PROCEEDINGS OF
INTERNATIONAL CONFERENCE ON NEW TRENDS IN APPLIED SCIENCES
https://proceedings.icontas.org/
International Conference on New Trends in Applied Sciences (ICONTAS'23), Konya, December 1-3, 2023.

COMPARISON OF VISION TRANSFORMERS AND CONVOLUTIONAL NEURAL NETWORKS FOR SKIN DISEASE CLASSIFICATION

Muhammet Fatih ASLAN
Electrical and Electronics Engineering, Karamanoglu Mehmetbey University, Karaman, Turkey,
https://orcid.org/0000-0001-7549-0137
mfatihaslan@kmu.edu.tr

ABSTRACT: Skin diseases are one of the most common diseases in humans. Due to its many various symptoms and types, computer vision studies have been frequently applied to its diagnosis and classification. Previous studies have frequently used machine learning methods and deep learning-based Convolutional Neural Networks (CNN) for skin disease diagnosis. Although deep learning-based applications have achieved great success in terms of detection accuracy, research continues to ensure the desired performance. However, Vision Transformer (ViT), recently proposed as a competitive alternative to CNNs, is gaining increasing popularity. This paper compares ResNet18 and ResNet50 networks, important CNN models, with ViT for classifying skin disease. The comparison is applied on a dataset containing a small number of samples. In the application performed on a dataset containing skin disease images, ViT provides 68.93% classification accuracy, while ResNet18 and ResNet50 classification accuracy is 61.65% and 61.17%, respectively. Other metrics calculated along with the accuracies also prove the superiority of ViT over ResNet models. However, ViT has a big disadvantage in terms of training time.

Key words: Convolutional neural networks, machine learning, ResNet18, ResNet50, skin disease, vision transformers

INTRODUCTION

Skin is one of the most important organs of the human body, covering the human body and protecting it from infections and harmful heat and light sources. If a disease occurs on the skin, the life of the person is negatively affected and the body remains unprotected against harmful infections that may come from outside (Verma et al., 2019). In addition, skin disease can turn into malignant tissues. Therefore, it is very important to protect skin health and treat it on time (Srinivasu et al., 2021). Despite these, unfortunately, skin disease is a serious global public health problem that is the most common worldwide (Karimkhani et al., 2017). It has been known as the most common disease worldwide since 1970 (Allugunti, 2022).

Skin diseases have a wide variety of symptoms, and symptoms can take a long time to change. The presence of the disease may not be recognized due to the texture of the skin and the hair on the skin. Additionally, most people neglect changes to their skin, which can lead to permanent skin deterioration or even worse, skin cancer (Wei et al., 2023). Although some types of skin diseases are very rare, most are common. Diagnosing and classifying these skin diseases is a very complex task (Swamy & Divya, 2021). Despite all this, dermatologists make the diagnosis with the naked eye in most non-invasive screening tests. This may cause skin diseases, which have many different types, to be overlooked (Sreekala et al., 2022). For this reason, a system that automatically detects skin disease has recently become an active field of study in order to reduce the workload of dermatologists, minimize human errors and provide more accurate diagnosis (Velasco et al., 2019).

Developing technology and data-based artificial intelligence applications play an important role in the analysis of skin diseases, as in many areas. The first artificial intelligence-based studies in this field are based on machine learning methods. Alquran et al. (2017) implemented a Support Vector Machine (SVM) based application for melanoma skin cancer detection and classification. They extracted texture (Co-occurrence Matrix (GLCM)), color and shape-dependent features from skin images and applied feature selection with Principal Component Analysis (PCA). At the end of the application, they distinguished normal and abnormal images with 92.1% accuracy. In another machine learning-based study, Janney et al. (2018) extracted texture and wavelet

features from skin images. They classified these features with SVM, Naive Bayes and Artificial Neural Network (ANN). At the end of the study, ANN provided the highest accuracy (89%).

The rapid development of deep learning-based techniques, thanks to the increasing parallel data processing capabilities of computers in the last decade, has brought convolutional neural networks (CNN) to the fore, especially in computer vision applications. In particular, CNNs have rapidly become the preferred method in medical image analysis due to the high performance they provide. In particular, its end-to-end structure and the elimination of the feature extraction step have made CNN models popular (Janney et al., 2018). Anand, Gupta, Koundal, et al. (2022) classified skin disease images by making some modifications to the original Xception CNN model and achieved a classification accuracy of 96.40%. Anand, Gupta, Nayak, et al. (2022) classified skin images with ResNet50, DenseNet121, VGG16 and ResNet18 models. Epoch, batch size and optimizers were modified to choose the best hyperparameters for the models. At the end of the study, ResNet models showed the highest performance with 90% accuracy.

While the advantages of CNN models in terms of performance and convenience are obvious, CNNs cannot capture long-term information or dependencies efficiently. Therefore, CNN performs an evaluation regardless of the content. In classifying image data, the architecture called Vision Transformers (ViT) (Dosovitskiy et al., 2020) has provided impressive results in various applications (Akinyelu et al., 2022; Aleissaee et al., 2023). ViTs have produced better results than CNNs on computer vision tasks in a wide range of applications (Li et al., 2022).

Although ViT-based image classification applications are still new, new studies in this field are constantly being developed by researchers. The number of ViT-based studies on skin disease is still low. Arshed et al. (2023) after performing pre-processing (normalization, resizing, etc.) on skin images, classified these images with both ViT and different CNN models. They applied fine-tuning, transfer learning, and data augmentation techniques before classification. At the end of the experimental study, ViT provided superior results with an accuracy of 92.14%. The ViT model was followed by the ResNet50 model with 82% accuracy. Aladhadh et al. (2022) first applied pre-processing such as brightness and contrast adjustment on skin images. They then expanded the dataset with data augmentation techniques. They fed this augmented data into the ViT model. At the end of the study, ViT achieved 96.14% classification accuracy, outperforming previous different models.

This paper prefers a dataset with a small number of samples for skin disease classification. The aim is to compare ViT and ResNet models (ResNet18 and ResNet50) on low-dimensional data. First, data augmentation steps are applied to skin images to diversify the data. The augmented data is then fed into both the ViT and ResNet models. Although the results show that ViT is more successful in terms of accuracy, ResNet models are superior in terms of computational cost and training time.

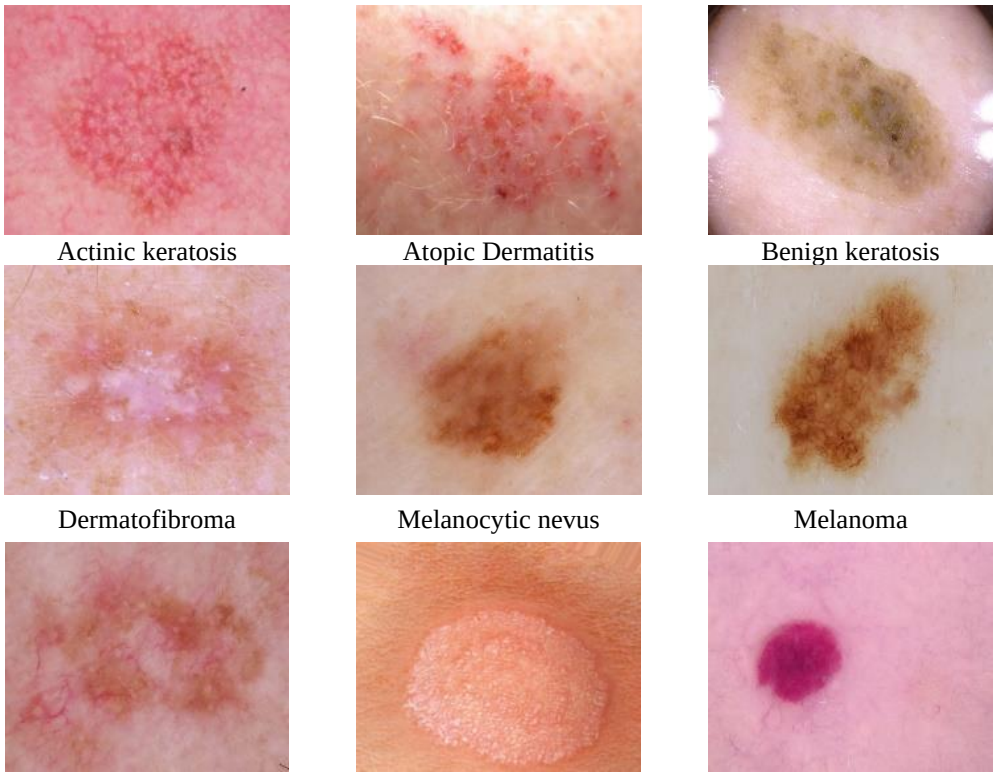

Figure 1. Some skin disease images in the dataset

MATERIALS AND METHODS

Dataset

This data set consists of skin images. These images are divided into training and validation. The names of the skin disease types in the dataset are as follows: Actinic keratosis, Atopic Dermatitis, Benign keratosis, Dermatofibroma, Melanocytic nevus, Melanoma, Squamous cell carcinoma, Tinea Ringworm Candidiasis, Vascular lesion. The sample numbers of these disease types are 80, 81,80, 80, 80, 80, 80, 56 and 80, respectively, for the training data. The sample numbers for validation data are 20, 21, 20, 20, 20, 20, 20, 20 and 20, respectively. As a result, the number of training data is 697, the number of validation data is 181 and the total number of classes is 9. Sample images of each skin disease are shown in Figure 1. Access to this dataset can be provided via the link below.

https://www.kaggle.com/datasets/riyaelizashaju/skin-disease-classification-image-dataset

Vision Transformer

The inability of CNN models to learn long-term dependencies during image classification is a disadvantage in terms of the performance of the network. Therefore, CNN models cannot comprehensively evaluate the input and resolve relationships between different regions of the image (Katar & Yildirim, 2023). ViT, developed as an alternative to this shortcoming of CNN, is a deep learning model that uses self-attention mechanisms to analyze relationships in different regions of the image. ViT was developed inspired by the transformer architecture (Vaswani et al., 2017), which broke new ground in the Natural Language Processing (NLP) problem.

Vision Transformers architecture adapts the technique applied by the transformer network in NLP methods to images. The image is divided into non-overlapping two-dimensional patches, and vectors are formed from the features of these linearly arranged patches. In addition to these features, positional information of the patches in the image is also added. These vectors and positional information are given as input to the transformer encoder. Self-attention modules contained in this encoder capture long-term dependencies (Maurício et al., 2023). Unlike CNNs, ViTs are generally more data-hungry, and often require the use of datasets containing millions of data (Cai et al., 2023)

Data Augmentation

Deep architectures such as CNN and ViT generally require large datasets to ensure successful and stable classification success. Obtaining a large amount of data is often difficult in real life. For this reason, computer science researchers often perform augmentation of data artificially in a computer environment. It is aimed to increase the success of the network and ensure data diversity with new synthetically generated data (Unlersen et al., 2022).

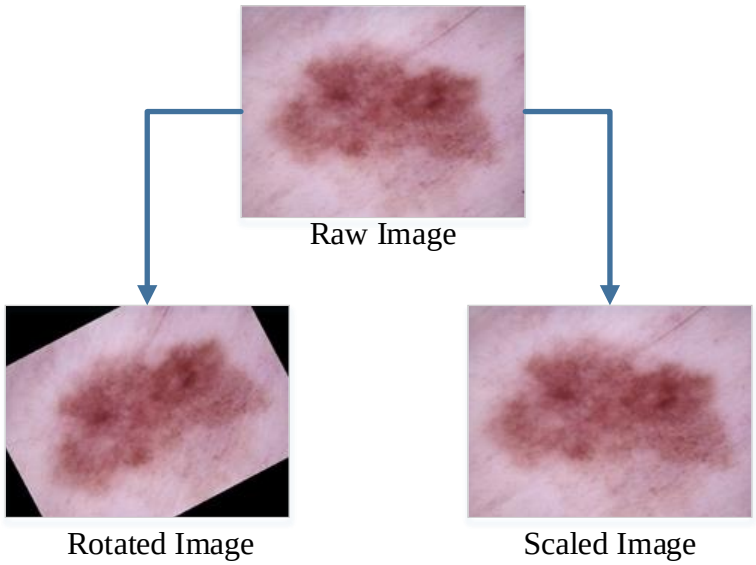

Figure 2. Data augmentation methods

Figure 2 shows a raw image of the dataset used and the new images resulting from the data augmentation steps applied to this image. As can be seen from Figure 2, two different data augmentation techniques are used in this study: rotation and scaling. These techniques are applied to all raw images at random values. The number of samples in the dataset before and after the data augmentation step is shown in Table 1.

Table 1. Changes in the number of data sets after data augmentation

	Number of Train Data	Number of Validation Data	Total
Before data augmentation	697	181	878
After data augmentation	1845	205	2050

RESULTS

In this section, the classification results obtained with ViT and ResNet models will be compared and analyzed.

Classification with ViT

The augmented data was first fed into the ViT model. Hyperparameter values of this model are as in Table 2. The ViT neural network used is the model proposed by Dosovitskiy et al. (2020). The patch size is 16 and the input image size is 384x384. Instead of training the network from scratch, the network trained on the ImageNet 2012 dataset is fine-tuned. This complex network contains more than 86 million parameters. Since the number of classes is 9, the number of fully connected layer outputs is changed to 9. 10% of all data is reserved for testing. Afterwards, training is started. The accuracy and loss graphs obtained after training are shown in Figure 3. As can be seen from the graph, as the number of iterations increases, the loss decreases and accuracy increases. This is valid for both training and test data. As a result, the network successfully performs the learning process during the training. At the end of the training, the accuracy of the test data is 68.93%. Total training time is 241 minutes.

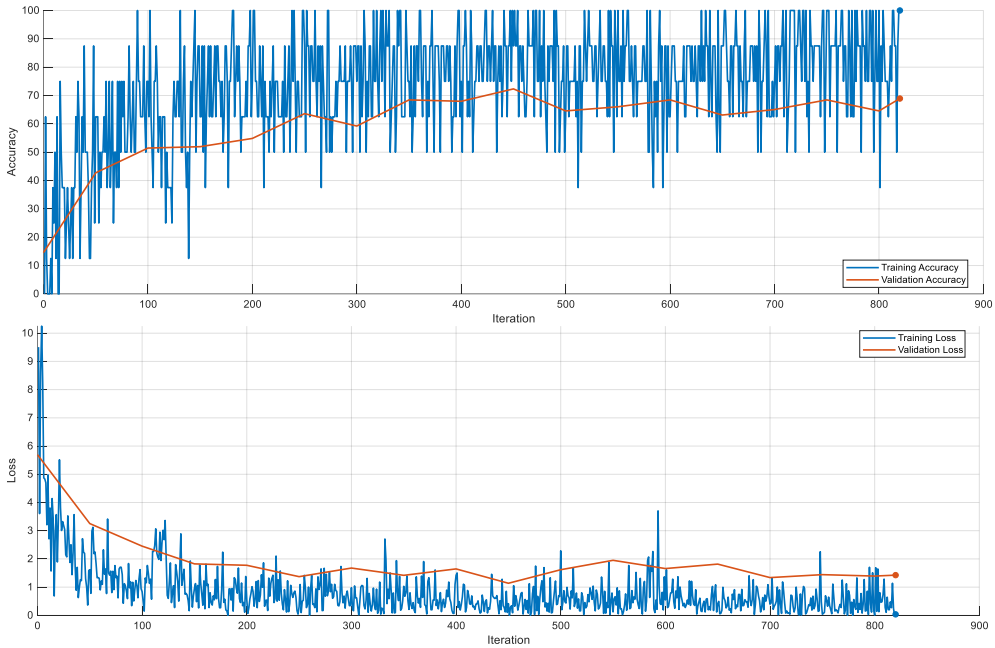

Figure 3. Accuracy and loss graphs obtained with ViT

The confusion matrix as a result of the classification obtained with the test data is shown in Figure 4. When the confusion matrix was examined, although 100% classification accuracy was achieved for some classes, no sample in the Melanoma class could be classified correctly. The similarity of images of skin diseases, lack of data diversity and low number of data are the reasons for failure in classification. Different performance metrics are also calculated based on the confusion matrix shown in Figure 4. These metric values are presented in Table 3.

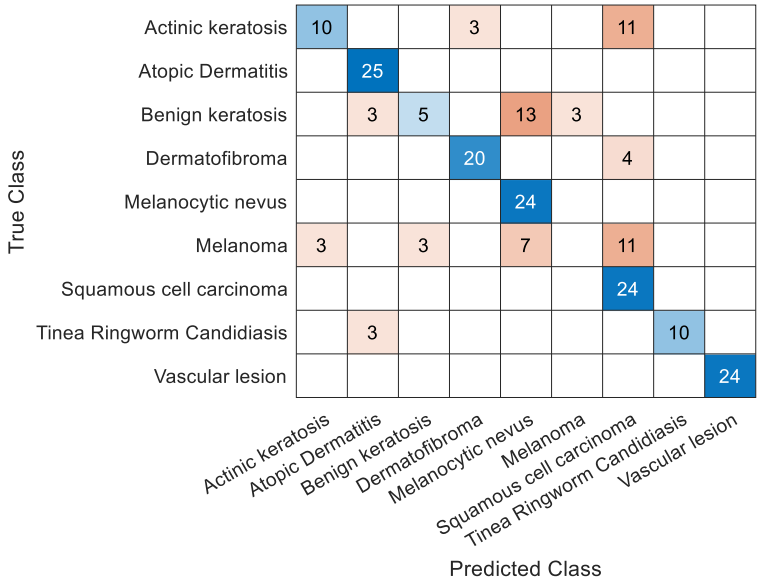

Figure 4. Confusion matrix of test data as a result of ViT

Classification with ResNet18 and ResNet50

The hyperparameters used in the ViT neural network are also valid in the ResNet model. These values are shown in Table 2. Using residual connections and thus solving the vanishing gradient problem has made ResNet models popular among CNN models. This study used two frequently used ResNet models, ResNet18 and ResNet50, to classify skin disease. After the data augmentation step, 10% of the entire data is reserved for testing. Then, ResNet18 and ResNet50 are trained with the augmented skin images. The accuracy and loss graphs obtained for both CNN models after training are shown in Figure 5 and Figure 6.

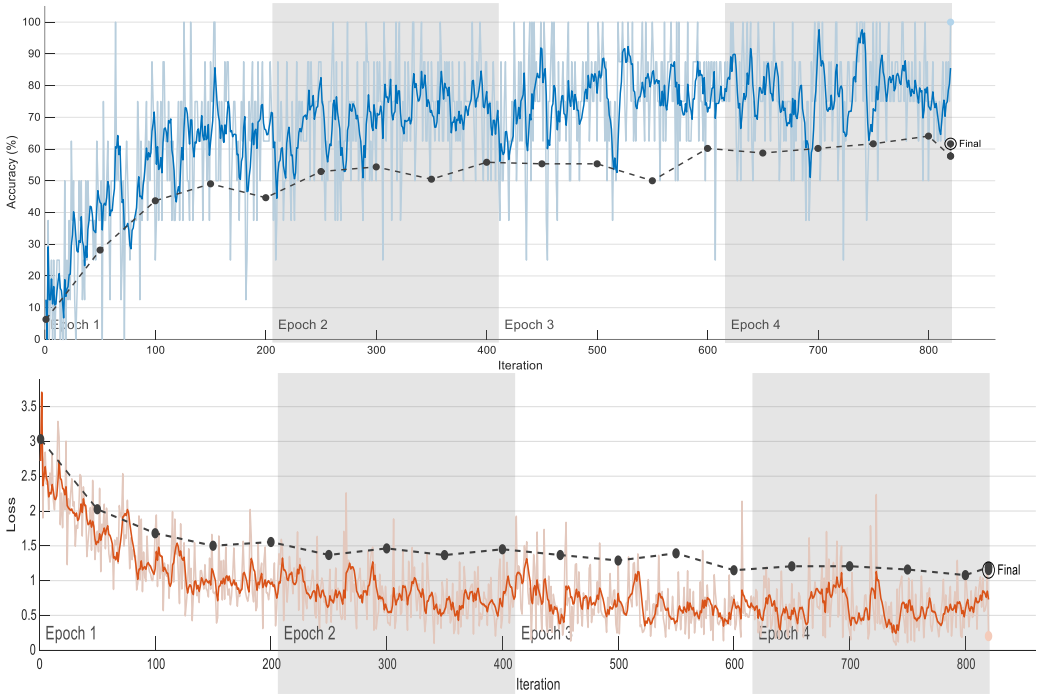

Figure 5. Accuracy and loss graphs obtained with ResNet18

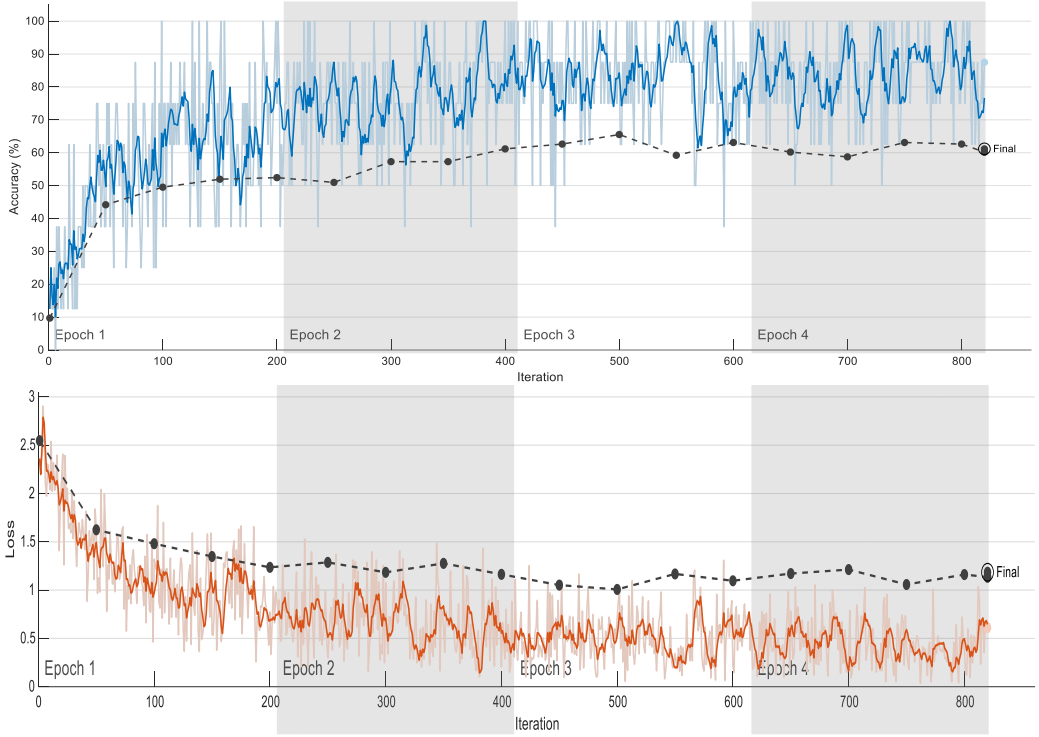

Figure 6. Accuracy and loss graphs obtained with ResNet50

When Figure 5 and Figure 6 are examined, as the number of iterations increases, the accuracy of the training and test data increases over time and the error decreases over time. Therefore, it appears that the network is learning. After training, the test data is classified with an accuracy of 61.65% and 61.17% for the ResNet18 and ResNet50 model. Figure 7 and Figure 8 show the confusion matrices for ResNet18 and ResNet50, respectively. When the classification success for each class is examined, similar to the ViT results, the most misclassified type is Melanoma for ResNet18. ResNet50 correctly classified only two samples of Benign Keratosis type. Achievement in different classes is highly variable. Table 3 contains the values of metrics showing the classification performance of both models. According to these results, both models provide similar accuracies. Training times for ResNet18 and ResNet50 are approximately 1 and 4 minutes, respectively.

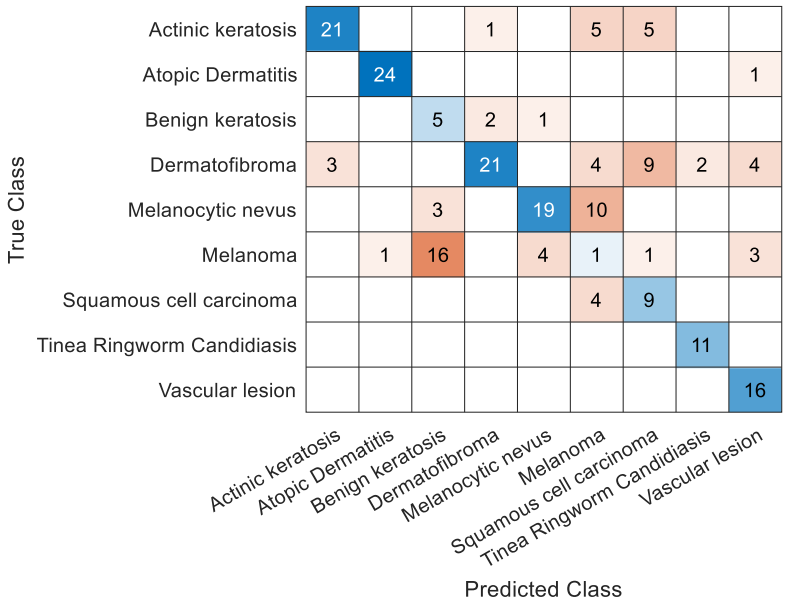

Figure 7. Confusion matrix of test data as a result of ResNet18

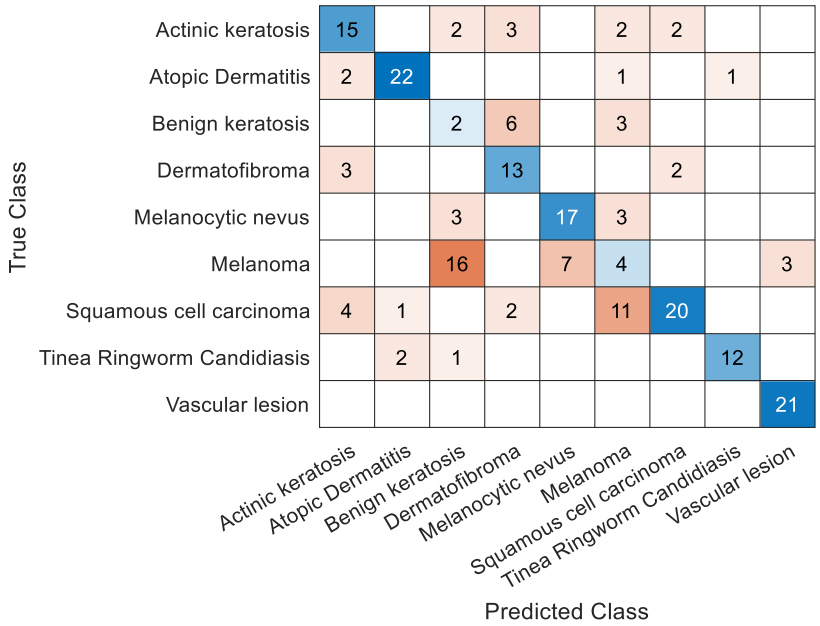

Figure 8. Confusion matrix of test data as a result of ResNet50

Table 2. Training options for ViT and ResNet models

Hyperparameter	Mini Batch Size	Epoch	Learning Rate
Value	8	4	0.0001

Table 3. Performance metrics for test data of ViT, ResNet18 and ResNet50 models

Metric	ViT	ResNet18	ResNet18
Precision	0.6773	0.6266	0.6263
Recall	0.6920	0.6727	0.6193
Accuracy	0.6893	0.6165	0.6117
Specificity	0.9609	0.9529	0.9517
F1score	0.6468	0.6190	0.6143

DISCUSSION

Early diagnosis of skin disease is important. However, the fact that there are many variants, the symptoms are diverse, and some diseases are very similar to each other make the dataset stand out in such studies. The dataset used in this study includes 9 species, but the number of samples is quite small. To successfully perform difficult classification tasks with deep architectures, the number of samples must be large. Experimental results show that ViT has a higher average classification success. However, in terms of training time, ResNet models show a clear superiority. Moreover, if the number of data is much larger, this difference may reach a level that will increase the preferability of ResNet models.

CONCLUSION

This study performed a comparison of ViT and ResNet models for a skin disease dataset containing few samples. First of all, the number of raw data was increased by data augmentation. Then, fine-tuning settings were made for deep networks. Afterwards, training and testing steps were carried out. The results showed that ResNet and ViT could not produce very successful results on a dataset containing a small number of samples. However, the classification accuracy of ViT is higher against both ResNet models.

REFERENCES

Akinyelu, A. A., Zaccagna, F., Grist, J. T., Castelli, M., & Rundo, L. (2022). Brain Tumor Diagnosis Using Machine Learning, Convolutional Neural Networks, Capsule Neural Networks and Vision Transformers, Applied to MRI: A Survey. *Journal of Imaging*, *8*(8), 205. https://www.mdpi.com/2313-433X/8/8/205

Aladhadh, S., Alsanea, M., Aloraini, M., Khan, T., Habib, S., & Islam, M. (2022). An Effective Skin Cancer Classification Mechanism via Medical Vision Transformer. *Sensors*, *22*(11), 4008. https://www.mdpi.com/1424-8220/22/11/4008

Aleissaee, A. A., Kumar, A., Anwer, R. M., Khan, S., Cholakkal, H., Xia, G.-S., & Khan, F. S. (2023). Transformers in Remote Sensing: A Survey. *Remote Sensing*, *15*(7), 1860. https://www.mdpi.com/2072-4292/15/7/1860

Allugunti, V. R. (2022). A machine learning model for skin disease classification using convolution neural network. *International Journal of Computing, Programming and Database Management*, *3*(1), 141-147.

Alquran, H., Qasmieh, I. A., Alqudah, A. M., Alhammouri, S., Alawneh, E., Abughazaleh, A., & Hasayen, F. (2017). The melanoma skin cancer detection and classification using support vector machine. 2017 IEEE Jordan Conference on Applied Electrical Engineering and Computing Technologies (AEECT),

Anand, V., Gupta, S., Koundal, D., Nayak, S. R., Nayak, J., & Vimal, S. (2022). Multi-class skin disease classification using transfer learning model. *International Journal on Artificial Intelligence Tools*, *31*(02), 2250029.

Anand, V., Gupta, S., Nayak, S. R., Koundal, D., Prakash, D., & Verma, K. D. (2022). An automated deep learning models for classification of skin disease using Dermoscopy images: a comprehensive study. *Multimedia Tools and Applications*, *81*(26), 37379-37401. https://doi.org/10.1007/s11042-021-11628-y

Arshed, M. A., Mumtaz, S., Ibrahim, M., Ahmed, S., Tahir, M., & Shafi, M. (2023). Multi-Class Skin Cancer Classification Using Vision Transformer Networks and Convolutional Neural Network-Based Pre-Trained Models. *Information*, *14*(7), 415. https://www.mdpi.com/2078-2489/14/7/415

Cai, G., Zhu, Y., Wu, Y., Jiang, X., Ye, J., & Yang, D. (2023). A multimodal transformer to fuse images and metadata for skin disease classification. *The Visual Computer*, *39*(7), 2781-2793. https://doi.org/10.1007/s00371-022-02492-4

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., & Gelly, S. (2020). An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Janney, B. J., Roslin, S. E., & Shelcy, M. J. (2018). A comparative analysis of skin cancer detection based on svm, ann and naive bayes classifier. 2018 International Conference on Recent Innovations in Electrical, Electronics & Communication Engineering (ICRIEECE),

Karimkhani, C., Dellavalle, R. P., Coffeng, L. E., Flohr, C., Hay, R. J., Langan, S. M., Nsoesie, E. O., Ferrari, A. J., Erskine, H. E., & Silverberg, J. I. (2017). Global skin disease morbidity and mortality: an update from the global burden of disease study 2013. *JAMA dermatology*, *153*(5), 406-412.

Katar, O., & Yildirim, O. (2023). An Explainable Vision Transformer Model Based White Blood Cells Classification and Localization. *Diagnostics*, *13*(14), 2459. https://www.mdpi.com/2075-4418/13/14/2459

Li, Y., Yuan, G., Wen, Y., Hu, J., Evangelidis, G., Tulyakov, S., Wang, Y., & Ren, J. (2022). Efficientformer: Vision transformers at mobilenet speed. *Advances in Neural Information Processing Systems*, *35*, 12934-12949.

Maurício, J., Domingues, I., & Bernardino, J. (2023). Comparing Vision Transformers and Convolutional Neural Networks for Image Classification: A Literature Review. *Applied Sciences*, *13*(9).

Sreekala, K., Rajkumar, N., Sugumar, R., Sagar, K. V. D., Shobarani, R., Krishnamoorthy, K. P., Saini, A. K., Palivela, H., & Yeshitla, A. (2022). Skin Diseases Classification Using Hybrid AI Based Localization Approach. *Computational Intelligence and Neuroscience*, *2022*, 6138490. https://doi.org/10.1155/2022/6138490

Srinivasu, P. N., SivaSai, J. G., Ijaz, M. F., Bhoi, A. K., Kim, W., & Kang, J. J. (2021). Classification of Skin Disease Using Deep Learning Neural Networks with MobileNet V2 and LSTM. *Sensors*, *21*(8), 2852. https://www.mdpi.com/1424-8220/21/8/2852

Swamy, K. V., & Divya, B. (2021, 16-17 Dec. 2021). Skin Disease Classification using Machine Learning Algorithms. 2021 2nd International Conference on Communication, Computing and Industry 4.0 (C2I4),

Unlersen, M. F., Sonmez, M. E., Aslan, M. F., Demir, B., Aydin, N., Sabanci, K., & Ropelewska, E. (2022). CNN–SVM hybrid model for varietal classification of wheat based on bulk samples. *European Food Research and Technology*, *248*(8), 2043-2052. https://doi.org/10.1007/s00217-022-04029-4

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

Velasco, J., Pascion, C., Alberio, J. W., Apuang, J., Cruz, J. S., Gomez, M. A., Molina Jr, B., Tuala, L., Thio-ac, A., & Jorda Jr, R. (2019). A smartphone-based skin disease classification using mobilenet cnn. *arXiv preprint arXiv:1911.07929*.

Verma, A. K., Pal, S., & Kumar, S. (2019). Classification of Skin Disease using Ensemble Data Mining Techniques. *Asian Pac J Cancer Prev*, *20*(6), 1887-1894. https://doi.org/10.31557/apjcp.2019.20.6.1887

Wei, M., Wu, Q., Ji, H., Wang, J., Lyu, T., Liu, J., & Zhao, L. (2023). A Skin Disease Classification Model Based on DenseNet and ConvNeXt Fusion. *Electronics*, *12*(2), 438. https://www.mdpi.com/2079-9292/12/2/438