JCTA Journal of Computing Theories and Applications ISSN:3024-9104

Research Article

Feature Fusion with Albumentation for Enhancing Monkeypox Detection Using Deep Learning Models

Nizar Rafi Pratama [1], De Rosal Ignatius Moses Setiadi [1,2,*], Imanuel Harkespan [1], and Arnold Adimabua Ojugo [3]

[1] Faculty of Computer Science, Univesitas Dian Nuswantoro, Semarang 50131, Indonesia; e-mail : nizarrafipratama@gmail.com; harkespan@dsn.dinus.ac.id
[2] Research Center for Quantum Computing and Materials Informatics, Faculty of Computer Science, Dian Nuswantoro University, Semarang 50131, Indonesia; e-mail : moses@dsn.dinus.ac.id
[3] Department of Computer Science, Federal University of Petroleum Resources Effurun, Nigeria; e-mail: ojugo.arnold@fupre.edu.ng
* Corresponding Author : De Rosal Ignatius Moses Setiadi

Abstract: Monkeypox is a zoonotic disease caused by Orthopoxvirus, presenting clinical challenges due to its visual similarity to other dermatological conditions. Early and accurate detection is crucial to prevent further transmission, yet conventional diagnostic methods are often resource-intensive and time-consuming. This study proposes a deep learning-based classification model by integrating Xception and InceptionV3 using feature fusion to enhance performance in classifying Monkeypox skin lesions. Given the limited availability of annotated medical images, data augmentation was applied using Albumentation to improve model generalization. The proposed model was trained and evaluated on the Monkeypox Skin Lesion Dataset (MSLD), achieving 85.96% accuracy, 86.47% precision, 85.25% recall, 78.43% specificity, and an AUC score of 0.8931, outperforming existing methods. Notably, data augmentation significantly improved recall from 81.23% to 85.25%, demonstrating its effectiveness in enhancing sensitivity to positive cases. Ablation studies further validated that augmentation increased overall accuracy from 82.02% to 85.96%, emphasizing its role in improving model robustness. Comparative analysis with other models confirmed the superiority of our approach. This research enhances automated Monkeypox detection, offering a robust and efficient tool for low-resource clinical settings. The findings reinforce the potential of feature fusion and augmentation in improving deep learning-based medical image classification, facilitating more reliable and accessible disease identification.

Keywords: Albumentation; Feature fusion; InceptionV3; Medical image classification; Monkeypox classification; Xception.

Received: January, 21st 2025
Revised: February, 18th 2025
Accepted: February, 20th 2025
Published: February, 21st 2025

1. Introduction

Monkeypox is a rare zoonotic disease caused by an Orthopoxvirus. The virus was first discovered in laboratory monkeys in 1958 and in humans in the Democratic Republic of the Congo in 1970[1]. Initially confined to Central and West Africa, monkeypox cases have now spread to non-endemic countries, including the Americas, Europe, and Asia, prompting the WHO to declare the disease a Public Health Emergency of International Concern (PHEIC) in May 2022 and August 2024[2]. Clinical manifestations of monkeypox include fever, lymphadenopathy, and skin lesions that often resemble chickenpox or measles. This poses challenges in clinical diagnosis, especially in areas with limited diagnostic facilities[3]. Therefore, rapid and accurate identification methods are needed to prevent further spread. In this context, emerging technologies, particularly machine learning (ML) and deep learning (DL), offer promising solutions to improve the speed and accuracy of monkeypox detection.

In the field of ML and DL, with its extraordinary capabilities in medical analysis[4], it has been proven effective in diagnosing various diseases using tabular, signal, image, and multi-modal data inputs[5]–[10]. Specifically in images, DL allows models to automatically extract important features from images, resulting in higher accuracy in disease identification[11], [12]. One of the most widely used approaches is Convolutional Neural Networks

(CNN), which can recognize complex patterns in images through convolutional layers tailored for specific classification tasks[13], [14]. In addition, transfer learning techniques are gaining popularity because they can leverage models that have been previously trained on large datasets, thereby improving model performance on similar tasks while reducing computational time and cost[15].

Previous studies have demonstrated the effectiveness of CNN architectures in diagnosing various skin conditions. Gouda et al.[16] used the InceptionV3 architecture to detect and classify skin cancer on a dataset of 3,533 skin lesion images, including 1,760 benign and 1,773 malignant lesions, and achieved an accuracy of 85.7%. Another study by Erdem et al. [17] used the Xception to classify a dataset of 10,000 skin lesion images with seven different classes (dermatofibroma, vascular lesions, actinic keratoses, basal cell carcinoma, benign keratosis, melanoma, and melanocytic nevus), achieving an accuracy of 88.92%. Recent advances in DL have shown that combining multiple architectures through feature fusion can improve classification performance. Feature fusion integrates the strengths of multiple models, creating richer feature representation and higher accuracy[18], [19]. Roseline et al.[20] successfully applied this approach by combining MobileNetV2 and Xception for skin cancer classification, resulting in an accuracy of up to 97.56%.

Data augmentation is an important technique in image processing that aims to expand training data distribution and improve the generalization deep learning models. Transformations such as flipping, rotation, translation, and brightness changes allow models to learn from a wider variety, reducing the risk of overfitting. Previous studies have shown that augmentation can improve the accuracy of medical image classification, especially when the dataset is limited[23]–[25]. Albumentation is an image augmentation library designed to improve efficiency and flexibility in medical image processing. By providing a variety of adaptive transformations, this library allows for richer data manipulation than conventional methods while still preserving the essential characteristics of the original image[26]. Several studies have shown that Albumentation-based augmentation can improve model performance[27]–[29].

This study proposes a novel approach by integrating Xception and InceptionV3 through feature fusion for monkeypox skin lesion image classification. Xception is known for its efficiency in extracting features using depthwise separable convolutions, resulting in high accuracy with fewer parameters[21]. Meanwhile, InceptionV3, through its inception module, excels in capturing complex patterns of different feature sizes in images[22]. By combining these two architectures, this study aims to leverage the complementary strengths of both models, thereby improving the accuracy and effectiveness in monkeypox lesion classification. Based on the literature review above, this study contributes to:

- Xception and InceptionV3 are used as base models to extract key features from monkeypox skin lesion images.
- The concatenation method is used to create a combined feature representation of both models to strengthen the performance of the proposed model.
- Adding augmentation data with albumentation improves the proposed model's performance and generalization.
- Evaluating the performance of the proposed model on the Monkeypox skin lesion dataset (MSLD) by conducting an ablation study and comparing it with related work based on standard classification metrics.

The rest of this paper is organized as follows: Section 2 discusses related research on medical image classification and the deep learning approach used. Section 3 describes the methodology, including model architecture, data preprocessing, and augmentation strategy applied. Section 4 presents the experimental results and model performance analysis based on key evaluation metrics. Section 5 discusses an ablation study to measure the impact of augmentation and compares it with previous methods. Section 6 concludes the research results and proposes further research directions.

2. Literature Review

2.1. Xception Transfer Learning

Xception is a deep convolutional neural network architecture based on depthwise separable convolutions, a more efficient variant of the convolution operation[21]. This architecture enables higher computational efficiency and fewer parameters while still

achieving high performance on complex image recognition tasks. The Xception model has demonstrated strong performance in various computer vision tasks, including fine-grained image classification, due to its ability to learn complex spatial patterns[17]. In this study, Xception serves as the first baseline model for feature extraction, capturing various visual patterns in images. More details of the Xception architecture are presented in Table 1.

Table 1. Xception Architecture.

Layer	Description	Main function
Stem (Initial Layers)	The initial convolutional layer reduces the image dimensions (default = 299×299×3) and extracts basic features for further processing.	Prepares the image for further processing by reducing its initial dimensions.
Depthwise Separable Convolutions	Uses depthwise convolution to process each channel separately and combines the results using pointwise convolution.	Reduces the number of parameters, enabling more efficient feature processing.
Residual Connections	Connects layers by introducing shortcut paths to facilitate information flow.	Prevents information loss during network propagation.
Fully Connected (Dense)	A fully connected layer that links information from all layers and produces the final output.	Organizes the learned information and generates the final prediction.
Softmax Output	Activation function for multi-class classification that outputs class probabilities.	Produces probability values for different output classes.

The Xception architecture introduces a more efficient approach to convolution by using depthwise separable convolutions. Unlike standard convolutions, this technique separates the convolution process into two steps: a depthwise convolution that processes each input channel separately. Then the results are combined using a pointwise convolution. This approach reduces the number of parameters in the model, increasing computational efficiency without compromising the quality of the results. In addition, Xception also adopts residual connections, which allow for better information flow between deeper and earlier layers, helping to prevent the loss of important information and addressing the vanishing gradients problem that can occur in deeper networks. After the features are processed, the information is processed by a fully connected layer to produce the final prediction. Then, a Softmax activation function is used to classify the results into relevant classes.

2.2. InceptionV3 Transfer Learning

InceptionV3 is a DL model designed to optimize computational resources while maintaining high accuracy[30]. The core concept of InceptionV3 is the Inception module, which applies multiple types of convolutional filters of different sizes to the input data, allowing the model to learn features at multiple scales. This modular architecture allows InceptionV3 to capture both small- and large-scale patterns in images efficiently, making it well-suited for image classification tasks[22], [31], [32]. In this study, InceptionV3 is used with Xception to extract complementary features, which are combined through feature fusion. More details of the InceptionV3 architecture are presented in Table 2.

InceptionV3 combines several techniques to improve the efficiency and performance of the model in image processing. The architecture starts with stem layers, a series of initial convolutions that reduce the dimensionality of the image and extract basic features. One of the key features of InceptionV3 is Inception Modules, which allow the model to process different filter sizes in a single layer to capture different scales of spatial features from the image. This facilitates the model in learning features with varying complexity. In addition, factorized convolutions are used to optimize the number of parameters and reduce the computational burden, maintaining efficiency without compromising performance. To make the model more stable and generalize better, InceptionV3 also includes auxiliary classifiers that help prevent overfitting during training. At the end of the architecture, fully connected layers combine the extracted features to make predictions, and the results are processed through a Softmax activation function to assign probabilities to the output classes.

Table 2. InceptionV3 Architecture.

Layer	Deskripsi	Fungsi Utama
Stem (Initial Layers)	The initial convolutional layer is used to reduce image dimensions (default = 299×299×3) and extract basic features for further processing.	Prepares the image for further processing by reducing its initial dimensions.
Depthwise Separable Convolutions	Uses depthwise convolution to process each channel separately and combines the results using pointwise convolution.	Reduces the number of parameters, allowing more efficient feature processing.
Residual Connections	Connects layers by introducing shortcut paths to facilitate information flow.	Prevents information loss during network propagation.
Fully Connected (Dense)	A fully connected layer that links information from all layers and generates the final output.	Organizes learned information and produces the final prediction.
Softmax Output	Activation function for multi-class classification that outputs class probabilities.	Produces probability values for different output classes.

2.3. Feature Fusion

Feature fusion is a technique used in deep learning to combine features extracted from different models or layers to create a richer and more informative feature representation[33]. This technique allows a model to combine information from different sources that carry complementary information to each other, such as spatial patterns, textures, or high-level characteristics of an image. This feature fusion is often performed through a concatenate operation, where feature maps from different models or layers are combined along a feature dimension to form a unified feature vector[34]. By combining information learned from different models, this technique can improve the model's ability to solve more complex object detection or classification problems. In addition, feature fusion can also improve the generalization ability of a model because the model does not rely solely on one type of feature or model to make predictions. By combining different features, the model can capture more comprehensive and more varied information, which in turn helps distinguish between very similar classes or identify more complex patterns. This technique has been used in various applications, such as object detection, image classification, and face recognition, where the information generated from multiple levels of data processing provides advantages in reducing prediction errors and improving overall model accuracy[35]–[37].

2.4. Related works

This section further discusses some related studies that inspired this research, including some that have been mentioned in the introduction. Gouda et al.[16] used the InceptionV3 architecture to detect and classify skin cancer. The dataset used consisted of 3,533 images of skin lesions divided into two classes, namely benign lesions and malignant lesions. This dataset was split into training and testing data with a ratio of 8:2, allowing the model to learn optimally. The results showed that the InceptionV3 model achieved an accuracy of 85.7%, indicating the potential of this architecture in medical image classification.

Ali et al.[38] conducted a study using pre-trained models such as VGG-16, ResNet50, and InceptionV3 to classify monkeypox and other diseases such as chickenpox and measles. In this study, the Monkeypox Skin Lesion Dataset (MSLD) was developed which consisted of 228 images classified into two classes: monkeypox and others. Among the tested models, ResNet50 showed the best performance with an accuracy of 82.96%, a precision of 87%, a recall of 83%, and an F1 score of 84%. Sahin et al. [39] used the same MSLD dataset but with a different approach, namely using the MobileNetV2 and EfficientNetB0 architectures. The results showed that MobileNetV2 was superior to EfficientNetB0.

Roseline et al. [20] focused their research on improving the accuracy of skin cancer detection and classification through the feature fusion method. The dataset used was SC, consisting of 288 images classified into two classes: cancer and non-cancer. In this study, they combined the MobileNetV2 and Xception architectures to take advantage of the advantages of each model. The results showed a significant increase in performance with an accuracy of 97.56%, a precision of 93.33%, a recall of 100%, and an F1 score of 96.55%. This feature fusion approach shows great potential, but has not been applied to monkeypox disease classification.

Based on the findings of previous studies, this study attempts to explore further the use of feature fusion from two powerful architectures, namely Xception and InceptionV3. This study aims to improve monkeypox disease classification performance by utilizing both architectures' advantages. This approach is expected to provide new contributions in deep learning-based disease classification, especially on the limited MSLD dataset.

3. Proposed Method

The proposed approach in this study aims to classify monkeypox disease by combining features from two models, namely Xception and InceptionV3, by utilizing the serial feature fusion technique. The diagram of the proposed method is shown in Figure 1. The stages carried out in this study are explained as follows:

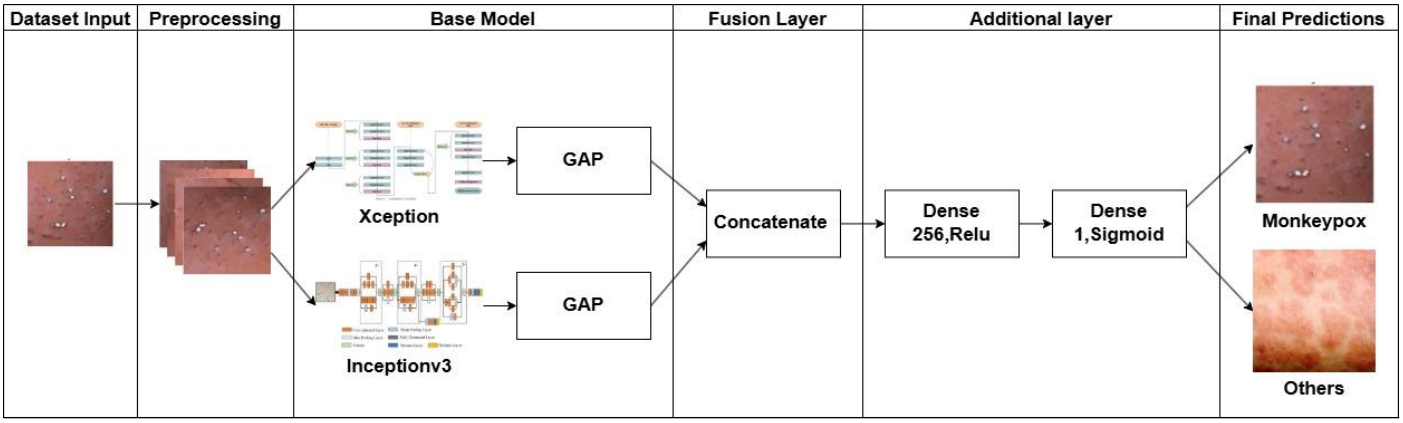

Figure 1. Proposed model.

3.1. Dataset

The data used in this study were obtained from the Kaggle site entitled "Monkeypox Skin Lesion Dataset," which was uploaded in 2022 by nafin59. This dataset is publicly available and can be accessed via the URL https://www.kaggle.com/datasets/nafin59/monkeypox-skin-lesion-dataset. The dataset consists of 228 images of monkeypox skin lesions covering various lesion conditions. The class division will be shown in Table 3. All images in this dataset have a resolution of 224×224 pixels and are used to detect and classify lesions related to monkeypox and those that are not.

Table 3. Dataset distribution.

Class	Number of images
Monkeypox	102
Others	126
Total	228

3.2. Preprocessing

The preprocessing stage is carried out to prepare the dataset for optimal use in the model training process. This process is designed to ensure that the images have adequate quality and variation so that the model can learn better from the data. The preprocessing steps include:

1. The images in the dataset already have a resolution of 224 × 224 pixels, so they do not require additional resizing before augmentation. This native resolution provides the flexibility to apply augmentations such as rotation, translation, and other transformations without sacrificing important details in the image. This resolution is maintained because it matches the size of the original dataset, so there is no loss of important visual details during preprocessing.
2. Data Augmentation is carried out on the training dataset with albumination[26], and the parameters and types of augmentation are presented in Table 4.
3. Image pixel values are normalized with a scale (1./255) to reduce data variability.

Table 4. Augmentation type used and its parameter values.

Augmentation Types	Values
Horizontal Flip	p = 0.5
Vertical Flip	p = 0.5
Rotation	±30° (p = 0.5)
Shift	±5% (p = 0.5)
Scale	±10% (p = 0.5)
Rotate	±15° (p = 0.5)
Random Brightness & Contrast	p = 0.3
Hue Shift	±10 (p = 0.3)
Saturation Shift	±20 (p = 0.3)
Value Shift	±10 (p = 0.3)
CLAHE	p = 0.2
Gaussian Blur	Kernel size (3,5) (p = 0.3)
Gaussian Noise	Variance (10.0, 50.0) (p = 0.3)

3.3. Model Configuration

The proposed model combines the strengths of two deep learning models, namely InceptionV3 and Xception. InceptionV3 is designed to capture multi-scale features from images by utilizing different filter sizes to detect patterns at different resolution levels. This is very effective in capturing visual variations that occur in images of monkeypox lesions with different sizes and textures. Meanwhile, Xception uses depthwise separable convolutions that are more efficient in capturing complex spatial features with fewer parameters, allowing for faster and more accurate detection of images containing fine details of monkeypox.

By default, InceptionV3 and Xception generate complex spatial feature maps in the last convolution layer. Therefore, both models are set with include_top=False, which removes the fully connected layer for classification so that it is only used as a feature extractor. The model input is set to 224×224 pixels, adjusted to the size of the original dataset, ensuring that the extracted features are representative enough. The features from both models are then converted into the same vector with Global Average pooling and combined using the Concatenate function, resulting in a combined feature vector. This vector is processed through a dense layer with 256 units and ReLU activation. After that, the output is forwarded to an output layer with one unit and a sigmoid activation function for binary classification between monkeypox and others.

Table 5. Proposed model configuration.

Function	Configuration
Feature Extraction	InceptionV3 (pretrained with ImageNet weights, include_top=False, input size 224×224×3, all layers frozen)
	Xception (pretrained with ImageNet weights, include_top=False, input size 224×224×3, all layers frozen)
Global average pooling Layers	GAP: Converts the 3D output from InceptionV3 and Xception into a 1D vector.
Feature Fusion	Concatenate function merges feature vectors from the transformed outputs of InceptionV3 and Xception.
Classification Layers	Dense Layer: A dense layer with 256 units and ReLU activation.
	Output Layer: A dense layer with 1 unit and sigmoid activation.
Training Configuration	Optimizer: Adam with a learning rate of 1×10^{-4}.
	Loss Function: Binary Crossentropy.
	Metrics: Accuracy.
	Epochs: 30.
	Batch size: 16.
	Validation: 5-fold stratified cross-validation.

Adam optimizer was chosen to optimize the model training because of its adaptive ability to handle the dynamics of the loss function during training. Adam allows the model to reach convergence faster without requiring much manual adjustment of the hyperparameters. A learning rate of 1×10^{-4} was chosen to maintain a balance between the learning rate and the stability of parameter updates, preventing overfitting or divergence during training. More details of the proposed model design are presented in Table 5.

3.4. Evaluation Metrics

A confusion matrix is an important tool for evaluating the performance of a classification model. It presents the number of True Positive (TP), False Positive (FP), True Negative (TN), and False Negative (FN) predictions, which provide insight into the types of errors the model makes. In the context of this study, TP means cases that are actually infected and predicted to be infected. TN means cases that are actually uninfected and predicted to be uninfected. FP means cases that are actually uninfected but predicted to be infected. Meanwhile, FN means cases that are actually infected but predicted to be uninfected. A confusion matrix is reported as aggregating all matrices generated during the Stratified K-Fold Cross-Validation process. The confusion matrix is visualized as a heatmap, which helps identify the pattern of model errors in distinguishing classes, which is presented in Figure 2.

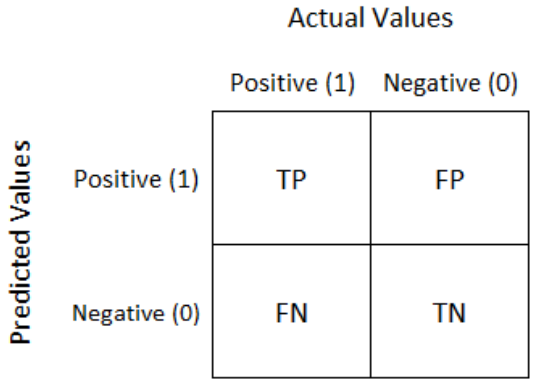

Figure 2. Confusion matrix visualization.

From the confusion matrix, various evaluation metrics can be calculated, such as: accuracy, precision, recall, specificity, and F1-score. Specificity value measures how well the model identifies uninfected individuals as defined in Equation (1). A high Specificity value indicates that the model rarely gives false positives, thus effectively distinguishing healthy individuals. Precision measures the accuracy of the model's positive predictions as presented in Equation (2). High precision indicates that when the model predicts someone is infected, the prediction is likely to be correct. Recall or sensitivity measures how many infected patients are successfully detected by the model. High recall means that the model rarely misses positive cases and is one of the important things in the medical field[40], [41], which is calculated in Equation (3). F1-Score is the harmonic mean between precision and recall, which is calculated by Equation (4). A high F1-score indicates an optimal balance between false positives and false negatives[42]. Accuracy indicates the proportion of correct predictions to the total predictions, see Equation (5). However, prioritizing accuracy can be misleading in imbalanced datasets, particularly in medical contexts.

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{1}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{2}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{3}$$

$$\text{F1} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4}$$

$$\text{Accuracy} = \frac{TP + TN}{TN + FP + TP + FN} \tag{5}$$

The Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) are crucial metrics in evaluating binary classification models, especially in the medical context. ROC curve shows the ability of the model to distinguish two conditions, for example, between infected and uninfected patients[43]. In the ROC graph, the y-axis represents the True Positive Rate (TPR), which describes the ability of the model to identify truly infected patients (recall)[44], while the x-axis represents the False Positive Rate (FPR), which shows how many healthy patients are wrongly classified as infected[43]. TPR and FPR are calculated using Equation (6) and (7), respectively.

$$\text{TPR} = \frac{TP}{TP + FN} \tag{6}$$

$$\text{FPR} = \frac{FP}{FP + TN} \tag{7}$$

The closer the ROC curve is to the upper left corner (0,1), the better the model distinguishes between the two conditions. AUC measures the area under the ROC curve and gives a value between 0 and 1, which describes the model's ability to distinguish between the two conditions. AUC is calculated as the numerical integral of the ROC curve, which is defined in Equation (8), and numerically, AUC is calculated using a numerical approach such as the trapezoidal rule, where the area is calculated by summing the contributions of each segment based on the coordinates (FPR, TPR) calculated in Equation (9).

$$\text{AUC} = \int_0^1 TPR(x)dx \tag{8}$$

$$\text{AUC} = \sum_{i=0}^{n} (\text{FPR}_i - \text{FPR}_{i-1}) \times \frac{\text{TPR}_i - \text{TPR}_{i-1}}{2} \tag{9}$$

The dx variable represents the change in FPR value along the x-axis in the ROC curve graph. With an AUC value close to 1, the model performs well in distinguishing infected and uninfected patients. Conversely, an AUC value close to 0.5 indicates that the model is no better than random guessing. The n value is the number of threshold points used to generate the coordinates (FPR, TPR) along the ROC curve. The larger the n value, the more accurate the AUC estimate is because more segments are used in calculating the area under the curve.

4. Results and Discussion

This study uses the Monkeypox Skin Lesion Dataset (MSLD) from the Kaggle site. This dataset consists of 228 images of monkeypox skin lesions, covering various lesion conditions with a resolution of 224×224 pixels, which allows the application of various efficient image processing techniques. The images in this dataset are in RGB color format, thus including richer color information for further analysis. This dataset was developed through web-scraping techniques from various sources, including news portals, websites, and publicly accessible case reports, resulting in a dataset with variations in lighting, background, and differences in the position and shape of the lesions. This diversity allows the model to be more adaptive in recognizing various Monkeypox skin lesion conditions. This dataset is divided into two classes, namely Monkeypox (102 images) and Others (126 images). Sample images contained in the MSLD dataset are presented in Figure 3.

As explained previously, in the preprocessing stage, the images in this dataset already have a resolution that meets the model's needs, which is 224 × 224 pixels, so no additional resizing process is required. Furthermore, data augmentation techniques are applied to the training data to increase image diversity and help the model learn better. The augmentation parameters used are listed in Table 4, and sample results are presented in Figure 4.

Furthermore, pixel value normalization is carried out by dividing each pixel value by 255, to reduce data variability and improve model performance.

Figure 3. Sample dataset of MSLD (**a**) Monkeypox class; (**b**) other class.

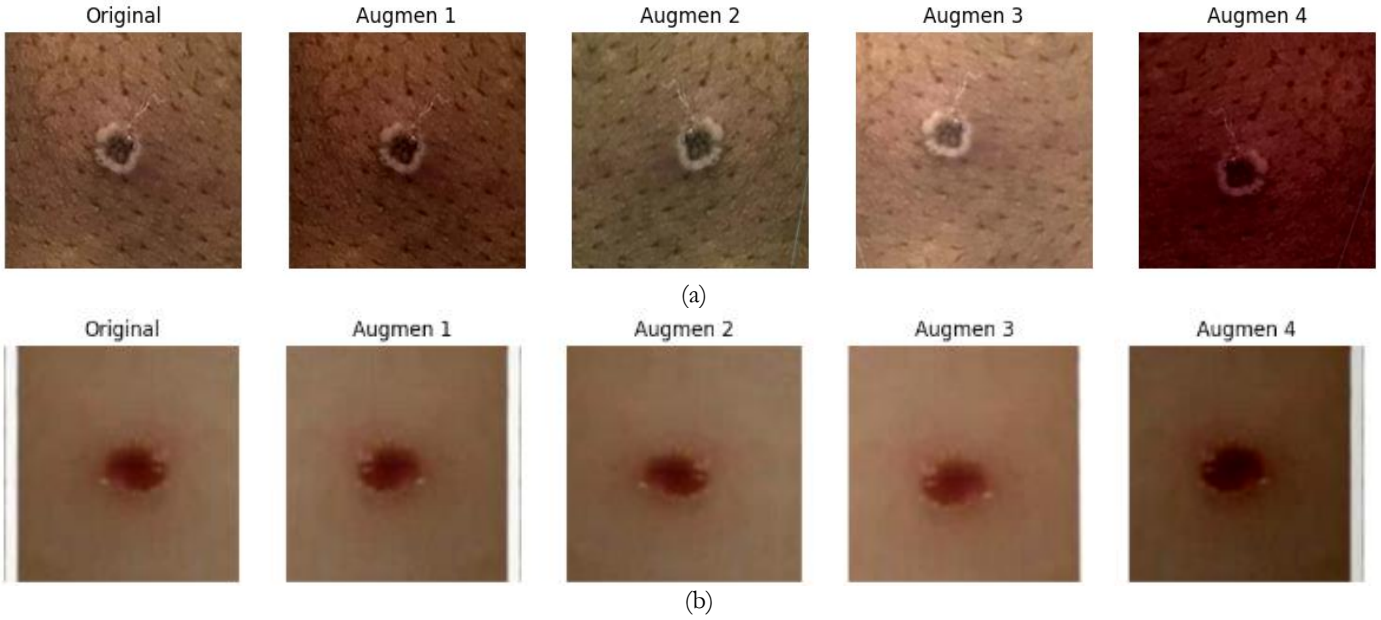

Figure 4. Sample dataset augmentation (**a**) Monkeypox class; (**b**) other class.

After preprocessing and data augmentation, the model was trained using a combination of Xception and InceptionV3 architectures with predetermined configurations. The training process was carried out for 30 epochs, with performance metrics monitoring on the training and validation datasets. To evaluate the stability of the model and its ability to generalize, accuracy and loss were compared between the training and validation data in each epoch. Figure 5a shows the development of training and validation accuracy during the training process. The significant trend of increasing training accuracy in the early epochs indicates that the model can learn from the data well. Meanwhile, Figure 5b illustrates the training and validation loss per epoch. The steady decrease in training loss indicates that the model optimizes parameters to reduce error.

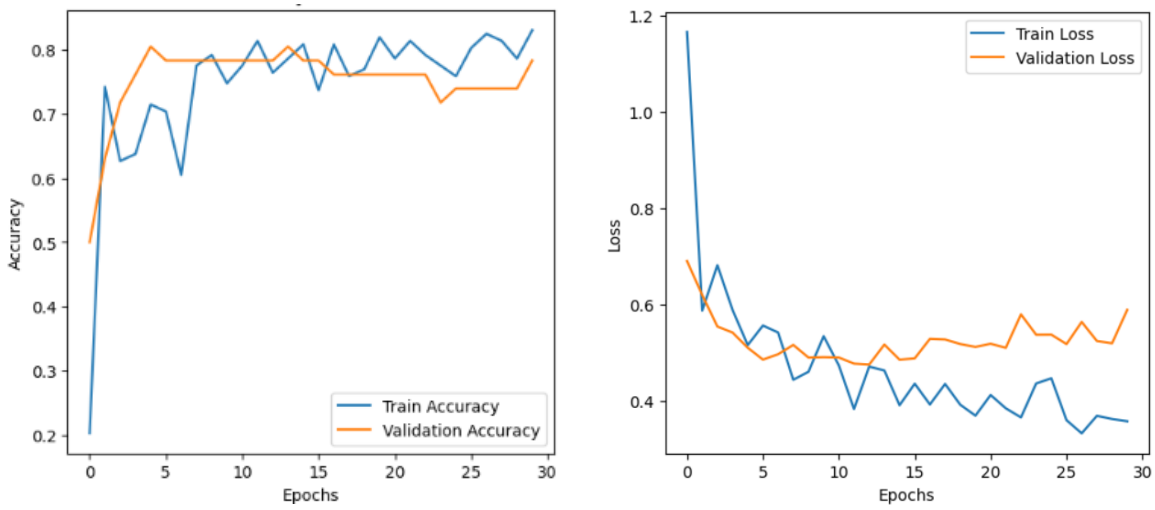

Figure 5. Training and validation accuracy and loss per epoch from a single fold in k-fold cross-validation: (a) training and validation accuracy; (b) training and validation loss.

Next, the model evaluation uses testing data, which is measured by a confusion matrix and classification report. Confusion matrix (Figure 6) and classification report (Figure 7) are presented based on a 5-fold validation aggregation. Based on the confusion matrix, the model performs very well in classifying Monkeypox and Others. Of the 90 Monkeypox samples, 80 were correctly classified (TP), while 10 were misclassified as Others (FP). For the Others class, 116 samples were correctly classified (TN), while 22 were misclassified as Monkeypox (FP).

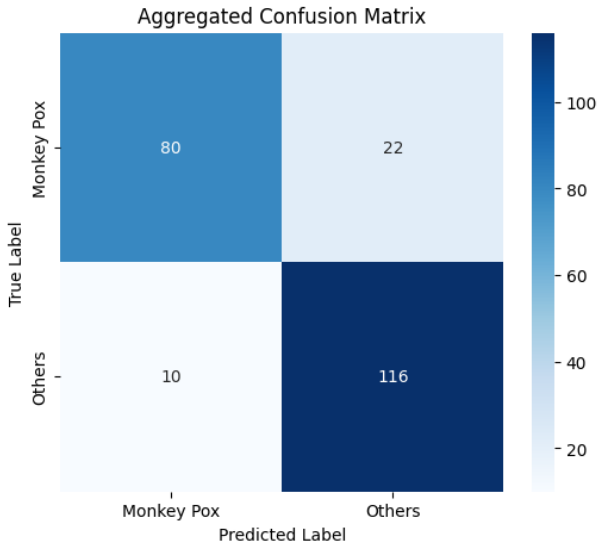

Figure 6. Confusion matrix results from aggregated 5-fold confusion matrix.

Aggregated Classification Report (all folds):

	precision	recall	f1-score	support
Monkey Pox	0.89	0.78	0.83	102
Others	0.84	0.92	0.88	126
accuracy			0.86	228
macro avg	0.86	0.85	0.86	228
weighted avg	0.86	0.86	0.86	228

Figure 7. Classification report of the proposed model from aggregated 5-fold confusion matrix.

From the classification report, it can be seen that the overall accuracy is 86%, as well as the macro average recall of 85% and F1-score of 86%, the model is proven to have good generalization even though there are indications of overfitting during training and validation. However, the lower recall for Monkeypox indicates that the model is more likely to miss some positive cases, which must be fixed to detect the disease more optimally. Model strengthening can be done by adjusting the decision threshold, increasing the weight of False Negative errors, and more specific data augmentation to overcome recurring misclassification. However, overall, the model still has very good generalization to data that has never been seen before. This shows that the feature fusion strategy of Xception and InceptionV3, as well as the data augmentation used, have helped improve the robustness of the model in handling variations in actual data.

Model performance was also analyzed using the ROC curve and AUC, as shown in Figure 8. The AUC value of 0.8931 indicates that the model has very good classification ability, with a high probability of distinguishing between infected and uninfected patients. AUC is a very important metric in medical datasets because it reflects the model's ability to recognize positive and negative cases comprehensively without being affected by the possibly imbalanced class distribution. An AUC value close to 1.0 indicates that the model has a minimal error rate in detecting truly infected patients (TP) while maintaining a low FP rate. This high performance on ROC-AUC indicates that although the recall of the Monkeypox class is slightly lower than the Others class, the model still has excellent overall detection capability. Considering this high AUC value, the model is reliable for medical classification applications, although improving the recall for Monkeypox cases still needs to be a focus in further development.

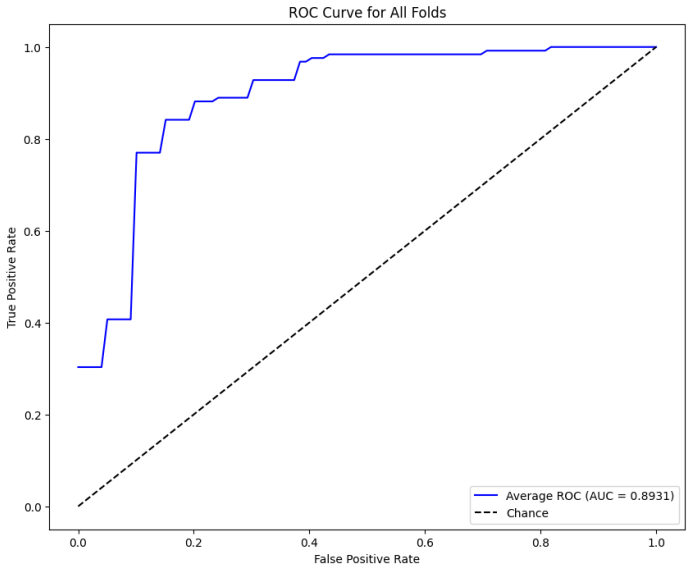

Figure 8. ROC and AUC results.

5. Ablation Study and Comparison

An ablation study was conducted to evaluate the impact of data augmentation on model performance. Based on Table 6, the proposed method significantly improves all metrics after augmentation. Without augmentation, the proposed model achieves an accuracy of 82.02%, which increases to 85.96% after augmentation. In addition, the precision increases from 82.03% to 86.47%, recall from 81.49% to 85.25%, and specificity from 76.29% to 78.43%, indicating an improvement in the model's ability to recognize positive cases while maintaining balance with the negative class.

Data augmentation also positively impacts the base model, with InceptionV3 improving from 81.58% to 82.46%, while Xception improves from 79.82% to 81.14% after augmentation. This ablation study confirms that augmentation improves accuracy and model generalization, with the most significant impact seen on the feature fusion model compared to the individual models. The 3.76% increase in recall of the proposed model indicates that augmentation helps the model recognize more positive cases that were previously missed, which is

very important in disease detection. Although the AUC of the proposed model slightly decreased from 0.9050 to 0.8931 after augmentation, the increase in recall indicates that the model is more sensitive to identifying monkeypox lesions, which is more important in medical applications than a slight decrease in classification balance.

Table 6. Ablation study.

Model	DA	Accuracy	Precision	Recall	F1-score	Specificity	AUC
InceptionV3	no	81.58	81.63	81.00	81.21	75.19	0.8964
Xception	no	79.82	79.67	79.41	79.52	75.38	0.8774
Proposed	no	82.02	82.03	81.49	81.68	76.29	**0.9050**
InceptionV3	yes	82.46	82.43	81.98	82.15	77.33	0.8769
Xception	yes	81.14	82.08	80.04	80.44	69.48	0.8774
Proposed	yes	**85.96**	**86.47**	**85.25**	**85.61**	**78.43**	0.8931

*DA: Data augmentation.

The model was compared with previous studies using the same dataset for further validation. Table 7 shows that the proposed model outperforms the method developed by Ali et al., with an accuracy of 85.96%, higher than 82.96% (Ali et al.). In addition, the proposed model also excels in precision (86.46%), recall (85.25%), and F1-score (86.61%), indicating a significant improvement in the balance between false positives and false negatives. Unlike previous studies, this model also comes with a specificity of 78.43% and an AUC Score of 0.8931, confirming its reliability in identifying Monkeypox and non-Monkeypox diseases more accurately.

Table 7. Comparison with related work.

Metrics	Model	
	Ali et al. [38]	Proposed
Accuracy	82.96	85.96
Precision	87	86.47
Recall	83	85.25
F1-score	84	85.61
Specificity	-	78.43
AUC	-	0.8931

These results confirm that combining feature fusion from Xception and InceptionV3 and data augmentation significantly improves the model performance in Monkeypox classification, making it more reliable than previous methods tested on the same dataset.

6. Conclusions

This study proposes a feature fusion-based deep learning model, combining Xception and InceptionV3 to improve the accuracy of Monkeypox skin lesion classification. With the application of data augmentation using Albumentation, the proposed model achieves an accuracy of 85.96%, a precision of 86.47%, a recall of 85.25%, a specificity of 78.43%, and an AUC of 0.8931, showing improvements compared to previous methods on the Monkeypox Skin Lesion Dataset (MSLD). The main findings of this study confirm that feature fusion can produce richer feature representations, while data augmentation increases the model's sensitivity to Monkeypox cases. The ablation study also proves that data augmentation significantly improves the model's generalization.

Although the model has shown good performance, there is still room for further optimization. Future studies can explore architecture fine-tuning strategies to improve the balance between recall and precision and evaluate additional feature selection methods to strengthen the extraction of skin lesion characteristics. In addition, using larger and more diverse datasets, including images from various clinical sources, can help improve the robustness of the model in dealing with variations in real conditions. The results of this study contribute to the development of a more accurate and efficient deep learning-based medical image

classification method, and can be applied in a medical decision support system to assist early detection of Monkeypox, especially in areas with limited conventional diagnostic facilities.

Author Contributions: Conceptualization: NRP. and DRIMS.; methodology, NRP. and DRIMS.; software: NRP.; validation: IH. and AAO.; Formal analysis: IH. and AAO.; investigation: IH. and AAO.; resources: NRP; data curation: IH. and AAO.; writing—original draft preparation: NRP.; writing—review and editing: DRIMS, IH. and AAO; visualization: NRP.; supervision: DRIMS.; project administration: DRIMS.; funding acquisition: N/A. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Conflicts of Interest: The authors declare no conflict of interest.

References

[1] I. D. Ladnyj, P. Ziegler, and E. Kima, "A human infection caused by monkeypox virus in Basankusu Territory, Democratic Republic of the Congo.," *Bull. World Health Organ.*, vol. 46, no. 5, pp. 593–7, 1972, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/4340218

[2] World Health Organization (WHO), "Mpox - WHO Fact Sheet," *World Health Organization (WHO)*. 2024. Accessed: Jan. 22, 2025. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/mpox

[3] E. Pérez-Barragán, S. Pérez-Cavazos, J. C. Rodríguez-Aldama, and R. A. Cruz-Flores, "Mimicking measles and syphilis: Mpox in PLHIV," *HIV Med.*, vol. 24, no. 7, pp. 851–853, Jul. 2023, doi: 10.1111/hiv.13462.

[4] D. S. Stamoulis and C. Papachristopoulou, "Artificial Intelligence in Radiology, Emergency, and Remote Healthcare: A Snapshot of Present and Future Applications," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 3, pp. 228–234, Oct. 2024, doi: 10.62411/faith.3048-3719-38.

[5] O. Jaiyeoba, O. Jaiyeoba, E. Ogbuju, and F. Oladipo, "AI-Based Detection Techniques for Skin Diseases: A Review of Recent Methods, Datasets, Metrics, and Challenges," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 3, pp. 318–336, Dec. 2024, doi: 10.62411/faith.3048-3719-46.

[6] T. R. Noviandy, G. M. Idroes, and I. Hardi, "An Interpretable Machine Learning Strategy for Antimalarial Drug Discovery with LightGBM and SHAP," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 2, pp. 84–95, Aug. 2024, doi: 10.62411/faith.2024-16.

[7] S. Fanijo, "AI4CRC: A Deep Learning Approach Towards Preventing Colorectal Cancer," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 2, pp. 143–159, Sep. 2024, doi: 10.62411/faith.2024-28.

[8] K. Pyar, "Segmentation Performance Analysis of Transfer Learning Models on X-Ray Pneumonia Images," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 1, pp. 64–74, Jun. 2024, doi: 10.62411/faith.2024-10.

[9] A. Pathirana *et al.*, "A Reinforcement Learning-Based Approach for Promoting Mental Health Using Multimodal Emotion Recognition," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 2, pp. 124–142, Sep. 2024, doi: 10.62411/faith.2024-22.

[10] F. S. Gomiasti, W. Warto, E. Kartikadarma, J. Gondohanindijo, and D. R. I. M. Setiadi, "Enhancing Lung Cancer Classification Effectiveness Through Hyperparameter-Tuned Support Vector Machine," *J. Comput. Theor. Appl.*, vol. 1, no. 4, pp. 396–406, Mar. 2024, doi: 10.62411/jcta.10106.

[11] M. B. Teferi and L. A. Akinyemi, "Deep Learning-Based Cross-Cancer Morphological Analysis: Identifying Histopathological Patterns in Breast and Lung Cancer," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 3, pp. 235–248, Oct. 2024, doi: 10.62411/faith.3048-3719-36.

[12] X. Jiang, Z. Hu, S. Wang, and Y. Zhang, "Deep Learning for Medical Image-Based Cancer Diagnosis," *Cancers (Basel).*, vol. 15, no. 14, p. 3608, Jul. 2023, doi: 10.3390/cancers15143608.

[13] M. S. Iqbal, L. Kotthoff, and P. Jamshidi, "Transfer Learning for Performance Modeling of Deep Neural Network Systems," *Proc. 2019 USENIX Conf. Oper. Mach. Learn. OpML 2019*, Apr. 2019, [Online]. Available: http://arxiv.org/abs/1904.02838

[14] F. Mazhar, N. Aslam, A. Naeem, H. Ahmad, M. Fuzail, and M. Imran, "Enhanced Diagnosis of Skin Cancer from Dermoscopic Images Using Alignment Optimized Convolutional Neural Networks and Grey Wolf Optimization," *J. Comput. Theor. Appl.*, vol. 2, no. 3, pp. 368–382, Jan. 2025, doi: 10.62411/jcta.11954.

[15] S. M. Anwar, M. Majid, A. Qayyum, M. Awais, M. Alnowami, and M. K. Khan, "Medical Image Analysis using Convolutional Neural Networks: A Review," *J. Med. Syst.*, vol. 42, no. 11, p. 226, Nov. 2018, doi: 10.1007/s10916-018-1088-1.

[16] W. Gouda, N. U. Sama, G. Al-Waakid, M. Humayun, and N. Z. Jhanjhi, "Detection of Skin Cancer Based on Skin Lesion Images Using Deep Learning," *Healthcare*, vol. 10, no. 7, p. 1183, Jun. 2022, doi: 10.3390/healthcare10071183.

[17] E. Erdem and T. Aydin, "Artificial Neural Network-Based Approaches to Improve Classification of Skin Lesions," in *2023 Medical Technologies Congress (TIPTEKNO)*, Nov. 2023, pp. 1–4. doi: 10.1109/TIPTEKNO59875.2023.10359238.

[18] G. Cheng and J. Han, "A survey on object detection in optical remote sensing images," *ISPRS J. Photogramm. Remote Sens.*, vol. 117, pp. 11–28, Jul. 2016, doi: 10.1016/j.isprsjprs.2016.03.014.

[19] L. R. Zuama, D. R. I. M. Setiadi, A. Susanto, S. Santosa, H.-S. Gan, and A. A. Ojugo, "High-Performance Face Spoofing Detection using Feature Fusion of FaceNet and Tuned DenseNet201," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 4, pp. 385–400, Feb. 2025, doi: 10.62411/faith.3048-3719-62.

[20] R. O. Ogundokun *et al.*, "Enhancing Skin Cancer Detection and Classification in Dermoscopic Images through Concatenated MobileNetV2 and Xception Models," *Bioengineering*, vol. 10, no. 8, p. 979, Aug. 2023, doi: 10.3390/bioengineering10080979.

[21] F. Chollet, "Xception: Deep Learning with Depthwise Separable Convolutions," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jul. 2017, vol. 2017-Janua, no. 3, pp. 1800–1807. doi: 10.1109/CVPR.2017.195.
[22] T. Hidayat, I. A. Astuti, A. Yaqin, A. P. Tjilen, and T. Arifianto, "Grouping of Image Patterns Using Inceptionv3 For Face Shape Classification," *JOIV Int. J. Informatics Vis.*, vol. 7, no. 4, Dec. 2023, doi: 10.30630/joiv.7.4.1743.
[23] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *J. Big Data*, vol. 6, no. 60, 2019, doi: 10.1186/s40537-019-0197-0.
[24] K. P. Zaw and A. Mon, "Enhanced Multi-Class Skin Lesion Classification of Dermoscopic Images Using an Ensemble of Deep Learning Models," *J. Comput. Theor. Appl.*, vol. 2, no. 2, pp. 256–267, Nov. 2024, doi: 10.62411/jcta.11530.
[25] M. A. Hambali and P. A. Agwu, "Adversarial Convolutional Neural Network for Predicting Blood Clot Ischemic Stroke," *J. Comput. Theor. Appl.*, vol. 2, no. 1, pp. 51–64, Jun. 2024, doi: 10.62411/jcta.10516.
[26] A. Buslaev, V. I. Iglovikov, E. Khvedchenya, A. Parinov, M. Druzhinin, and A. A. Kalinin, "Albumentations: Fast and Flexible Image Augmentations," *Information*, vol. 11, no. 2, p. 125, Feb. 2020, doi: 10.3390/info11020125.
[27] N. Buławka, H. A. Orengo, and I. Berganzo-Besga, "Deep learning-based detection of qanat underground water distribution systems using HEXAGON spy satellite imagery," *J. Archaeol. Sci.*, vol. 171, p. 106053, Nov. 2024, doi: 10.1016/j.jas.2024.106053.
[28] B. J. Filia *et al.*, "Improving Batik Pattern Classification using CNN with Advanced Augmentation and Oversampling on Imbalanced Dataset," *Procedia Comput. Sci.*, vol. 227, pp. 508–517, 2023, doi: 10.1016/j.procs.2023.10.552.
[29] A. Tatar, M. Haghighi, and A. Zeinijahromi, "Experiments on image data augmentation techniques for geological rock type classification with convolutional neural networks," *J. Rock Mech. Geotech. Eng.*, vol. 17, no. 1, pp. 106–125, Jan. 2025, doi: 10.1016/j.jrmge.2024.02.015.
[30] C. Szegedy *et al.*, "Rethinking the Inception Architecture for Computer Vision," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Jun. 2016, pp. 2818–2826. doi: 10.1109/CVPR.2016.308.
[31] H. T. Adityawan, O. Farroq, S. Santosa, H. M. M. Islam, M. K. Sarker, and D. R. I. M. Setiadi, "Butterflies Recognition using Enhanced Transfer Learning and Data Augmentation," *J. Comput. Theor. Appl.*, vol. 1, no. 2, pp. 115–128, Nov. 2023, doi: 10.33633/jcta.v1i2.9443.
[32] F. M. Firnando, D. R. I. M. Setiadi, A. R. Muslikh, and S. W. Irianandа, "Analyzing InceptionV3 and InceptionResNetV2 with Data Augmentation for Rice Leaf Disease Classification," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 1, pp. 1–11, May 2024, doi: 10.62411/faith.2024-4.
[33] W. Song, S. Li, L. Fang, and T. Lu, "Hyperspectral Image Classification With Deep Feature Fusion Network," *IEEE Trans. Geosci. Remote Sens.*, vol. 56, no. 6, pp. 3173–3184, Jun. 2018, doi: 10.1109/TGRS.2018.2794326.
[34] P. Xu, F. Li, and H. Wang, "A novel concatenate feature fusion RCNN architecture for sEMG-based hand gesture recognition," *PLoS One*, vol. 17, no. 1, p. e0262810, Jan. 2022, doi: 10.1371/journal.pone.0262810.
[35] Y. Zhang, Y. Chen, C. Huang, and M. Gao, "Object Detection Network Based on Feature Fusion and Attention Mechanism," *Futur. Internet*, vol. 11, no. 1, p. 9, Jan. 2019, doi: 10.3390/fi11010009.
[36] D. Duarte, F. Nex, N. Kerle, and G. Vosselman, "Multi-Resolution Feature Fusion for Image Classification of Building Damages with Convolutional Neural Networks," *Remote Sens.*, vol. 10, no. 10, p. 1636, Oct. 2018, doi: 10.3390/rs10101636.
[37] X. Lu, X. Duan, X. Mao, Y. Li, and X. Zhang, "Feature Extraction and Fusion Using Deep Convolutional Neural Networks for Face Detection," *Math. Probl. Eng.*, vol. 2017, no. 1, Jan. 2017, doi: 10.1155/2017/1376726.
[38] S. N. Ali *et al.*, "Monkeypox Skin Lesion Detection Using Deep Learning Models: A Feasibility Study," *ArXiv*. pp. 2–5, Jul. 06, 2022. [Online]. Available: http://arxiv.org/abs/2207.03342
[39] V. H. Sahin, I. Oztel, and G. Yolcu Oztel, "Human Monkeypox Classification from Skin Lesion Images with Deep Pre-trained Network using Mobile Application," *J. Med. Syst.*, vol. 46, no. 11, p. 79, Oct. 2022, doi: 10.1007/s10916-022-01863-7.
[40] D. R. I. M. Setiadi, K. Nugroho, A. R. Muslikh, S. W. Irianandа, and A. A. Ojugo, "Integrating SMOTE-Tomek and Fusion Learning with XGBoost Meta-Learner for Robust Diabetes Recognition," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 1, pp. 23–38, May 2024, doi: 10.62411/faith.2024-11.
[41] D. R. I. M. Setiadi, H. M. M. Islam, G. A. Trisnapradika, and W. Herowati, "Analyzing Preprocessing Impact on Machine Learning Classifiers for Cryotherapy and Immunotherapy Dataset," *J. Futur. Artif. Intell. Technol.*, vol. 1, no. 1, pp. 39–50, Jun. 2024, doi: 10.62411/faith.2024-2.
[42] R. Trevethan, "Sensitivity, Specificity, and Predictive Values: Foundations, Pliabilities, and Pitfalls in Research and Practice," *Front. Public Heal.*, vol. 5, Nov. 2017, doi: 10.3389/fpubh.2017.00307.
[43] K. Hajian-Tilaki, "Receiver Operating Characteristic (ROC) Curve Analysis for Medical Diagnostic Test Evaluation.," *Casp. J. Intern. Med.*, vol. 4, no. 2, pp. 627–35, 2013, [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/24009950
[44] P. Singh, N. Singh, K. K. Singh, and A. Singh, "Diagnosing of disease using machine learning," in *Machine Learning and the Internet of Medical Things in Healthcare*, Elsevier, 2021, pp. 89–111. doi: 10.1016/B978-0-12-821229-5.00003-3.