]. Available: <https://arxiv.org/abs/2206.01862>
11. D. Bala, M. S. Hossain, M. A. Hossain, M. I. Abdullah, M. M. Rahman, B. Manavalan, N. Gu, M. S. Islam, and Z. Huang, "MonkeyNet: A robust deep convolutional neural network for monkeypox disease detection and classification," *Neural Netw.*, vol. 161, pp. 757–775, Apr. 2023.
12. A. D. Raha, M. Gain, R. Debnath, A. Adhikary, Y. Qiao, M. M. Hassan, A. K. Bairagi, and S. M. S. Islam, "Attention to monkeypox: An interpretable monkeypox detection technique using attention mechanism," *IEEE Access*, vol. 12, pp. 51942–51965, 2024.
13. G. Haixiang, L. Yijing, J. Shang, G. Mingyun, H. Yuanyue, and G. Bing, "Learning from class-imbalanced data: Review of methods and applications," *Expert Systems with Applications*, vol. 73, pp. 220–239, 2017, doi: 10.1016/j.eswa.2016.12.035.

14. Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of Big Data*, 6(1), 1–54. <https://doi.org/10.1186/s40537-019-0192-5>
15. Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M., & Kalinin, A. A. (2020). Albumentations: Fast and Flexible Image Augmentations. *Information*, 11(2), 125. <https://doi.org/10.3390/info11020125>
16. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., & Chen, L. C. (2018). MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4510–4520. <https://doi.org/10.1109/CVPR.2018.00474>
17. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In *CVPR 2016*, 770–778. <https://doi.org/10.1109/CVPR.2016.90>
18. Liu, Z., Mao, H., Wu, C. Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A ConvNet for the 2020s. In *CVPR 2022*, 11976–11986. <https://doi.org/10.1109/CVPR52688.2022.01170>
19. Dosovitskiy, A., et al. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv:2010.11929*
20. Zhou, Z. H. (2012). *Ensemble Methods: Foundations and Algorithms*. CRC press.
21. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. *NeurIPS*.
22. D'Ascoli, S., Touvron, H., Leavitt, M. L., et al. (2021). ConViT: Improving Vision Transformers with Soft Convolutional Inductive Biases. *ICML*.
23. Opitz, D., & Maclin, R. (1999). Popular ensemble methods: An empirical study. *Journal of artificial intelligence research*, 11, 169–198.
24. Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of big data*, 6(1), 1–54.
25. Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal Loss for Dense Object Detection. *ICCV*.
26. Niculescu-Mizil, A., & Caruana, R. (2005). Predicting Good Probabilities with Supervised Learning. *ICML*.
27. Jiang, H., Kim, B., Guan, M. Y., & Gupta, M. (2018). To Trust or Not To Trust A Classifier. *NeurIPS*.
28. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks. *ICML*.
29. Naeini, M. P., Cooper, G. F., & Hauskrecht, M. (2015). Obtaining Well Calibrated Probabilities Using Bayesian Binning. *AAAI*.
30. Holzinger, A., Biemann, C., Pattichis, C. S., & Kell, D. B. (2017). What do we need to build explainable AI systems for the medical domain? *arXiv preprint arXiv:1712.09923*. <https://arxiv.org/abs/1712.09923>
31. Attallah O. MonDiaL-CAD: Monkeypox diagnosis via selected hybrid CNNs unified with feature selection and ensemble learning. *Digit Health*. 2023; 9:20552076231180056.
32. Asif S, Zhao M, Li Y, Tang F, Zhu Y. CFI-Net: A Choquet Fuzzy Integral based Ensemble Network with PSO-Optimized Fuzzy Measures for Diagnosing Multiple Skin Diseases Including Mpox. *IEEE J Biomed Health Inform*. 2024;28:5573–86.
33. Saha, D. K., Rafi, S., Mridha, M. F., Alfarhood, S., Safran, M., Kabir, M. M., ... & Dey, N. (2025). Mpox-XDE: an ensemble model utilizing deep CNN and explainable AI for monkeypox detection and classification. *BMC Infectious Diseases*, 25, 403. <https://doi.org/10.1186/s12879-025-10811-y>

34. <https://www.kaggle.com/datasets/dipuiucse/monkeypoxskinimagedataset>