BMC Infectious Diseases

RESEARCH Open Access

Mpox-XDE: an ensemble model utilizing deep CNN and explainable AI for monkeypox detection and classification

Dip Kumar Saha[1], Sadman Rafi[2], M. F. Mridha[2*], Sultan Alfarhood[3], Mejdl Safran[3*], Md Mohsin Kabir[4] and Nilanjan Dey[5]

Abstract

The daily surge in cases in many nations has made the growing number of human monkeypox (Mpox) cases an important global concern. Therefore, it is imperative to identify Mpox early to prevent its spread. The majority of studies on Mpox identification have utilized deep learning (DL) models. However, research on developing a reliable method for accurately detecting Mpox in its early stages is still lacking. This study proposes an ensemble model composed of three improved DL models to more accurately classify Mpox in its early phases. We used the widely recognized Mpox Skin Images Dataset (MSID), which includes 770 images. The enhanced Swin Transformer (SwinViT), the proposed ensemble model Mpox-XDE, and three modified DL models-Xception, DenseNet201, and EfficientNetB7-were used. To generate the ensemble model, the three DL models were combined via a Softmax layer, a dense layer, a flattened layer, and a 65% dropout. Four neurons in the final layer classify the dataset into four categories: chickenpox, measles, normal, and Mpox. Lastly, a global average pooling layer is implemented to classify the actual class. The Mpox-XDE model performed exceptionally well, achieving testing accuracy, precision, recall, and F1-score of 98.70%, 98.90%, 98.80%, and 98.80%, respectively. Finally, the popular explainable artificial intelligence (XAI) technique, Gradient-weighted Class Activation Mapping (Grad-CAM), was applied to the convolutional layer of the Mpox-XDE model to generate overlaid areas that effectively highlight each illness class in the dataset. This proposed methodology will aid professionals in diagnosing Mpox early in a patient's condition.

Keywords Monkeypox, Deep learning, Mpox, Detection, Ensemble model, XAI

Introduction

Mpox, an ailment that has been historically endemic to Africa, experienced its most severe outbreak in 2022. The disease has now spread to numerous regions worldwide, posing a public health hazard [1]. Mpox is a type of viral infection disease that can be spread by significant physical interactions [2]. Additinally, it can be transmitted with tainted objects or with animals that are polluted. A skin rash or mucosal lesions that might last two to four weeks are common signs of mpox. These symptoms are often accompanied by fever, back, discomfort of low energy, headache, muscle aches, and enlarged lymph nodes [3]. These are early stage symptoms. Typically, a rash begins

*Correspondence:
M. F. Mridha
firoz.mridha@aiub.edu
Mejdl Safran
mejdl@ksu.edu.sa
[1] Department of CSE, Stamford University Bangladesh, Siddeswari, Dhaka, Bangladesh
[2] Department of CSE, American International University-Bangladesh, Kuratoli, Dhaka, Bangladesh
[3] Department of Computer Science, College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia
[4] Division of Computer Science and Software Engineering, Mälardalens University, 722 20 Västerås, Sweden
[5] Department of CSE, Techno International New Town, New Town, West Bengal, India

appearing after several days. The rash initially presents as unpleasant, flat, red papules. The elevated regions ultimately evolve into blisters, subsequently filled with fluid. The blisters solidify and desquamate between two and four weeks [4]. The signs of Mpox do not manifest in all individual infected with the disease [5]. There is a cause for concern for global public health as of November 2023, with 116 countries reporting 92,783 illnesses and 171 deaths [6]. Now it is getting more serious day by day. Owing to the rarity of Mpox, skilled medical personnel may initially believe that rash illnesses similar to measles or chickenpox are the reason. However, enlarged lymph nodes are typically able to distinguish between other forms of pox and Mpox [7]. To diagnose Mpox, medical professionals obtain a tissue sample from a patient who is actively infected. After that, the samples are forwarded to laboratories for polymerase chain reaction (PCR) analysis [8]. It is well known that this type of test is costly and takes some time to obtain findings. There is currently no vaccine against viruses that can treat Mpox sufficiently to recover humans [9]. Therefore, the development of a proper system that can detect the mpox with an artificial intelligence (AI) system is urgently needed. There are several different works on this case. To identify Mpox, a paper proposed a DL MiniGoggleNet model [10]. Additionally, another study proposed new fractional-order SEIR model [11] in response to growing outbreaks and increasing numbers of Mpox cases. Additionally, for skin diseases, a database known as "Mpox skin lesion dataset (MSLD)", comprising 228 different types of skin lesions in total was used [12].

Numerous research studies published in peer-reviewed journals have demonstrated the remarkable efficacy of artificially intelligent (AI) methodologies through the detection of Mpox models. For example, convolutional neural networks (CNNs) have recently gained popularity in detection and classification tasks as a result of their dependence on automated feature extraction techniques. Transformer-based models, such as SwinVit and DL models, including Xception, DenseNet201, and EfficientNetB7, have recently demonstrated exceptional performance in computer vision tasks, which is a result of the extensive training data they require. Additionally, XAI techniques, such as Grad-Cam have been used. A dearth of research exists to devise a more precise and effective algorithm for the identification of Mpox in the MSID dataset. Owing to the lack of research, these datasets pose serious and difficult classification problems. Our proposed Mpox-XDE method has been thoroughly tested on this MSID dataset to improve Mpox detection via the ensemble model.

Several studies have used statistical examination in conjunction with machine learning, deep learning and image processing techniques to extract data regarding skin diseases, specially Mpox conditions [13, 14]. A research proposes to enhance Real-ESRGAN for retinal fundus image super-resolution by incorporating a structural segmentation map of a pre-trained U-Net model with a structural prior, preserving retinal vessel details. Experimental results confirm that the improved approach reduces structural distortions and outperforms other approaches visually, rendering retinal fundus images clearer and more resolution-enhanced for enhanced clinical diagnosis [15]. A paper proposed a novel hybrid model, namely NIMEQ-SACNet, in the context of diabetic retinopathy, with the integration of an optimization algorithm to a capsule network with Self-Attention. The introduced model significantly improved the performance in binary, 5 stage and 7 stage classifications of VTDR. Integrating quantum computing methodologies into Binary Grey Wolf Optimization and using the proposed optimizer for fine-tuning the parameters of SACNet resulted in excellence depicted through many parameters [16]. By utilizing gene expression profiling from TCGA, a current study incorporates WGCNA, Lasso, and machine learning algorithms for precise diagnosis and staging of colon cancer. RF model performed well with a high classification accuracy rate of 99.81% and an F1 score of 0.9968 for cancer diagnosis. Additionally, it uncovers eight prognosis genes for a 73.04% recall rate for colon cancer staging, which have potential markers of targeted therapy as well as early detection [17]. To improve biomarker selection and phenotype prediction and overcome the issues of high noise and low reproducibility in genomic data, a new self-paced learning absolute network-based logistic regression model (SLNL) combining feature network information with a self-paced learning process was suggested in another study. Several experimental datasets show that SLNL selects fewer but more informative biomarkers with improved interpretability while achieving superior or competitive prediction performance over six state-of-the-art competitors [18]. In order to successfully remove noise while maintaining crucial ECG properties, this work suggests an advanced ECG denoising technique that combines S-transform (ST), bi-dimensional empirical mode decomposition (BEMD), and non-local means (NLM) filtering. The suggested method works better than the current wavelet-based and NLM approaches, attaining higher SNR and SSIM while reducing RMSE and PRD, as shown by experimental findings on the MIT-BIH arrhythmia database. This makes it ideal for processing ECG signals [19]. For the classification of breast cancer, a hybrid Extreme Learning Machine model combining a transfer learning and a special optimizer was suggested. For the normal, benign, and malignant classes, the method reached exceptional accuracy rates of 96.54%, 97.24%, and 98.01%, respectively. An important step in improving patient outcomes and early breast cancer identification was the combination of image data, deep feature extraction, and enhanced ELM classification [20]. In another work, introduces a novel cluster center transformer that enhances dental plaque segmentation by grouping pixels of similar intensity and texture so that transformers can better capture local contours and edges in low-contrast images. Experimental comparisons on the dental plaque dataset show that the proposed method achieves state-of-the-art performance with IoU and pixel accuracy rates of 60.91% and 76.81%, respectively, significantly improving segmentation accuracy in unconstrained environments [21].

A classification problem in artificial intelligence is the detection of the Mpox virus via pictures of skin. AI has been growing significantly in the medical field and clinical setting. Using machine learning (ML) and DL algorithms to create intelligent, automated AI-based solutions can be advantageous for the medical industry [22, 23]. Researchers have dedicated a great deal of effort to creating increasingly complicated frameworks that can be used for a wide range of image-related applications. CNNs are thought to be the most advanced technique for analysing visual perception and are presently accessible. Many models, such as Xception [24], DenseNet201 [25], EfficientNetB7 [26], and RestNet [27] are used for this factor. However, it can be a great form of work in the case of an ensemble model. Applying these models to Mpox detection could yield competitive or even higher performance compared with traditional CNN-based techniques. Particularly helpful for handling the fine features found in datasets of photos with Mpox conditions is their scalability to large datasets and high resolutions. Therefore, for more precise and clear visualizations, more work is needed. The main motivation of our study is to create an ensemble model thst is built with three main base models: Xception, DenseNet201 and EfficientNetB7. We also use the transformer model SwinVit in this study. Making an ensemble model by using those models, we have undertaken substantial research with our proposed system to better anticipate Mpox on the MSID dataset.

For the classification of Mpox, this study developed an ensemble model consisting of three modified DL models. We employed an XAI approach to successfully detect the affected areas of Mpox, creating a more robust and adaptive model. For the widely known Mpox dataset, three enhanced DL models-Xception, DenseNet201, and EfficientNetB7-the improved transformer SwinViT, and the proposed ensemble model Mpox-XDE were utilized. To build the ensemble model, we combined the three DL models using a dense layer, a Softmax layer, a flattened layer and a 65% dropout rate. In the final layer, four neurons were used to classify the dataset into four categories. The Softmax activation function was employed for multiclass classification. A single model may occasionally result in overfitting and distorted outcomes. To get around these issues, the average TL model uses grouped weights. Three TL models were chosen for the ensemble model out of the four that were put into practice. The multichannel data vectors are flattened using the flattened layer. Before being sent to the following layer, the data is transformed into a 1-D array. Following the addition of the dense layer with 512 neurons and the activation function relu, the final layer uses 4 neurons to classify each of the four disease groups. The softmax activation function was implemented in the model's last layer because of the multi-class detection issue. Lastly, a global average pooling layer was used to classify the actual class. Additionally, the superimposed areas, generated by Grad-CAM, effectively highlighted each illness class across both datasets. This popular XAI method, Grad-CAM, will assist specialists in determining the affected areas of a patient's skin. The following are the research contributions of this paper:

1. We propose a state-of-the-art ensemble method for Mpox classification, called Mpox-XDE, which consists of three enhanced DL models.
2. To generate the ensemble model, the three DL models were combined via a Softmax layer, a dense layer, a flattened layer, and a 65% dropout rate.
3. The advanced XAI method, Grad-CAM, is used in the convolutional layers of the Mpox-XDE model to detect areas affected by Mpox.

A review of the proposed literature is introduced in Related works section, and the methodology is presented in Methods and materials section. Methods and materials section is separated into six subsections: image data preprocessing, data splitting, DL models and the proposed method, hyperparameter tuning for the proposed ensemble MPox-XDE model, the architecture of the SwinViT model, and the formulation of the XAI method (Grad-CAM). The experimental results and evaluation, which are similarly split down into seven sections, are thoroughly analyzed in Experimental evaluation and results section: specification of the environment setup, dataset description, performance evaluation metrics, results analysis, output analysis via Grad-CAM, comparative study with previous research, and analysis of computational complexity. Finally, Conclusion and future work section ends the study with suggestions for future work and limitations.

Related works

Researchers endeavored to develop a system capable of rapidly and accurately identifying several dermatological conditions, specially Mpox. This section covers current research that classified and detected Mpox using CNNs and XAI techniques. Sitaula et al. [22] evaluated 13 pretrained DL models fro recognizing the Mpox virus. The top-performing models were combined and a highest accuracy of 87.13% was achieved. Additionally, Raha et al. [28] introduced an attention-based MobileNetV2 model for Mpox detection optimized for deployment on edge devices. The model makes use of spatial and channel attention strategies to increase accuracy. To enhance early-stage diagnosis, they included a wider spectrum of skin illnesses comparable to MSID. Two XAI techniques were applied to provide interpretability and insights into the diagnostic logic of the model. The model outperformed the baseline models Transfer learning using pretrained CNN frameworks was applied to identify Mpox from skin lesion images with high accuracy. Among them, the ResNet50V2-based model achieved the highest accuracy of 98.74%, which evidences the potential of DL methods for early detection of Mpox. The proposed models hold great promise, especially in resource-constrained settings, and there are plans for deploying the DL-based model as a web application for greater accessibility [29].

The Mpox Skin Images Dataset (MSID) was initially presented by Bala et al. [30]. It is a publicly accessible set of skin pictures from Mpox cases that are intended to aid in the creation and evaluation of DL models for the early detection of Mpox. On the basis of a modified DenseNet-201 architecture, they developed an innovative deep convolutional neural network called MonkeyNet, which achieved 93.19% and and 98.91% diagnostic accuracy with original dataset and augmented dataset respectively. Grad-CAM was also used in this study to emphasize the affected skin areas, improve model interpretability and aid medical personnel in detecting and enhancing early illness diagnosis more accurately, which may have pandemic consequences. DL, in particular the architecture of DenseNet, has been applied to several manners for detecting and classifying lumpy skin disease virus (LSDV). The proposed model achieved a final accuracy of 99.11% on the increased dataset and 94.23% on the original dataset by incorporating CBAM and SA. This paper brought to the fore the capability of DL with regard to the challenge by the scarcity of publicly available LSDV datasets and provided a strong solution for early detection and classification [31]. Lin et al. [32] outlined a novel macrophage vesicle-based bionic self-adjuvanting monkeypox vaccine by co-mixing MV-related intracellular proteins and EV antigen (B6R), which significantly enhanced antigen presentation and adaptive immunity. Experimental evidence shows that this approach boosts the activation of antigen-presenting cells almost fourfold compared to single-antigen approaches, shows potent immune-protective efficacy in mouse models, and presents a promising path for pre-clinical monkeypox vaccine design. In another study, Azar et al. [33] developed a DenseNet201-based deep neural network model for identifying Mpox from skin images. The model gained accuracy for 97.63% of the scenarios and better F1-scores than those of previous studies. The use of the LIME and Grad-CAM techniques increases the model's transparency and interpretability, giving information about the process of decision making and helping clinicians understand the basis of the diagnosis. Pal et al. [34] reported the use of DL for accurate Mpox detection from skin lesion images. The study compared several DL models by using a public dataset for training as well as testing. Among them, Inception v3 achieved the highest classification accuracy at 96.56%. Ozsahin et al. [35] explored a computer-aided DL framework for detecting and classifying Mpox and chickenpox skin abnormalities in patients. They employed a two-dimensional CNN to analyse digital skin images. Their proposed CNN model achieved a test accuracy of 99.60% that outperforming DL models for skin lesion detection. This study was highlighted the model's potential for rapid and accurate Mpox detection. Additionally, another study by Pramanik et al. [36] suggested a combined ML approach for identifying Mpox from skin images. It uses three established models that have been adjusted to work with a particular Mpox dataset. When tested on a publicly available Mpox dataset with fivefold cross-validation, the framework gained 93.39% accuracy. This method shows the effectiveness of combining multiple models to improve detection accuracy.

In addition to XAI, DL and other techniques have been applied to image detection and classification, particularly in Mpox and related diseases. Using skin lesion images, Nayak et al. [37] explored DL techniques for Mpox diagnosis via five pretrained neural networks. With a validation accuracy exceeding 95% for all the changed models, ResNet-18 earned the maximum accuracy of 99.49%. These findings suggest that optimized DL models, such as the ResNet-18, could be effectively deployed on performance-limited devices such as smartphones, which would provide vital tools in the fight against the Mpox virus. Healthcare practitioners benefit from the improved interpretability of the forecasts made possible by the incorporation of the LIME and Grad-CAM approaches. In the study of Ahsan's et al. [38], a modified VGG16 model was proposed to diagnose Mpox, achieving an accuracy of 97.18% in two studies. Use of Local LIME further enhances the comprehensibility, providing deeper insights into the features associated with Mpox detection. To classify diseases Ali et al. [39] developed the Mpox Skin Lesion Dataset (MSLD), which includes skin abnormalities photos of Mpox, chickenpox, and measles obtained from various online sources; employed pretrained several DL models, and developed an ensemble of those models. Among them, ResNet50 achieved the maximum accuracy of 82.96% among all of the models. A prototype web tool was created for online Mpox screening. Although the initial results are promising, a broader and more varied dataset is needed to improve the model and its accuracy. Yang et al. [40] introduced the AICOM-MP, an AI-based Mpox detection tool developed under an self-sufficient mobile medical facility initiative, aimed at enabling healthcare access in nations with the least amount of development via low-end mobile devices, which focuses on minimizing gender, racial, and age biases, performing accurate binary classification with minimal computing power, and delivering reliable results regardless of image quality or background. The tool has demonstrated top-tier performance and is available as a web service, with its source code and dataset open-sourced for integration by health AI professionals.

Almufareh et al. [41] recommended a noninvasive, computer vision-based approach that uses DL algorithms to analyse skin lesion photos to diagnose Mpox virus. The model was tested via MSID and MSLD datasets. Regarding sensitivity, specificity, and balanced accuracy, the proposed approach showed positive outcomes, indicating that it could find broader use, especially in impoverished areas with limited laboratory equipment. In another study performed by Uysal et al. [42], data augmentation and preprocessing techniques were employed. Several DL models were employed for initial Mpox identification. The two top-performing DL models were combined with an LSTM model to build a hybrid model that improved the classification results. This model achieved 87% test accuracy. Additionally, the study of Kundu et al. [43] presented a federated learning (FL) framework using MobileNetV2, Vision Transformer (ViT), and ResNet50 DL models, which addresses accurate diagnosing issues of Mpox by enabling secure data sharing and classification. The ViT-B32 model achieved a 97.90% accuracy rate, highlighting the framework's potential for effective and secure Mpox detection. Conversely, Jaradat et al. [44] identified the finest DL model for detecting Mpox with 98.16% accuracy. Testing with various datasets confirmed its effectiveness, achieving a top accuracy of 94%. Demir et al. [45] proposed an automated detection model using a new dataset of 910 images across five categories which are healthy, Mpox, chickenpox, smallpox and zoster zona. The model uses different techniques, and a special classifier. It achieved a 91.87% classification accuracy, demonstrating potential for rapid Mpox detection. In addition, Gupta et al. [46] developed a blockchain-based framework for the early detection and classification of Mpox via transfer learning. The framework was tested on a dataset of 1,905 images, gaining a classification accuracy of 98.80% via models such as Xception, VGG19, and VGG16. The effectiveness of the model suggests its potential for diagnosing other skin diseases, like measles and chickenpox. The framework, SkinMarkNet, utilizing transfer learning models were proposed for the classification of Mpox lesions. With data augmentation strategies to address the scarcity of annotated data, the model was able to achieve 90.615% accuracy. This thus proved that advanced DL architecture combined with data augmentation strategies might improve early diagnosis and intervention in clinical settings [47]. The research performed by Dahiya et al. [48] introduced a DL model that uses the Yolov5 method for detecting human Mpox accurately. By utilizing convolutional neural networks, transfer learning, and hyperparameter optimization, the model gained a classification accuracy of 98.18% for Mpox skin lesions.

Despite all the major progress in Mpox detection using different DL models, the most common limitation with many related works is the absence of efficient ensemble methods to yield higher performances. Besides this, while interpretability techniques like Grad-CAM and LIME are employed, their application has been confined to individual models, further reducing the reliability of insights across diverse conditions. In addition, most of these studies are done in isolation and may be limited to less generalizability across broader and more complex clinical scenarios. This research will cover those gaps by presenting Mpox-XDE, an ensemble of different enhanced DL models with improved interpretability by Grad-CAM, developing a more comprehensive diagnostic framework that improves both classification accuracy and clinical decisions.

Methods and materials

This section outlines our introduced method for accurately detecting and classifying the MSID dataset's four classes. Figure 1 illustrates the process's primary steps. Our methodology compares the performance of the suggested ensemble of DL models for Mpox image classification with that of popular CNN models via an XAI technique. For this investigation, we used SwinViT as the transformer model and evaluated its performance against other DL models, including Xception, DenseNet201, and EfficientNetB7. Before training the models using these architectures, preprocessing and customized model layers were performed on the dataset to optimize the models' performance. These actions are required to classify the MSID dataset via a reliable and precise approach.

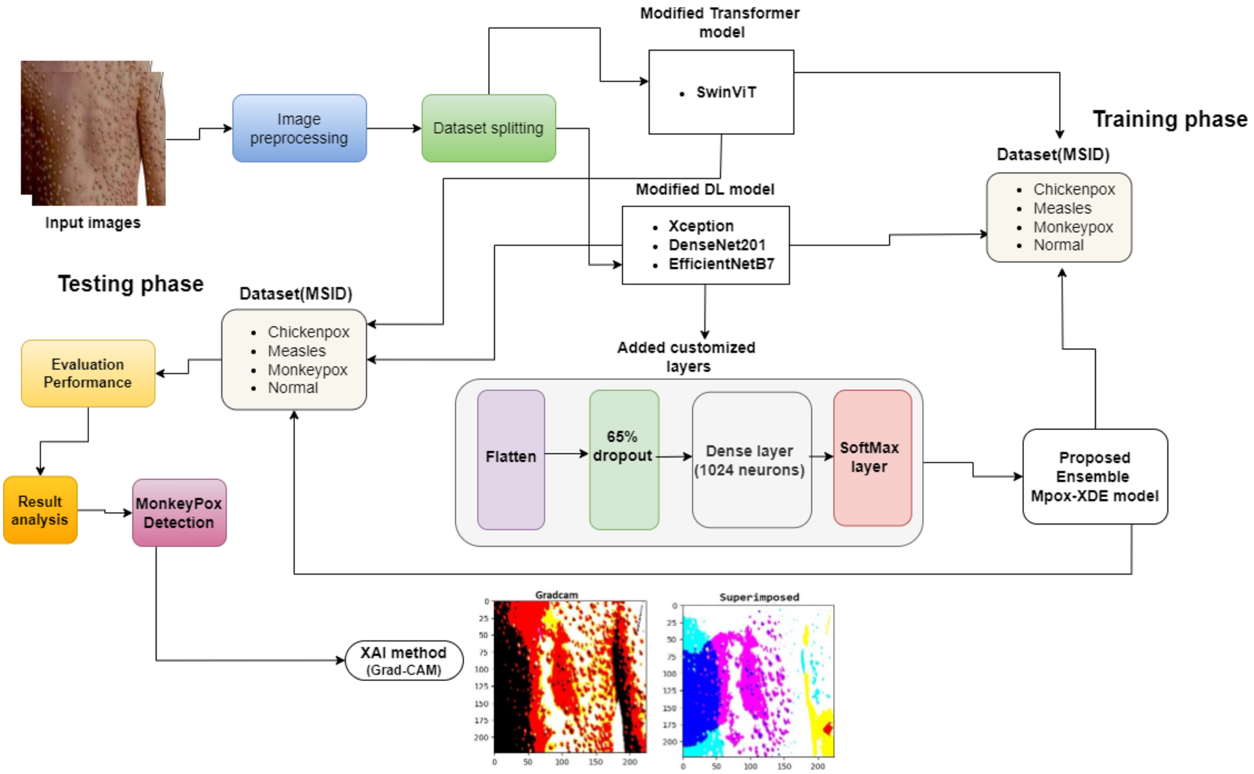

Fig. 1 The suggested methodology's architecture involves the following steps: **a** Loading the input images. **b** Image preprocessing and dataset splitting. **c** 80% of the dataset (MSID) is trained via modified SwinViT, Xception, DenseNet201, EfficientNetB7, and the proposed ensemble model namely Mpox-XDE. **d** Then all the DL models are passed into flatten layers, a dropout rate of 0.65 is calculated, and a dense layer and softmax layer are added. **e** Test 10% of data (MSID) applying transformer, DL, and proposed ensemble models. **f** Evaluate the performance matrix and analyse the results. **g** Finally, the XAI method (Grad-CAM) is applied to successfully detect the affected area of Mpox

Image data preprocessing

The preprocessing stage improves the performance of the DL models. It aids in pixel value normalization and standardization, noise reduction, and feature enhancement, all of which contribute to the model's increased applicability. Preprocessing also facilitates the identification of significant patterns in the images that can be used to reliably forecast outcomes. Additionally, it helps overcome obstacles such as class differences. There were 770 photos in this collection, however not all of them had adequate focus. Additionally, there are many different-sized photos. To ensure uniformity, we used a cropping and resizing algorithm that shrinks the image to 224 × 224 pixels while keeping any notable illness spots. This size fits all of the models that are used in the procedure.

Dataset splitting

Splitting the dataset is an essential step in maintaining its balance. Before initiating the training process, we separated the dataset (MSID) into three sets: training, validation, and testing. The percentages are as follows: 10% are used for validation, 80% are used for training, and the remaining 10% are used for testing the complete dataset. For this investigation, a total of 770 images from the popular Mpox Skin Images Dataset (MSID) dataset were used. In the training phase, we used 614 images, and in the validation and testing phases, we used 79 and 77 images. Table 1 displays the distribution of every class.

Architecture of DL models

Xception

In Xception, a deep convolutional architecture with depthwise separable convolution layers is employed. We

Table 1 Distribution of the Dataset (MSID)

Class	Training	Validation	Testing	Total
Chickenpox	85	11	11	107
Measles	72	10	09	91
Mpox	223	28	28	279
Normal	234	30	29	293
Total	**614**	**79**	**77**	**770**

propose a state-of-the-art deep CNN architecture [49] on the basis of discovery, where the inception modules use depthwise separable convolutions. Within CNN feature maps, a concept known as Xception [50] generates decoupled spatial relations and cross-channel correlations on the basis of the CNN model. When data with a 1 × 3 convolution size as input are pooled, distinct 3 × 3 convolution sizes are obtained. These convolution sizes function in different areas of the output channels before acquiring forward schemes. The Xception module is more resilient since it can fully decouple maps via cross-channel relations and spatial abilities.

$$T_l(l+1)^y(P_1,P_2) \;=\; \sum_{(d_1,d_2)} m_t^y(d_1,d_2).\, e_t^y(P_1,P_2) \tag{1}$$

$$T_l(l+2)^y \;= C_p\left(m^y_{(l+1)}.\, k_{(l+1)}\right) \tag{2}$$

It reduces the number of layers and parameters in the network, making it lighter. Below the discharge of this correlation are equations (1) and (2). Here, the feature map of the l transformation layers is denoted by T_l, and the spatial indices of the feature map f_m and the depth one kernel y are shown as (d_1, d_2) and (P_1, P_2), respectively. Spatial convalescence of kernel k is observed in feature map m. $C_p(.)$ indicates the complex technique. There are 60 convolution layers and 18 modules in the basic Xception model. Twelve of these layers are connected to a residual layer, which improves precision and speeds up the merging process. The network's core flow section manages feature development and feature optimization.

DenseNet201

Two transition blocks, three dense blocks, and an initial convolutional layer make up the DenseNet-201 architecture [51]. Remarkably effective feature extraction and representation are made possible by the densely coupled convolutional layers that comprise each dense block. By adding average pooling layers and fewer channels, transition blocks, on the other hand, make navigation between dense blocks easier.

We improved the DenseNet201 model and produced four variants by integrating deeper network layers. The enhanced block is constructed with numerous layers connected by long connections. Using stride2, we used a 7 × 7 convolution layer in the first phase. The Maxpool layer was employed before passing the Desnseblock. From start to finish, our augmented layer was composed of a dense block, batch normalization (BN), ReLU, 1 x 1 Conv, and BN, ReLU, 3 × 3 Conv. The dense network approach performs better in maximum picture classification applications and allows for internal feature reuse, which reduces the gradient demise issue rate.

EfficientNetB7

The EfficientNetB7 [52] model performs better for medical imaging tasks. To improve the accuracy and better classification model, we have used modified architecture of EfficientNetB7 model. We choose the EfficienNetB7 model and optimize it by adding more layers. The depthwise block has 3 × 3 conv, fewer parameters, and takes full advantage of modern accelerators. Figure 2 illustrates how the first block constructed conv 1 × 1 for the expansion conv 1 × 1 in MB Conv and the depthwise conv 3 × 3. After that, an average pooling layer was added to improve the model's robustness.

The EfficientNetB7 model has a complex design that consists of multiple sections, including height, weight, and resolution scaling. Balancing these parameters is a mathematical procedure that leads to precise and effective neural network topologies. EfficientNetB7 performance is based on compound scaling, which manages the number of layers and channels, and the height and width of the image resolution, where H denotes the height, W denotes the width and C represents the number of channels per layer in the convolution block. In the deptwise convolution, we have employed four channels which is improved the accuracy of this model.

Proposed ensemble method (MPox-XDE)

We propose an ensemble [53] model called Mpox-XDE, which is based on the above DL models to obtain better outcomes. This section will describe the structure of the proposed ensemble model. Our ensemble model aggregates predictions from Xception, DenseNet201, and EfficientNetB7. Combining those three models, we create an ensemble of DL models that performs exceptionally well in M-Pox classification. Figure 2 presents the architecture of the Mpox-XDE model. First, we trained the state-of-the-art ensemble model using the input data. After that, the three DL models mentioned above were used to integrate the preprocessed images. The model description above included information on every layer and convolution block. We have since added a few layers to our model, elevating it to the state of the art and improving our ability to detect MPox. A flattening layer follows the DenseNet201 transition layer and the Xception and EficientNetB7 models' average pooling layers. Following the Average pooling layers of the DL models, the flattened layer flattens the multidimensional input tensors. Prior to being sent to the following layer, the data is transformed into a 1-D array. The flattened layer is supplemented with a dropout layer that has a 65% rate. In every training cycle, this dropout layer eliminates 65% of the model's neuron connections. However, the overfitting problems here might be lessened by the global average pooling layer that was employed in the last layer.

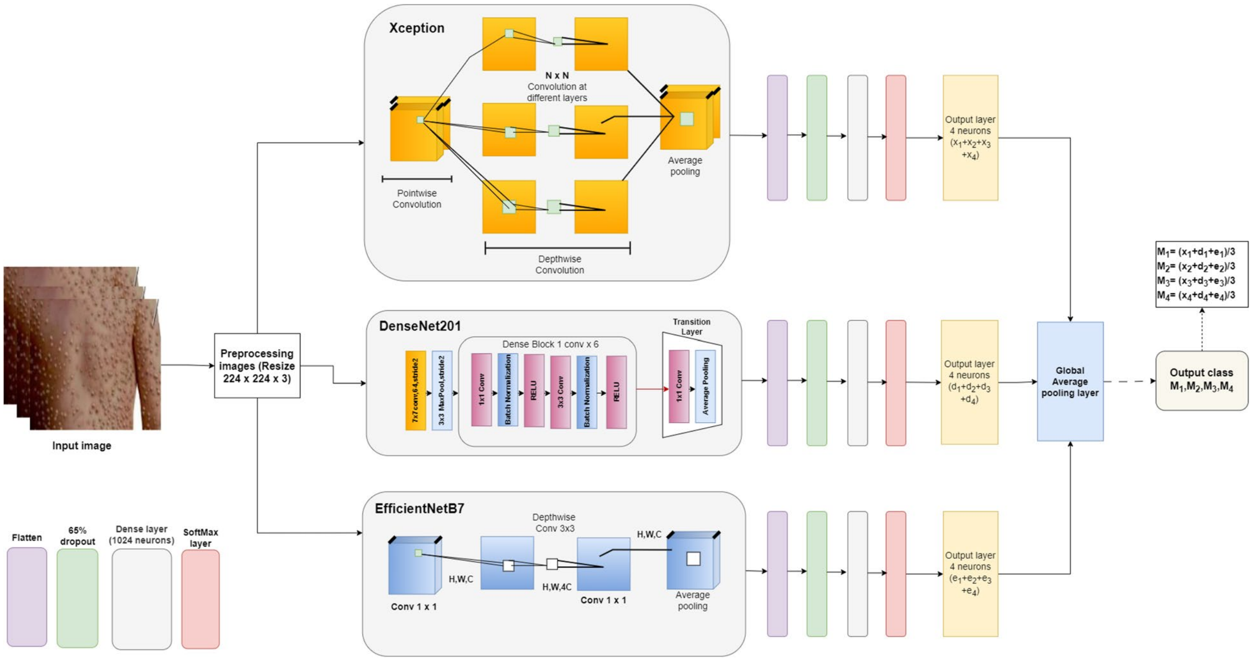
Xception
N x N
Convolution at different layers
Pointwise Convolution
Depthwise Convolution
Average pooling
Input image
Preprocessing images (Resize 224 x 224 x 3)
DenseNet201
Dense Block 1 conv x 6
Transition Layer
7x7conv,64,stride2
3x3 MaxPool,stride2
1x1 Conv
Batch Normalization
RELU
3x3 Conv
Batch Normalization
RELU
1x1 Conv
Average Pooling
EfficientNetB7
Depthwise Conv 3x3
H,W,C
H,W,4C
H,W,C
Conv 1 x 1
Conv 1 x 1
Average pooling
Flatten
65% dropout
Dense layer (1024 neurons)
SoftMax layer
Output layer 4 neurons (x_1+x_2+x_3 +x_4)
Output layer 4 neurons (d_1+d_2+d_3 +d_4)
Output layer 4 neurons (e_1+e_2+e_3 +e_4)
Global Average pooling layer
M_1= (x_1+d_1+e_1)/3
M_2= (x_2+d_2+e_2)/3
M_3= (x_3+d_3+e_3)/3
M_4= (x_4+d_4+e_4)/3
Output class M_1,M_2,M_3,M_4

Fig. 2 The architecture of the proposed ensemble model entails the following steps: **a** Load the input images **b** train the images via the Xception, DenseNet201, and EfficientNetB7 models **c** Then pass all model into CNN and flattern layers **d** Calculate the dropout of 65% of the data and evaluate the dense layer **e** Calculate the output layers **f** Pass all the models into the average pooling layer **g** Detect the disease class

Next, a dense layer is added. Since the dropout layer is supported by a dense layer of 1024 neurons and serves to reduce the ensemble model's complexity, there is no possibility of overfitting in this case. Here, the activation function used is RELU. In the final layer, four neurons are employed to classify the four classes of the dataset. Here, the Softmax activation function is used for for multiclass classification Finally, a global average pooling layer is added to predict the actual class. To average three models, we employed the global average pooling layer following the output layers of four neurons. The final layers of three models are averaged using the average pooling layer. When three models are averaged, the model's complexity increases since it has more layers, which enables better feature extraction and allows it to solve more complicated problems. These findings indicate that compared with a single model, the ensemble model yields more accurate classification and Mpox detection results.

The model becomes more sophisticated when using this pooling layer since it has more layers, which enables better feature extraction and helps it solve more complicated issues. These findings suggest that the ensemble model outperforms a single model in classification. As a consequence of this study, both experts and professionals will be able to identify Mpox from photos more precisely.

$$l = -\sum_{c=1}^{M} y_{o,c} log(p_{o,c}) \tag{3}$$

$$\Gamma(\vec{V})_i = \frac{e^{v_i}}{\sum_{j=1}^{C_l} e^{v_j}} \tag{4}$$

Equation 3 shows that the DL models use categorical cross-entropy as their loss function, with Adam as the optimizer and $2e - 4$ as the assigned learning rate. Equation 4 represents the Softmax function as *gamma*. The input vector, represented by V in a multi-class classifier, has an exponential function, which is represented by the phrase standard exponential function $v(i)$. Conversely, the output vector consists of C_l classes, where the normal exponential function is denoted by the exponential function of $v(j)$.

Hyperparameter tuning for the proposed ensemble MPox-XDE model

Our proposed Mpox-XDE model was tuned using hyperparameters, and it was applied to classify Mpox. It allows for optimizing model performance and generalization capacity. By adjusting hyperparameters such as learning rate, number of hidden layers, and dropout rate, models can optimized to increase accuracy and reduce issues such as underfitting or overfitting. We completed a classification task using a customized approach that combined the three DL models, Xception, DenseNet201, and EfficientNetB7. We adopted a strategy that required extensive model-wide parameter modifications.

We analyzed to find the best Mpox-XDE hyperparameters for the classification problem. The factors we evaluated were the kernel, learning rate, dropout, batch size, layer number, neuron size, stride, and padding. This experiment was designed to assess how different factors affect our model. Table 2 presents the Mpox-XDE model's hyperparameter-improving experiment. For the Mpox detection challenge, we employed the optimized learning rate of $2e-4$ and kernel of [1 x 1]. Along with the 32 batch size, the stride of 3 × 3, and the 0.65 dropout rate. To obtain a correct solution for the multiclass classification task, we employed the Adam optimizer.

Architecture of SwinVit

We executed the SwinViT architecture, the latest advance in computational vision that poses a challenge to CNN design. SwinViT's [54] fundamental principle is to organize the input image into nonoverlapping local windows, which are referred to as "shifted windows". The model is able to effectively encompass both local and global characteristics because of the independent handling of each window. The model is able to effectively aggregate data at various spatial resolutions as a result of the architecture's hierarchical transformer layer stack. Here, the fundamental idea behind the Swin transformer is to create so-called shifted windows-a set of local windows that don't overlap at various scales-from the input image. The model is able to obtain data that are both local and global efficiently since each window is handled independently. A detailed explanation of the SwinViT method [55] is shown in Fig. 3. The SwinViT approach was completed in this research using two steps.

Table 2 Experiment to tune hyperparameters for the proposed ensemble model(Mpox-XDE)

Parameter	Search area	Optimized value
Kernel	[1x1, 3x3, 7x7]	[1x1]
Padding	[same, valid]	same
Pool	[Max]	Max
Learning rate	$[1e-4, 1e-3, 2e-4]$	$2e-4$
Objective	[Multiclass classification]	four class
Layers_number	[9, 12, 19, 23]	12
Neurons_size	[128, 256, 512, 1024]	1024
Optimizer	[Adam, RMSprop]	Adam
Dropout rate	[0.50, 0.65]	0.65
Stride	[3 x 3, 7 x 7]	[3 x 3]
Batch size	[16, 32, 64, 128]	32

Step 1: The resolution of the actual image is initially scaled to 224 × 224 in order to accommodate the size requirements of SwinViT, which are pretrained and fine-tuned. Furthermore, starting with the initial patch size in the original method, the input image with measurements like $H \times W \times 3$ is divided into smaller sections having sizes of 4×4. This means that each image patch has a size of $4 \times 4 \times 3 = 48$. Using several SwinViT units with modified attention, this patch straight embedded of size C is passed through at the number of tokens at approximately $\frac{H}{5} \times \frac{W}{5}$ is retained.

Step 2: Every pair of neighboring two-by-two patches is concatenated with its features via the patch merging layer. A linear layer is subsequently applied to the concatenated 5C dimensional features. Consequently, the amount of patches is quadrupled, and the output depth of the linear layer is found to be 2C. Additionally, SwinViT blocks are enhanced by using the transform features, and the quantity of Stage 2 output patches is kept at $\frac{H}{10} \times \frac{W}{10}$. The formulas for computing consecutive SwinViT blocks applying the shifted window partitioning technique are given by equations (5) through (8). Here, OC^{l} and OC^{li} stand for the TCR module for block l and the W-HAN's output features respectively. Similarly, for block $l+1$, OC^{l+1} and OC^{li+1} reflect the outputs of the TCR and TD-HAN. Prior to multihead self-attention and TCR in every SwinViT block, layer normalization was applied.

$$\mathrm{OC}^{li} = WHAN(YM\,(\mathrm{OC}^{l-1})) + \mathrm{OC}^{l-1} \tag{5}$$

$$\mathrm{OC}^{l} = TCR(YM\,(\mathrm{OC}^{li})) + \mathrm{OC}^{li} \tag{6}$$

$$\mathrm{OC}^{li+1} = TDHAN(YM\,(\mathrm{OC}^{l})) + \mathrm{OC}^{l} \tag{7}$$

$$\mathrm{OC}^{l+1} = TCR(YM\,(\mathrm{OC}^{li+1})) + \mathrm{OC}^{li+1} \tag{8}$$

SwinViT blocks are enhanced by using the transform features, and the quantity of step 2 output patches is kept at $\frac{H}{10} \times \frac{W}{10}$. Using several SwinViT units with modified attention, this patch straight embedded of size C is passed through at the number of tokens at approximately $\frac{H}{5} \times \frac{W}{5}$ is retained. A straight encoding stage projects the raw-valued feature to any dimension after the RGB values of individual pixels are concatenated to form each token's feature. Each pair of neighboring patches has its features concatenated by the patch merging layer. The amount of patches is thus tripled, and the output layer of the linear phase is found to be $2C$.

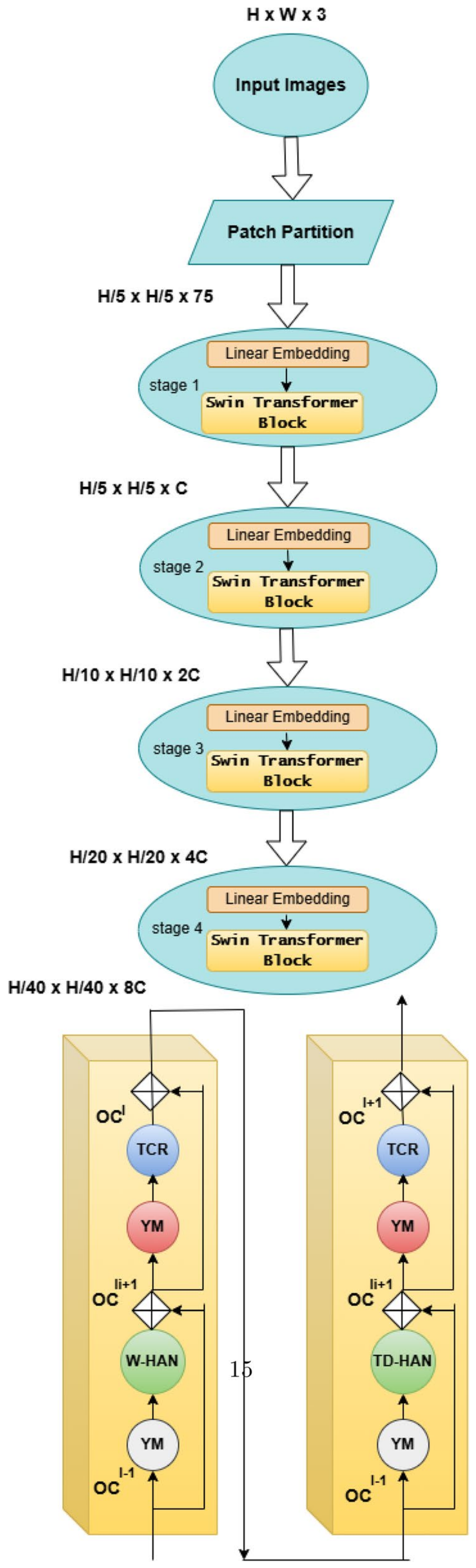

Fig. 3 Architecture of the SwinViT: Output characteristics are represented by the letters OC, channel by C, height by H, width by W, and layer normalization by YM; TCR for multilayer perceptron; Window-based and shifting window-based multi-head self-attention are referred to as W-HAN and TD-HAN, respectively

Grad-CAM

Grad-CAM uses the gradients of the proposed ensemble model Convolution layer target outputs to create a heatmap which reveals important areas for prediction-making. This method offers visual elucidations for the network's verdicts, improving the model's interpretability. The Grad-CAM algorithm involves the following steps:

1. Forward flow: A forward flow is performed over the network to retrieve the class scores.
2. Calculate Gradients: Compute the gradients of the target class score compared with the convolutional layer's feature mappings.
3. Neuron Significance Weights: Using global average pooling, extract the neuron significance weights from the gradients.
4. Produce superimposed image: By applying the ReLU function and weighting the feature maps with the neuron significance weights, a superimposed image was created.

By using F^z as the convolutional layer's z-th feature map and g^s as the final class s result prior to the Softmax layer, [56] computes the neuron significance score β_z^s. The equation follows as:

$$\beta_z^s = \frac{1}{M} \sum_p \sum_q \frac{\partial g^s}{\partial F_{pq}^z} \tag{9}$$

where M is the entire quantity of pixels in the feature map F^z. The Grad-CAM R^s for class s is then computed as [56]:

$$R^s = ReLU\left(\sum_z \beta_z^s F^z\right) \tag{10}$$

This ensures that only the characteristics that have a positive impact on the target class's score are added to the heatmap. The weighted combination of superimposed images is subjected to the ReLU function. By efficiently removing the negative values from the feature maps, this ReLU application highlights significant regions that have a favourable influence on class detection.

Grad-CAM plays a crucial role in advancing the Mpox-XDE model's interpretability by revealing visually how the model identifies and classifies Mpox and related diseases. The technique generates heatmaps overlaid on input images of skin to indicate the most significant areas responsible for the model's predictions. Grad-CAM was utilized in this research to the convolutional layers of the Mpox-XDE ensemble model, combining Xception, DenseNet201, and EfficientNetB7, to deliver an interpretable layer-wise activation map. The major aim of utilizing Grad-CAM was to confirm that the classification decision of the model for Mpox, chickenpox, measles, and normal skin was made on meaningful image features instead of noise or artifacts. The Grad-CAM approach achieves this by computing the gradients of the target class scores with respect to the feature maps of the convolutional layers and subsequently using those gradients to assign importance weights to different regions of the image to create a heatmap identifying the most salient regions that have a significant influence on the decision of the model. The result of the study established that Grad-CAM could accurately identify affected areas of skin with a high success rate, particularly for Mpox and measles cases. The superimposed heatmaps indicated that the model selectively pointed to areas of lesions instead of unnecessary background features, thus confirming the credibility of the AI-based classification Moreover, this interpretability characteristic is important to medical professionals as it helps validate the model's predictions and ensures AI-based diagnostics align with clinical observations. By offering an open visual explanation of how the decision-making process was performed, Grad-CAM not only increases confidence in automated dermatology analysis but also enables early and precise Mpox detection. This explainable AI breakthrough bridges the gap between deep learning models and the clinical field, more transparent and reliable AI-facilitated dermatological diagnosis.

Experimental evaluation and results

Specifications of the environment setup

A considerable portion of the processing power needed for image evaluation and categorization can be supplied by Graphics Computing Units (GPUs) [57]. To assist the processing activity, a GPU installation is more expensive and requires additional hardware. Consequently, Google Colab was utilized [58] platform, which offers access to strong cloud GPUs, to train our model. The configuration and parameters of the environment used for our investigation are listed in Table 3. The Google Colab comes with a v3 TPU processor, 14.3 GB of RAM, 21.9 GB of disk space, and two TensorCores. These criteria can be used to train DL models in a large-scale computer system.

Table 3 Specifications and parameters needed to carry out this study

Environment	Parameters
Platform	Google Colab
Language version	Python 3.0
Graphics card	TPU, 13.7 GB
Available RAM	14.3 GB
Disk space	21.9 GB

Dataset description

The Mpox Skin Images Dataset (MSID) [30] was utilized in this research for Mpox screening. There are 770 photos in this dataset, including those showing the chickenpox, measles, mpox, and normal classes. From this dataset, we used 77 photos for testing, 79 images for validation, and 614 images for training our models. Every image was selected as a class label, and Fig. 4 shows instances from the dataset with their corresponding labels[1].

Measures for evaluating performance

This section analyses how well the system's examination performed [59]. A classifier is employed to forecast some classifications that could be true or false. First of all, true Positive (T_p) and true Negative (T_n) methods demonstrate that all predictions are accurate, true, or untrue. Additionally, there can be a situation where the forecasting is true in concept but not in reality, or vice versa. False positive (F_p) and false negative (F_n) are the terms utilized to explain these two circumstances. Furthermore, by calculating more precise metrics from the confusion matrix (C_m), we may be able to assess the classification performance of our models.

Accuracy (A_c): This metric is described as the entirety of the test cycle and the dimension of the correctly identified samples.

$$A_c = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \tag{11}$$

Precision (P_r): The precision is stated as the number of tests and predicted outcomes that are accepted as positive.

$$P_r = \frac{T_p}{T_p + F_p} \tag{12}$$

Recall (R_c): The positive samples number that can be reliably identified and appropriately classified as positive is represented by the R_c.

[1] https://www.kaggle.com/datasets/dipuiucse/monkeypoxskinimagedataset

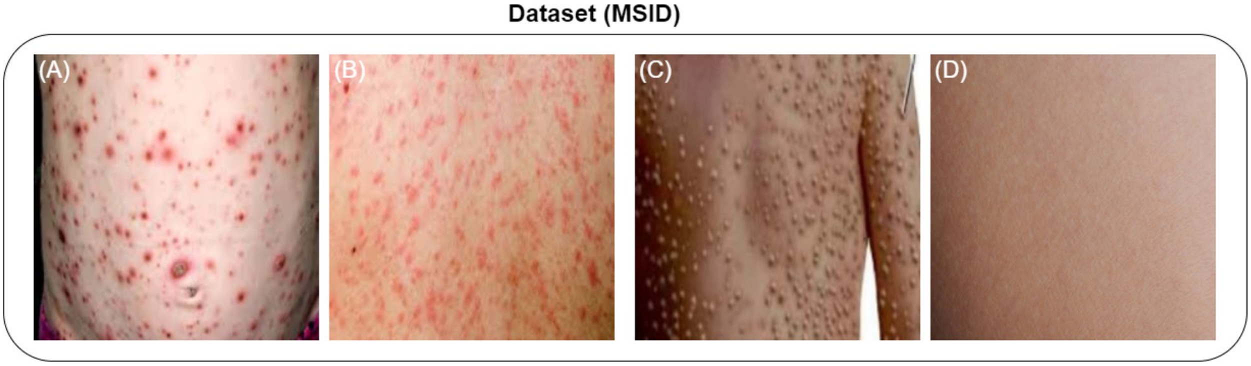

Fig. 4 Sample images from the Mpox Skin Images Dataset (MSID). **A** Chickenpox, **B** Measles, **C** Mpox, and **D** Normal

$$R_c = \frac{T_p}{T_p + F_n} \tag{13}$$

F1 score: The F1 score functions as an individual evaluation and is produced by linearly combining A_c and R_c.

$$\text{F1 score} = \frac{2 \cdot P_r \cdot R_c}{P_r + R_c} \tag{14}$$

Accuracy and the F1-score may not always be enough to assess models for predictions. Consequently, the receiver operating characteristics (ROC) curve is used as a secondary criterion in the evaluation process. The AUC is obtained from a curve called the ROC curve. The true positive rate (Tpr) and false positive rate (Fpr) were compared to determine the ROC curve. The recall is the only component of the TPR, and this formula establishes the Fpr.

$$Fpr = \frac{F_p}{F_p + T_n} \tag{15}$$

Results analysis

Here, we present the results from each of the evaluated models employed in this study. This result suggested that the Mpox-XDE design was used to assess the models' solidity. Additionally, this research used Xception, EfficientNetB7, DenseNet201, and SwinViT to assess and contrast the effectiveness of our suggestions. The Mpox-XDE performs significantly better than the other models do, particularly with respect to the M-Pox detection covered in Experimental evaluation and results section of the experimental findings. Only DL models may not adjust as well to new data and may yield biased conclusions because they are trained on pretrained data. Overfitting can be readily resolved since our MPox-XDE model is more adaptable to modifications in the data or model architecture because it uses group weights. Compared to DL models, ensemble models frequently exhibit higher accuracy. Integrating the predictions of many models can reduce individual model mistakes and improve overall accuracy, a finding supported by the results of this study. Finally, using numerous graphs and charts, we provide a performance study of the recommended architectures.

Classifier performance matrix of all the evaluated models

Each method's confusion matrix has the following data. Similarly, a table-like representation of the model's classifier performance is shown. The classification results of

Table 4 Classwise classification results for the proposed Mpox-XDE, Xception, EfficientNetB7, SwinViT, and DenseNet201

Method	Class	Precision %	Recall %	F1-score %
Proposed Mpox-XDE	Chickenpox	99.99	94.00	96.79
	Measles	99.97	99.80	99.00
	Mpox	99.99	99.95	99.75
	Normal	96.50	99.80	98.72
Xception	Chickenpox	99.67	90.00	94.00
	Measles	99.99	88.40	94.50
	Mpox	98.69	99.50	99.70
	Normal	92.70	97.00	96.25
EfficientNetB7	Chickenpox	99.77	82.45	88.33
	Measles	99.99	84.50	91.30
	Mpox	93.50	99.30	97.00
	Normal	94.40	99.50	97.00
SwinViT	Chickenpox	99.50	83.70	91.00
	Measles	99.97	99.30	99.99
	Mpox	99.99	88.77	94.40
	Normal	92.00	99.00	93.38
DenseNet201	Chickenpox	93.00	99.00	96.48
	Measles	99.76	88.00	93.99
	Mpox	99.50	84.26	92.00
	Normal	85.69	99.00	92.35

The F1-score, precision, and recall of the transformer and DL models using dataset (MSID) in the training phase as well as the evaluation metric values for each class, are displayed

Table 5 Classwise classification results for the proposed Mpox-XDE, Xception, EfficientNetB7, SwinViT, and DenseNet201

Method	Class	Precision %	Recall %	F1-score %
Proposed Mpox-XDE	Chickenpox	99.99	92.00	96.13
	Measles	99.98	99.99	99.99
	Mpox	99.99	99.98	99.99
	Normal	96.00	99.99	98.00
Xception	Chickenpox	99.99	88.00	93.00
	Measles	99.99	89.00	94.00
	Mpox	97.00	99.98	98.32
	Normal	91.00	99.97	95.28
EfficientNetB7	Chickenpox	99.98	80.25	89.20
	Measles	99.97	83.58	91.00
	Mpox	92.00	99.97	96.14
	Normal	94.00	99.96	97.43
SwinViT	Chickenpox	99.51	82.78	90.00
	Measles	99.99	99.74	99.86
	Mpox	99.99	89.52	93.87
	Normal	86.00	99.85	93.26
DenseNet201	Chickenpox	92.00	99.00	97.00
	Measles	99.98	88.49	93.21
	Mpox	99.98	83.67	91.00
	Normal	84.00	99.70	92.80

The F1-score, precision, and recall of the transformer and DL models using dataset (MSID) in the testing phase as well as the evaluation metric values for each class, are displayed

the training phase by SwinViT, DenseNet201, EfficientNetB7, and our proposed system Mpox-XDE, as shown in Table 4. The ensemble model performs remarkably well in Mpox classification, as Table 4 illustrates. For the chickenpox and Mpox classes a precision of 99.99% was achieved by the proposed ensemble model. The classification results of the testing phase by all of our DL, transformer, and our suggested system Mpox-XDE are given in Table 5. Here also, the proposed ensemble model performs remarkably well in Mpox classification. This is a notable finding in our study. In addition, Xception and EfficientNetB7 both effectively classify the Mpox illness. However, the balanced dataset is related to all the models' results throughout the training and testing stages. In detecting the Mpox class, the Mpox-XDE model has achieved a greater precision of 99.99%.

Figure 5 displays the confusion matrix of the five models when the dataset (MSID) is utilized in the testing phase. As shown in Fig. 5E, the proposed ensemble model Mpox-XDE performs better in terms of Mpox classification. Only one normal class image is misclassified as a chickenpox class image. As a result, the proposed method achieves an excellent level of classification accuracy. Mpox-XDE performed better than the transformer and all the DL variants.

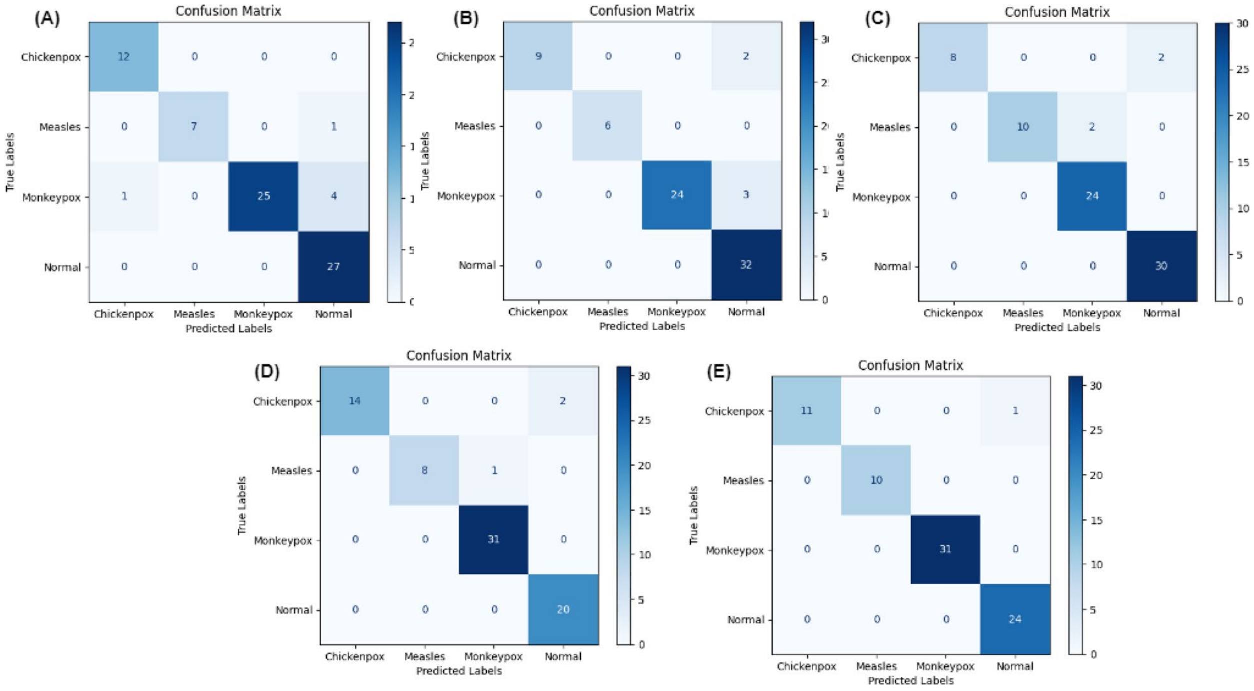

Fig. 5 Confusion matrix in the testing phase for right and wrong estimates made when using dataset (MSID) is used for the **A** DenseNet201, **B** SwinViT, **C** EfficientNetB7, **D** Xception, and **E** proposed Mpox-XDE models

Classifier parameter results

The ROC curve, which is shown in Fig. 6, is a supplementary compliance indicator for the introduced models. The highest documented micro average AUC score for the proposed ensemble model is 1.00 which is outstanding, the Xception model has the second-most nearly ROC curve. The suggested model performed exceptionally well in our suggested methodology. The detailed report using dataset (MSID) in the testing phase of classification considering the mean accuracy, recall, and F-1 score for each approach is shown in Table 6. Mpox-XDE performed better in this case than any other model we tested. A precision of 98.90%, recall of 98.80%, and F1 score of 98.80% indicated exceptional success using this method. Additionally, Xception achieves the second highest precision of 96.50% in this research. In this instance, Mpox-XDE outperforms every other model that we explored.

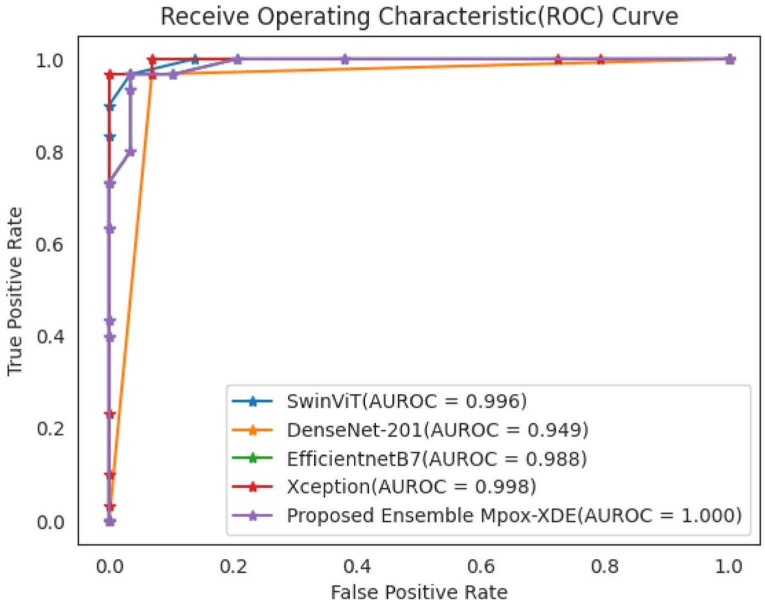

Fig. 6 ROC curves for all evaluated methods for Mpox classification in the testing phase on dataset. The microareas under the ROC curve for the Mpox-XDE, SwinViT, Xception EfficientNetB7, and MobileNetV3 presented

Table 6 Extensive report on classification using dataset (MSID) in testing phase

Method	Precision %	Recall %	F1-score %	Sensitivity %	Specificity %
Proposed Mpox-XDE	98.90	98.80	98.80	97.80	99.20
Xception	96.50	96.10	96.10	95.50	97.70
EfficientNetB7	95.70	94.70	94.90	96.40	94.70
SwinViT	94.30	93.45	93.40	95.20	93.70
DenseNet201	93.50	92.20	92.20	91.00	95.70

(The models proposed Mpox-XDE, Xception, EfficientNetB7, SwinViT and DenseNet201 were compared based on average precision, recall, and F1-score values)

Table 7 Loss and Accuracy for all evaluated models using dataset (MSID) in this research

	Mpox-XDE	Xception	EfficientNetB7	DenseNet201	SwinViT
Training accuracy %	98.90	97.60	96.00	94.00	95.80
Testing accuracy %	**98.70 (Proposed)**	96.10	94.70	92.20	93.40
Training loss %	0.29	0.75	1.48	2.40	1.83
Validation loss %	0.31	0.93	1.90	3.10	2.55

Loss and accuracy of predictions

By using the methods we suggested, this section displays the loss and accuracy for the dataset (MSID) of each of the five models. For the training, testing, and validation stages, Table 7 is presented the ideas of accuracy and loss. The Mpox-XDE model achieves higher accuracy of 98.90% and 98.70% in the training and testing sets, respectively. With a 0.31% validation loss, the Mpox-XDE model using the dataset exhibited the best testing accuracy, at 98.70%. Compared with the other methods, DenseNet201 had the lowest test accuracy of 92.20%. The proposed ensemble model produces excellent results due to the balanced dataset and optimal ensemble model used in this study.

Figure 7 represents both the loss and accuracy of the proposed ensemble model for the training phases. Figure 7A illustrates the training loss over 80 epochs. In this case, the training loss decreased as the number of epochs increased. That indicates that the model executes perfectly on the balanced dataset. Additionally, Fig. 7B presents the training accuracy for the MSID datasets on the same number of epochs. As the number of epochs increased, the accuracy of the suggested ensemble model improved. However, Fig. 7B demonstrates that the training accuracy has increased and reached higher levels after completing 45 epochs.

Dataset diversity and biasness

Although the Mpox-XDE model has worked excellently in Mpox, measles, chickenpox, and normal skin conditions classification, its accuracy is necessarily linked to the comprehensiveness of the training dataset employed. The Mpox Skin Images Dataset (MSID) with 770 images provides a foundation for the model's learning, but potentially cannot fully express the abundant richness of variation seen in real-world clinical cases. Factors such as differences in skin tones, lesion characteristics, imaging parameters, and population heterogeneity can significantly influence the model's generalization. Class imbalances in data sets are one of the main defects with medical image classification. If a model is trained on some disease or demographic classes that are poorly represented, its accuracy will be reduced, as will its dependence on predictions being unreliable. Mpox-XDE's ensemble approach reduces overfitting and enhances robustness but can still allow bias to affect clinical utility if it is derived from a sparse data set.

Output analysis of Grad-CAM

Analysing the appearance of regions in an image that a DL model focuses on can be performed with the help of the Grad-CAM analysis [60]. There is a significant achievement in generating a Grad-CAM map and superimposed area that highlights the affected areas. In this work, we employed Grad-CAM analysis to investigate the Grad-CAM maps produced by our suggested Mpox-XDE model-a combination of three enhanced DL models-for Mpox-infected skin pictures. The proposed ensemble model detects the affected area very smoothly via the Grad-CAM. This might facilitate medical personnel's

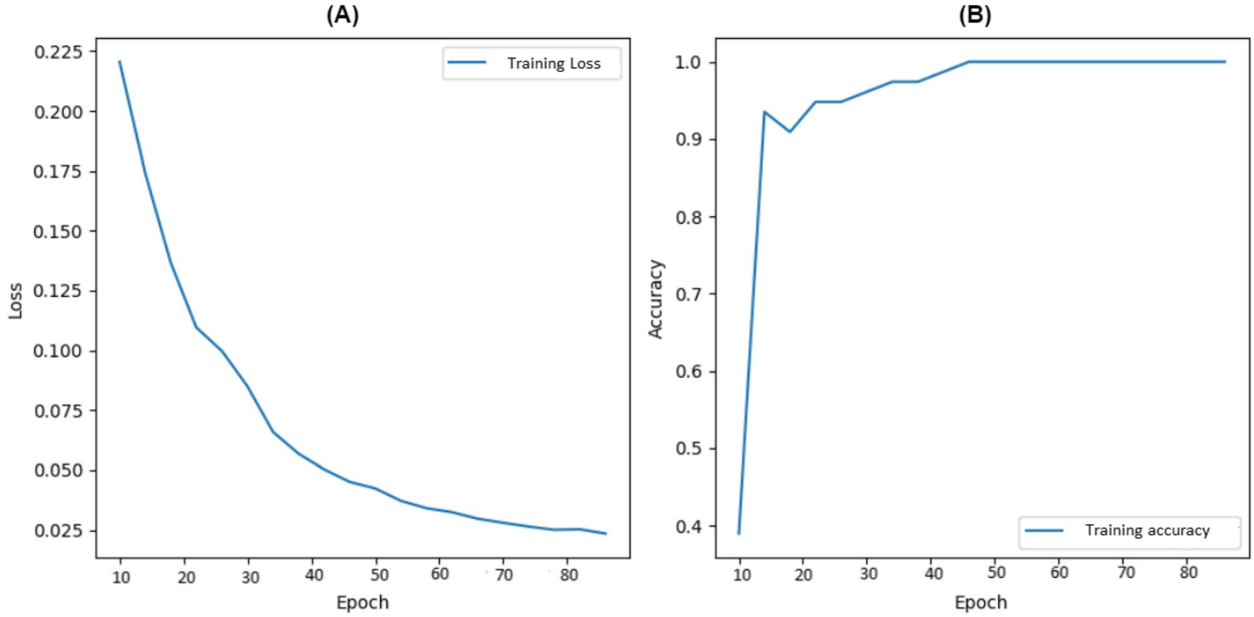

Fig. 7 Loss and accuracy for the proposed Mpox-XDE model. **A** Training loss, and **B** accuracy using Datset(MSID)

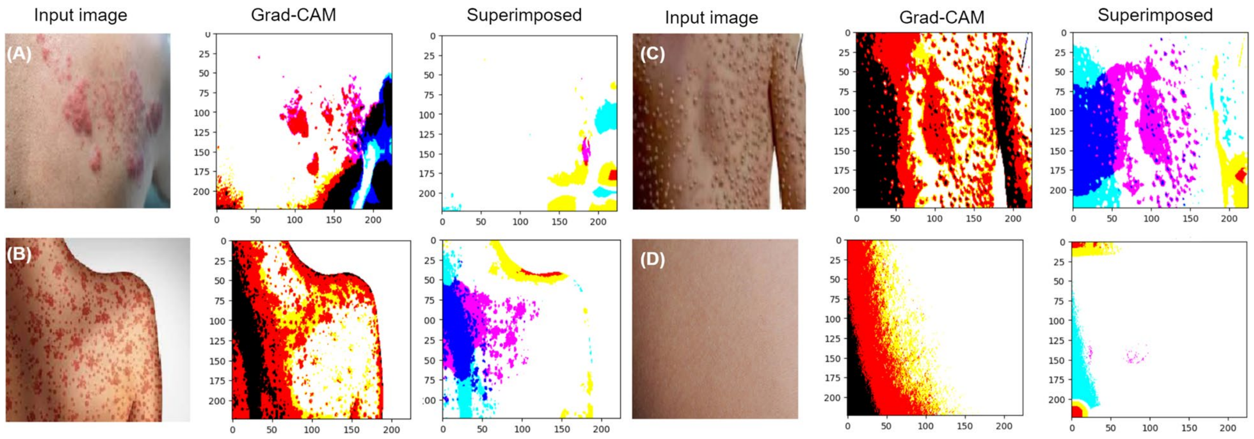

Fig. 8 Some sample images from the input images (MSID) along with their Grad-CAM and superimposed images generated by the proposed ensemble model. **A** Chickenpox, **B** Measles, **C** Mpox, and **D** Normal

easy diagnosis of the Mpox-affected area. Figure 8 shows the original input photos, their Grad-CAM, and their overlay images. Figure 8 shows that for each class our proposed ensemble model was used to detect the affected area. For the Measels and Mpox classes, the Grad-CAM showed remarkable performance. Significantly, the contaminated area of the skin pictures contaminated with Mpox is correctly identified by our proposed model. This Grad-CAM stage will be useful for the specialists in determining whether areas of the patient's skin are affected.

Table 8 also compares the most recent research on MSID dataset using XAI techniques published in several publications on the detection and classification of Mpox. In terms of accuracy, this research demonstrated that our DL-based ensemble method using the XAI technique performs better than any prior study does. The proposed Mpox-XDE model significantly contributes to our research and can be crucial in quickly detecting Mpox cases during current outbreaks.

Analysis of computational complexity

A crucial balance between computational effectiveness and model intricacy is emphasized by a comparison between the proposed Mpox-XDE and all the models investigated in this study. Compared with the other DL models, Mpox-XDE and EfficientNetB7 contained more parameters-28.7 million and 24.6 million, respectively. In addition, the transformer model has the highest number of parameters of 43.5 million. The computational complexity of each examined model is shown in Table 9. The architecture, number of layers, and settings of SwinViT, EfficientNetB7, and the suggested ensemble models require additional training time and disk space. Additionally, the Mpox-XDE models outperformed SwinViT in this investigation and used less

Table 8 The effectiveness of the proposed method was contrasted with a previous study that used XAI approaches to increase the Mpox class detection accuracy on MSID and related datasets

Previous study	Dataset	Applying XAI method (Yes/No)	Purpose	Best Method	Accuracy %
Demir et al. [45]	Private	No	Classification	MNPDenseNet	91.87%
Sitaula et al. [61]	Private	Yes	Detection	Ensemble model	87.13%
Ali et al. [44]	MSLD	No	Classification	ResNet50	82.96%
Raha et al. [28]	MSID, MSLD	Yes	Detection	MobileNetV2	98.19% (MSID)
Pramanik et al. [36]	MSLD	No	Detection	Ensemble	93.39%
Bala et al. [30]	MSID	Yes	Detection and classification	MobileNet	93.19% (original) 98.91% (augmented)
Attallah et al. [62]	MSID, MSLD	No	Detection and classification	MonDiaL-CAD	97.1% (MSID)
Asif et al. [63]	MSID, MSLD	Yes	Detection and classification	CFI-Net	94.81% (MSID)
Proposed ensemble model	**MSID**	**Yes**	**Classification and detection**	**Mpox-XDE**	**98.70% (MSID)**

Table 9 Complexity analysis of all evaluated methods

Points	Mpox-XDE	SwinViT	Xception	DenseNet201	EfficientNetB7
Parameters (millions)	28.7	43.5	22.8	13	24.6
RAM used (GB)	9.5	12.9	6.7	3.1	8
GPU used (GB)	14.4	21	10.6	6.3	12
Training time (hours)	4.8	6	4.5	2	4

processing resources. As we developed an ensemble architecture of three DL models, with other DL models, computational power and model performance are at odds. Researchers and practitioners must carefully consider the resulting compromise in light of their unique needs and limitations.

Conclusion and future work

The outbreak of Mpox throughout the world in 2022 put up a high demand for improved diagnostic tools. In this work, we have developed an ensemble model for Mpox detection, known as Mpox-XDE, combining three DL models. Those are Xception, DenseNet201, and EfficientNetB7. This work approach also employed an explainable AI tool called Grad-CAM to visualize better affected skin areas, thus presenting more interpretable results to medical experts. The proposed Mpox-XDE system makes use of the popular dataset, which is MSID. During the testing phase, the proposed ensemble method achieves impressive percentages of precision 98.90%, recall 98.80%, and f1-score 98.80%. Using the Mpox-XDE model, this study concludes with an outstanding testing accuracy of 98.70%.

Despite these promising results, some limitations exist: the dataset was relatively small and, despite preprocessing, may lead to overfitting; this in turn reduces the model's capability for generalization on new data. On the other hand, the more complex architecture of Mpox-XDE increases the time it takes for training and reduces the interpretability of how the specific layers contribute to the predictions. Future work will, therefore, be directed to address these challenges. Increasing the dataset with more diverse samples can make the model robust and more reliable. In the future, we will work on another real-time Mpox dataset to evaluate the best performance of our methodology. Simultaneously, simplifying the architecture of the model or trying some optimization techniques may reduce computational requirements with no performance loss. Refining explainable AI will lead to better transparency and trust gain among healthcare professionals. Application of this framework in the diagnosis of other diseases could contribute more to global healthcare solutions based on AI.

Acknowledgements
The authors extend their appreciation to King Saud University for funding this research through Researchers Supporting Project Number (RSPD2025R890), King Saud University, Riyadh, Saudi Arabia.

Authors' contributions
DKS Writing - original draft, Writing - review & editing, Methodology, Validation, Software, Investigation, Formal analysis, Conceptualization. SR Validation, Data curation, Investigation, Formal analysis. MFM Supervision, Project administration. SA Formal analysis, Funding acquisition, Project administration. MS and MKB Writing - review & editing, Validation. ND Project administration, Supervision.

Funding
This research was supported by the Researchers Supporting Project Number (RSPD2025R890), King Saud University, Riyadh, Saudi Arabia.

Data availability
The dataset used in this research is publicly available at https://www.kaggle.com/datasets/dipuiucse/monkeypoxskinimagedataset.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 21 November 2024 Accepted: 17 March 2025
Published online: 25 March 2025

References

1. Molla J, Sekkak I, Ortiz AM, Moyles I, Nasri B. Mathematical modeling of mpox: A scoping review. One Health. 2023;16: 100540.
2. Amer F, Khalil HE, Elahmady M, ElBadawy NE, Zahran WA, Abdelnasser M, et al. Mpox: Risks and approaches to prevention. J Infect Public Health. 2023;16(6):901–10.
3. Fatima N, Mandava K. Monkeypox-a menacing challenge or an endemic? Ann Med Surg. 2022;79: 103979.
4. Kang JH. Febrile illness with skin rashes. Infect Chemother. 2015;47(3):155–66.
5. Adalja A, Inglesby T. A novel international monkeypox outbreak. North Independence Mall West: American College of Physicians; 2022.
6. Zinnah MA, Uddin MB, Hasan T, Das S, Khatun F, Hasan MH, et al. The Re-Emergence of Mpox: Old Illness, Modern Challenges. Biomedicines. 2024;12(7):1457.
7. Koenig KL, Beÿ CK, Marty AM. Monkeypox 2022 Identify-Isolate-Inform: A 3I Tool for frontline clinicians for a zoonosis with escalating human community transmission. One Health. 2022;15: 100410.

8. Reed KD, Melski JW, Graham MB, Regnery RL, Sotir MJ, Wegner MV, et al. The detection of monkeypox in humans in the Western Hemisphere. N Engl J Med. 2004;350(4):342–50.
9. Reynolds MG, McCollum AM, Nguete B, Shongo Lushima R, Petersen BW. Improving the care and treatment of monkeypox patients in low-resource settings: applying evidence from contemporary biomedical and smallpox biodefense research. Viruses. 2017;9(12):380.
10. Alcalá-Rmz V, Villagrana-Bañuelos KE, Celaya-Padilla JM, Galván-Tejada JI, Gamboa-Rosales H, Galván-Tejada CE. Convolutional neural network for monkeypox detection. In: International conference on ubiquitous computing and ambient intelligence. Belfast: Springer; 2022. pp. 89–100.
11. Almuzini M, Batiha IM, Momani S. A study of fractional-order monkeypox mathematical model with its stability analysis. In: 2023 International Conference on Fractional Differentiation and Its Applications (ICFDA). Ajman: IEEE; 2023. pp. 1–6.
12. Ali SN, Ahmed MT, Paul J, Jahan T, Sani S, Noor N, et al. Monkeypox skin lesion detection using deep learning models: A feasibility study. 2022. arXiv preprint arXiv:2207.03342.
13. Mengistu AD, Alemayehu DM. Computer vision for skin cancer diagnosis and recognition using RBF and SOM. Int J Image Process (IJIP). 2015;9(6):311–9.
14. Islam MN, Gallardo-Alvarado J, Abu M, Salman NA, Rengan SP, Said S. Skin disease recognition using texture analysis. In: 2017 IEEE 8th control and system graduate research colloquium (ICSGRC). Shah Alam: IEEE.; 2017. pp. 144–8.
15. Jia Y, Chen G, Chi H. Retinal fundus image super-resolution based on generative adversarial network guided with vascular structure prior. Sci Rep. 2024;14(1):22786.
16. Bilal A, Liu X, Shafiq M, Ahmed Z, Long H. NIMEQ-SACNet: A novel self-attention precision medicine model for vision-threatening diabetic retinopathy using image data. Comput Biol Med. 2024;171: 108099.
17. Su Y, Tian X, Gao R, Guo W, Chen C, Chen C, et al. Colon cancer diagnosis and staging classification based on machine learning and bioinformatics analysis. Comput Biol Med. 2022;145: 105409.
18. Huang H, Wu N, Liang Y, Peng X, Shu J. SLNL: A novel method for gene selection and phenotype classification. Int J Intell Syst. 2022;37(9):6283–304.
19. Bing P, Liu W, Zhai Z, Li J, Guo Z, Xiang Y, et al. A novel approach for denoising electrocardiogram signals to detect cardiovascular diseases using an efficient hybrid scheme. Front Cardiovasc Med. 2024;11:1277123.
20. Bilal A, Imran A, Liu X, Liu X, Ahmad Z, Shafiq M, et al. BC-QNet: A quantum-infused ELM model for breast cancer diagnosis. Comput Biol Med. 2024;175: 108483.
21. Song W, Wang X, Guo Y, Li S, Xia B, Hao A. Centerformer: a novel cluster center enhanced transformer for unconstrained dental plaque segmentation. IEEE Trans Multimedia. 2024:26:10965–78.
22. Dev S, Wang H, Nwosu CS, Jain N, Veeravalli B, John D. A predictive analytics approach for stroke prediction using machine learning and neural networks. Healthc Analytics. 2022;2: 100032.
23. AlSaad R, Malluhi Q, Janahi I, Boughorbel S. Predicting emergency department utilization among children with asthma using deep learning models. Healthc Analytics. 2022;2: 100050.
24. Panthakkan A, Anzar S, Jamal S, Mansoor W. Concatenated Xception-ResNet50-A novel hybrid approach for accurate skin cancer prediction. Comput Biol Med. 2022;150: 106170.
25. Sanghvi HA, Patel RH, Agarwal A, Gupta S, Sawhney V, Pandya AS. A deep learning approach for classification of COVID and pneumonia using DenseNet-201. Int J Imaging Syst Technol. 2023;33(1):18–38.
26. Kartal MS, Polat Ö. Segmentation of skin lesions using U-Net with efficientNetB7 backbone. In: 2022 Innovations in Intelligent Systems and Applications Conference (ASYU). Antalya: IEEE; 2022. pp. 1–5.
27. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. IEEE; 2016. pp. 770–8.
28. Raha AD, Gain M, Debnath R, Adhikary A, Qiao Y, Hassan MM, et al. Attention to Monkeypox: An Interpretable Monkeypox Detection Technique Using Attention Mechanism. IEEE; 2024.
29. Thorat R, Gupta A. Transfer learning-enabled skin disease classification: the case of monkeypox detection. Multimedia Tools Appl. 2024;83:1–19.
30. Bala D, Hossain MS, Hossain MA, Abdullah MI, Rahman MM, Manavalan B, et al. MonkeyNet: A robust deep convolutional neural network for monkeypox disease detection and classification. Neural Netw. 2023;161:757–75.
31. Mujahid M, Khurshaid T, Safran M, Alfarhood S, Ashraf I. Prediction of lumpy skin disease virus using customized CBAM-DenseNet-attention model. BMC Infect Dis. 2024;24(1):1181.
32. Lin W, Shen C, Li M, Ma S, Liu C, Huang J, et al. Programmable Macrophage Vesicle Based Bionic Self-Adjuvanting Vaccine for Immunization against Monkeypox Virus. Adv Sci. 2025;12(1):2408608.
33. Sorayaie Azar A, Naemi A, Babaei Rikan S, Bagherzadeh Mohasefi J, Pirnejad H, Wiil UK. Monkeypox detection using deep neural networks. BMC Infect Dis. 2023;23(1):438.
34. Pal M, Mahal A, Mohapatra RK, Obaidullah AJ, Sahoo RN, Pattnaik G, et al. Deep and transfer learning approaches for automated early detection of monkeypox (Mpox) alongside other similar skin lesions and their classification. ACS Omega. 2023;8(35):31747–57.
35. Uzun Ozsahin D, Mustapha MT, Uzun B, Duwa B, Ozsahin I. Computer-aided detection and classification of monkeypox and chickenpox lesion in human subjects using deep learning framework. Diagnostics. 2023;13(2):292.
36. Pramanik R, Banerjee B, Efimenko G, Kaplun D, Sarkar R. Monkeypox detection from skin lesion images using an amalgamation of CNN models aided with Beta function-based normalization scheme. PLoS ONE. 2023;18(4): e0281815.
37. Nayak T, Chadaga K, Sampathila N, Mayrose H, Gokulkrishnan N, Prabhu S, et al. Deep learning based detection of monkeypox virus using skin lesion images. Med Nov Technol Devices. 2023;18: 100243.
38. Ahsan MM, Uddin MR, Farjana M, Sakib AN, Momin KA, Luna SA. Image Data collection and implementation of deep learning-based model in detecting Monkeypox disease using modified VGG16. 2022. arXiv preprint arXiv:2206.01862.
39. Ali SN, Ahmed MT, Paul J, Jahan T, Sani S, Noor N, et al. Monkeypox skin lesion detection using deep learning models: A feasibility study. 2022. arXiv preprint arXiv:2207.03342.
40. Yang T, Yang T, Liu A, An N, Liu S, Liu X. AICOM-MP: an AI-based monkeypox detector for resource-constrained environments. Connect Sci. 2024;36(1):2306962.
41. Almufareh MF, Tehsin S, Humayun M, Kausar S. A transfer learning approach for clinical detection support of monkeypox skin lesions. Diagnostics. 2023;13(8):1503.
42. Uysal F. Detection of monkeypox disease from human skin images with a hybrid deep learning model. Diagnostics. 2023;13(10):1772.
43. Kundu D, Rahman MM, Rahman A, Das D, Siddiqi UR, Alam MGR, et al. Federated deep learning for monkeypox disease detection on gan-augmented dataset. IEEE Access. 2024;12:32819–29.
44. Jaradat AS, Al Mamlook RE, Almakayeel N, Alharbe N, Almuflih AS, Nasayreh A, et al. Automated monkeypox skin lesion detection using deep learning and transfer learning techniques. Int J Environ Res Public Health. 2023;20(5):4422.
45. Demir FB, Baygin M, Tuncer I, Barua PD, Dogan S, Tuncer T, et al. MNP-DenseNet: automated monkeypox detection using multiple nested patch division and pretrained densenet201. Multimedia Tools Appl. 2024;1–23.
46. Gupta A, Bhagat M, Jain V. Blockchain-enabled healthcare monitoring system for early Monkeypox detection. J Supercomput. 2023;79(14):15675–99.
47. Akram A, Jamjoom AA, Innab N, Almujally NA, Umer M, Alsubai S, et al. SkinMarkNet: an automated approach for prediction of monkeyPox using image data augmentation with deep ensemble learning models. Multimedia Tools Appl. 2024;1–17.
48. Dahiya N, Sharma YK, Rani U, Hussain S, Nabilal KV, Mohan A, et al. Hyper-parameter tuned deep learning approach for effective human monkeypox disease detection. Sci Rep. 2023;13(1):15930.
49. Lei X, Pan H, Huang X. A dilated CNN model for image classification. IEEE Access. 2019;7:124087–95.
50. Shaheed K, Mao A, Qureshi I, Kumar M, Hussain S, Ullah I, et al. DS-CNN: A pre-trained Xception model based on depth-wise separable convolutional neural network for finger vein recognition. Expert Syst Appl. 2022;191: 116288.

51. Mofrad FB, Valizadeh G. DenseNet-based transfer learning for LV shape Classification: Introducing a novel information fusion and data augmentation using statistical Shape/Color modeling. Expert Syst Appl. 2023;213: 119261.
52. Raza R, Zulfiqar F, Khan MO, Arif M, Alvi A, Iftikhar MA, et al. Lung-EffNet: Lung cancer classification using EfficientNet from CT-scan images. Eng Appl Artif Intell. 2023;126: 106902.
53. Hossain S, Chakrabarty A, Gadekallu TR, Alazab M, Piran MJ. Vision transformers, ensemble model, and transfer learning leveraging explainable AI for brain tumor detection and classification. IEEE J Biomed Health Inform. 2023;28(3):1261–72.
54. Liu Z, Lin Y, Cao Y, Hu H, Wei Y, Zhang Z, et al. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proceedings of the IEEE/CVF international conference on computer vision. IEEE; 2021. pp. 10012–2.
55. Saha DK, Joy AM, Majumder A. YoTransViT: A transformer and CNN method for predicting and classifying skin diseases using segmentation techniques. Inform Med Unlocked. 2024;47: 101495.
56. Abhishek A, Jha RK, Sinha R, Jha K. Automated detection and classification of leukemia on a subject-independent test dataset using deep transfer learning supported by Grad-CAM visualization. Biomed Signal Process Control. 2023;83: 104722.
57. Corral JMR, Civit-Masot J, Luna-Perejón F, Díaz-Cano I, Morgado-Estévez A, Domínguez-Morales M. Energy efficiency in edge TPU vs. embedded GPU for computer-aided medical imaging segmentation and classification. Eng Appl Artif Intell. 2024;127:107298.
58. Kuroki M. Using Python and Google Colab to teach undergraduate microeconomic theory. Int Rev Econ Educ. 2021;38: 100225.
59. Saha DK. An extensive investigation of convolutional neural network designs for the diagnosis of lumpy skin disease in dairy cows. Heliyon. 2024;10(14).
60. Jahmunah V, Ng EY, Tan RS, Oh SL, Acharya UR. Explainable detection of myocardial infarction using deep learning models with Grad-CAM technique on ECG signals. Comput Biol Med. 2022;146: 105550.
61. Sitaula C, Shahi TB. Monkeypox virus detection using pre-trained deep learning-based approaches. J Med Syst. 2022;46(11):78.
62. Attallah O. MonDiaL-CAD: Monkeypox diagnosis via selected hybrid CNNs unified with feature selection and ensemble learning. Digit Health. 2023;9:20552076231180056.
63. Asif S, Zhao M, Li Y, Tang F, Zhu Y. CFI-Net: A Choquet Fuzzy Integral based Ensemble Network with PSO-Optimized Fuzzy Measures for Diagnosing Multiple Skin Diseases Including Mpox. IEEE J Biomed Health Inform. 2024;28:5573–86.

Publisher's Note