LuminSkin:
Skin Cancer Classification using Vision Transformers

Draft Report

Final Draft

Manav garg
manav.garg@hotmail.com
https://github.com/manavvgarg

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 variabilities 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.

Keywords: Vision Transformers, PyTorch, Skin Image Analysis, Skin Cancer, AI in Medical

CONTENTS

CHAPTER 5: RESULTS AND DISCUSSION

5.1 Training Results on Augmented and Non-Augmented Datasets

5.2 Performance Metrics and Observations

5.3 Interpretation of Results

5.4 Error Analysis

CHAPTER 6: CONCLUSION AND FUTURE WORK

6.1 Summary of Findings

6.2 Limitations of Current Study

REFERENCES

LIST OF FIGURES

LIST OF ACRONYMS

Chapter 1

Introduction

1.1 OVERVIEW OF SKIN CANCER

Skin cancers are one of the most expanding and common cancers all around the world. They are accounted for by millions of cases yearly. This disease encompasses four main types: basal cell carcinoma, squamous cell carcinoma, Merkel cell carcinoma, and melanoma. Both basal and squamous cell carcinomas are less aggressive but must be treated early to stop the tissue damage and their loco-regional spread. Merkel cell carcinoma is very rare but very aggressive. Of course, melanoma is the worst type of skin cancer because it holds an extremely high potential for metastasizing and spreading very fast; especially if left to go undiagnosed or untimely treated. Melanoma begins in the melanocytes, or pigment-producing cells, which can spread through circulation along the lymphatic and vascular channels to other portions of the body, becoming thereby metastatic.

Early detection is the only means by which mortality in melanoma patients can be decreased. Early treatment can radically improve the prognosis, and when the melanoma is diagnosed early, there is an opportunity to treat it, and more than 90% survival in case of localization of the melanoma. So again, with the progression of melanoma, treatment becomes extremely challenging and survival rates come drastically down; therefore, accuracy of diagnosis should also be improved. Though in many clinical cases, benign lesions can at times appear very much like malignant ones, both visually with irregular contours and shading, texture changes, all these challenges pose a question before automated diagnostic systems that help clinicians identify the correct type.

Advances in machine learning and computer vision lend promise to these approaches for diagnosis. Skin lesions might thus be categorized using sophisticated image-based features with the aid of machine learning and deep learning architectures that will transform these analyses into something far less subjective than visual assessment. Such computerized image classification systems based on vast amounts of databases may, therefore, help overcome the issues involved in traditional diagnostic methodologies and lead towards more accurate, fast, and accessible solutions for screening skin cancer. The area continues to advance through various research in this aspect with the aim of trying to improve diagnostic algorithms and ultimately achieve better patients' outcomes in skin cancers management.

1.2 EVOLUTION OF COMPUTER VISION IN MEDICAL DIAGNOSIS

Application to medical diagnosis has been one area where the field has made significant strides over the last several decades. The ability of computers to analyze and interpret visual information automatically has proven useful in allowing health care professionals to make more accurate diagnoses within ever-tightening time constraints.

These included applications related to medical image analysis, mainly X-rays, CT scans, and MRI scans. Researchers started to investigate early in the 1970s and 1980s how computer algorithms could be applied for the detection and classification of abnormalities in those images, such as tumors, fractures, or certain pathologies. Today, the preliminary work forms the basis for CAD, widely applied in various clinical practices.

The advances in imaging technology pushed forward the complexity of computer vision algorithms applied to medical images. The very recent modern techniques in deep learning have revolutionized this area, making computers automatically learn and extract complex features on large sets of medical images. Improvement in the accuracy and speed of diagnosis, in particular, it allows for the identification of subtle patterns that would otherwise go unnoticed by a human observer.

Other than just medical images, computer vision has also been applied for several other applications. For instance, to analyze skin lesions, digital cameras-as used in smartphones-are used by algorithms that detect signs of possible skin cancer. Computer vision is used to analyze retinal images for possible conditions that may lead to blindness such as diabetic retinopathy.

Integration with Other Healthcare Data: It is really exciting to consider how the integration with other healthcare data, such as electronic health records, patient history, and lab test results, will be mediated within this important medical diagnosis domain. By integrating visual data with other clinical information, computer algorithms can make a more integrative, holistic assessment of patient health, which thereby makes treatment plans much more personalized and effective.

Future applications could very well include the further penetration of these technologies into the clinical practice of healthcare professions. This is because computer vision holds the potential to process huge amounts of data, uncover patterns not discernible to human clinicians, and play a critical role in medical diagnosis and patient care in the years to come.

Thus, such systems are not infallible, and their use in medical diagnosis would need proper validation and oversight. Ethical considerations about data privacy and the potential for biases in algorithmic decision-making have to be carefully worked through. As with any form of medical technology, it will remain the realm of the trained healthcare professional, who can expand their own clinical acumen to take advantage of the power offered by computer vision to diagnose and treat.

1.3 OVERVIEW OF VISION TRANSFORMERS

Vision Transformers (ViTs) represent a new class of representation that solves the hidden problems of classical CNNs on image data. While CNNs have achieved considerable success and created a hierarchical structure with localized receptive fields, fixed kernel sizes present constraints on modeling global dependencies across an image--an aspect crucial in applications such as medical imaging. However, as they employ the mechanism of self-attention, they can capture long-range dependencies as well as global context throughout the image, which makes them better at complex visual tasks.

At its heart, it lies in the idea of image patching. Instead of treating an image as some single, single block of pixels, ViTs split it into smaller pieces of fixed sizes, to later be linearly embedded into a sequence of tokens. This allows the model to treat an image as a sequence of discrete units, much like transformers process NLP inputs - sentences as sequences of words. Each patch in the image receives specific positional encoding, thereby preserving information about the spatial relationship of the patch within the overarching sequence. The architecture then passes on the embedded sequence through a transformer architecture that casts multi-head self-attention across all patches; this allows ViTs to weigh relevance for each patch in relation to the other patches. This processing technique allows ViTs to capture complex global relationships within images, hence alleviating the limited contextual awareness that exists within CNNs.

The multi-head self-attention mechanism used in ViTs computes attention scores between each patch and every other patch and identifies crucial spatial relationships that may have been missed by CNN. That is, one self-attention layer generates multiple different attention heads that focus on different aspects of image patches. In doing this, the model captures diverse features at different scales. This characteristic is particularly useful for medical image analysis, were subtle, spatially distributed patterns might be necessary to diagnose properly. For instance, with skin cancer detection, tiny changes in texture, color, and boundary irregularities could be spread throughout an entire lesion. By considering such patches as a group and by capturing relations at the lesion level, ViTs could outperform CNNs that make decisions based on local feature extraction and thus may miss the global patterns.

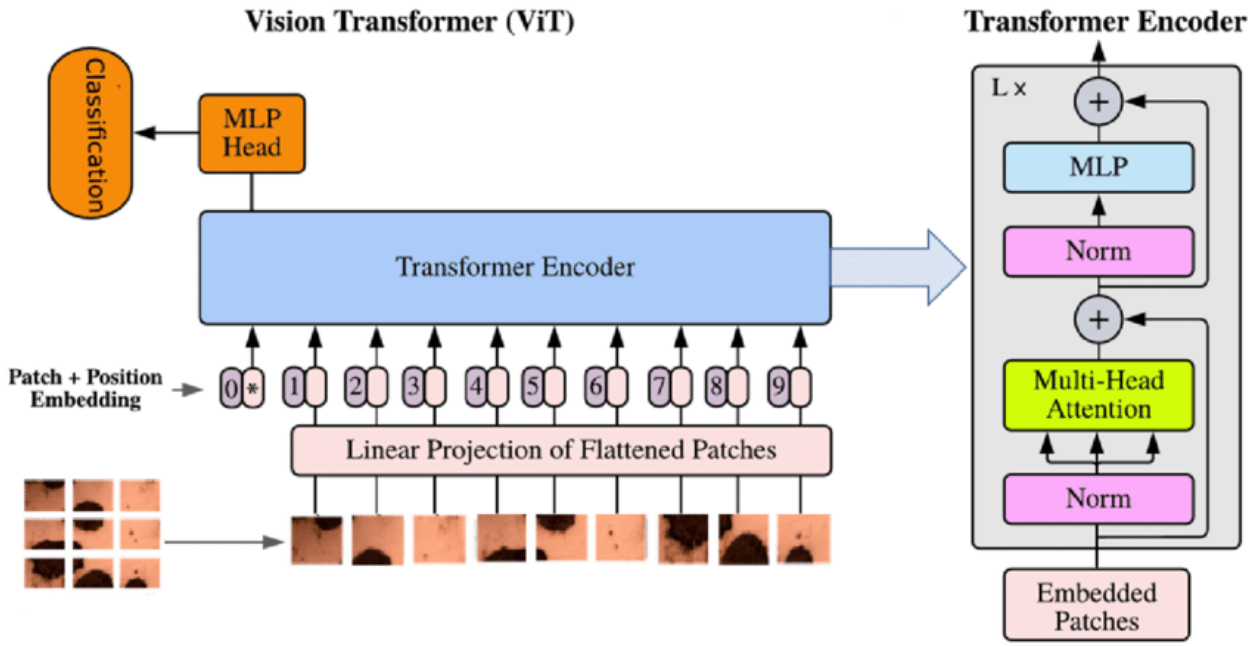


Fig. 1. Vision Transformer Architecture

Currently, Vision Transformers already proved their state-of-the-art performance on multiple standard image-classification benchmarks, including ImageNet, thereby setting up new standards and, in turn, challenging the long dominance of CNNs. Regarding their medical use, ViTs should also open promising avenues toward more diagnostically useful tools, particularly in application

areas like dermatology, radiology, and pathology, where the complete spatial context of an image may be critical. For example, ViTs in dermatology can capture minute features in an image of skin cancer and differentiate with high accuracy as to whether they are benign or malignant lesions. A skin texture that only displays fine-grained details, asymmetry of lesions, and an irregular border is amenable to tasks in which subtle visual distinction is paramount for ViT's applications.

In addition, ViTs are scalable to large datasets without the necessity of feature engineering to be domain-specific; therefore, these models could be used on diverse types of medical image datasets. Since medical images are usually high-resolution scans, the self-attention mechanism has the advantage of capturing extensive dependencies by being able to pick up fine details across large images, like in dermoscopic images, for example, when trying to identify early signs of melanoma or detecting small abnormalities in mammograms. It is demonstrated by researchers that pre-trained models of ViT can be fine-tuned on special datasets, like the ISIC Archive for skin cancer, and achieve high accuracy using just relatively small additional training.

However, there are some challenges with Vision Transformers. ViTs, while having their benefits, are typically data-intensive, requiring an often-large dataset to be effectively trained, due to inductive biases like translation invariance built into the CNN structure. This requirement for high amounts of data is a potential drawback in medical usage where annotated images could be scarce. The self-attention mechanism also has a high computational cost because the mechanism quadratically scales with the number of patches, and therefore, the computation cost of ViTs tends to be more computationally intensive than traditional CNNs. Researchers are aggressively pursuing hybrid models that inject the best of CNNs and ViTs, such as CvTs: a model that extends the transformer structure using convolutional layers to reduce redundancy and their data requirements.

Vision Transformers present an entirely new framework for computer vision, overcoming some of the inherent limitations of CNNs in modeling global dependencies on images. Their distinctive self-attention mechanism, patch-based processing, and ability for more fine-grained detail analysis make ViTs highly promising for medical image analysis. The potential of enhancing diagnostic accuracy, particularly in skin cancer detection, makes this even more significant with increased computational resources and large annotated medical datasets accessible.

1.4 MOTIVATION FOR VISION TRANSFORMERS

The motivation behind the advent of Vision Transformers and subsequent application in image classification, particularly in cases of medical diagnosis, has been from the limitations presented by conventional Convolutional Neural Networks in tasks related to image analysis, thus not being well suited for some heavy-weighted complex image analysis tasks. Although CNNs have immensely played a crucial role in computer vision, localized receptive fields inherent in their architecture bar them from capturing long-range dependencies and global context within an image; this proves a serious constraint in the medical fields, where diagnoses often rely on

recognizing subtle patterns scattered throughout the entire image. For instance, in the tasks of diagnosing skin cancer, to determine whether a lesion is benign or malignant, subtle changes of texture, color, and structure must be detected spread across and around that particular region. Such global characteristics are missed by CNNs, with all their power, due to the localized task they were trained on for pattern identification.

Majorly, Vision Transformers derive their transformative capabilities from their use of self-attention mechanisms that allow modeling all-to-all relationships between parts of an image. This approach turns out to be very valuable for tasks that require having a holistic view of an image, thus allowing ViTs to weigh every part of an image relative to others even at significant distances. Such ability is critical in medical imaging, where the spatial context of regions can be very important for finding disease-specific patterns, such as asymmetric and jaggy boundaries of melanoma or subtle variations within radiographic images. ViTs are a means by which one can learn such global dependencies without careful handcrafted features and yield better precision and robustness in diagnosis.

A related motivation arises by generalizing Vision Transformers across various domains for medical applications and imaging modalities. That is to say that ViTs could be pre-trained upon huge general datasets and then fine-tuned on domain-specific medical datasets relying much less on large, expensive annotated medical data. Their flexibility makes them one of the various solutions to data scarcity something health care does not have an abundance of. By their ability to be pre-trained on big datasets and fine-tuned to specific medical contexts with relatively little additional training, ViTs can be applied to a wider range of applications in radiology, dermatology, and other specialty domains.

Finally, clinicians are soon going to face even higher and more detailed resolutions in imaging technologies requiring precision along with computational efficiency in their automated analysis. ViTs are particularly well-suited to these high-dimensional tasks since the self-attention mechanism scales naturally with image size and thereby allows for rich analysis of even large, high-resolution images. The derivation of hybrid models that combine CNNs and transformers further speaks to the motivation to draw on the strengths of both architectures to develop models that perform efficiently with global awareness in complex data-intensive applications of healthcare. Vision Transformers are, therefore, a promising area of research that can be leveraged for improving automated diagnostic tools to support healthcare professionals in assessing timeliness and accuracy for better patient outcomes in areas where early detection is of paramount interest.

1.5 CHALLENGES

Though ViTs have achieved great success in a broad array of image classification benchmarks, their application to specific domains such as medical image analysis, for instance, skin cancer detection, faces several hurdles. One main hurdle is data dependency. For the generalization to

depend heavily on large datasets, ViTs do not work well without them since, in contrast to CNNs, ViTs are deficient in some innate inductive biases, translation invariance in particular. In medical image techniques, large-sized annotated datasets are difficult to obtain and costly, especially in rare diseases or highly specialized fields such as dermatology where expert annotations require a significant amount of money. Therefore, lesser availability of labelled data for ViTs may hamper generalization and robustness in classification tasks of skin cancer.

Another is computational complexity: the self-attention mechanism of ViTs, while globally contextualizing the model, quadratically scales with the number of patches. In high-resolution medical images, this can result in rather heavy computational requirements, and therefore it is very challenging to use such models in real-time or resource-constrained environments like clinical settings, an inhibitor to scalability, particularly for high-dimensional data as typical with medical imaging.

Another challenge is interpretability and explainability. In contrast to CNNs with clearly defined filters that correspond to visually interpretable features, self-attention in ViTs behaves in a tougher-to-interpret manner. Since it's essential for applications like skin cancer detection-the application of transparent and understandable model outputs for proper clinical decisions-the lack of interpretability may, therefore, impede trust and clinical adoption. Knowing which precise regions or features of an image influenced the classification made by a ViT is crucial for medical experts, even if attention weights themselves do not always offer clear insight into why certain classifications are made by ViTs.

Overfitting is also a problem, especially when fine-tuning ViTs on small, specialized datasets. With high capacity, ViTs suffer from noise in smaller datasets, impairing generalization ability over other data. Overfitting often requires sophisticated techniques at regularization or a hybrid model approach-a combination of CNNs and transformers to use the strengths of each while reducing vulnerability to biases in small dataset sizes.

Further research is done toward the minimization of requirements for data, improvement in interpretation, and optimization of computationally expensive procedures in order to make Vision Transformers of practical utility in skin cancer detection as well as in other applications involving medical imaging.

1.6 PROJECT STATEMENT

The central issue addressed by this research project is the design for the very first time of an efficient and reliable Vision Transformer (ViT)-based model for skin cancer detection, focusing on distinguishable positivity-segregation ACSD regarding benign and malignant skin lesions. Skin cancer, being the most common and among the deadliest tumors globally, requires speedy and accurate clinical diagnostic sources by which early detection can be improved for therapy and prognosis. Traditional diagnostic techniques require a large consideration to visualize lesions, which incurs a high demand for the use of advanced computer vision techniques in dermatological diagnostics.

The contribution is directed toward utilizing a Vision Transformer for overly catching global dependencies and fine-grained features over the full dimension of images to address drawbacks associated with Convolutional Neural Networks (CNNs) when addressing complex patterns associated with skin cancers. With a very precise classification of different forms of skin lesions--especially melanoma--the ViT model will be fine-tuned within the ambit of the ISIC Archive-a well-annotated clinical skin dataset-in tandem with emphasis on high classification accuracy while keeping false positives at minimal.

Data augmentation, efficient training regimes, and visualization technology will be used to allow ViTs to work with a famished data training set; furthermore, the interpretability is what the project will address. In essence, this project aims at creating an automated, scalable, yet interpretable diagnostic tool for skin cancer detection that can benefit dermatologists making timely and superior diagnoses.

1.7 OBJECTIVES

- Develop a particularly fine-tuned skin-cancer-detection ViT model, focusing on high accuracy in distinguishing between benign and malignant skin lesions.
- Leverages self-attention mechanisms in the ViTs for skin lesions to capture global dependencies and subtle patterns and thereby improve diagnostic precision beyond the traditional CNN-based methods.
- Train and test it using a clinically annotated skin cancer dataset, such as the ISIC Archive, so that it delivers clinically relevant accuracy across types of skin lesions, including melanoma.
- Use data augmentation techniques, since medical datasets are scarce and have limited sizes. This should be utilized to make the model more robust and reduce potential side effects of overfitting.
- Optimize the ViT model's computational efficiency such that it would scale to afford real-time applications in clinical environments.
- Ease interpretation of the model through availability of visualizations such as the attention map visualization of the areas in images that caused a prediction by the model, and thus ease clinician trust and transparency.
- Measure the performance of the model with metrics such as sensitivity, specificity, accuracy, and AUC for high reliability in diagnostics.

Chapter 2

Literature Review

2.1 CNN – BASED APPROACHES FOR SKIN CANCER DETECTION

Convolutional Neural Networks CNNs have been the pillar of many studies concerning skin cancer detection because such networks are capable of automatically abstracting hierarchies of features from images. Popular architectures for CNN use techniques called VGG16, VGG19, and ResNet, with these architectures successfully being used to classify and segment skin lesions. These networks have layers of convolutions designed to pick up visual features at multiple levels of complexity: from edges and textures in the more superficial layers to shapes and patterns at deeper layers. CNNs' deep structure have made possible the localization and recognition of lesions based on patterns not within human's naked eyes. Such is the direct consequence of this property that it has improved the accuracy for diagnosis.

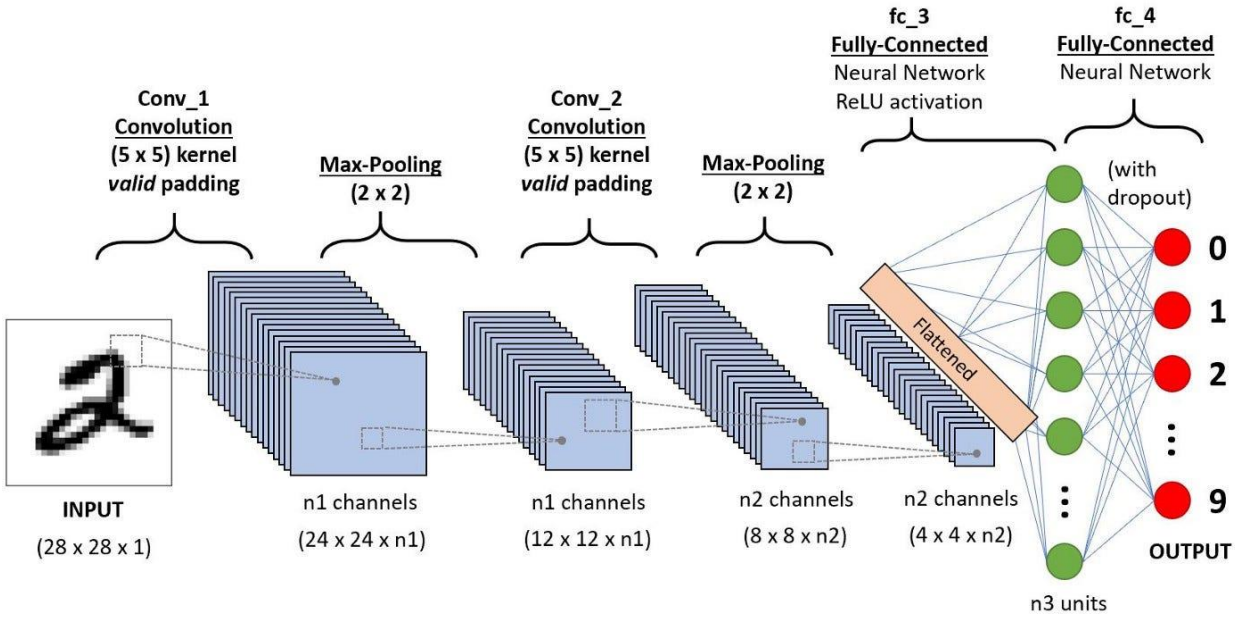


Fig 2. Popular CNN architecture

A very common strategy adopted in the process of skin cancer studies involving CNNs is transfer learning. Transfer learning refers to the leveraging of pre-trained CNNs that have first been trained on big data sources, such as ImageNet, and fine-tuning these CNNs on smaller, domain-specific datasets, such as those comprising images of skin cancer lesions. Fine-tuning of pre-trained CNNs does improve performance in cases of limited data; this is because the model could have transferred general features learned from the larger dataset to be adapted towards identifying characteristics specific to skin cancer. Research applied transfer learning with VGG

architectures, such as VGG16 and VGG19, and reported accurately classifying skin lesions when fine-tuned on dedicated skin-cancer datasets. One of the studies used more than 98% accuracy using VGG16 and VGG19 models with custom-made designed layers to suit the special characteristics of skin lesions. That is, this just highlights the superiority of transfer learning in enhancing classification results.

Even though the CNN-based methods have been successful, they remain existing with a plethora of limitations: for example, dependency on data and lack of feature extraction. CNNs consume a huge amount of labeled training data to achieve high accuracy without overfitting. Achieving large, annotated datasets is costly and time-consuming in medical applications. Moreover, CNNs use localized receptive fields in the convolutional layers, meaning that they might not well capture long-range dependencies and context in the image. This is particularly crucial for skin cancer detection, where small and distributed features, such as asymmetry and aberrant borders, may occupy larger parts of an image and demand global context for detection. Although deeper CNNs, such as ResNet, alleviated this problem by incorporating residual connections to preserve more context at different levels, they yet do not completely overcome the challenge of capturing these subtle, high-level features in skin cancer images.

However, in efforts to counter these challenges, researchers have also investigated ensemble approaches where multiples are merged together in an effort to enhance robustness and lower error. Ensembles may also be helpful to capture the diversity of features and maybe more reliable classifications, but these come at a cost in increased computational complexity and may not even capture the global context required for more complex diagnoses. Although CNNs form an excellent foundation for the automated detection of skin cancer, their limitations form a critical underpinning for alternative approaches such as Vision Transformers that offer better capability to model global dependencies and complex spatial relationships within images.

2.2 EMERGENCE OF VISION TRANSFORMERS IN IMAGE ANALYSIS

The newly conceived computer vision advancement, known as Vision Transformers (ViTs), promotes the transformer architecture, fundamentally developed for natural language processing originally, to compete within the scope of image analysis. This concept was introduced by Dosovitskiy et al. in their groundbreaking paper from 2020, demonstrating how, in many classification tasks, ViTs seem to be capable of performing at a level surpassing that of traditional CNNs, especially when it is the case of training on large-scale datasets like ImageNet. It is distinct because of its structure and mechanism centered around the attention principle, which helps in distinguishing it from CNNs and offers superiority in tasks demanding overall spatial as well as contextual understanding.

Contrary to the function-based dependencies of CNNs on local receptive fields and hierarchical extraction of feature information, the Vision Transformer processes images as a sequence of non-overlapping patches. Each image is segmented into patches; therefore, these are linearly embeddable where the patching of an image is considered as tokens for a sequence just like tokens from words in the sentence for some NLP task. The structure also allows ViTs to be applied to self-attention mechanisms over all patches, so that the model can learn relationships between pixels spread throughout the image rather than being limited to local neighborhoods.

This would enable ViTs to capture the global context much better than CNNs, especially in medical images where the strong, space-distributed features often contain essential diagnostic information.

The global dependence capturing capability of ViTs has deep significance in the area of medical image analysis. Differences here could be minor, such as irregular borders, asymmetry, and texture variations, and thus are crucial to detect in benign versus malignant lesions. In such applications of skin cancer detection, global attention within ViTs can drastically improve the classification accuracy. Their attention on the global paradigm of vision allows them to integrate relatively dispersed visual cues, which may result in a more holistic structure in the appearance of the lesion and reduce the chance of misclassification based on isolated, localized features. This is more so especially in dermatology and other fields that fine granular visual analysis is necessary for proper diagnosis.

Ever since the advent of Vision Transformers, a lot of work has been done to adapt the model and fine-tune its architecture for various medical image analysis applications. There are various adaptations - hybrid models combining ViTs with CNN layers, aiming to fully exploit both the global attention and localized feature extraction capabilities. These hybrids have given promising results in improving performance across tasks that require a balance of global and local contexts. Besides that, other variants of ViT such as Swin Transformers and DeiT (Data-efficient Image Transformers) have been brought forth to mitigate some data requirements and computational overheads while promising high accuracy on image classification tasks.

This potential in medical image analysis extends the ViTs from the confines of dermatology to radiology and pathology also, where similar subtle visual cues and spatial context make it particularly challenging. Studies have therefore shown the capability of ViTs to exceed or at least match the performance of CNNs in classifying complex medical images, thus marking a step forward in automated diagnostics. Capturing complete image context efficiently, Vision Transformers appear to be a promising alternative to CNNs with promises of progress toward greater accuracy, interpretability, and clinical applicability in a range of fields within medicine.

2.3 RECENT ADVANCEMENTS AND CHALLENGES

Although ViTs have shown very good promise on image classification and medical imaging applications, significant challenges are associated with their application, mainly concerning data requirements, computational demands, and model interpretability. A typical ViT depends on huge amounts of labeled training data to function best. The reason for such reliance on extensive data is due to the fact that the model has a high number of parameters and is without built-in inductive biases, such as locality and translational invariance, which CNNs have. As a result, ViTs are prone to overfitting when used with smaller datasets, a common limitation in medical image analysis that involves expensive and time-consuming data collection and annotation.

These limitations motivate several adaptations that make ViTs more efficient on domains with few observations, such as medical imaging. One distinct innovation is the Tokens-to-Token Vision Transformer (T2T-ViT), an improvement on tokenization via progressive aggregation of neighboring patches into tokens, which reduces the number of tokens the model must process.

This aggregation not only reduces the computational overhead but also enhances feature representations by incorporating the local context into any token. T2T-ViT has also proven effectiveness in reducing the count of parameters and efficiency, hence more feasible for use with smaller datasets.

Another more recent technique to optimize ViTs on datasets is the Re-attention method. This method enhances the diversity of attention maps that are derived through the self-attention process, enabling the model to learn a more valuable representation of the input data without significantly increasing the computational cost of doing so. Producing diverse attention patterns, Re-attention can help better capture local and global dependencies and improve the performance of ViT in low-data regimes. Other techniques include data augmentation and self-supervised learning, which can further enhance the generalizability of a model and allow for the application of ViTs to smaller datasets for learning robust, transferable features from uninformed data.

Despite the above techniques and methods that show much promise, there are several challenges still present and particularly within medical imaging applications. For instance, interpretability is highly necessary in healthcare applications since clinicians must comprehend and rely on the decisions made by these models. However, because complex attention mechanisms are part of ViTs, inner workings become difficult to interpret. Tools for visualization such as attention map visualization and saliency maps are, therefore, being integrated to provide insight into where an image is affecting the predictions of the model, thus making the model more transparent for clinical use.

Conversely, ViTs are computationally expensive and consume much more memory and processing resources than a typical CNN, which makes them not that accessible in some resource-limited deployment scenarios. To this end, researchers are looking at more lightweight variants of transformers, such as DeiT - Data-efficient Image Transformers - and hybrid approaches that combine CNN with transformers, trying to strike the balance of the performance-related efficiency trade-off.

These advancements mark considerable progress toward bringing ViTs into real-life application in medical imaging, but more research is needed to achieve optimized functionality of ViTs on small data sizes, increase interpretability, and decrease computational cost. This is a continuing line of research adapting ViTs for specific health-care applications, where they will considerably enhance diagnostic capability.

Chapter 3

Architecture Overview

3.1 BASE VISION TRANSFORMER ARCHITECTURE

The ViT architecture thus provides a powerful alternative to CNNs in the task of image classification. Being based on CNNs where this architecture captures spatial relationships locally, the transformer architecture may be exploited to create global dependencies across an entire image, thus enabling ViTs to identify intricate relations among pixels over long distances-advantageous for medical imaging and others requiring fine-grained analysis.

1. Image Patching and Tokenization: In the ViT architecture, the first stage is to divide an input image into a sequence of smaller non-overlapping patches, in terms of pixels. For example, an input image of size 256 x 256 pixels can be divided into 16 x 16 patches, in other words, each of size 16 x 16 pixels, which, as a result, leads to a sequence of 256 patches. As soon as flattened, each patch gets turned into a 1D vector. This is done by transforming the patch into a sequence of linearly arranged pixel values. The patch vectors are then forwarded into a learnable linear projection layer that transforms each patch into an embedding dimension, for instance, into 1024 dimensions acting like tokens within the transformer model.

2. Class Token and Position Embeddings: One of the quite distinctive features of the ViT architecture is that there is a particular learnable embedding introduced as the class token. It added at the beginning of the sequence of patch embeddings, and it essentially plays the role of a global representation of the image and is eventually used for classification. Also, each patch embedding in ViT includes adding position embeddings to them. Since transformers are inherently position-agnostic, these position embeddings encode the spatial location of each patch within the image, which then helps the model understand the spatial structure of the input.

3. Transformer Encoder Blocks: After being tokenized, the patch embeddings, accompanied by the class token, are passed into the transformer encoder. The transformer architecture consists of two main parts in each of its encoder blocks:

- **Multi-Head Self-Attention (MHSA):** MHSA enables the model to learn the relationships among different patches by assigning attention weights, which allows it to focus on important patches across the image. The term "multi-head" indicates that a model contains multiple attention heads that can simultaneously look at different aspects of the image, capturing various interactions of features. For instance, a 16-head ViT model can process 16 different relationships in parallel, thereby taking a global view that is comprehensive.

- **Feed-Forward Neural Network (FFN):** This is another position-wise feed-forward network in each transformer block after the attention mechanism. Each token undergoes linear transformations along a series, which promotes nonlinear patterns by the model

and obtains stronger representations of advanced features from the data.

There are two modules: MHSA and FFN. The components are repeated across multiple layers or transformer blocks with residual connections and layer normalization for stability during training and easier gradient flow.

4. Class Token Aggregation: From this processed sequence of patch embeddings, by feeding it to the transformer encoder, the transformer encoder output corresponding to the class token is extracted. Since the class token has acquired all the information coming from all the patches of the image due to the self-attention mechanism, it serves as the global representation of the whole image.

5. Classification Head: The class token is passed to the classification head. The classification head is a fully connected layer. Therefore, the classification head maps the representation of that class token to the output logits-the probability scores for every one of the classes involved in classification. If there is skin cancer classification involved, such as benign versus malignant, the classification head usually uses a SoftMax layer that provides normalized class probabilities.

6. Training and Optimization: The ViT model is trained with a suitable loss function for the considered classification task, such as cross entropy loss in binary or multiclass classification and uses optimization algorithms like Adam with techniques like learning rate scheduling and regularization to stabilize and optimize the training process.

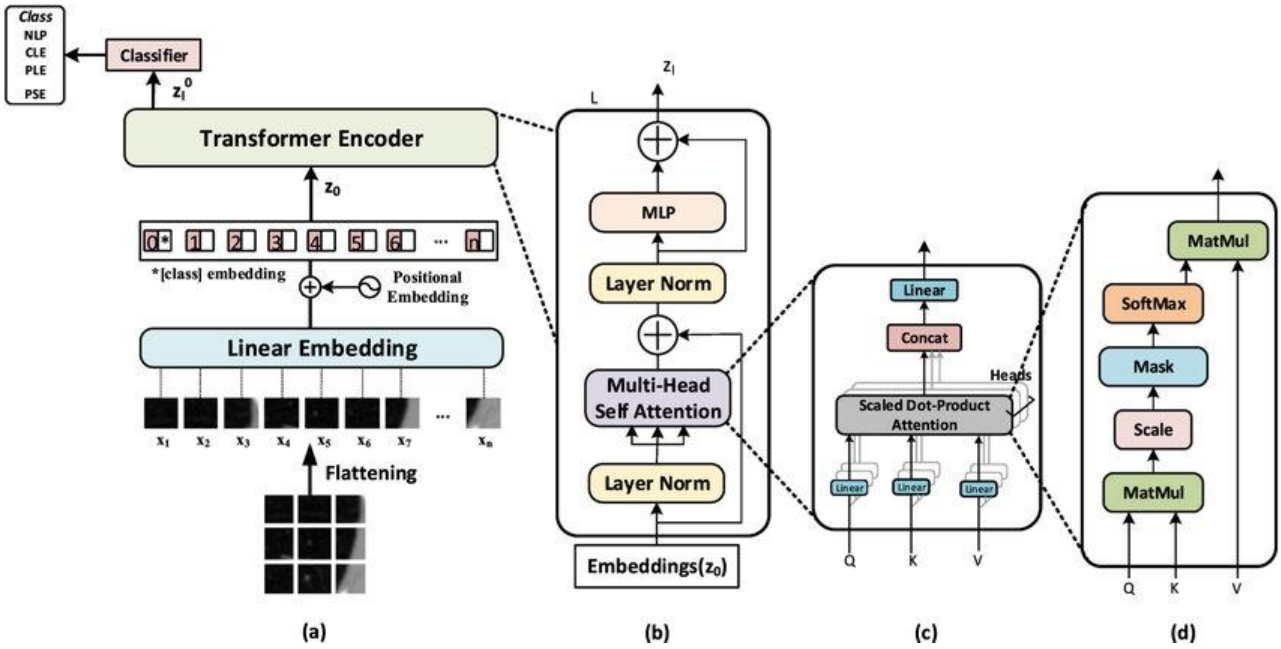


Fig. 3. The vision transformer architecture. (a) The main architecture of the ViT; (b) transformer encoder module; (c) the multi-head self-attention block; (d) the self-attention head.

3.2 DEEP VISION TRANSFORMER ARCHITECTURE

Deep Vision Transformer stands for the model developed as an extension to the base model, the ViT, where the deep vision transformer gains an architecture enhancement of the basic one, the so-called standard model, from some natural inherited limitations within the Vision Transformer. DeepViT tries to optimize self-attention mechanism and transformer blocks depth so that it can exploit higher quality attention maps, making the model more suitable for a variety of data-intensive applications, for example, medical image analysis, along with other needs for detailed feature learning.

1. Input Image Patching and Tokenization: DeepViT, like the standard ViT, divides the input image into a grid of patches of fixed sizes. Then, each of these is flattened and passed through a linear projection layer in order to be transmuted into an embedding. This process allows the model to treat every patch as a token in some sequence, so you get a list of tokens that will then be processed with the transformer architecture. Take an example: take an image as input to your model, say 256x256 in size. You divide it into 32x32 patches, so you get the whole image represented in a sequence of tokens.

2. Class Token and Positional Encoding: This class token and position embedding DeepViT uses, which is able to preserve the structure and global representation of the image. At training time, this class token is learned and appended as a list of patch tokens where it can aggregate information across all patches. Positional encoding is then added to each token to offer spatial context to the model, so that it can understand the relative positions of patches-a thing which is typically important for tasks in images.

3. Self-Attention Enhanced with Re-Attention DeepViT: It suggests an optimization variant of attention as referred to as Re-Attention and focuses directly on the self-attention layers at deeper levels of transformer architectures.

- **Re-Attention Mechanism:** In the conventional ViT, each self-attention layer computes an attention map, whose output explains the relationships among patches. For deeper models, such attention maps could be noisy and lack insight. Re-Attention progressively improves attention maps across layers. By repeatedly computing multiple levels of attention maps and nudging each layer towards producing clearer attention maps, DeepViT reduces redundancy and improves overall quality in feature representations.

- **Multi-Head Attention (MHA):** Similar to ViT, in DeepViT, the multi-head attention is also used where separate heads are focusing on different aspects of the relation across patches. Re-Attention enables every head to look at an extremely diverse range of features that significantly increase the ability of the model to identify both local and global patterns, which is very beneficial in complex tasks.

4. Deeper Transformer Blocks with Layer Stacking: To make the best use of Re-Attention, DeepViT is built with even more transformer blocks than it would be in a typical ViT model. Each transformer block has two major components:

- **Layer Normalization and Multi-Head Self-Attention:** In the multi-head self-attention layer, it takes the sequence of patch tokens (with class token included) for processing. It applies layer normalization before processing with this layer, which really stabilizes training as well as improves gradient flow, especially in deeper architectures.

- **Feed-Forward Network (FFN):** Following each self-attention layer, the tokens are fed into a position-wise feed-forward network that adds nonlinear transformations to assist the model in its capability to represent complex patterns.

With more extensive depth in feature extraction created by a deeper stack of transformer blocks, DeepViT creates views that are deeply rich for visions at multiple levels; this allows for capturing complex information across the scales.

5. Class Token Aggregation and Classification Head: Once the patch tokens pass through all of the transformer blocks, the class token would have accumulated information from all parts of the sequence. This would be left at the end with an aggregate representation of all the features in the image. A fully connected classification head then maps the output of the class token to final predictions that represent the model's confidence over each class.

6. Training and Optimization: Appropriate loss function is employed, such as cross-entropy loss for classification, with an optimization method like Adam or AdamW and associated learning rate scheduling. DeepViT suffers from the overfitting problem because of its deep architecture, and careful learning rate scheduling along with regularization techniques should be implemented.

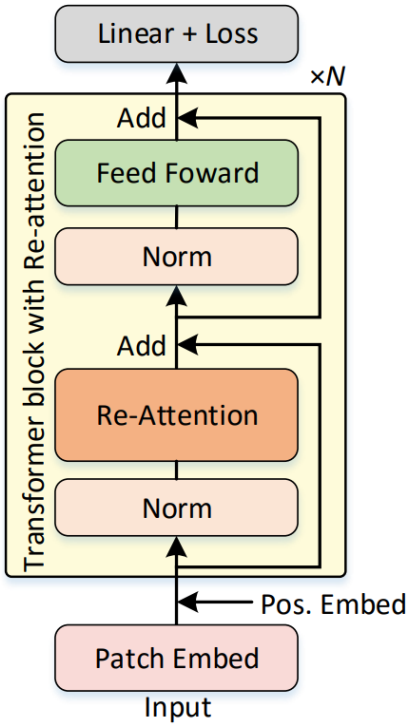


Fig. 4. Transformer Block in Deep Vision Transformers

3.3 Architectural Differences Between Base Vision Transformer and Deep Vision Transformer

1. **Self-Attention Mechanism:**

   - **Base ViT:** Employing basic self-attention, that is token-wise with sparse refinement across layers, thereby prone to noise for deeper networks.

   - **DeepViT:** Utilizes a Re-Attention mechanism which refines attention maps across layers such that clarity is enhanced as well as without redundancy in its relationship with respect to tokens, thus improving the possibility of token relationships for deeper architectures.

2. **Model Depth:**

   - **Base ViT:** Generally designed to be relatively shallow (6–12 layers), sufficient for pretty fine-grained structures.

   - **DeepViT:** With a very deeper structure, stacking 12+ transformer blocks to capture more complex, multi-scale features and especially useful for high-resolution or intricate datasets.

3. **Attention Map Quality:**

   - **Base ViT:** Attention maps degrade as depth increases with a few limitations on the model's ability to keep many clear feature relationships in complex images.

   - **DeepViT:** Re-Attention refines the attention maps layer by layer, improving feature extraction while keeping clear spatial relations in attentions across deeper layers.

4. **Tokenization and Embedding:**

   - **Base ViT:** The model uses simple tokenization by dividing images into fixed-size patches with encoding without any further optimization.

   - **DeepViT:** Deep ViT uses optimized tokenization along with strong positional encodings that don't have the loss of spatial information for high-resolution inputs.

5. **Positional Encoding and Class Token:**

   - **Base ViT:** Positional encoding is added to tokens at the very start but not reinforced in any higher layers, which may result in less spatial clarity for deep networks.

- **DeepViT:** Positional encoding is often encouraged to be strengthened over layers by which the spatial structures are much more preserved in deeper models, especially multi-layered architecture.

6. **Diversity of Multi-Head Attention:**

- **Base ViT:** Normalized multi-head self-attention in which each head could only pick a few different patterns among patches

- **DeepViT:** Re-Attention further diversifies multi-head attention in which every head identifies more inherently different and contextually diverse relationships, and therefore such models come up with more diverse feature representations.

7. **Feature Representation and Learning Capacity:**

- **Base ViT:** Good for global features but can miss subtle distinctions as it is shallower in fewer transformer layers.

- **DeepViT:** It enhances depth and re-attention strength to the capacity of the model to learn hierarchical features, making it better at picking out complex patterns and minute details.

8. **Computational Complexity**

- **Base ViT:** Less computational cost; Very much suited for general-purpose tasks and can be managed with average resources.

- **DeepViT:** Deeper structure and further refinements in attention make this computationally expensive in terms of memory as well as computational power.

9. **High-Resolution Images Performance:**

- **Base ViT:** It is likely to be great on standard datasets but does not have depth or refinement in attention sufficient to perform any of the detailed medical or high-resolution tasks.

- **DeepViT:** That is explicitly designed to cater for high-resolution and intricate images. For making fast precise feature extraction for tasks involving medical imaging, it is the best suited model.

10. **Generalization and Overfitting**

- **Base ViT:** More susceptible to overfitting as depth is shallow. This can lead to potential under-generalization for complex images.

- **DeepViT:** More depth with properly tuned attention helps in reducing overfitting and increases generalization when the depth is high, and multiple layers help to capture complex features.

11. Suitability for Applications:

- **Base ViT:** Use in general classification of images if it does not demand detailed feature extraction in higher resolution.

- **Deep ViT:** DeepViT will be used on specialized applications such as medical image analysis where subtle feature details are crucial and clarity in space will be required for the classification to be accurate.

Chapter 4

Methodology

4.1 DATASET DESCRIPTION

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.

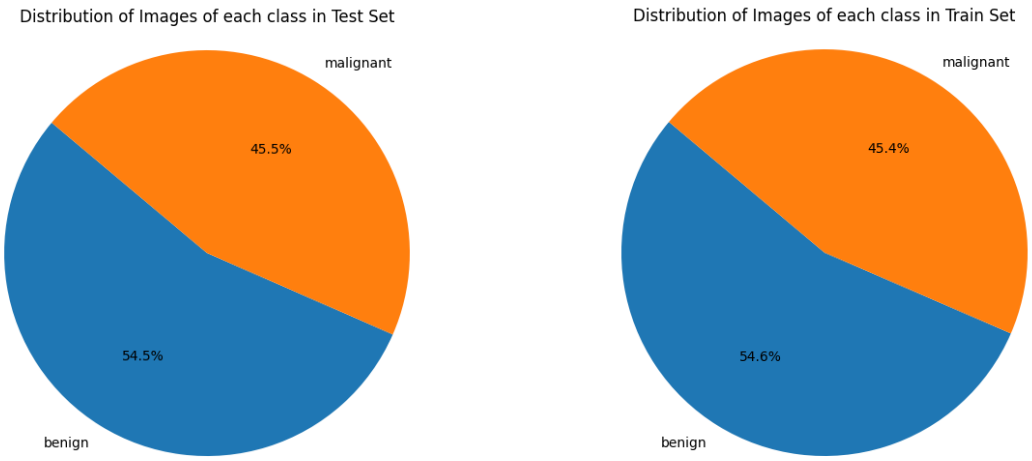


Fig. 5. Distribution of Images before Data Augmentation, Test set [LEFT] and Train set [RIGHT]

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.

4.2 DATA PREPROCESSING TECHNIQUES

Data preprocessing, therefore, forms an important step for the optimization of a dataset in view of the training of machine learning models, especially ViTs, since they require input based on requirements. In this study, a number of data preprocessing techniques were done to standardize and prepare the ISIC Archive dataset for training, which improves model performance and generalizability.

First, all images need to be resized to a fixed resolution: 256x256 pixels. Along with standardizing the dimensions of the images, this resizes also align well with the input requirement of most pre-trained ViTs as they tend to be optimized for an input size of 256x256. Resizing therefore ensures that images end up with the same dimensions and hence makes training much more efficient as well as reducing computational complexity.

Next normalization of all images was performed to equate their pixel intensity. Normalization is achieved by applying the images to the commonly used mean value, which is [0.485, 0.456, 0.406] along with a standard deviation of [0.229, 0.224, 0.225] since they are pre-trained models on ImageNet. Each channel, namely, Red, Green, and Blue will be scaled to a comparable scale, thereby reducing the variance in the dataset, thus helping to improve the convergence of the models. It helps align with the statistics of ImageNet. This means that such normalization is one of the features that enables the use of transfer learning, as it is a good adaptation of layers pre-trained on the model for the specific skin cancer dataset.

Other data augmentation techniques applied to the training set included random rotations, horizontal flips, and zoom adjustments to increase variability in the dataset. These augmentations make the model robust enough to withstand slight variations of the same image to achieve more authentic real-world conditions. This would help avoid overfitting, since in medical datasets, we deal with rather small datasets, the model will be a bit better at generalizing on new data.

Therefore, the dataset is ready for supervised learning by numerical encoding of class labels. A number is assigned to each lesion category. This can be either benign or malignant. In this way, categorical data will be encoded to a numerical format, which will be processed by the model for the classification task. The reason is that it will allow the model to distinguish classes during training and in making the right predictions.

This dataset, after preprocessing, is divided into train and test sets with a ratio of 80:20. It means 80% images for training purposes, and 20% are retained for testing the model's performance on unseen examples. This provides a trade-off between having enough information to train properly and yet having enough data to test the generality of the model. With an 80-20 split, machine learning often employs it for reasons of balance between the sufficiency of training data and availability of test data.

4.3 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 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.

- **Random Horizontal Flip:** This looks to flip images horizontally with a probability of 0.5. It introduces variety in the dataset so that features can easily be identified by the model whether they are upward or downward-oriented, especially in cases where lesion patterns appear symmetrically.

- **Random Rotation (±30°):** The images are randomly rotated by ±30° so that the model can be invariant to orientation. Skin lesions will come in different orientations and at times may not be well aligned so that this rotation would ensure the feature extraction for patterns would occur despite the variations in orientation.

- **Color Jittering:** Applying random brightness, contrast, saturation, and hue alterations of the images to simulate photos taken at various lighting conditions. This will help make the model less sensitive to lighting differences and skin color differences that are typically dramatic in clinical settings. Thus, the model pays more attention to structural and textural characteristics than to color-related characteristics of the lesion.

- **Randomly resized crops (Scale 0.8-1.0):** This augmentation randomly crops images and scales between 0.8 and 1.0 to vary the features scale. This will make the model more generalizable to images of all sizes, reducing overfitting in lesion recognition at a close-up or distant view.

- **The Random Affine and Perspective Transformations:** Affine transformations (such

as translation, scaling, rotation, etc.) and change of perspective simulate changes in scale and viewpoint. Such transformations make the model much more robust against slight positional and shape variations, thus further increasing its ability to generalize. The slight variance in imaging angles does not make it seem overly sensitive to certain orientations or placements within the frame.

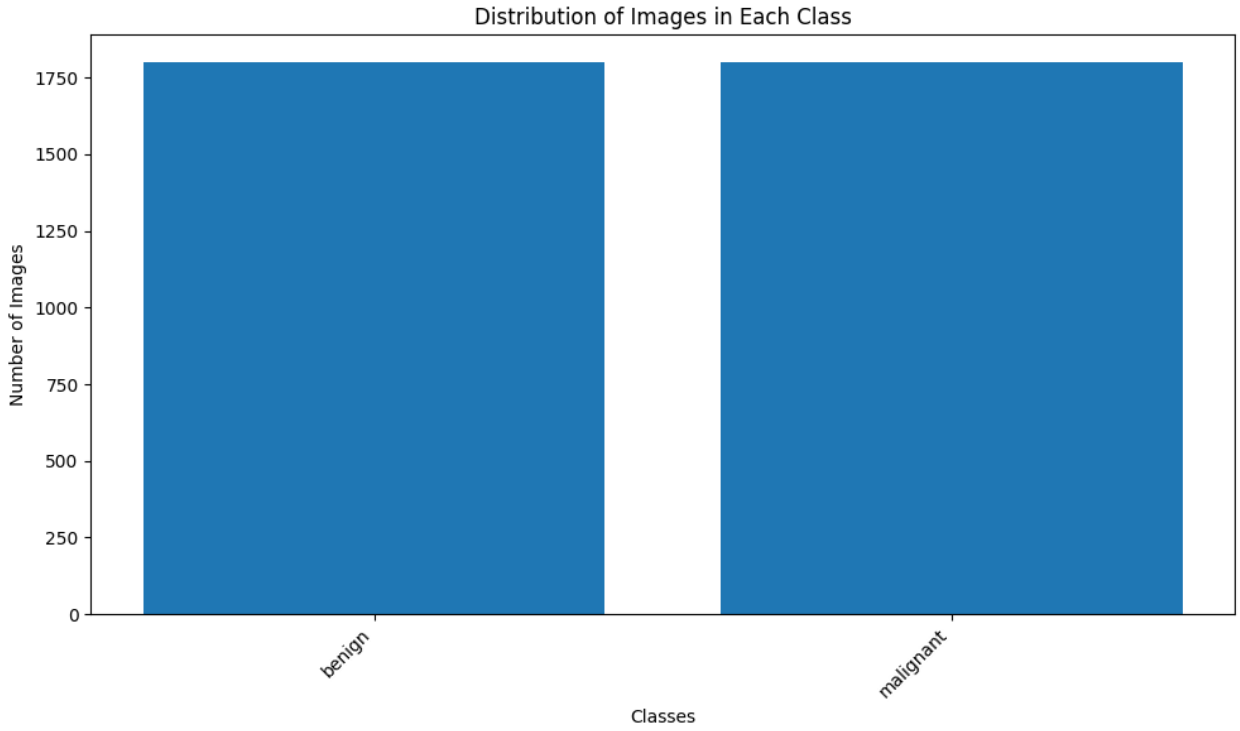


Fig. 6. Distribution of Images after Data Augmentation

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.

Fig. 7(a). Visualizing samples from "BENIGN" augmented dataset

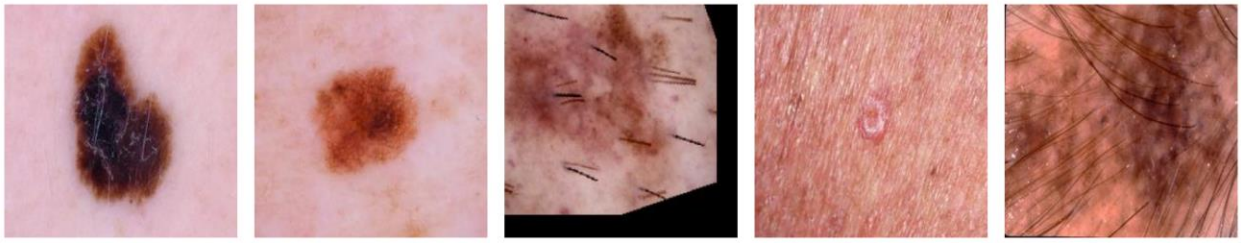


Fig. 7(b). Visualizing samples from "MALIGNANT" augmented dataset

4.4 VISION TRANSFORMER MODEL CONFIGURATION

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 hyperparameters 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.

4.5 TRAINING AND EVALUATION METRICS

The 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.

Training: 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 backpropagation, 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.

Important assessment metrics were used to monitor performance throughout training and validation:

- **Training Loss and Accuracy:** These metrics indicate how well the model learns to perform during training. The train loss contains the closeness of the prediction obtained from the model in comparison to the true labels, whereas the train accuracy is the percentage for images correctly predicted by the model within the training set. The tracking of these metrics helps in ascertaining whether the model is actually learning and converging to a solution or not.

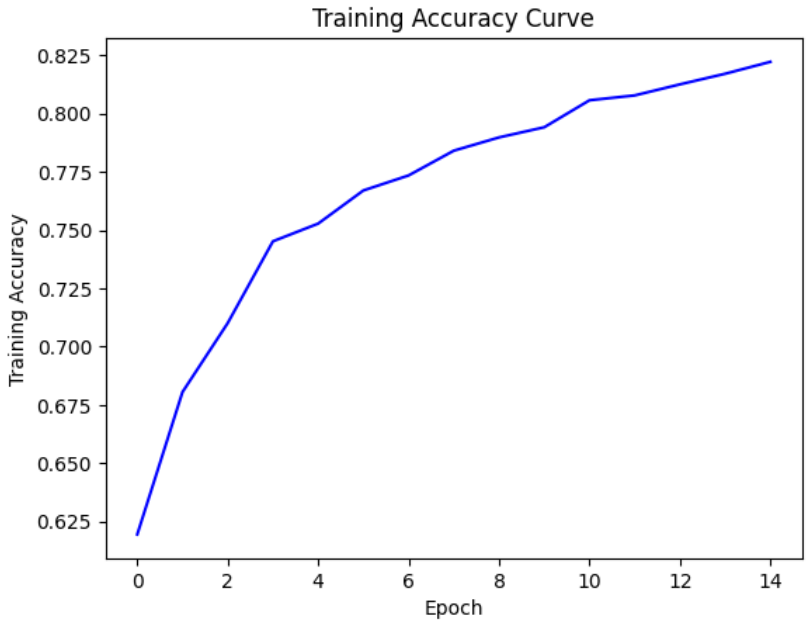


Fig. 8(a). Training accuracy curve on augmented dataset [Base Vision Transformer]

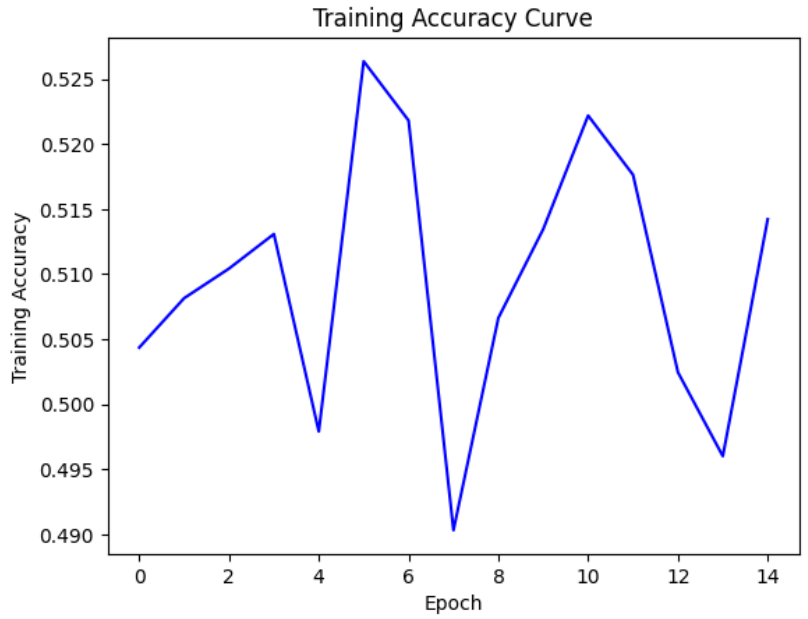


Fig. 8(b). Training accuracy curve on non-augmented dataset [Base Vision Transformer]

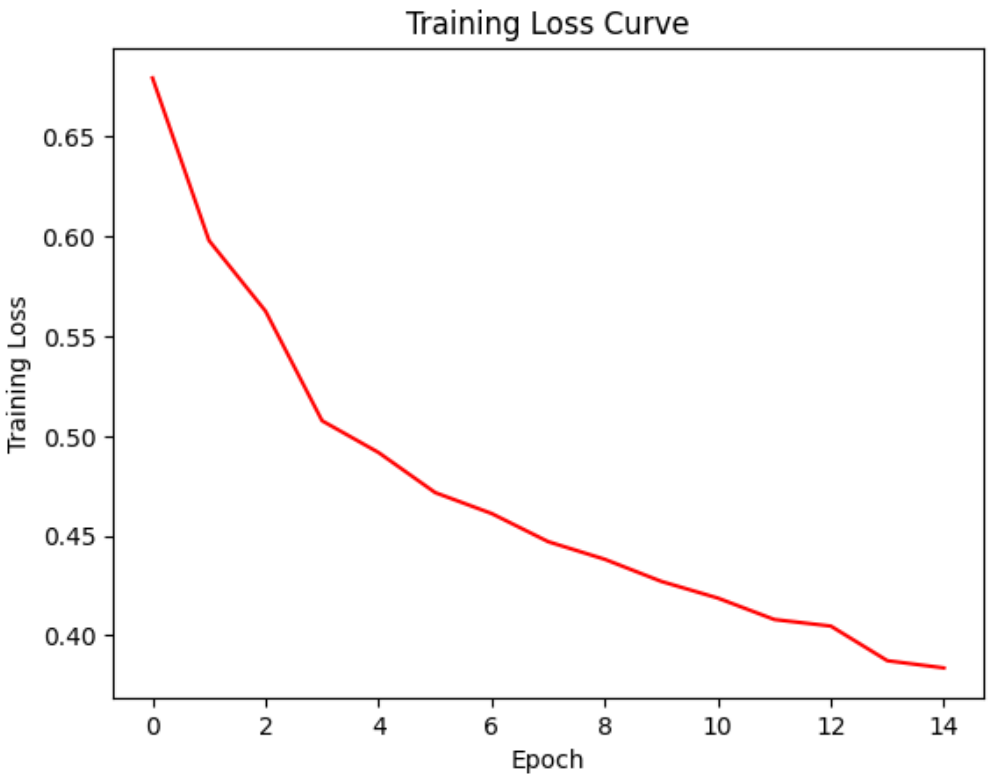


Fig. 8(c). Training loss curve on augmented dataset [Base Vision Transformer]

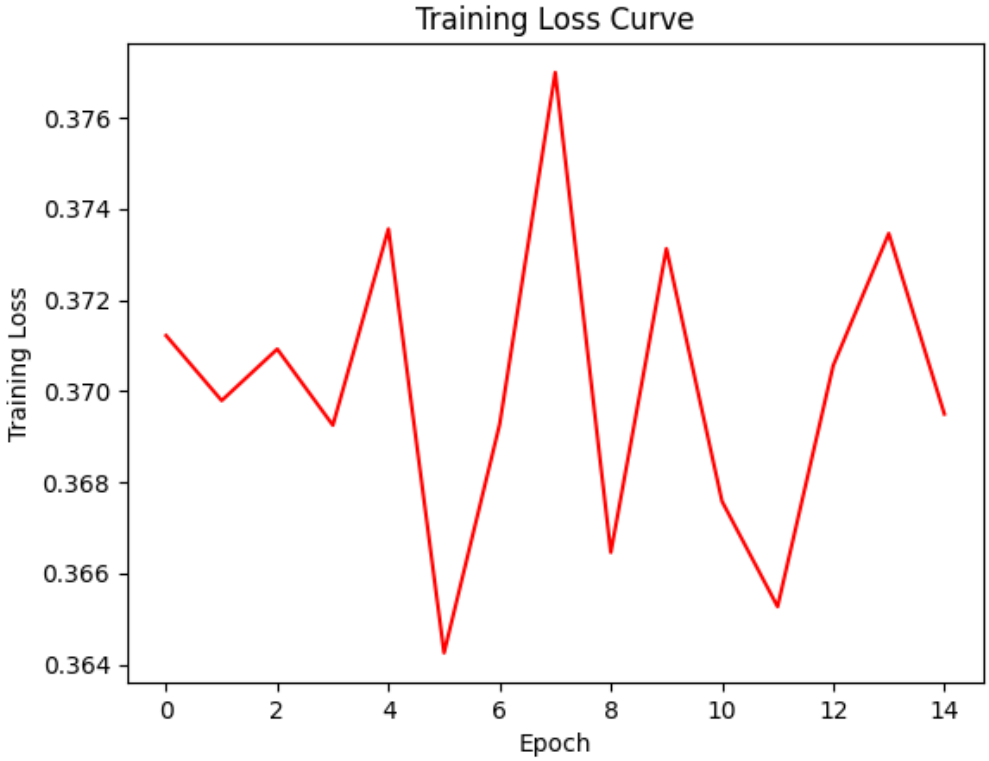


Fig. 8(d). Training loss curve on non-augmented dataset [Base Vision Transformer]

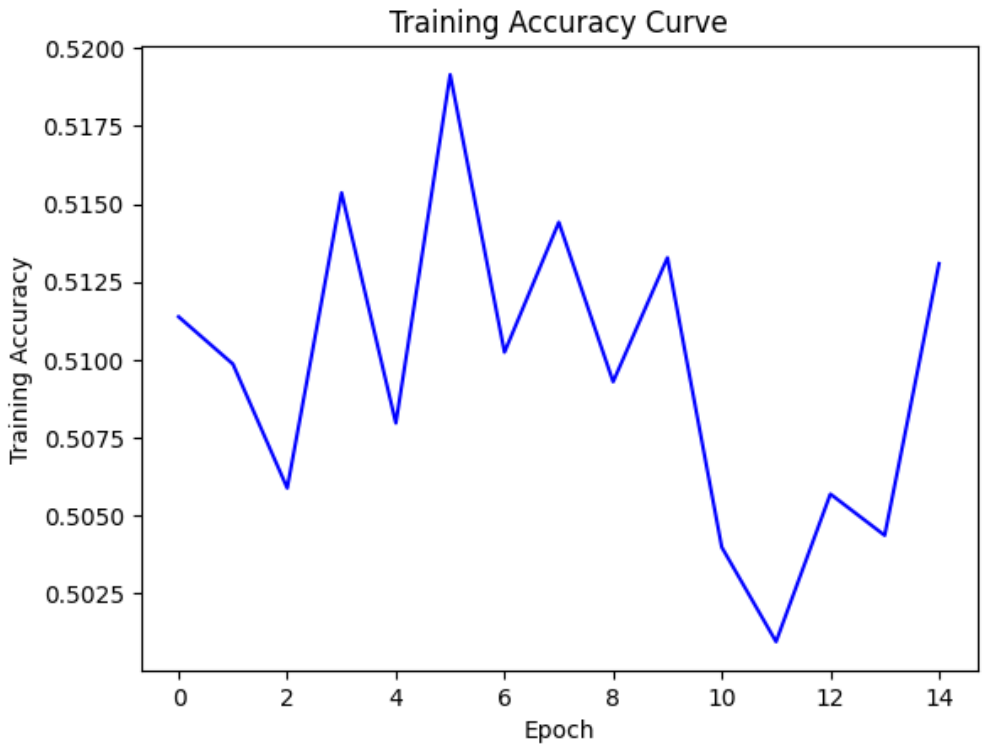


Fig. 9(a). Training accuracy curve on augmented dataset [Deep Vision Transformer]

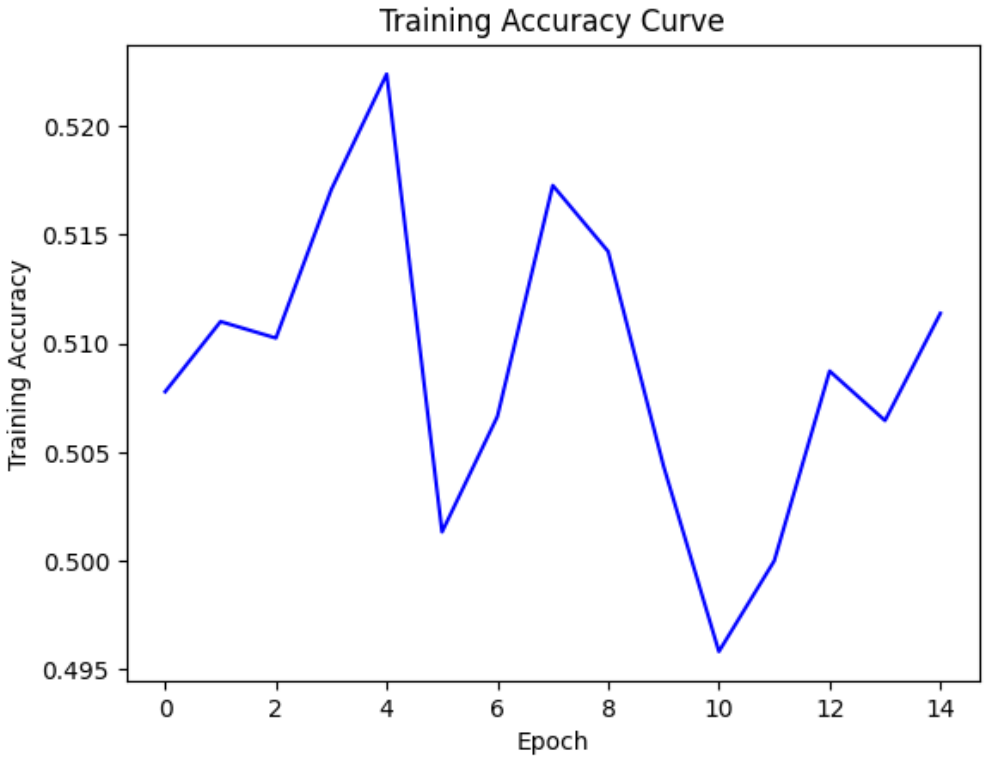


Fig. 9(b). Training accuracy curve on non-augmented dataset [Deep Vision Transformer]

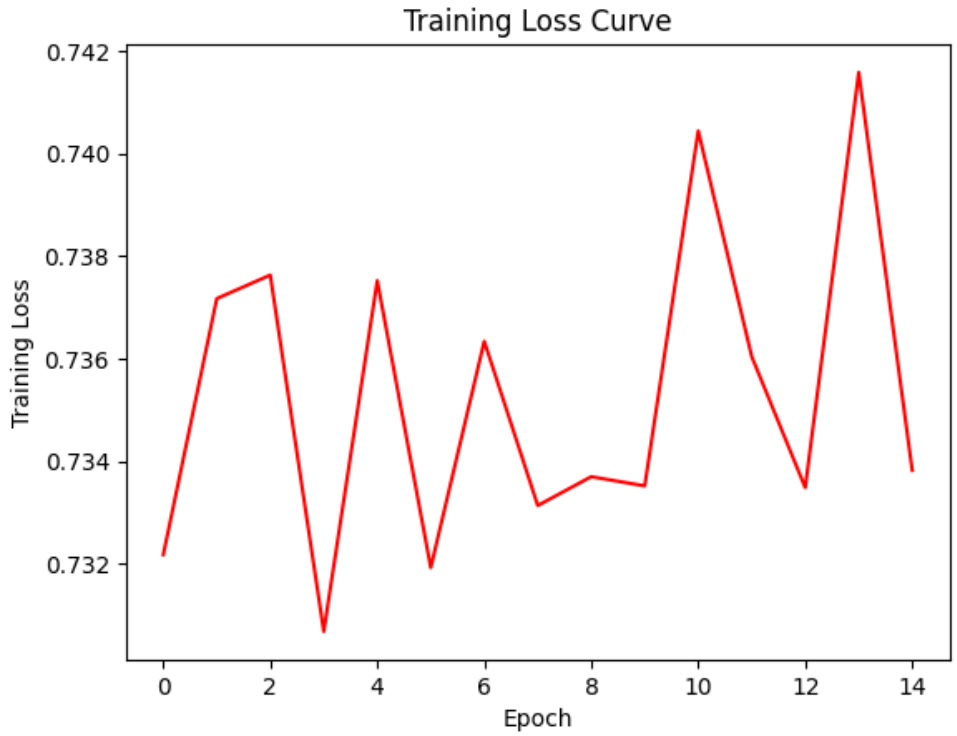


Fig. 9(c). Training loss curve on non-augmented dataset [Deep Vision Transformer]

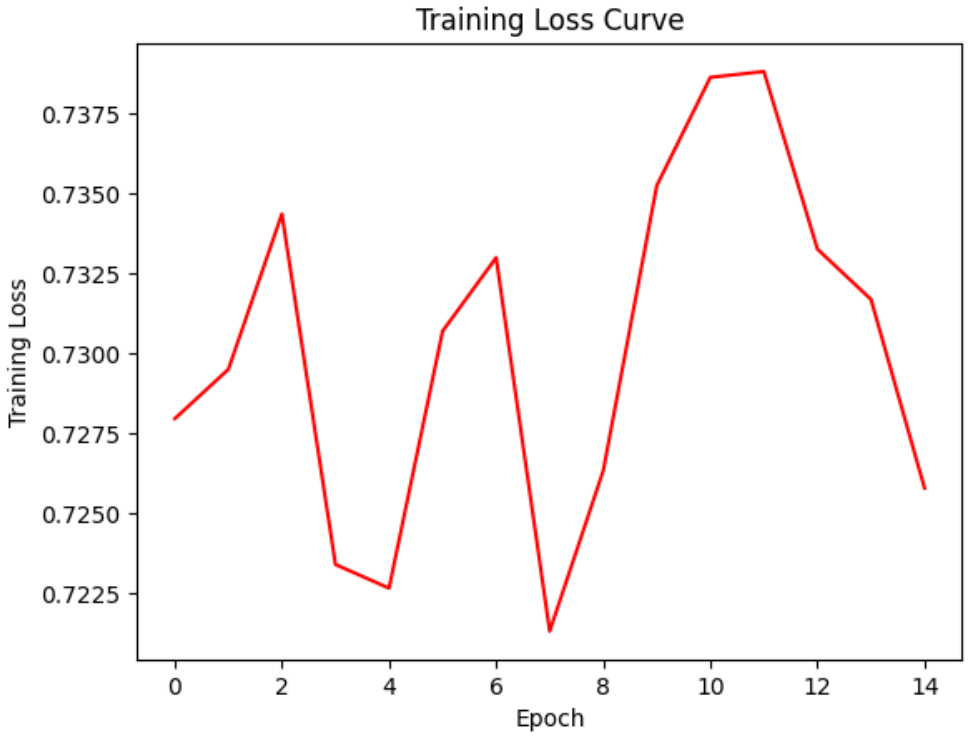


Fig. 9(d). Training loss curve on non-augmented dataset [Deep Vision Transformer]

- **The validation accuracy and loss:** The number of correct predictions made about the validation dataset, and the validation loss quantifies the error over this data. A big gap between the training accuracy and validation accuracy may raise an alarm for overfitting; that is, the model seems to work well on training data but fails to generalize for the new data.

- **Confusion Matrix:** False Positive, False Negative, True Positive, True Negative This is one of the resources to evaluate the performance of a model. This resource contains false positives, false negatives, true positives, and true negatives. It narrates an elaborate classification breakdown on how the model classified one class against other classes. The confusion matrix finds the misclassifications; it is very useful for imbalanced datasets where some classes are underrepresented. Using the confusion matrix, we determine the accuracy of how the model is actually distinguishing between the benign and malignant lesions and indicates where that it needs more effort.

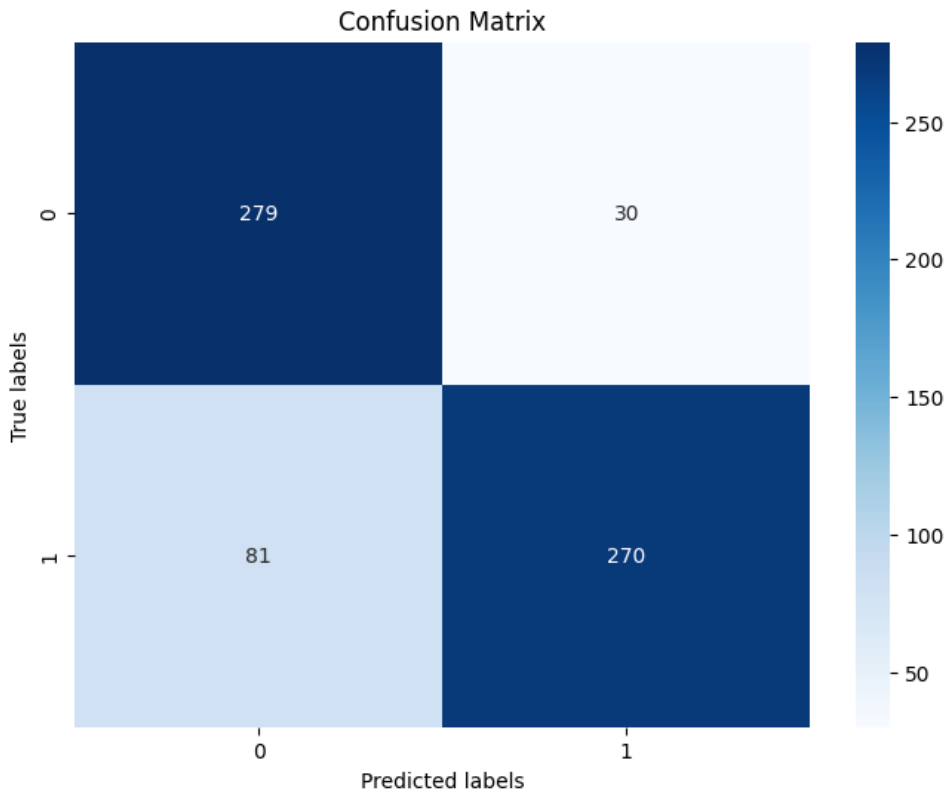


Fig. 10(a). Confusion Matrix on augmented dataset [Base Vision Transformer]

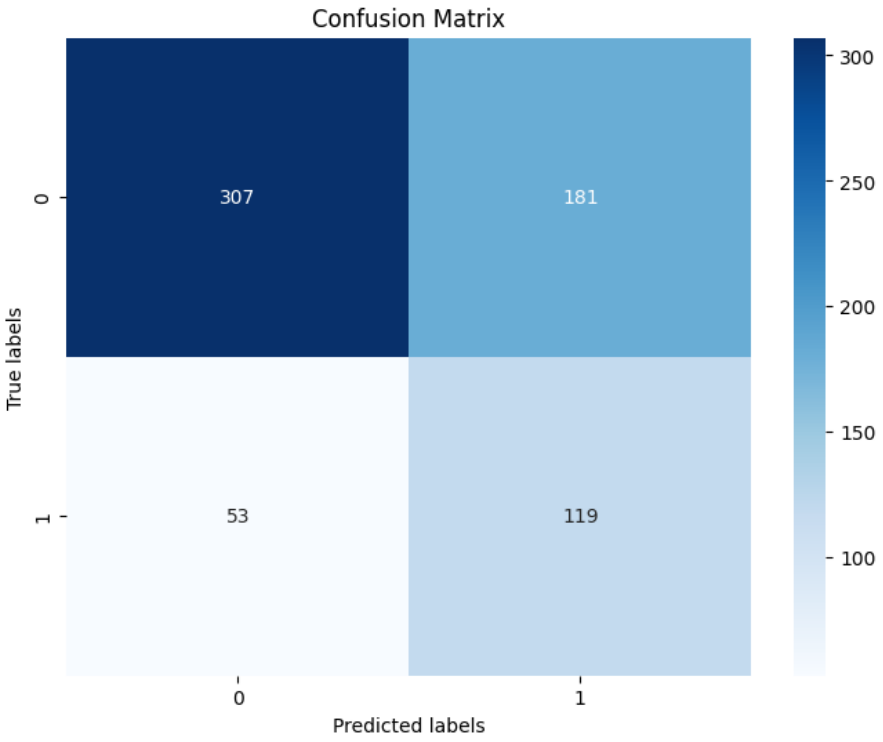


Fig. 10(b). Confusion Matrix on non-augmented dataset [Base Vision Transformer]

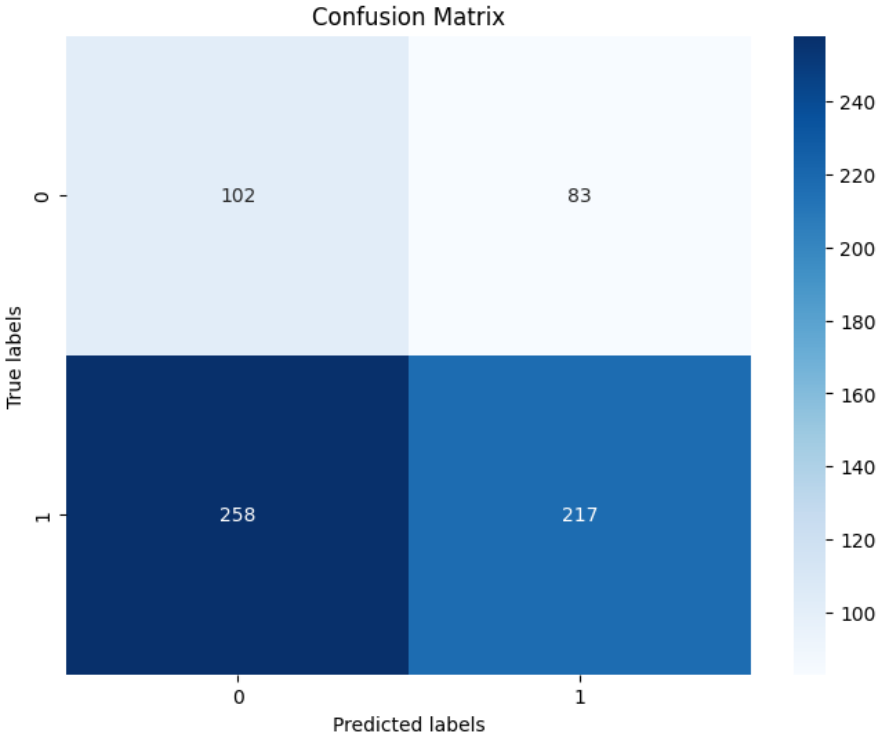


Fig. 10(c). Confusion Matrix on augmented dataset [Deep Vision Transformer]

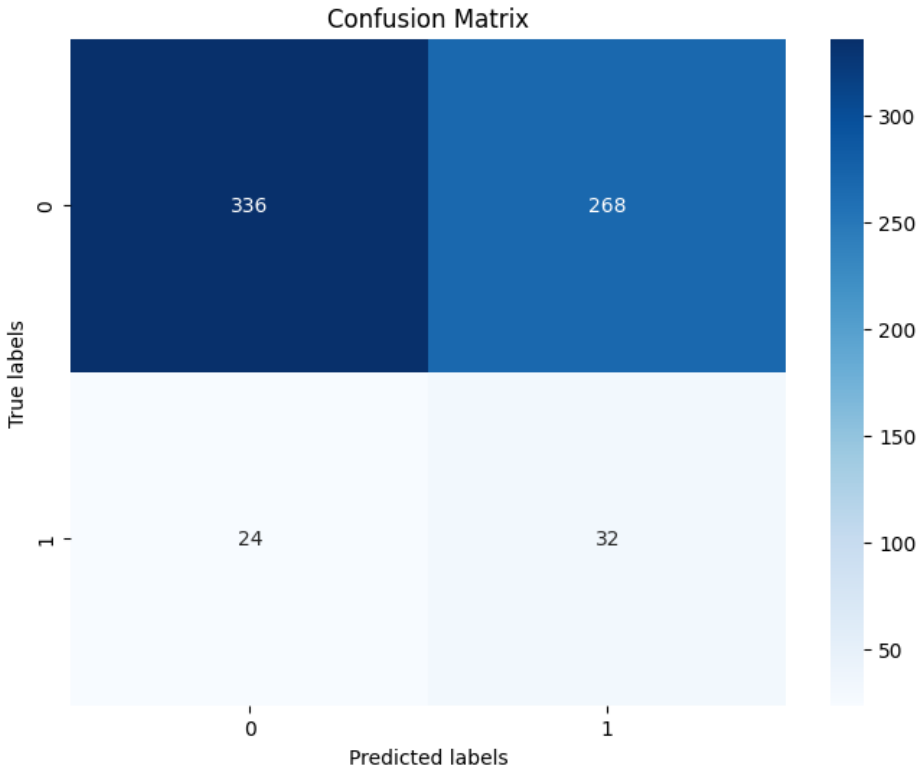


Fig. 10(d). Confusion Matrix on non-augmented dataset [Deep Vision Transformer]

These, altogether, give a comprehensive view to the performance of the model, thereby giving better insight regarding its strength and weakness in skin cancer detection. Evaluation at all phases of training is done regularly, ensuring there are timely adjustments that can be made so that the model is learning well and generalizing to unseen data.

Chapter 5

Results and Discussion

5.1 TRAINING RESULTS ON AUGUMENTED AND NON-AUGUMENTED DATASETS

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.

Data augmentation is also an important component of eliminating model bias, which could otherwise arise due to a limited or unbalanced training data set. This increased dataset sufficiently responds to the issues by providing the model with variations of existing images, hence giving it a broader and more balanced representation of different kinds of skin lesions. Therefore, the model becomes better positioned to detect minimal differences between malignant and benign lesions, which is critical in proper detection of skin cancer.

Overall, the experiment clearly indicates that the augmentation technique can effectively enhance the performance of Vision Transformer models in skin cancer classification. Data augmentation not only perfected the model's ability to learn complex features but also enabled it to overcome the restraints from the 'pure' non-augmented data, thus achieving much higher classification accuracy and better generalization. This work underscores the importance of including data augmentation in the training pipeline, especially in tasks that involve medical image analysis-where gainful access to large, diverse datasets is challenging.

5.2 PERFORMANCE METRICS AND OBSERVATIONS

To analyze the performance of the ViT model, a thorough assessment was conducted using multiple performance metrics, such as training and validation accuracy, loss, and analysis of confusion matrices, on a wide range of performances. All of these would give insight regarding how well it learned and generalized upon its training dataset, especially when it came to the differentiation between benign and malignant skin lesions.

More notably, this model performed much better with respect to the overall accuracy when tested on the augmented dataset. Regarding both the training as well as validation phases, this model revealed a marked reduction in the training loss accompanied by an increase in the accuracy. It can be noted from the two data sets that the training loss of the augmented dataset was extremely small, meaning because of diversity introduced from data augmentation, the model was allowed to converge faster and more effectively. On the other hand, the model trained on the non-augmented dataset gave a greater training loss, which is expected as it might be due to overfitting on less data because of the challenges in optimizing the weights and getting correct predictions that the model had while training.

Of course, in terms of classification accuracy, the model trained on the augmented dataset achieved a testing accuracy of 89.73%, whereas the non-augmented version achieved only 64.55%. It is therefore very evident that dataset augmentation would be important to determine its generalizing capability. The test capability of the model against the variability in the samples improved its performance and enhanced its accuracy.

This was further supported by a more explicit confusion matrix which incorporated true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN). The augmented model shows a clear advantage in correctly classifying benign and malignant cases with 546 out of 660 test samples being correctly classified. That represents an impressive correct classification rate of 82.73% which points out the proficiency of the model in distinguishing between the two classes. The rest were misclassified as benign or malignant. This is a challenge in the detection of skin cancer since some lesions are small, subtle, and often pretty similar in appearance.

A closer look into the confusion matrix showed that the FN rate was significantly lower for the augmented set compared to the non-augmented one, which means that the model would less likely give false negatives when malignant lesions are concerned due to data augmentation. This is a very critical improvement because false negatives in detecting skin cancer can result in dire consequences and hence undiagnosed malignant cases, probably delayed in treatment. Lastly, the rate of false positives was rather low as well, indicating that the model did not incorrectly classify too many of the benign lesions as malignant, thereby improving reliability and diagnostic utility.

5.3 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.

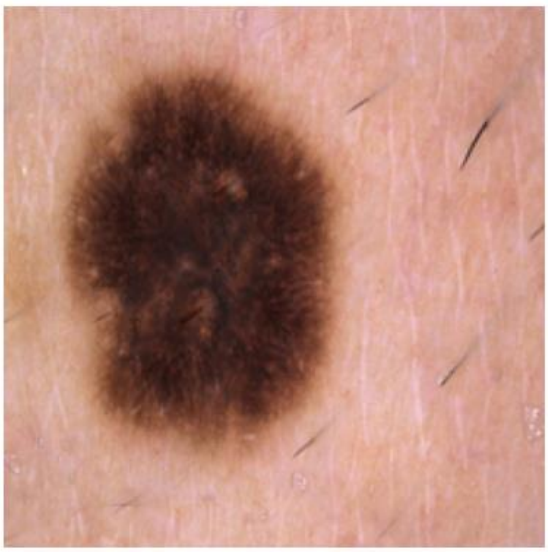
5.4 ERROR ANALYSIS

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.



Predicted Class: Malignant
Correct Class: Benign

Predicted Class: Benign
Correct Class: Malignant

Fig. 11. Misclassification TYPE 1 [LEFT] and Misclassification TYPE 2 [RIGHT]

Conclusion Error analysis reveals that despite the encouraging results of the Vision Transformer models, they fail to differentiate between benign and malignant lesions that are visually very similar to each other with low magnitude. Better data collection, features extracted, or fine-tuned models will eventually improve the accuracy and reliability of skin cancer detection systems.

Chapter 6

Conclusion and Future Work

6.1 SUMMARY OF FINDINGS

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.

6.2 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 generalizability 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. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. An image is worth 16x16 words: Transformers for image recognition at scale, 2020.

2. Tai, C. A., Janes, E., Czarnecki, C., & Wong, A. Double-condensing attention condenser: Leveraging attention in deep learning to detect skin cancer from skin lesion images, 2023.

3. Faghihi, A., Fathollahi, M., & Rajabi, R. Diagnosis of skin cancer using VGG16 and VGG19 based transfer learning models, 2024.

4. Hassani, A., Walton, S., Shah, N., Abuduweili, A., Li, J., & Shi, H. Escaping the big data paradigm with compact transformers, 2022.

5. Himel, G. M. S., Islam, M. M., Al-Aff, K. A., Karim, S. I., & Sikder, M. K. U. Skin cancer segmentation and classification using vision transformer for automatic analysis in dermatoscopy-based non-invasive digital system, 2024.

6. Touvron, H., Cord, M., Douze, M., Massa, F., Sablayrolles, A., & Jégou, H. Training data-efficient image transformers with distillation through attention, 2021.

7. Yuan, L., Chen, Y., Wang, T., Yu, W., Shi, Y., Jiang, Z., et al. Tokens-to-token ViT: Training vision transformers from scratch on ImageNet, 2021.

8. Zhou, D., Kang, B., Jin, X., Yang, L., Lian, X., Jiang, Z., et al. DeepViT: Towards deeper vision transformer, 2021.

9. Masood A, Al-Jumaily AA. Review article computer aided diagnostic support system for skin cancer: a review of techniques and algorithms. Int J Biomed Imaging 2013.

10. Esteva1 A, Kuprel1 B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Nature 2017.

11. Dorj U-O, Lee K-K, Choi J-Y, Lee1 M. The skin cancer classification using deep convolutional neural network. Multimedia Tools and Applications; 2018.

12. Han SS, Kim MS, Lim W, Park GH, Park I, Chang SE. Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm. J Investig Dermatol 2018.

13. Brinker1 TJ, Hekler1 A, Utika JS, Grabe N, Schadendorf D, Klode J, Berking C, Steeb T, von Kalle1 AHEC. Skin cancer classification using convolutional neural networks: systematic review. J Med Internet Res 2018.

14. Demir, A., Yilmaz, F. and Kose, O., 2019, October. Early detection of skin cancer using deep learning architectures: resnet-101 and inceptionv3. In 2019 medical technologies congress (TIPTEKNO).

15. Kondaveeti, H.K. and Edupuganti, P., 2020, December. Skin cancer classification using transfer learning. In 2020 IEEE International Conference on Advent Trends in Multidisciplinary Research and Innovation (ICATMRI).

16. Sedigh, P., Sadeghian, R. and Masouleh, M.T., 2019, November. 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).

17. 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.

18. Rezaoana, N., Hossain, M.S. and Andersson, K., 2020, December. 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)

19. Hekler A, Utikal JS, Enk AH, Hauschild A, Weichenthal M, Maron RC, et al. Superior skin cancer classification by the combination of human and artificial intelligence. Eur J Cancer. 2019.

20. Ali K, Shaikh ZA, Khan AA, Laghari AA. Multiclass skin cancer classification using EfficientNets – a first step towards preventing skin cancer. Neuroscience Informatics. 2022.

LIST OF FIGURES

LIST OF ACRONYMS

1. ViT – Vision Transformer
2. CNN – Convolutional Neural Network
3. ISIC – International Skin Imaging Collaboration
4. MHA – Multi-Head Attention
5. FFN – Feed-Forward Network
6. T2T-ViT – Tokens-to-Token Vision Transformer
7. DeiT – Data-efficient Image Transformer
8. GPU – Graphics Processing Unit