Bitlis Eren Üniversitesi Fen Bilimleri Dergisi
BİTLİS EREN UNIVERSITY JOURNAL OF SCIENCE
ISSN: 2147-3129/e-ISSN: 2147-3188
VOLUME: 13 NO: 3 PAGE: 844-850 YEAR: 2024
DOI:10.17798/bitlisfen.1505636

# Automatic Classification of Melanoma Skin Cancer Images with Vision Transform Model and Transfer Learning

Alper Talha KARADENIZ*

*Department of Computer Engineering, Faculty of Engineerin and Architecture, Kahramanmaras Sutcu Imam University, Kahramanmaras, TR*
*(ORCID: 0000-0003-4165-3932)*

**Keywords:** Melanoma, Deep learning, Vision Transformer, Classification

**Abstract**
Melanoma is one of the most aggressive and lethal forms of skin cancer. Therefore, early diagnosis and correct diagnosis are very important for the health of the patient. Cancer diagnosis is made by field experts and this increases the possibility of error. Today, with the developing deep learning technology, it has been seen that automatic detection of Melanoma skin cancer can be performed with high accuracy by computer systems.

One of the latest technologies developed in the field of deep learning is the Vision Transformer (ViT) model. This model was produced by Google and has achieved very successful results in the field of classification. This study aims to detect melanoma skin cancer with high accuracy using the ViT model.

In the study, the melanoma skin cancer dataset consisting of 9600 training and 1000 test images in the Kaggle library was used. In order to use the data set more effectively, some pre-processing methods were first applied. Model performance was evaluated using the transfer learning approach together with the ViT model on this data set. Training and experimental testing of the model was carried out with Python language on the Colab platform.

As a result of the experimental studies carried out on the test data set, it was seen that the model reached 93.5% accuracy rate. This rate is competitive and promising when compared to existing models in the literature.

## 1. Introduction

Melanoma, a malignant tumor arising from melanocytes, is one of the most aggressive and lethal forms of skin cancer. Early diagnosis and accurate diagnosis are of great importance for effective treatment and improving patient survival rates. Traditional methods include visual examinations and dermoscopy performed by dermatologists. However, these methods rely heavily on the expertise and experience of the practitioner. This increases the potential for human error and calls for more reliable diagnostic tools. The development of computer-aided diagnostic systems is crucial to meet this need [1].

In recent years, advances in machine learning and deep learning technologies are opening new avenues for automatic detection and classification of melanoma from skin lesion images. These technologies show promise in improving diagnostic accuracy and efficiency. Deep learning models are comparable to human experts, especially when trained on large data sets [2].

Vision Transformer (ViT) is a model developed by Google that achieves revolutionary results in the field of computer vision processing. ViT processes these tokens with the Transformer encoder by dividing the images into fixed-size patches and embedding each patch into a linear token. This method effectively captures spatial relationships and contextual information within the image. It shows superior performance, especially in large-scale data sets [3]. However, the success of Transformer-based models

* Corresponding author: alperkaradeniz5@gmail.com

*Received:27.06.2024, Accepted:29.07.2024*

in natural language processing has encouraged the application of these models to image processing tasks. Vision Transformer (ViT), developed by Dosovitskiy et al., has emerged as an alternative to CNNs in image classification tasks. ViT has shown state-of-the-art results, outperforming traditional CNN architectures, especially when pre-trained on large datasets [4].

Many studies have been conducted on automatic diagnosis and classification of melanoma skin cancer. Among traditional approaches, image processing and machine learning methods stand out. In the literature, deep learning methods have been widely used in the classification of skin lesions and successful results have been obtained.

Kaur et al demonstrated that they achieved dermatologist-level accuracy in skin cancer classification using deep CNNs. In the study, dermoscopic images containing different cancer samples were obtained from the International Skin Imaging Cooperation data repositories (ISIC 2016, ISIC2017 and ISIC 2020). The proposed DCNN classifier achieved 81.41%, 88.23% and 90.42% accuracy on ISIC 2016, 2017 and 2020 datasets, respectively [5].

Prakhar Shobhit et al. He applied the Vision Transformer Attention method on the melanoma skin cancer data set and achieved a 91% accuracy rate as a result of the classification [6]. Another study proposes an alternative method based on ready-to-use ViT to identify various skin cancer diseases. To evaluate its performance, the proposed method was compared with 11 CNN-based transfer learning methods that are known to perform better than deep learning techniques in the literature. The proposed model achieved an accuracy of 92.14%, outperforming CNN-based transfer learning models [7].

In our study, after pre-processing the data set, classification of melanoma skin cancer was carried out by fine-tuning it with the google/vit-large-patch32-384 transfer learning method suggested by the ViT model. The rest of this study is as follows: In the second section, the data set and the methods used are mentioned. In the third section, the proposed model is explained. The fourth section includes the results of experimental studies. The last section contains the discussion and conclusion section.

## 2. Material and Method

### 2.1. Dataset

The Kaggle melanoma skin cancer dataset was used in the study. The dataset contains two different classes: benign and malignant. The dataset includes 9600 training images and 1000 test images. Each image in the data set was prepared with a size of 300x300 pixels. The purpose of creating the data set is to help deep learning methods to be developed for the detection of melanoma skin cancer [8]. An example image of the data set is shown in Figure 1 [9].

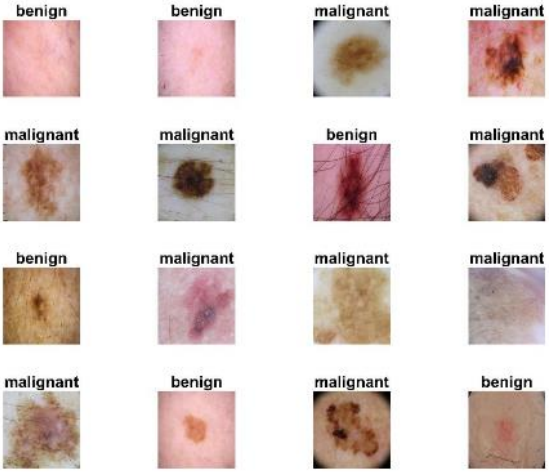

**Figure 1.** Sample image of the dataset.

### 2.2. Vision Transformer Model

In this study, the Vision Transformer (ViT) architecture was used for melanoma classification.

Vision Transformers (ViTs) have been produced by Google as an alternative to convolutional neural networks (CNN) in various computer vision processes. After first attracting attention with its superior performance in natural language processing, ViTs have also achieved success in image studies [10].

ViT has demonstrated state-of-the-art results in image classification, object detection, and segmentation tasks, often outperforming traditional CNNs.

ViT transforms an input image into a fixed-size, non-overlapping array of patches, each linearly embedded in a token. These tokens are then processed by a standard Transformer encoder, which uses a self-attention mechanism to capture global dependencies and contextual information across the entire image. This mechanism allows ViT to understand spatial relationships within the image more effectively than the local receptive fields typically used in CNNs. Positional coding is added to patch placements to preserve spatial information, allowing the model to distinguish the position of each patch within the original image [11].

ViT is pre-trained on extensive datasets, especially ImageNet, and then fine-tuned for specific tasks. It shows superior performance on various large-scale image classification criteria. The model's scalability and ability to capture long-range dependencies make it highly effective for a variety of computer vision applications, including object detection and image segmentation. Leveraging the global context understanding provided by the self-attention mechanism, ViT demonstrates the potential to revolutionize the field of computer vision by often exceeding the performance of traditional CNN architectures. This innovative use of Transformer architecture in visual tasks promises to be improved, highlighting a significant shift in the design of image classification models [12].

The working architecture of ViT is shown in figure 2 [13].

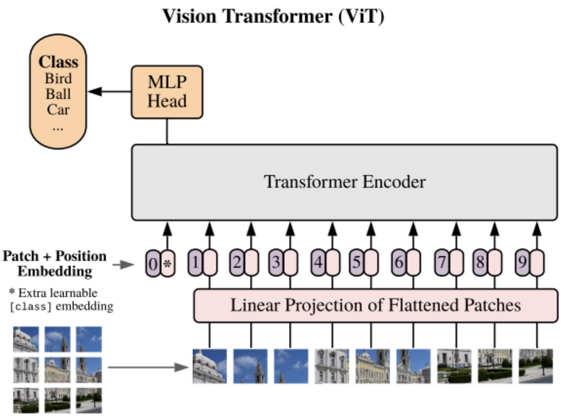
Vision Transformer (ViT)
Class
Bird
Ball
Car
...
MLP Head
Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0
*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches

**Figure 2.** ViT Architecture.

### 2.3. Transfer Learning Approach

Large data sets and powerful computers are often needed for a complete model training process. Also, the training process takes a long time. Determining the architecture of the model to be created and changing its hyperparameters are quite difficult operations, as well as requiring large amounts of data and powerful computing power. The transfer learning approach is therefore widely applied in the literature. The process of adapting a pre-trained model to a specific task or data collection is known as transfer learning. New data is trained only in the final stage, using the weights of a previously trained model. Thus, the pre-trained model is prepared for a new task. There are many advantages to using a pre-trained model. It allows you to use the latest models without having to start from scratch and reduces transaction costs [14] [15].

## 3. Model

In the study, classification was performed on the Melonoma image dataset. In the first stage, it was aimed to increase the training success by applying pre-processing on the images. In the second stage, pre-trained weights were given to the pre-processed data set during the transfer learning phase with the ViT model. Transfer learning of the ViT model was prepared by pre-training on the ImageNet dataset. The ImageNet dataset contains more than 14 million images and 21,000 classes. A wide variety of learning is provided as a result of training on such a large data set. After applying many preprocessing methods for skin cancer classification, transfer learning was used on images with fine tuning [16]. The classification process was carried out in 3 stages by coding the gogle/vit-large-patch32-384 method in the ViT model with the Python programming language in the Colab environment. This model is part of the ViT method produced by Dosovitskiy et al. This particular variant of the model uses a basic configuration with a patch size of 32x32 and an input resolution of 384x384 pixels [13].

The model architecture is shown in figure 3.

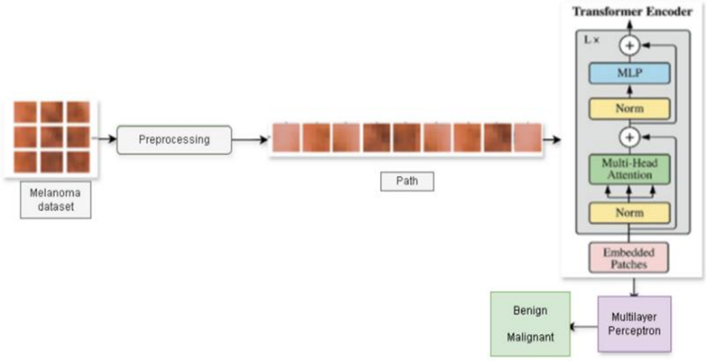

**Figure 3.** Model Architecture.

## 4. Experimental Tests

### 4.1. Performance Metrics

There are two different classes in our data set: malignant and benign. Confusion matrix was used to measure classification performances. The values obtained from this matrix are True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN).

The performance criteria used in the study are Accuracy (Acc), F-Score, Recall (Rec) and Precision (Pre). The formulas of these performance metrics are shown in equations 1-4 [17] [18].

$$Rec = \frac{TP}{TP+FN} \tag{1}$$

$$F1 = 2 * \frac{Pre*Rec}{Pre+Rec} \quad (2)$$

$$Acc = \frac{TP+TN}{TP+TN+FN+FP} \quad (3)$$

$$Pre = \frac{TP}{TP+FP} \quad (4)$$

### 4.2. Preprocessing

Pre-processing was applied to the data set before the training process. In the first stage, all images were resized to 384x384 and brought to the appropriate input size for the ViT model used in the study. In the second stage, Variance Stabilization, which is frequently used in classifying medical images, was applied to the data set. Variance Stabilization is the process of applying transformations to fix the variance of the data before analysis. This preprocessing method increases the performance of statistical analyzes and models by making the data more homogeneous [19]. Finally, the data set was normalized and made suitable for training. Normalization process is one of the basic methods used to increase model performance in deep learning.

### 4.3. Training

In the study, the data set was divided into 9600 and 1000 images as training and test data sets, respectively. Hyperparameters were tuned by trial and error. The hyperparameters used in coding were determined as batch_size=32, epoch=10, lr=1e-5, optimizer=man, loss function=CrossEntropyLoss, GELU activation function. The gogle/vit-large-patch32-384 model was used in the pre-training (Transfer Learning) phase of the training. In this model, patch sizes are set to 32x32. There is a 24-layer transformer encoder in the encoder part of the model. Transformer's attention mechanism, ViTSdpaSelfAttention consists of 1024 inputs and 1024 outputs, ViTSelfOutput consists of 1024 inputs and 1024 outputs, ViTIntermediate consists of 1024 inputs and 4096 outputs, and the ViTOutput layer consists of 4096 inputs and 1024 outputs. The layer that comes after the Encoder is the LayerNorm layer. Finally, there is a linear layer with 2 class outputs from 1024 features. In the study carried out with Python code in the Colab environment, training times were significantly reduced by using "CUDA" and "GPU".

### 5. Results and Discussions

In this study, automatic classification of melanoma skin cancer was performed using the Google Vision Transformer (ViT) model. In the first stage, all images are set to 384x384 (Resize). Secondly, Variance Stabilization, which is frequently used for medical images, was applied. Then, experimental tests were carried out by applying the ViT model and transfer learning to the data set.

As a result of the experimental tests performed on the test data of the study, an accuracy rate of 93.5% was achieved. As a result of the experimental tests of the classification process, the Precision, Recall and F1 Score values were calculated as 95.2%, 92.1% and 93.5%, respectively. We can say that these results are quite successful for image classification studies. The complexity matrix is given in figure 4 to see the success of the experimental test results in detail.

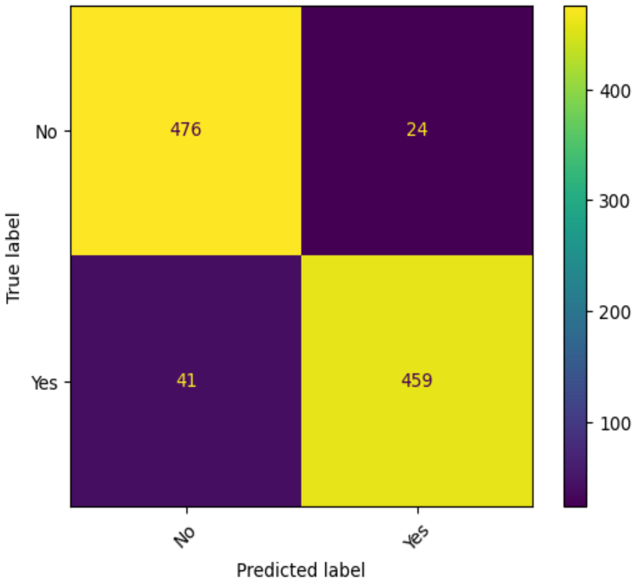

**Figure 4.** Confusion Matrix.

It is known in the literature that deep learning methods are widely used in skin cancer diagnosis and provide successful results. In particular, the study by Kaur et al. achieved very good levels of accuracy using deep CNNs. These results suggest that the ViT model can outperform traditional CNN-based models in diagnosing skin cancer.

A comparison table with popular studies in the literature on automatic classification of melanoma skin cancer is given below.

**Table 1.** Comparative Analysis for Automated Classification of Melanoma Skin Cancer.

| Paper | Year | Model | Acc |
|---|---|---|---|
| [20] | 2022 | Transfer learning with AlexNet | %87.1 |
| [21] | 2022 | VGG16, ResNet50, and Xception | %86.5 in VGG16, %81.6 in Resnet50 and 90.9 in Xception |
| [5] | 2022 | DCNN | %90,42 |
| [7] | 2023 | Transfer learning with ViT and CNN | %92.14 with ViT, %82 in ResNet50 |
| [6] | 2023 | Vision Transformer Model | %91 |
| [22] | 2024 | Transfer learning with ViT, Swin transformers and CNN | %88.6 with ViT and %88.8 with ResNet50 |
| This study | 2024 | Preprocessing, Transfer learning with ViT and Fine Tuning | %93.5 |

Table 1 shows that the study gives better results than popular studies in the literature.

Instead of operating on pixels like CNN methods, ViT transforms the input image into a fixed-size, non-overlapping, linear array of patches embedded in a token. It is then processed by a standard Transformer encoder, which uses a self-attention mechanism to capture global dependencies and contextual information across the entire image. Positional coding is added to patch placements to preserve spatial information, allowing the model to distinguish the position of each patch within the original image. This mechanism allows ViT to understand spatial relationships within the image more effectively than the local receptive fields typically used in CNNs [11]. Additionally, ViT is pre-trained on extensive datasets, especially ImageNet, and then fine-tuned and used for specific tasks. These mentioned structures make the ViT model more successful than other methods in the literature in image classification applications.

## 6. Conclusions

The Kaggle melanoma skin cancer dataset used in our study consists of benign and malignant samples and consists of 9600 training and 1000 test images. The breadth and diversity of the data set increased the generalization ability of the model. In addition, the applied pre-processing played an important role in achieving high accuracy rates. As a result of experimental tests performed on test data, it has proven that it can be used for the diagnosis of Melenoma skin cancer by reaching an accuracy rate of 93.5%.

In future studies, it is recommended to expand the data set and compare it with different deep learning models to further improve the performance of the model. Additionally, further research is needed to test the model in real-world applications and make it suitable for use in clinical settings. In conclusion, this study demonstrates that the ViT model is a promising tool in skin cancer diagnosis and lays a solid foundation for future research.

## References

[1] R. Deepa, G. ALMahadin, and A. Sivasamy, “Early detection of skin cancer using AI: Deciphering dermatology images for melanoma detection,” *AIP Adv.*, vol. 14, no. 4, 2024.

[2] I. H. Sarker, “Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions,” *SN Comput. Sci.*, vol. 2, no. 6, p. 420, 2021.

[3] K. Al-Hammuri, F. Gebali, A. Kanan, and I. T. Chelvan, “Vision transformer architecture and applications in digital health: a tutorial and survey,” *Vis. Comput. Ind. Biomed. Art*, vol. 6, no. 1, p. 14, 2023.

[4] A. Sriwastawa and J. A. Arul Jothi, “Vision transformer and its variants for image classification in digital breast cancer histopathology: A comparative study,” *Multimed. Tools Appl.*, vol. 83, no. 13, pp. 39731–39753, 2024.

[5] R. Kaur, H. GholamHosseini, R. Sinha, and M. Lindén, “Melanoma classification using a novel deep convolutional neural network with dermoscopic images,” *Sensors*, vol. 22, no. 3, p. 1134, 2022.

[6] P. Shobhit and N. Kumar, “Vision Transformer and Attention-Based Melanoma Disease Classification,” in *2023 4th International Conference on Communication, Computing and Industry 6.0 (C216), IEEE*, 2023, pp. 1–6.

[7] M. A. Arshed, S. Mumtaz, M. Ibrahim, S. Ahmed, M. Tahir, and M. Shafi, “Multi-class skin cancer classification using vision transformer networks and convolutional neural network-based pre-trained models,” *Information*, vol. 14, no. 7, p. 415, 2023.

[8] S. Ghosh, S. Dhar, R. Yoddha, S. Kumar, A. K. Thakur, and N. D. Jana, “Melanoma Skin Cancer Detection Using Ensemble of Machine Learning Models Considering Deep Feature Embeddings,” *Procedia Comput. Sci.*, vol. 235, pp. 3007–3015, 2024.

[9] S. R. Waheed et al., “Melanoma skin cancer classification based on CNN deep learning algorithms,” *Malaysian J. Fundam. Appl. Sci.*, vol. 19, no. 3, pp. 299–305, 2023.

[10] Z. Chen et al., “Vision transformer adapter for dense predictions,” arXiv Prepr. arXiv2205.08534, 2022.

[11] A. Parvaiz, M. A. Khalid, R. Zafar, H. Ameer, M. Ali, and M. M. Fraz, “Vision transformers in medical computer vision—A contemplative retrospection,” *Eng. Appl. Artif. Intell.*, vol. 122, p. 106126, 2023.

[12] X. Su et al., “Vitas: Vision transformer architecture search,” *in European Conference on Computer Vision, Springer*, 2022, pp. 139–157.

[13] A. Dosovitskiy et al., “An image is worth 16x16 words: Transformers for image recognition at scale,” arXiv Prepr. arXiv2010.11929, 2020.

[14] G. Mesnil et al., “Unsupervised and transfer learning challenge: a deep learning approach,” in *Proceedings of ICML Workshop on Unsupervised and Transfer Learning, JMLR Workshop and Conference Proceedings*, 2012, pp. 97–110.

[15] S. Ghosal and K. Sarkar, “Rice Leaf Diseases Classification Using CNN With Transfer Learning,” in *2020 IEEE Calcutta Conference (CALCON), IEEE*, 2020, pp. 230–236.

[16] A. Rahmouni, M. A. Sabri, A. Ennaji, and A. Aarab, “Skin Lesion Classification Based on Vision Transformer (ViT),” in *The International Conference on Artificial Intelligence and Smart Environment, Springer*, 2023, pp. 472–477.

[17] A. T. Karadeniz, Y. Çelik, and E. Başaran, “Classification of walnut varieties obtained from walnut leaf images by the recommended residual block based CNN model,” *Eur. Food Res. Technol.*, pp. 1–12, 2022.

[18] E. Başaran, Z. Cömert, and Y. Celik, “Timpanik Membran Görüntü Özellikleri Kullanılarak Sınıflandırılması,” *Fırat Üniversitesi Mühendislik Bilim. Derg.*, vol. 33, no. 2, pp. 441–453, 2021.

[19] S. M. Lin, P. Du, W. Huber, and W. A. Kibbe, “Model-based variance-stabilizing transformation for Illumina microarray data,” *Nucleic Acids Res.*, vol. 36, no. 2, pp. e11–e11, 2008.

[20] T. M. Ghazal, S. Hussain, M. F. Khan, M. A. Khan, R. A. T. Said, and M. Ahmad, “Detection of benign and malignant tumors in skin empowered with transfer learning,” *Comput. Intell. Neurosci*., vol. 2022, no. 1, p. 4826892, 2022.

[21] A. Bassel, A. B. Abdulkareem, Z. A. A. Alyasseri, N. S. Sani, and H. J. Mohammed, “Automatic malignant and benign skin cancer classification using a hybrid deep learning approach,” *Diagnostics,* vol. 12, no. 10, p. 2472, 2022.

[22] G. H. Dagnaw, M. El Mouhtadi, and M. Mustapha, “Skin cancer classification using vision transformers and explainable artificial intelligence,” J. *Med. Artif. Intell*., vol. 7, 2024.