Düzce University
Journal of Science & Technology

Research Article

Diagnosis of the Skin Cancer by Vision Transformers

Uğur DEMİROĞLU [a,*]

[a] *Department of Software Engineering, Faculty of Engineering, Architecture and Design, Kahramanmaraş İstiklal University, Kahramanmaraş, TÜRKİYE*
** Corresponding author's e-mail address: ugur.demiroglu@istiklal.edu.tr*
DOI: 10.29130/dubited.1572317

ABSTRACT

Skin cancer, one of the most frequent cancers, requires early identification for efficient treatment and better survival. Early diagnosis relies on precise and quick skin lesion categorization into benign and malignant categories. This work uses Vision Transformers (ViTs) to classify skin cancer photos by modeling long-range relationships and capturing complicated visual patterns. ViTs, originally created for natural language processing, have showed great potential in picture classification tasks because to their self-attention processes, outperforming CNNs. A public collection of 270 skin lesion images—240 malignant and 30 benign—was used in this study. Preprocessing included scaling and normalizing the dataset to 384x384x3 and splitting it into 80% training and 20% testing sets. Transfer learning optimised a pre-trained ViT model for this job. To improve accuracy and avoid overfitting, hyperparameters were carefully selected for network training. Parallel computing accelerated training to 30 minutes and 20 seconds. Vision Transformers classify medical images well, according to the study. The ViT model outperformed numerous other methods with 98.15% accuracy on the test set. A confusion matrix analysis showed great sensitivity in detecting malignant lesions and low misclassification of benign patients. These findings show that ViTs can capture detailed medical picture aspects, making them useful for dermatological diagnoses. This study shows that Vision Transformers can improve diagnosis accuracy and lays the groundwork for their use in other medical imaging fields. In the battle against skin cancer and other illnesses that need early diagnosis, ViTs' scalability, efficiency, and precision are important. Future research might integrate ViTs with other deep learning architectures to improve robustness and flexibility. This study adds to the data supporting sophisticated AI in medical diagnostics and lays the groundwork for automated, reliable, and efficient healthcare solutions.

Keywords: *Skin cancer, Categorized images, Malignant and benign classes, Classification, Vision transformers*

Vision Transformers ile Cilt Kanseri Tanısı

ÖZ

En sık görülen kanserlerden biri olan cilt kanseri, etkili tedavi ve daha iyi sağ kalım için erken teşhis gerektirir. Erken teşhis, cilt lezyonlarının iyi huylu ve kötü huylu kategorilere hassas ve hızlı bir şekilde sınıflandırılmasına dayanır. Bu çalışma, uzun menzilli ilişkileri modelleyerek ve karmaşık görsel desenleri yakalayarak cilt kanseri fotoğraflarını sınıflandırmak için Görüntü Dönüştürücülerini (ViT'ler) kullanır. Başlangıçta doğal dil işleme için oluşturulan ViT'ler, kendi kendine dikkat süreçleri sayesinde CNN'leri geride bırakarak resim sınıflandırma görevlerinde büyük potansiyel göstermiştir. Bu çalışmada 240 kötü huylu ve 30 iyi huylu olmak üzere 270 cilt lezyonu görüntüsünün genel koleksiyonu kullanılmıştır. Ön işleme, veri kümesinin 384x384x3'e ölçeklenmesini ve normalleştirilmesini ve %80 eğitim ve %20 test kümelerine bölünmesini içeriyordu. Transfer öğrenme, bu iş için önceden eğitilmiş bir ViT modelini optimize etti. Doğruluğu artırmak ve aşırı uyumu önlemek için, ağ eğitimi için hiperparametreler dikkatlice seçildi. Paralel hesaplama, eğitimi 30 dakika 20 saniyeye hızlandırdı. Çalışmaya göre, Görüntü Dönüştürücüler tıbbi görüntüleri iyi sınıflandırıyor. ViT modeli, test setinde %98,15 doğrulukla diğer birçok yöntemi geride bıraktı. Bir karışıklık matrisi analizi, kötü huylu lezyonları tespit etmede büyük

Geliş: 23/10/2024, Düzeltme: 07/12/2024, Kabul: 12/12/2024

hassasiyet ve iyi huylu hastaların düşük yanlış sınıflandırılması gösterdi. Bu bulgular, ViT'lerin ayrıntılı tıbbi resim yönlerini yakalayabildiğini ve bu sayede dermatolojik teşhisler için yararlı hale geldiğini gösteriyor. Bu çalışma, Görüntü Dönüştürücülerin teşhis doğruluğunu artırabileceğini ve diğer tıbbi görüntüleme alanlarında kullanımları için temel oluşturduğunu gösteriyor. Cilt kanseri ve erken teşhis gerektiren diğer hastalıklarla mücadelede, ViT'lerin ölçeklenebilirliği, verimliliği ve hassasiyeti önemlidir. Gelecekteki araştırmalar, sağlamlığı ve esnekliği artırmak için ViT'leri diğer derin öğrenme mimarileriyle entegre edebilir. Bu çalışma, tıbbi teşhislerde sofistike yapay zekayı destekleyen verilere katkıda bulunuyor ve otomatik, güvenilir ve verimli sağlık çözümleri için temel oluşturuyor.

Anahtar Kelimeler: *Cilt kanseri, Kategorize edilmiş görüntüler, Kötü huylu ve iyi huylu sınıflar, Sınıflandırma, Vision transformers*

I. INTRODUCTION

Cancer is a disease characterized by the uncontrolled proliferation and spread of aberrant cells. [1] These aberrant cells can develop tumors, which can infiltrate adjacent tissues or spread throughout the body via the bloodstream or lymphatic system. There are numerous varieties of cancer, each with its own distinct features and treatment choices. The causes of cancer are diverse and vary by type. The majority of malignancies are thought to be caused by a combination of hereditary and environmental causes. Skin cancer is one of the most prevalent types of cancer. It occurs when abnormal skin cells multiply uncontrollably. There are three forms of skin cancer: basal cell carcinoma, squamous cell carcinoma, and melanoma. Basal cell carcinoma and squamous cell carcinoma are commonly known as non-melanoma skin malignancies. They are typically curable, but if neglected, they can spread to other parts of the body. Melanoma is the most deadly form of skin cancer and can be fatal if not diagnosed and treated promptly. It is caused by the unregulated proliferation of melanocytes, which create melanin, the pigment that gives skin its color [2].

Early detection of cancer is critical for successful treatment and better results. Detecting cancer in its early stages allows for more effective treatment options, which frequently result in improved survival rates. Early detection may also lessen the need for severe therapies such as surgery, chemotherapy, or radiation therapy. Furthermore, early discovery can help prevent cancer from spreading to other parts of the body, considerably increasing the likelihood of a complete cure. As a result, regular screenings and checkups are critical for early cancer identification and improved general health. For the diagnosis of the diseases from medical images, computerized techniques have gained much attention recently.

Classification systems for medical images serve an important role in healthcare since they allow for accurate and efficient diagnosis and treatment. These strategies use algorithms to classify medical images based on their visual qualities. For example, they can be used to distinguish between benign and malignant tumors, identify different types of diseases, and track disease progression over time. By automating the classification process, medical personnel can save time while improving diagnostic accuracy, resulting in better patient outcomes. Furthermore, classification approaches can be used to help with treatment planning and monitoring, ensuring that patients receive appropriate care. This paper uses the Vision Transformers (ViT) approach for classification [3]. ViT has had a considerable impact on the field of computer vision, namely picture classification. Unlike typical convolutional neural networks (CNNs), ViT employs transformers, a sequence-to-sequence modeling architecture initially designed for natural language processing. ViT divides images into patches, flattens them into vectors, and then feeds them via a transformer encoder. This method enables ViT to handle photos of varying sizes and resolutions efficiently. Furthermore, ViT has shown competitive performance with CNNs, particularly on large-scale datasets. Their scalability, flexibility, and solid theoretical underpinning make them an appealing option for a variety of computer vision tasks, including medical picture processing and distant sensing. Additional information regarding the approach can be found in the following sections.

Based on the evidence presented above, computerized techniques are extremely important in the early detection of cancers. This hypothesis has piqued the interest of experts, leading to useful studies on the subject. It would be useful to give a brief literature survey. Subba and Sunaniya used an attention-based GoogLeNet-style CNN to improve brain tumor classification in [4] and Elbedoui et al. investigated deep learning methods for dermoscopic image-based skin cancer detection in [5]. More specifically, valuable studies focused on skin cancer detection implementing the Vision Transformers method are listed in Table 1, briefly.

This study significantly advances the domain of medical image analysis by investigating the use of Vision Transformers (ViTs) for classifying skin cancer pictures. The study clearly illustrates that ViTs, through their self-attention mechanism, can capture intricate patterns in high-dimensional picture data, exceeding the accuracy and computing efficiency of typical Convolutional Neural Networks (CNNs). The authors get a remarkable classification accuracy of 98.15% using a publicly accessible skin cancer dataset, underscoring the model's efficacy in differentiating between benign and malignant tumors. The approach of the work include dataset preparation, training the ViT model with optimal hyperparameters, and utilizing parallel computing to improve the training process. The research presents ViTs as a scalable and adaptable option for medical imaging applications, highlighting their capacity to enhance diagnostic precision and enable early skin cancer diagnosis. This study situates its findings within the wider context of deep learning progress, citing relevant research that combines Vision Transformers with other designs to improve performance. The research highlights the transformational potential of ViTs in dermatological diagnostics and sets the stage for future advancements in automated, AI-driven healthcare solutions.

This study is organized in the following way. Section 2 gives information about the used dataset including the images. Section 3 recalls the Vision Transformers which is the main method used in this study. A case study is given in Section 4 presenting the results obtained and the last section has the conclusions.

Table 1. *Studies focused on skin cancer detection using the vision transformers*

Study	Year	Classifying Method
Elbedoui et al [5]	2024	Efficient andaccurate pre-trained models, EfficientNet-B0-V2 and Vision Transformers ViT-b16
Mejri et al [6]	2024	VGG-16(Visual Geometry Group) and ResNet-50 model (ResidualNetwork)
Hameed et al [7]	2024	Vision Transformers
Kumar et al [8]	2024	Combines the MobileNetV2and Vision Transformer (ViT) architectures
Dagnaw et al [9]	2024	Two pretrained ViTs, three Swin transformers, five pretrained CNNs, and three visual-based explainable artificial intelligence (XAI) models.The ViT-base, ViT-large, Swin-tiny, Swin-base, and Swin-small transformer models and VGG19, ResNet18,ResNet50, MobileNetV2, and DenseNet201 pretrained CNN models
Yang et al [10]	2024	Inception ResNet, ViT-Base
Remya et al [11]	2024	Vision Transformer model with transfer learning, channel attention mechanism, and ROI for the accurate detection

Table 1 (cont). *Studies focused on skin cancer detection using the vision transformers*

Abou et al [12]	2024	Pre-trained ImageNet architectures and Vision Transformer models
Ashfaq et al [13]	2023	Various deep neural networks (DNN), Vision Transformer (ViT)
Islam et al [14]	2023	Combines DenseNet121 and Xception, two cutting-edge pretrained CNNs, to extract useful features
Desale et al [15]	2024	A combination of techniques such as piecewise linear bottom hat filtering, adaptive median filtering, Gaussian filtering, and an enhanced gradient intensity method is used for pre-processing
Kusumastuti et al [16]	2023	Convolutional Neural Networks (CNN),EfficientNet, and Vision Transformer (VIT) methods

II. THE SKIN CANCER DATASET

The dataset used in this study is a skin cancer dataset, categorized into malignant and benign images, and was downloaded from Kaggle [17]. This dataset contains images of skin lesions classified into two categories: malignant and benign. It is organized into two folders:

- *Malignant:* This folder contains images of skin lesions diagnosed as malignant, indicating the presence of skin cancer. These images can be used to train models that detect and differentiate malignant skin conditions.
- *Benign:* This folder contains images of non-cancerous skin lesions that are not immediately life-threatening. These images are essential for training models to accurately distinguish between harmful and harmless skin conditions.

This dataset can be used for tasks such as image classification, deep learning model training, and medical image analysis. Researchers and practitioners can leverage this dataset to develop models that assist in the early detection and diagnosis of skin cancer. The dataset consists of 270 images, with 30 benign and 240 malignant cases, totaling 7MB in size. The images have a resolution of 96 DPI, with a 24-bit depth, and all images have a width and height of at least 200 pixels. The image format is JPG, and the benign/malignant status is used to label cancer types.

The dataset, which is under a public license, is widely used in medical, cancer, and computer vision research fields. It is freely accessible for download and continuous use in these areas. Fig. 1 shows sample images from benign and malignant cases. The dataset used in this study was chosen due to its suitability for the research objective: classifying skin cancer images into benign and malignant categories. This publicly available dataset, sourced from Kaggle, provides a reliable and diverse set of skin lesion images specifically categorized to facilitate machine learning tasks in medical diagnostics. It includes 270 images, with 240 malignant and 30 benign samples, ensuring a clear representation of both classes required for training and testing classification models. The dataset's organization into well-defined categories aligns with the goals of this study, enabling the Vision Transformer (ViT) model to effectively learn the distinctions between malignant and benign lesions. Additionally, the dataset’s manageable size (7MB) and standardized image resolution (minimum 200 pixels in width and height, with 24-bit depth) make it computationally efficient for model training without requiring extensive preprocessing. Its public license further promotes reproducibility and accessibility, ensuring that findings from this research can be validated and extended by other researchers. These characteristics collectively make the dataset an ideal choice for exploring the potential of ViTs in skin cancer diagnosis.

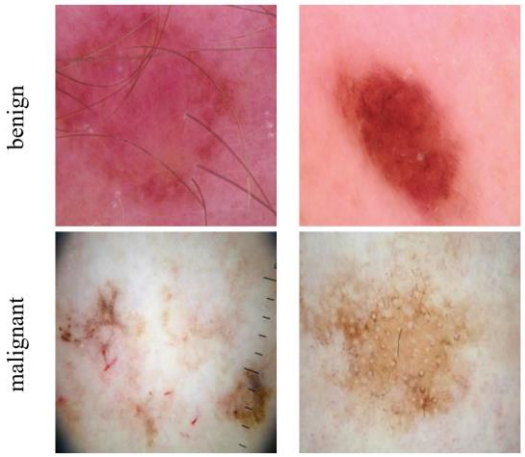


Figure 1. *Sample images from the dataset.*

III. THE VISION TRANSFORMERS

The ViT is a deep learning architecture originally developed for Natural Language Processing (NLP), designed to overcome the limitations of Recurrent Neural Networks (RNNs) in learning long-range dependencies. The Transformer can directly model the relationships between all words in an input sequence using the self-attention mechanism. As a result, Transformer models can achieve faster and more accurate results compared to RNNs. The Transformer architecture is composed of the following components.

ViT, or Vision Transformer, is a model for image classification that uses a Transformer-like architecture on image patches. An image is divided into fixed-size patches, each of which is then linearly embedded, position embeddings are added, and the resulting sequence of vectors is fed into a standard Transformer encoder. To perform classification, the standard approach of adding an extra learnable "classification token" to the sequence is used. The ViT architecture is given in Fig. 2 [18].

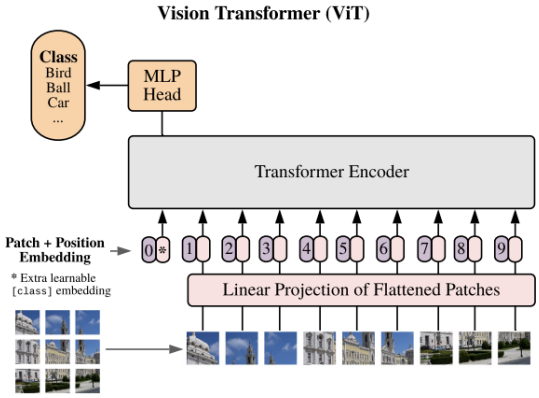


Figure 2. *The ViT architecture.*

Since their introduction in NLP, transformers have gained significant attention, achieving highly successful results and often replacing models like Long Short-Term Memories (LSTMs). The most notable feature of transformers is the self-attention mechanism, which analyzes relationships between all words in a sequence. This mechanism allows the model to treat the input as a whole. In contrast, convolutional layers in Convolutional Neural Networks (CNNs) process information based on the filter size, which is referred to as locality in the literature. This locality limits CNNs from capturing global semantic features, whereas transformers, by processing the entire input holistically, can extract strong global semantic features.

Unlike traditional CNNs, which eliminate the need for manually crafted features, the Vision Transformer (ViT) differentiates itself by leveraging a self-attention mechanism to gather global contextual information from the entire image. This approach divides an input image into fixed-size patches, embeds each patch into high-dimensional vectors, and processes these vectors using a transformer encoder. This methodology enhances ViT's ability to capture complex long-term dependencies and subtle relationships between different regions of the image.

In this study, a pre-trained Vision Transformer (ViT) neural network will be used for the classification of brain MRI images. ViT is a neural network model that employs a transformer architecture to encode image inputs into feature vectors. The network consists of two main components: the backbone and the head. The backbone is responsible for the encoding step, where it takes input images and extracts feature vectors. The head is responsible for making predictions by mapping the encoded feature vectors to prediction scores. In this study, the pre-trained ViT network has already learned strong feature representations for images. Using transfer learning, the model can be fine-tuned for better performance on specific tasks. The block diagram of the ViT network can be found in Fig. 3 [19].

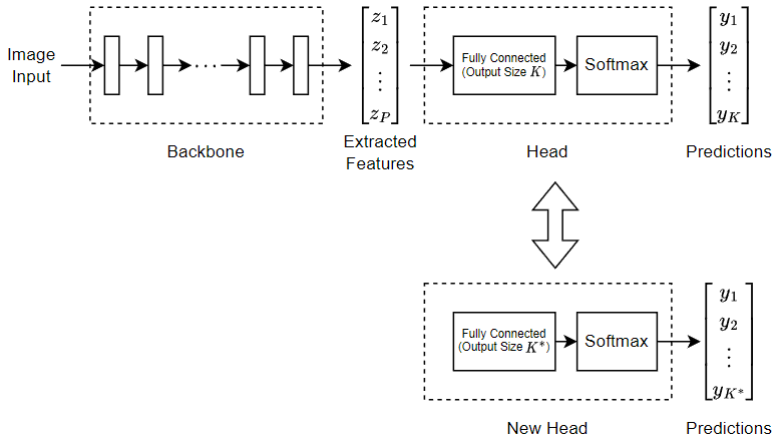


Figure 3. *Block diagram of the ViT network.*

This diagram outlines the architecture of a ViT network that makes predictions for K classes and illustrates how the network is structured to enable transfer learning for a new dataset with K classes. The diagram highlights the process of adapting the pre-trained ViT model to the specific task by fine-tuning the network for the new dataset. The backbone of the ViT extracts feature representations from the input images, while the classification head is modified to output predictions for the new K classes. This adjustment enables the network to learn and generalize effectively on the new dataset while leveraging the pre-trained knowledge.

IV. APPLICATION

80% of the dataset is used for training, with the remaining 20% reserved for non-training test data. The images were scaled and normalized to 384x384x3 and processed as color images for both training and testing. The training parameters are set at Mini Batch Size = 16, Maximum Epochs = 5, Iterations Per Epoch = 13, and Validation Frequency = 3. The network was trained using Stochastic Gradient Descent with Momentum (SGDM) as the optimizer, which was combined with a stochastic solver. It would be useful to describe these terms. The training of the Vision Transformer (ViT) model in this study was optimized using carefully selected parameters to balance performance and computational efficiency. A mini-batch size of 16 was employed, allowing for the processing of smaller subsets of the data at a time, which facilitates more frequent weight updates and effective utilization of computational resources. Training was conducted over 5 epochs, with each epoch comprising 13 iterations, and a validation frequency of 3 iterations was set to monitor performance and detect overfitting. The learning rate was initialized at 1e-4 to ensure stable and gradual convergence, while shuffling the data at each epoch

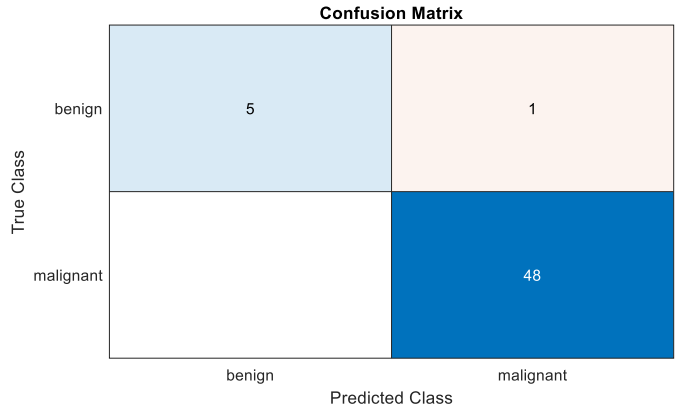
improved the model's ability to generalize by preventing it from memorizing the data order. Training was accelerated using parallel computing with 16 workers on a high-performance system featuring an NVIDIA RTX A4000 GPU. The optimization algorithm used was Stochastic Gradient Descent with Momentum (SGDM), which enhanced convergence by incorporating a momentum term. SGDM updates the model's weights iteratively based on the gradient of the loss function while adding a fraction of the previous weight update to the current step. This approach reduces oscillations, accelerates convergence, and helps navigate the loss landscape effectively, particularly in regions with small gradients. Together, these parameters and the SGDM algorithm enabled the ViT model to achieve efficient training and impressive classification accuracy of 98.15%, underscoring its capability in diagnosing skin cancer images accurately.

The use of an 80/20 train-test split instead of cross-validation in this study was mainly influenced by the dataset's particular attributes and the computing efficiency needed for training Vision Transformers (ViTs). Although cross-validation offers a more reliable assessment of a model's generalizability through the utilization of multiple training and testing folds, this methodology can be computationally intensive, especially when applied to deep learning models such as Vision Transformers, which demand considerable time and resources for training. This study utilized a dataset of 270 photos, with an 80/20 split allocating 216 images for training, therefore enabling the ViT model to discern significant patterns. The remaining 54 photos were allocated for testing, enabling an objective assessment of the model's performance. Conversely, cross-validation would have subdivided the dataset into smaller training and validation subsets, thus restricting the data available for training in each fold, which might negatively impact the performance of a data-intensive model such as ViTs. The computing burden of cross-validation is substantial. Given constrained computer resources, repeatedly training the ViT model over many folds would have considerably augmented training duration and resource expenditure without guaranteeing commensurate insights. The 80/20 division, coupled with meticulous validation throughout training (e.g., via validation frequency), facilitated a compromise between efficiency and thorough assessment of the model. Although cross-validation may provide a little more dependable assessment of generalizability, the 80/20 split was used as a pragmatic and efficient method for this investigation. Future research may integrate cross-validation should bigger datasets or enhanced computational resources be accessible, enabling a more thorough examination of the ViT model's performance resilience.

Table 2. *Specifications of the computer used in the experiment.*

Processor	12th Gen Intel(R) Core(TM) i9-12900F 2.40 GHz
Cores, Processors	16, 24
Installed RAM	64.0 GB (63.7 GB usable)
GPU	NVIDIA RTX A4000
DirectX version	12 (FL 12.1)
GPU Memory	47.9 GB (16.0 GB Dedicated, 31.9 GB Shared)
SSD Capacity	477 GB

Training was completed in 30 minutes and 20 seconds. Training accuracy was attained at 0.9815. Fig. 4 depicts the confusion matrix that was obtained.



Fiure 4. *The confusion matrix.*

Upon examining the Confusion Matrix, it is observed that out of 54 test samples, 6 are benign and 48 are malignant images. Among the benign images, 5 were correctly predicted, while 1 was misclassified. Similarly, all 48 malignant images were correctly predicted. The progress of training iterations, iteration time, mini-batch performance, test performance, and errors are presented in Table 3. This detailed breakdown provides insights into the model's performance during training and testing, showing how well it generalizes across different data points.

Table 3. *Training iterations.*

Epoch	Iteration	Time Elapsed (hh:mm:ss)	Mini-batch Accuracy	Validation Accuracy	Mini-batch Loss	Validation Loss
1	1	00:00:36	43.75%	90.74%	1.8579	0.3866
1	3	00:01:40	81.25%	88.89%	1.4480	0.9417
1	6	00:02:57	100.00%	88.89%	0.0013	1.0018
1	9	00:04:23	87.50%	88.89%	0.6754	0.5734
1	12	00:05:49	93.75%	98.15%	0.1009	0.1070
2	15	00:07:09	68.75%	96.30%	1.0546	0.1256
2	18	00:08:33	75.00%	92.59%	0.9555	0.2944
2	21	00:09:54	87.50%	94.44%	0.4395	0.2519
2	24	00:11:20	87.50%	96.30%	0.2401	0.1465
3	27	00:12:45	93.75%	98.15%	0.0954	0.0846
3	30	00:14:06	93.75%	98.15%	0.0842	0.0959
3	33	00:15:29	100.00%	98.15%	0.0029	0.1047
3	36	00:16:52	81.25%	98.15%	0.4746	0.0795
3	39	00:18:16	100.00%	98.15%	0.0046	0.0520
4	42	00:19:40	100.00%	96.30%	0.0020	0.0456
4	45	00:21:03	87.50%	98.15%	0.4422	0.0501
4	48	00:22:24	100.00%	98.15%	0.0002	0.1098
4	50	00:23:18	100.00%		0.0302	
4	51	00:23:47	87.50%	96.30%	0.6107	0.1627
5	54	00:25:11	93.75%	98.15%	0.0553	0.1036
5	57	00:26:34	100.00%	98.15%	0.0050	0.0773
5	60	00:27:55	100.00%	98.15%	0.0042	0.0613
5	63	00:29:20	100.00%	98.15%	0.0041	0.0572
5	65	00:30:19	100.00%	98.15%	0.0015	0.0454

Fig. 5 shows the training and error graphics. As shown in the figure, the test accuracy of the training process utilizing the ViT network was 98.15%. This shows the effectiveness of the method.

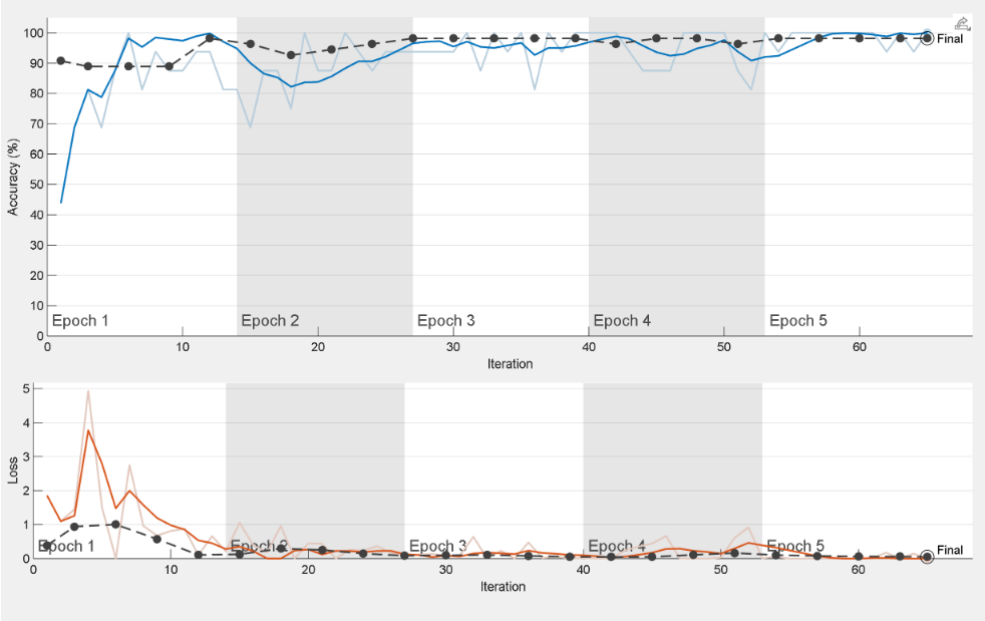


Figure 5. *Training and loss visualization.*

The selection of a modest dataset including 270 photos for this investigation was determined by the accessibility of publically available, pre-categorized skin lesion photographs. Although the dataset size may seem limited, it was adequate to illustrate the efficacy of Vision Transformers (ViTs) in differentiating between benign and malignant skin lesions. The main objective of this work was to assess the independent performance of ViTs on a realistic dataset instead of attaining optimal results by considerable dataset augmentation. Nonetheless, it is recognized that an expanded dataset frequently improves model generalizability and robustness. Techniques such as data augmentation (e.g., rotation, flipping, and scaling) or sophisticated methods like Generative Adversarial Networks (GANs) may be utilized to artificially expand the dataset size. These strategies may provide more diversity into the training data, potentially enhancing the model's capacity to generalize to unfamiliar pictures. For example, augmentation can present the model with various orientations and lighting conditions, but GANs can produce wholly new yet realistic samples that replicate the original data's underlying distribution. The choice to exclude augmentation or GANs in this investigation was intentional, concentrating on evaluating the baseline performance of the ViT model on a conventional, unaltered dataset. This method guarantees that the outcomes are only ascribed to the model's design and training regimen, rather than being affected by supplemented input. Future study may investigate the influence of these strategies on categorization accuracy. Combining ViTs with enriched datasets or GAN-generated pictures is expected to sustain or maybe improve the existing accuracy of 98.15%, hence further substantiating the model's relevance in dermatological diagnoses.

IV. CONCLUSION

This study demonstrates the efficacy of Vision Transformers (ViTs) in categorizing skin cancer photos, with an accuracy of 98.15%. This study distinguishes itself from previous research by concentrating on the usage of Vision Transformers (ViTs) as an independent methodology, rather than employing hybrid approaches that integrate ViTs with other architectures. Kumar et al. used ViTs with MobileNetV2 to augment performance, whilst Desale et al. employed intricate preprocessing methods, such as gradient-based intensity modifications, to increase classification results. This work illustrates that Vision Transformers (ViTs), when fine-tuned and trained with optimum parameters, may achieve competitive outcomes independently, without the need for substantial preprocessing or hybrid designs. This study diverges from Mejri et al., who employed pretrained CNNs such as VGG-16 and ResNet-50, and Hameed et al., who integrated several CNN and transformer models for improved segmentation and classification. This work demonstrates the practicality of a pre-trained ViT model by simplifying the

architecture and training procedure, particularly on smaller datasets. The dataset employed herein comprises 270 photos, somewhat less than those utilized in other comparative research; yet, the findings obtained are comparable to or superior to those from studies employing bigger datasets, such as the ISIC dataset utilized by Hameed et al. A further critical distinction pertains to computational efficiency. This work maximized training by parallel computing and simple preprocessing, in contrast to others that need substantial resources for training intricate hybrid models or doing complicated data augmentation. The training was accomplished in approximately 30 minutes and 20 seconds, underscoring the efficacy of ViTs in contrast to CNN-based approaches, which often need extended training durations owing to their dependence on localized feature extraction. This research highlights the independent capability of Vision Transformers as a scalable, efficient, and extremely precise instrument for skin cancer diagnostics. It distinguishes itself by attaining superior performance with less computer resources and dataset prerequisites, demonstrating the wider use of ViTs in medical imaging. Future research may examine comparisons on more extensive datasets and explore the incorporation of explainable AI methodologies to better substantiate ViTs in therapeutic applications.

V. REFERENCES

[1] B. Banushi, S. R. Joseph, B. Lum, J. J. Lee, and F. Simpson. "Endocytosis in cancer and cancer therapy," *Nature Reviews Cancer*, *23*(7), 450-473, 2023.

[2] M. Naqvi, S. Q. Gilani, T. Syed, O. Marques, and H. C. Kim. "Skin cancer detection using deep learning—a review," *Diagnostics*, 13(11), 1911, 2023.

[3] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, and D. Tao. "A survey on vision transformer," *IEEE transactions on pattern analysis and machine intelligence*, 45(1), 87-110, 2022.

[4] A. B. Subba, and A. K. Sunaniya. "Computationally optimized brain tumor classification using attention based GoogLeNet-style CNN," *Expert Systems with Applications*, 125443, 2024.

[5] K. Elbedoui, H. Mzoughi, and M. B. Slima. "Deep Learning Approaches for Dermoscopic Image-Based Skin Cancer Diagnosis. In *2024 IEEE 7th International Conference on Advanced Technologies, Signal and Image Processing (ATSIP), 2024*, Vol. 1, pp. 1-7, IEEE.

[6] S. Mejri, and A. E. Oueslati. Dermoscopic Images Classification Using Pretrained VGG-16 and ResNet-50 Models", in *2024 IEEE 7th International Conference on Advanced Technologies, Signal and Image Processing (ATSIP)*, 2024, Vol. 1, pp. 342-347, IEEE.

[7] M. Hameed, A. Zameer, and M. A. Z. Raja. "A Comprehensive Systematic Review: Advancements in Skin Cancer Classification and Segmentation Using the ISIC Dataset," *CMES-Computer Modeling in Engineering & Sciences*, 140(3), 2024.

[8] M. R. Kumar, S. Priyanga, J. S. Anusha, V. Chatiyode, J. Santiago, and P. Revath. "Synergistic Skin Cancer Classification: Vision Transformer alongside MobileNetV2," in *2023 4th International Conference on Intelligent Technologies (CONIT), 2024*, pp. 1-7, IEEE.

[9] G. H. Dagnaw, M. El Mouhtadi, and M. Mustapha. "Skin cancer classification using vision transformers and explainable artificial intelligence," *Journal of Medical Artificial Intelligence*, 2024.

[10] G. Yang, S. Luo, and J. Li. "Advancing skin cancer classification across multiple scales with attention-weighted transformers," in *Fourth Symposium on Pattern Recognition and Applications (SPRA 2023)* (Vol. 13162, pp. 30-35). SPIE, 2024.

[11] S. Remya, T. Anjali, and V. Sugumaran. “A Novel Transfer Learning Framework for Multimodal Skin Lesion Analysis,” *IEEE Access*, 2024.

[12] M. Abou Ali, F. Dornaika, I. Arganda-Carreras, H. Ali, and M. Karaouni. “Naturalize Revolution: Unprecedented AI-Driven Precision in Skin Cancer Classification Using Deep Learning,” *BioMedInformatics*, 4(1), 638-660, 2024.

[13] M. Ashfaq, and A. Ahmad. “Skin Cancer Classification with Convolutional Deep Neural Networks and Vision Transformers Using Transfer Learning. In Advances in Deep Generative Models for Medical Artificial Intelligence (pp. 151-176),” *Cham: Springer Nature* Switzerland, 2023.

[14] N. Islam, J. T. Raya., M. T. Maisha, and D. M. Farid. “Feature Fusion with Attention Mechanism for Skin Cancer Classification,” in *2023 6th International Conference on Electrical Information and Communication Technology (EICT)*, 2023, pp. 1-6, IEEE.

[15] R. P. Desale, and P. S. Patil. “An efficient multi-class classification of skin cancer using optimized vision transformer,” *Medical & Biological Engineering & Computing*, 62(3), 773-789, 2024.

[16] R. Kusumastuti, and A. Sunyoto. “Skin Cancer Classification Using EfficientNetV2 and ViT B16,” in *2023 6th International Conference on Information and Communications Technology (ICOIACT)*, 2023, pp. 395-400, IEEE.

[17] C. Kaggle. (2024, Oct 20). *Dataset* [Online]. Available: https://www.kaggle.com/datasets/shashanks1202/skin-cancer-dataset

[18] C. Paperswithcode. (2024, Oct 20). *Method* [Online]. Available: https://paperswithcode.com/method/vision-transformer

[19] C. Mathworks. (2024, Oct 20). *Matlab* [Online]. Available: https://www.mathworks.com/help/deeplearning/ug/train-vision-transformer-network-for-image-classification.html