Electronic Research Archive

ERA, 34(2): 738–776.
DOI: 10.3934/era.2026034
Received: 16 November 2025
Revised: 06 January 2026
Accepted: 14 January 2026
Published: 22 January 2026

http://www.aimspress.com/journal/ERA

Research article

Attention-enhanced hybrid deep learning framework for Monkeypox skin lesion classification

Zhonghua Zhang* and Siying Zheng

School of Science, Xi'an University of Science and Technology, Xi'an 710600, China

* **Correspondence:** Email: wwwzhangzhonghua@163.com.

Abstract: Monkeypox, a re-emerging zoonotic disease, has become a growing global public health concern due to its rapid transmission and high visual similarity to other dermatological conditions such as chickenpox and measles. This resemblance complicates early clinical diagnoses, particularly in resource-limited settings where laboratory testing capabilities are scarce. Deep learning methods capable of detecting monkeypox from skin images offer a promising alternative to manual inspection and PCR-based diagnoses. However, existing approaches are often limited by poor dataset quality, weak generalization ability, and insufficient model interpretability. To address these challenges, this paper proposes an attention-enhanced hybrid deep learning framework for the automated classification of monkeypox skin lesions. Specifically, the proposed model employs DenseNet-121 and EfficientNet-B4 as parallel convolutional feature extractors and integrates Convolutional Block Attention Modules (CBAM) to adaptively emphasize lesion-related features while suppressing background interference. Experimental results demonstrate that the proposed framework outperforms conventional deep learning baselines, thereby achieving an accuracy exceeding 91% and a Cohen's Kappa score above 0.88 on the primary dataset. Furthermore, the model exhibits a strong generalization capability, thereby maintaining classification accuracies above 90% across multiple external public datasets. To enhance the transparency and clinical reliability, explainable artificial intelligence (XAI) methods are employed to visualize the model's decision-making process. In addition, quantitative interpretability metrics are used to assess the reliability and consistency of the generated explanations, thus highlighting the model's potential for practical clinical applications.

Keywords: Monkeypox; image classification; deep learning; CBAM

1. Introduction

Monkeypox is an orthopoxvirus disease caused by the monkeypox virus, a zoonotic pathogen first isolated in 1958 from cynomolgus monkeys in Copenhagen [1]. The first confirmed human case was reported in 1970 in the Democratic Republic of the Congo [2]. Since then, human infections have steadily increased, with most early cases originating in Africa. However, the geographic distribution of monkeypox has substantially expanded in recent years. In 2003, the first human case outside Africa was reported in the United States, and in May 2022, cases were identified in the United Kingdom, followed by a widespread transmission across Europe and other regions worldwide. In response to the escalating situation, the Director-General of the World Health Organization (WHO) declared monkeypox a Public Health Emergency of International Concern (PHEIC) in July 2022 and again in August 2024, underscoring the seriousness of the global outbreak.

Clinically, monkeypox typically presents after an incubation period of 5–21 days and progresses through two main stages: a prodromal phase and an exanthematous phase. During the prodromal phase, infected individuals commonly experience fever, headache, myalgia, back pain, and lymphadenopathy, which usually persist for 1–5 days. This is followed by the exanthematous phase, which is characterized by a centrifugal rash pattern in which lesions first appear on the face and subsequently spread to the extremities, with frequent involvement of the palms and soles [3]. Human-to-human transmission primarily occurs through direct contact with lesions, respiratory droplets, or contaminated fomites.

At present, laboratory confirmation of monkeypox mainly relies on polymerase chain reaction (PCR) assays or ultrastructural examination of lesion samples using electron microscopy. Although PCR-based testing is considered the diagnostic gold standard, it requires specialized equipment, costly reagents, and trained personnel, which significantly limits its availability in resource-constrained settings such as rural or remote regions [4]. In addition, improper sample collection and inadequate biosafety protocols may increase the risk of false-negative results and occupational exposure. These limitations, together with the clinical similarity between monkeypox and other viral exanthematous diseases such as chickenpox and measles, make early and accurate diagnoses particularly challenging. Consequently, there is an urgent need for rapid, accessible, and reliable diagnostic alternatives.

Recent advances in deep learning have demonstrated a strong potential in automated medical image analysis, including applications in diabetic retinopathy grading, pulmonary nodule detection, and dermatological lesion classification [5–8]. Owing to their powerful representation learning capability, deep learning models can automatically extract discriminative visual features from skin images and differentiate monkeypox lesions from visually similar conditions. The adoption of such techniques has the potential to substantially reduce the diagnostic time, improve the screening efficiency, and minimize the exposure risk for healthcare workers, thereby complementing or partially alleviating the reliance on laboratory based diagnostic methods.

Despite recent advances, existing deep learning-based approaches for monkeypox detection still face important limitations. Many studies rely on small or noisy datasets with limited diversity, thus raising concerns about robustness and generalizability. Evaluations are often confined to binary or low-class classification settings, with insufficient validation across independent datasets. Moreover, model interpretability remains underexplored, despite its critical importance for clinical trust and real-world deployment. These issues highlight the need for deep learning solutions that are accurate, robust, generalizable, and interpretable.

To address these challenges, we propose an attention-enhanced hybrid deep learning framework

for monkeypox skin lesion classification. Rather than introducing new attention modules or backbone architectures, this work adopts a controlled hybrid design that integrates dual-backbone feature fusion with attention-based refinement, thus enabling the systematic analysis of attention contributions in static dermatological images. By combining multi-dataset generalization evaluation with quantitative and qualitative explainability analyses, the proposed framework offers practical insights into the design of reliable and interpretable models for monkeypox diagnoses. Therefore, the core contribution of this work lies not in proposing novel foundational modules, but in adopting a controlled hybrid design paradigm to systematically evaluate the value of attention mechanisms for static dermatological image classification. The main contributions of this work are summarized as follows: (i) a hybrid attention-based classification framework that enhances the focus on diagnostically relevant lesion regions and improves the robustness to visual variability; (ii) comprehensive generalization evaluation across multiple public datasets; and (iii) the integration of explainable artificial intelligence techniques to provide transparent and clinically meaningful model interpretations.

The remainder of the paper is organized as follows: Section 2 provides a review of the existing literature on monkeypox diagnoses based on deep learning; Section 3 outlines the datasets used in this study, introduces the adopted data preprocessing techniques, presents the baseline models for comparison, and elaborates on the experimental design framework, including the strategies and methods employed to construct the hybrid model; Section 4 offers a comprehensive analysis and discussion of the experimental results for both the baseline and hybrid models, along with a validation of the hybrid model's generalization capability; Section 5 examines the interpretability of the hybrid model and evaluates its decision-making process using quantitative metrics; Section 6 discusses the main findings of the study; and finally, Section 7 concludes the paper and suggests directions for future research.

2. Literature review

The rapid global spread of monkeypox has motivated extensive research on automated and accurate diagnostic approaches using artificial intelligence and computer vision. With the increasing availability of labeled skin lesion images and public datasets such as the Monkeypox Skin Image Dataset (MSID) and the Monkeypox Skin Lesion Dataset (MSLD), deep learning-based image classification has become the predominant research paradigm. Existing studies can be broadly categorized according to the evolution of methodological strategies, from early CNN-based models to recent hybrid, attention-enhanced, and explainable frameworks.

In the early stage of monkeypox image classification research, most studies relied on individual convolutional neural network (CNN) architectures, which are often combined with transfer learning. Ahsan and Uddin et al. [9] proposed a Generalized and Regularized Transfer Learning Approach (GRA-TLA) and evaluated ten pre-trained CNN models for both binary and multi-class monkeypox classification. Their results showed that the Xception model achieved accuracies of 77% (multi-class) and 88% (binary), while ResNet-101 achieved the highest accuracy range of 84–99%. The authors further employed Local Interpretable Model-agnostic Explanations (LIME) to interpret prediction results. However, the use of self-collected datasets introduced variability in the image quality and illumination, thus potentially limiting model generalization.

With the release of public datasets, comparative benchmarking studies became more common. Almufareh et al. [10] evaluated InceptionV3, ResNet50-V2, MobileNetV2, and EfficientNet-B4 on the MSID and MSLD datasets. On MSID, MobileNetV2 achieved a balanced accuracy of 96.55%,

with a 93% specificity and a 100% sensitivity, while on MSLD, InceptionV3 achieved a 94% balanced accuracy, 100% specificity, and 88% sensitivity. These studies confirmed the effectiveness of CNN-based transfer learning while highlighting dataset-dependent performance variability.

To overcome the limitations of single model approaches, subsequent studies introduced hyperparameter optimization and ensemble learning. Altun et al. [11] proposed a hybrid framework that combined transfer learning with hyperparameter optimization on MSID and MSLD, which achieved an average F1-score of 98%, AUC of 99%, accuracy of 96%, and recall of 97% using an optimized MobileNetV3-s model. Dahiya et al. [12] applied multiple optimization strategies, including Bayesian optimization and Learning without Forgetting (LwF), and reported a maximum accuracy of 98.18% on a Roboflow dataset containing 971 images.

Ensemble learning approaches further improved the robustness by aggregating complementary CNN representations. Sitaula and Shahi [13] systematically evaluated thirteen pre-trained CNN architectures and showed that an ensemble of Xception and DenseNet-169 achieved an 87.13% accuracy, 85.44% precision, 85.47% recall, and 85.40% F1-score on the MSID dataset. Despite the improved stability, ensemble pipelines often increased the computational cost and still struggled to capture subtle lesion-level features.

Recent studies have focused on enhancing the feature discrimination through attention mechanisms and hybrid architectures. Uysal [14] integrated CNN feature extraction with long short-term memory (LSTM) and achieved an 87% accuracy with a Cohen's Kappa score of 0.8222 on MSID, which indicated modest performance gains at the expense of increased computational complexity. Raha et al. [15] improved the MpoxNet framework by introducing spatial and channel attention mechanisms and applying Grad-CAM and LIME for the interpretability analysis. The enhanced MpoxNet achieved a 98.19% accuracy on the original MSID, 92.28% on an extended MSID, and 93.33% on MSLD. Similarly, Sun et al. [16] proposed a lightweight ConvNeXt-based MpoxNet that incorporated a Convolutional Block Attention Module (CBAM), which achieved a 95.28% accuracy, 96.40% precision, 93.00% recall, and 95.80% F1-score, while significantly reducing model parameters.

Beyond CNN-centric designs, transformer-based and explainable hybrid frameworks have recently gained attention for their ability to capture global contextual information and improve robustness. Khan and Iqbal [17] proposed RS-FME-SwinT, a feature map enhancement framework that integrated a customized Swin Transformer with residual and spatial CNN blocks. Their model achieved a 97.80% accuracy, 96.82% sensitivity, 98.06% precision, and 97.44% F1-score, and demonstrated a strong discrimination capability for monkeypox skin images. More recently, Deng et al. [18] proposed a deep learning screening framework that integrated multi-scale feature learning and attention mechanisms for efficient monkeypox screening. Their model reported a 98.12% accuracy, 98.34% F1-score, 98.34% precision, 96.17% MCC, and 99.83% AUC, and indicated a strong reliability and suitability for large-scale screening scenarios. Explainable nature-inspired hybrid models have also been explored. Shateri et al. [19] combined Xception-based feature extraction with NGBoost classification, which was optimized using the African vultures optimization algorithm (AVOA). Their framework achieved a 97.53% accuracy, 97.72% F1-score, and 97.47% AUC, while employing LIME and Grad-CAM to enhance the interpretability. In parallel, Hossain et al. [20] introduced a deep learning framework informed by a systematic survey of existing monkeypox diagnostic methods and evaluated the robustness across multiple datasets. Their approach consistently achieved over a 95% accuracy and incorporated SHapley Additive exPlanations (SHAP)-based explainability to analyze model behavior across different skin tones and anatomical regions, thus

highlighting the importance of robustness and fairness in real-world deployment.

Despite the promising performance reported by recent studies that employed attention mechanisms, transformer-based architectures, or ensemble strategies for monkeypox and skin lesion classifications, several limitations remain insufficiently addressed. Many studies insufficiently addressed the class imbalance, which led to biased predictions toward majority classes [21]. Existing attention-based models, such as those that incorporated CBAM or related modules, primarily focused on improving the classification accuracy within a single dataset, while systematic evaluations under controlled experimental settings were often lacking. Transformer-based approaches demonstrated strong representation learning capabilities but typically involved a higher computational complexity and limited analyses of generalization across heterogeneous datasets. Ensemble methods further improved the robustness but introduced an increased model complexity and a reduced interpretability.

In contrast to these approaches, the present study emphasizes a controlled hybrid design that integrates complementary convolutional features with lightweight gated aggregation, thus ensuring fair comparisons across model variants. Moreover, the proposed framework is evaluated on multiple public datasets to explicitly assess cross-dataset generalization. Finally, quantitative interpretability analyses are incorporated to complement the performance evaluation, thus providing more reliable insights into the model decision behavior beyond qualitative visualizations alone.

3. Materials and methodology

3.1. Dataset sources

This study utilized the MSID dataset. It is an open-source collection available on the Kaggle platform [22] and is comprised of a total of 770 images, which are categorized as follows: 293 normal images, 279 images of monkeypox, 91 images of measles, and 107 images of varicella. The distribution of them is as illustrated in Figure 1.

It is necessary to emphasize that all images in this dataset are sourced from open-source and shared datasets, and these images are subject to relatively permissive licensing terms. None of the images contain any personally identifiable information, and no human subjects were involved in the data collection process. Therefore, there is no need to conduct an ethical review when using this dataset.

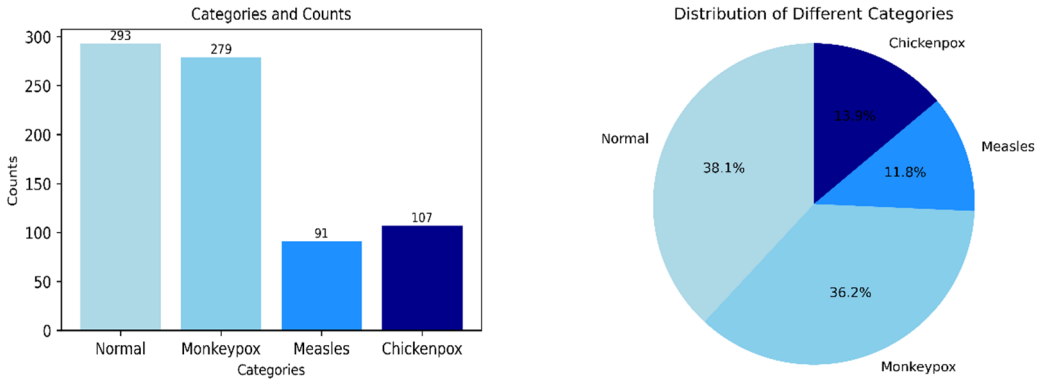


Figure 1. Distribution of monkeypox skin image dataset.

3.2. Dataset preprocessing

The MSID dataset exhibits a significant class imbalance across its four diagnostic categories: Normal, Monkeypox, Measles, and Chickenpox, comprising 770 skin images in total. Balancing this distribution is critical for effective neural network training [23]. To establish a reliable evaluation framework, a meticulous preprocessing pipeline was implemented prior to any model training.

Dataset splitting was performed on the original images to prevent augmentation artefacts from leaking into the evaluation subsets. For each class, 80 original images were randomly and exclusively allocated to form the validation and test sets, with 40 images assigned to each. All remaining images for a class constituted the initial training pool. Following this initial split, a dedicated deduplication step was conducted. As MSID is a web-scraped collection that lacks patient identifiers, we employed perceptual hashing (pHash, with a Hamming distance threshold of ≤ 5) and a structural similarity analysis (SSIM ≥ 0.95) to identify near-duplicate images across the subsets. This analysis revealed 8 cross-subset duplicate pairs, all within the Monkeypox category, comprising 4 train-test, 2 val-test, and 2 train-val pairs. To eliminate data leakage while preserving the integrity of the test set, a conservative deduplication strategy was applied. For any pair of near duplicate images identified across different data splits, the duplicate was removed from the earlier-phase set, with priority given to keeping the test set intact. Consequently, six Monkeypox images were removed from the training set and two images were removed from the validation set, while the test set remained unchanged.

To address persistent class imbalance and to enhance the model's generalization, data augmentation was exclusively applied to the final deduplicated training set. By employing techniques including horizontal flipping, random rotation, and color jittering, the augmentation intensity was calibrated per class to yield a perfectly balanced training set with 320 samples per category. As detailed in Table 1, the complete preprocessing workflow, final dataset composition, and augmentation specifications resulted in a robust 8:1:1 split for training, validation, and testing, respectively. This rigorous protocol ensures the independence of evaluation subsets while laying a solid foundation for model development by improving the class balance and robustness.

Table 1. Detailed distribution of the MSID dataset before and after data augmentation.

Class	Original images	Training	Validation	Test	Augmented images (training only)	Final training samples
Normal	293	213	40	40	+107	320
Monkeypox	279	199→193	40→38	40	+127	320
Measles	91	11	40	40	+309	320
Chickenpox	107	27	40	40	+293	320
Total	**770**	**444**	**158**	**160**	**+836**	**1280**

The specific parameters for data augmentation are provided in Table 2. The comparison images before and after augmentation are presented in Figures 2 and 3, respectively.

Table 2. Data enhancement parameter values.

Data enhancement type	Data enhancement parameter values
equalize	P = 0.5
horizontal_flip	P = 0.6
random_rotation	P = 0.5 angle_limit = 30
scaling	P = 0.5 scale_range: (0.8, 1.2)
brightness_contrast	P = 0.5 brightness_limit = 0.1 contrast_limit = 0.5
hue_saturation_value	P = 0.5 hue_shift_limit = 20 sat_shift_limit = 30 val_shift_limit = 20
shift_scale_rotate	P = 0.7 shift_limit = 0.1 scale_limit = 0.05 rotate_limit = 60
rgb_shift	P = 0.2 r_shift_limit = 5 g_shift_limit = 5 b_shift_limit = 5

Figure 2. Monkeypox skin image data set before data enhancement.

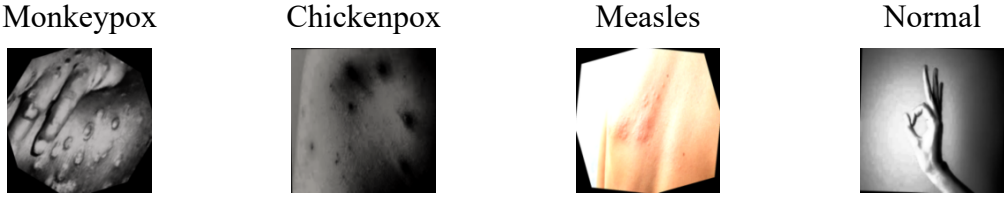


Figure 3. Monkeypox skin image data set after data enhancement.

To improve the efficiency of the learning algorithm, we applied min-max normalization to the dataset to scale the RGB pixel values in a unified range. Let x be the original pixel value in each color channel. The general equation [24] for min-max normalization is as follows:

$$x' = \frac{x - min(x)}{max(x) - min(x)}, \tag{1}$$

where x' is the normalized value.

For an 8-bit RGB image, the pixel values for each color channel range from 0 to 255, and Eq (1) simplifies to the following:

$$x' = \frac{x}{255}. \tag{2}$$

3.3. Basic methods

CNNs are deep learning architectures dedicated to processing image data. Their core structure includes convolutional layers, pooling layers, and fully connected layers. Convolutional layers extract local features of the image through local receptive fields and weight sharing. Pooling layers are used to reduce the spatial dimensions of feature maps, thus decreasing the computational load. Fully connected layers integrate these features and output classification results. The basic CNN architecture is illustrated in Figure 4.

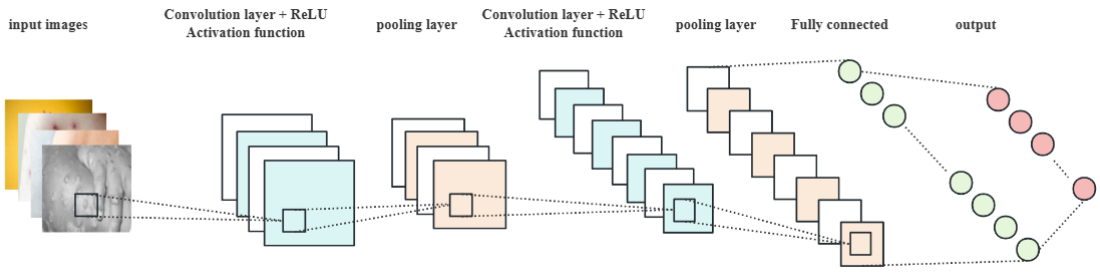


Figure 4. Basic convolutional neural network architecture diagram.

Building upon the foundational CNN architecture, researchers have developed diverse modified models to enhance performance, increase efficiency, or address specific task demands. For this study, we selected eight baseline models, namely VGG19, MobileNetV3-Large, ResNet-50, ResNet-101, InceptionV3, Xception, EfficientNet-B4, and DenseNet-121, all of which demonstrate strong performances in image recognition and classification tasks.

In the following, the above eight baseline models were employed to identify and classify the monkeypox skin image dataset. Based on a series of evaluation metrics, the top two models that exhibited the best performances were selected. To further enhance the models' performance, each of the two models were integrated with a GRU and a CBAM, respectively, to construct hybrid models aimed at achieving better classification accuracies. Furthermore, Grad-CAM and LIME techniques were employed to gain a deeper understanding of the decision-making processes of the selected deep learning models. These methods provide valuable insights into the internal mechanisms of the models, enhance interpretability, and offer robust diagnostic reasoning capabilities.

A brief introduction to the models is as follows.

3.3.1. VGG19

The Visual Geometry Group (VGG) network, a deep convolutional neural network model proposed by the VGG at the University of Oxford, achieved outstanding results in the 2014 ImageNet Large Scale Visual Recognition Challenge (ILSVRC). Among the various VGG architecture variants, VGG19 [25] is a notable representative. It consists of 16 convolutional layers, 5 max-pooling layers, 3 fully connected layers, and a final Softmax layer for classification. In this study, we utilized the pre-trained VGG19 model, which was pre-trained on the ImageNet dataset, to transfer its shallow features to the MSID dataset. Additionally, we fine-tuned the convolutional layers to adapt the model to the new dataset and generate a customized feature map.

3.3.2. MobileNetV3-Large

MobileNetV3, which is built upon MobileNet and MobileNetV2, is the third generation of lightweight deep learning architecture. It optimizes the network structure using neural architecture search (NAS) techniques to reduce computational resource consumption while maintaining a high accuracy. MobileNetV3-Large [26] is a variant of this architecture, designed to balance a low computational cost with a high accuracy.

3.3.3. ResNet

Residual Network (ResNet), proposed by He et al. [27] in 2015, introduces a residual learning mechanism that enables direct signal propagation across layers via skip (or shortcut) connections. This innovation effectively addresses gradient vanishing and network degradation in deep neural network training. The ResNet framework can scale to very deep architectures, such as ResNet-50, ResNet-101, and ResNet-152, which have shown superior performances in various image recognition tasks, particularly on large-scale datasets such as ImageNet and COCO.

3.3.4. InceptionV3

InceptionV3, which was introduced by Szegedy et al. [28] in 2015, is the third iteration of Google's Inception models, also referred to as GoogLeNet V3. It is a deep CNN specifically designed for image recognition tasks, particularly in large-scale visual recognition challenges. By optimizing the architecture of the Inception module, it reduces both the number of parameters and computational overhead, thus enhancing the model’s computational efficiency.

3.3.5. Xception

Xception was proposed by Chollet [29] in 2017. Its core innovation is the introduction of depthwise separable convolution, which decomposes the standard convolution operation into two independent steps: depthwise convolution and pointwise convolution. Xception achieved strong performances on multiple benchmark datasets, particularly in large-scale image classification tasks such as ImageNet. It efficiently extracts multi-scale image features through its depthwise separable convolution layers while maintaining a low computational complexity.

3.3.6. EfficientNet-B4

As a part of the EfficientNet family, the EfficientNet-B4 model was introduced by Tan and Le [30] in 2019. It enhances both the efficiency and accuracy by systematically scaling the network's depth, width, and input image resolution using a composite scaling method. Additionally, EfficientNet-B4 incorporates an automatic data augmentation strategy during training that dynamically searches for the optimal augmentation approach.

3.3.7. DenseNet-121

The DenseNet-121 model was introduced by Gao et al. [31] in 2017. It is a member of the DenseNet family, with '121' representing the number of layers in the network. The key feature of DenseNet-121 is its dense connectivity, where each layer directly connects to all preceding layers. This property facilitates feature reuse and allows for the extraction of more comprehensive and representative features.

3.3.8. Gated recurrent unit (GRU)

The gated recurrent unit (GRU), introduced by Cho et al. in 2014 [32], is a lightweight variant of recurrent neural networks originally designed to model sequential data. In the present study, the MSID dataset consists of independent still images without longitudinal patient follow-up or temporal ordering. As a result, the GRU is not used to model temporal disease progression or lesion evolution.

Instead, the GRU is employed as a gated feature aggregation module. The fused deep feature vector extracted from the convolutional backbones is reshaped into a fixed-length pseudo sequence of feature tokens and processed by the GRU to capture structured dependencies among these tokens. Through its gating mechanisms, the GRU selectively integrates discriminative feature components while suppressing redundant responses.

Compared with long short-term memory networks, the GRU provides a more compact architecture with fewer parameters and lower computational cost, which is advantageous for relatively small-scale medical image datasets. The GRU incorporates two gating mechanisms. The reset gate modulates the influence of previous hidden states on the current input, while the update gate controls the balance between retained information and newly incorporated features. Through these mechanisms, the GRU enables adaptive feature integration rather than explicit temporal modeling within the proposed framework. The specific formulas are as follows:

1) reset gate: $r_t = \sigma(W_{rx}x_t + U_r h_{t-1} + b_r)$,
2) update gate: $z_t = \sigma(W_{zx}x_t + U_z h_{t-1} + b_z)$,
3) candidate hidden state: $\hat{h}_t = \tanh(W_{hx}x_t + U_h(r_t \odot \mathrm{h}_{\mathrm{t\text{-}1}}) + b_h)$,
4) updated hidden state: $h_t = z_t \odot \mathrm{h}_{\mathrm{t\text{-}1}} + (1 - z_t) \odot \hat{\mathrm{h}}_{\mathrm{t}}$,

where σ represents the sigmoid function and is used to control the information flow valued between 0 and 1, the tanh function is used to convert the data into the candidate hidden state ranging in [-1, 1], $\odot$ is the Hadamard product, and W, U, b are learnable parameters [32].

3.3.9. Convolutional block attention module (CBAM)

CBAM is a typical attention mechanism designed to enhance the feature representation capability of CNNs. This mechanism sequentially integrates channel attention modules and spatial attention modules, thus enabling the network to adaptively focus on significant channels and discriminative spatial regions in the input feature maps. The primary reason for selecting CBAM in this study is its ability to perform adaptive weight allocation on both the channel dimension and the spatial dimension of feature maps. This effectively highlights lesion-related key fine-grained features while suppressing the interference from irrelevant background information. It prompts the model to focus more precisely on regions of diagnostic value, thereby enhancing the model's ability to differentiate similar diseases. The architecture of CBAM is illustrated in Figure 5.

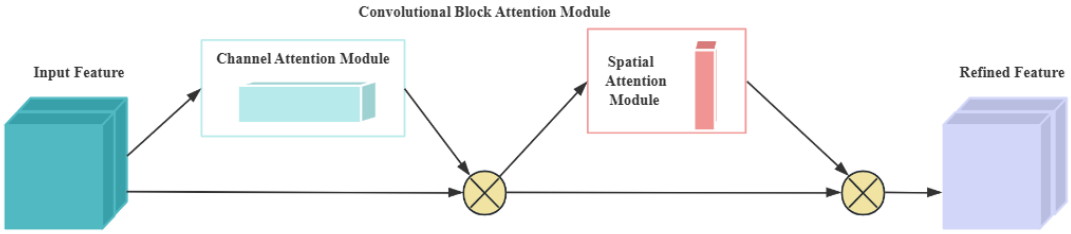


Figure 5. CBAM structure diagram.

The CBAM module is described in detail below.

1) Channel attention module (CAM)

The purpose of the channel attention module is to illustrate which channels are more important. Given an input feature map $\mathrm{F} \in \mathrm{R}^{C\times H\times W}$, the module outputs a channel attention map $\mathrm{M}_c \in \mathrm{R}^{C\times 1\times 1}$. The calculation process is as follows:

(i) Global pooling: Apply global average pooling (AvgPool) and global max pooling (MaxPool) to F and produce aggregated feature descriptors F_{avg} and F_{max}, respecgively;

(ii) Feature transformation: Process both F_{avg} and F_{max} through a shared multilayer perceptron (MLP) to generate transformed features $\mathrm{M}_{c,avg}$ and $\mathrm{M}_{c,max}$, respectively;

(iii) Attention generation: $\mathrm{M}_{c,avg}$ and $\mathrm{M}_{c,max}$ are added together, and the final channel attention map M_c is obtained by the sigmoid activation function.

The channel attention module is formally defined as follows:

$$\mathrm{M}_c(\mathrm{F}) = \sigma(\mathrm{MLP}(\mathrm{AvgPool}(\mathrm{F})) + \mathrm{MLP}(\mathrm{MaxPool}(\mathrm{F}))), \tag{3}$$

where σ denotes the sigmoid activation function, and MLP represents the shared multilayer perceptron [33].

2) Spatial attention module (SAM)

The purpose of the spatial attention module is to identify which spatial positions in the feature map are important. Given the channel-attention-processed feature map F', the module outputs a spatial attention map $\mathrm{M}_s \in \mathrm{R}^{1\times H\times W}$. The computation proceeds as follows:

(i) Channel-wise pooling: perform the average pooling and maximum pooling operations of the channel dimension on F' to obtain F'_{avg} and F'_{max}, respectively;

(ii) Feature transformation: Concatenate F'_{avg} and F'_{max}, then pass them through a convolutional layer $f_{7\times7}$ with a 7 × 7 kernel;

(iii) Activation: Generate the final spatial attention map M_s via a sigmoid activation function.

The spatial attention module is formally defined as follows:

$$M_s(F') = \sigma(f_{7\times7}([F'_{avg}; F'_{max}])), \tag{4}$$

where σ denotes the sigmoid activation function, and $f_{7\times7}$ represents the convolutional layer with a 7 × 7 kernel.

The overall computation of the CBAM is summarized as follows:

$$\begin{aligned} F' &= M_c(F) \otimes F, \\ F'' &= M_s(F') \otimes F', \end{aligned} \tag{5}$$

where $\otimes$ denotes element-wise multiplication, M_c and M_s are the channel and spatial attention maps, respectively, and F'' the final output of the CBAM module [33].

Using Eq (5), the CBAM module enhances the important channels and spatial positions in the feature map, thus improving the performance of the model.

3.4. Experimental design

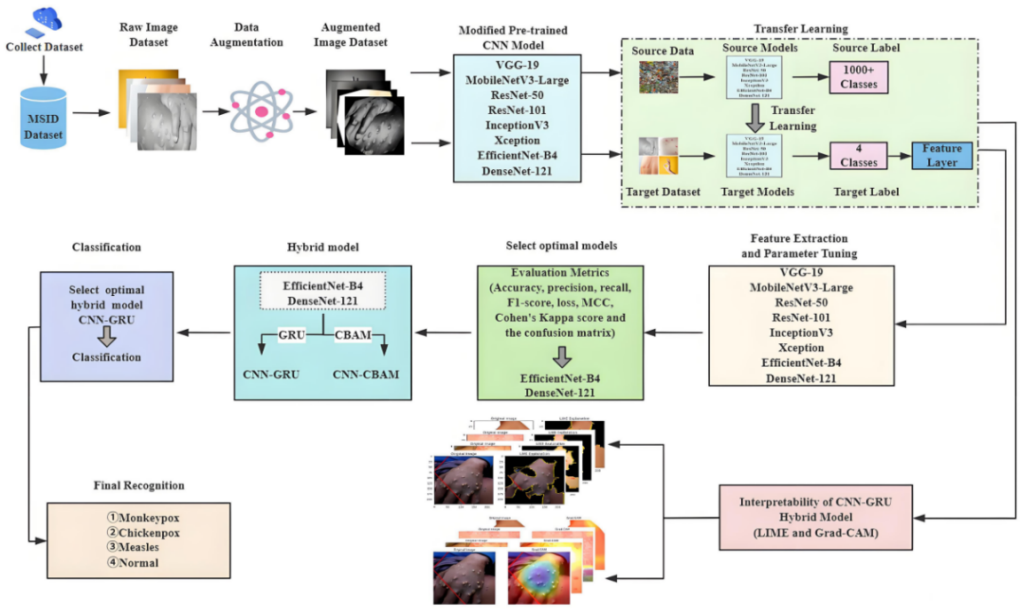


Figure 6. Flow chart of experimental method.

We utilized the eight ImageNet pre-trained deep CNN models to perform classification tasks on the MSID dataset and adapt their shallow features to the MSID dataset. By modifying the convolutional layers, we generated a new feature map that is more suited to the characteristics of the target dataset. Since the MSID dataset consists of four categories, we adjusted the number of neurons in the final fully connected layer to four to match the number of categories [14]. Based on performance evaluation

metrics, we selected the two best-performing models for fusion with GRU and CBAM. Then, two hybrid models were constructed to enhance the detection efficiency for monkeypox lesions and identify the optimal model. The detailed method flowchart is presented in Figure 6.

3.4.1. CNN-GRU hybrid model

In this section, we describe a CNN-GRU hybrid architecture designed for monkeypox skin lesion classification. To leverage complementary feature representations, two ImageNet-pretrained CNNs, DenseNet-121 and EfficientNet-B4, were employed as parallel feature extractors. Given an input image, each backbone produced a high-level feature map, which was subsequently transformed into a compact feature embedding via Global Average Pooling. The resulting embeddings were concatenated to form a fused feature vector of 2816 dimensions. Since the MSID dataset consists of static dermoscopic images without patient-level temporal information, the proposed model does not aim to capture lesion progression over time. Instead, the GRU was adopted as a feature aggregation mechanism to model structured dependencies within the fused representation. Specifically, the 2816-dimensional feature vector was reshaped into a fixed-length pseudo-sequence of tokens with dimensions $T \times d$, where $T \times d = 2816$. This pseudo-sequential formulation allows the GRU to selectively integrate informative feature components through its gating mechanism, thus enabling effective refinement of the fused CNN features without introducing temporal assumptions. Compared with conventional fully connected classification heads, the GRU-based aggregation provides adaptive feature weighting with a compact parameter footprint, thus facilitating a fair comparison with non-recurrent alternatives in subsequent ablation studies. An overview of the proposed CNN-GRU hybrid architecture is illustrated in Figure 7.

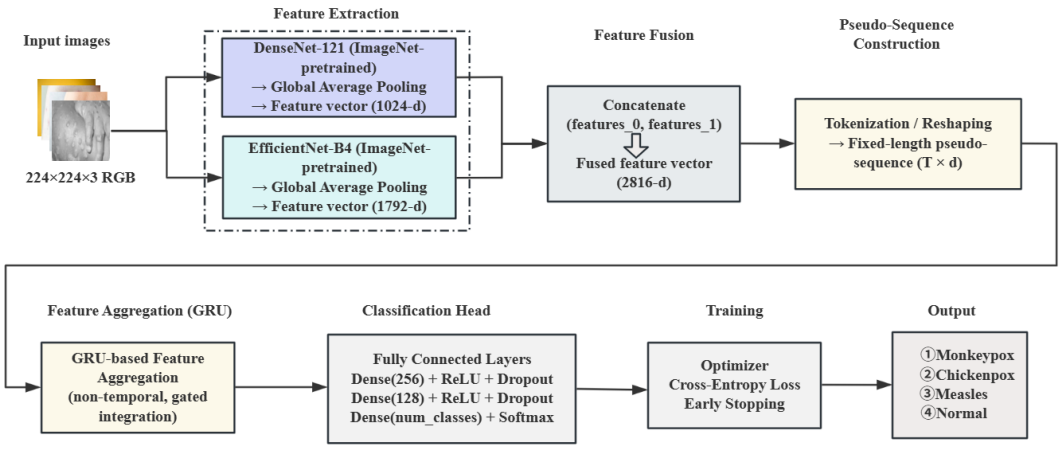


Figure 7. The framework of the CNN-GRU hybrid model (GRU as Feature Aggregator).

3.4.2. CNN-CBAM hybrid model

For clarity, throughout the remainder of this paper, the term "CNN-CBAM" specifically refers to the dual-backbone DenseNet-121+EfficientNet-B4 architecture augmented with CBAM modules in both branches. The CBAM module enables the model to focus on more relevant features by introducing

channel and spatial attention mechanisms, thus improving both its efficiency and robustness. It was applied to the outputs of both networks to enhance the feature representation, followed by global average pooling to reduce the feature dimensions. DenseNet-121 and EfficientNet-B4 are used to extract features with their top layers removed to obtain feature outputs. The features of DenseNet-121 and EfficientNet-B4 were concatenated and processed through two FC layers. Each FC layer was followed by a dropout layer with a rate of 0.2. The model was compiled using the Adam optimizer and a cross-entropy loss function, with a learning rate of 0.0001, 100 epochs, and a batch size of 16. The final FC layer produced classification results by using the softmax activation function. To prevent overfitting, an early stopping callback was employed to halt training when the validation loss ceased to improve. The structure of the CNN-CBAM hybrid model is detailed in Figure 8.

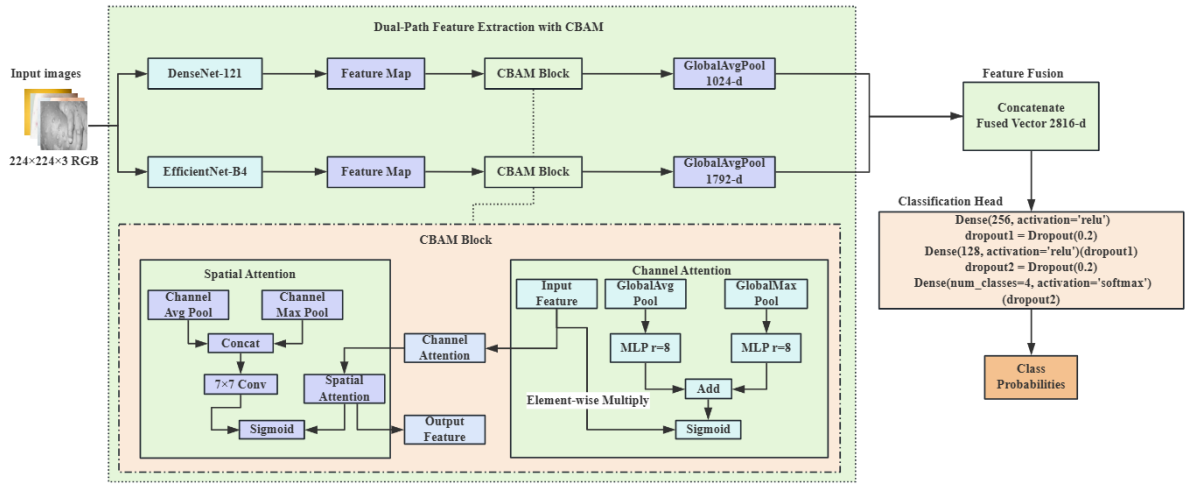


Figure 8. The framework of the CNN-CBAM hybrid model.

3.5. Evaluation metrics

We employed several metrics, including accuracy, precision, recall, F1-score, loss, Matthews correlation coefficient (MCC), Cohen's Kappa score (κ), and the confusion matrix, to evaluate the performance of the monkeypox skin image classification model. The specific formulas for these metrics are provided below.

$$\text{Accuracy} = p_0 = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{6}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{7}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{8}$$

$$\text{F1-score} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}} \tag{9}$$

$$\text{Loss} = -\sum_{i=1}^{n} y_i \log(p_i) \tag{10}$$

$$\text{MCC} = \frac{\text{TP} \times \text{TN} - \text{FP} \times \text{FN}}{\sqrt{(\text{TP} + \text{FP}) \cdot (\text{TP} + \text{FN}) \cdot (\text{TN} + \text{FP}) \cdot (\text{TN} + \text{FN})}} \tag{11}$$

$$\begin{aligned} P_{positive} &= \frac{(\text{TP} + \text{FP})(\text{TP} + \text{FN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})^2} \\ P_{negative} &= \frac{(\text{FN} + \text{TN})(\text{FP} + \text{TN})}{(\text{TP} + \text{TN} + \text{FP} + \text{FN})^2} \end{aligned} \tag{12}$$

$$p_e = p_{positive} + p_{negative} \tag{13}$$

$$\kappa = \frac{p_0 - p_e}{1 - p_e} \tag{14}$$

Among them, TP represents the number of positive samples that were predicted as positive, FN represents the number of positive samples that were predicted as negative, FP represents the number of negative samples that were predicted as positive, TN represents the number of negative samples that were predicted as negative, Loss represents the value of cross entropy loss, n represents the number of classification categories, p_i represents the probability of being predicted as the i-th class, and y_i is the label value. When the predicted class is consistent with the real label, y_i is 1; otherwise, it is 0.

3.6. Experimental setup

The experiment was conducted in a Windows 10 environment using the Tesla P100-PCIE-16 GB GPU provided by the Kaggle platform as the primary computational resource. The model was built and trained using the TensorFlow and Keras libraries. To better meet the task's specific requirements, transfer learning was used to fine-tune the pre-trained model, thus enhancing the training efficiency and accelerating the overall process. Throughout the experiment, the Adam optimizer and the cross-entropy loss function were utilized for all models. The hyperparameter settings are listed in Table 3.

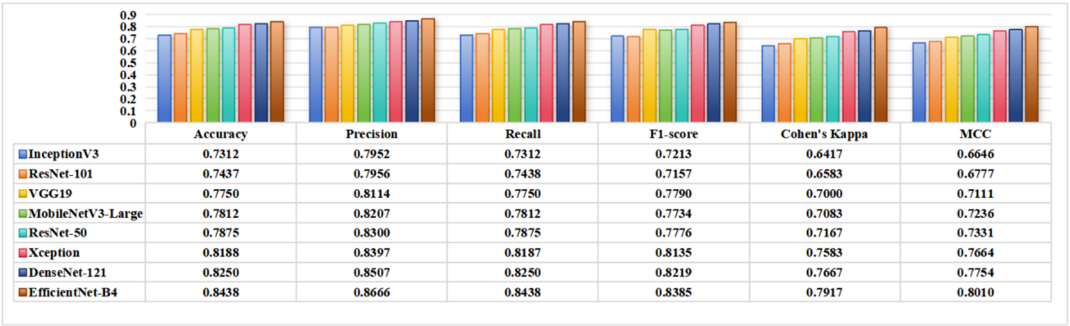
Table 3. Model hyperparameter settings.

Model	Epoch	Batchsize	Learning rate	Dropout	L2	Patience
ResNet-50	100	16	0.0001	0.2	0.01	20
ResNet-101	100	32	0.001	0.2	0.01	10
VGG-19	100	32	0.0001	0.3	0.01	20
InceptionV3	100	32	0.0001	0.2	0.01	10
Xception	100	32	0.001	0.2	0.01	10
MobileNetV3-Large	100	8	0.0001	0.2	0.01	20
DenseNet-121	100	32	0.001	0.3	0.01	20
EfficientNet-B4	100	32	0.001	0.2	0.01	10

4. Results

4.1. Experimental results and analysis of baseline model

In this section, we employ eight baseline models for classification on the MSID dataset: VGG19, MobileNetV3-Large, ResNet-50, ResNet-101, InceptionV3, Xception, EfficientNet-B4, and DenseNet-121. Figure 9 presents a performance comparison of these eight baseline models on the MSID dataset. The EfficientNet-B4 model demonstrated an outstanding performance and achieved a precision of 0.8666 and an accuracy of 0.8438. It not only highlights its robustness in accurately identifying positive cases across various skin conditions, but also reflects the capacity to correctly classify negative cases and maintain a high overall accuracy under different conditions. Furthermore, the EfficientNet-B4 model outperformed the others in recall, F1 score, and loss, which indicate its ability to effectively capture most positive samples while minimizing false positives. Additionally, it achieved a Cohen's Kappa value of 0.7917 and an MCC of 0.8010, further confirming its superior classification accuracy and consistency.



	Accuracy	Precision	Recall	F1-score	Cohen's Kappa	MCC
InceptionV3	0.7312	0.7952	0.7312	0.7213	0.6417	0.6646
ResNet-101	0.7437	0.7956	0.7438	0.7157	0.6583	0.6777
VGG19	0.7750	0.8114	0.7750	0.7790	0.7000	0.7111
MobileNetV3-Large	0.7812	0.8207	0.7812	0.7734	0.7083	0.7236
ResNet-50	0.7875	0.8300	0.7875	0.7776	0.7167	0.7331
Xception	0.8188	0.8397	0.8187	0.8135	0.7583	0.7664
DenseNet-121	0.8250	0.8507	0.8250	0.8219	0.7667	0.7754
EfficientNet-B4	0.8438	0.8666	0.8438	0.8385	0.7917	0.8010

Figure 9. Baseline model evaluation results.

In comparison, the DenseNet-121 model exhibited a slightly inferior performance but still achieved notable results with a precision of 0.8507 and an accuracy of 0.8250. Additionally, it overtook the other six models in recall and F1-score. The architectural features of EfficientNet-B4 and DenseNet-121 aligned well with the MSID dataset, which enabled a better adaptation to its complexities and superior performance.

Figures 10 and 11 present the confusion matrices and receiver operating characteristic (ROC) curves for the eight evaluated models on the test set in a four-class classification task. The confusion matrices are structured as 4 × 4 grids, where rows represent ground-truth labels and columns denote predicted labels. Each cell indicates the frequency of samples classified into corresponding actual-predicted category pairs. An optimal performance is exhibited by the maximized diagonal values (approaching the ideal count of 40 per class, reflecting correct classifications) and the minimized off-diagonal entries (indicating reduced misclassifications). The area under the ROC curve (AUC) serves as a critical metric to evaluate the model's discriminative capacity. AUC values that approach 1.0 signify superior predictive accuracy. A comparative analysis of Figures 10 and 11 demonstrates that EfficientNet-B4 and DenseNet-121 outperformed other models in confusion matrices and ROC metrics. This highlights that they possess enhanced classification accuracy, robust sensitivity in detecting positive samples, and effective suppression of false positives.

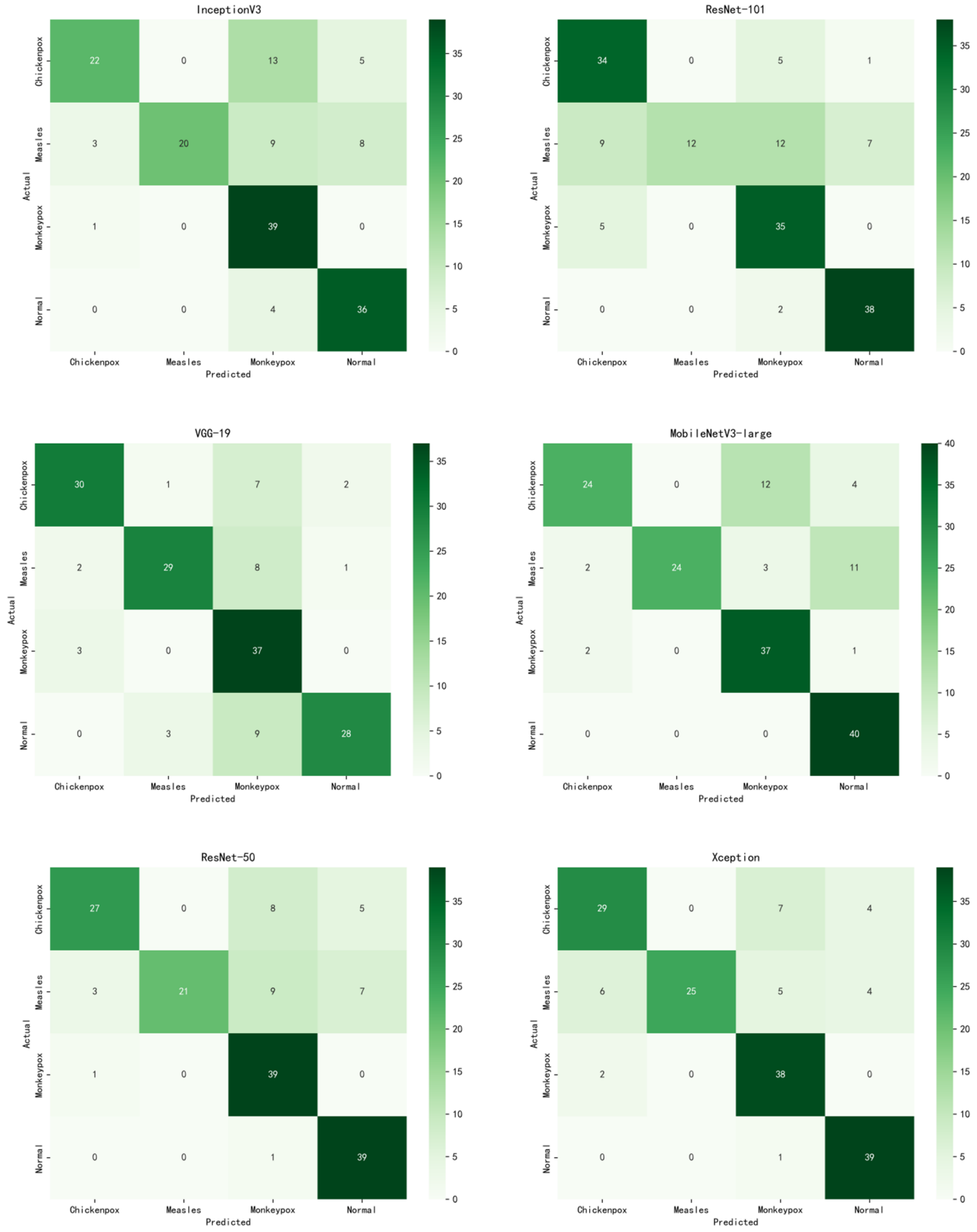


Figure 10. Confusion matrix of baseline model.

Continued on next page

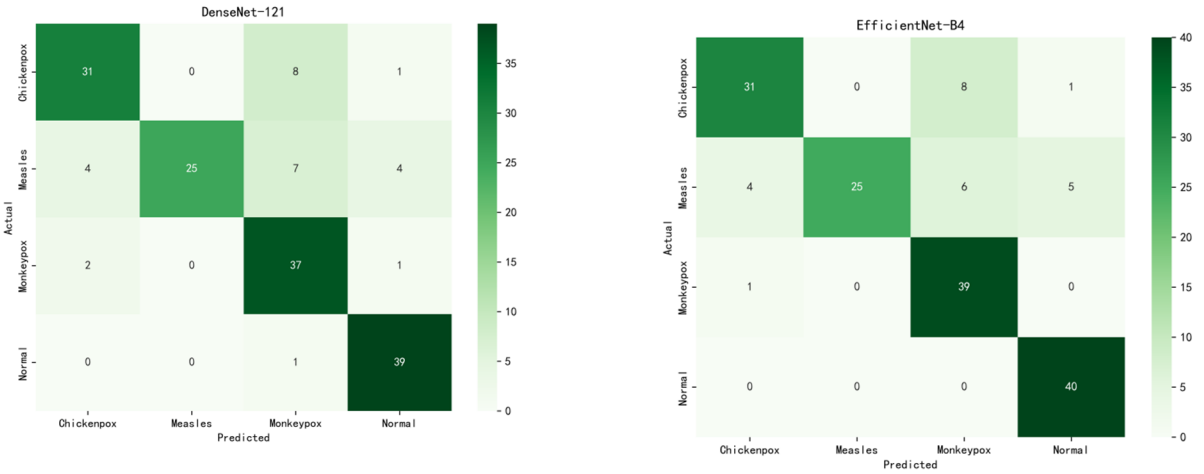


Figure 10. Confusion matrix of baseline model.

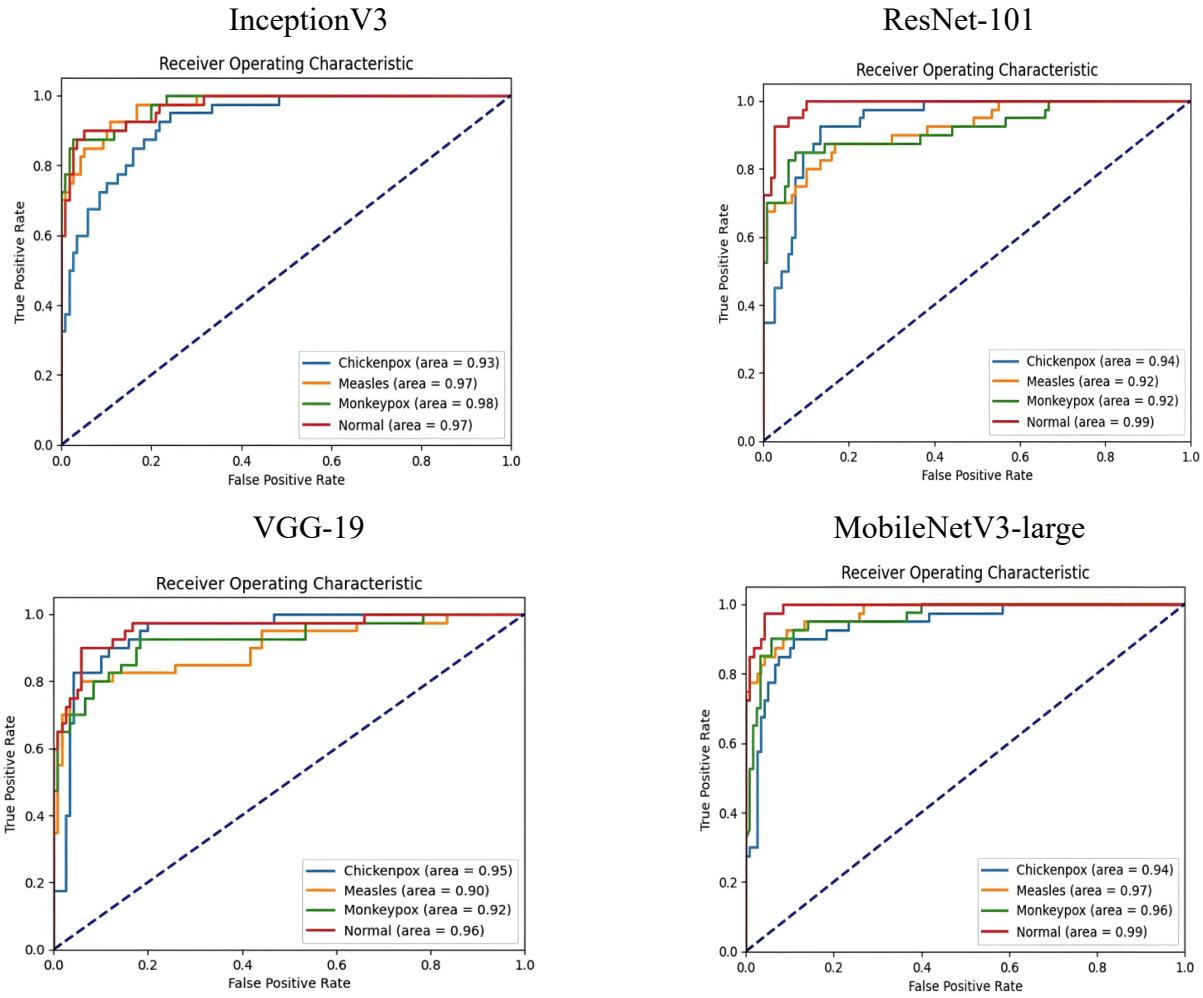


Figure 11. ROC curve of baseline model.

Continued on next page

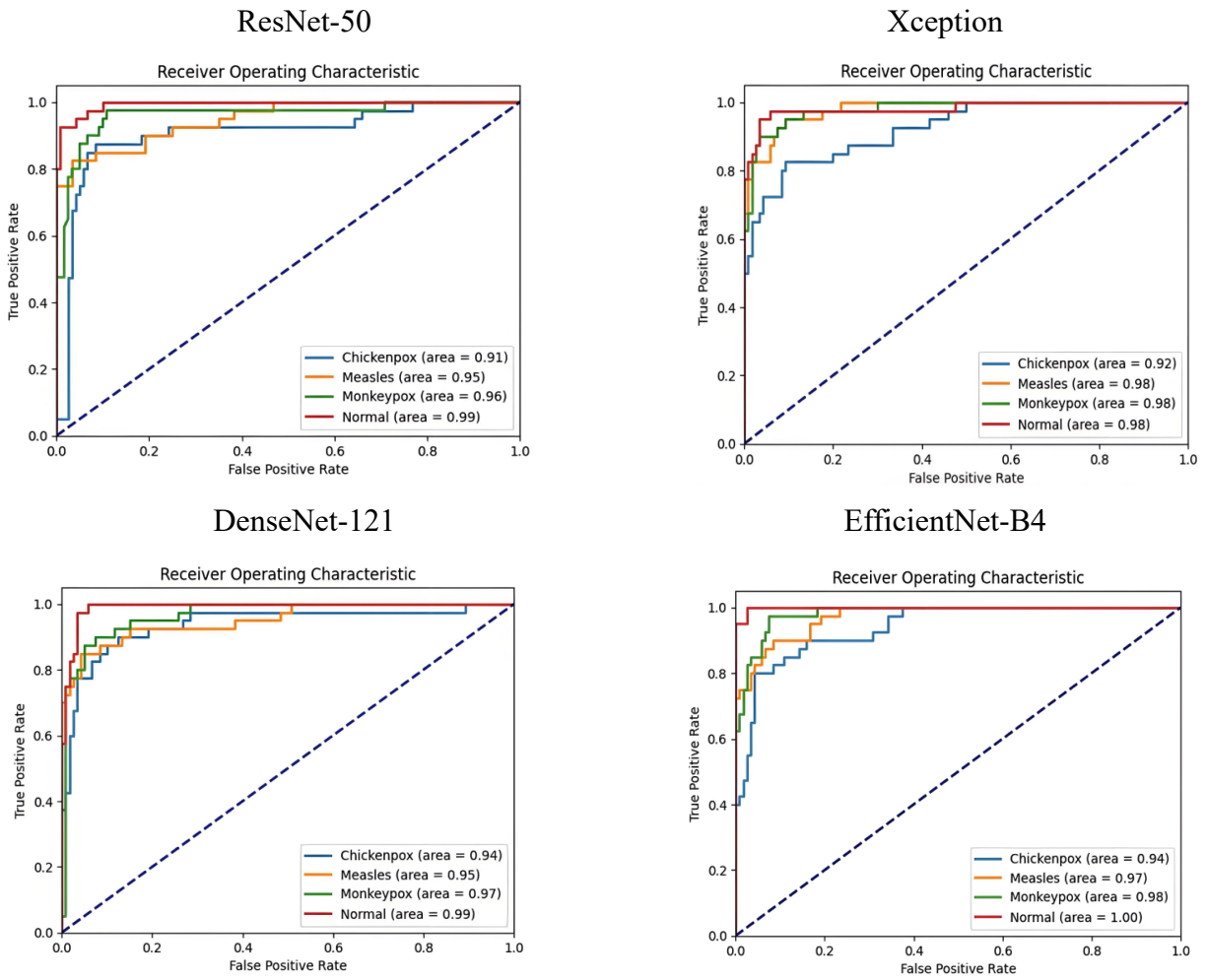
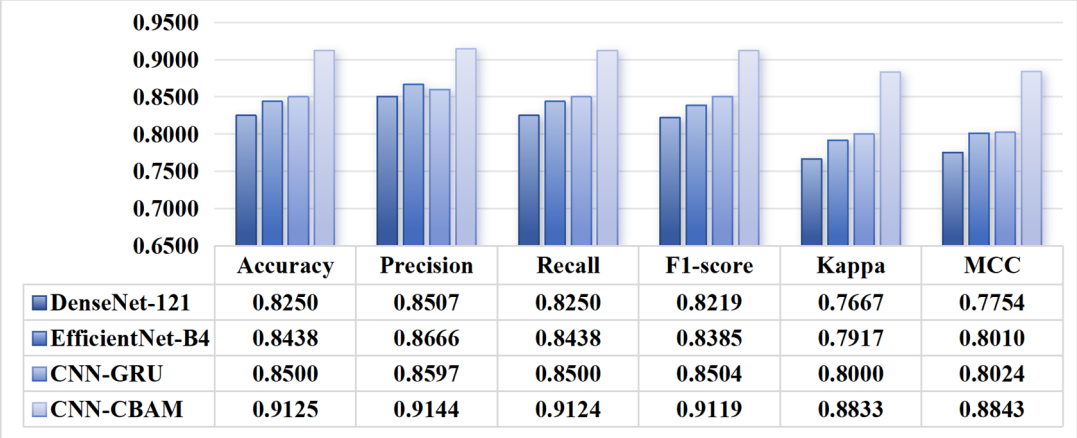


Figure 11. ROC curve of baseline model.

4.2. Hybrid model experimental results and analysis

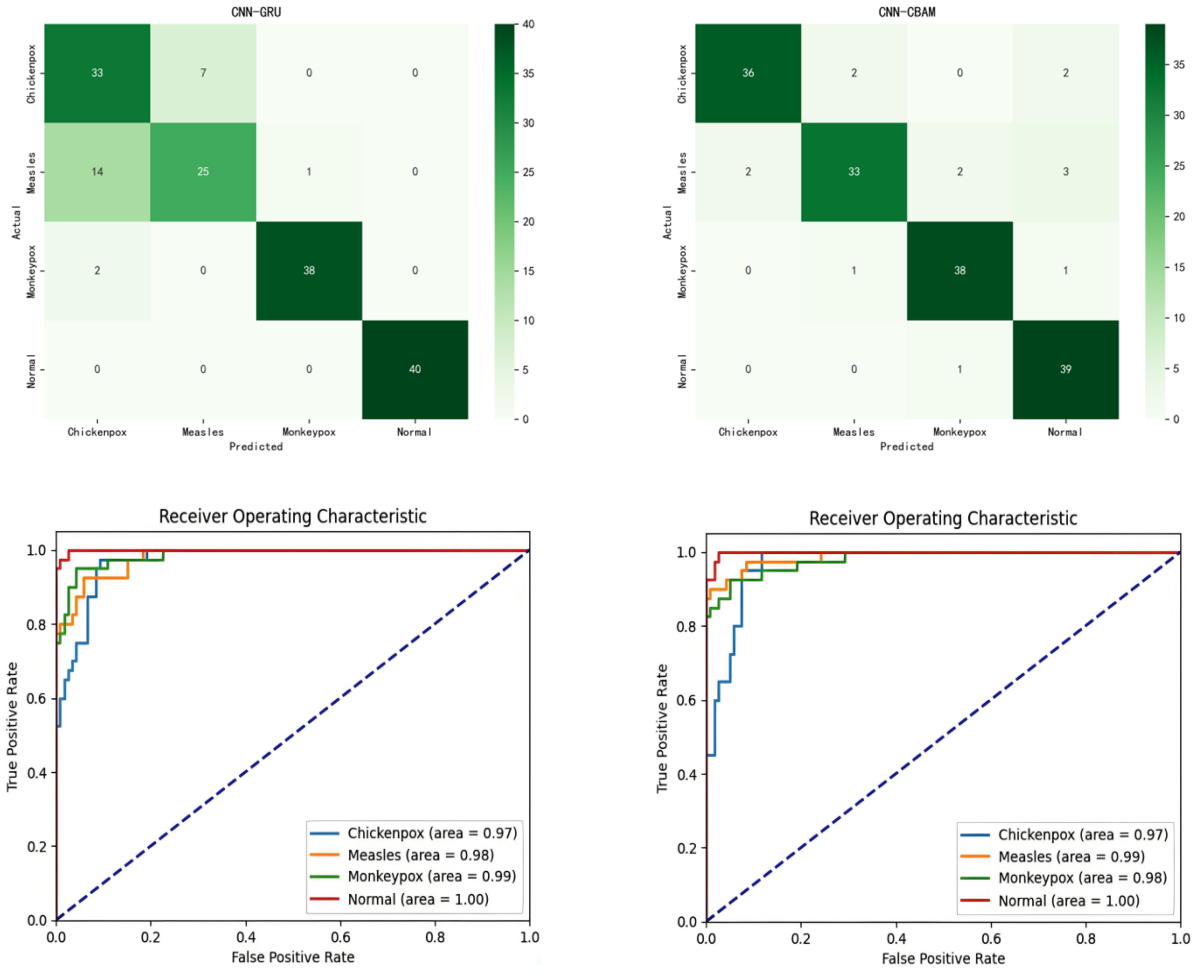
Based on a comprehensive evaluation of multiple performance metrics, both the EfficientNet-B4 and DenseNet-121 models exhibited superior overall performances. In this section, we integrate these two models with GRU and CBAM, respectively, to develop two hybrid models to further enhance the prediction accuracy. After parameter optimization and training, we compared the baseline EfficientNet-B4 and DenseNet-121 models with the two hybrid models, as shown in Figure 12. The CNN-CBAM hybrid model achieved an accuracy of 0.9125 and a precision of 0.9144. These suggest that the model effectively identified positive instances and demonstrated strong performances in predicting positive classes. The recall of 0.9124 and the F1 score of 0.9119 indicate that the model can capture most positive samples while maintaining a high level of accuracy and completeness. Additionally, both Cohen's Kappa and MCC exceed 0.88, thus indicating a strong agreement between the predicted labels and the true labels. It shows the model's enhanced ability to learn optimal parameters during training, thus resulting in higher prediction accuracy.



	Accuracy	Precision	Recall	F1-score	Kappa	MCC
DenseNet-121	0.8250	0.8507	0.8250	0.8219	0.7667	0.7754
EfficientNet-B4	0.8438	0.8666	0.8438	0.8385	0.7917	0.8010
CNN-GRU	0.8500	0.8597	0.8500	0.8504	0.8000	0.8024
CNN-CBAM	0.9125	0.9144	0.9124	0.9119	0.8833	0.8843

Figure 12. Evaluation results of the model.

Figure 13 presents the confusion matrices and ROC curves of two hybrid models, CNN-GRU and CNN-CBAM, for the four-class classification task that involve chickenpox, measles, monkeypox and normal skin conditions.



CNN-GRU CNN-CBAM

Figure 13. Confusion matrix and ROC curve of hybrid model.

An analysis of the confusion matrices shows that the CNN-GRU model achieved an overall satisfactory performance. It demonstrated a high accuracy, particularly in identifying monkeypox, with 38 out of 40 samples correctly classified and normal skin conditions with perfect classifications. However, significant mutual misclassification occured between chickenpox and measles, with 7 and 14 misclassified cases, respectively. This indicates that the model has a limited ability to distinguish between these two visually similar diseases. In contrast, the CNN-CBAM model exhibited a more balanced performance. The number of correctly predicted cases increased for both chickenpox, with 36 out of 40 samples and measles with 33 out of 40 samples. Mutual misclassification between these two classes was substantially reduced. A high accuracy was maintained for monkeypox, with 38 out of 40 samples correctly identified, and a near-perfect performance was achieved for normal skin conditions, with 39 out of 40 samples correctly classified. These results suggest that the CBAM attention mechanism effectively enhances the inter-class discriminability by strengthening the learning of spatial and channel features. Further support for the above conclusions comes from the analysis of ROC curves. The CNN-GRU model yielded high AUC values ranging from 0.97 to 1.00. The CNN-CBAM model achieved comparable or superior AUC values for classes such as measles where the AUC reached 0.99. Moreover, its ROC curves were closer to the upper-left corner, thus reflecting a better balance between sensitivity and specificity.

Overall, the CNN-CBAM model demonstrated a more stable classification performance than the CNN-GRU model. It significantly reduces confusion between similar disease classes. The performance improvement is attributed to the CBAM module which adaptively focuses on the key features for diagnosis. This validates the effectiveness of the attention mechanism in hybrid CNN architectures.

4.3. Ablation study on feature aggregation strategies

To further investigate the role of recurrent aggregation in static image classification, we conducted an ablation study that compared a GRU-based aggregation head with a simpler multilayer perceptron (MLP) head. Both models shared the same convolutional feature extractors (DenseNet-121 and EfficientNet-B4) and used identical fused feature representations, thus ensuring a fair comparison. In the GRU-based configuration, the fused feature vector was reshaped into a fixed-length pseudo-sequence and processed using a GRU module that acted as a gated feature aggregator. In contrast, the MLP-based model directly fed the fused features into a lightweight fully connected classifier with a comparable parameter scale. The quantitative results are summarized in Table 4. The MLP-head model achieved a higher overall performance, thereby obtaining an accuracy of 0.8688, a weighted F1-score of 0.8682, a Cohen's Kappa of 0.8250, and an MCC of 0.8268. In comparison, the GRU-based aggregation model achieved an accuracy of 0.8500 and slightly lower agreement metrics.

These results indicate that, for static skin lesion images without an inherent temporal structure, recurrent aggregation does not provide additional performance benefits over a simpler feedforward classifier. This finding is consistent with the characteristics of the MSID dataset and supports the use of lightweight non-recurrent heads for static image-based diagnoses. Therefore, the GRU module is interpreted as an exploratory gated aggregation mechanism rather than a superior alternative to MLP-based classification. Importantly, this negative result is itself informative, as it empirically demonstrates that sequence-based aggregation modules do not necessarily benefit static dermatological image classification, thus providing practical guidance for future model design.

Table 4. Ablation study on feature aggregation strategies for fused CNN features.

Model	Accuracy	Precision	Recall	F1-score	Kappa	MCC
DenseNet + EffNet + GRU-spatial-token	0.8500	0.8597	0.8500	0.8504	0.8000	0.8024
DenseNet + EffNet + MLP-head	0.8688	0.8737	0.8688	0.8682	0.8250	0.8268

4.4. Statistical robustness and significance analysis

4.4.1. Performance comparison across multiple random seeds

To evaluate the robustness of the proposed framework with respect to random initialization, all models were independently trained and evaluated using five different random seeds. For each model, the mean and standard deviation of accuracy, macro-F1 score, and MCC were computed on the MSID test set, thus providing a comprehensive assessment of both predictive performance and stability across runs.

As summarized in Table 5, the proposed CNN-CBAM framework consistently outperformed the EfficientNet-B4 baseline across all evaluation metrics. Specifically, CNN-CBAM achieved a higher average accuracy, macro-F1 score, and MCC, while simultaneously exhibiting lower standard deviations compared with the baseline. This reduced variability across random seeds indicates an improved robustness to initialization and more stable training behavior.

Table 5. Performance comparison across five random seeds (mean ± standard deviation).

Model	Accuracy (mean ± std)	Macro-F1 (mean ± std)	MCC (mean ± std)
EfficientNet-B4	0.8313 ± 0.0145	0.8259 ± 0.0163	0.7867 ± 0.0178
CNN-CBAM	0.9119 ± 0.0088	0.9104 ± 0.0128	0.8837 ± 0.0101

These results suggest that the integration of dual convolutional backbones with attention mechanisms enables CNN-CBAM to learn more discriminative and reliable feature representations, thus improving both performance and consistency in monkeypox skin lesion classification.

4.4.2. Bootstrap-based confidence interval analysis

To further quantify the statistical uncertainty associated with the reported performance, bootstrap-based 95% confidence intervals (CIs) were computed on the MSID test set using 1000 bootstrap resamples. This analysis estimates the variability of the model's performance under repeated sampling from the same test population and complements the multi-seed evaluation by focusing on test-set uncertainty.

The resulting confidence intervals are reported in Table 6. Across all three metrics, the confidence intervals of CNN-CBAM were consistently shifted toward higher values compared with those of EfficientNet-B4. Although a partial overlap between the confidence intervals was observed, this behavior is expected given the limited size of the test set. Importantly, the overall trend of the bootstrap confidence intervals well aligns with the improvements observed in the multi-seed evaluation. Taken together, these results provide additional evidence that the superior performance of CNN-CBAM is

not attributable to a specific random split or initialization, but rather reflects a consistent advantage under test-set resampling variability.

Table 6. Bootstrap-based 95% confidence intervals on the MSID test set.

Model	Accuracy (95% CI)	Macro-F1 (95% CI)	MCC (95% CI)
EfficientNet-B4	[0.7875, 0.8587]	[0.7692, 0.8406]	[0.7143, 0.8073]
CNN-CBAM	[0.8250, 0.9262]	[0.8155, 0.8944]	[0.7903, 0.8906]

4.4.3. Paired statistical comparison using McNemar's test

In addition to the above analyses, a paired McNemar's test was conducted to directly compare CNN-CBAM and EfficientNet-B4 on identical test samples. McNemar's test focuses on the disagreement patterns between two classifiers and is appropriate to evaluate paired nominal outcomes. The results of the McNemar's test are reported in Table 7, where b denotes the number of samples correctly classified by EfficientNet-B4 but misclassified by CNN-CBAM, and c denotes the number of samples correctly classified by CNN-CBAM but misclassified by EfficientNet-B4. As shown in the table, CNN-CBAM correctly classified more samples that were misclassified by EfficientNet-B4 than vice versa (c = 28 vs. b = 15), thus indicating a consistent performance advantage on a per-sample basis. While the observed difference does not reach statistical significance under the conventional threshold when using the continuity-corrected McNemar's test ($p = 0.0673$), the direction of the effect is consistent with the results obtained from both the multi-seed evaluation and the bootstrap confidence interval analysis. The lack of statistical significance is likely attributable to the limited size of the MSID test set, which restricts the statistical power of paired hypothesis testing.

Table 7. Paired McNemar test results comparing CNN-CBAM and EfficientNet-B4.

Comparison	CNN-CBAM vs EfficientNet-B4
b	15
c	28
p-value	0.0673

Overall, the combined evidence from multi-seed experiments, bootstrap confidence intervals, and paired statistical testing supports the robustness and effectiveness of the proposed CNN-CBAM framework.

4.5. Generalization proof of the CNN-CBAM model

In clinical practice, models must cope with skin images that originate from diverse sources and are captured under varying acquisition conditions. Therefore, training results solely based on the four-class MSID dataset are insufficient to demonstrate practical clinical applicability; rigorous evaluation of generalization ability is a fundamental prerequisite to translate a model from the "laboratory" to clinical deployment. Existing studies are largely restricted to validation on a single dataset. This work assessed the generalizability of the CNN-CBAM through cross-dataset and cross-task evaluations, thus addressing the current shortfall in the monkeypox-detection literature where systematic verification of model adaptability to heterogeneous data sources is lacking and strengthening the overall model-

validation framework. For the generalization experiments, we employed the MSLD and the MSLD v2.0. The datasets are described in detail below.

4.5.1. Monkeypox skin lesion dataset (MSLD)

The original dataset is comprised of 228 images, of which 102 are labeled as "monkeypox" and the remaining 126 are labeled as "other" (non-monkeypox cases, i.e., chickenpox and measles). The images were partitioned into training, validation, and test sets with an approximate ratio of 70:20:10. To preserve the authenticity of the data, data augmentation was only applied to the training set; the validation and test sets only underwent standard normalization [34].

4.5.2. Mpox skin lesion dataset version 2.0 (MSLD v2.0)

The MSLD v2.0 [35] is comprised of six categories: Monkeypox (284 images), Chickenpox (75 images), Measles (55 images), Cowpox (66 images), hand-foot-and-mouth disease (HFMD, 161 images), and healthy skin (114 images), resulting in a total of 755 original skin lesion images collected from 541 distinct patients. The dataset was curated and endorsed by professional dermatologists and approved by appropriate regulatory authorities, thus ensuring high clinical reliability.

Importantly, MSLD v2.0 provides explicit patient identifiers embedded in image filenames, following the format DiseaseCode_PatientNumber_ImageNumber. Based on this information, the dataset authors released an official five-fold cross-validation protocol (FOLDS) in which images are partitioned at the patient level. In this study, we strictly followed the official fold definitions, thus ensuring that images from the same patient did not appear across training, validation, and test sets, thereby preventing patient-level data leakage.

The dataset further provides a separate directory (FOLDS_AUG) that contains augmented images exclusively corresponding to the training subsets of each fold. These augmented samples were generated using a variety of transformations, including rotation, translation, reflection, color jittering, noise injection, and scaling. In our experiments, only the augmented training images were used for model optimization, while the validation and test sets solely consisted of original, non-augmented images. This strategy ensures fair evaluation and avoids contamination of evaluation subsets by augmented data.

By adhering to the official patient-level splitting protocol and restricting data augmentation strictly to the training sets, the experimental design on MSLD v2.0 ensures rigorous evaluation and reliable assessment of the model's generalization capability. The detailed results are shown in Figure 14.

The confusion matrices for the CNN-CBAM model revealed key characteristics of its classification behavior in both the binary and multi-class tasks. On the MSLD binary dataset, the "Monkeypox" class demonstrated correct classification for 19 samples, with 1 misclassified as "Other". Conversely, the "Other" class showed correct classification for 22 samples, with 3 misclassified as "Monkeypox". These results indicate a generally good overall recognition capability for the model. In the MSLD v2.0 multi-class task, the confusion matrix presented a more complex classification structure. For individual classes, "Chickenpox" and "Cowpox" achieved correct classifications for 28 and 34 samples, respectively, without significant extreme misclassifications. However, "HFMD", "Normal", and "Measles" achieved correct classifications for 79, 47, and 25 samples, respectively, yet exhibited some degree of multi-class confusion. As the largest class, "Monkeypox" was correctly

identified for 115 samples, but also suffered from misclassifications that originated from other categories.

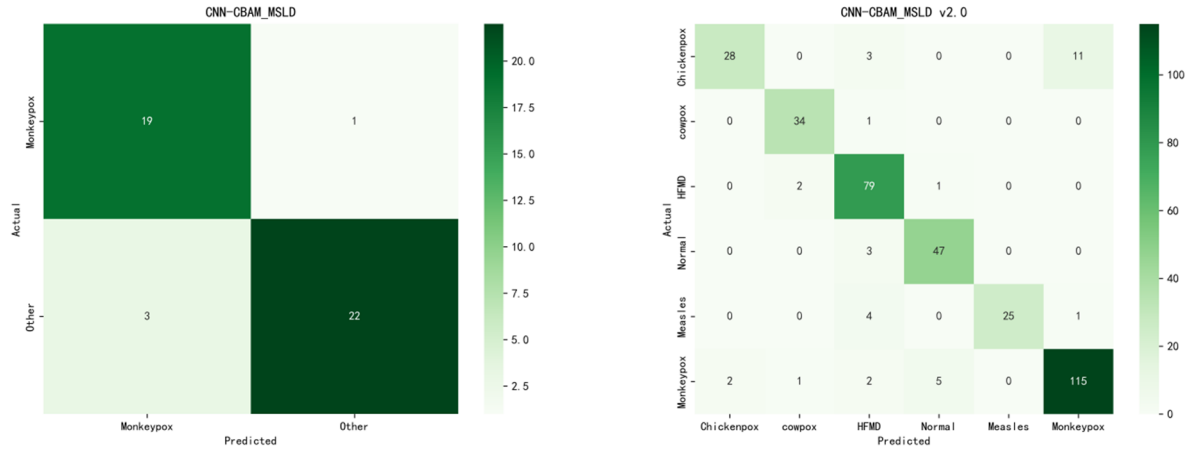
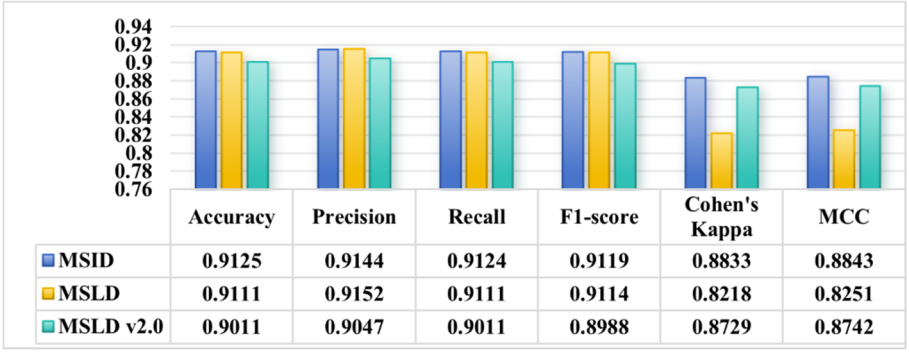


Figure 14. Confusion matrices of CNN-CBAM on binary and six-class datasets.

Overall, due to the inherently high complexity of the six-class task and the relatively minor feature differences between some classes, this, to some extent, interfered with the model's ability to precisely distinguish between each class. Nevertheless, based on the metric of correctly classified samples within each class, the model still demonstrated a reliable discriminative capability and strong classification performance in the six-class scenario.

As shown by the experimental results in Figure 15, the CNN-CBAM model demonstrated a strong generalization performance across different types of datasets.



	Accuracy	Precision	Recall	F1-score	Cohen's Kappa	MCC
MSID	0.9125	0.9144	0.9124	0.9119	0.8833	0.8843
MSLD	0.9111	0.9152	0.9111	0.9114	0.8218	0.8251
MSLD v2.0	0.9011	0.9047	0.9011	0.8988	0.8729	0.8742

Figure 15. Comparison of performance metrics of the CNN-CBAM model on MSID, MSLD, and MSLD v2.0 datasets.

In the MSID four-class task, the model achieved high comprehensive evaluation scores, with an accuracy of 0.9125, precision of 0.9144, recall of 0.9124, F1-score of 0.9119, Cohen's Kappa of 0.8833, and MCC of 0.8843. These metrics collectively indicate that the model effectively learns discriminative features and achieves accurate class separation in this relatively complex multi-class scenario. In the MSLD binary task, the model performed consistently well, with only minor fluctuations in individual metrics. It maintained a high overall performance, including an accuracy

of 0.9111 and precision of 0.9152, thus confirming that CNN-CBAM retains a stable classification capability even in simpler binary tasks. Furthermore, when applied to the more challenging MSLD v2.0 six-class dataset, the model exhibited a moderate decline in performance; for instance, the accuracy decreased to 0.9011 and the precision decreased to 0.9047, although they remained within a relatively desirable range. This outcome suggests that the model is still capable of extracting key features and maintaining a competent classification performance even when confronted with higher-complexity multi-class problems.

In summary, the CNN-CBAM model delivers a consistent and reliable classification performance across binary, four-class, and six-class tasks. Multiple evaluation metrics jointly validate its adaptability to varying data distributions and classification scenarios, thus underscoring its strong generalization capability.

4.6. Comparison of results on MSID dataset

Next, we employed multiple evaluation metrics to assess the performance of different methods on the MSID dataset. The results are shown in Figure 16, from the overall trend, the CNN-CBAM proposed method demonstrated the best performance in the vast majority of metrics. It achieved an accuracy of 0.9125, a recall of 0.9124, an F1-score of 0.9119, a Cohen's Kappa of 0.8833, and an MCC of 0.8843, all of which are higher than those of the other two methods. In terms of precision, although it was slightly lower than Uysal's result of 0.9300, it still reached 0.9144, which is a high level. The Uysal method showed an outstanding performance in precision, thereby reaching 0.9300, and had a decent F1-score of 0.9000. However, it was significantly lower than the CNN-CBAM proposed method in terms of accuracy, recall, Cohen's Kappa, and MCC. The Sitana and Shahi method recorded the lowest values across all metrics, thus indicating relatively weak performances in this task. In summary, the CNN-CBAM proposed method possesses significant advantages across multiple key evaluation dimensions of the classification task, thus reflecting its superiority in model performance.

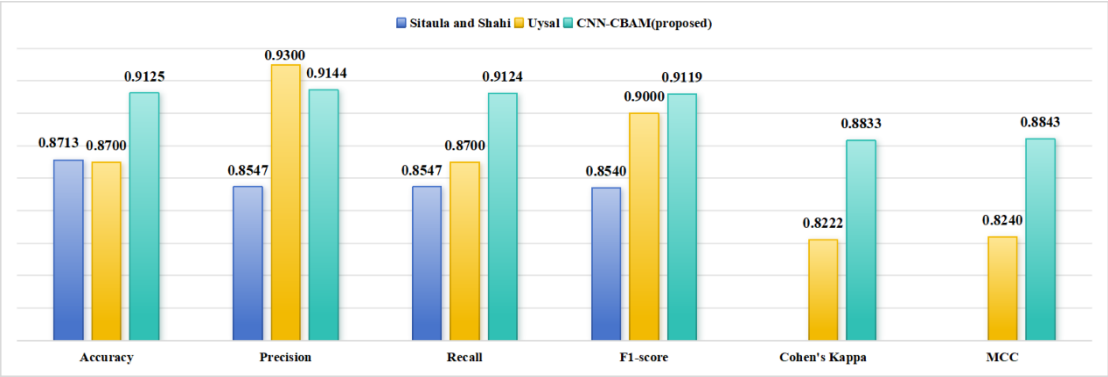


Figure 16. Comparison of results on MSID dataset.

5. Interpretability of hybrid models

To gain a deeper insight into the decision-making process of hybrid models, we employed two interpretability techniques: Grad-CAM and LIME. These methods were used to reveal the model's internal decision-making mechanisms by identifying key input features and activation patterns, thus

enhancing both the interpretability and operational transparency of the model.

5.1. Local interpretable model-agnostic explanations

LIME is a method designed to interpret predictions made by machine learning models. It is particularly useful for complex models, such as deep neural networks. The method operates by generating perturbed samples around a specific input data point. These perturbed samples are created by making small modifications to the original input. The model's predictions for these modified instances are observed and recorded. Based on the observations, LIME trains a simple and interpretable model, such as a linear regression or a decision tree. This interpretable model approximates the behavior of the original complex model in the local region surrounding the input data point. By focusing on local approximations, LIME provides insights into how the model makes predictions for specific instances, thus enhancing the interpretability and the transparency. The local explanatory model [36] is optimized by the following:

$$\xi(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g), \tag{15}$$

where $\xi(x)$ represents the explanation generated by LIME, x denotes the target sample to be explained (e.g., a skin lesion image), the original complex model is represented by f, g refers to the interpretable surrogate model (typically a linear model or decision tree), the kernel function π_x defines the local weighting of samples, the loss function L measures the approximation error of the surrogate model g within the local neighborhood of f, the regularization term $\Omega(g)$ constrains the complexity of g, and G represents the hypothesis space of interpretable models (e.g., all sparse linear models).

The comparison evaluation results of the CNN-CBAM model and the CNN-GRU model are shown in Figure 17.

The interpretability analysis figures indicate that the yellow boundaries in the CNN-CBAM model's LIME explanations closely surround the dense and typical lesion areas within the images, aligning well with the medical morphology of lesions. These boundaries accurately marked the most critical skin regions for the model's prediction of various lesion types, thus suggesting the model can accurately capture core visual features associated with lesions. In contrast, the regions delineated by the yellow boundaries in the LIME explanations for the CNN-GRU model showed slightly less alignment with typical lesion areas compared to CNN-CBAM. The marked regions covered the core lesion features less precisely, thereby exhibiting some deviation from key lesion areas, which reflects a slightly lower accuracy in capturing critical visual lesion features. Furthermore, the attention heatmaps of the CNN-CBAM model displayed concentrated and complete white areas, which indicate high attention weights that cover the main distribution regions of the lesion images. This clearly illustrates that the model highly focuses its attention on key regions related to lesions while paying less attention to irrelevant areas. Conversely, the attention heatmaps of the CNN-GRU model showed a relatively scattered distribution of high-weight white areas, thereby covering the main lesion regions less completely and concentratively than CNN-CBAM. This suggests the CNN-GRU model exhibits slightly insufficient focus on key regions.

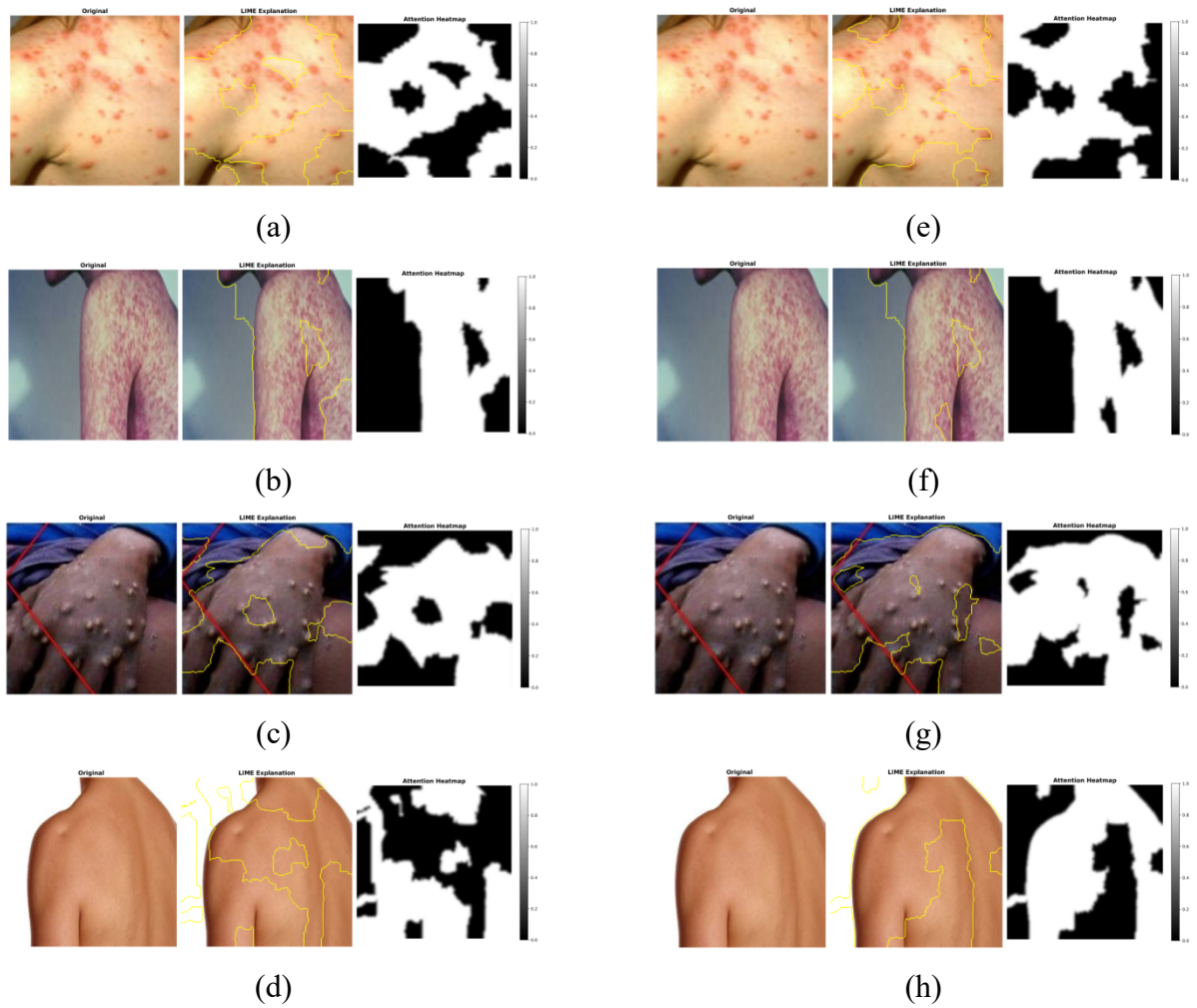


Figure 17. LIME's interpretability analysis chart. Figures (a)–(d) illustrate the interpretability analysis using LIME for the CNN-CBAM model, while Figures (e)–(h) present the corresponding LIME interpretability analysis for the CNN-GRU model. Each subplot consists of three components: the original image, the LIME key regions, and the attention heatmap. In the LIME explanations, the yellow boundaries clearly outline the regions deemed most critical by LIME for the lesion prediction. The attention heatmaps utilize a gradient from black to white to represent attention weights. White areas indicate regions that receive the highest attention from the model, whereas black areas represent regions that receive the lowest attention.

5.2. Quantitative evaluation of LIME interpretability analysis

For the quantitative evaluation of LIME interpretability, we selected incremental and decremental curves as well as stability metrics to comparatively assess the CNN-CBAM and CNN-GRU models.

5.2.1. Incremental and decremental curves

Incremental and decremental curves serve as core metrics to evaluate the fidelity of interpretability methods [37]. The underlying principle involves systematically modifying the input

images by progressively removing or adding pixels and observing the corresponding changes in the model's prediction probabilities. A high-quality explanation should accurately identify the features most critical to the model's decision-making process. The evaluation was based on two complementary types of curves:

• Decremental curves: This curve evaluates the impact of feature removal on model predictions. The process starts from the original image, and pixels are grayed out one by one in descending order of importance as determined by LIME. Ideally, if the explanation method correctly identifies the key features, removing the top few important features should lead to a rapid decline in the model's predicted probability for the target class. The decrease in predicted probability is calculated as follows:

$$\Delta P_{dec}(k) = P(y \mid I_0) - P(y \mid I_k). \tag{16}$$

• Incremental curves: This curve evaluates the impact of feature introduction on model predictions. The process starts with a blank or blurred image, and pixels are restored one by one in descending order of importance as determined by LIME. A high-quality explanation should enable the model's predicted probability to rapidly increase after incorporating only a small number of important features [37]. The increase in predicted probability is calculated as follows:

$$\Delta P_{inc}(k) = P(y \mid I_k) - P(y \mid I_0). \tag{17}$$

In the above formulas, I_0 represents the starting image, I_k denotes the image after modifying the top k important features, and $P(y \mid I_k)$ is the model's predicted probability for class y.

To enable a quantitative comparison of the curve shapes, the AUC [38] is calculated as follows:

• AUC of decremental curve: A lower AUC value indicates a steeper drop in the model's prediction as key features are removed, thus reflecting a higher explanation fidelity.

$$AUC_{dec} = \frac{1}{T \cdot P(y \mid I_0)} \sum_{k=1}^{T} P(y \mid I_k). \tag{18}$$

• AUC of incremental curve: A higher AUC value indicates a faster recovery of the model's prediction as key features are added, thus reflecting a higher explanation fidelity.

$$AUC_{inc} = \frac{1}{T \cdot P(y \mid I_T)} \sum_{k=1}^{T} P(y \mid I_k), \tag{19}$$

where T is the total number of steps, $P(y \mid I_0)$ is the model's predicted probability of the original image, $P(y \mid I_T)$ is the model's predicted probability of the complete image, and $P(y \mid I_k)$ is the model's predicted probability at the k-th step.

5.2.2. Stability evaluation

A stability evaluation aims to measure the sensitivity of interpretability methods to minor changes in the input. Ideally, a robust interpretability method should be resistant to input perturbations that are imperceptible to the human eye, thus ensuring that the generated explanations remain relatively stable. The specific evaluation procedure is as follows. First, multiple perturbed versions are generated by applying Gaussian noise to the original image, which is mathematically expressed as follows:

$$I_{pert} = clip(I + \alpha \cdot N, 0, 1), \tag{20}$$

where I denotes the original image, N represents noise following a Gaussian distribution $N \sim \mathcal{N}(0,\sigma^2)$, and α regulates the noise intensity.

Subsequently, LIME explanations were generated for both the original image and each perturbed image [39]. To quantify the stability of the explanations, the structural similarity index (SSIM) was calculated between the original explanation heatmap X and each perturbed explanation heatmap Y. The SSIM formula is defined as follows:

$$SSIM(X,Y) = \frac{(2\mu_X\mu_Y + C_1)(2\sigma_{XY} + C_2)}{(\mu_X^2 + \mu_Y^2 + C_1)(\sigma_X^2 + \sigma_Y^2 + C_2)}, \tag{21}$$

where μ_X and μ_Y denote the means of X and Y, respectively, σ_X^2 and σ_Y^2 represent the variances of X and Y, respectively, σ_{XY} is the covariance between X and Y, $C_1 = (0.01 \cdot L)^2$ and $C_2 = (0.03 \cdot L)^2$ are stability constants, and L is the dynamic range of pixel values.

Finally, after conducting N perturbation experiments, the average SSIM was computed as the stability score for the method:

$$\text{Stability} = \frac{1}{N}\sum_{i=1}^{N} SSIM(M_{original}, M_{perturbed_i}), \tag{22}$$

where $M_{original}$ denotes the original explanation, and $M_{perturbed_i}$ represents the explanation from the i-th perturbation [40].

The quantitative evaluation results are shown in Figure 18.

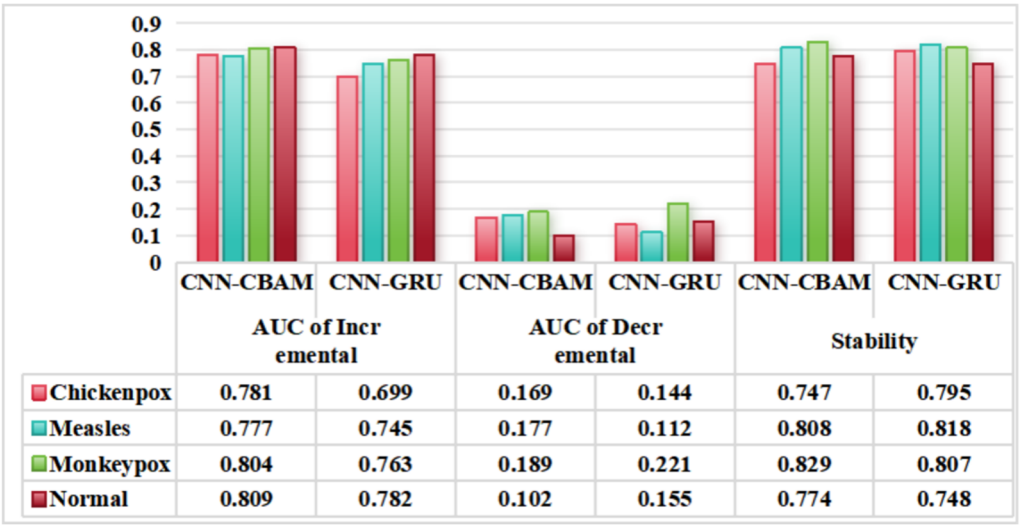


Figure 18. Quantitative comparison of LIME interpretability for hybrid models.

Figure 18 presents a quantitative comparison of LIME explainability between the CNN-CBAM and CNN-GRU models. Overall, each model demonstrated distinct advantages across different evaluation metrics. For the Chickenpox, Measles, Monkeypox, and Normal categories, CNN-CBAM generally exhibited a higher AUC of the incremental values. For instance, it reached 0.781 in Chickenpox, 0.777 in Measles, 0.804 in Monkeypox, and 0.809 in Normal, which are generally better than the corresponding results of CNN-GRU. This indicates that CNN-CBAM has a relatively stronger capability in identifying important features. For the AUC of the Decremental metric, a lower value

means the model has less dependence on irrelevant features. CNN-GRU achieved 0.144 and 0.112 for Chickenpox and Measles, respectively, which are lower than 0.169 and 0.177 of CNN-CBAM, thus showing better performances. However, in the Monkeypox and Normal categories, the AUC of Decremental of CNN-CBAM were 0.189 and 0.102, respectively, thus outperforming 0.221 and 0.155 of CNN-GRU. This shows that the sensitivity of the models to noisy features varies across different categories. Regarding the Stability metric, a higher value indicates more stable explanation results of the model. In Chickenpox and Measles, CNN-GRU had a higher stability, with 0.795 and 0.818, respectively. In contrast, CNN-CBAM performed better for Monkeypox and Normal, with 0.829 and 0.774, respectively. This suggests that the two models each excel in explanation consistency, which may be affected by the characteristics of the datasets.

In summary, CNN-GRU demonstrates better robustness against irrelevant features and higher stability in certain categories, while CNN-CBAM generally excels in identifying critical features across most datasets.

5.3. Gradient-weighted class activation mapping

Grad-CAM is a visualization technique designed to interpret the decisions made by convolutional neural networks. The core idea involves computing the gradient of the output for a target class with respect to the feature map of the final convolutional layer. This process generates a heatmap for the target class. The heatmap highlights the regions of the input image that contribute most significantly to the model's prediction. As a result, it provides an intuitive explanation of the model's decision-making process. The heatmap [41] can be generated through the following steps:

1) Compute the gradient of the target class c with respect to the feature map A^k from the final convolutional layer: $\frac{\partial y^c}{\partial A^k}$;

2) Calculate the weight coefficients for the feature map: $\alpha_k^c = \frac{1}{Z}\sum_i\sum_j \frac{\partial y^c}{\partial A_{i,j}^k}$;

3) Generate the Grad-CAM activation map: $L_{Grad-CAM}^c = \mathrm{ReLU}(\sum_k \alpha_k^c A^k)$,

where y^c represents the model's output score for the target class c, A^k denotes the k-th feature map of the last convolutional layer, $\frac{\partial y^c}{\partial A_{i,j}^k}$ represents the gradient of the class score y^c with respect to the activation $A_{i,j}^k$, and Z is the normalization constant denoting the spatial dimensions of the feature map.

The comparative evaluation results of Grad-CAM interpretability between the CNN-CBAM model and the CNN-GRU model are shown in Figure 19.

Overall, the high-attention regions in the CNN-GRU model heatmaps were relatively concentrated and cover a small area. This indicates the model primarily focuses on a localized small part of the image for classification decisions, thus showing limited overall coverage and attention to lesion regions in the image. This may lead to the model "neglecting" certain rash areas, thus making it difficult to fully understand which rash features the model uses to identify lesion types. In contrast, the high-attention regions in the CNN-CBAM model heatmaps cover a broader area, and the gradient of colors suggests the model assigns some attention to more regions in the image. This means the model can focus on more lesion areas in the image, utilizing the distribution of lesions and other features more comprehensively for classification. This aids in a clearer understanding of the model's decision

basis. In the overlay images of the CNN-GRU model, since the high-attention regions cover a relatively small area, the correspondence between the model's focused areas and the actual lesion regions in the original image is not tight. Many rash areas are not covered by the high-attention regions in the overlay images, which makes it hard to intuitively determine which specific rashes the model uses for its decisions. For the CNN-CBAM model, the broader coverage of high-attention regions in the heatmaps resulted in a tighter correspondence between the model's focused areas and the lesion regions in the original image when overlaid. This alignment better matches the human perception of lesion distribution characteristics, thus providing more intuitive interpretability. In summary, the CNN-CBAM model's Grad-CAM heatmaps more comprehensively focus on lesion-related areas in the image, and its overlay images better correspond the model's attention to the lesion features in the original image. Therefore, the Grad-CAM interpretability analysis for the CNN-CBAM model is superior.

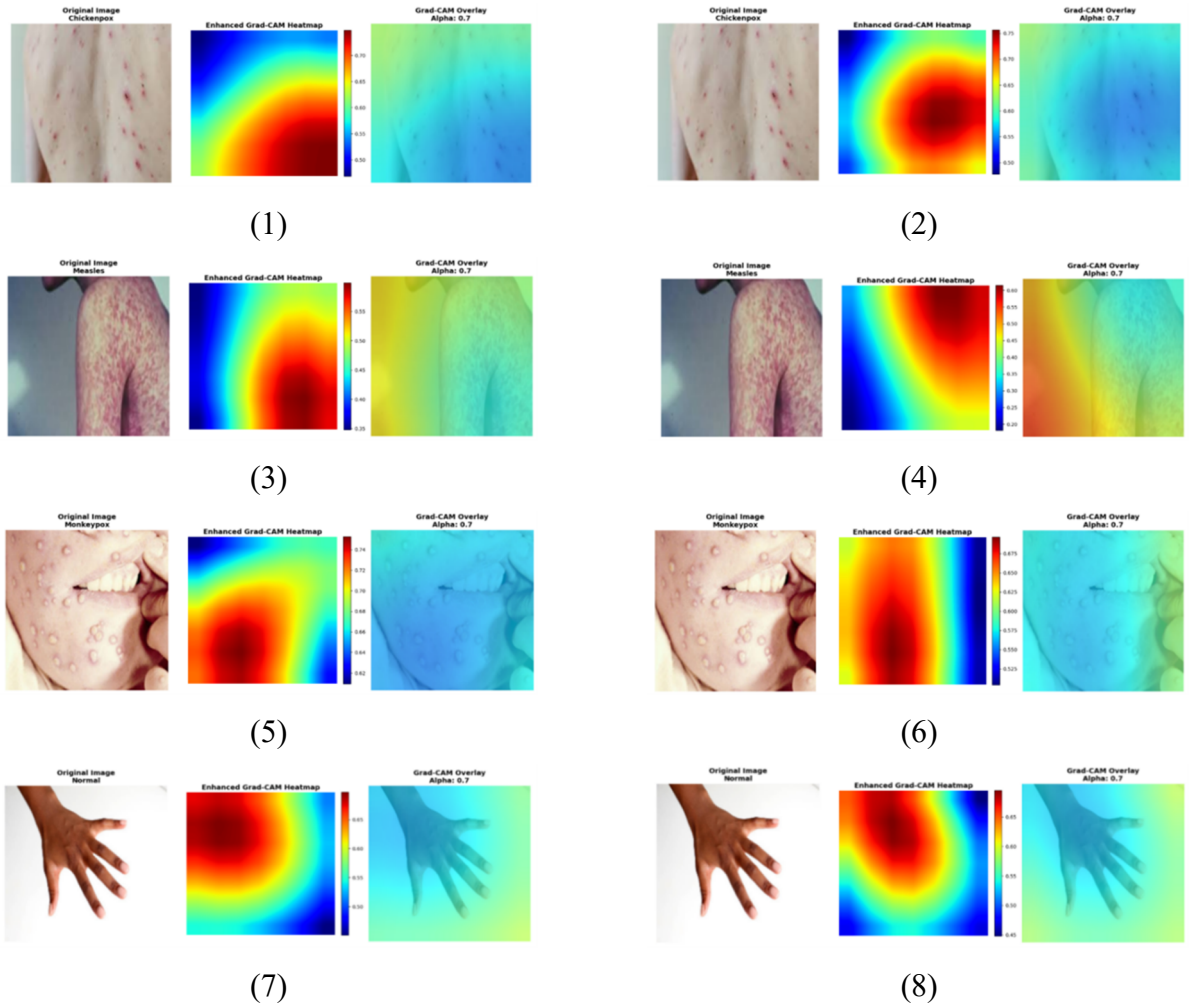


Figure 19. Grad-CAM interpretability analysis map. Figures (1), (3), (5), and (7) present interpretability analysis for the CNN-CBAM model, while Figures (2), (4), (6), and (8) present interpretability analysis for the CNN-GRU model. Each subplot consists of three components: the original image, the enhanced Grad-CAM heatmap, and the Grad-CAM overlay. The enhanced Grad-CAM heatmap uses color to indicate the model's attention to different regions of the image.The Grad-CAM overlay is the result of overlaying the heatmap onto the original image with an alpha value of 0.7, resulting in 70% transparency.

5.4. Quantitative evaluation of grad-CAM interpretability

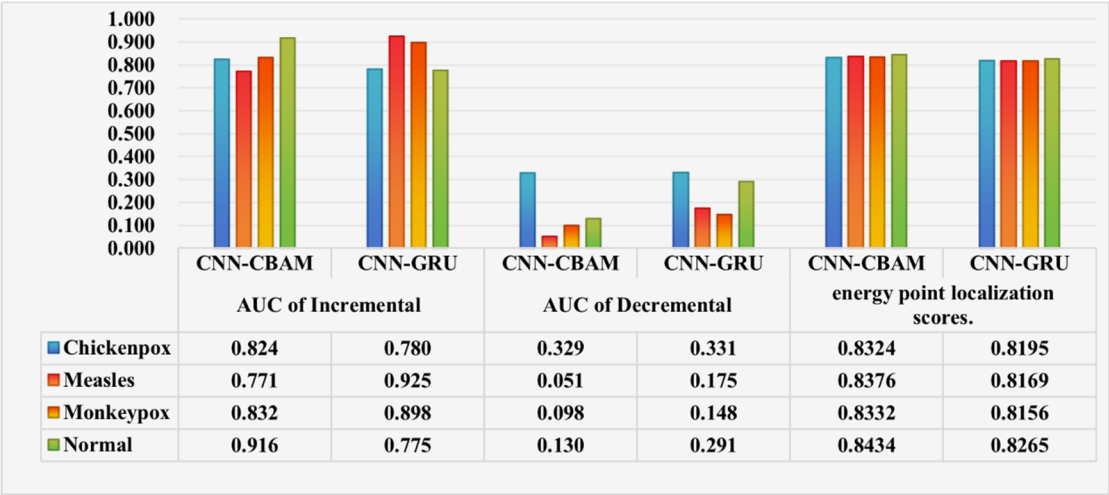
For the quantitative evaluation of Grad-CAM interpretability, we employed insertion and deletion curves and the point localization game as core metrics to achieve a quantitative measure of the accuracy of heatmaps. The definitions of the incremental and decremental curves are consistent with the previous text.

5.4.1. Energy point localization game

The energy point localization game evaluates the localization capability of Grad-CAM explanations by measuring the extent to which the heatmap focuses the "importance" on the target object [42]. This metric verifies whether the explanation genuinely focuses on key regions consistent with human annotations [43]. The energy point localization score is as follows:

$$E_{loc} = \frac{\sum_{(i,j)\in B} M(i,j)}{\sum_{(i,j)\in I} M(i,j)}, \tag{23}$$

where $M(i,j)$ represents the Grad-CAM heatmap value at position (i,j), B denotes the bounding box region of the target object, I refers to the entire image region, $\sum_{(i,j)\in B} M(i,j)$ is the total heatmap energy within the bounding box, and $\sum_{(i,j)\in I} M(i,j)$ is the total heatmap energy of the entire image.



	CNN-CBAM	CNN-GRU	CNN-CBAM	CNN-GRU	CNN-CBAM	CNN-GRU
	AUC of Incremental		AUC of Decremental		energy point localization scores.	
Chickenpox	0.824	0.780	0.329	0.331	0.8324	0.8195
Measles	0.771	0.925	0.051	0.175	0.8376	0.8169
Monkeypox	0.832	0.898	0.098	0.148	0.8332	0.8156
Normal	0.916	0.775	0.130	0.291	0.8434	0.8265

Figure 20. Quantitative comparison chart of Grad-CAM for CNN-CBAM and CNN-GRU.

Figure 20 presents the quantitative evaluation results of Grad-CAM, this study systematically analyzed the interpretability performance of the two models across three dimensions: incremental curve AUC, decremental curve AUC, and energy point localization scores. For the AUC of the incremental metric, where higher values indicate better performance, the CNN-CBAM model generally outperformed CNN-GRU in most categories. CNN-CBAM achieved higher AUC increase values in the chickenpox and normal categories, reaching 0.824 and 0.916, respectively, which are clearly superior to the corresponding values of 0.780 and 0.775 obtained by CNN-GRU. This indicates

that the high-response regions generated by Grad-CAM in CNN-CBAM more effectively contribute to the model's prediction outcomes in these categories, thus reflecting a stronger ability to identify key lesion-related features. Although CNN-GRU attained a higher value of 0.925 in the measles category compared with 0.771 for CNN-CBAM, the overall trend suggests that CNN-CBAM demonstrates a more stable and generally superior feature capture capability across most categories. In terms of the AUC of the Decremental metric, CNN-CBAM achieved lower AUC reduction values in the measles, monkeypox, and normal categories, with values of 0.051, 0.098, and 0.130, respectively, which are markedly better than those obtained by CNN-GRU at 0.175, 0.148, and 0.291, respectively. This finding indicates that CNN-CBAM is able to more precisely focus its attention on regions that are genuinely relevant to the classification decision in these categories, thereby reducing the influence of noise or background regions. In the chickenpox category, the AUC reduction values of the two models were very similar, with CNN-CBAM achieving 0.329 and CNN-GRU achieving 0.331. The small difference suggests that both models exhibit a comparable sensitivity to irrelevant features in this category. Finally, for the energy point localization scores, CNN-CBAM consistently maintained a lead across all four disease categories, with scores generally approximately 0.01 to 0.02 higher than those of CNN-GRU. For example, in the monkeypox category, it achieved 0.8332, thus reflecting that the explainability results of the CNN-CBAM model exhibit higher consistency and reliability.

6. Discussion

The experimental results provide strong evidence for the effectiveness of the proposed attention-enhanced hybrid deep learning framework for monkeypox skin lesion classification. Among all evaluated models, the CNN-CBAM architecture consistently outperformed both single-backbone convolutional neural networks and the CNN-GRU hybrid model across multiple performance metrics. This performance improvement is mainly attributable to the complementary feature extraction achieved through the parallel integration of DenseNet-121 and EfficientNet-B4, together with the adaptive feature refinement enabled by the CBAM module.

DenseNet-121, through dense connectivity and efficient feature reuse, facilitates the extraction of fine-grained texture information that is essential to visually distinguish similar dermatological lesions. In contrast, EfficientNet-B4 adopts a compound scaling strategy across network depth, width, and input resolution, thus allowing it to learn robust and discriminative representations at multiple spatial scales. The parallel integration of these two backbones enables effective fusion of multi-level and multi-scale information, resulting in more stable and discriminative feature embeddings. On this basis, the CBAM module further enhances the representation quality by dynamically emphasizing diagnostically relevant channels and spatial regions while suppressing irrelevant background information, thus improving the performance on images with complex textures and heterogeneous backgrounds.

The comparison between CNN-CBAM and CNN-GRU provides valuable empirical insights into feature aggregation strategies for static medical image analyses. Although the GRU module was explored as a gated feature aggregation mechanism, experimental results indicated that recurrent modeling did not yield stable or significant performance gains for static skin lesion images that lacked genuine temporal or longitudinal structure. In contrast, the attention-based CNN-CBAM achieved a superior classification performance using a more concise and computationally efficient architecture. This finding suggests that spatial and channel attention mechanisms are better suited than recurrent units to enhance the feature discriminability in image-based monkeypox diagnoses and offers practical

guidance for future model design.

The generalization capability is a critical requirement for clinical deployment. To evaluate robustness under different data distributions, the CNN-CBAM framework was further validated on the MSLD binary classification dataset and the more challenging MSLD v2.0 six-class dataset. Despite substantial differences in the class composition, imaging conditions, and disease categories, the model consistently achieved classification accuracies above 90%. Although a moderate performance decline was observed in the six-class task due to an increased inter-class similarity, the overall results demonstrated a strong adaptability and stability across varying levels of classification complexity and data heterogeneity.

The robustness of these performance improvements was further supported by comprehensive statistical analyses. Experiments conducted with multiple random seeds, bootstrap-based confidence interval estimation, and paired statistical testing collectively indicated that the performance advantage of CNN-CBAM over strong baseline models was not attributable to specific random initializations or a single data split. These analyses strengthen the credibility of the experimental findings and provide more reliable evidence for the reported improvements.

In terms of model positioning, this work differs from recent related studies in both design philosophy and evaluation focus. For example, the study by Hossain and colleagues emphasized validating the model reliability across multiple datasets using comprehensive deep learning frameworks and interpretability analyses, while the work by Shateri and collaborators integrated nature-inspired optimization algorithms with explainable models to improve the classification performance. In contrast, the present study did not introduce new backbone architectures or optimization techniques. Instead, it adopted a controlled hybrid architectural design to systematically analyze the relationships among attention mechanisms, feature fusion strategies, and model robustness under unified experimental conditions. In addition, this work extended the interpretability evaluation by incorporating quantitative metrics to assess the consistency and reliability of Grad-CAM and LIME explanations, thus contributing further perspectives on the objectivity and reproducibility of explainability analyses.

Model interpretability is a fundamental prerequisite to establish clinical trust in medical artificial intelligence systems. Analyses based on Grad-CAM and LIME showed that CNN-CBAM consistently focused on clinically meaningful lesion regions and produced more coherent and stable attention distributions than CNN-GRU. Combined with quantitative evaluations such as insertion and deletion curves, stability analyses, and localization metrics, the generated explanations demonstrated a high consistency and reliability across different categories and samples, thereby enhancing the transparency of the model's decision-making process.

Despite these encouraging results, the present study is limited by the scale and diversity of existing public monkeypox skin lesion datasets. Future work will focus on expanding the dataset coverage in terms of skin tone diversity, geographic representation, and imaging devices. Additionally, further research will explore the integration of multimodal clinical information, as well as advanced attention mechanisms and self-supervised learning strategies, to improve robustness under data-limited conditions. Prospective validation studies and clinical deployment evaluations will be conducted to assess the practical utility and generalizability of the proposed framework in real-world healthcare settings.

7. Conclusions

This study presents an attention-enhanced hybrid deep learning framework for automated Mpox skin lesion classification. By combining dual-backbone feature fusion with attention-based refinement, the proposed approach achieved robust performance across multiple public datasets. Importantly, this work demonstrated that architectural design choices and systematic interpretability evaluation play a more critical role than introducing increasingly complex modules. The findings of this study demonstrated that, in resource-constrained medical image analysis, the adoption of a controlled hybrid architecture design coupled with systematic generalization and interpretability evaluation constitutes an effective approach to developing clinically trustworthy models.

Use of AI tools declaration

The authors declare they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

The authors declare there are no conflicts of interest.

References

1. S. Parker, R. M. Buller, A review of experimental and natural infections of animals with monkeypox virus between 1958 and 2012, *Future Virol.*, **8** (2013), 129–157. https://doi.org/10.2217/fvl.12.130
2. J. G. Breman, Kalisa-Ruti, M. V. Steniowski, E. Zanotto, A. I. Gromyko, I. Arita, Human monkeypox, 1970-79, *Bull. World Health Organ.*, **58** (1980), 165–182.
3. *World Health Organization, Diagnostic testing for the monkeypox virus (MPXV): interim guidance*, Report of Hong Kong SARS Expert Committee, 2024. Available from: https://www.who.int/zh/publications/i/item/WHO-MPX-Laboratory-2024.1.
4. E. Petersen, A. Kantele, M. Koopmans, D. Asogun, A. Yinka-Ogunleye, C. Ihekweazu, et al., Human monkeypox: epidemiologic and clinical characteristics, diagnosis, and prevention, *Infect. Dis. Clin. North Am.*, **33** (2019), 1027–1043. https://doi.org/10.1016/ j.idc.2019.03.001
5. J. Egger, C. Gsaxner, A. Pepe, K. L. Pomykala, F. Jonske, M. Kurz, et al., Medical deep learning—a systematic meta-review, *Comput. Methods Programs Biomed.*, **221** (2022), 106874. https://doi.org/10.1016/j.cmpb.2022.106874
6. N. Tsiknakis, D. Theodoropoulos, G. Manikis, E. Ktistakis, O. Boutsora, A. Berto, et al., Deep learning for diabetic retinopathy detection and classification based on fundus images: a review, *Comput. Biol. Med.*, **135** (2021), 104599. https://doi.org/ 10.1016/ j.compbiomed.2021.104599
7. R. Javed, T. Abbas, A. H. Khan, A. Daud, A. Bukhari, R. Alharbey, et al., Deep learning for lungs cancer detection: a review, *Artif. Intell. Rev.*, **57** (2024), 197. https://doi.org/10.1007/s10462 -024-10807-1
8. O. S. Kareem, A. M. Abdulazee, D. Q. Zeebaree, Skin lesions classification using deep learning techniques: review, *Asian J. Res. Comput. Sci.*, **9** (2021), 1–22. https://doi.org/10.9734/ajrcos/2021/ v9i130210

9. M. M. Ahsan, M. R. Uddin, M. S. Ali, M. K. Islam, M. Farjana, A. N. Sakib, et al., Deep transfer learning approaches for Monkeypox disease diagnosis, *Expert Syst. Appl.*, **216** (2023), 119483. https://doi.org/10.1016/j.eswa.2022.119483
10. M. F. Almufareh, S. Tehsin, M. Humayun, S. Kausar, A transfer learning approach for clinical detection support of monkeypox skin lesions, *Diagnostics*, **13** (2023), 1503. https://doi.org/10.3390/diagnostics13081503
11. M. Altun, H. Gürüler, O. Özkaraca, F. Khan, J. Khan, Y. Lee, Monkeypox detection using CNN with transfer learning, *Sensors*, **23** (2023), 1783. https://doi.org/10.3390 /s23041783
12. N. Dahiya, Y. K. Sharma, U. Rani, S. Hussain, K.V. Nabilal, A. Mohan, et al., Hyper-parameter tuned deep learning approach for effective human monkeypox disease detection, *Sci. Rep.*, **13** (2023), 15930. https://doi.org/10.1038/s41598-023-43236-1
13. C. Sitaula, T. B. Shahi, Monkeypox virus detection using pre-trained deep learning-based approaches, *J. Med. Syst.*, **46** (2022), 78. https://doi.org/10.1007/s10916-022-01868-2
14. F. Uysal, Detection of monkeypox disease from human skin images with a hybrid deep learning model, *Diagnostics*, **13** (2023), 1772. https://doi.org/10.3390/ diagnostics 13101772
15. A.D. Raha, M. Gain, R. Debnath, A. Adhikary, Y. Qiao, M. M. Hassan, et al., Attention to monkeypox: an interpretable monkeypox detection technique using attention mechanism, *IEEE Access*, **12** (2024), 51942–51965. https://doi.org/10.1109/ACCESS.2024.3385099
16. J. Sun, B. Yuan, Z. Sun, J. Zhu, Y. Deng, Y. Gong, et al., MpoxNet: dual-branchdeep residual squeeze and excitation monkeypox classification network with attention mechanism, *Front. Cell. Infect. Microbiol.*, **14** (2024), 1397316. https://doi.org/10.3389/ fcimb.2024.1397316
17. S. H. Khan, R. Iqbal, RS-FME-SwinT: a novel feature map enhancement framework integrating customized SwinT with residual and spatial CNN for monkeypox diagnosis, preprint, arXiv:2410.01216. https://doi.org/10.48550/arXiv.2410.01216
18. J. Deng, J. Liu, C. Kong, B. Zang, Y. Hu, M. Zou, Using novel deep learning models for rapid and efficient assistance in monkeypox screening from skin images, *Front. Med.*, **11** (2024), 1443812. https://doi.org/10.3389/fmed.2024.1443812
19. A. Shateri, N. Nourani, M. Dorrigiv, H. Nasiri, An explainable nature-inspired framework for monkeypox diagnosis: Xception features combined with NGBoost andAfrican vultures optimization algorithm, preprint, arXiv:2504.17540. https://doi.org/ 10.48550/arXiv.2504.17540
20. M. S. Hossain, M. Ahmed, M. S. Rahman, From survey to solution: a deep learning framework for reliable monkeypox diagnosis using skin images, *Array*, **28** (2025), 100554. https://doi.org/10.1016/j.array.2025.100554
21. W. Chen, K. Yang, Z. Yu, Y. Shi, C. P. Chen, A survey on imbalanced learning: latest research, applications and future directions, *Artif. Intell. Rev.*, **57** (2024), 137. https://doi.org/10.1007/s10462-024-10759-6
22. D. Bala, M. S. Hossain, M. A. Hossain, M. I. Abdullah, M. M. Rahman, B. Manavalan, et al., MonkeyNet: A robust deep convolutional neural network for monkeypox disease detection and classification, *Neural Networks*, **161** (2023), 757–775. https://doi.org/10.1016/j.neunet.2023.02.022
23. G. E. Batista, A. L. Bazzan, M. C. Monard, Balancing training data for automated annotation of keywords: a case study, in *WOB*, (2003), 10–18.
24. S. Rao, P. Poojary, J. Somaiya, P. Mahajan, A comparative study between various preprocessing techniques for machine learning, *Int. J. Eng. Appl. Sci. Technol.*, **5** (2020), 2455–2143.

25. W. Rawat, Z. Wang, Deep convolutional neural networks for image classification: a comprehensive review, *Neural Comput.*, **29** (2017), 2352–2449. https://doi.org/10.1162/neco_a_00990
26. A. Howard, M. Sandler, G. Chu, L. C. Chen, B. Chen, M. Tan, et al., Searching for mobilenetv3, in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, (2019), 1314–1324.
27. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2016), 770–778.
28. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, Z. Wojna, Rethinking the inception architecture for computer vision, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2016), 2818–2826.
29. F. Chollet, Xception: deep learning with depthwise separable convolutions, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, (2017), 1251–1258.
30. M. Tan, Q. Le, Efficientnet: rethinking model scaling for convolutional neural networks, in *International Conference on Machine Learning*, **97** (2019), 6105–6114. https://doi.org/10.48550/arXiv.1905.11946
31. G. Huang, Z. Liu, L. V. Der Maaten, K. Q. Weinberger, Densely connected convolutional networks, in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, (2017), 2261–2269. https://doi.org/10.1109/CVPR.2017.243
32. K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, et al., Learning phrase representations using RNN encoder-decoder for statistical machine translation, preprint, arXiv:1406.1078. https://doi.org/10.48550/arXiv.1406.1078
33. S. Woo, J. Park, J. Y. Lee, I. S. Kweon, CBAM: convolutional block attention module, in *Proceedings of the European Conference on Computer Vision (ECCV)*, **11211** (2018), 3–19. https://doi.org/10.1007/978-3-030-01234-2_1
34. Monkeypox Skin Lesion Dataset. Available from: https://www.kaggle.com/datasets/nafin59/monkeypox-skin-lesion-dataset.
35. Mpox Skin Lesion Dataset Version 2.0 (MSLD v2.0). Available from: https://www.kaggle.com/datasets/joydippaul/Mpox-skin-lesion-dataset-version-20-msld-v20.
36. M. T. Ribeiro, S. Singh, C. Guestrin, "Why should I trust you?" Explaining the predictions of any classifier, in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, (2016), 1135–1144. https://doi.org/10.1145/2939672.2939778
37. V. Petsiuk, A. Das, K. Saenko, RISE: randomized input sampling for explanation of black-box models, preprint, arXiv:1806.07421. https://doi.org/10.48550/arXiv.1806.07421
38. W. Samek, A. Binder, G. Montavon, S. Lapuschkin, K. R. Müller, Evaluating the visualization of what a deep neural network has learned, *IEEE Trans. Neural Networks Learn. Syst.*, **28** (2016), 2660–2673. https://doi.org/10.1109/TNNLS.2016.2599820
39. D. A. Melis, T. Jaakkola, Towards robust interpretability with self-explaining neural networks, *Adv. Neural Inf. Process. Syst.*, **31** (2018). https://doi.org/10.48550/arXiv.1806.07538
40. A. Ghorbani, A. Abid, J. Zou, Interpretation of neural networks is fragile, in *Proceedings of the AAAI Conference on Artificial Intelligence*, **33** (2019), 3681–3688. https://doi.org/10.1609/aaai.v33i01.33013681
41. R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-CAM: visual explanations from deep networks via gradient-based localization, in *2017 IEEE International Conference on Computer Vision (ICCV)*, (2017), 618–626. https://doi.org/10.1109/ICCV.2017.74

42. B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, (2016), 2921–2929. https://doi.org/10.1109/CVPR.2016.319
43. J. Zhang, S. A. Bargal, Z. Lin, X. Shen, J. Brandt, S. Sclaroff, Top-down neural attention by excitation backprop, *Int. J. Comput. Vis.*, **126** (2018), 1084–1102. https://doi.org/10.1007/ s11263-017-1059-x

AIMS Press

©2026 the Author(s), licensee AIMS Press. This is an open access article distributed under the terms of the Creative Commons Attribution License (https://creativecommons.org/licenses/by/4.0)