JTAM (Jurnal Teori dan Aplikasi Matematika)
http://journal.ummat.ac.id/index.php/jtam
p-ISSN 2597-7512 | e-ISSN 2614-1175
Vol. 10, No. 3, July 2026, pp. 783-796

Computational Analysis of Xception and ConvMixer Architecture in Classification of Skin Disease Images Using Geometric Transformation

Maya Isafa Sam Saputri[1*], Sugiyarto Surono[1], Aris Thobirin[1]
[1]Department of Mathematics, Universitas Ahmad Dahlan, Indonesia
2200015023@webmail.uad.ac.id

Article History:
Received : 09-12-2025
Revised : 31-01-2026
Accepted : 02-02-2026
Online : 01-07-2026

Keywords:
Deep Learning;
Xception;
Convermixer.

ABSTRACT

This research seeks to evaluate and contrast the effectiveness of two deep learning models, Xception and ConvMixer, for classification of skin disease images. An experimental methodology was employed using the Massive Skin Disease. The data is divided into training, validation, and test data with a ratio of 80:10:10. The pre-processing stage includes resizing, normalization, and the application of geometric augmentation to improve visual variation in the training data. Both models were trained using equalized parameters so that comparisons were made objectively. The models were assessed through several evaluation metrics, including loss, accuracy, precision, recall, and F1-score metrics in a multi-class classification scheme. The results showed that Xception obtained a test accuracy of 99,70%, while ConvMixer achieved 94,60%. Additionally, Xception exhibits faster convergence and more stable inter-class performance consistency, while ConvMixer excels in compute time efficiency. This study contributes in the form of a comparative analysis of two modern architectures with training parameters that are equalized in the classification of skin diseases. However, the study is still limited to the use of a partial class and a single dataset, so further testing is needed to ensure the generalization capabilities of the model over a wider range of scenarios.

https://doi.org/10.31764/jtam.v10i3.37037

This is an open access article under the CC–BY-SA license

A. INTRODUCTION

The advancement of deep learning technology in recent years has led to siginificiant progress across various domains, particularly in digital image processing and medical data analysis (Mienye & Black, 2024). As a branch of machine learning, deep learning utilizes multilayer artificial neural networks to automatically learn complex and nonlinear feature representations from raw data, eliminating the need for manual feature engineering (Khoei et al., 2023). This approach has been demonstrated to enhance both the accuracy and efficiency of the system, especially in large-scale visual and textual data processing (Wang et al., 2023). As the availability of datasets continues to grow and computational power increases, deep learning is increasingly becoming a fundamental component in the development of modern artificial intelligence technologies, including applications in the medical field (Abulwafa, 2022; Supriyono et al., 2024).

One of the deep learning architectures that is very influential in image processing is the Convolutional Neural Network (CNN) (Krichen, 2023; Rangel et al., 2024). CNNs are specifically designed to take advantage of the spatial characteristics and local correlations between pixels

in images through convolutional, pooling, and nonlinear activation operations (Krichen, 2023; Raj & Kos, 2025). Through this mechanism, CNN is able to build a multi-level representation of features, ranging from basic features such as edges and textures to high-level features such as complex shapes and patterns of objects (Bracci et al., 2023; Tempfli & Sándor, 2024). This ability allows CNN to mimic the way the human visual system works in recognizing objects, where each layer of the network learns increasingly abstract information. Therefore, CNN is becoming the primary approach widely used in various image processing applications, including classification, detection, and analysis of medical images (Younesi et al., 2024).

Although CNN has been shown to be highly effective across various image processing applications, its performance largely depends on the architectural design of the network employed (Krichen, 2023; Rangel et al., 2024). Conventional CNN architectures such as LeNet, AlexNet, and VGG have a large number of parameters, requiring high computing resources and risking overviting, especially when the amount of training data is limited (Bhatt et al., 2021; Krichen, 2023). In addition, the deeper a network is, the more complex the training process will be, which can affect the stability and efficiency of the model (Rangel et al., 2024). Therefore, various architectural innovations are constantly being developed to overcome these limitations, such as the use of residual connections on ResNet and the implementation of modular blocks on inception, which aim to improve computing efficiency while maintaining strong feature representation capabilities (Bhatt et al., 2021).

One of the notable architectural advancements in CNN is Xception, proposed by François Chollet as an extension of the Inception architecture (Chollet, 2017). Xception was developed to enhance computational efficiency while preserving strong feature extraction capabilities, enabling it to achieve competitive results across various image classification tasks. The architecture is recognized for its relatively simple yet powerful design and its ability to lower model complexity compared to traditional CNN models. Owing to these advantages, Xception has been extensively applied in numerous image processing domains, including medical applications, and has demonstrated an effective balance between accuracy and computational cost (Chollet, 2017; Nawar et al., 2024; Sathya et al., 2024).

In addition to Xception, recent developments in deep learning architecture design have given birth to the ConvMixer model, which combines the concept of patch-based processing like the Vision Transformer with the computing efficiency of CNN (Trockman & Kolter, 2022). ConvMixer leverages patch-based representation and convolutional operations to combine spatial and channel information, resulting in an architecture that is lightweight, easy to train, and stable during the convergence process. Compared to conventional CNNs, ConvMixer is able to provide competitive and stable performance in the training process. Because of these characteristics, ConvMixer began to be widely used as an alternative to modern architecture in various image classification tasks, including in the medical field (Solano et al., 2023; Üzen & Fırat, 2024).

Although the Xception and ConvMixer architectures have been widely used in various image processing, studies that directly compare the performance of the two architectures in the classification of skin disease images are still relatively limited. For the most part, previous researchers still focused on using conventional CNN architectures or only evaluating one specific type of architecture without conducting a comprehensive comparison (Sarı & Keser,

2025). In addition, the use of data augmentation, particularly geometric augmentation, has often not been studied in depth as a factor affecting the performance and generalization capabilities of the model (Awaluddin et al., 2023; Hao et al., 2023; Mumuni & Mumuni, 2022). Therefore, a study is needed that analyzes and compares Xception and ConvMixer directly using equalized trainer parameters and the application of geometric augmentation in order to obtain a more objective picture of the advantages of each architecture in the classification of skin diseases.

Based on the research background that has been described, this study aims to analyze and compare the architectural performance of Xception and ConvMixer in the classification of skin disease images by utilizing data augmentation based on geometric transformation. This study is focused on evaluating the performance of the two architectures using equalized training parameters so that comparisons can be carried out objectively. The performance of the model was assessed using several evaluation metrics, including loss, accuracy, precision, recall, F1-score, and computational time. Therefore, this study is intended to offer a clearer understanding of the effectiveness and efficiency of Xception and ConvMixer, while also serving as a reference for the development of a more accurate and dependable image-based skin disease diagnostic system.

B. METHODS

This research adopts an experimental design with a quantitative approach to compare the performance of two deep learning architectures, Xception and ConvMixer, for classifying skin disease images. The experiment was carried out by applying both models to the same dataset, using a data sharing scheme, augmentation method, and equalized training parameters so that comparisons could be carried out objectively and fairly. The results of the experiment were analyzed based on model performance evaluation metrics, namely loss, accuracy, precision, recall, and F1-score. The stages of the methodology of research carried out in general can be seen in Figure 1.

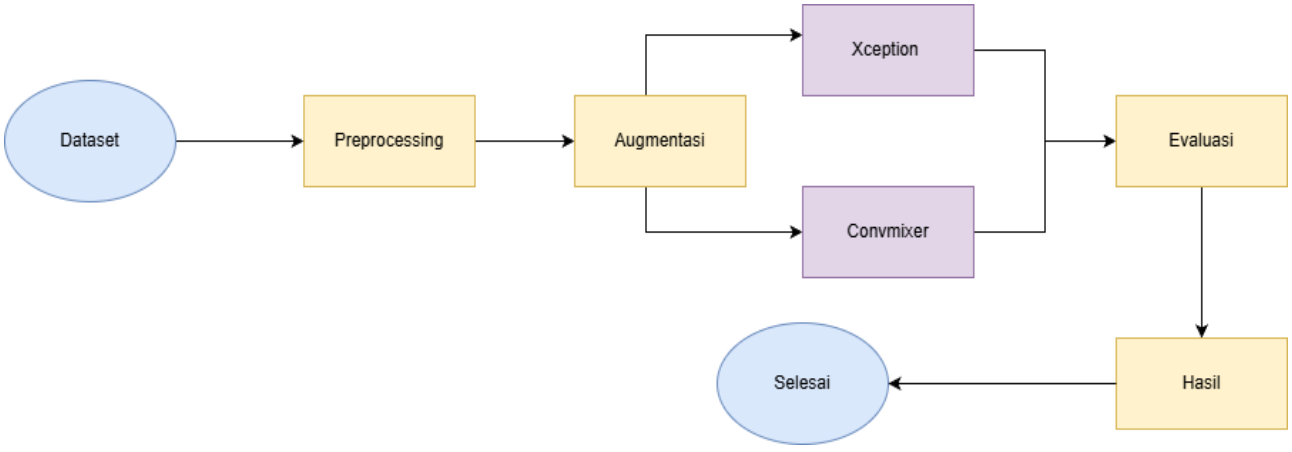


Figure 1. Research flowchart

1. Dataset

The dataset used in this study comes from Kaggle with the name Massive Skin Disease Balanced Dataset developed by Muhammad Abdul Sami. This dataset is a skin disease classification dataset designed for the needs of deep learning research and image-based dermatology diagnosis. Overall, this dataset consists of 262,874 images of skin diseases categorized into 34 different disease classes, covering different types of dermatological conditions. Out of a total of 34 classes available, this study only used 14 classes with a total of 104,430 images that were selected based on the availability of representative data and their relevance to the research objective, namely the classification of skin disease images using a deep learning approach. The selection of this subset of classes aims to maintain the balance of data between classes and ensure the quality and consistency of the imagery used in the training and testing of the model. The 14 classes used in this study include: Pa Cutaneous Larva Migrans, Poison Ivy and Other Contact Dermatitis, Psoriasis Pictures and Related Diseases, Rashes, Scabies and Other Infestations and Bites, Seborrheic Keratoses and Other Benign Tumors, Systemic Disease, Tinea Ringworm and Other Fungal Infections, Urticaria Hives, Vascular Tumors, Vasculitis Photos, Vi Chickenpox, Vi Shingles, and Warts Molluscum and Other Viral Infections.

2. Preprocessing

Following the class selection stage, the dataset is partitioned into three primary subsets: training, validation, and testing data, with a proportion of 80:10:10. The splitting process is performed randomly while maintaining a balanced class distribution within each subset. The training set is used to train the model, the validation set monitors performance during the training phase, and the test set is applied to objectively evaluate the model's final results. Furthermore, all images are resized according to the input size requirements of each architecture, 299 × 299 pixels for Xception and 244 × 244 pixels for ConvMixer. In addition, pixel values are normalized by converting their range from [0–255] to [0, 1] to improve training stability and computational efficiency. The dataset distribution is shown in Table 1.

Table 1. Data Sharing Distribution

Data Spliting	Amount
Train	83539
Validation	10438
Test	10453

3. Augmentation

Data augmentation is a method employed to expand the diversity of training data by applying various transformations to images while preserving their original labels or classes. The purpose of this approach is to introduce greater visual variation into the training dataset, enabling the model to recognize objects under different display conditions. In this research, the augmentation strategy emphasizes geometric transformations, which modify the position, orientation, and scale of images without altering the identity of the objects they contain. Therefore, geometric augmentation contributes significantly to enhancing the model's generalization ability and minimizing the risk of overfitting. Some commonly used types of

image augmentation include rotation, flipping, zooming, cropping, changing brightness levels and adding noise(J. Wang & Perez, 2017)(Alomar & Aysel, 2023). In this study, the types and parameters of data augmentation used are presented in Table 2.

Table 2. Augmentation

Types of Augmentation	Value
Rotation	20°
Width shift	0,2
High Shift	0,2
Shear	0,2
Zoom	0,2
Horizontal flip	True

4. Model Architecture

a. Xception

Xception (Extreme Inception) is a CNN architecture proposed by François Chollet in 2017 as an extension of the Inception model (Chollet, 2017). The core concept of Xception is to substitute the conventional convolution operations used in Inception with Depthwise Separable Convolution, a technique that divides spatial filtering and channel-wise mixing into two distinct stages. Through this strategy, Xception can decrease the number of parameters while enhancing computational efficiency, without compromising its ability to extract meaningful features (Chollet, 2017; Muhammad et al., 2021). In addition, Xception views standard convolution as a linear mapping between channels that can be broken down into depthwise convolution and pointwise convolution (Lukasz et al., 2017). Overall, Xception has a simple but very robust structure, consisting of an entry flow, middle flow, and exit flow, each using a depthwise separable convolution block with a residual connection (Chollet, 2017). Mathematically, point wise can be defined as:

$$PointwiseConv(W, y)_{(i,j)} = \sum_m^M W_m \cdot y_{(i,j,m)} \quad (1)$$

Then Depthwise Convolution is formulated as:

$$DepthwiseConv(W, y)_{(i,j)} = \sum_{k,l}^{K,L} W_{(k,l)} \odot y_{(i+k,j+l)} \quad (2)$$

The combination of the two operations forms a depthwise separable convulution which is expressed as:

$$SepConv(W_p, W_d, y)_{(i,j)} = PointwiseConv_{(i,j)}(W_p, DepthwiseConv_{(i,j)}, (W_d, y)) \quad (3)$$

This formulation shows that depthwise separable convolution is capable of approaching standard convolution performance with much lower computational complexity, so it is

widely used in efficient CNN architectures such as Xception (Chollet, 2017; Muhammad et al., 2021).

b. Convmixer

ConvMixer is a convolution-based deep learning architecture that combines the patch embedding concept of Vision Transformer with the convolutional efficiency of CNN (Trockman & Kolter, 2022). This architecture is designed to produce models that are simple, lightweight and stable in the training process, while still being able to provide competitive performance on image classification tasks. In this study, ConvMixer was used as a comparative model of xception to evaluate the effectiveness of a patch-based convolution approach on the classification of skin disease images.

At the initial stage, the input image is transformed into a patch-based representation through a patch embedding process that applies a convolution operation with a stride equal to the kernel size. As a result, the image is partitioned into several patches, which are subsequently processed by the ConvMixer block. Each ConvMixer block comprises a depthwise separable convolution for spatial information mixing and a pointwise convolution to integrate information across channels. Furthermore, every layer is followed by a Gaussian Error Linear Unit (GELU) activation function and a skip connection to enhance the stability of the training process (Ibrahem et al., 2025; Iijima & Kiya, 2022; Lin et al., 2024). Mathematically, the image is input with the size HxW and the number of channels C expressed as:

$$x \boldsymbol{\epsilon} \mathbb{R}^{H \times W \times C} \tag{4}$$

Then, the image is re-represented as a set of two-dimensional patches that have been flattened as:

$$X_p \boldsymbol{\epsilon} \mathbb{R}^{N \times (P^2 C)} \tag{5}$$

By number of patches:

$$N = \frac{HW}{P^2} \tag{6}$$

The embedding process of each patch can be expressed as:

$$z_0 = [x^1 E;\ x^2 E; \ldots;\ x^N E] +\ E_{pos} \tag{7}$$

Where E is the projection matrix and E_pos is the positional embedding that is added to maintain the patch position information (Zhai et al., 2021). Convmixer uses the Gausian activation function of linear error units (GELU) which is mathematically formulated as (Hendrycks & Gimpel, 2016):

$$GELU(x) = xP(X \leq x) = x\phi(x) = x \cdot \frac{1}{2}[\,1 + \text{erf}(\text{x}/\ \sqrt{2})] \tag{8}$$

This formulation demonstrates that ConvMixer combines the advantages of a patch-based approach with the simplicity of convolutional operations, resulting in a computationally efficient and stable architecture during the training process. These characteristics make ConvMixer relevant to be directly compared with Xception in the skin disease image classification experiment conducted in this study.

5. Setup Experium

In this study, the Xception and ConvMixer architectures were trained using the same configuration of training parameters to ensure fair and objective performance comparisons. Parameter equalization is done so that the resulting performance differences really come from the characteristics of each architecture, not from differences in training settings. The training parameters used in this study are presented in Table 3.

Table 3. Data Sharing Distribution

Parameters	Approach
Learning Rate	$1x10^{-4}$
Epoch	20
Batch Size	32
Optimizer	Adam

The entire training process was carried out on the same dataset and augmentation scheme for both models. With this setting, the performance evaluation obtained can reflect the true ability of Xception and ConvMixer to objectively classify skin disease images and can be replicated by other researchers.

6. Evaluation matrix

In this research, model performance was evaluated by computing loss, accuracy, precision, recall, and F1-score metrics. Since the dataset consists of 14 labeled classes, a multiclass confusion matrix was employed. The application of a multiclass confusion matrix enables a more thorough assessment of model performance across each class (Ainurrohmah & Wiyanti, 2023; Terven et al., 2025). The loss function applied in this study is categorical cross-entropy, which is defined as follows:

$$\mathcal{L}_{CCE}(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{C} y_{i,j}\log(\hat{P}_{i,j}) \tag{9}$$

Next, the average accuracy is calculated using the equation:

$$Average\ accuracy = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{TP_i+TN_i}{TP_i+TN_i+FP_{i+FN_i}}\right) \tag{10}$$

The precision for each class is defined as:

$$Precision_i = \frac{TP_i}{TP_i+FP_i} \tag{11}$$

Recall for each class is defined as:

$$Recall_i = \frac{TP_i}{TP_i + FN_i} \tag{12}$$

While the F1-score is calculated as the harmonic mean of precision and recall, which is formulated as:

$$F1 - Score = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{13}$$

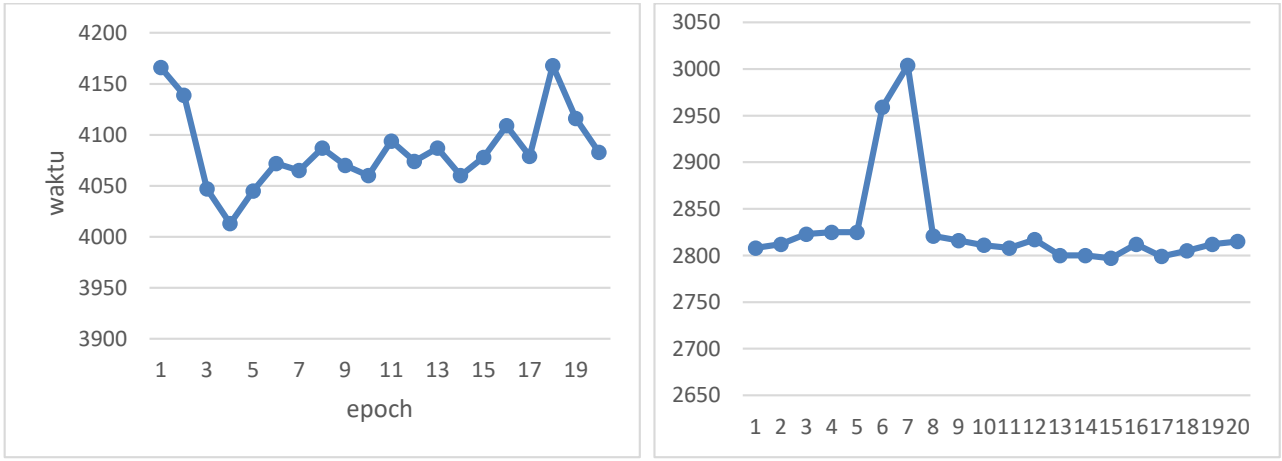
C. RESULT AND DISCUSSION

In the augmentation process, each epoch does not see the same image combination, because each batch produced is the result of a random transformation of the original image. So that augmentation succeeds in increasing the diversity of data trains without the need to increase the number of image files. In Figure 2 it can be seen that how the result of an image is augmented.



Figure 2. Augmentation

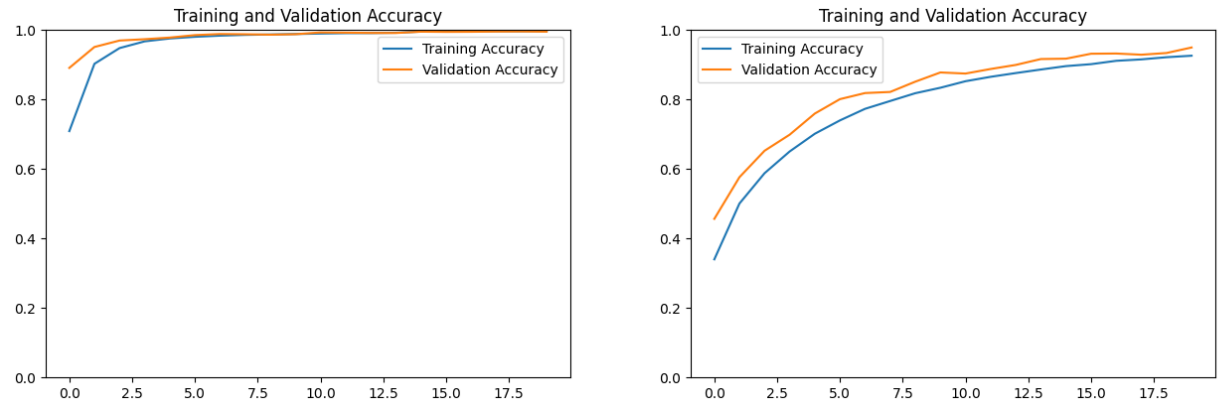
From the results of the training, the results of the length of training time in each epoch are presented in Figure 3, the accuracy value of each epoch can be seen in Figure 4, the value of loss in Figure 5 of both architectures. Then the values of precision, recall, and f-1score of each image shown in figure 6. And the best accuracy values can be seen in Table 4.



(a) (b)

Figure 3. (a) time per Xception epoch; (b) time per Convmixer epoch

Figures 3a and Figures 3b show the comparison of training time per epoch between the Xception and ConvMixer architectures. It can be seen that ConvMixer consistently requires a shorter training time than Xception. This is due to the simpler and lighter structure of the ConvMixer, as it uses a patch-based convolution approach with relatively fewer operations compared to the deeper convolution blocks on Xception. The lower computational complexity makes ConvMixer more efficient in terms of training time. These results show that ConvMixer has advantages in terms of computing efficiency, making it more suitable for use in environments with limited resources or fast processing needs.

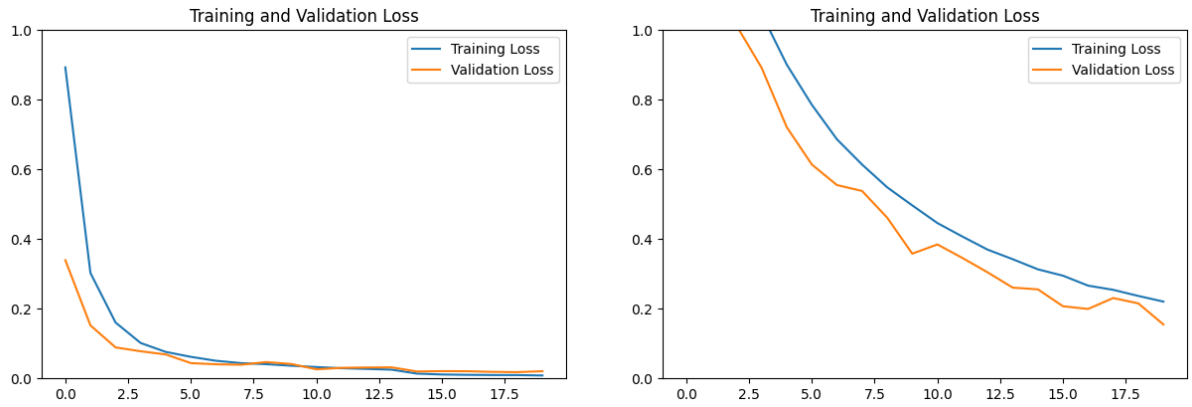


(a) (b)

Figure 4. (a) Xception Accuracy Training and Vaidasi; and (b) ConvMixer Accuracy Training and Training

Figure 4a shows that Xception was able to achieve high accuracy values in a relatively short time, characterized by a very sharp increase in the early epochs. This shows that the Xception architecture is very effective in extracting discriminatory features from skin disease images. This capability has to do with the use of depthwise separable convolution which allows for more efficient processing of spatial and inter-channel information. In contrast, in Figure 4b it can be seen that the ConvMixer has a more gradual increase in accuracy. This pattern indicates that the patch-based approach requires more epoch to be able to adjust to the complexity of

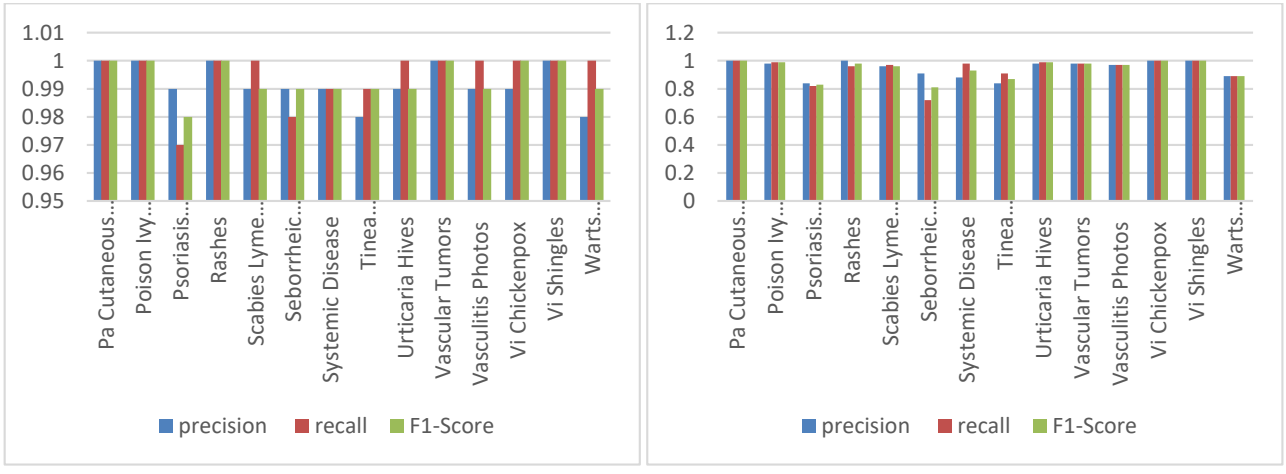
textures and visual patterns in dermatological imagery. However, the small gap between the training accuracy and validation curves in both models suggests that the applied data augmentation successfully reduces the risk of overfitting as well as improves the model's generalization capabilities.



(a) (b)

Figure 5. (a) Training and Vaidasi Loss Xception; and (b) Training and Vaidasi Loss Xception

In Figure 5a it can be seen that the loss value of Xception decreases very rapidly in the initial epochs and then becomes stable in subsequent epochs. This pattern shows that Xception has strong learning capabilities and can achieve convergence conditions faster. In contrast to Xception, Figure 5b shows that ConvMixer experiences a more gradual decrease in loss. Although the convergence is slower, a steady decrease in losses indicates that the learning process is proceeding consistently. This difference shows a trade-off between convergence speed and architectural simplicity, where Xception is faster to achieve optimal conditions, whereas ConvMixer requires longer training times.



(a) (b)

Gambar 6. (a) Classification Report Xception; and (b) Classification Report Convmixer

Figure 6a shows that Xception produces relatively uniform precision, recall, and F1-score values across the entire class. This uniformity indicates that the model has good generalization capabilities and does not only work optimally in certain classes. This is especially important in

medical classifications that are multiclass, as each class has the same level of importance. Figure 6b shows that ConvMixer is also capable of achieving high precision, recall, and F1-score, but the variation between classes tends to be greater than that of Xception, especially in recall values. This suggests that ConvMixer still needs further adjustments, such as increasing the number of epochs or modifying the architecture configuration, in order to recognize all classes more consistently.

Table 4. Accuracy Values

Architecture	Train Accuracy	Validation Accuracy	Test Accuracy
Xception	0.997	0.995	0.997
Convmixer	0.926	0.949	0.946

Table 4 shows that the Xception architecture produces higher accuracy values than ConvMixer in training, validation, and test data. These results indicate that Xception is more effective in extracting discriminatory features from skin disease imagery in the experimental scenarios used. However, ConvMixer still shows competitive performance with a fairly high accuracy value, so it still has the potential to be used as a lighter architectural alternative, especially in applications that prioritize computing efficiency.

The results of this study are in line with the research conducted by Sarı et al. which showed that deep learning-based methods and CNN are very effective for the classification of skin diseases. In their study, Sarı et al. reported that the combination of CNN architecture with feature selection using the Relief algorithm and SVM classification was able to achieve an accuracy of 92.10%, which suggests that the image of skin diseases has a visual pattern that can be optimally learned by the machine learning model. The findings support the results of this study, where the Xception architecture results in higher accuracy on training, validation, and test data, thus demonstrating that the end-to-end deep learning approach with modern CNN is capable of delivering highly competitive performance. The difference lies in the modelling strategy, where Sarı et al. explicitly utilize feature extraction and selection, while this study uses automated learning of features through a comparison of two modern architectures, namely Xception and ConvMixer. Thus, this study not only confirms the effectiveness of deep learning as demonstrated by Sarı et al., but also expands the study by evaluating the influence of network architecture design on accuracy, computational efficiency, and performance consistency between classes.

Although the results show promising performance, this study still has some limitations. First, out of a total of 34 classes available in the dataset, this study used only 14 classes, so the results obtained did not fully represent the overall variation of skin diseases. Second, the number of epoch and training parameters used is still limited and a thorough exploration of hyperparameter tuning has not been carried out. Third, the evaluation is only carried out on one dataset, so the generalization ability of the model against other datasets cannot be ascertained.

D. CONCLUSION AND SUGGESTIONS

Based on the results of the study, the Xception architecture showed superior performance to ConvMixer in the classification of skin disease images, with a test accuracy of 0.997 for Xception and 0.946 for ConvMixer. This suggests that Xception is more effective at extracting complex visual features and resulting in more stable classifications between classes. Practically, Xception is more suitable for use in image-based skin disease diagnosis systems that require a high level of accuracy and reliability. However, this study has several limitations, including the use of only 14 out of 34 available classes, limitations in exploring training parameters, and evaluation that is still limited to one dataset. Therefore, the results of this study cannot be generalized widely without further testing. For further research, it is recommended that all classes in the dataset be used, hyperparameter tuning is carried out to obtain a more optimal model configuration, and tests are carried out on other datasets. In addition, the development of a hybrid approach that combines CNN as a feature extractor with feature selection and classification methods such as SVM, as well as strengthening data augmentation strategies, has the potential to improve the accuracy and generalization capabilities of the model.

ACKNOWLEDGEMENT

The title for the thank you to the institution or the person who has contributed during the research and references is not numbered.

REFERENCES

Abulwafa, A. (2022). A Survey of Deep Learning Algorithms and its Applications. *Nile Journal of Comunication & Computer Science*, *3*(1), 28–49. https://doi.org/10.21608/njccs.2022.139054.1000

Ainurrohmah, & Wiyanti, D. T. (2023). Analisis Performa Algoritma Decision Tree , Naïve Bayes , K-Nearest Neighbor Untuk Klasifikasi Zona Daerah Risiko Covid-19 Di Indonesia Performance Analysis Of Decision Tree , Naïve Bayes , K-Nearest Neighbor Algorithm For Covid-19 Risk Zone Classificati. *Jurnal Teknologi Informasdi Dan Ilmu KOmputer(JTIIK)*, *10*(1), 115–122. https://doi.org/10.25126/jtiik.2023105935

Alomar, K., & Aysel, H. I. (2023). Data Augmentation in Classification and Segmentation : A Survey and New Strategies. *Journal Of Imaging*, *9*(2), 46. https://doi.org/10.3390/jimaging9020046

Awaluddin, B. A., Chao, C. T., & Chiou, J. S. (2023). Investigating Effective Geometric Transformation for Image Augmentation to Improve Static Hand Gestures with a Pre-Trained Convolutional Neural Network. *Mathematics*, *11*(23), 4783. https://doi.org/10.3390/math11234783

Bhatt, D., Patel, C., Talsania, H., Patel, J., Vaghela, R., Pandya, S., Modi, K., & Ghayvat, H. (2021). Cnn variants for computer vision: History, architecture, application, challenges and future scope. *Electronics (Switzerland)*, *10*(20), 2470. https://doi.org/10.3390/electronics10202470

Bracci, S., Mraz, J., Zeman, A., Leys, G., & de Beeck, H. O. (2023). The representational hierarchy in human and artificial visual systems in the presence of object-scene regularities. *PLoS Computational Biology*, *19*(4), Article e1011086. https://doi.org/10.1371/journal.pcbi.1011086

Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition(CVPR 2017)*, 1800–1807. https://doi.org/10.1109/CVPR.2017.195

Hao, X., Liu, L., Yang, R., Yin, L., Zhang, L., & Li, X. (2023). A Review of Data Augmentation Methods of Remote Sensing Image Target Recognition. *Remote Sensing*, *15*(3), 827. https://doi.org/10.3390/rs15030827

Hendrycks, D., & Gimpel, K. (2016). *Gaussian Error Linear Units (GELUs).* ArXiv.

Ibrahem, H., Salem, A., & Kang, H. (2025). Pixel shuffling is all you need : spatially aware convmixer for dense prediction tasks. *Pattern Recognition*, *158*, 111068.

https://doi.org/10.1016/j.patcog.2024.111068

Iijima, R., & Kiya, H. (2022). *An Encryption Method of ConvMixer Models without Performance Degradation.* ArXiv.

Khoei, T. T., Slimane, H. O., & Kaabouch, N. (2023). Deep learning : systematic review , models , challenges , and research directions. *Neural Computing and Applications*, *35*(31), 23103–23124. https://doi.org/10.1007/s00521-023-08957-4

Krichen, M. (2023). Convolutional Neural Networks: A Survey. *Computers*, *12*(8), 151. https://doi.org/10.3390/computers12080151

Lin, H., Imaizumi, S., & Kiya, H. (2024). Privacy-Preserving ConvMixer Without Any Accuracy Degradation Using Compressible Encrypted Images. *Information*, *15*(11), 723. https://doi.org/doi.org/10.3390/info15110723

Lukasz, K., Gomez, A. N., & Chollet, F. (2017). *Depthwise Separable Convolutions for Neural Machine Translation*. ArXiv.

Mienye, I. D., & Swart, T. G. (2024). A Comprehensive Review of Deep Learning : Architectures , Recent Advances , and Applications. *Informatics*, *15*(12), 755. https://doi.org/doi.org/10.3390/info15120755

Muhammad, W., Aramvith, B., & Onoye, T. (2021). Multi-scale Xception based depthwise separable convolution for single image super- resolution. *Plos One*, *16*(8), e0249278. https://doi.org/doi.org/10.1371/journal.pone.0249278

Mumuni, A., & Mumuni, F. (2022). Data augmentation: A comprehensive survey of modern approaches. *Array*, *16*, 100258. https://doi.org/10.1016/j.array.2022.100258

Nawar, S., Joty, T. A., & Hashem, M. M. A. (2024). A Lightweight Deep Learning Architecture for Efficient Multimodal Medical A Lightweight Deep Learning Architecture for Efficient Multimodal Medical Image Segmentation Using Attention Mechanism. *ICCA '24: Proceedings of the 3rd International Conference on Computing Advancements*, *October*, 970–977. https://doi.org/doi.org/10.1145/3723178.3723307

Raj, R., & Kos, A. (2025). An Extensive Study of Convolutional Neural Networks: Applications in Computer Vision for Improved Robotics Perceptions. *Sensors*, *25*(4), 1033. https://doi.org/doi.org/10.3390/s25041033

Rangel, G., Cuevas-Tello, J. C., Nunez-Varela, J., Puente, C., & Silva-Trujillo, A. G. (2024). A Survey on Convolutional Neural Networks and Their Performance Limitations in Image Recognition Tasks. *Journal of Sensors*, *2024*(1), 2797320. https://doi.org/10.1155/2024/2797320

Sarı, M. O., & Keser, K. (2025). Classification of skin diseases with deep learning based approaches. *Scientific Reports*, *15*(1), 27506. https://doi.org/https://doi.org/10.1038/s41598-025-13275-x

Sathya, R., Mahesh, T. R., Bhatia Khan, S., Malibari, A. A., Asiri, F., Rehman, A. ur, & Malwi, W. Al. (2024). Employing Xception convolutional neural network through high-precision MRI analysis for brain tumor diagnosis. *Frontiers in Medicine*, *11*(1), 1487713. https://doi.org/https://doi.org/10.3389/fmed.2024.1487713

Solano, A., Dietrich, K. N., Martínez-Sober, M., Barranquero-Cardeñosa, R., Vila-Tomás, J., & Hernández-Cámara, P. (2023). Deep Learning Architectures for Diagnosis of Diabetic Retinopathy. *Applied Sciences (Switzerland)*, *13*(7), 4445. https://doi.org/10.3390/app13074445

Supriyono, Prasetya, A., Suyono, & Kurniawan, F. (2024). Telematics and Informatics Reports Advancements in natural language processing : Implications , challenges , and future directions. *Telematics and Informatics Reports*, *16*, 100173. https://doi.org/10.1016/j.teler.2024.100173

Tempfli, L., & Sándor, C. (2024). HierNet: Image Recognition with Hierarchical Convolutional Networks. *International Conference on Agents and Artificial Intelligence*, *2*(Icaart), 147–155. https://doi.org/10.5220/0012321100003636

Terven, J. R., Cordova-esparza, D. M., Ramirez-pedraza, A., Chavez-urbiola, E. A., & Romero-gonzalez, J. A. (2025). Loss Functions And Metrics In Deep Learning. *Springer Artificial Intelligence Review*, *58*(6), 195. https://doi.org/https://doi.org/10.1007/s10462-025-11198-7

Trockman, A., & Kolter, J. Z. (2022). Patches Are All You Need? *ArXiv*. https://doi.org/https://doi.org/10.48550/arXiv.2201.09792

Üzen, H., & Fırat, H. (2024). A hybrid approach based on multipath Swin transformer and ConvMixer for white blood cells classification. *Health Information Science and Systems*, *12*(1), 33.

https://doi.org/10.1007/s13755-024-00291-w

Wang, J., & Perez, L. (2017). *The Effectiveness of Data Augmentation in Image Classification using Deep Learning.* ArXiv. https://doi.org/arXiv:1712.04621v1

Wang, Y., Han, Y., Wang, C., Song, S., Tian, Q., & Huang, G. (2023). *Computation-efficient Deep Learning for Computer Vision :* ArXiv. https://doi.org/arXiv:2308.13998v1

Younesi, A., Ansari, M., Fazli, M., Ejlali, A., Shafique, M., & Henkel, J. (2024). A Comprehensive Survey of Convolutions in Deep Learning: Applications, Challenges, and Future Trends. *IEEE Access*, *12*(pp), 41180–41218. https://doi.org/10.1109/ACCESS.2024.3376441

Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). *An Image Is Worth 16x16 Words: Transformers For Image Recognition At Scale.* ArXiv. https://doi.org/arXiv:2010.11929v2