European Journal of Technique

journal homepage: https://dergipark.org.tr/en/pub/ejt

Vol.15, No.2, 2025

Research Article

Comparative Analysis of Transfer Learning and Vision Transformer Models for Skin Cancer Classification Using Enhanced Dermoscopic Images

Yasin OZKAN[1*]

[1*]Department of Computer Technologies, Zonguldak Bulent Ecevit University, Zonguldak, Turkey (e-mail: yasin.ozkan@beun.edu.tr).

ARTICLE INFO

Received: May., 28. 2025
Revised: Oct., 31. 2025
Accepted: Dec., 23. 2025

Keywords:
Classification
Deep learning
Medical imaging
Skin cancer
Transfer learning
Vision transformer

Corresponding author: *Yasin OZKAN*

ISSN: *2536-5010* | e-ISSN: *2536-5134*

DOI: https://doi.org/10.36222/ejt.1708219

ABSTRACT

In recent years, deep learning has achieved remarkable advancements in medical image analysis, particularly through Convolutional Neural Networks (CNNs) and Transformer-based architectures. This study aims to evaluate and compare the performance of five transfer learning models (DenseNet169, InceptionV3, MobileNetV2, VGG16 and Xception) and a Vision Transformer (ViT) model for the classification of skin cancer using the "Skin Cancer: Malignant vs. Benign" dataset .In the first phase, the ViT model achieved the highest overall performance with 93.79% recall, 92.22% precision, 93.00% F1-score and 92.42% accuracy. Although InceptionV3 and MobileNetV2 demonstrated strong recall values, they did not match the overall accuracy of ViT. In the second phase, image enhancement techniques—grayscale conversion, thresholding, Canny edge detection, dilation, and erosion were applied to emphasize lesion boundaries and improve contrast. Using these enhanced images, the ViT model again achieved the best performance, with 95.49% recall, 94.17% precision, 94.83% F1-score, and 94.39% accuracy. These results indicate that the ViT architecture provides superior accuracy and reliability in complex and enhanced medical images. Furthermore, the study demonstrates that incorporating image preprocessing techniques can significantly enhance the performance of deep learning models in medical imaging applications.

1. INTRODUCTION

Ultraviolet (UV) rays are the type of electromagnetic radiation outside of visible sunlight that has significant effects on human health. UV rays are classified into three different wavelength ranges: UVA, UVB and UVC. UVA can cause deep damage to the skin, leading to premature aging and DNA damage, while UVB rays directly cause mutations in DNA and contribute to the development of skin cancer [1]. UVC rays do not reach the Earth's surface because they are largely absorbed by the ozone layer in the atmosphere. Excessive exposure to UV rays is the most important environmental factor that increases the risk of skin cancer. Skin cancer mainly occurs in three main types: basal cell carcinoma (BCC), melanoma and squamous cell carcinoma (SCC) [2]. While BCC and SCC generally have lower mortality rates, melanoma is more aggressive and can be fatal if not diagnosed early. The fact that UVB radiation, in particular, causes direct DNA damage in skin cells and that this damage accumulates and leads to carcinogenesis plays a key role in understanding the pathogenesis of skin cancer [3]. Genetic predisposition, skin type, excessive exposure to sunlight, use of solarium and some environmental factors are among the prominent risk factors for skin cancer. Skin cancers diagnosed early can usually be successfully treated with surgical intervention and other treatment modalities, but cases diagnosed late can adversely affect the treatment process and prognosis [4].

Early detection of skin cancer is a vital factor that directly determines the chances of a cure and the impact on the patient's future health. In recent years, deep learning and machine learning methods have emerged as technologies with the capacity to transform this process. Machine learning can be used as a powerful tool in the early detection of diseases such as skin cancer, especially by analyzing large data sets.

Machine learning is used as a powerful tool for early detection of diseases such as skin cancer, especially by analyzing large data sets [5]. Deep learning, on the other hand, offers very successful results, especially in the analysis of dermatological images. By automatically classifying skin lesions, deep learning networks can detect subtle changes that expert dermatologists may miss. In this way, patients can be accurately diagnosed at earlier stages [6].

In this study, existing deep learning-based approaches in the literature are considered to be effective in terms of classification accuracy and speed, but some challenges still remain in terms of accurate diagnosis and early stage detection. If skin cancer, especially aggressive types such as melanoma, is treated with early diagnosis, treatment success increases and patient survival rate improves. Deep learning-based algorithms enable fast and accurate evaluation of dermatologic images. This means lower error rates and shorter response times in the diagnostic process [7]. Furthermore, these technologies allow doctors to work as decision support systems, thus contributing to fast, accurate and reliable results in the diagnosis of skin cancer [8].

Although techniques such as transfer learning and convolutional neural networks (CNNs) used in existing studies provide accurate classification, some classification errors can lead to confusion, especially between benign and malignant lesions with similar shapes. To address this problem, deep learning techniques and transfer learning models need to be further optimized. Deep learning techniques are of great importance in skin cancer classification and diagnosis. Accurate classification of skin cancer lesions enables faster and more accurate results by utilizing the power of deep learning algorithms. In this context, deep learning-based approaches for skin cancer classification have been reviewed in detail in the literature.

In [9], a system for automatic classification of skin cancer and benign tumor lesions was developed. The study aims to reduce the time loss in the diagnosis process due to the similar shapes between skin cancer and benign lesions. The proposed model consists of three hidden layers, each with 16, 32 and 64 output channels respectively. Various optimization algorithms such as SGD, RMSprop, Adam and Nadam were used in the model and the best performance with a learning rate of 0.001 was obtained with Adam optimization. Adam optimization achieved 99% accuracy by classifying skin lesions into four classes from the ISIC dataset. In [10], a model was developed to classify skin cancer types. The model used image processing, deep learning and data augmentation techniques to classify 9 different types of skin cancers. The accuracy rate of the CNN model was obtained as 79.45%. In [11], a DCNN-based method was developed to detect skin lesions. The method uses various techniques for contrast enhancement, lesion boundary extraction and deep feature extraction. An accuracy of 98.4% and 94.8% was achieved in PH2 and ISIC 2017 datasets, respectively. In [12], the use of image classification algorithms to identify skin cancer types was investigated. There are different stages of skin cancer and survival rates vary at each stage (first stage 99%, fifth stage 20%). In this study, CNN is used to identify the different shapes and textures of skin cancer lesions. The proposed algorithm is applied on a dataset of 10,000 images of seven different types of lesions. In [13], a system for skin cancer detection and classification is proposed. In this study, skin cancer lesions are classified using the MNIST HAM-10,000 dataset. The proposed system detects and classifies skin cancer into different classes using CNN. Image processing and deep learning techniques are used to remove noise and improve the resolution of skin cancer dermoscopy images. The number of images is increased by various image augmentation techniques. In addition, the classification accuracy was further improved by transfer learning (using the ResNet model). The weighted average accuracy of the CNN model was 88%, recall was 74% and F1-score was 77%. The transfer learning approach provided 90.51% accuracy. In [14], a deep learning-based model was developed for the diagnosis of skin cancer types. Data on four skin cancer types were collected and the dataset was increased with image augmentation techniques. The CNN-based model achieved 95.98% accuracy on test data and outperformed models such as GoogleNet and MobileNet by 1.76% and 1.12%, respectively. In [15], a new CNN model called TurkerNet is proposed for skin cancer detection. The model aims to improve the classification performance by minimizing the training parameters. TurkerNet was tested with benign and malignant skin cancer images and achieved 92.12% accuracy. In [16], two hybrid CNN models are proposed to classify dermoscopy images into benign or melanoma lesions. These models combine features extracted from the first and second CNN and feed them to an SVM classifier. In the tests on the ISBI 2016 dataset, the proposed models outperformed the existing CNN models by achieving 88.02% and 87.43% accuracy, respectively. In [17], SVM, ResNet50 and MobileNet models were compared for skin cancer diagnosis using HAM10000 dataset. SVM was implemented with Histogram of Oriented Gradient (HOG) features and PCA, and SMOTE was used to stabilize the dataset. The results showed that SVM performed the best with 99.15% accuracy. In [18], the combination of human and CNN architectures in skin cancer classification was investigated. Using 11,444 dermoscopic images, the independent classifications of 112 dermatologists and CNNs were combined using gradient boosting. The results showed that the combination of human and machine achieved 82.95% accuracy. This is 1.36% higher than the CNN's 81.59% accuracy. In [19], a CNN-based model was proposed to improve accuracy in skin cancer diagnosis. The model developed using the HAM10000 dataset classifies skin lesions with convolutional, pooling and dense layers. To overcome data imbalance, a data augmentation strategy is applied and the model is trained with Adam optimization. The model achieved 97.78% accuracy, 97.9% precision, 97.9% recall and 97.8% F2 score. In [20], a model was developed for the classification of skin cancer types. The model trained for seven classes on the HAM10000 dataset was compared with five pre-trained CNNs and four ansamble models. The results showed an accuracy of 93.20% for independent models and 92.83% for ansamble models. In [21], a CNN model was developed for skin cancer detection. This model was built in Python using Keras and TensorFlow libraries. Trained with different network architectures and layer structures, the model achieved early convergence by utilizing Transfer Learning techniques. The model was tested on the ISIC dataset and achieved high accuracy rates in classifying skin cancer types. In [22], two methods are proposed for skin cancer detection: one is using a three-layer CNN and the other is a Support Vector Machine (SVM) model with the default RBF kernel. The features extracted by image processing techniques were used to classify the image as Benign or Malignant. The SVM classifier achieved 79.39% accuracy and 0.81 AUC, while the CNN model achieved 84.39% accuracy after 100 epochs. The CNN

model was presented as a web application using Streamlit. In [23], a CNN for skin cancer detection is proposed. As training data, 97 samples (50 benign and 47 malignant) from ISIC were used. To overcome the lack of data, synthetic skin cancer images were generated with Generative Adversarial Network (GAN). While the CNN model trained without synthetic images provided 53% accuracy, the accuracy of the model increased to 71% when augmented with these images. In [24], a DCNN model was developed to accurately classify skin cancer lesions. The model improved accuracy by using preprocessing and data augmentation techniques. Compared to transfer learning models such as VGG-16, AlexNet, DenseNet, MobileNet and ResNet on HAM10000 dataset, the proposed model achieved more reliable results with 93.16% training accuracy and 91.93% testing accuracy. In [25], a CNN model trained on the HAM10000 dataset is proposed. The model classifies skin lesions as cancerous or non-cancerous, allowing doctors and laboratory technicians to quickly learn three high-probability diagnoses.

The literature review demonstrates that deep learning and image processing techniques have substantially advanced the diagnosis, detection, and classification of skin cancer. These methods enable early and accurate identification of malignant lesions, thereby improving prognosis and treatment planning. The superior accuracy of deep learning algorithms, combined with the enhanced feature extraction and segmentation capabilities of image processing, has significantly strengthened automated diagnostic systems. However, despite their promising performance, current studies emphasize the need for further research on model optimization, robustness, and image enhancement strategies to address the inherent variability in dermoscopic images.

In this study, a two-phase experimental framework was designed to evaluate and enhance the performance of deep learning-based skin cancer classification. In the first phase, five transfer learning architectures (DenseNet169, InceptionV3, MobileNetV2, VGG16, and Xception) and a Vision Transformer (ViT) model were trained and compared on the original skin cancer dataset. In the second phase, the dataset was preprocessed using various edge detection filters to enhance lesion boundaries, and the same models were re-evaluated on these improved images. The results revealed that both the enhanced images and the ViT architecture achieved the highest classification accuracy, outperforming conventional CNN-based models. Unlike previous studies, this work not only highlights the efficiency of ViT in medical image analysis but also demonstrates the significant contribution of image preprocessing to overall model performance. Nevertheless, the findings suggest that additional efforts are required to optimize the computational efficiency and memory utilization of ViT-based systems for real-time clinical deployment. The main contributions of this study can be summarized as follows:

(1) a comprehensive comparative evaluation of multiple transfer learning architectures and the Vision Transformer model on both raw and edge-enhanced dermoscopic images

(2) the introduction of an image preprocessing pipeline combining edge detection and enhancement techniques, which significantly improved classification performance

(3) empirical verification of ViT's superiority over CNN-based models in skin lesion classification

The study is structured as follows: in Section 2, the dataset, the development of the hybrid classification framework, the proposed image processing approach, the transfer learning architectures, and the Vision Transformer implementation are described in detail; in Section 3, the results and discussion present model performances on original and enhanced images and compare them with existing studies; and finally, in Section 4, the conclusion summarizes the main findings and outlines future directions for efficient, real-time clinical applications.

2.MATERIALS AND METHODS

Skin cancer is a treatable disease with early detection and accurate diagnosis can greatly improve the treatment process. While traditional diagnostic methods are based on visual examinations by dermatologists and assessments with tools such as dermoscopy, these methods can be time-consuming and subjective [5]. Therefore, artificial intelligence and deep learning techniques play an important role in the early detection of skin cancer. CNNs, in particular, are widely used in the automatic analysis of medical images, but large data sets are needed for high accuracy [26]. At this point, transfer learning techniques provide better results with limited labeled data and benefit from the previous experience of the model [27].

With these developments, artificial intelligence techniques used in skin cancer diagnosis have become more diversified. In particular, the recent successes of the Vision Transformer (ViT) model in the field of visual classification have attracted attention. ViT divides images into fixed-size chunks and processes these chunks using a transformer architecture, thus learning global dependencies more efficiently [28]. This feature is especially useful for accurately classifying small and complex lesions in the diagnosis of diseases that require visual inspection, such as skin cancer [29]. Another important advantage of ViT is that it considers global relationships over the whole image rather than local features. This approach enables the model to perform a more precise classification, resulting in high accuracy in early detection of skin cancer.

In this study, an innovative classification method is developed by combining image processing approaches, transfer learning techniques and ViT methods to classify skin cancers. The study explains in detail how these methods are applied respectively.

2.1.Dataset

The dataset used in this study is a balanced image dataset compiled to distinguish between malignant and benign skin lesions [30]. The balanced nature of the dataset allows the model to learn both classes accurately. The number of images used for classification in the transfer learning models and ViT methods used in the study are presented in detail in Table 1.

TABLE I
NUMBER OF IMAGES BELONGING TO THE CLASSES IN THE DATASET

Classes in the data set	Test(%20)	Train (%80)	Total
Benign	360	1440	1800
Malign	300	1197	1497
Total	660	2637	3297

Furthermore, Figure 1 shows sample images of the benign and malignant classes in the dataset. Images in the benign class generally have smooth edges and homogeneous color distribution, while images in the malignant class have more

irregular edges and color variations. These images allow the model to learn the distinguishing features of both classes.

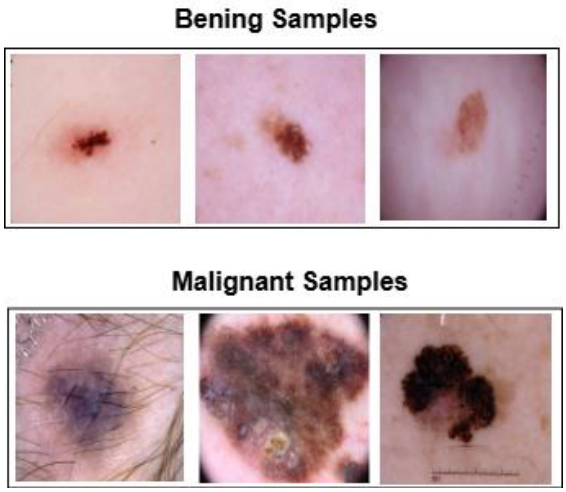

Figure 1. Examples of benign, malignant classes in the "Skin Cancer: Malignant vs. Benign" dataset.

A total of 3297 original images in the dataset are split into 80% training data and 20% test data in order to optimize the training process of the model. These ratios aim both to train the model efficiently during the learning phase and to provide enough test data to evaluate its accuracy.

2.2.Model Development

In the field of image classification, CNN is a fundamental technique for extracting visual features and improving the success of classification tasks based on these features. CNNs are particularly powerful in detecting edges, textures and patterns in images because they learn local features by performing convolution on each image segment [31]. This allows CNNs to generally operate with high accuracy. However, in areas where data sets are limited, such as healthcare, training with large datasets can be time-consuming and the accuracy of models trained with limited data can be degraded. In this context, Transfer Learning comes into play. Transfer learning accelerates the learning process and improves accuracy rates by reusing the knowledge of a model that has been previously trained on large data sets for a new task. This method has the potential to achieve high performance with limited data, especially in medical imaging applications such as skin cancer.

Transfer learning can improve the accuracy of the model in specialized and complex tasks such as skin cancer classification, as it enables the adaptation of features from larger datasets to current tasks. However, in recent years, new model architectures such as ViT have attracted attention by exhibiting superior performance in image classification tasks. Unlike the ability of conventional CNNs to extract local features, ViT processes images into small parts and processes each part with an attention mechanism, thus learning global contexts more effectively [28]. This feature allows the model to more efficiently acquire general information and make more accurate classifications, especially in large datasets. These advantages of ViT contribute to rapid accurate diagnoses in sensitive tasks such as medical imaging. In this context, both Transfer Learning and ViT models stand out as important tools to improve classification accuracy and achieve more efficient results. Figure 2 presents an overview of the metadology used in the study.

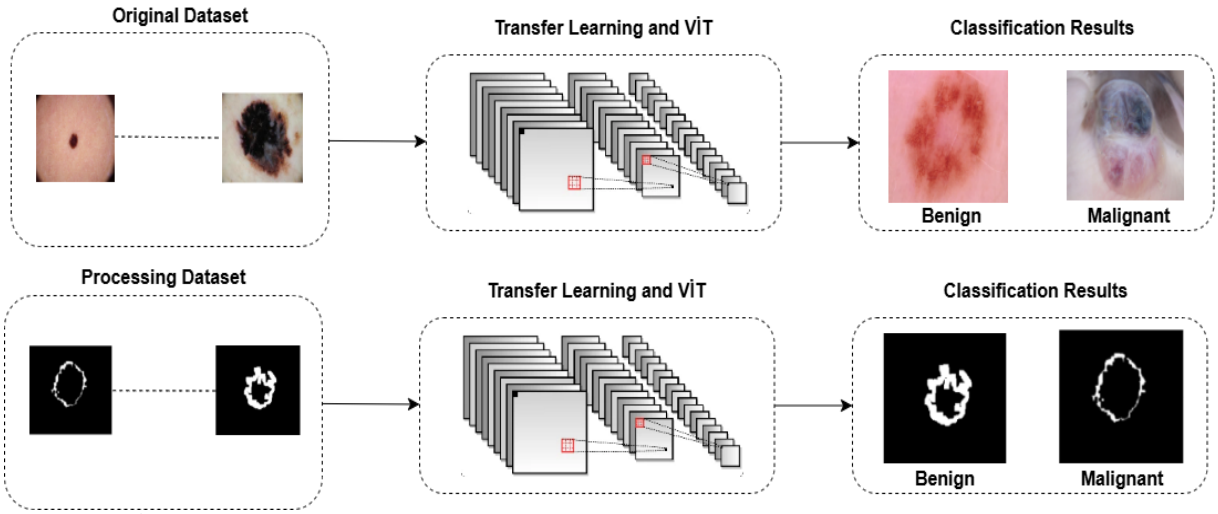

Figure 2. Overview of the metodology used in the study.

2.2.1.Image Processing

Image processing is a field that involves the digital analysis, processing and interpretation of digital images. Images can be defined as data, usually two-dimensional, obtained as a result of the reflection or transmission of light on a surface. While the visual perception of the human eye processes these light reflections through a biological mechanism, computers treat these images as digital data and analyze them with various algorithms. Image processing offers a wide range of applications not only in aesthetics but also in many disciplines such as engineering, medicine, biometrics, space exploration, industrial automation and security [32]. The basic processes in this field involve applying mathematical and algorithmic operations to the numerical representations of pixels in an image, analyzing the features in the image (e.g., edges, textures, colors), and transforming these features when necessary [33]. Image processing techniques are used to remove image distortions, optimize contrast and brightness levels, perform edge detection, object recognition and more complex operations. In addition, significant advances have been made in the field of image processing in recent years with the integration of artificial intelligence and deep learning methods. By learning from large data sets, these approaches improve the accuracy and efficiency of image analysis and recognition [31]. In this context, image processing has become not only a theoretical field, but also an increasingly preferred technology in practical applications.

2.2.1.1.Proposed Hybrid Image Processing Approach

Medical image processing has become one of the fundamental components of modern healthcare systems, contributing significantly to early diagnosis, treatment planning, and continuous disease monitoring. The accuracy and interpretability of diagnostic models largely depend on the quality and clarity of the input images. Therefore, preprocessing and enhancement techniques play a vital role in preparing medical data for deep learning-based analysis. In this study, a comprehensive image preprocessing pipeline was implemented to improve image clarity, enhance lesion boundaries, and ensure that critical features were effectively represented for subsequent model training.

The preprocessing workflow consisted of six main stages: color-to-grayscale conversion, inverse thresholding, Canny edge detection, morphological dilation, morphological erosion, and visualization of the intermediate results. Initially, each color image in the dataset was converted to grayscale to remove chromatic variations and emphasize intensity-based information. This transformation simplifies image representation by retaining only luminance components, enabling algorithms to focus on structural rather than color features. Grayscale conversion also reduces computational complexity and noise, thereby facilitating more efficient feature extraction in later stages.

Following grayscale conversion, an inverse thresholding operation was applied to segment the image into binary regions based on pixel intensity levels. Pixels below a defined threshold value were set to white (255), while those above were assigned as black (0). This binary separation effectively highlights regions of interest, such as skin lesions, and removes irrelevant background details. Thresholding is particularly valuable in medical imaging, where contrast enhancement can make pathological structures more discernible and measurable.

After segmentation, edge information was extracted using the Canny edge detection algorithm. This multi-stage process includes Gaussian filtering to suppress noise, computation of image gradients to identify areas with rapid intensity change, and application of dual thresholds to determine true edges. The Canny algorithm's precision in localizing boundaries makes it an ideal choice for identifying lesion margins in dermoscopic and histopathological images. Detecting these boundaries with high fidelity is crucial for downstream tasks such as segmentation, morphological analysis, and classification.

To further refine the detected edges, morphological dilation was performed using a 5×5 kernel. Dilation enlarges bright regions (white pixels) in the image, which helps to close small gaps and make the boundaries more continuous. This step improves the connectivity of edge structures and enhances the overall visibility of lesions. Subsequently, morphological erosion was applied as a complementary operation to dilation. Erosion reduces the size of bright regions, removing small artifacts and irregularities introduced during dilation. Together, these two morphological transformations help maintain sharp and well-defined contours while eliminating unnecessary noise and smoothing over-segmented areas.

After all preprocessing operations were completed, the results of each transformation stage were visualized to illustrate the cumulative effect of the pipeline. The comparative visualization demonstrated a clear progression in image enhancement — from raw color images to refined representations with distinct lesion boundaries. This structured enhancement process provides valuable insights into how successive operations contribute to improved visual and analytical interpretability.

Overall, the proposed preprocessing framework establishes a robust foundation for advanced image analysis techniques such as segmentation, feature extraction, and deep learning-based classification. By systematically emphasizing contrast and structural details, the approach enhances the discriminative capacity of learning models. Figure 3 presents the complete flowchart of the proposed hybrid preprocessing and analysis approach, summarizing the logical sequence and interdependencies among the applied operations.

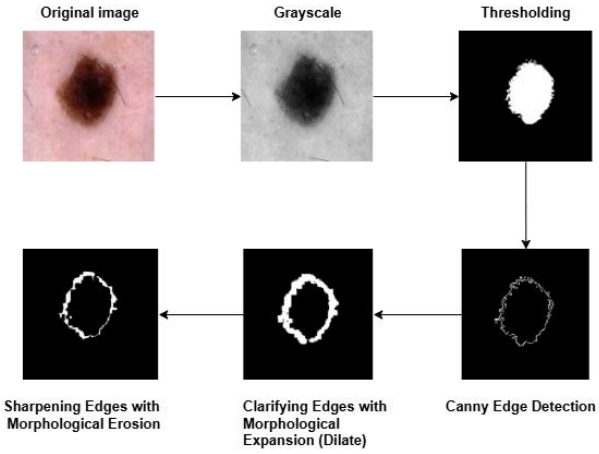

Figure 3. Flowchart of the proposed hybrid image processing approach

Different image processing techniques are used in turn in the hybrid image processing approach shown in Figure 2. Following clarification of the picture features, the images were registered in order to compare the performance of various ViT models and transfer learning models (DenseNet169, InceptionV3, MobileNetV2, VGG16, Xception).

2.3.Transfer Learning Architectures Used in the Study

In the field of deep learning, various architectures have been developed to meet the requirements of large datasets and high processing power. These architectures have achieved significant success, especially in tasks such as processing and classifying visual data. At this point, transfer learning techniques allow previously trained models to be reused for another task, reducing training times and improving performance. Various deep learning architectures have been used for transfer learning and each offers different advantages. In this paper, we will discuss in detail the features, applications and impact on transfer learning of widely used models such as DenseNet169, InceptionV3, MobileNetV2, VGG16 and Xception. These models have an important place in deep learning with their different building blocks and optimization strategies.

DenseNet169 is a remarkable and high-performing architecture in deep learning. This model is a part of the DenseNet family and makes the learning process more efficient by establishing dense connections between each layer. In particular, DenseNet169 aims to achieve high accuracy with fewer parameters by using depth-separable

convolutions and short connections. The main advantage of the model is that each layer directly receives all feature maps from previous layers, thus creating richer and deeper representations. This structure both prevents the gradient loss problem and makes the model work more efficiently [34]. DenseNet169 has achieved successful results especially in areas such as image classification, object recognition and medical image processing and is frequently preferred for transfer learning applications.

InceptionV3 is a model developed by Google and has an important place in the field of deep learning. This model, which is the evolution of the Inception architecture, offers a structure that allows the efficient extraction of multi-scale features. By combining structures with different layer depths and filter sizes, InceptionV3 optimizes computational cost while increasing the overall efficiency of the model. An important feature of the model is the combination of 1x1, 3x3 and 5x5 filters in the inception block, allowing for a wider range of information. In addition, the "auxiliary classifier" structure of the model enables faster learning in deep layers [35]. InceptionV3 is frequently used in transfer learning applications, exhibiting strong performances in tasks such as image recognition and classification.

MobileNetV2 is a deep learning model developed especially for mobile and low processing power devices. The most prominent feature of this model is that it computes more efficiently by using a structure called "inverted residuals". By adopting depth-separable convolutions, MobileNetV2 first applies each filter on a single channel and then combines these channels to achieve high accuracy with lower parameter counts [36]. Furthermore, the linear bottleneck structure used in the output layers of the model offers advantages in terms of speed and efficiency, especially on mobile devices. MobileNetV2 has become a popular choice for transfer learning applications due to its low computational power requirement and high efficiency, especially in tasks such as image recognition on mobile devices.

VGG16 is a model developed by the Visual Geometry Group (VGG) of the University of Oxford, which has an important place in the field of deep learning. The basic structure of this model consists of convolutional layers of increasing depth and consists of 16 layers. VGG16 is characterized by its simple structure and strong performance. An important advantage of the model is that it improves its generalization ability by training on large datasets. In particular, it allows for more efficient processing of images without increasing the number of parameters by using fixed filter sizes in each layer [37]. VGG16 is a frequently preferred model as it can achieve high accuracy rates and can be easily adapted for transfer learning.

Xception is a model that has attracted attention in the field of deep learning and is considered as a more advanced version of the Inception architecture. This model, which can be defined as "Extreme Inception", replaces traditional convolution operations with "depthwise separable convolutions". This structure reduces computational costs by applying each filter in a single channel, while increasing the parameter efficiency of the model. Xception performs particularly well in deep and complex network structures, achieving more accurate results with fewer parameters [38]. The model is highly effective for transfer learning applications, provides high accuracy on large datasets, and is currently used in many visual recognition applications.

2.4. Vision Transformer

In recent years, ViT architectures have emerged as a transformative paradigm in computer vision, representing a major shift from traditional CNN-based frameworks [28]. Unlike CNNs, which rely on local convolutional filters to extract spatially constrained features, ViT employs a global self-attention mechanism originally developed for Natural Language Processing (NLP) tasks. This innovation allows ViT to model long-range dependencies and global contextual relationships across the entire image, overcoming one of the key limitations of CNNs.

ViT divides each image $x \in R^{H \times W \times Cx}$ into a sequence of non-overlapping patches of size P×P. Each flattened patch x_p^i is linearly embedded into a vector representation using a learnable projection matrix EEE and its mathematical representation is presented in Equation (1).

$$z_0^i = x_p^i E \quad (1)$$

To preserve spatial information, positional encodings are added to these embeddings before being processed by the Transformer encoder. The self-attention mechanism, which forms the core of the Transformer, models the relationships between all image patches as in Equation (2).

$$Attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d_k}})V \quad (2)$$

where Q,K,V are the query, key, and value matrices, respectively, and d_k is the dimension of the key vectors.

To enhance representational richness, ViT uses Multi-Head Self-Attention (MSA), where multiple attention heads are computed and combined in parallel, and is calculated as in Equation (3).

$$MSA(X) = [head_1; head_2; \dots; head_h]W_0 \quad (3)$$

This multi-head design enables the model to capture diverse dependencies among image patches. By leveraging this architecture, ViT achieves a comprehensive understanding of the global structure of images, surpassing the locality limitations of CNNs. Its effectiveness becomes particularly evident when trained on large-scale datasets, where the self-attention mechanism allows ViT to outperform traditional convolutional models in both accuracy and generalization.

3. RESULT AND DİSCUSSİON

In recent years, deep learning techniques have transformed the field of medical imaging. CNN is still one of the most widely used methods for analyzing medical images, often with successful results on data types such as radiological images, pathology specimens. However, the emergence of Transformer-based models such as ViT is reshaping visual data processing paradigms. ViT provides a significant improvement, especially in the classification of small and complex images, thanks to the attentional mechanisms used to understand the broader context in visual data [28].

In contrast, traditional CNN-based architectures still provide distinct advantages in environments with limited computational resources. Although these models are not capable of learning the global context, they are highly

efficient in more localized and targeted learning processes. Therefore, when deciding which model to use in medical imaging and diagnostic applications, factors such as the computational cost of the model, the size of the dataset and the application requirements should be taken into account. In the future, making Transformer-based models more efficient could further expand their applications in the medical field. However, this process will require significant investments in terms of both data and computing resources. In this context, considering the advantages and limitations of each model, optimal solutions should be developed for the effective use of medical imaging technologies.

In this study, we compare the classification performance of ViT models with five different transfer learning architectures such as DenseNet169, InceptionV3, MobileNetV2, VGG16 and Xception in classifying skin cancer types. First, the performance metrics obtained on the original images are presented in Table 2.

TABLE II
PERFORMANCE METRICS OF VIT MODELS WITH 5 DIFFERENT TRANSFER LEARNING ARCHITECTURES ON ORIGINAL IMAGES.

Transfer Learning Models	TP	TN	FN	TN	Recall (%)	Precision (%)	F1-score (%)	Accuracy (%)
DenseNet169	282	78	32	268	89.81	78.33	83.68	83.33
Inceptionv3	292	68	25	275	92.11	81.11	86.26	85.91
MobileNetV2	285	75	25	275	91.94	79.17	85.07	84.85
Vgg16	268	82	37	263	87.87	76.57	81.83	81.69
Xception	259	91	30	270	89.62	74.00	81.06	81.38
ViT	332	28	22	278	93.79	92.22	93.00	92.42

The primary metrics used to assess a classification model's performance are True Negative (TN), True Positive (TP), False Negative (FN), and False Positive (FP) values. These values help us understand how accurate or inaccurate the model predicts. Precision measures the rate at which the model correctly predicts the positive class, while Recall refers to the rate at which the model finds true positive examples. F1 Score provides a balanced combination of precision and recall, assessing the model's ability to both predict correctly and correctly detect positive classes. Accuracy indicates the overall performance of the model, i.e. the rate at which the model makes correct predictions across all classifications. These metrics are used to measure the effectiveness of the model and to more accurately assess the performance of the model, especially in imbalanced data sets [39]. The mathematical expressions of the metrics are presented in Equations (4), (5), (6) and (7).

$$Accuracy = \frac{(TN + TP)}{(FP + TN + FN + TP)} \quad (4)$$

$$\text{Precision} = \frac{TP}{(FP + TP)} \quad (5)$$

$$\text{Recall} = \frac{TP}{(FN + TP)} \quad (6)$$

$$\text{F1} - \text{Score} = 2 \text{ x} \frac{\text{Recall x Precision}}{(\text{Recall} + \text{Precision})} \quad (7)$$

Table 2 shows the classification performance of ViT models with 5 different transfer learning architectures. The ViT model achieved the best performance in all metrics with Recall (93.79%), Precision (92.22%), F1-score (93.00%) and Accuracy (92.42%). This result shows that ViT can classify complex medical data with high accuracy thanks to its ability to learn in a global context.

However, Inceptionv3 and MobileNetV2 models stood out with high Recall values of 92.11% and 91.94% respectively. However, Inceptionv3 stands out with higher F1-score (86.26%) and Accuracy (85.91%), while MobileNetV2 offers an advantage with its usability in environments with low computational resources. The DenseNet169 and Xception models, on the other hand, despite their high Recall values, lagged behind in Precision and F1-score metrics, and thus, although they are more efficient, they underperformed in terms of resource utilization. These findings suggest that each model offers advantages in different usage scenarios.

In the second part of the study, the classification performances of the same transfer learning architectures and ViT models are compared on the dataset obtained from the enhanced images. Table 3 presents the performance metrics obtained from the enhanced images.

TABLE III
PERFORMANCE METRICS OF VIT MODELS WITH 5 DIFFERENT TRANSFER LEARNING ARCHITECTURES ARE COMPARED BASED ON THE RESULTS OBTAINED FROM THE ENHANCED IMAGES.

Transfer Learning Models	TP	TN	FN	TN	Recall (%)	Precision (%)	F1-score (%)	Accuracy (%)
DenseNet169	289	71	25	275	92.04	80.28	85.76	85.45
Inceptionv3	295	65	22	278	93.06	81.94	87.15	86.82
MobileNetV2	287	73	23	277	92.58	79.72	85.67	85.45
Vgg16	274	86	26	274	91.33	76.11	83.03	83.03
Xception	264	96	19	281	93.29	73.33	82.12	82.58
ViT	339	21	16	284	95.49	94.17	94.83	94.39

According to the data presented in Table 3, ViT shows a prominent performance in all metrics. In particular, the high recall (95.49%) and precision (94.17%) rates, F1-score (94.83%) and accuracy (94.39%) show that the ViT model is very successful in accurate classification and can effectively distinguish both classes (positive and negative). Figure 4 shows the training and loss graph for the most successful result (ViT) in the study.

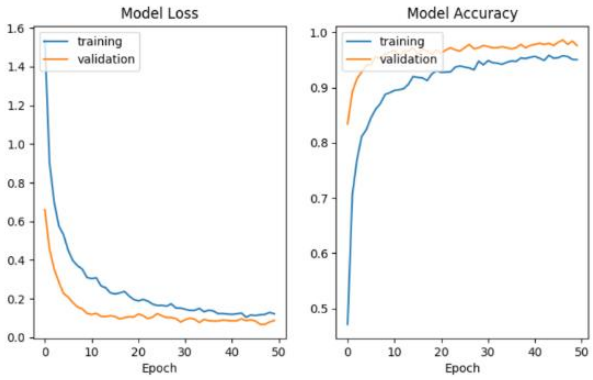

Figure 4. Loss and accuracy plots of the enhanced images in the ViT model.

The performance of the InceptionV3 model is close to ViT, but with a slight regression in recall and precision (recall 93.06%, precision 81.94%). This suggests that the model is not as effective as ViT in reducing false negatives, but its overall accuracy (accuracy 86.82%) is high. Other models, such as DenseNet169, MobileNetV2 and Xception, perform poorly compared to ViT. For example, DenseNet169's recall rate (92.04%) is lower than ViT, while it outperforms ViT in metrics such as F1-score and accuracy. Xception, on the other hand, has a high recall (93.29%) but a very low precision (73.33%), indicating that the model makes more false positive classifications. The VGG16 model, on the other hand, performed poorly compared to the other models, especially in terms of precision (76.11%) and F1-score (83.03%), indicating that the model over-identified false positives and thus made misclassifications. Overall, the ViT model achieved the highest success in the classification task with high recall, precision and F1-score values and significantly outperformed the other models. This shows that ViT is a strong candidate for classification in enhanced images and has the ability to cope with challenges such as class imbalance. Figure 5 shows the accuracy graph of the achievements of the models applied on the original and enhanced images.

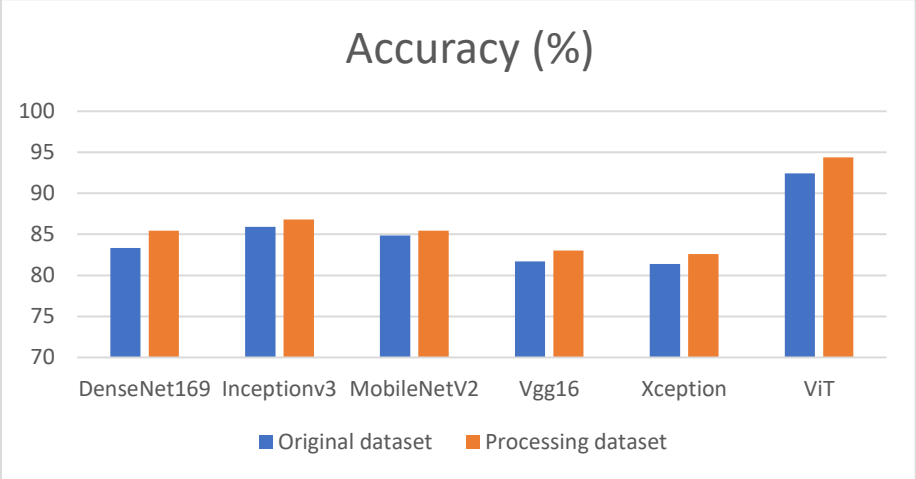

Figure 5. Accuracy results obtained from the original and enhanced images.

Analysis of the two experimental phases revealed that CNN-based architectures underperformed compared to ViT models. While CNNs effectively captured local spatial patterns, their limited receptive field restricted their ability to model global dependencies within images. In contrast, ViT leveraged its self-attention mechanism to integrate long-range contextual relationships, achieving more discriminative and holistic feature representations.

Additionally, the applied image processing techniques—such as grayscale conversion, thresholding, Canny edge detection, dilation, and erosion—significantly enhanced model performance. These operations improved image contrast and lesion boundary clarity, leading to notable gains in recall, precision, and overall accuracy. The results indicate that preprocessing steps play a crucial role in improving the effectiveness of deep learning models, particularly in complex medical image datasets.

Finally, to assess the competitiveness of the proposed method, its performance was compared with recent deep learning-based studies using skin cancer datasets. As presented in Table 4, the proposed ViT-based framework achieved superior accuracy and reliability, confirming its advantage over conventional CNN approaches and its potential as a robust tool for medical image classification.

TABLE IV
AN OVERVIEW OF THE LITERATURE ON THE CLASSIFICATION OF SKIN CANCER IMAGES.

Study	Data Set	Method	Accuracy (%)
[40]	HAM10000	EfficientNet B0-B7, Transfer Learning, Fine-Tuning	87.91
[41]	HAM10000	CNN and ViT	94,30
[42]	5846 clinical images were collected.	CNN	91.5

[15]	Skin cancer dataset	CNN	92.12
[43]	Skin cancer dataset	CNN and Transfer Learning	89.09
Proposed model	Skin cancer dataset	Image processing, Transfer Learning and ViT	94.39

Table 4 presents various literature studies on the classification of skin cancer images and the accuracy rates obtained with the methods used in each study. In the study by [40], EfficientNet B0-B7 models, transfer learning and fine-tuning techniques were used to classify the HAM10000 dataset. The accuracy rate obtained with this method was reported as 87.91%. [41], using the same dataset, preferred the combination of CNN and ViT and obtained an accuracy rate of 94.30% with this method. [42] used only CNN method in their study on 5846 clinical images and achieved an accuracy rate of 91.5%.

In [15], [43] and the proposed study, three different skin cancer classification studies were conducted on skin cancer dataset. In the study by [15], an accuracy rate of 92.12% was obtained using CNN only. [43] achieved 89.09% accuracy using a combination of CNN and transfer learning methods on the same dataset. In the proposed study, image processing, transfer learning and ViT techniques were applied separately on the Skin cancer dataset and a high accuracy rate of 94.39% was obtained.

Research in the field of skin cancer classification shows that deep learning methods are effective in improving accuracy rates. High success has been achieved with techniques such as CNN, transfer learning and ViT. In our proposed work, 94.39% accuracy rate was achieved by applying each of these methods separately on the Skin Cancer Dataset. This result reveals that the combination of multiple methods provides higher success, especially in complex classification tasks such as skin cancer. The proposed method makes a significant contribution to the field by providing one of the highest accuracy rates in the literature.

4.CONCLUSION

Skin cancer is a disease of great importance in the health field, requiring early diagnosis and accurate classification. In recent years, the increasing use of deep learning techniques in this field has led to high accuracy rates in skin cancer classification tasks. In this study, we compared the performance of the models in the classification of skin cancer types through a two-stage analysis on original images and enhanced images. In the first part, the performances of five different transfer learning models and ViT are compared. As a result of this comparison, the ViT model showed the highest success in the analysis performed on the original images, reaching an accuracy rate of 92.42%. In the second stage analysis on the enhanced images, it was evident that the image processing techniques improved the performance of the ViT model. At this stage, ViT achieved the highest performance with an accuracy of 94.39%. It was observed that image processing techniques, especially edge detection and other enhancement methods, significantly improve the accuracy of the model and are effective in dealing with challenges such as class imbalance. The performance of ViT was further strengthened in classifications with the enhanced images, with high results in metrics such as recall (95.49%), precision (94.17%) and F1-score (94.83%). The results show that the combination of image processing techniques and the ViT model provides high accuracy rates in skin cancer classification. The contribution of the ViT model to the field of medical imaging is especially evident in early diagnosis processes. The ability of this model to perform more accurate and faster classifications can make a great contribution to clinical decision support systems. As the diagnosis of early-stage skin cancer can directly impact treatment success, ViT's high accuracy rates in this area can enable patients to be treated earlier and more accurately. Therefore, optimizing future studies to make the ViT model more suitable for clinical applications could lead to significant improvements in medical imaging and early diagnosis.

REFERENCES

[1] B. K. Armstrong and A. Kricker, "The epidemiology of UV induced skin cancer," *J. Photochem. Photobiol. B: Biol.*, vol. 63, no. 1-3, pp. 8-18, 2001.

[2] International Agency for Research on Cancer, *Working Group on the Evaluation of Carcinogenic Risks to Humans. Human papillomaviruses*, vol. 90, 2011.

[3] L. Hogue and V. M. Harvey, "Basal cell carcinoma, squamous cell carcinoma, and cutaneous melanoma in skin of color patients," *Dermatol. Clin.*, vol. 37, no. 4, pp. 519-526, 2019.

[4] C. Reggiani et al., "Update on non-invasive imaging techniques in early diagnosis of non-melanoma skin cancer," *G Ital Dermatol Venereol*, vol. 150, no. 4, pp. 393-405, 2015.

[5] A. Esteva et al., "Dermatologist-level classification of skin cancer with deep neural networks," *Nature*, vol. 542, no. 7639, pp. 115-118, 2017.

[6] H. A. Haenssle et al., "Man against machine: diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists," *Ann. Oncol.*, vol. 29, no. 8, pp. 1836-1842, 2018.

[7] A. A. Adegun and S. Viriri, "Deep learning-based system for automatic melanoma detection," *IEEE Access*, vol. 8, pp. 7160-7172, 2019.

[8] R. T. Sutton et al., "An overview of clinical decision support systems: benefits, risks, and strategies for success," *NPJ Digit. Med.*, vol. 3, no. 1, p. 17, 2020.

[9] Y. N. Fu'adah, N. C. Pratiwi, M. A. Pramudito, and N. Ibrahim, "Convolutional neural network (CNN) for automatic skin cancer classification system," in *IOP Conf. Ser.: Mater. Sci. Eng.*, vol. 982, no. 1, p. 012005, Dec. 2020.

[10] N. Rezaoana, M. S. Hossain, and K. Andersson, "Detection and classification of skin cancer by using a parallel CNN model," in *Proc. IEEE WIECON-ECE*, pp. 380-386, Dec. 2020.

[11] T. Saba, M. A. Khan, A. Rehman, and S. L. Marie-Sainte, "Region extraction and classification of skin cancer: A heterogeneous framework of deep CNN features fusion and reduction," *J. Med. Syst.*, vol. 43, no. 9, p. 289, 2019.

[12] W. O'Keefe et al., "A CNN approach for skin cancer classification," in *Proc. Int. Conf. Inf. Technol. (ICIT)*, pp. 472-475, Jul. 2021.

[13] R. Garg, S. Maheshwari, and A. Shukla, "Decision support system for detection and classification of skin cancer using CNN," in *Proc. ICICV 2020*, pp. 578-586, Springer, 2021.

[14] M. S. Junayed, N. Anjum, A. Noman, and B. Islam, "A deep CNN model for skin cancer detection and classification," 2021.

[15] T. Tuncer et al., "A lightweight deep convolutional neural network model for skin cancer image classification," *Appl. Soft Comput.*, p. 111794, 2024.

[16] D. Keerthana et al., "Hybrid convolutional neural networks with SVM classifier for classification of skin cancer," *Biomed. Eng. Adv.*, vol. 5, p. 100069, 2023.

[17] L. I. Mampitiya, N. Rathnayake, and S. De Silva, "Efficient and low-cost skin cancer detection system implementation with a comparative study between traditional and CNN-based models," *J. Comput. Cogn. Eng.*, vol. 2, no. 3, pp. 226-235, 2023.

[18] A. Hekler et al., "Superior skin cancer classification by the combination of human and artificial intelligence," *Eur. J. Cancer*, vol. 120, pp. 114-121, 2019.
[19] M. M. Musthafa et al., "Enhanced skin cancer diagnosis using optimized CNN architecture and checkpoints for automated dermatological lesion classification," *BMC Med. Imaging*, vol. 24, no. 1, p. 201, 2024.
[20] S. S. Chaturvedi, J. V. Tembhurne, and T. Diwan, "A multi-class skin cancer classification using deep convolutional neural networks," *Multimed. Tools Appl.*, vol. 79, no. 39, pp. 28477-28498, 2020.
[21] H. Nahata and S. P. Singh, "Deep learning solutions for skin cancer detection and diagnosis," in *Machine Learning with Healthcare Perspective*, pp. 159-182, 2020.
[22] R. Tanna and T. Sharma, "Binary classification of melanoma skin cancer using SVM and CNN," in *Proc. Int. Conf. AIMV*, pp. 1-4, Sept. 2021.
[23] P. Sedigh, R. Sadeghian, and M. T. Masouleh, "Generating synthetic medical images by using GAN to improve CNN performance in skin cancer classification," in *Proc. ICRoM*, pp. 497-502, Nov. 2019.
[24] M. S. Ali et al., "An enhanced technique of skin cancer classification using deep convolutional neural network with transfer learning models," *Mach. Learn. Appl.*, vol. 5, p. 100036, 2021.
[25] S. Mohapatra et al., "Skin cancer classification using convolution neural networks," in *Proc. ICADCML 2020*, pp. 433-442, Springer, 2020.
[26] G. Litjens et al., "A survey on deep learning in medical image analysis," *Med. Image Anal.*, vol. 42, pp. 60-88, 2017.
[27] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. Knowl. Data Eng.*, vol. 22, no. 10, pp. 1345-1359, 2009.
[28] A. D. Alexey, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint*, arXiv:2010.11929, 2020.
[29] H. Touvron et al., "Training data-efficient image transformers & distillation through attention," in *Proc. ICML*, pp. 10347-10357, Jul. 2021.
[30] C. Fanconi, "Skin Cancer: Malignant vs. Benign, Processed Skin Cancer pictures of the ISIC Archive," [Online]. Available: https://www.kaggle.com/datasets/fanconic/skin-cancer-malignant-vs-benign. [Accessed: Dec. 15, 2024].
[31] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436-444, 2015.
[32] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 2nd ed., Prentice Hall, 2002.
[33] M. Sonka, V. Hlavac, and R. Boyle, *Image Processing, Analysis, and Machine Vision*, Springer, 2013.
[34] G. Huang et al., "Densely connected convolutional networks," in *Proc. CVPR*, pp. 4700-4708, 2017.
[35] C. Szegedy et al., "Rethinking the inception architecture for computer vision," in *Proc. CVPR*, pp. 2818-2826, 2016.
[36] M. Sandler et al., "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. CVPR*, pp. 4510-4520, 2018.
[37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint*, arXiv:1409.1556, 2014.
[38] F. Chollet, "Xception: Deep learning with depthwise separable convolutions," in *Proc. CVPR*, pp. 1251-1258, 2017.
[39] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Inf. Process. Manag.*, vol. 45, no. 4, pp. 427-437, 2009.
[40] K. Ali et al., "Multiclass skin cancer classification using EfficientNets–a first step towards preventing skin cancer," *Neurosci. Inform.*, vol. 2, no. 4, p. 100034, 2022.
[41] C. Xin et al., "An improved transformer network for skin cancer classification," *Comput. Biol. Med.*, vol. 149, p. 105939, 2022.
[42] S. Jinnai et al., "The development of a skin cancer classification system for pigmented skin lesions using deep learning," *Biomolecules*, vol. 10, no. 8, p. 1123, 2020.
[43] V. Anand et al., "An enhanced transfer learning based classification for diagnosis of skin cancer," *Diagnostics*, vol. 12, no. 7, p. 1628, 2022.

BIOGRAPHIES

Yasin OZKAN received his Ph.D. in Computer Engineering in 2024. His research interests focus on artificial intelligence, deep learning, and machine learning. He has worked on developing advanced models for image analysis and classification tasks, with a particular emphasis on healthcare applications. His academic background combines theoretical foundations with practical implementations, aiming to contribute to intelligent system development through innovative AI solutions.