Toward image classification based on Vision Transformer: Case study of skin lesion.

Sabri My Abdelouahed[1][0000-1111-2222-3333], Rahmouni Abla[1], Ennaji Asmae[2], and Aarab Abdellah[2]
[1] Computer science department, Faculty of sciences, University Sidi Mohamed Ben Abdallah, Morocco.
[2] Physics department, Faculty of sciences, University Sidi Mohamed Ben Abdallah, Morocco.
abdelouahed.sabri@usmba.ac.ma

Abstract: In recent years, Vision Transformer (ViT) has emerged as a groundbreaking model in the realm of image classification, leveraging self-attention mechanisms to process images effectively. This chapter presents a comprehensive investigation into the application of the Vision Transformer model in the domain of skin lesion classification. The initial section introduces the fundamental concept of attention mechanisms and their relevance to computer vision tasks. It then traces the evolution of transformers from their origins in natural language processing to their adaptation for image analysis. The main focus of this study is the Vision Transformer, which has redefined the landscape of image classification by directly processing images through self-attention mechanisms, surpassing traditional convolutional neural networks in performance. Through a detailed case study on skin lesion classification, we demonstrate the Vision Transformer's efficacy in medical image analysis, showcasing its potential for accurate and robust diagnosis. By exploring the key components and mechanisms driving the Vision Transformer's success, this chapter sheds light on its significance in image classification and its potential applications in other medical diagnostic tasks.

Key words: Image classification, Skin lesion classification, Deep learning, Attention mechanism, Transformers, Vision Transformer

1. Introduction

In the field of computer vision, image classification has been a pivotal research area with numerous applications, from object recognition to medical diagnostics. Traditional convolutional neural networks (CNNs) have long been the backbone of image classification tasks, exhibiting impressive performance across various domains. However, recent advancements in attention mechanisms and transformer-based models have sparked a paradigm shift in the way we approach visual recognition tasks.

Attention mechanisms, originally introduced in the natural language processing (NLP) domain, enable models to focus on relevant parts of the input while downplaying irrelevant

information [1]. Their ability to capture long-range dependencies and contextual information has proved highly effective in language-related tasks. With the success of attention mechanisms in NLP, researchers began exploring their application in computer vision, paving the way for novel approaches to image classification [2].

The transformer architecture, introduced in the seminal work "Attention Is All You Need," demonstrated how self-attention mechanisms could efficiently process sequential data without the need for recurrent or convolutional layers [1]. The transformer's success in NLP tasks triggered a cascade of research exploring its adaptation for computer vision problems.

In this chapter, we present the evolution of the attention mechanisms to the state-of-the-art Vision Transformer (ViT) and its application in image classification and specifically for skin lesion classification. The ViT represents a significant breakthrough, as it directly applies self-attention mechanisms to process images without relying on traditional convolutional layers. This innovative approach has opened new possibilities and challenged the dominance of CNNs in image classification tasks.

The specific case study of skin lesion classification serves as an exemplary use case for the Vision Transformer's effectiveness in medical image analysis. Skin cancer is one of the most prevalent types of cancer globally, and early and accurate diagnosis is crucial for improving patient outcomes [3]. With the ever-growing availability of medical imaging data, there is a pressing need for advanced and reliable classification systems.

In this chapter, we aim to provide a comprehensive understanding of the transition from attention mechanisms to the Vision Transformer, highlighting the key components and mechanisms that have driven this paradigm shift. Through a thorough analysis of the Vision Transformer's architecture and its application to skin lesion classification, we showcase its potential impact on medical diagnostics.

As we delve into the world of Vision Transformers and their role in image classification, we hope to shed light on the future possibilities of attention-based models and their potential in revolutionizing medical imaging and beyond.

In this research undertaking, our goal is to create a robust skin lesion classification model using the Vision Transformer (VIT) architecture. Our primary objective is to leverage the exceptional capabilities of VIT in unraveling complex image recognition tasks and achieve unparalleled accuracy in categorizing different types of skin lesions. We will work on two

different datasets, one smaller and the other larger, to ensure the generalization capability of the developed models.

The following objectives will guide our research:

1. Develop and evaluate VIT-based models for skin lesion classification: We will create four models based on different aspects of VIT, including attention mechanisms, self-attention, transforms, and the full ViT model. These models will be trained and evaluated on the two datasets, aiming to achieve superior classification accuracy compared to traditional approaches.
2. Compare the performance of VIT-based models with traditional approaches: We will conduct a comparative analysis to evaluate the classification accuracy of the VIT-based models against traditional approaches such as convolutional neural networks (CNNs). This analysis will provide insights into the advantages of utilizing VIT in skin lesion classification.
3. Provide insights and recommendations for future research and practical implementation: Through our research, we aim to contribute valuable insights and recommendations for future advancements in the field of skin lesion classification. We will offer guidance for further research, practical implementation of VIT-based models in clinical settings, and potential challenges that need to be addressed.

By accomplishing these objectives, we aspire to transform the landscape of skin lesion analysis, empowering dermatologists with accurate and efficient diagnostic tools and opening avenues for future breakthroughs at the intersection of technology and healthcare.

2. From CNN to ViT for image classification

In recent years, image recognition and classification have witnessed remarkable advancements due to the advent of deep learning techniques. Deep learning models, particularly Convolutional Neural Networks (CNNs), have demonstrated impressive performance in various computer vision tasks. However, the traditional approaches to image recognition have certain limitations that have sparked the exploration of alternative architectures. Inspired by the success of transformer models in Natural Language Processing (NLP), researchers have started exploring the application of transformers in computer vision tasks. This has led to the emergence of the Vision Transformer (VIT)

model, which represents a groundbreaking approach to image recognition and classification.

Image recognition and classification have relied on a variety of traditional handcrafted feature extraction methods [4]. These methods aim to extract distinctive features from images that can be used to identify and classify objects. Let's explore some commonly used techniques:

- Textural features [5]: The most known texture descriptors are the Gray Level Co-Occurrence Matrix (GLCM), Gabor filter, Local Binary Pattern (LBP) and Histogram of Oriented Gradient (HOG).
- Color features [4]: In order to calculate the color that the lesion contains, according to the literature the melanoma are described by the presence of six different colors that are, white, red, light brown, dark brown, blue-gray and black.
- Shape features [6-7]: Shape can be one of the most useful features that can be used in image classification. Shape features can be similarity, perimeter, surface, skeletonization, …

Deep learning has emerged as a dominant approach in the field of computer vision, revolutionizing image classification and achieving remarkable success across diverse domains. Deep learning, especially Convolutional Neural Networks (CNNs), offered a breakthrough by automatically learning hierarchical representations from raw pixel data, mitigating the need for manual feature engineering. The key components of deep learning models include convolutional layers, pooling layers, activation functions, and fully connected layers, which together create a powerful architecture capable of capturing intricate features and patterns. Deep learning models have demonstrated their prowess in large-scale image classification challenges, such as the ImageNet competition, surpassing human-level performance and achieving unprecedented accuracy rates [8]. The ability of deep learning to learn complex representations and hierarchical features has made it a cornerstone of modern computer vision systems and especially in medical images [9].

Skin lesion classification presents a critical challenge in the field of medical image analysis. Differentiating between benign and malignant skin lesions is vital for early diagnosis and appropriate treatment planning. Dermatologists rely on their expertise to assess skin lesions visually, but the sheer volume of cases and the potential for human error

necessitate automated and reliable classification systems. Transfer learning is another significant advantage of deep learning in skin lesion classification. Pre-trained CNN models, such as VGG, ResNet, and Inception, trained on large-scale generic image datasets, can be fine-tuned on smaller medical image datasets to achieve impressive performance. This transfer of knowledge allows researchers and clinicians to work with limited data while benefiting from the knowledge captured by models trained on diverse image datasets [4, 6].

However, recent advancements in deep learning, particularly the introduction of Vision Transformer (ViT), have shown promising potential in medical image analysis. This literature review aims to explore the application of ViT in medical image classification, focusing on its advantages, challenges, and performance compared to traditional CNN-based approaches.

The seminal work by Dosovitskiy et al. introduced the Vision Transformer model, showcasing its success in natural image classification tasks [2]. Their approach utilized self-attention mechanisms and transformer architecture, achieving state-of-the-art performance on standard image benchmarks such as ImageNet. While the initial application was on natural images, the potential of Vision Transformer in medical image classification was evident. In the paper [10], Chen et al. explored the application of Vision Transformer in classifying skin lesion images for melanoma detection. The authors compared the performance of ViT against traditional CNN models like ResNet and DenseNet. The results demonstrated that the Vision Transformer outperformed CNNs in terms of accuracy and robustness, even with a relatively smaller dataset. The self-attention mechanisms of ViT facilitated effective feature extraction from skin lesion images, leading to improved diagnostic capabilities. Gabriel et al. in [11] presented an application of ViT in classifying chest X-ray images for detecting common thoracic diseases. The authors proposed a modified ViT architecture to handle the unique challenges of medical images, such as high resolution and class imbalance. Their results revealed that the ViT-based model achieved higher sensitivity and specificity in comparison to traditional CNN-based models, making it a potential candidate for assisting radiologists in clinical diagnosis.

3. Vision transformer (ViT)

In this section, we set forth on an exploration of the revolutionary Vision Transformer (VIT) architecture. VIT has garnered considerable attention as a potent tool that harnesses

transformer capabilities to revolutionize visual perception and automate tasks like skin lesion classification. By employing self-attention mechanisms and other essential components, VIT presents a promising avenue for enhancing accuracy and efficiency in dermatological practices. Throughout this segment, we will delve deep into the architecture and components of VIT, unveiling its distinctive attributes and advancements. We will closely examine how VIT adeptly analyzes and processes visual information, paving the way for significant advancements in medical image analysis and classification tasks. To underscore the real-world impact of VIT, we will explore its practical applications and notable achievements, highlighting the significant positive influence it has had on the field of dermatology. By showcasing the transformative outcomes and concrete results, we aim to emphasize the profound contribution of VIT in advancing medical image analysis and classification tasks in dermatological practice.

3.1. Attention mechanism

The attention mechanism is a fundamental component in many deep learning models, particularly in sequence-to-sequence tasks such as machine translation, text summarization, and image captioning. It allows the model to focus on specific parts of the input sequence or image while generating an output. At a high level, the attention mechanism measures the relevance or importance of different elements in the input sequence and assigns weights to them. These weights indicate the attention or focus that the model should give to each element when making predictions or generating outputs. By dynamically weighing the contribution of different parts of the input, the attention mechanism enables the model to selectively attend to the most relevant information. Imagine you have a group photo of your first school. Typically, the photo shows a group of children arranged in rows, with the teacher sitting somewhere among them. Now, if someone were to ask you, "How many people are there?", you would easily answer by counting the number of heads in the photo. You don't need to consider any other details in the picture. However, if someone asked a different question, such as "Who is the teacher in the photo?", your brain instinctively knows what to do. It will focus on identifying the characteristics of an adult in the photo while disregarding the other features. This ability of our brain to selectively focus on specific elements is known as "Attention," and it is something our brain excels at.

Equation (1) presents the calculation formula of the weights in the attention mechanism.

$$c_i = \sum_{j}^{N} \alpha_{ij} \cdot h_j \tag{1}$$

In the attention mechanism, the context vector, denoted as c_i, as shown above, for the output word y_i is generated by taking a weighted sum of the annotations. The annotations refer to the intermediate representations of the input sequence obtained from an encoder in a sequence-to-sequence model. And the weights ij are computed by a softmax function given by the following equation:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{N} \exp(e_{ik})} \tag{2}$$

$$e_{ij} = a(s_{i-1}, h_i) \tag{3}$$

The output score, e_{ij}, is obtained from a feedforward neural network function, denoted as "a" This network aims to capture the alignment between the input at position j and the output at position i.

3.2. Self-attention

Self-attention is a powerful mechanism for capturing relationships and dependencies between elements in a sequence, and in Vision Transformers, it plays a crucial role in modeling the spatial relationships between image patches to enable effective image understanding and recognition [1].

Known as scaled dot-product attention, self-attention a mechanism used in deep learning models to capture dependencies between different elements in a sequence. It is a fundamental component of transformer-based architectures, which have been widely used in natural language processing and computer vision tasks [1]. In self-attention, each element in the input sequence (e.g., a patch in an image) interacts with all other elements to compute an attention score. These scores represent the importance or relevance of each element to the others in the sequence. The attention scores are then used to compute a weighted sum of the values of the elements, creating a context vector that contains information from the entire sequence.

Given a query vector Q, a set of key vectors K, and a set of value vectors V, the attention score between the query Q and each key K is calculated using the dot product between Q and K. The dot products are then scaled by the square root of the dimension of the query (or key) vectors to avoid extremely large gradients during training. The resulting scaled attention scores are then used as weights to compute a weighted sum of the value vectors V, giving us the context vector. The scaled dot-product attention can be mathematically described as follows:

$$O = \text{Attention}(Q,K,V) = EV = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{4}$$

where d is the dimension of K

In the Vision Transformers (ViT), the attention mechanism is used to process image patches and model the relationships between different patches. In ViT, the input image is divided into non-overlapping fixed-size patches, and each patch is then linearly transformed into query, key, and value vectors. These vectors are used to perform self-attention, allowing the model to attend to relevant patches while processing each patch. The output context vectors are then used for further processing in the transformer layers to model global context information in the image and enable effective image recognition or other computer vision tasks.

3.3. Vision Transformer (ViT)

Vision Transformer (ViT) is a state-of-the-art deep learning architecture that applies the Transformer model, originally designed for natural language processing tasks, to computer vision tasks. It was introduced in the landmark paper "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" by Dosovitskiy et al. in 2020 [2]. The primary idea behind Vision Transformers is to treat images as sequences of patches and process them using the Transformer's attention mechanism.

The attention mechanism and, specifically, self-attention play a crucial role in the success of Vision Transformers. By capturing long-range dependencies and enabling efficient context modeling, self-attention allows Vision Transformers to achieve state-of-the-art performance in various computer vision tasks [1].

In the context of Vision Transformers, the self-attention mechanism is used to process the image patches. Instead of using traditional convolutional layers, the Vision Transformer breaks down the input image into smaller fixed-size non-overlapping patches and flattens them into a sequence. Each patch is then linearly embedded into a lower-dimensional representation.

Figure 1 presents the architecture of ViT in case of skin lesion classification. The ViT uses as input a set of patches of the input image instead of the whole image. The patches are transformed into a two-dimensional to learn the relationship between each patch through multi-head self-attention. Considering the operation of ViT models in detail, the image X ∈ RH×W×C is reshaped into patches xp ∈ RN×(P2·C) and then mapped to D dimensions (image size: (H, W,C), patch size : $(P, P), N = \frac{HW}{P^2}$). After these patches pass

through a learnable linear projection, two-dimensional patch embeddings are derived as an output. The positional Epos ∈ R(N+1)×D is added to the patch embedding E, which concatenates the [cls] token$Z_0^0 = x$cls, to preserve position information. The embeddings pass through layers composed of multi-head self-attention, an MLP block, and Layer Normalization (LN) by the number of blocks. Among the patch embeddings derived from the transformer encoder, only the [cls] token is used as an input to the MLP head to perform the image classification task [12].

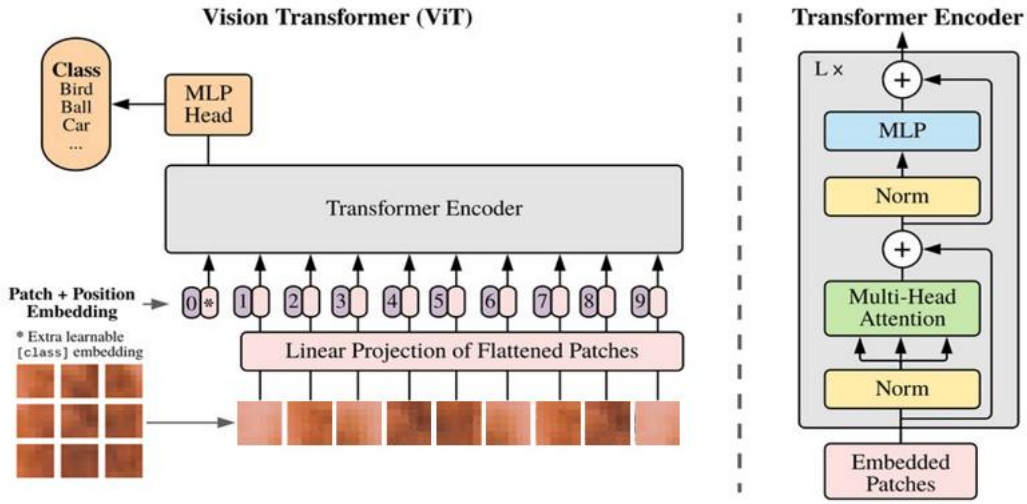


Figure 1 – Architecture of original ViT. [12]

4. Methodology

In this chapter, we aim to explore the capabilities of the Vision Transformer (ViT) architecture by conducting experiments on two different datasets. We will evaluate the performance of ViT in image classification tasks and compare it to other models, including attention-based models, self-attention-based models, and transformer-based models. By examining the performance of ViT on these diverse datasets and comparing it to established models, we seek to gain insights into its effectiveness and potential advantages in different scenarios.

4.1. Training data preparation

In deep learning, training data preparation refers to the process of organizing and preprocessing the data that will be used to train a deep learning model. This stage is crucial

for the success of the model as the quality and structure of the training data can greatly impact the model's performance.

Here are some common steps involved in training data preparation for deep learning:

- Data Collection: Gathering relevant data for your deep learning task. This could involve web scraping, accessing existing datasets, or acquiring data through other means.
- Data Cleaning: Removing or correcting any errors, inconsistencies, or outliers in the data. This may involve handling missing values, dealing with noise, or eliminating irrelevant data points.
- Data Formatting: Ensuring that the data is in a suitable format for training the deep learning model. This could involve converting the data into a specific file format, such as CSV or JSON, or structuring the data in a particular way based on the input requirements of the model.
- Data Splitting: Dividing the data into different sets for training, validation, and testing. Typically, the dataset is divided into training data (used to train the model), validation data (used to tune hyperparameters and evaluate performance during training), and testing data (used to assess the final model's performance).
- Data Preprocessing: Applying various transformations or preprocessing techniques to the data to make it more amenable for deep learning. This might include steps like normalization, feature scaling, one-hot encoding, or handling text data through tokenization and embedding.
- Data Augmentation: Generating additional training examples by applying random transformations to the existing data, such as rotation, translation, scaling, or flipping. Data augmentation can help increase the model's robustness and improve its generalization ability.
- Batch Generation: Dividing the training data into smaller batches or mini batches to enable efficient computation during the training process. This helps in performing gradientbased optimization algorithms, such as stochastic gradient descent (SGD), on subsets of the data rather than the entire dataset at once.

By following these steps, the training data can be appropriately prepared to ensure that the deep learning model receives clean, well-structured, and representative data to learn

from. This, in turn, increases the chances of the model learning meaningful patterns and producing accurate predictions or classifications when applied to new, unseen data.

4.2. Evaluation metrics

In this work, the performance measures employed include Recall, Specificity, Precision, and Accuracy. These measures are commonly used in assessing the effectiveness and performance of classification models.

- Recall: also known as Sensitivity or True Positive Rate, measures the proportion of true positive predictions out of all actual positive instances. It indicates the model's ability to correctly identify positive cases. And we can calculate it using the following formula:

$$Recall = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}} \tag{5}$$

- Specificity: it measures the proportion of true negative predictions out of all actual negative instances. It represents the model's capability to accurately identify negative cases:

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}} \tag{6}$$

- Precision: quantifies the accuracy of positive predictions by measuring the proportion of true positive predictions out of all positive predictions. It focuses on the correctness of positive classifications:

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}} \tag{7}$$

- Accuracy: provides an overall measure of the model's performance by calculating the proportion of correctly classified instances (both true positives and true negatives) out of the total number of instances. It gives an indication of the model's overall effectiveness in making correct predictions:

$$\text{Accuracy} = \frac{\text{True Positives} + \text{True Negatives}}{\text{True Positives} + \text{True Negatives} + \text{False Positives} + \text{False Negatives}} \tag{8}$$

4.3. Dataset

In our work we train our model on two distinct datasets the first one is PH2 and the second is ISIC. Examples of Images from both datasets are presented in the following figure.

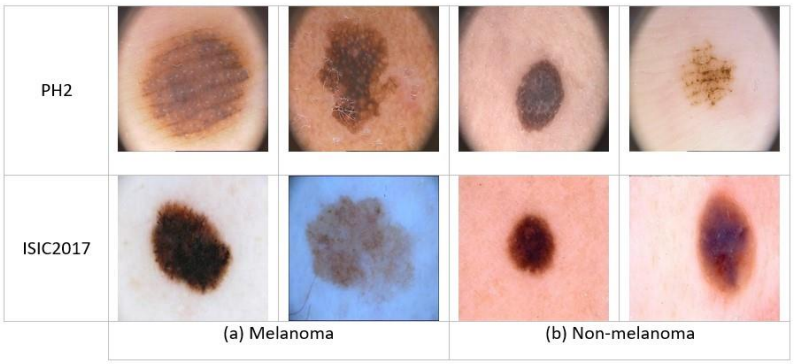


Figure 1 – Example of melanoma and non-melanoma skin cancer from PH2 database and ISIC 2017 challenge

The PH2 dataset contains 200 images: 160 non-melanomas (80 common nevi, 80 atypical nevi), and 40 melanomas skin cancer. These images are in RGB (red, green, blue) color system and have a resolution of 764*575 pixels. In Figure 3.1, the first row displays examples of the database.

The ISIC2017 dataset contains 2000 images: 1626 non-melanomas (254 seborrheic keratoses, 1372 atypical nevi), and 374 melanomas skin cancer. These images are RGB (red, green, blue) color system and have a resolution of 767*1022 pixels. The second row of Figure. 3.1 illustrates examples of the database.

The two databases are classified by experts and contain the segmentation ground-truth. As is customary and to evaluate our proposed approach, the dataset is randomly divided into training and test sets using k-fold cross-validation (5-folds in this study). That preserves the fairness of the performance of our proposed approach.

5. Results and analysis

In this section, we evaluate the performance of our custom-built ViT model on two datasets for skin lesion classification. Through rigorous testing and analysis, we assess the effectiveness and limitations of our architecture. We also conduct a comparative study to identify patterns and draw distinctions between different approaches. Finally, we discuss the results, interpret trends, and explore the implications of our findings for enhancing

diagnostic capabilities in dermatology. This examination provides valuable insights into the potential of our ViT model in skin lesion classification.

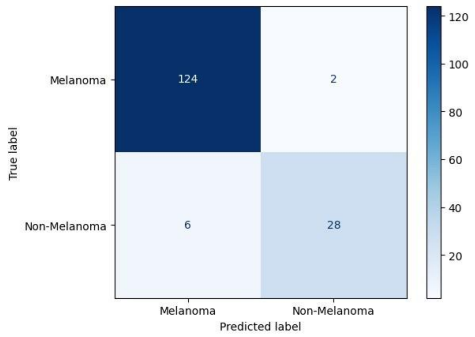
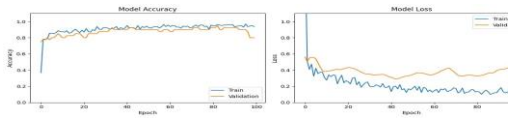
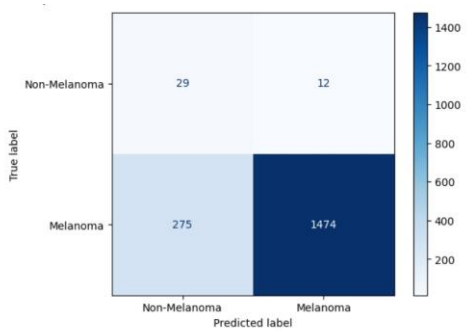
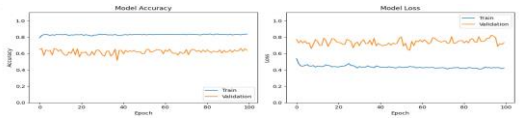
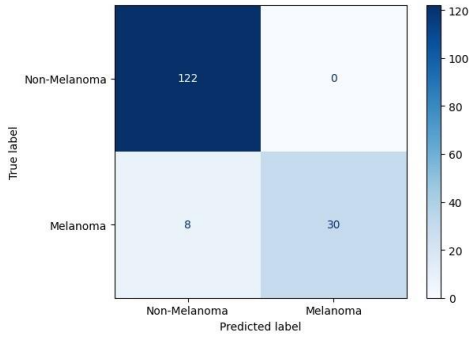
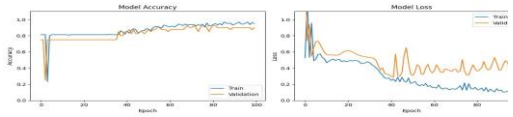
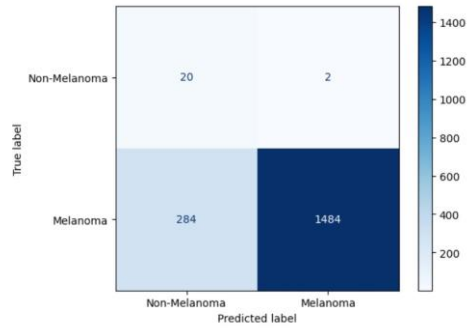
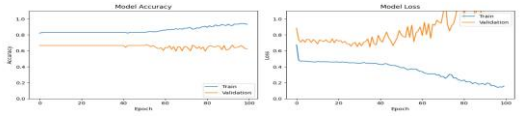
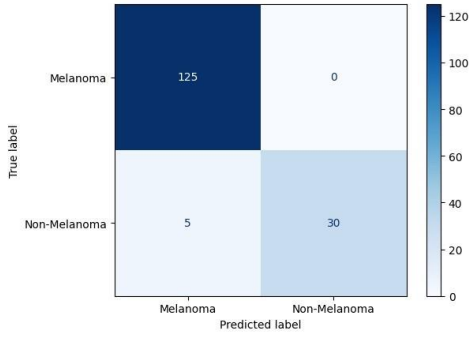
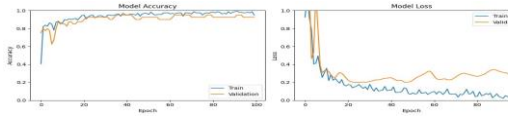
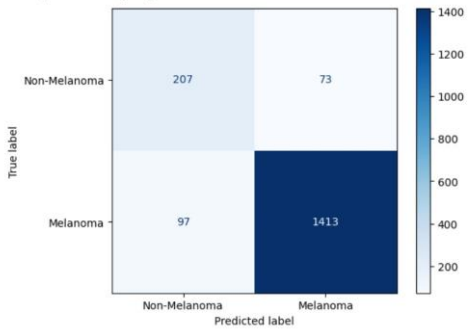
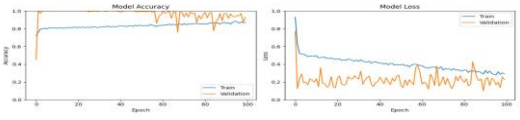
We undertake the performance evaluation of our implemented architectures using two distinct datasets: PH2 and ISIC2017. An extensive analysis was conducted to determine the effectiveness of the implemented architectures in tackling the difficulties associated with the classification of skin lesions in both datasets. Let's discuss the outcomes of three distinct architectures: the self-attention model, the based-transformer model, and the Vision Transformer (ViT).

We will use the training and test accuracies, the confusion matrices, and the Training and validation loss and accuracy curves to evaluate the proposed schemes.

Table 1 presents the simulation results of the Attention mechanism, self-attention, and Vision transformer using the two datasets: PH2 and ISIC 2017.

From the results presented in Table 1 we can conclude that the ViT model demonstrated superior performance in both datasets, surpassing the attention mechanism and self-attention based- models in accuracy. These results highlight the ViT architecture's potential in accurately classifying skin lesions. Further refinement and optimization may be needed for the self-attention and based-transformer models to improve their performance in this domain.

Table 1: The simulation results of the Attention mechanism, self-attention, and Vision transformer for the datasets PH2 and ISIC 2017

	PH2	ISIC2017
Attention mechanism	 Accuracy: 95% 	 Accuracy: 83% 
self-attention	 Accuracy: 95% 	 Accuracy: 84% 
Vision Transformer	 Accuracy: 96.875 	 Accuracy: 90% 

While Vision Transformer (VIT) models have shown promising results in various computer vision tasks, including image classification. In skin lesion classification, VIT can be seen as a good solution in comparison with the well know deep learning architecture. However, when working with VIT models for skin lesion classification, there are several limitations and challenges encounter that's needs to be treated to have a robust and powerful classification model.

- Lack of Sufficient Data: Training deep learning models like VIT requires a large amount of labeled data. However, collecting a diverse and well-annotated dataset for skin lesions can be challenging due to the need for expert dermatologist annotations and the rarity of certain types of lesions.
- Data Imbalance: Skin lesion datasets often suffer from class imbalance, where non-Melanoma class have significantly fewer samples than Melanoma. This can negatively impact the performance of VIT models, as they may struggle to learn and generalize well on underrepresented classes.
- Fine-Grained Localization: VIT models, by design, are not explicitly designed for pixel level localization tasks. Skin lesion classification typically requires identifying the lesion region within an image that are indicative of a particular lesion type. Incorporating localization capabilities into VIT models or combining them with additional algorithms for lesion segmentation can be a challenge.
- Interpretability: VIT models are generally regarded as "black boxes" due to their complex architectures and large number of parameters. Interpreting the decision-making process of VIT models for skin lesion classification can be challenging, making it difficult to understand which features or characteristics the model is relying on for its predictions.
- Generalization to Unseen Lesions: VIT models may struggle with generalizing to skin lesion types that are significantly different from the ones present in the training data. Skin lesions can vary in appearance, size, and texture, and it is important to ensure that the VIT model can handle novel lesions that were not seen during training.
- Computation and Memory Requirements: VIT models are typically computationally intensive and require substantial memory resources, particularly

when dealing with high-resolution medical images. Training and deploying VIT models may require specialized hardware or significant computational resources.

Addressing these limitations and challenges often requires a combination of strategies, such as data augmentation techniques, transfer learning, architectural modifications, ensemble methods, and domain-specific knowledge to improve the performance and applicability of VIT models for skin lesion classification tasks.

6. Conclusion

In this research work, we explored the application of the Vision Transformer (ViT) model in the domain of image recognition and classification, specifically focusing on skin lesion analysis.

Our contributions to the field of image recognition and classification are two-fold. Firstly, we demonstrated the effectiveness of the ViT model in handling complex visual tasks, such as distinguishing between melanoma and non-melanoma skin cancer. The ViT model's ability to capture both local and global features through self-attention mechanisms proved to be advantageous in accurately classifying skin lesion images. Secondly, we shed light on the significance of hyperparameter tuning in deep learning models. By carefully selecting hyperparameter values, such as learning rate, patch size, and number of transformer layers, we achieved improved model performance and mitigated common issues like overfitting and underfitting. This research provides valuable insights and guidelines for practitioners in selecting appropriate hyperparameters for similar image classification tasks. The implications and practical applications of our findings are significant. Accurate and reliable skin lesion classification is crucial for early detection and diagnosis of skin cancer. By leveraging the ViT model, healthcare professionals and dermatologists can enhance their diagnostic capabilities, leading to timely interventions and improved patient outcomes. Additionally, our research made valuable contributions to the field of computer vision and deep learning by applying various architectural approaches alongside the ViT model. We explored and compared the performance of different architectures, including self-attention, based-attention, and based-transformer, on two distinct datasets. This comprehensive analysis demonstrated the efficacy and versatility of the ViT model in comparison to other architectures, highlighting its superior capabilities in

image recognition and classification tasks. Our work showcases the power and potential of the ViT model in advancing the field of computer vision.

In summary, our research elucidates the potential of the Vision Transformer model in the realm of image recognition and classification, particularly in the context of skin lesion analysis. By optimizing hyperparameters and harnessing self-attention mechanisms, we achieved promising outcomes in accurately classifying skin lesion images. The implications for healthcare and the broader field of computer vision are significant. Further research and exploration of the ViT model hold promising prospects for advancing image analysis and classification techniques across various domains.

References

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. Advances in neural information processing systems, 30. https://doi.org/10.48550/arXiv.1706.03762

[2] D Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, el al. . (2010). An image is worth 16x16 words: Transformers for image recognition at scale. Computer Vision and Pattern Recognition. https://doi.org/10.48550/arXiv.2010.11929

[3] "Skin Cancer Foundation". https ://www.skincancer.org/skin-cancer-information/skin cancer-facts/. Consulted on: 31/07/2023

[4] Filali, Y., EL Khoukhi, H., Sabri, M.A. et Aarab, A. Efficient fusion of handcrafted and pre-trained CNNs features to classify melanoma skin cancer. MULTIMEDIA TOOLS AND APPLICATIONS (2020). https://doi.org/10.1007/s11042-020-09637-4

[5] Y. Filali, H. El Khoukhi, M. A. Sabri, A. Yahyaouy and A. Aarab, "Texture Classification of skin lesion using convolutional neural network," 2019 International Conference on Wireless Technologies, Embedded and Intelligent Systems (WITS), April 03-04, Fez, Morocco, 2019, pp. 1-5. doi: 10.1109/WITS.2019.8723791

[6] Youssef Filali, My Abdelouahed Sabri and Abdellah Aarab. (2020) Efficient skin cancer diagnosis based on deep learning approach using lesions skeleton. International Journal of Cloud Computing 2021 10:5-6, 565-578. Vol. 10, No. 5-6 . DOI: 10.1504/IJCC.2021.120395. Published Online: 10 Jan 2022. http://dx.doi.org/10.1504/IJCC.2021.120395

[7] Youssef Filali, Hasnae El Khoukhi, My Abdelouahed Sabri, Ali Yahyaouy and Abdellah Aarab. "New and Efficient Features for Skin Lesion Classification based on Skeletonization". Journal of Computer Science. September 2019, Volume 15, Issue 9. pp 1225.1236. DOI: 10.3844/jcssp.2019.1225.1236.

[8] A. Krizhevsky, I. Sutskever and G. E. Hinton, "Imagenet classification with deep convolutional neural networks", Proc. Adv. Neural Inf. Process. Syst., pp. 1097-1105, 2012.

[9] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, et al. 2017. A survey on deep learning in medical image analysis. Medical Image Analysis. Volume 42. Pages 60-88. ISSN 1361-8415. https://doi.org/10.1016/j.media.2017.07.005.

[10] Richard J. Chen, Rahul G. Krishnan. 2022. Self-Supervised Vision Transformers Learn Visual Concepts in Histopathology. Computer Vision and Pattern Recognition. https://doi.org/10.48550/arXiv.2203.00585.

[11] Gabriel Iluebe Okolo, Stamos Katsigiannis, Naeem Ramzan. 2022. IEViT: An enhanced vision transformer architecture for chest X-ray image classification. Computer Methods and Programs in Biomedicine. Volume 226. 107141. ISSN 0169-2607. https://doi.org/10.1016/j.cmpb.2022.107141.

[12] Lee, Y., Kang, P. (2022). AnoViT: Unsupervised anomaly detection and localization with vision transformer-based encoder-decoder. IEEE Access, 10, 46717-46724.