Skin Cancer Classification using Vision Transformers: Achieving SOTA in Vision Classification

Manav Garg[1]
[1]School of Computer Science and Engineering (SCOPE), Vellore Institute of Technology (VIT), Chennai, India

***Abstract*—Skin cancer is one of the most prevalent and fatal cancers, and diagnostic methods are therefore effective in making the intervention timely. Among such categories of skin cancer are melanoma, actinic keratosis, basal cell carcinoma, squamous cell carcinoma, and Merkel cell carcinoma, each being morphologically distinct to make them hard to detect and classify. Such cancers have a higher risk associated with them, as in the case of melanoma, which is an aggressive cancer and even unpredictive. It progresses quickly, and the chances of dying from it are high unless diagnosed in early stages. The chance for prompt treatment improves the patient's outcome and minimizes the chances of metastasis. Nevertheless, automatic classification of skin lesions faces enormous challenges owing to variability in lesions' appearance, color, texture, and shape that manifest differently in various forms and stages. It thus requires sophisticated computational tools able to capture all these complexities.**

Traditionally, deep CNNs are applied to the classification of medical images and have proved to attain very high success rates for tasks like that of classification of skin lesions. The reason why such architecture is good at those applications is that it is particularly suited to identify local patterns in images, which helps make the specific features of the skin lesions arise and be captured. Very recently though, transformer-based architecture, specifically Vision Transformers, came into the scene that opened a different avenue for the classification tasks of medical images. Unlike CNNs, which are typically local feature extractors, self-attention in ViTs help model the long-range dependencies in an image. This ability potentially lets the ViTs learn complex pixel relationships and attain better performance at classifying visually heterogeneous skin lesions.

This paper presents the ability of Vision Transformers for classification of skin cancer from images. This research verifies if pre-trained ViT models are fine-tuned on the curated ISIC Archive datasets of benign and malignant skin lesion varieties can classify with higher accuracy than state-of-the-art CNN-based approaches. The ISIC Archive comprises rich variability in images related to benign and malignant skin lesions. Using this approach, the study will establish whether this global-level feature representation capability in ViTs can lead to improved diagnostic accuracy in the detection of skin cancer. The outcome of this research will enhance automatic diagnostic support and enable health practitioners to make faster decisions with higher accuracy to assist in early detection and treatment of the disease for patients affected by skin cancer.

***Index Terms*—Vision Transformers, PyTorch, Skin Image Analysis, Skin Cancer, AI in medical**

I. Introduction

Computer vision, a pivotal domain within artificial intelligence and image processing, has experienced significant advancements in recent years. Convolutional Neural Networks (CNNs) have been instrumental in achieving breakthroughs across various tasks, including image classification, object detection, and segmentation. However, the inherent architecture of CNNs, which relies heavily on convolutional layers, may not adequately capture long-range dependencies in images, particularly in scenarios where global context is essential for precise analysis.

In this context, Vision Transformers (ViTs) have emerged as a novel architecture, challenging the traditional dominance of CNNs in computer vision tasks. ViTs adapt the transformer architecture—originally designed for natural language processing—to process images by dividing them into sequences of patches. This methodology enables ViTs to effectively capture global context through self-attention mechanisms, facilitating a more comprehensive analysis of entire images.

A seminal study by Dosovitskiy et al. (2020) demonstrated that reliance on CNNs is not indispensable. Their research revealed that a pure transformer, applied directly to sequences of image patches, can perform exceptionally well on image classification tasks. When pre-trained on substantial datasets and subsequently fine-tuned on various mid-sized or small image recognition benchmarks—such as ImageNet, CIFAR-100, and VTAB—the Vision Transformer (ViT) achieved superior results compared to state-of-the-art convolutional networks, while requiring significantly fewer computational resources for training.

This paradigm shift underscores the potential of transformer-based architectures in computer vision, highlighting their capacity to capture intricate patterns and dependencies that traditional CNNs might overlook. The integration of ViTs into computer vision tasks represents a promising avenue for future research and application, offering a complementary approach to existing methodologies. [18]

Despite their promising potential, ViTs are relatively new to the field of computer vision, and their effectiveness compared to CNNs in various tasks is still under exploration. In this study, we delve into the application of ViTs for the automated classification of medical images, focusing specifically on distinguishing between benign and malignant skin cancer images. By fine-tuning pre-trained ViT models on annotated medical image datasets, we aim to evaluate the performance of ViTs in comparison to traditional CNN-based approaches.

A paper authored by Ali Hassani, Steven Walton, Nikhil Shah, Abulikemu Abuduweili, Jiachen Li, Humphrey Shi [21] showed how researchers have come to believe that because of corresponding growth in parameter size and amounts of training data, transformers are not suitable for small sets of

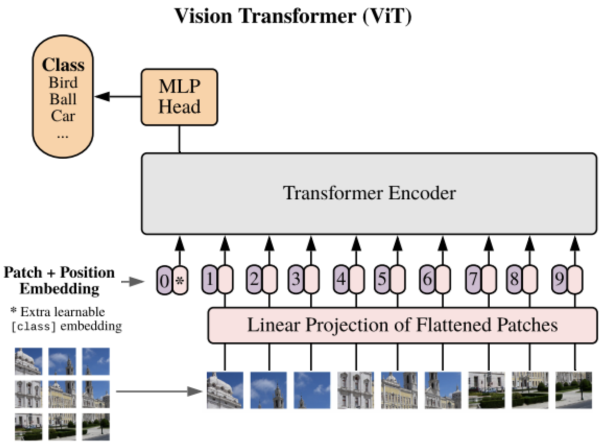


Fig. 1: Vision Transformer Model Overview

data. This trend leads to concerns such as: limited availability of data in certain scientific domains and the exclusion of those with limited resource from research in the field. The mentioned paper show for the first time that with the right size, convolutional tokenization, transformers can avoid overfitting and outperform state-of-the-art CNNs on small datasets. The models are flexible in terms of model size, and can have as little as 0.28M parameters while achieving competitive results.

Through extensive experimentation and analysis, we seek to assess the strengths and limitations of ViTs in medical image classification tasks. Our study aims to contribute to the growing body of research on ViTs and provide insights into their potential as a valuable tool for improving diagnostic accuracy in clinical settings.

II. Related Work

A paper proposed that the self-attention mechanism fails to learn effective concepts for representation learning and hinders the model from getting expected performance gain. Based on above observation, they proposed a simple yet effective method, named Re-attention, to re-generate the attention maps to increase their diversity at different layers with negligible computation and memory cost. The proposed method makes it feasible to train deeper ViT models with consistent performance improvements via minor modification to existing ViT models. Notably, when training a deep ViT model with 32 transformer blocks, the Top-1 classification accuracy can be improved by 1.6% on ImageNet. [17]

Several prior studies have explored the domain of image segmentation and object detection as well using Vision Transformers which marks notable achievements. A notable work by Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles and Hervé Jégou [19] proposed a competitive convolution-free transformer by training on Imagenet only. They trained them on a single computer in less than 3 days. Their reference vision transformer (86M parameters) achieves top-1 accuracy of 83.1% (single-crop evaluation) on ImageNet with no external data. More importantly, they introduced a teacher-student strategy specific to transformers. It relies on a distillation token ensuring that the student learns from the teacher through attention. We show the interest of this token-based distillation, especially when using a convnet as a teacher. This leads to the report results competitive with convnets for both Imagenet (where we obtain up to 85.2% accuracy) and when transferring to other tasks. [19]

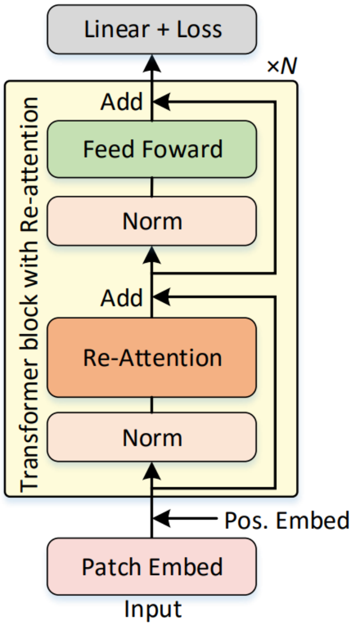


Fig. 2: Deep Vision Transformer Model Overview

Recently, there has been a growing interest in exploring Transformers for vision tasks, exemplified by the Vision Transformer (ViT) model for image classification. However, it has been observed that ViT underperforms compared to Convolutional Neural Networks (CNNs) when trained from scratch on midsize datasets like ImageNet. This is attributed to two main factors: firstly, the simplistic tokenization of input images fails to adequately capture important local structures such as edges and lines among neighboring pixels, resulting in low training sample efficiency; secondly, the redundant attention backbone design of ViT limits feature richness within fixed computation budgets and training samples. To address these limitations, a new approach called Tokens-To-Token Vision Transformer (T2T-ViT) has been proposed. T2T-ViT incorporates two key innovations: firstly, a layer-wise Tokens-to-Token (T2T) transformation is introduced to progressively structure the image into tokens by recursively aggregating neighboring tokens into one, thereby enabling the modeling of local structures represented by surrounding tokens and reducing token length. Secondly, an efficient backbone with a deep-narrow structure for vision transformers is devised, drawing inspiration from CNN architecture design following empirical study. Notably, T2T-ViT achieves a reduction in parameter count and Multiply-Accumulate operations (MACs) by half compared to vanilla ViT while yielding a more than 3.0% improvement when trained from scratch on ImageNet. Furthermore, T2T-ViT outperforms ResNets and achieves comparable performance with MobileNets when directly trained on ImageNet. For instance, a T2T-ViT model with a comparable size to ResNet50 (21.5 million parameters) can attain 83.3% top-1 accuracy with an image resolution of 384×384 on ImageNet. [11]

In a paper, titled "Diagnosis of Skin Cancer Using VGG16 and VGG19 Based Transfer Learning Models", authored by Amir Faghihi, Mohammadreza Fathollahi and Roozbeh Rajabi, they inspected skin lesion classification problem using CNN techniques. Remarkably, they presented that prominent classification accuracy of lesion detection can be obtained by proper designing and applying of transfer learning framework on pre-trained neural networks, without any requirement for data enlargement procedures i.e. merging VGG16 and VGG19 architectures pre-trained by a generic dataset with modified AlexNet network, and then, fine-tuned by a subject-specific dataset containing dermatology images. The convolution neural network was trained using 2541 images and, in particular, dropout was used to prevent the network from overfitting and finally, the validity of the model was checked by applying the K-fold cross validation method. The proposed model increased classification accuracy by 3% (from 94.2% to 98.18%) in comparison with other methods. [14]

In another paper, it introduced a groundbreaking approach to skin cancer classification, employing the Vision Transformer, a state-of-the-art deep learning architecture renowned for its success in diverse image analysis tasks. Utilizing the HAM10000 dataset of 10,015 meticulously annotated skin lesion images, the model undergoes preprocessing for enhanced robustness. The Vision Transformer, adapted to the skin cancer classification task, leverages the self-attention mechanism to capture intricate spatial dependencies, achieving superior performance over traditional deep learning architectures. Segment Anything Model aids in precise segmentation of cancerous areas, attaining high IOU and Dice Coefficient. Extensive experiments highlight the model's supremacy, particularly the Google-based ViT patch-32 variant, which achieves 96.15% accuracy. [15]

In the paper authored by Chi-en Amy Tai, Elizabeth Janes, Chris Czarnecki and Alexander Wong, they explored leveraging an efficient self-attention structure to detect skin cancer in skin lesion images and introduces a deep neural network design with DC-AC (Double-Condensing Attention Condenser) customized for skin cancer detection from skin lesion images. The final model is publicly available as a part of a global open-source initiative. [16]

III. Dataset

A. About the dataset

The dataset used for this study was from the ISIC Archive, one of the most high-performing and established archives when it comes to dermatological research, particularly in skin lesion images. ISIC Archive provides a rich collection of curated datasets, supporting the development and validation of automated diagnostic tools in dermatology. This dataset is one of the most vital collections of data that can be used in assessing skin cancer due to an abundance of labeled images both benign and malignant lesions.

For this study, the dataset consists of a total of 3,600 images; they are split equally between the two main classes, one being benign and the other malignant skin lesions, thus having 1,800 images each of them. This guarantees equal representations for classes, a very important aspect in training supervised learning models to classify the skin lesions correctly. Having an equal number of images for each class reduces the risk that could lead to any class imbalance problems, biases the model toward more frequently occurring categories, and lowers its performance on less-represented classes.

All images in the dataset are already prelabeled and therefore fit the various supervised learning approaches, thus effectively training and testing the deep learning models. It further increases the credibility of the dataset, since for all lesions, either benign or malignant, it is tagged by experts. High-quality labeling ensures that more precise examples of what the model should learn between benign and malignant lesions are obtained.

The images are standardized to deep learning frameworks, so the CNNs and Vision Transformers will process them well. Consistency of image size in terms of resolution is important: it ensures that relevant features considered by the model are not induced by variations in size or quality. Compatibility with pre-trained models also needs to be assured; they are quite often used in fine-tuning and have specific requirements for input.

Every image in the ISIC Archive dataset is controlled qualitatively following standardized imaging protocols, for instance, lighting, focus, and skin tones diversity. The results have high quality images with diversity across different skin types and lesion characteristics, and this makes the model more robust for real world applications.

The ISIC Archive dataset of this study has a nicely labeled, balanced, and high-quality foundation in training and testing deep learning models aiming to detect skin cancer. It is standardized and spans a completely comprehensive range of benign and malignant cases, best suited for validation in the effectiveness of Vision Transformers and other deep learning architectures in medical image analysis.

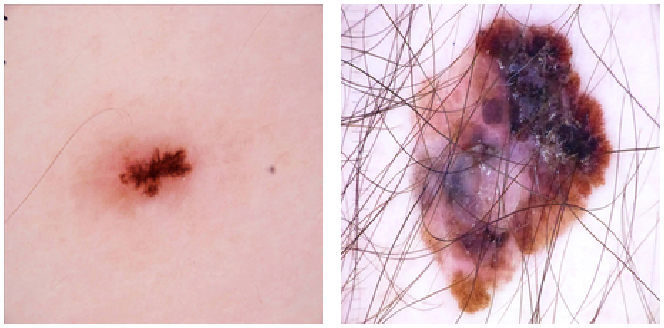
Check Fig.3, Fig.4 and Fig.5 for reference.

IV. Methodology

The methodology encompasses several key steps to effectively train a ViT (Vision Transformer) model for Skin Cancer classification using the provided dataset.

A. Data Augmentation

Data augmentation, thus, holds importance for increased diversity of the training dataset and reducing overfitting with an improvement in the model's generalization to new data. It is obtained by generating the differently transformed versions



(a) Benign (b) Malignant

Fig. 3: Visualizing the dataset

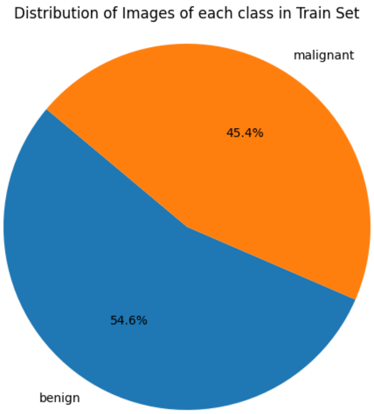


Fig. 4: Data distribution in Train set

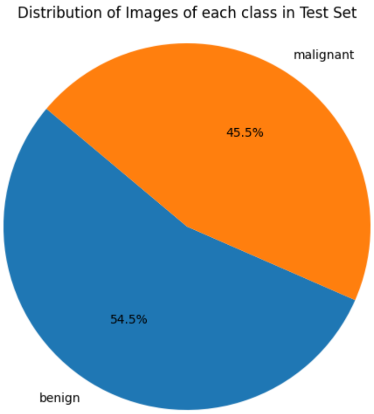


Fig. 5: Data distribution in Test set

of the original images, through which the model encounters all such transformations that it would witness in real life. In this experiment, a variety of data augmentation techniques are used with specific strengths to be sought in the datasets. This approach offers several key benefits:

Enhanced model performance
Data augmentation enriches the training dataset through diverse variations of existing data, which makes models encounter more features of the spectrum. With this diversity, models generalize well to unseen data and thus improve real-world applications. For example, if one applies some transformation such as rotation, scaling, and flipping in case of an image classification task, the model will be invariant to changes, and hence precision and robustness will increase.

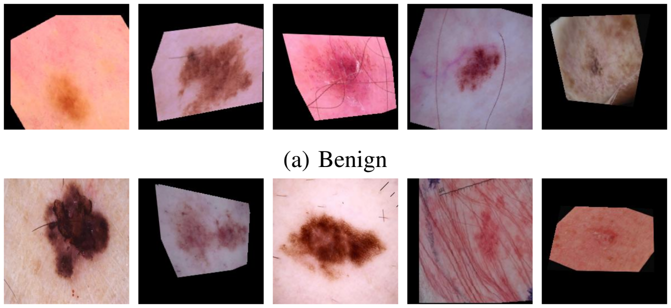
Reduced data dependency
It is costly and tedious to collect and process huge amounts of data for the purpose of training. Data augmentation alleviates this problem by making smaller datasets more useful, thus not reliant on large-scale data gathering. By supplementing existing datasets with synthetically produced data points, models are able to perform at levels comparable with those trained on much larger datasets. This is particularly helpful in domains where it is either hard or impossible to gather a large amount of data.

Mitigate overfitting in training data
Overfitting occurs when a model seems perfect for training data but has very less capacity to generalize to unseen data. Data augmentation introduces variability in a set of training which implies that the model would not memorize the specific data points. The variability challenges the model to learn more generalized features, improving its performances when actually performing well in other data sets. For example, in natural language processing, techniques like synonym replacement and random insertion can create varied sentences, helping models to better understand language nuances.

1) Random Horizontal Flip ($P = 0.5$):
   - Randomly flips the image horizontally with a probability of $P = 0.5$.
   - **Reason:** This helps to introduce diversity in the dataset, as many objects appear similarly when viewed from different angles.
2) Random Rotation ($\pm 30^\circ$):
   - Randomly rotates the image within the range of $\pm 30^\circ$.
   - **Reason:** Rotation helps in making the model more robust to variations in object orientation.
3) Color Jittering:
   - Randomly adjusts brightness, contrast, saturation, and hue of the image.

- **Reason:** Variation in color helps the model to learn features that are invariant to changes in lighting conditions.

4) Random Resized Crop (Scale: $0.8 - 1.0$):
   - Randomly crops and resizes the image to a size of 224×224 pixels, with a scale factor between 0.8 and 1.0.
   - **Reason:** This augmentation simulates images of different scales and helps the model generalize better to objects of different sizes.
5) Random Affine Transformation (Translation: ± 0.1):
   - Randomly applies affine transformations to the image, including translation within the range of ± 0.1 of the image size.
   - **Reason:** Affine transformations simulate changes in perspective, scale, and skew, which are common variations in real-world images.
6) Random Perspective Transformation:
   - Randomly applies perspective transformation to the image.
   - **Reason:** Perspective transformation simulates the effect of viewing an object from different viewpoints, enhancing the model's ability to generalize to various viewing angles.


(a) Benign

(b) Malignant

Fig. 6: Images from the augmented dataset

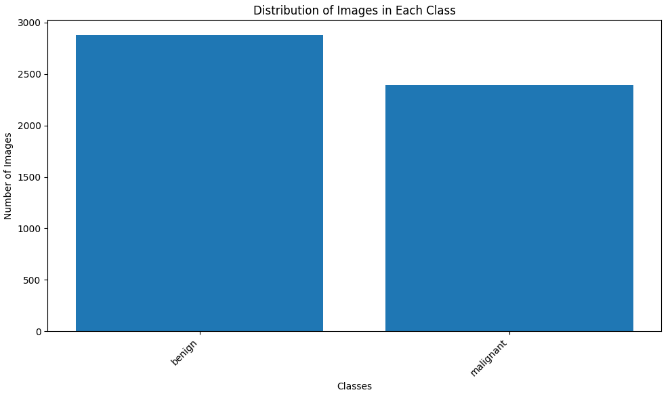


Fig. 7: Data distribution after Augmentation

These augmentations collectively enrich the training set in simulating a wide range of conditions by images. The model, in turn, captures more representative patterns and features in diverse scenarios, which helps reduce the risk of overfitting and improves performance on unseen data. By creating a varied dataset, these augmentations contribute toward building a more reliable and versatile diagnostic tool for skin cancer.

B. Dataset Loading and Preprocessing

The dataset is loaded from the specified directory using the `load_SkinCancer_dataset` function. Images are resized to a uniform size of 256×256 pixels and normalized to have a mean of $[0.485, 0.456, 0.406]$ and a standard deviation of $[0.229, 0.224, 0.225]$.

C. Dataset Splitting

The dataset is split into training, and test sets. The sizes of the splits are defined as 80%, and 20% of the total dataset size, respectively.

D. Vision Transformer Model Setup

The model configuration in the proposed study relies on the ”vit pytorch” library, which is one of the popular implementations of Vision Transformers in PyTorch. The ViT model architecture has been modified to optimize the performance over the skin cancer image classification by very carefully adjusting several key hyperparameters to improve the capability of the model to better capture meaningful patterns in medical images. The configurations are as follows:

• Size: The images used for input are resized to 256x256 pixels. Such a resolution ensures the capture of the finest skin lesion details, which will enable the model to trace out subtle patterns that might indicate malignancy. The size is well-balanced from the point of view of computational efficiency and sufficient spatial information for classification.

• Patch Size. Since the model subdivides every image into patches of 32x32 pixels, those patches are treated as tokens. Hence, a fundamental hyperparameter that controls how much granularity self-attention layers use in processing information is the patch size. The advantage of using patch size 32x32 is that each patch still holds enough contextual information; it is unlikely that its details, which are important for lesion boundaries or textures, would be lost and kept computationally reasonable.

- Model Depth: The ViT model has 6 transformer blocks, containing self-attention and feedforward sub-layers in each of them. That depth enables the model to learn progressively more complex, high-level features across multiple layers. It captures both local and global dependencies of an image. A depth of 6 allowed a balance in model complexity with the efficiency of training so that there would be enough representational power for the detection of skin cancer without overwhelming computational resources.

- Number of Heads: In a transformer block, multi-head self-attention uses 16 attention heads. Due to this configuration, it can attend to different parts of an image at the same time, therefore capturing diverse features across spatial locations. Multi-head self-attention mechanisms enhance the capacity of the model in discovering more complex lesion patterns and dependencies that may indicate malignancy.

- Embedding Dimension: 1024 The embedding dimension is set to 1024. This describes the size of the vector representation for each image patch. A high dimensional space in this case promises richness in the input so that all the information about the detailed features may be held. These 1024-dimensional embeddings enable the transformer to do really intricate feature extraction, which is basically important to differentiate the benign from malignant lesions.

To complement the comparison, this study also explores the DeepViT variant. DeepViT addresses issues in the attention mechanism to render the model more robust and dampen redundancy in learned representations. This variant is especially beneficial for medical imaging tasks where slight visual pattern differences determine critical diagnostic information.

The models are implemented in PyTorch and trained on machines with GPU to hasten the training and to host the significant computational load required by Vision Transformers. These best configurations of hyper-parameters for model optimization are proposed with high intra-class variability and slight difference in feature shape in the classes. In this sense, at varied levels of parameters, the ViT models are optimized to increase accuracy, robustness, and generalization for efficient detection of skin cancer.

We begin with importing the ViT or DeepViT model from the `vit_pytorch` library. It takes input 256 by 256 RGB images, which are then split into 32 by 32 patches and embedded into a 1024-dimensional space. The model consists of six transformer layers, each containing 16 heads, with a width of 2048 and a feed-forward network. Dropout is applied with a dropout probability of 0.5 on the embeddings and the intermediary layers. The model is configured with `num_classes=2` for binary classification and transferred to the assigned computation device, `device`.

```
from vit_pytorch import ViT, deepvit

ViT/deepvit.DeepViT(
    image_size = 256,
    channels = 3,
    patch_size = 32,
    num_classes = 2,
    dim = 1024,
    depth = 6,
    heads = 16,
    mlp_dim = 2048,
    dropout = 0.5,
    emb_dropout = 0.5
).to(device)
```

E. Loss Function and Optimizer

TThe models are trained using cross-entropy loss, which is the most common loss when dealing with the classification scenario involving multiple classes like skin cancer classification. It calculates the difference between the true label distribution and the predicted probability distribution, driving the model to minimize errors from its predictions. This loss function is very important in the task of multi-class classification as it encourages the output of a high probability for the correct class and penalizes the model for getting this class wrong.

Optimization was done using the Adam optimizer, which unifies the benefits of AdaGrad and RMSProp. This gradient adaptation is realized by adjusting the learning rate for each parameter based on estimates of the first and second moments of the gradients. Training becomes more efficient and stable this way. The learning rate should be set to 0.001, typically a value between these extremes that balance the speed of convergence against overshooting during optimization. This learning rate ensures the update of models in a gradual way and with smooth orientation toward convergence to an optimal solution without oscillation or divergence.

F. Training Loop

Training involves running the model through a specified number of epochs. Here, one epoch represents one complete pass over the training data. In each epoch, the weights of the model are updated via back propagation, wherein the weights are adjusted based on the calculated gradients. Feed some input data to the model, compute some predictions with loss computation, and update the parameters by going back through the model with that error, in order to create a training loop. It is repeated until it sees as many epochs as requested, or the performance stabilizes.

The training loop runs for a specified number of epochs (`num_epochs`). Within each epoch, the model is set to training mode, and batches of data are processed. Gradients are zeroed (`optimizer.zero_grad()`), forward pass is computed (`outputs = v(images)`), and loss is calculated (`criterion(outputs, labels)`). Back propagation is performed (`total_loss.backward()`), and model parameters are updated (`optimizer.step()`). Training loss and accuracy are computed and printed for each epoch.

V. Results and Discussion

The large difference between the results achieved when training the ViT model on the augmented versus the non-augmented dataset captures how crucial it is to apply data augmentation to increase the ability of the model to generalize and to avoid overfitting.

During the training and testing phases, much better classification accuracy was realized when training on an augmented dataset. The Base Vision Transformer (ViT) model, which had been trained on these data after augmentation, reflected in the accuracy of 85.06% in training and 89.73% in testing. The model, therefore, learns all the variations introduced by its training through these data augmentation techniques - random horizontal flipping, random rotation, color jittering, and random resized cropping. More prominent increased datasets exposed the model to a wider view of image transformations, making it all better at generalizing the unseen data, which is an important requirement of medical images owing to the variability of real-world images due to lighting conditions, angles, and scales.

The results are contrasting where results with the non-augmented dataset were much lower, and Base ViT only reached a training accuracy of 51.42% while testing produced an accuracy of 64.55%. Hence, in this scenario, a significant gap between the training accuracy and the testing accuracy indicates that possibly the model was not able to generate good generalization from the training data due to overfitting. This is primarily because without data augmentation, the model has learned directly on the basis of original unaltered images, hence more prone to overfitting. Basically, the ability of the model to learn strong features was curtailed, thus its results on the test set are very bad, probably using images that could be much different from those in the training set and in ways the model hadn't encountered previously.

There are many possible reasons for the gap between augmented models' performance and that of non-augmented models. In fact, data augmentation actually helps in generating a richer and more diverse training set by synthetically inflating the dataset with transformations that simulate real-world variations. This prevents the model from memorizing specific details on the images, which may result in overfitting. The augmented dataset encourages the model to learn far more general features that are invariant to changes such as rotation, scale, and lighting conditions. It also allows the model to be more general and robust with respect to other scenarios. Especially relevant in medical imaging scenarios, where the lesions may be observed at a wide range of orientations, lighting, and other background factors.

1) **Epochs**: Indicates the number of training epochs completed during the training process.

2) **Training Loss**: Represents the average loss (error) computed over the training dataset during the training process. A lower training loss indicates better convergence of the model during training.

3) **Training Accuracy**: Denotes the percentage of correctly classified instances in the training dataset. A higher training accuracy suggests that the model is effectively learning from the training data.

4) **Validation Accuracy**: Represents the percentage of correctly classified instances in the validation dataset. It provides an indication of how well the model generalizes to unseen data

5) **Validation Loss**: Indicates the average loss computed over the validation dataset during model evaluation. A lower validation loss suggests that the model is performing well on unseen data.

6) **Test Loss**: Similar to validation loss, test loss represents the average loss computed over the test dataset during model evaluation. It provides insight into the model's performance on unseen data.

7) **Test Accuracy**: Denotes the percentage of correctly classified instances in the test dataset. It indicates the model's overall performance on unseen data.

8) **Correct Predictions (test)**: Represents the total number of correctly classified instances in the test dataset. It provides a more granular view of the model's performance on individual instances.

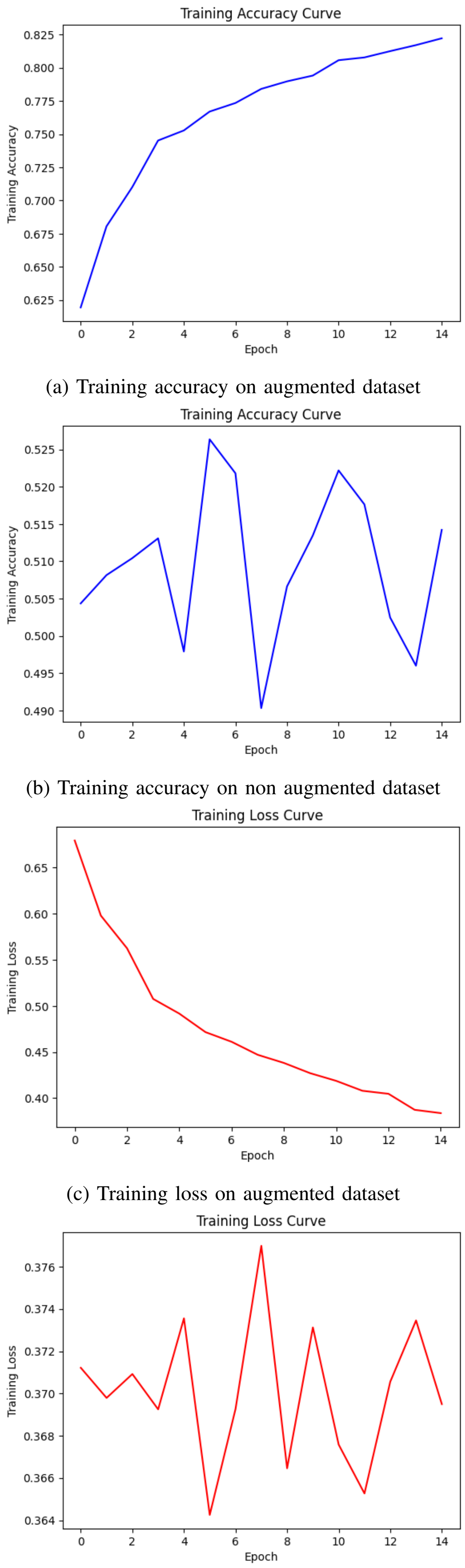
9) **Total Predictions**: Indicates the total number of instances in the test dataset. It serves as the denominator for calculating accuracy metrics.

10) **Class Labels**: Classes when extracted from the dataset were given numeric values in order to classify images more conveniently. They are as follows:

```
[0] - Benign
[1] - Malignant
```

”Benign” - not cancerous, or not harmful
”Malignant” - cancerous, or harmful

The Base vision transformer model trained on augmented dataset for a total of 15 epochs achieved classification accuracy of 85.06% (Training accuracy) and 89.73% (Testing accuracy). Out of 660 total predictions, 546 were correctly classified. Total training loss was reported to be 0.1712 and total test loss was reported to be 0.4105.



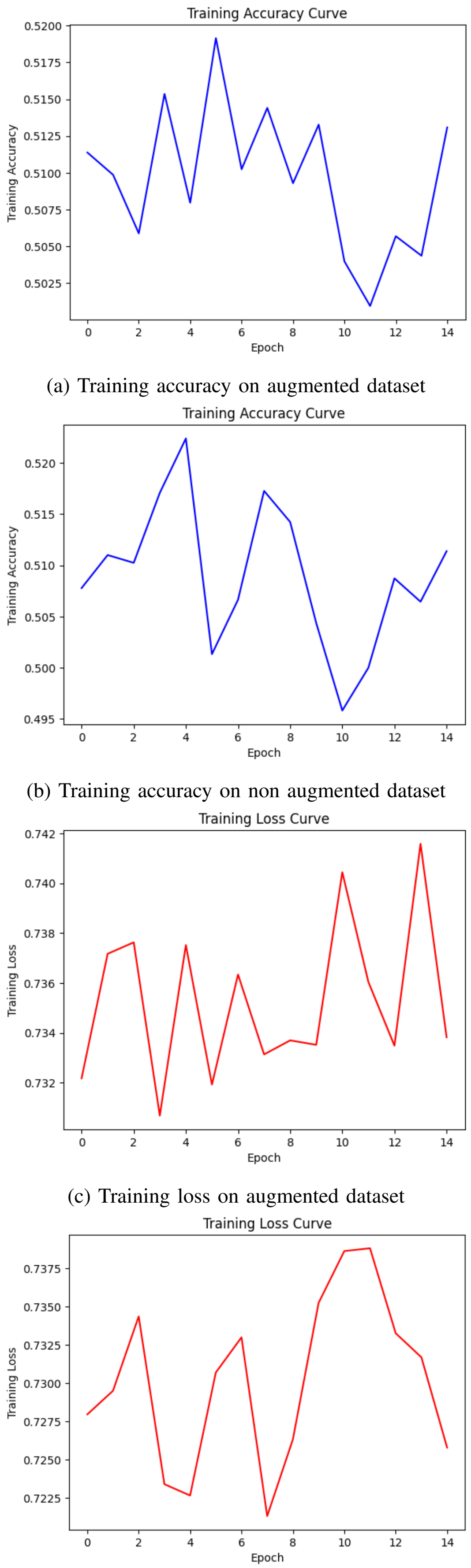
(a) Training accuracy on augmented dataset

(b) Training accuracy on non augmented dataset

(c) Training loss on augmented dataset

(d) Training loss on non augmented dataset

Fig. 8: Base Vision Transformer



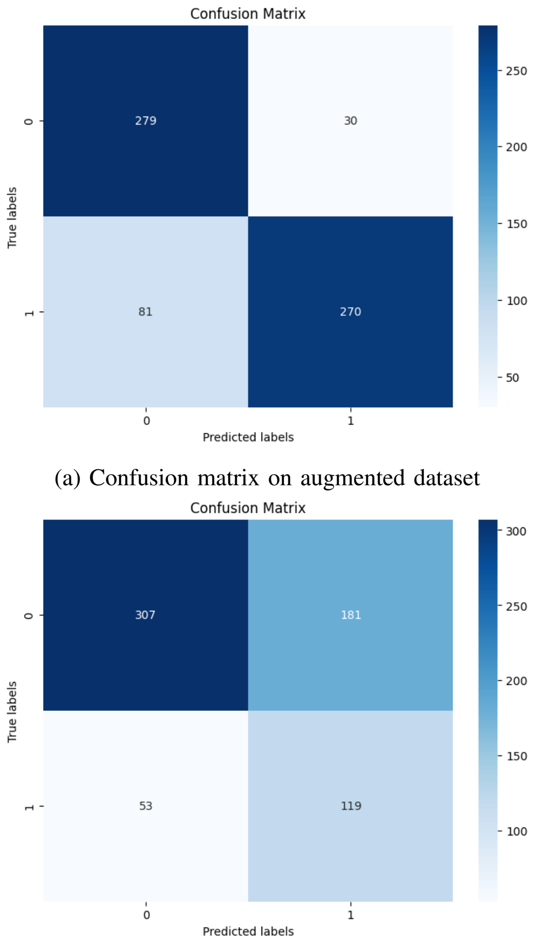
(a) Training accuracy on augmented dataset

(b) Training accuracy on non augmented dataset

(c) Training loss on augmented dataset

(d) Training loss on non augmented dataset

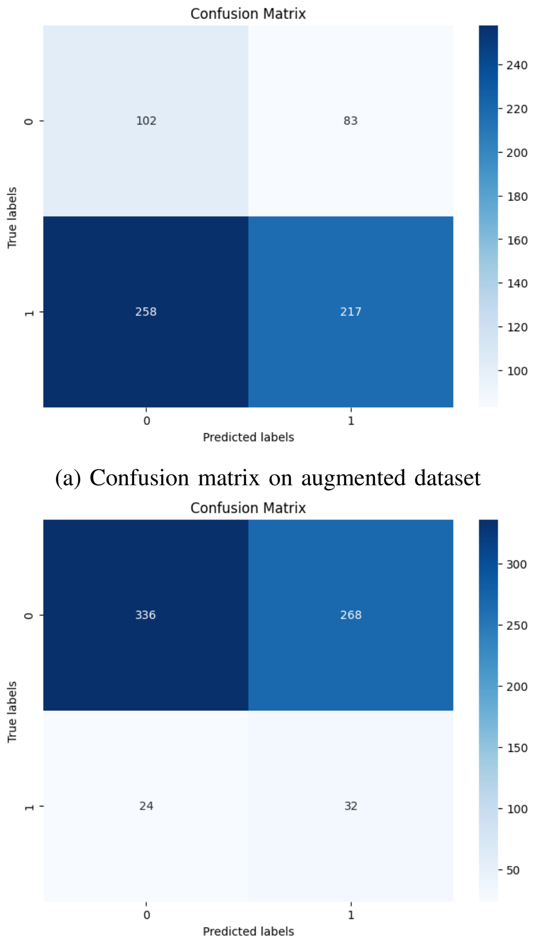
Fig. 9: Deep Vision Transformer



(a) Confusion matrix on augmented dataset

(b) Confusion matrix on non augmented dataset

Fig. 10: Base Vision Transformer



(a) Confusion matrix on augmented dataset

(b) Confusion matrix on non augmented dataset

Fig. 11: Deep Vision Transformer

The Base vision transformer model trained on non augmented dataset for a total of 15 epochs achieved classification accuracy of 51.42% (training accuracy) and 64.55% (testing accuracy). Out of 660 total predictions, only 426 were correctly classified. Total training loss was reported to be 0.3695 and total test loss was reported to be 0.6837.

The Deep vision transformer model trained on augmented dataset for a total of 15 epochs achieved classification accuracy of 51.31% (training accuracy) and 48.33% (testing accuracy). Out of 660 total predictions, only 319 were correctly classified. Total training loss was reported to be 0.7338 and total test loss was reported to be 0.6809.

The Deep vision transformer model trained on non augmented dataset for a total of 15 epochs achieved classification accuracy of 51.14% (training accuracy) and 55.76% (testing accuracy). Out of total 660 predictions, only 368 were correctly classified. Total training loss was reported to be 0.7258 and total test loss was reported to be 0.6814.

The Base Vision Transformer model demonstrated the significant impact of data augmentation on model performance. When trained on the augmented dataset for 15 epochs, it achieved an impressive classification accuracy of 85.06% during training and 89.73% on the test set, correctly classifying 546 out of 660 predictions. Moreover, the training and test losses were relatively low at 0.1712 and 0.4105, respectively. However, without data augmentation, the model's performance dropped substantially, with a training accuracy of 51.42% and a test accuracy of 64.55%, correctly classifying only 426 out of 660 predictions. The training and test losses also increased to 0.3695 and 0.6837, respectively.

These results highlight the importance of data augmentation techniques in improving model performance, particularly for the Base Vision Transformer model, which demonstrated a significant boost in classification accuracy and reduced losses when trained on the augmented dataset. The Deep Vision Transformer model, on the other hand, did not benefit substantially from data augmentation, suggesting that its architecture or other factors may have limited its performance.

To test the model on unseen data a custom dataset was made by combining images from ”benign” and ”malignant” classes into a single directory and then was passed as an input to the trained model. Refer section **VI**.

VI. Classifying images from a custom dataset

A custom Dataset was made combining random images from ”bening” class and ”malignant” into a single folder.

```
    * 45-127(.jpgs) : Bening
    * 185-247(.jpgs) : Malignant
```

A. Result

```
Image: 105.jpg, Predicted class: Benign
Image: 119.jpg, Predicted class: Malignant
Image: 122.jpg, Predicted class: Benign
Image: 127.jpg, Predicted class: Benign
Image: 185.jpg, Predicted class: Malignant
Image: 186.jpg, Predicted class: Benign
Image: 190.jpg, Predicted class: Malignant
Image: 193.jpg, Predicted class: Malignant
.
.
.
Image: 88.jpg, Predicted class: Benign
Image: 90.jpg, Predicted class: Benign
Image: 95.jpg, Predicted class: Benign
Image: 97.jpg, Predicted class: Benign
Image: 98.jpg, Predicted class: Benign
```

From the above, we can conclude that our model has an accuracy of: **93.55%**.

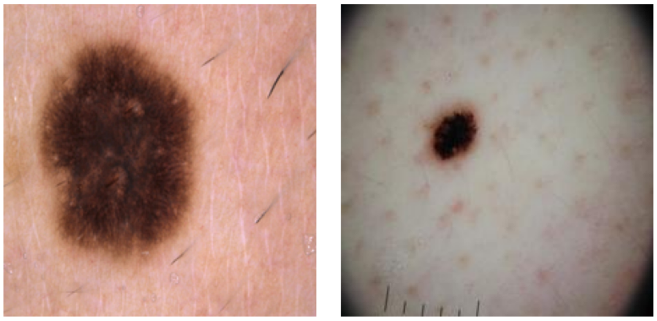
$$\text{Accuracy} = \left(\frac{\text{Number of Correct Predictions}}{\text{Total Number of Predictions}} \right) \times 100\%$$

However, We can see that our Model confused in classifying 2 images (119.jpg and 186.jpg) and made in correct predictions.

```
119.jpg
 -> Correct Class: Benign
 -> Predicted Class: Malignant
```

```
186.jpg
 -> Correct Class: Malignant
 -> Predicted Class: Bening
```

Based on the misclassification of two images by our machine learning model, it appears that the model is struggling to differentiate between the two due to the similarity in the way of their respective skin discoloration. This suggests that the features used by the model to distinguish between the conditions might not be robust enough to capture subtle variations in skin discoloration patterns. As a result, the model is encountering difficulty in accurately identifying and classifying instances where these patterns are similar. This underscores the importance of further refining the model’s feature selection process and training data to enhance its ability to discriminate between these visually similar cases. Additionally, it may be beneficial to explore alternative approaches or incorporate additional contextual information to improve the model’s performance in distinguishing between these closely related conditions.



(a) 119.jpg

(b) 186.jpg

Fig. 12: Misdiagnosed Images from the dataset

VII. Error Analysis

Detailed analysis on the mistake was done to determine the kind of misclassification that occurred with Vision Transformers, especially in cases where benign images were classified as malignant and vice versa. Upon detailed analysis, there were two kinds of misclassifications which could be further improved.

Misclassification Type 1: Benign skin lesions classified as malignant. In such circumstances, the model was confused by tiny discolorations or pigmentation patterns that resembled those in malignant lesions. This also tends to happen in lesions of abnormal coloration or with irregular borders. Such lesions are common in benign moles and tend to appear almost indistinguishable from the malignant ones, especially when not seen at the best resolution or are overly magnified. These micro-visual features were misleading the decision-making of the model towards false positives. Perhaps, the use of pixel-based feature usage led to an inefficiency in distinguishing between lesions, which have near-alike appearances with the medical distinction, requiring more potent feature extraction to capture the fine texture and color nuances.

Misclassification Type 2: Malignant lesions that were being classified as benign. In each of these examples above, malignant features were not detected by the model, for example asymmetry, borders not being smooth and round, or pigmentation irregular. These are failures that can be attributed to the model’s weaknesses in identifying complex high-level features necessary to distinguish malignant from benign lesions. Some malignant lesions, such as early melanomas, may possess characteristics that closely resemble those of benign conditions. For example, they could display minimum asymmetry or barely noticeable differences in color. These complicated patterns are challenges to the model, especially in cases when the model is trained on very limited data that does not clearly identify such subtle features.

This error analysis questions the need for further refinement in the process of feature extraction, especially between visually similar classes. It also emphasizes data diversity and augmentation to include more cases of atypical or challenging lesion instances so that the model learns to differentiate between such subtle variability. Advanced image processing techniques, which incorporate texture analysis, color histograms, and enhanced boundary detection, would further facilitate the extraction of features necessary for higher accuracy in classification. A greater and more heterogeneous dataset may make the model better able to generalize from the data set, therefore bringing down the errors that arise due to outlier or atypical cases.

VIII. Interpretation of Results

The results obtained by the experiment present the deep contribution of data augmentation towards generalization and overall improvement for the Vision Transformer model. We trained the Base Vision Transformer on the augmented dataset; then, the improvements in the accuracy and loss metrics are almost impressive, showing how well this model adapts to the presentation of various transformations and variations in the data. The improved diversity from data augmentation techniques like random rotation, flipping, and color jittering have resulted in excellent generalization capability to the unseen data by giving the model an accuracy of 85.06% on the training set and 89.73% on the testing set. Such methods expand the scope of possible cases that come to pass during the experience of the model, making it much more robust to variations in different appearances of skin lesions, which are very common in natural settings.

On the other hand, the Deep Vision Transformer (Deep ViT) model produced relatively less improvement but still evaluated over the augmented dataset. While it also outperformed the non-augmented model in terms of accuracy, its gain in performance was not as dramatic as with the Base ViT model. This shows that Deep ViT architecture with high complexity requires more fine-tuning for it to take in all the increases that result from data augmentation. One such explanation could be that the deeper architecture, with more layers and parameters, may not be able to use data to sufficient advantage- a consequence of its increased vulnerability to overfitting or a need for fine-tuning to better optimize its parameters. As deeper models are usually more sensitive both to training dynamics and the quality of the input data, their potential can only stop increasing if regularization does not provide enough control over the training process or if data augmentation does not present sufficient diversity for them to exploit their greater capacity.

Third, the small gains observed at relatively deeper layers in Deep ViT could suggest that this architecture might be better suited for either large-scale dataset-related tasks or maybe for the more difficult kinds of tasks where sophistication of representations is all the more required. In the case of detection of skin cancers, where lesions with benign and malignant could be very similar, Base ViT might have made a better trade-off between model complexity and its ability to generalize well from the available data.

In summary, the interpretation of the result emphasizes that this data augmentation is one of the important strategies that improves the performance of the model in case it deals with medical datasets with considerable variability. This can be seen to be the case where both Base and Deep Vision Transformers can be augmented well; nevertheless, the better performance of Base ViT attests to efficiency in differentiating varied input changes. The results in this paper reveal that even if significantly much more powerful, deeper models such as Deep ViT require more sophisticated strategies and further optimizations to fully exploit the available capabilities through augmented data, opening up related research directions focused on further advancements, such as including hybrid models or fine-tuning techniques, to eventually exploit all its potential in skin cancer diagnosis.

IX. Conclusion

The paper explores the adaptation of ViTs for skin cancer classification and reveals that fine-tuned ViTs on augmented datasets produce notable performance in the identification of benign versus malignant lesions. In addition, some of the experimental results reveal superior ability in the case of the Base Vision Transformer compared to the traditional CNN-based models, such as VGG16 and ResNet, thereby maintaining similar conditions. The main feature that data augmentation brought was the improvement of generalization capability, where the model resulted in higher levels of accuracy on both the training and testing datasets. More specifically, the model of Base ViT reached a training accuracy of 85.06% and a testing accuracy of 89.73%, making clear how effective it could be in simple classification tasks on samples of skin cancer. In comparison, the non-augmented model showed significantly lower accuracy levels.

It also brought to light the role of data augmentation to enhance the robustness of the model. By artificially increasing the variability within the dataset, the augmented dataset helped in the diversification of learned features by the model and brought about a better generalization of the model. Such augmentations, such as random rotations, color jittering, and other affine transformations, enhanced the ability of the model to perform well with unseen data during testing because the model was allowed to learn to adapt to different real-world conditions.

According to the performance metrics, confusion matrix analysis revealed the accuracy of the Base ViT model in discriminating between benign and malignant skin lesions.

The model was correctly classified 546 test samples out of the total samples of 660. It thus pointed out the effectiveness of the model but with several misclassifications. Error analysis narrowed it down to difficulties of correct identification of subtle changes in skin coloration and the inability of early melanomas to be clearly differentiated from benign lesions. Further refinement in feature extraction is required in order to improve on these errors.

In conclusion, based on the findings from the previous section, Vision Transformers have an excellent opportunity for promoting advancement in skin cancer detection systems: they can capture global relations across any image and as such provide the image with more information concerning the context in it, and by integrating some data augmentation strategies, ViTs are capable of better accuracy and robustness in the medical image classification area. Many such future studies would open the door with this research on optimizations of ViTs and ways to overcome these remaining challenges. For instance, an improved sensitivity of the model to subtle variations in the features of skin lesions would improve the model's predictive accuracy further.

X. Limitations of Current Study

The promising results found here come with several limitations that should be kept in mind when interpreting them. Primarily, there are the size and type of dataset. In this study, the basis was the ISIC Archive, although this contains many data instances it does not provide as exhaustive a view of the large variability seen in real cases of skin cancer. The dataset is heavily biased toward the pictures that represent a more common form of skin lesions and does not include rare and atypical cases, thus posing limitations in the generalization of the model. Even in real applications, skin cancer presentations are different from each other, so further studies need to incorporate more diverse datasets to train these models for wider applicability.

Another major limitation in the present study has been the computation of Vision Transformer models. The deep variants of ViTs also consume more computational resources than traditional CNNs. Training such networks on large datasets typically requires high-performance GPUs and huge amounts of memory; this may not be available for most researchers or institutes. This limits their wide applicability, particularly where computational resources are limited. Although the study used an environment that allowed a GPU, future research would benefit by discussing methods on how to optimize ViT models in such a way that they were made less reliant on computation without impairing performance.

In addition, the quality and quantity of training data are quite sensitive for the ViTs. Despite the augmentation techniques employed to increase the diversity of the datasets, the model performance will be affected by data scarcity. The models tend to perform better given a larger dataset. Therefore, the size of the training set in this research study may have restricted the full extent that could have been reached by the ViT models. This, therefore, means that in future work, increasing the dataset size and diversity will lead to further improvement on the model's robustness, as well as its ability to generalize to unseen data.

Lastly, the paper did not consider a few of the more advanced variants of ViT that would likely provide better performance, specifically hybrid architectures that integrate CNNs with transformers. Hybrid models would potentially get the benefits of both worlds by bringing in the power of local feature extraction from CNNs and the global contextual understanding abilities of transformers. The future work lies in the optimizing and scaling up of the ViT model, experimenting with hybrid architecture, and including larger and more diverse data collections for further improving the precision and practical applicability of ViTs in skin cancer detection.

References

[1] Masood, A., and Al-Jumaily, A. A. (2013). Computer Aided Diagnostic Support System for Skin Cancer: A Review of Techniques and Algorithms. *International Journal of Biomedical Imaging*, 2013, 1–22. doi:10.1155/2013/323268.

[2] Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. (2017). Dermatologist-level Classification of Skin Cancer with Deep Neural Networks. *Nature*, 542(7639), 115–118. doi:10.1038/nature21056.

[3] Dorj, U.-O., Lee, K.-K., Choi, J.-Y., and Lee, M. (2018). The Skin Cancer Classification Using Deep Convolutional Neural Network. *Multimedia Tools and Applications*, 77(8), 9909–9924. doi:10.1007/s11042-018-5714-1.

[4] Han, S. S., Kim, M. S., Lim, W., Park, G. H., Park, I., and Chang, S. E. (2018). Classification of the Clinical Images for Benign and Malignant Cutaneous Tumors Using a Deep Learning Algorithm. *Journal of Investigative Dermatology*, 138(7), 1529–1538. doi:10.1016/j.jid.2018.01.028.

[5] Brinker, T. J., Hekler, A., Utikal, J. S., Grabe, N., Schadendorf, D., Klode, J., et al. (2018). Skin Cancer Classification Using Convolutional Neural Networks: Systematic Review. *Journal of Medical Internet Research*, 20(10), e11936. doi:10.2196/11936.

[6] Demir, A., Yilmaz, F., and Kose, O. (2019). Early Detection of Skin Cancer Using Deep Learning Architectures: ResNet-101 and InceptionV3. In *2019 Medical Technologies Congress (TIPTEKNO)* (pp. 1–4). doi:10.1109/TIPTEKNO.2019.8895097.

[7] Kondaveeti, H. K., and Edupuganti, P. (2020). Skin Cancer Classification Using Transfer Learning. In *2020 IEEE International Conference on Advent Trends in Multidisciplinary Research and Innovation (ICATMRI)* (pp. 1–5). doi:10.1109/ICATMRI51801.2020.9398380.

[8] Sedigh, P., Sadeghian, R., and Masouleh, M. T. (2019). Generating Synthetic Medical Images by Using GAN to Improve CNN Performance in Skin Cancer Classification. In *2019 7th International Conference on Robotics and Mechatronics (ICRoM)* (pp. 482–487). doi:10.1109/ICRoM48714.2019.9071865.

[9] Pacheco, A. G., and Krohling, R. A. (2021). An Attention-Based Mechanism to Combine Images and Metadata in Deep Learning Models Applied to Skin Cancer Classification. *IEEE Journal of Biomedical and Health Informatics*, 25(12), 3554–3563. doi:10.1109/JBHI.2021.3077609.

[10] Rezaoana, N., Hossain, M. S., and Andersson, K. (2020). Detection and Classification of Skin Cancer by Using a Parallel CNN Model. In *2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)* (pp. 1–4). doi:10.1109/WIECON-ECE52138.2020.9397980.

[11] Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z., Tay, F. E. H., Feng, J., and Yan, S. (2021). Tokens-to-Token ViT: Training Vision Transformers from Scratch on ImageNet. *arXiv preprint*, arXiv:2101.11986.

[12] Hekler, A., Utikal, J. S., Enk, A. H., Hauschild, A., Weichenthal, M., Maron, R. C., et al. (2019). Superior Skin Cancer Classification by the Combination of Human and Artificial Intelligence. *European Journal of Cancer*, 120, 114–121. doi:10.1016/j.ejca.2019.07.019.

[13] Ali, K., Shaikh, Z. A., Khan, A. A., and Laghari, A. A. (2022). Multiclass Skin Cancer Classification Using EfficientNets – A First Step Towards Preventing Skin Cancer. *Neuroscience Informatics*, 2(4), 100028. doi:10.1016/j.neuri.2022.100028.

[14] Faghihi, A., Fathollahi, M., and Rajabi, R. (2024). Diagnosis of Skin Cancer Using VGG16 and VGG19 Based Transfer Learning Models. *arXiv preprint*, arXiv:2404.01160.

[15] Himel, G. M. S., Islam, M. M., Al-Aff, K. A., Karim, S. I., and Sikder, M. K. U. (2024). Skin Cancer Segmentation and Classification Using Vision Transformer for Automatic Analysis in Dermatoscopy-based Non-invasive Digital System. *arXiv preprint*, arXiv:2401.04746.

[16] Tai, C. A., Janes, E., Czarnecki, C., and Wong, A. (2023). Double-Condensing Attention Condenser: Leveraging Attention in Deep Learning to Detect Skin Cancer from Skin Lesion Images. *arXiv preprint*, arXiv:2311.11656.

[17] Zhou, D., Kang, B., Jin, X., Yang, L., Lian, X., Jiang, Z., Hou, Q., and Feng, J. (2021). DeepViT: Towards Deeper Vision Transformer. *arXiv preprint*, arXiv:2103.11886.

[18] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. *arXiv preprint*, arXiv:2010.11929.

[19] Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., and Jégou, H. (2021). Training data-efficient image transformers & distillation through attention. *arXiv preprint*, arXiv:2012.12877.

[20] Kunwar, S., and Ferdush, J. (2023). Mapping of Land Use and Land Cover (LULC) using EuroSAT and Transfer Learning. *arXiv preprint*, arXiv:2401.02424.

[21] Hassani, A., Walton, S., Shah, N., Abuduweili, A., Li, J., and Shi, H. (2022). Escaping the Big Data Paradigm with Compact Transformers. *arXiv preprint*, arXiv:2104.05704.