University of Essex

School of Computer Science and Electronics Engineering

CE901 - MSc Project and Dissertation

An Improved Vision-Transformer Network for Skin Cancer Classification

A thesis submitted in fulfillment of the requirements for the degree of Master of Science

Gayathri Mol Shajimon

Reg# 2201829

(gi22846@essex.ac.uk)

Supervisor: **Haider Raza**

August 24, 2023
Colchester

Declaration of Authorship

I, Gayathri Mol Shajimom, declare that this thesis titled, "An Improved Vision-Transformer Network for Skin Cancer Classification" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed: _____

Date: _____

Abstract

Automated methods for early skin cancer detection are crucial as there is a considerable increase in the chances of patient recovery and well-being. The International Skin Imaging Collaboration (ISIC) 2017 dataset is used in this work to evaluate the effectiveness of Vision Transformers (ViTs), a novel and sophisticated approach to the classification of skin cancer lesions. The concept of Vision Transformers has developed as a result of the evolution that the area of computer vision has undergone. Self-attention mechanism gives ViTs a special ability to recognise detailed patterns and characteristics in images. ViTs have the unique capability to grasp significant dependencies within images, outperforming traditional Convolutional Neural Network (CNN) models in this regard. The main benefit of ViTs is their ability to understand and recognize complex visual structures. Due to this capability, ViTs are suited for tasks that require comprehensive image understanding, such as the classification of skin cancer lesions. ViTs can therefore provide precise and trustworthy automated diagnosis, revolutionising the field of medical image analysis.

Using the ISIC 2017 dataset, a detailed study was carried out to demonstrate the viability of ViTs. ISIC 2017 dataset contains a variety of skin lesion images and is frequently used in skin cancer research. The performance of our proposed model has been compared with the Inception-ResNet-V2 + Soft Attention (IRV2 + SA) method and our our ViT-based model and model demonstrated improved performance for binary classification tasks in terms of accuracy, precision, recall, and AUC-ROC score. These findings support the idea that ViTs are well adapted for challenging tasks like classifying skin lesions. The results of this work show that ViTs have great potential for the classification of skin cancer lesions and they are a strong alternative to current CNN architectures. ViTs are positioned as a possible answer to the difficulties associated with skin cancer diagnosis because of their outstanding performance and capacity for learning complex image structures. The use of ViTs in medical image processing hopes to result in earlier and more precise diagnoses, thereby improving patient health and

chances of curability.

Acknowledgement

I would like to acknowledge my sincere gratitude to my supervisor, **Dr. Haider Raza**, for his valuable guidance, support, and insights throughout the research. His immense knowledge and encouragement have helped me a lot in shaping the direction of this work.

I am also thankful to the faculty members and staff at the University of Essex and the Department of Computer Science and Electronics Engineering, for their continuous support and the facilitative academic environment that has enriched my learning experience.

Lastly, I extend my gratitude to all the resources, references, and materials that have contributed to the depth of this work.

Contents

1 Introduction **10**

1.0.1 Aim . . . 10

1.0.2 Objective . . . 10

1.0.3 Dissertation Overview . . . 13

2 Literature Review **14**

2.1 Skin Cancer and its Causes . . . 14

2.2 Skin Cancer in different color tones . . . 14

2.3 Skin Cancer Treatments . . . 15

2.4 Introduction of automatic skin cancer detection . . . 16

2.4.1 Computer based approaches . . . 17

2.4.2 Necissity of Preprocessing . . . 17

2.5 Convolutional Neural Network (CNN) . . . 17

2.6 Transfer learning . . . 18

2.7 Transformers . . . 20

2.8 Vision Transformers(ViTs) . . . 21

2.9 Skin cancer lesion datasets . . . 23

2.9.1 Publicly Available Datasets . . . 23

2.9.2 Closed (restricted) Datasets . . . 24

2.9.3 **ISIC 2017 DataSet** . . . 24

3 Methodology **26**

3.1 Problem Statement . . . 26

3.2 Exploratory Data Analysis . . . 26

3.2.1 Melanoma and Non-Melanoma Images . . . 27

3.2.2 Age distribution . . . 27

3.2.3 Gender distribution 28
3.3 Image Preprocessing 28
3.4 Data Augmentation 30
3.5 Loss Functions 31
3.5.1 Binary Cross Entropy Loss 31
3.5.2 Focal Loss 31
3.6 Hair Removal Algorithm 32
3.7 Batch Normalization vs Layer Normalization 34
3.7.1 Batch Normalization 34
3.7.2 Layer Normalization 34
3.7.3 Batch Normalization vs Layer Normalization 35
3.8 **Proposed System** 35
3.9 **Hardware and Software Requirements** 37
3.10 **Data Partitioning** 38
3.11 **Model Training** 39
3.12 **Model Testing and Evaluation** 39

4 Result **40**
4.1 **Evaluation matrices** 40
4.2 **Experiment Results** 41

5 Discussion **44**

6 Conclusions **46**

A Appendix **48**

B Another Appendix **49**

List of Figures

2.1 Melanoma Cases in Different Groups [1] 15
2.2 Representation of (a)traditional machine learning and (b) transfer learning[2] 19
2.3 Existing methods of single-head and multi-head (transformer) designs for vision tasks are revealed by a taxonomy of the self-attention design space. Efforts are also made to incorporate convolution-based architecture expertise to improve ViTs, such as through multi-scale and hybrid designs[3] 21
2.4 Model proposed by Yang et. al[4], which is current state-of-the-art model using Vision Transformer 22
2.5 Data distribution - Train Dataset 25
2.6 Distribution of cases in each class for Task 1 and Task 2 25

3.1 Melanoma Images 27
3.2 Non-Melanoma images 27
3.3 Age distribution in the Train and Test dataset 28
3.4 Age distribution of melanoma cases in Train and Test dataset 28
3.5 Gender distribution in Train and Test dataset 29
3.6 Gender distribution of melanoma cases in Train and Test dataset 29
3.7 Resultant images before and after hair removal 34
3.8 Skin Cancer Detection System Block Diagram 36
3.9 General ViT architecture used for skin cancer classification. 37
3.10 Proposed 9-layer ViT Model Architecture with layer normalization and regularization 38

List of Tables

1.1 Comparison of Benign and Melanoma Features 11

4.1 Test dataset results for data partitioning scheme SS1 under setting S1 (i.e., training with 5-fold CV) 41

4.2 Test dataset results for data partitioning scheme SS2 under setting S1 (i.e., training with 5-fold CV) 41

4.3 Comparison with state-of-the-art results on test dataset for data partitioning scheme SS1 under setting S2 (i.e., training on whole training data), where DA - Data Augmentation, FL - Focal Loss, BCE - Binary Cross Entropy 42

4.4 Comparison with state-of-the-art results on test dataset for data partitioning scheme SS2 under setting S2 (i.e., training on whole training data), where DA - Data Augmentation, FL - Focal Loss, BCE - Binary Cross Entropy 42

CHAPTER 1

Introduction

This section gives an overall idea about the aim, objective and research and analysis we have conducted to achieve the goal.

1.0.1 Aim

The aim of this project is to implement an innovative solution using Vision Transformers by utilizing its self-attention mechanism, to achieve state-of-the-art performance and thereby a better recall and accuracy in classifying cancerous and non-cancerous lesions using the ISIC 2017 dataset.

1.0.2 Objective

The following are the main objectives of this experiment:

- To do a background study of the evolution of automatic skin cancer detection.
- To explore the ISIC 2017 dataset and study the important features of the images and data provided in order to efficiently utilize them for the experiment.
- To implement Skin Cancer Lesion classification using ViTs in ISIC 2017 dataset and to achieve state-of-the-art result
- To submit this research paper in BIBM 2023.

Skin cancer is a predominant form of cancer and poses a significant public health concern worldwide due to the incidence, mortality rates and morbidity. Long-term exposure to ultraviolet radiation from the sun is considered to be the primary etiologic agent in the growth of skin cancer [5]. Compared to other cancers, skin cancer detection is challenging as it can sprout anywhere in the body. Among different types of skin cancers, Melanoma is considered the most aggressive and deadliest[6]. The unique features of melanoma such as uneven distribution, asymmetrical shape, scalloped or notched borders, and uneven distribution of colours will be helpful in distinguishing it from other skin cancer types. Early identification of Melanoma is critical as it has the tendency to spread quickly to other parts of the body and is resistant to traditional treatments. Early detection of melanoma improves the chances of successful treatment and increases the possibility of patient survival. To handle this paramount need, the integration of Artificial Intelligence (AI) and machine learning has emerged as a reassuring approach to melanoma detection. Researchers have made significant progress in building automated and accurate methods to assist doctors in quickly and reliably diagnosing melanoma lesions by leveraging AI algorithms and computer vision techniques[7]. Recent studies show these new technologies outperform dermatologists in various multi-class skin cancer classifications [8].

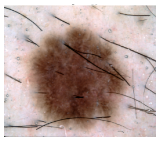
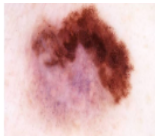
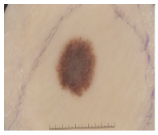
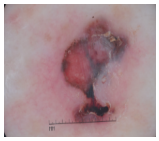
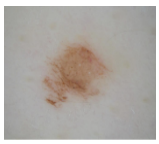
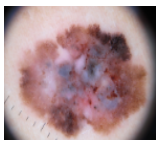
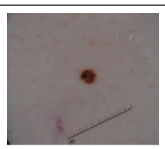
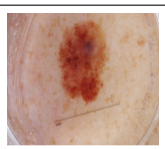
Benign	**Melanoma**	**Feature**	**Characteristics**
		Asymmetry	Half of the mole does not match with the other half
		Border	When the border or edges of the mole are ragged or irregular
		Color	Color of the mole varies
		Diameter	The diameter of the mole larger than 1/4 inch for melanoma[9]

Table 1.1: Comparison of Benign and Melanoma Features

The automatic detection of skin cancer has made significant developments, moving from traditional image processing methods to cutting-edge deep learning models. Early efforts focused on using manually created features and rule-based algorithms. However, these methods had trouble treating complicated skin lesions successfully, which led to the development of data-driven procedures as a more viable route [10]. The development of Convolutional Neural Networks (CNNs), which revolutionised the ability to capture detailed details within skin lesions, was a huge advance [11]. Exceptional performance was displayed by models like VGGNet [12], ResNet [13], and InceptionNet [14], which helped this field. Even though they were effective, CNNs relied on already existing spatial hierarchies, which led to the creation of the Vision Transformer (ViT).

ViTs introduced a novel method for capturing both local and global dependence by utilising self-attention techniques. It is well-suited for the investigation of a variety of skin lesions [15] classification tasks. When used on large skin lesion datasets such as those from the International Skin Imaging Collaboration (ISIC), ViT particularly showed outstanding performance in terms of accuracy, sensitivity, specificity, and Area Under the Curve (AUC). This success underlines ViT's key role in advancing automated skin cancer identification. ViT's ability to improve diagnostic precision was demonstrated by its integration with huge datasets. To further improve ViT's effectiveness in skin cancer diagnosis, researchers are focusing on improving both its structural architecture and training procedures. This dedication to ongoing study shows an understanding of the importance of ViT and a determination to push the limits of its potential.

The transition from conventional approaches to cutting-edge deep learning models is evidence of the extraordinary advancements made in the field of automated skin cancer diagnosis. The possibility of even more precise and trustworthy automated skin cancer identification is looming as the focus continues to shift toward improving ViT and its applications. This not only raises the possibility of improved diagnostic procedures but also emphasises the crucial part AI-powered medical technology plays in the development of healthcare.

In this paper, our aim is to explore the application of ViT in the field of skin cancer lesion detection, specifically in the ISIC 2017 dataset. The ISIC dataset offers a wide range and comprehensive collection of dermoscopic pictures, enabling researchers to

create and test reliable skin cancer detection models[16]. We investigated the potential advantages and limitations of ViTs over current state-of-the-art Inception ResNetV2[17] comparing their performance. Our study intends to add to the growing body of knowledge in automated skin cancer detection by utilising ViT and the ISIC 2017 dataset. ViTs are based on the transformer architecture which was initially proposed for natural language processing tasks. In the field of machine translation, sentiment analysis and text generation, transformers have achieved groundbreaking success. This identical idea was applied to tasks in computer vision, leading to the emergence of ViTs. ViTs can directly analyze images, eliminating the requirement for pre-established spatial structures[18]. The findings from this study hold the promise to enhance the accuracy, efficiency, and accessibility of skin cancer diagnosis, ultimately benefiting patient outcomes and reducing healthcare system burdens.

1.0.3 Dissertation Overview

This section explains our aim, and objective and an introduction to the experiment we have done to get the state-of-the-art result. The organization of the remaining chapters are structured as below:

- Chapter 2 - Contains a background study, which highlights the recent advancements in deep learning in the field of skin cancer detection and the description of the dataset.
- Chapter 3 - contains the methodology and proposed system which includes preprocessing techniques, the system architecture and solutions.
- Chapter 4 - contains the result and impact which includes all the experimental results and comparison with current state-of-the-art results.
- Chapter 5 - contains discussions, which include the challenges and suggestions for future improvements
- Chapter 6 - This section contains the summary and our final results at the end of the experiment.

CHAPTER 2

Literature Review

2.1 Skin Cancer and its Causes

Skin cancer is considered to be a common type of cancer particularly in the fair-skinned population[5]. According to the author, Ultraviolet radiation(UVR) acts as the primary etiological agent behind the development of skin cancer. UVR possesses the ability to damage DNA and induce genetic mutation, which ultimately results in the formation of skin cancer. Also, they have mentioned that the current UVR-based treatment modalities (such as phototherapy) can also increase the risk of developing skin cancer. Skin cancer is mainly divided into two, cancers derived from melanocytes called melanoma and another type is which derived from epidermal cells called non-melanoma cancer [19]. Craythorne et.al [19] also mentioned that 95% of skin cancers belong to the category of non-melanoma cancers, thus making only a few percentages of skin cancers diagnosed to be melanoma. Furthermore, they have also mentioned that skin cancer can be developed genetically and also due to environmental factors like long-time exposure to UVR rays.

2.2 Skin Cancer in different color tones

Skin cancer in different skin colours is considered for study in [1]. According to their study authors conveys that dark skin has the capability of filtering UVR due to increased

level of epidermal melanin. But on the other hand white skin transmits more UVR due to less melanized melanosomes. Also, they have mentioned that black men have a higher risk than black women in melanoma incidence. Furthermore, when considering the ratio of the incident rate of melanoma in dark-skinned and white-skinned in the United States is 1:16. They have also mentioned that in dark skin tone, melanoma tends to develop in non-sun exposed sites such as plantar, palmar and mucosal surfaces. Even though the melanoma incidence in dark skin tone is less, the mortality rate is high due to poorer diagnoses including delays in diagnosis and treatment, and also the presence of frequent thick primary lesions and naturally more aggressive acral tumours.

((a)) Melanoma in black-skinned male ((b)) Melanoma in white-skinned male

Figure 2.1: Melanoma Cases in Different Groups [1]

2.3 Skin Cancer Treatments

Early diagnosis of melanoma is essential to stop the frequent spread and for better curability. The visual diagnosis of cancer lesions depends on the expertise of the physician. The accuracy of diagnosis via visual inspection is less reliable as there is a higher chance of misdiagnosis of cancer lesions in the early stages for example melanoma and benign nevi[20]. Traditional medicine has played a significant role in human history in identifying a variety of illnesses, including skin cancer. Technologies and conventional approaches have both been used to diagnose ailments throughout history. Dermoscopy is a cutting-edge method that Dorrell et al. [21] suggested for quickly evaluating suspicious lesions. It is a cost-effective and time-saving tool for dermatologists. Dermoscopy enables a more accurate visual inspection of the structures and patterns of lesions, with a specificity of 0.81 and a sensitivity of 0.71 in visual

assessment without magnification. Dermoscopy can be used by medical professionals to perform tests with elevated accuracy, specificity, and sensitivity of 0.9. Despite of these much developments, dermoscopy has not improved melanoma survival rates, according to [22]. For a clear diagnosis, dermoscopy is frequently used in conjunction with more traditional methods like biopsy and histopathology analysis. Dermoscopy mastery requires substantial training and remains as a subjective process.

Various techniques used for skin cancer detection include - multiphoton tomography and high-frequency ultrasound[21]. A study conducted Almaraz-Daminan et al. [22] highlights how the experience of experts can lead to different opinions and diagnoses. The different methods used for identifying skin cancer produced different results, this is one of the main reasons behind the misdiagnosis. These uncertainties lead to the need for a standardized and computerized approach. The purpose of this strategy is to create a system that is portable, quickly diagnoses diseases, and is affordable and comfortable for patients thereby reducing the need for biopsies. This method should also continue to detect skin cancer with excellent levels of sensitivity and accuracy. From the above studies, it is clear that traditional approaches and developing technologies coexist in the effort to strengthen skin cancer detection procedures. Dermoscopy is a potential visual assessment tool, but its limits are highlighted by the reliance on human competence and the subjective aspect of the procedure. In order to improve accuracy, fill knowledge gaps, and eventually advance the prediction and treatment of skin cancers, there is a high requirement for a computer-aided, standardised, and user-friendly approach.

2.4 Introduction of automatic skin cancer detection

In recent times, there have been significant advancements towards skin cancer detection research. These advancements include areas such as traditional medicine, computer-aided image processing, deep learning, transfer learning, and ViTs. Computer-aided image processing is one promising field, as it allows for the analysis of a large amount of image data, which can be helpful in detecting melanoma cases quickly. Also, automatic skin cancer detection tools will be helpful for dermatologists to take a second opinion or to confirm the disease more accurately.

2.4.1 Computer based approaches

The applications of computer-aided approaches to skin cancer detection have shown remarkable potential in this field. Jain et.al[23] introduced an inventive approach by applying ABCD (Asymmetry, Border, Colour, Diameter) dermoscopic rule, image processing techniques such as resizing, gamma correction and compensation for non-uniform illumination etc for feature extraction and image segmentation. As a result, a reliable lesion image analysis technique for melanoma detection was created. Also, a study conducted by Saba et.al [24] engaged in a thorough comparison of handcrafted and non-handcrafted features, including deep learning techniques like CNN and DCNN as well as texture-based transformations like Gabor wavelets and well-established rules like the ABCD rule and Menzies scoring. This study demonstrated the effectiveness of a well-trained computerised system for autonomous skin cancer diagnosis.

2.4.2 Necissity of Preprocessing

In order to make further progress, it is necessary to remove obstacles from images including hair, glare, and shading while utilising feature extraction methods to identify colours, textures, and shapes. These improvements provide a more sophisticated method of segmenting images than segmentation techniques like Watershed, Otsu, and modified Otsu. Preprocessing of images plays an important role in the performance of the model, so that we can ensure the quality of the images given for training in the computer-aided classifiers, which overcomes the limits of manual diagnosis. This convergence of cutting-edge methods, supported by a paradigm focused on computers, has the potential to revolutionise the field of skin cancer detection by enabling precise and effective diagnostics while reducing the reliance on human intervention.

2.5 Convolutional Neural Network (CNN)

CNN has become a powerful tool in the field of image classification, because of its ability to recognise complex patterns in visual data. This capacity, which enables CNNs to properly interpret the complexity of image, becomes more apparent when they are trained on substantial datasets. A comprehensive CNN model with fully-connected

layers was introduced by Fuadah et al. [25] and was intended to perform well in multi-class classification settings. The ISIC dataset, which includes different classes including dermatofibroma, nevus pigments, melanoma, and squamous cell carcinoma, was used to evaluate this model. A training dataset of 4,000 photos was created using data augmentation techniques to balance the dataset and ensure robustness. The results were impressive since they demonstrated the effectiveness of contemporary computer-aided approaches like CNNs and their amazing accuracy of 0.99. Esteva et al. [11] used CNN models to categorise photos of skin cancer, providing additional evidence of the effectiveness of CNNs. Their study focused on two crucial scenarios: differentiating malignant melanoma from benign nevi and differentiating benign nevi from keratinocyte carcinomas. They did this by using a massive database of 129,450 clinical pictures. The attained results, which were notable for being closely in line with dermatologists' diagnostic proficiency, highlighted CNN's ability to overcome diagnostic gaps. Subramanian et al. [26] highlighted the potential of CNNs in the classification of skin cancer by demonstrating the performance of their model on the HAM10000 dataset. Notably, their model was able to keep the false negative rate to 0.1 while maintaining accuracy and precision over 0.8. This accomplishment highlights CNN's capacity to reduce false negatives and maintain diagnostic accuracy, making it a powerful tool in the field of skin cancer categorization. As a result of these studies and combined findings, CNNs have become an essential component of modern computer-aided diagnostic efforts, and their capacity to analyse complex visual data has propelled them to the forefront of skin cancer detection research.

2.6 Transfer learning

Transfer learning has become a popular and promising area in machine learning because of the numerous application possibilities. The main objective of transfer learning is to improve performance by transferring the knowledge contained in different but related domains to the target learner on the target domain[27]. This will reduce the requirement for a large amount of target domain data for training the model. In order to address the challenge of limited data in specific domains, researchers are increasingly utilizing high-performing models that have been trained on more accessible and substantial

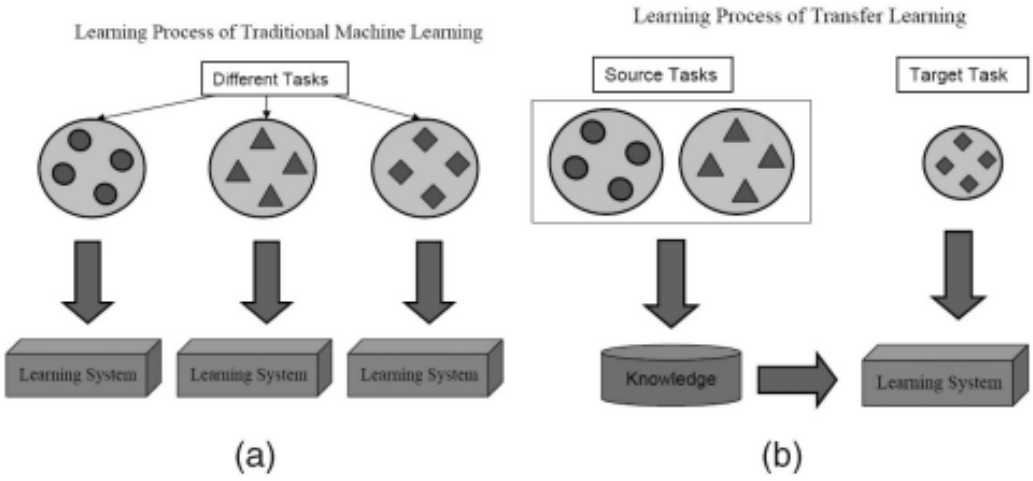


Figure 2.2: Representation of (a)traditional machine learning and (b) transfer learning[2]

datasets from various domains [28]. When there is an issue with data scarcity or limited data, researchers leverage the knowledge acquired by this model and adapt the same to their respective domains. One study, conducted by Dorj et al. [29], focused on skin cancer detection using a dataset of 3753 images, they explored the classification of four different types of skin cancer such as actinic keratoses, basal cell carcinoma, squamous cell carcinoma, and melanoma. Employing transfer learning, a technique that leverages pre-trained neural network models, they tackled the challenge of limited data availability using this method. To be more precise, they employed an error-correcting output code (ECOC) Support Vector Machine (SVM) for classification and the AlexNet convolutional neural network (CNN) to extract features from the images. Despite having access to a relatively small dataset comprising 768 images for training and 190 for testing they achieved remarkable results. The model attained accuracy, sensitivity, and specificity values of 0.94, 0.98, and 0.91 respectively, showcasing its efficacy in diagnosing various forms of skin malignancies.

In an independent study, Ali et al. [8] concentrated on improving the method of skin cancer diagnosis using convolutional neural networks. They started by adding a preprocessing step to the HAM10000 dataset that involved hair removal. Then they carried out resizing and data augmentation to increase the diversity of the dataset. Their research's main focus was on developing and testing different CNN models using the EfficientNet family, ranging from B0 to B7. Surprisingly, they found that models of intermediate complexity, more notably EfficientNet B4 and B5, produced the

best outcomes. Impressive Receiver Operating Characteristic Area Under the Curve (ROC AUC) scores of 0.98 was attained by these models, demonstrating that increased model complexity does not automatically correlate to better performance. Overall, both studies demonstrate the potential of utilising transfer learning and carefully chosen preprocessing methods to address data constraints in medical image analysis. They offer important insights on the effectiveness of applying cutting-edge deep learning techniques to certain domains, where getting large datasets might be difficult [**?**], [**?**]. Researchers have explored the potentials of ViTs for skin cancer classification given their unique ability to uncover long-range dependencies and complexities in images in ways CNNs are limited in. This is a significant advantage when compared to the limitation faced by CNN in handling such complexities effectively.

2.7 Transformers

Transformers are a type of deep-learning model that uses a self-attention mechanism to obtain contextual information from the input sequence which helps them to excel in various NLP and Computer Vision tasks. The two key ideas that contributed to the development of the conventional Transformer model include - 1) **Self-attention** - helps to capture long-term sequence elements on the other hand it is challenging for recurrent models to encode such relationships.2) *Pretraining* - is the second key idea for pretraining on large labelled or unlabelled corpus in a supervised or self-supervised. manner[3]. Vision models with self-attention are broadly divided into two categories such as the model with 1. ***Single-head self-attention***[3].: refers to the original self-attention mechanism which got introduced in the transformer architecture. To compute the weighted representation of each element in the sequence a single attention mechanism is used. For each element in the sequence, this single attention mechanism computes a single set of weighted context vectors. Three basic steps in computation include - Query, Key and Value. The similarity score will be first calculated by taking the dot product of Key vectors and query vectors. Based on the similarity scores query vectors retrieve the information from key vectors. The softmax function is then used to normalise the resultant scores in order to obtain the attention weights. The final contextualised representations are created by computing a weighted sum of value

vectors using these weights. Some examples include: BERT, GPT etc

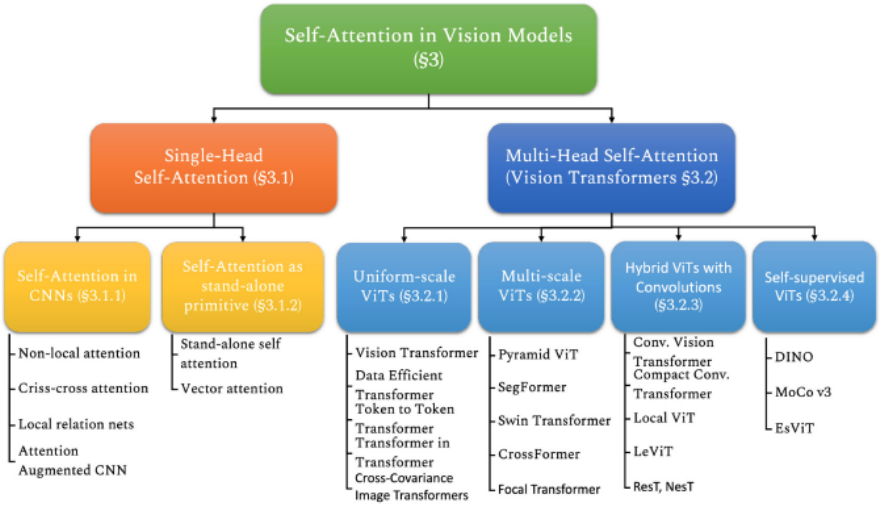


Figure 2.3: Existing methods of single-head and multi-head (transformer) designs for vision tasks are revealed by a taxonomy of the self-attention design space. Efforts are also made to incorporate convolution-based architecture expertise to improve ViTs, such as through multi-scale and hybrid designs[3]

2. ***Multi-head Self-Attention***[3].: this extends the single-head mechanism by employing multiple parallel attention mechanisms in the self-attention layer. Each head is expected to learn different relationships focusing on different parts of the input sequence. The final output is produced by concatenating and linearly transforming the result of the self-attention layer. The model is capable of capturing complicated patterns and contextual information because of the multi-head self-attention that aids in the capturing of many dependencies and relationships within the input sequence. Since each head can focus on different portions of the input the model can better capture both local and global dependencies. Some of the examples include the Original Transformer model, T5, BERT with Multiple Heads, DeiT, Image GPT, ViT-GAN, and TransUNet.

2.8 Vision Transformers(ViTs)

Vision Transformers (ViTs) are a class of deep learning models that use the transformer architecture, which was originally created for natural language processing, later experimented in computer vision tasks and produced state-of-the-art results[18]. ViTs

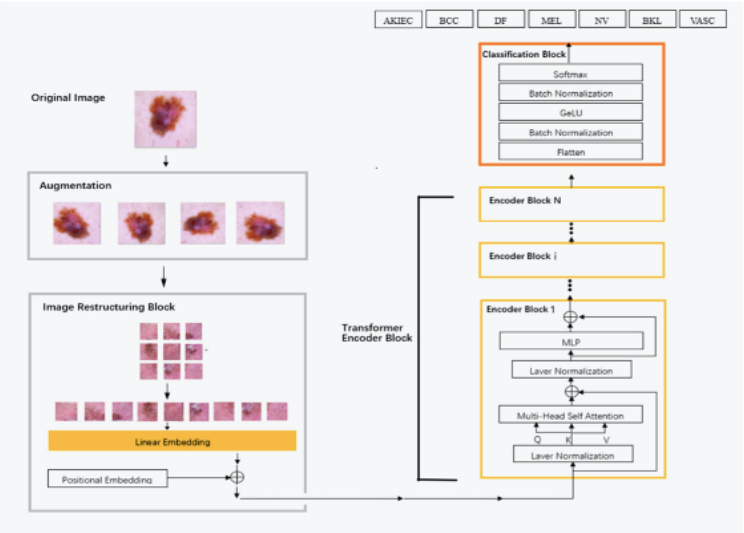


Figure 2.4: Model proposed by Yang et. al[4], which is current state-of-the-art model using Vision Transformer

divide input images into fixed-size patches, linearly alter them, and then feed those transformed images through self-attention processes to capture both global and local contextual information. This method eliminates the requirement for manually creating convolutional processes and has displayed outstanding performance in a variety of image classification and segmentation tasks. [**?**]. Yang et. al[4] proposes a novel ViT model for skin cancer lesion classification. Different techniques were used, such as rebalancing classes, splitting images into patches, tokenizing them, and using a transformer encoder with self-attention for classification. Using a pre-trained image net that was fine-tuned on the HAM10000 dataset, their results surpass that of Datta et. al[17] with an accuracy of 0.94. This shows that the attention layers enhanced the model's performance. We have implemented the architecture proposed by Yang et. al[4] in ISIC 2017 as one result for comparison. We have made some architectural changes in the classification block by replacing batch normalization with layer normalization and adding some additional regularization. The architecture proposed by them is illustrated in Fig. 2.4. Xin et. al[30] proposed a new ViT model for skin cancer image feature extraction and lesion classification using multi-scale patch embedding, overlapping sliding windows, and constructive learning. Applying this concept to the HAM10000 dataset, an accuracy of 0.94 and a precision of 0.94 was achieved.

In conclusion, the development of ViTs has resulted in a paradigm shift in the classi-

fication of skin cancer. These results highlight ViTs' superior capacity to comprehend complex picture structures and linkages, creating a new standard for performance and accuracy in the classification of skin cancer lesions. These findings are expected to open the door for improved skin cancer detection and treatment as the medical industry continues to adopt cutting-edge technologies.

2.9 Skin cancer lesion datasets

2.9.1 Publicly Available Datasets

1. ISIC Archive: The ISIC archive gallery comprises a diverse collection of clinical and dermoscopic skin lesion datasets from across the globe[31],

2. HAM10000[32]: It is a collection of 10,000 training images aimed at identifying pigmented skin lesions, featuring dermatoscopic images gathered from various populations and acquired through different modalities.

3. BCN20000[33]: consists of 19,424 dermoscopic images depicting skin lesions that were acquired between 2010 and 2016 at the Hospital ClÃnic in Barcelona. This dataset is suitable for tasks involving lesion identification, including segmenting, detecting, and classifying skin lesions.

4. MED-NODE Dataset[34]: This dataset offers 170 clinical images representing both melanoma and nevi cases. It is conveniently accessible for research purposes.

5. Asan Dataset[35]: Featuring a compilation of 17,125 clinical images depicting various skin diseases prevalent among Asian populations.

6. Dermnet NZ: Renowned for its comprehensive array of clinical, dermoscopic, and histological images depicting diverse skin diseases. Researchers can utilize these images for academic research purposes.

7. The Cancer Genome Atlas[36]: A substantial repository boasting 2,871 cases of pathological skin lesion slides. This dataset is freely accessible to the research community.

2.9.2 Closed (restricted) Datasets

1. Dermofit Image Library[37] A curated collection featuring 1300 high-resolution images categorized into 10 classes of skin lesions. A licensing agreement is required, entailing a one-time fee of €75. An academic license option is available as well.

2. PH2 Dataset[38]: This dataset offers a compilation of 200 dermoscopic images, encompassing 40 melanoma and 160 nevi cases. Access is granted after a brief online registration process.

3. Derm7pt[39]: A dataset encompassing 1,011 dermoscopic images, including 252 melanoma and 759 nevi cases, complete with 7-point checklist criteria.

2.9.3 ISIC 2017 DataSet

In this work, we have used ISIC 2017 dataset for the classification for the classification of skin cancer. The dataset provided for this task contains 2000 images in *.jpg* format and with image name in the format $ISIC\_ < image\_id > .jpg$ where image id is a unique 7-digit number. The ISIC 2017 consist of 3 parts: 1) lesion segmentation; 2) lesion dermoscopic feature extraction; and 3) lesion classification. In this project, we focused mainly on lesion classification.

In lesion classification, the aim is to perform binary classification, which involves 3 types of lesion classes, which is given as follows:
a) Melanoma - Malignant skin tumour (melanocytic)
b) Nevus - Benign cancer (melanocytic)
c) Seborrheic keratosis - A common noncancerous (benign) skin growth (non-melanocytic)
Following are the two binary classification tasks:

1. Task 1: 'Melanoma' vs 'Nevus and Seborrheic Keratosis'
2. Task 2: 'Seborrheic Keratosis' vs 'Nevus and Melanoma'

The number of images in each class for Task 1 and Task 2 is illustrated in Fig.2.6

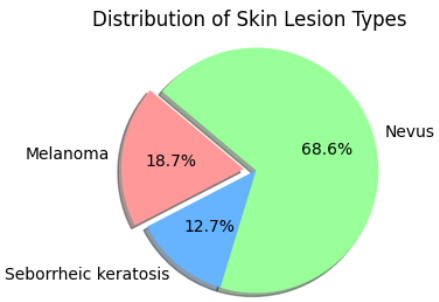


Figure 2.5: Data distribution - Train Dataset

((a)) Task 1

((b)) Task 2

Figure 2.6: Distribution of cases in each class for Task 1 and Task 2

Training Dataset

The training image consists of 2000 images, where 3 classes are distributed as follows: Melanoma - 374 images Seborrheic keratosis - 254 images; and Benign nevi - 1372 images. The distribution of the classes is illustrated in Fig. 2.5

Along with 2000 lesion images, a CSV file containing metadata of patients was also available, which contains the fields - $image\_id$, $age\_approximate$, sex. In addition to that, there is separate validation and test images are provided along with their labels. We have used training and validation datasets for model training and test data for final evaluation. Validation and test data contain 150 and 600 images respectively.

CHAPTER 3

Methodology

3.1 Problem Statement

In this experiment, we aim to implement an innovative technique that is capable of producing significant performance in automated Skin Cancer Lesion Classification. ViTs possess the unique capability to capture complex patterns and long-range dependencies within the images. By utilizing these features we aim to create a robust and more accurate skin cancer classification model that surpasses the performance of the current state-of-the-art model[17], thereby assisting dermatologists in the early and precise diagnosis of skin cancer lesions.

3.2 Exploratory Data Analysis

Exploratory Data Analysis(EDA) is the foundation of the investigation of data, which helps us to identify patterns, anomalies and deep insights into the data. This section deals with the EDA done as part of digging deeper into the characteristics of data.

The data used is mainly from metadata provided in the dataset in CSV format. Which contains information about the image id, approximate age and gender of the patient.

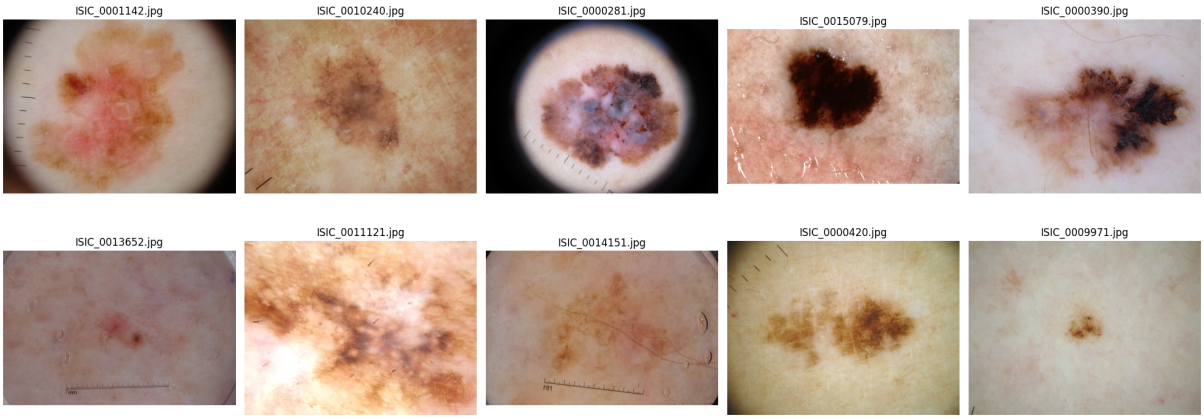


Figure 3.1: Melanoma Images

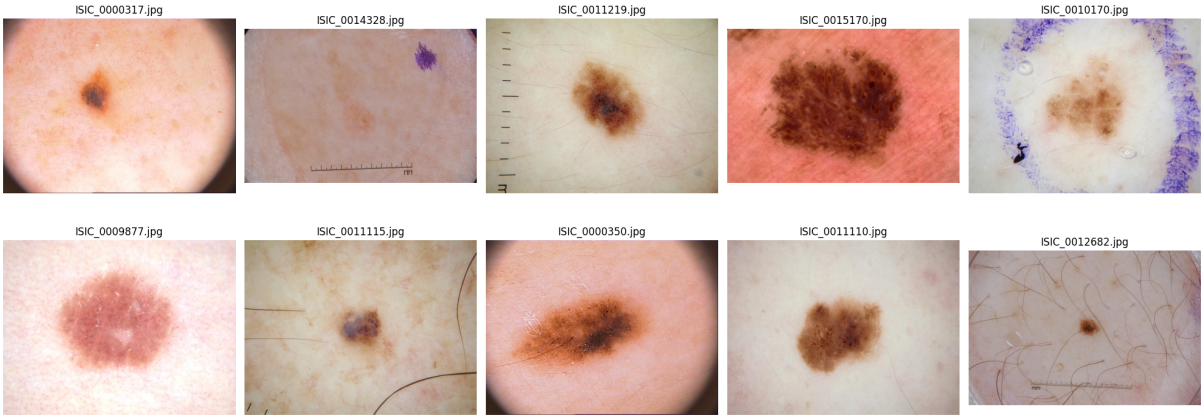


Figure 3.2: Non-Melanoma images

3.2.1 Melanoma and Non-Melanoma Images

Some of the melanoma and non-melanoma images has printed in Fig.3.1 and Fig. 3.2. From the figure we can compare the some of the important features that can be used to differentiate melanoma and non-melanoma images such as uneven distribution, asymmetrical shape, scalloped or notched borders, and uneven distribution of colour.

3.2.2 Age distribution

The age distribution across the train and test dataset has been illustrated in Fig.3.3. From the figure, we can see that the dataset contains images of people of all ages with most of them between the age group of 15 - 70, similarly, the test dataset also followed a similar trend. We had done some analysis to find the age group of melanoma cases which is illustrated in Fig. 3.4. From the figure, we can see that most of the melanoma

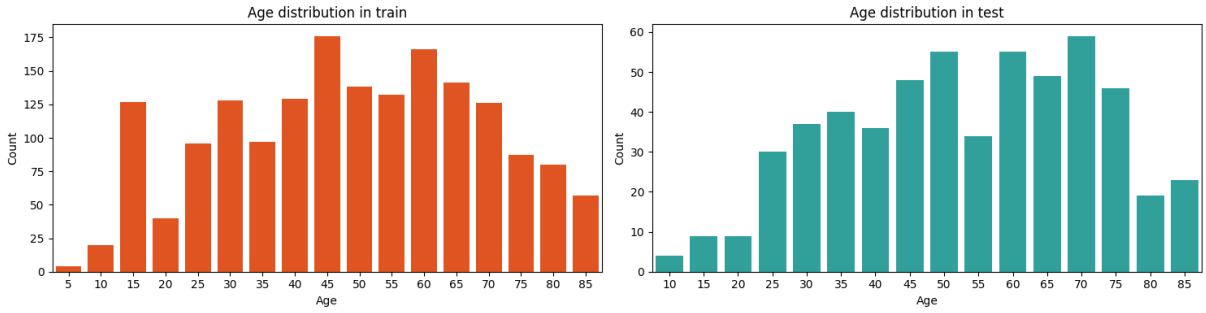


Figure 3.3: Age distribution in the Train and Test dataset

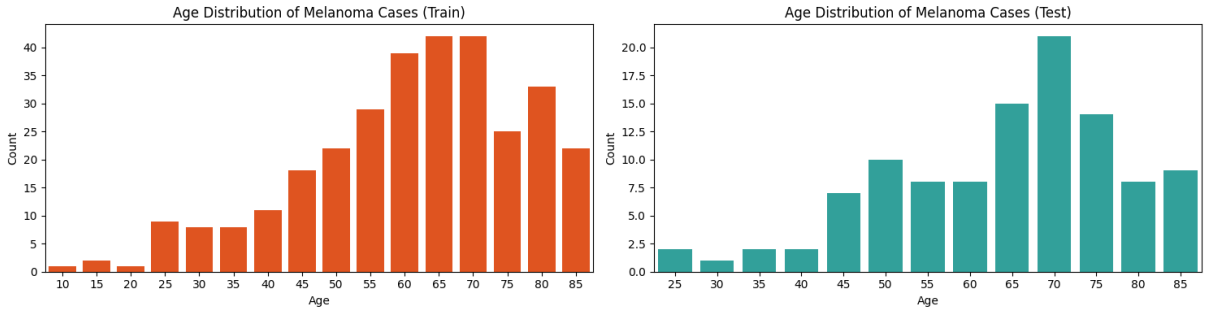


Figure 3.4: Age distribution of melanoma cases in Train and Test dataset

cases reported are between the age group of 45 - 85 and the reported cases are smaller in cases less than 40 years, but in the test dataset, we can see that most of the reported melanoma cases fall in the age group of 40 - 85 where the most number of cases reported for the age of 70

3.2.3 Gender distribution

The distribution of gender in the train and test dataset is illustrated in Fig. 3.5. From the figure, we can see that the number of male and female is equally distributed with about 200 genders unknown. A similar trend is followed in the test dataset as well. But when analysing the Fig. 3.6 we can see that number of melanoma cases reported is for males compared to females. From the figure we can see that number of females is half the number of males. A similar pattern is followed in the test dataset.

3.3 Image Preprocessing

We have tried different preprocessing steps to improve the performance of the model like hair removal (Section 3.6), resizing of image, data augmentation etc. Even though

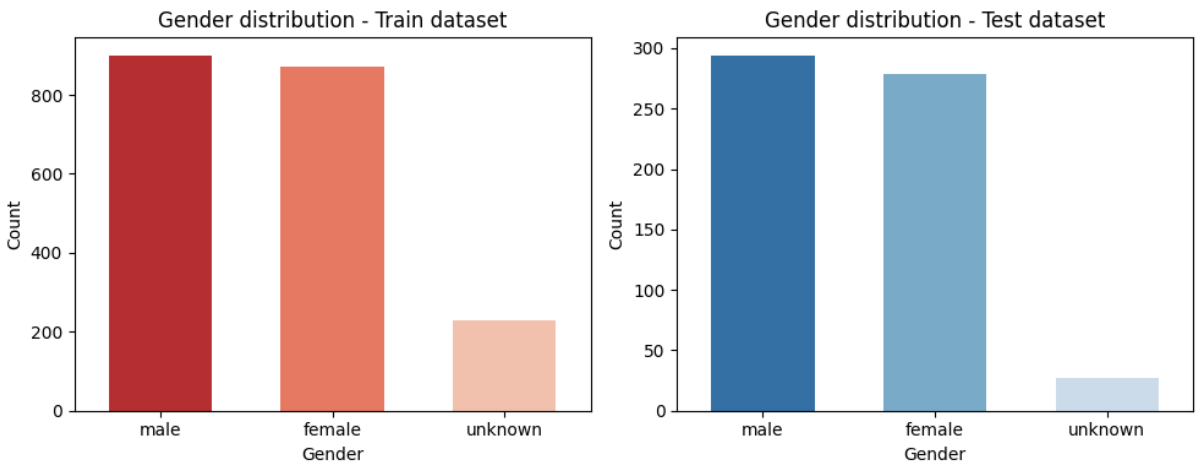


Figure 3.5: Gender distribution in Train and Test dataset

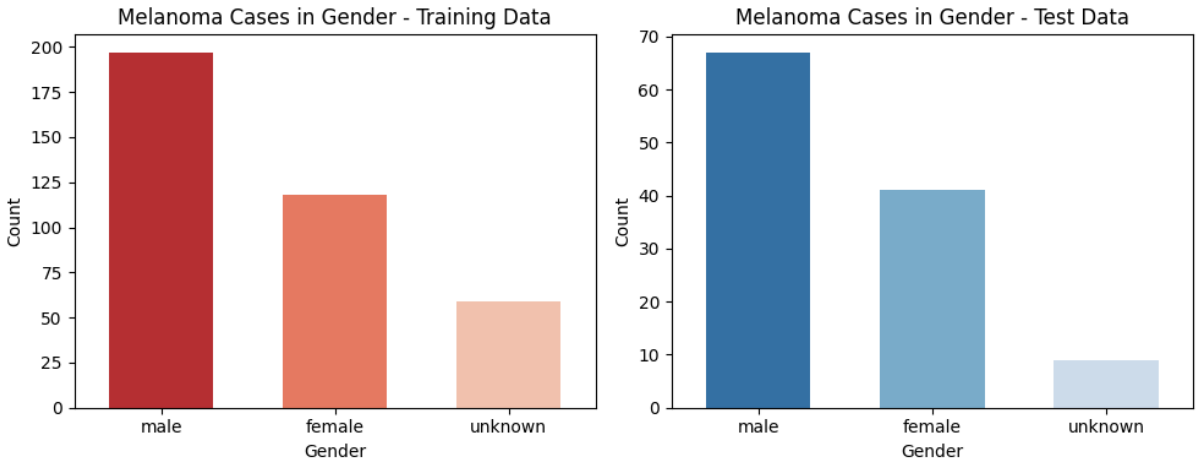


Figure 3.6: Gender distribution of melanoma cases in Train and Test dataset

hair removal shows better performance in Task 2, it doesn't produce any improvement in Task 1. So we used resizing and data augmentation as part of image preprocessing. Data augmentation has been explained in section 3.4.

we have resized the image to 224×224 pixels without reducing the quality of the image. After resizing, the computation time for training has been reduced significantly from 1.5 hours to 20 minutes. In order to maintain the quality of the image, we have used the Lanczos resampling filter method which is proposed by Vandame B [40] and Moraes et. al[41]. The Lanczos kernel defined in equation 3.1:

$$L(x) = \begin{cases} \text{sinc}(x) \cdot \text{sinc}\left(\frac{x}{a}\right) & \text{if } |x| \leq a \\ 0 & \text{otherwise} \end{cases} \tag{3.1}$$

where $a$is 'filter size' which is a user-defined parameter. The sinc(x) function is given by $\frac{\sin(\pi x)}{\pi x}$. The Lanczos resampling technique minimises artefacts that could result from scaling while preserving the visual integrity of the images. During training and inference, it makes sure that the images keep their quality and give accurate representations of the model.

3.4 Data Augmentation

In order to get performance, a large number of positive and negative samples are often required for training the model. As our dataset is relatively small, we have used data augmentation techniques to address this issue. A wide range of augmentation has been done to increase the dataset size.

The dataset was augmented using a variety of data augmentation approaches, each of which has increased the diversity of images. In order to avoid the possibility of non-uniform distortions, we have excluded non-uniform transformations like skewing and stretching. The following are augmentation techniques we have used in the experiment:

- rotation range=180
- width shift range=0.1
- height shift range=0.1
- zoom range=0.2
- horizontal flip=True
- vertical flip=True
- fill mode='nearest'

The above data augmentation techniques greatly increased the number of images in the dataset for each class, making it about five times bigger than it was before. We have intentionally kept the data imbalance in our dataset, as we are using focal loss as the loss function, which is known to produce better performance with imbalanced datasets.

3.5 Loss Functions

The choice of the loss function is very important while designing complex image segmentation as they instigate the learning process of the algorithm[42]. Since 2012 researchers have done various experiments in various domains to improve the results of their datasets. In our experiment, we used two loss functions and compared the performance both with a combination of batch normalization and layer normalization. The two loss functions we used include Binary cross Entropy loss and Focal loss.

3.5.1 Binary Cross Entropy Loss

Cross-entropy, sometimes abbreviated as CE, is a metric for measuring the difference between two probability distributions that correspond to a single random variable or a set of occurrences. It finds vast application in classification tasks, as it allows the evaluation of the dissimilarity between predicted and actual results. Given its suitability for pixel-level classification tasks like image segmentation, Cross-Entropy proves effective.

Binary Cross-Entropy (BCE) is a variant of the general Cross-Entropy loss and is particularly useful for binary classification scenarios[42]. It is represented by the formula:

$$\text{BCE}(y, \hat{y}) = -\left(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y})\right)$$

In the above equation, y depicts the ground truth value, while $\hat{y}$ represents the predicted result. This formula quantifies the divergence between the actual and predicted probabilities associated with binary classification outcomes.

3.5.2 Focal Loss

Focal Loss is generally used to handle the extreme imbalance issue between the two classes[43]. For the binary classification task, the focal loss starts from cross-entropy loss. The cross-entropy loss function is defined in the equation.3.2:

$$CE(p, y) = \begin{cases} -\log(p) & \text{if } y = 1, \\ -\log(1 - p) & \text{otherwise.} \end{cases} \tag{3.2}$$

In the above $y \in \{\pm 1\}$ specifies the ground-truth class and $p \in [0, 1]$ is the modelâs estimated probability for the class with the label $y = 1$. For notational convenience, we define p_t as an equation. 3.3:

$$p_t = \begin{cases} p & \text{if } y = 1, \\ 1 - p & \text{otherwise.} \end{cases} \tag{3.3}$$

We can rewrite the cross-entropy loss $CE(p, y)$ as $CE(p_t) = -\log(p_t)$. For handling the problem of imbalance, a new parameter α has been added in the above equation, which is termed as a balanced cross-entropy loss that is the base for focal loss. The equation 3.4 shown below defines the cross-entropy loss with p_t:

$$CE(p_t) = -\alpha_t \log(p_t). \tag{3.4}$$

For focal loss, [43] authors added new parameter γ to the above equation to handle extreme imbalance. In cross-entropy loss, negative samples dominate the gradient and contribute to the majority of the loss, leading to easier classification. α in the above equation for balanced cross-entropy balances the importance of positive and negative examples, it does not differentiate between easy or hard examples. As an alternative, they suggest reshaping the loss function to down-weight easy examples and concentrating training on difficult negatives by introducing γ in the above equation as follows 3.5::

$$FL(p_t) = -\alpha_t (1 - p_t)^{\gamma} \log(p_t) \tag{3.5}$$

where $-(1-p_t)^{\gamma}$ is the modulating factor, in which γ is a tunable focusing parameter and $\gamma \geq 0$.

3.6 Hair Removal Algorithm

The algorithm used OpenCV's image processing functions to construct a hair removal method. This method employs a sequential approach to eliminate structures that resemble hair from an image.

The input image is first transformed to grayscale. The processing necessary to identify hair-like features is made simpler by the fact that grayscale images only include

intensity information. The image is prepared for further processes by this conversion.

The morphological procedure that will be used later has a kernel defined next. In this instance, a kernel with 17x17 dimensions is employed. Small matrices known as kernels are used in image processing to conduct a variety of operations on the image. The selected kernel will have an impact on how hair structures in the image are identified.

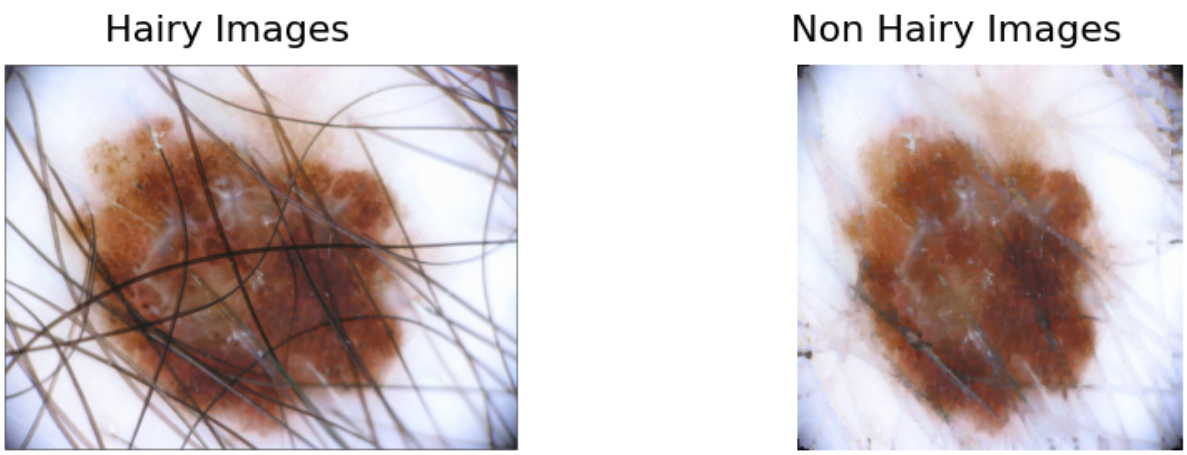
The morphological black-hat operation (cv2.MORPH_BLACKHAT), which forms the basis of the method, is used extensively. Morphological dilation and closure are combined in this procedure. Because of its ability to separate microscopic black structures from a bright background, it is ideal for accentuating structures that resemble hair. The resulting picture emphasises regions where there are hair-like features.

Thresholding is done on the output image after the black-hat process. A pixel's colour is set to 255 (white) for values greater than or equal to 10, and 0 (black) for values lower than 10. The areas where hair-like features are found are successfully identified in this binary thresholded image. Following that, inpainting is applied to these areas.

The thresholded binary image is used as a mask when applying the inpainting method (cv2.inpaint) to the original RGB image. Inpainting is a technique used to reconstruct damaged or missing portions of an image using the surrounding details. In this particular instance, the inpainting procedure reconstructs the regions where hair-like structures were discovered using data from the original image, thereby erasing the hair-like regions.

This method employs a number of image-processing processes to eliminate hair-like objects after receiving an RGB image as input. The method successfully removes hair by first applying morphological techniques to improve these structures, and then inpainting over them with the original picture data. The algorithm's effectiveness could be influenced by elements like image quality and the properties of the hair-like features.

Hair removal has been implemented in both Task 1 and Task 2, we got an improvement in accuracy in Task 2 but it doesn't give any performance improvement in Task 1. After analysis, we came to the conclusion that this algorithm can reduce the image quality and is prone to removal of some important features of the images that is important for the classification of melanoma images.



((a)) Before Hair Removal ((b)) After Hair Removal

Figure 3.7: Resultant images before and after hair removal

3.7 Batch Normalization vs Layer Normalization

3.7.1 Batch Normalization

Batch normalisation is a normalisation method used on the inputs of every layer in a neural network. It computes the mean and variance of the inputs throughout the mini-batch during training and then uses these statistics to normalise the inputs. Additionally, the normalised data are scaled and shifted using learnable parameters (gamma and beta), enabling the network to modify the normalisation to suit its requirements. Batch normalisation, which is frequently used before to the activation function, can assist in stabilising and accelerating training by lowering internal covariate shift, or the alteration in the distribution of network activations brought on by parameter updates. It also has a regularisation impact, which lessens the need for additional regularisation strategies like dropout. However, Batch Normalisation has significant drawbacks, such as its reliance on batch size and potential inefficiency with very tiny batches. In addition, when used with recurrent networks, it can cause difficulties.

3.7.2 Layer Normalization

Layer normalisation is another method used to standardise the inputs to each layer of a neural network. But Layer Normalisation computes statistics across all features (units)

in a single data sample, as opposed to doing so across a mini-batch. It employs learnable parameters for scaling and moving the normalised values, similar to batch normalisation. Ordinarily, layer normalisation is used either before or after the activation function. When compared to Batch Normalisation, it has the benefit of being less sensitive to changes in batch size. Layer Normalisation is hence appropriate in situations where modest batch sizes are employed. Furthermore, Layer Normalisation is better able to tolerate fluctuations in spatial dimensions, which is important for systems like Vision Transformers (ViTs) that operate on image patches with different spatial dimensions. Additionally, it does not experience the "train-test discrepancy" that Batch Normalisation can and has a tendency to be more consistent during training. Due to its adaptability, layer normalisation is also preferable in situations requiring repeated elements or self-attentional mechanisms.

3.7.3 Batch Normalization vs Layer Normalization

The decision between Batch Normalisation and Layer Normalisation depends on the design in the context of Vision Transformers, such as ViT B16. ViTs transform image patches into different spatial dimensions as a consequence. The calculation of statistics for Batch Normalisation may be complicated by this difference in geographic dimensions. It is more suited for ViTs because Layer Normalisation can normalise across features in a patch. Additionally, Layer Normalisation is typically favoured since it is compatible with recurrent structures and self-attention mechanisms if the ViT B16 architecture has them. Overall, Layer Normalisation is frequently chosen in ViTs because it is better suited to these particular designs due to its capacity to manage fluctuations in spatial dimensions, suitability for small batch sizes, and stability during training.

3.8 Proposed System

The Vision Transformer (ViT) model specifically the ViT-B16 variation is used in the proposed system for classifying skin cancer lesions. The skin cancer detection system block diagram is illustrated in Fig. 3.8. By incorporating self-attention processes, the ViT architecture pioneers a fresh method of image classification and equips the model to recognise intricate patterns and distant connections in the input images. The ViT

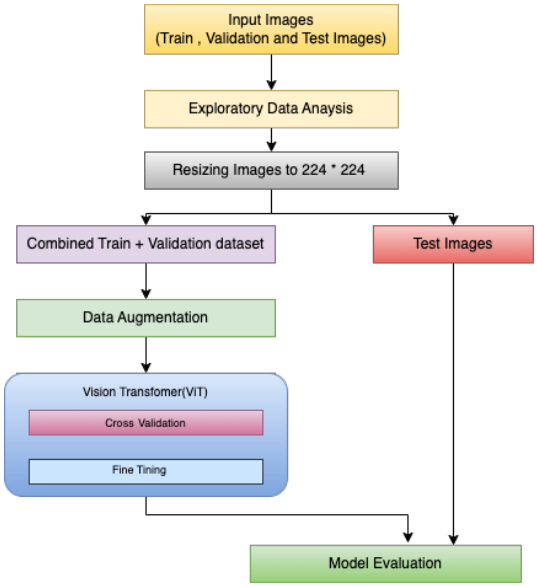


Figure 3.8: Skin Cancer Detection System Block Diagram

architecture is illustrated in Fig. 3.9, which processes skin cancer input images by first dividing them into fixed-size patches (16*16). These patches are then linearly projected into high-dimensional vectors, capturing local image details. Positional embeddings are added to provide spatial information, and the patch embeddings are fed into a stack of transformer encoder layers. These layers use multi-head self-attention to capture global relationships and pass the embeddings through feed-forward neural networks for non-linearity. The final transformer encoder layer's output is globally averaged and passed through a fully connected layer with a softmax activation to obtain class probabilities for the input image. The foundation of the classification architecture in our implementation was the pre-trained ViT-B16 model. The pre-trained model was modified with additional layers to make it specifically suitable for the task of classifying skin cancer lesions. These layers included a flattened layer, layer normalization, drop out and dense layers. A 1-dimensional feature vector was created from the output of the ViT model using the flattened layer. The activations were normalised using layer normalisation, which also lessened the impact of internal covariate shift. The final classifier, the dense layers, assigned the collected features to the various subclasses of skin lesions (i.e., melanoma and non-melanoma).

The proposed architecture was trained using the Adam optimizer and optimized for focal loss. During training, the training data were fed to the model in batches, utilizing data generators for efficient memory utilization. Early stopping was employed

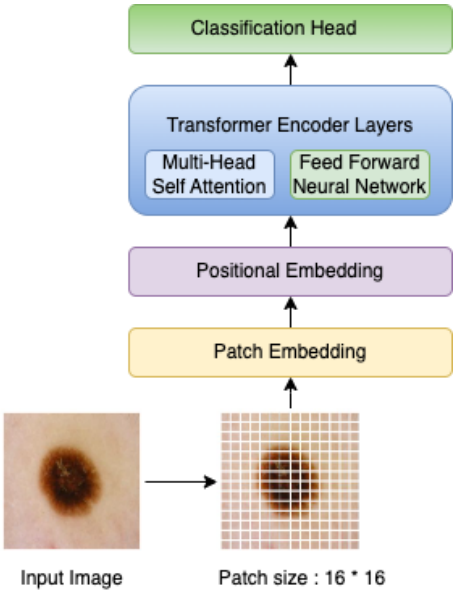


Figure 3.9: General ViT architecture used for skin cancer classification.

to prevent overfitting, while a learning rate scheduler dynamically adjusted the learning rate throughout the training process. The suggested model illustrates its capability to accurately categorise skin cancer lesions through this architecture. Beyond the capabilities of traditional CNNs, the integration of the ViTs enables the capture of complex patterns and long-range dependencies. The proposed architecture is illustrated in Fig. 3.10. Data preprocessing for this experiment includes resizing images into fixed size, which is 224 * 224 in our case. Initially, we trained the model without resizing, but it took a minimum of 90 minutes to complete the training, but after resizing the computation time was reduced to 20 minutes. Data augmentation is the second preprocessing step we have implemented. In the data augmentation step, we excluded techniques such as stretching and skewing as this may cause a change in the shape of images which is an important feature classifying melanoma images, as the asymmetric border is a key feature in classifying melanoma images. We implemented hair removal and trained the image without hair. However there was no significant improvement in the result, so we trained without reducing the image quality.

3.9 Hardware and Software Requirements

The hardware used in the experiments includes an NVIDIA Tesla T4 GPU available on Google Colab with the following specifications: driver version 525.105.17 and CUDA

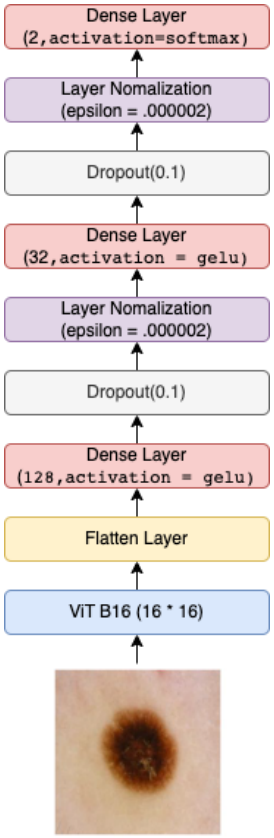


Figure 3.10: Proposed 9-layer ViT Model Architecture with layer normalization and regularization

version 12.0. The GPU has a memory capacity of 15,360MiB. We implemented our model using ViT-Keras 0.1.2 with TensorFlow version 2.12.0 as the backend, along with CUDA 12.0 for GPU acceleration. Our system is implemented with Python 3.10.6 on Google Colab.

3.10 Data Partitioning

The training dataset consists of 2000 images, where 3 classes are distributed as follows: Melanoma - 374 images; Seborrheic keratosis - 254 images; and Benign nevi - 1372 images. The class distribution is illustrated in Fig. 2.5. In addition to training data, separate validation (150 images) and test (600 images) are available along with their labels. In our study, we considered two different model evaluation settings: 1) S1: cross-validation (CV) and 2) S2: normal. Under each evaluation set, we have further divided it into two different data partitioning schemes: a) SS1: partitioned the training set into 5-folds and performed 5-fold CV and evaluated the model performance on the

test set; b) SS2: merged the training and validation data then partitioned in into 5-folds and performed 5-fold CV and evaluated the model on the test set.

3.11 Model Training

We used the Adam Optimizer to train the model with a learning rate of 0.00002 without exponential decay, and default values for hyperparameters $\beta 1$ and $\beta 2$, which regulate the exponential moving averages of the gradients and squared gradients, respectively. The model has been trained for 30 epochs with early stopping criteria. Early stopping criteria were configured with a patience of 8. After training model evaluation was carried out on the unseen test dataset. As suggested by [44], we chose γ = 2 and α=.7 for handling the imbalance for focus loss. .

3.12 Model Testing and Evaluation

For setting S1, 5-fold cross-validation (CV) was done for different combinations of binary classification with and without DA under two different data partitioning schemes (i.e., SS1 and SS2), given in Table. 4.1 and Table. 4.2. For setting S2, normal training and test partitioning were done for different combinations of binary classification with and without DA under two different data partitioning schemes (i.e., SS1 and SS2), given in Table. 4.3 and Table. 4.4.

CHAPTER 4

Result

4.1 Evaluation matrices

In both setting S1 and S2, the test set (unseen data) was used to evaluate the model's performance. Due to the imbalance in the data, we have used a range of evaluation metrics such as recall, sensitivity, accuracy, and AUC-ROC score. Additionally, in the skin cancer detection model, detecting false negatives and achieving high recall scores is vital to ensure early identification of potential cases. This enhances the model's ability to capture a significant portion of actual positive cases, minimizing the risk of missing critical diagnoses. Furthermore, we also calculated, the AUC-ROC score as it quantifies the model's ability to distinguish between benign and malignant cases across various classification thresholds. A high AUC-ROC indicates effective discrimination, aiding in the selection of optimal models and ensuring accurate diagnostic outcomes.

The equations for accuracy, recall, and specificity are given below 4.1:

$$\begin{aligned} \text{accuracy} &= \frac{TP + TN}{TP + TN + FP + FN}, \\ \text{sensitivity/recall} &= \frac{TP}{TP + FN} \end{aligned} \tag{4.1}$$

Table 4.1: Test dataset results for data partitioning scheme SS1 under setting S1 (i.e., training with 5-fold CV)

Model Name	Layer Normalization + FL				Batch Normalization + BCE				Batch Normalization + FL			
	Recall	Precision	Acc	AUC	Recall	Precision	Acc	AUC	Recall	Precision	Acc	AUC
Nev vs Seb (DA)	**0.90**	**0.91**	**0.90**	**0.94**	0.87	0.89	0.87	0.92	0.74	0.85	0.74	0.86
Nev vs Seb	**0.87**	0.87	**0.87**	**0.89**	0.75	0.85	0.74	0.86	**0.87**	**0.92**	0.75	0.81
Seb vs [Mel & Nev] (DA)	**0.89**	**0.89**	**0.89**	**0.89**	0.81	0.89	0.81	0.89	**0.89**	**0.89**	**0.89**	**0.89**
Seb vs [Mel & Nev]	**0.86**	**0.87**	**0.84**	0.85	0.68	0.86	0.68	0.85	0.708	0.86	0.708	0.84

Table 4.2: Test dataset results for data partitioning scheme SS2 under setting S1 (i.e., training with 5-fold CV)

Model Name	Layer Normalization + FL				Batch Normalization + BCE				Batch Normalization + FL			
	Recall	Precision	Acc	AUC	Recall	Precision	Acc	AUC	Recall	Precision	Acc	AUC
Mel vs [Nev & Seb] (DA)	**0.83**	0.82	**0.83**	**0.83**	**0.83**	**0.83**	**0.83**	0.81	0.79	0.81	0.80	0.76
Mel vs [Nev & Seb]	**0.83**	**0.80**	**0.81**	**0.78**	0.76	0.78	0.76	0.74	0.55	0.73	0.55	0.61
Seb vs [Mel & Nev] (DA)	**0.87**	**0.88**	**0.87**	**0.88**	0.79	0.88	0.79	0.87	0.81	0.85	0.81	0.81
Seb vs [Mel & Nev]	**0.87**	**0.88**	**0.87**	**0.88**	0.80	0.83	0.80	0.77	0.80	0.83	0.80	0.77

4.2 Experiment Results

We have conducted two sets of experiments S1 and S2, for both experiments, we used data partition schemes SS1 and SS2 as described in section 3.6. Setting S1 (i.e., 5-fold CV) results are presented in Table. 4.1 and Table. 4.2. Setting S2 (i.e., normal training test partitioning) results are presented in Table. 4.3 and Table. 4.4.

In the case of setting S1 under data partitioning scheme SS1, the results are given in Table.4.1, ViT with layer normalization plus focal loss provided the best recall score in three pairwise binary classifications (i.e., Nev vs Seb with DA is 0.90, Nev vs Seb without DA is 0.87, and Seb vs [Mel & Nev] with DA is 0.89. In the case of setting S1 under data partitioning scheme SS2, the results are given in Table.4.2, a similar pattern was observed, where ViT with layer normalization plus focal loss provided the best recall score in all four pairwise binary classifications (i.e., Nev vs Seb with DA is 0.83, Nev vs Seb without DA is 0.83, Seb vs [Mel & Nev] with DA is 0.87, and Seb vs [Mel & Nev] is 0.87.

In the case of setting S2 under data partitioning scheme SS1, the results are given

Table 4.3: Comparison with state-of-the-art results on test dataset for data partitioning scheme SS1 under setting S2 (i.e., training on whole training data), where DA - Data Augmentation, FL - Focal Loss, BCE - Binary Cross Entropy

Model Name	Nevus VS Seb				[Seb] vs [Melanoma & Nevus]			
	Sens	Spec	ACC	AUC	Sens	Spec	ACC	AUC
IRV2 (SA)	0.95	0.71	0.90	0.94	0.93	0.69	0.90	0.94
ViT B16 (BN)*	0.93	0.38	0.72	0.83	0.69	**0.83**	0.71	0.86
ViT B16 (LN)*	0.93	0.58	0.88	0.86	**0.93**	0.58	0.88	0.86
ViT B16 (BN + BCE + DA)*	0.95	0.68	0.89	0.68	0.94	0.76	**0.91**	0.91
ViT B16 (LN + BCE + DA)*	0.94	0.71	0.90	0.90	0.95	0.67	**0.91**	0.91
ViT B16 (BN + FL + DA)*	**0.95**	0.68	0.89	0.93	0.92	0.70	0.89	0.89
ViT B16 (LN + FL + DA)	0.93	**0.84**	**0.92**	**0.95**	**0.96**	0.59	**0.91**	0.93

Table 4.4: Comparison with state-of-the-art results on test dataset for data partitioning scheme SS2 under setting S2 (i.e., training on whole training data), where DA - Data Augmentation, FL - Focal Loss, BCE - Binary Cross Entropy

Model Name	[Melanoma] vs [Nevus & Seb]				[Seb] vs [Melanoma & Nevus]			
	Sens	Spec	ACC	AUC	Sens	Spec	ACC	AUC
ResNet50	0.63	0.89	0.84	0.86	0.87	0.84	0.84	0.95
RAN50 2	0.62	0.91	0.85	0.85	0.88	0.86	0.86	0.94
SEnet50 3	0.62	0.90	0.85	0.86	0.86	0.87	0.86	0.95
ARL-CNN	0.66	0.89	0.85	0.88	0.87	0.87	0.87	0.96
ViT B16 (BN)*	0.42	0.87	0.77	0.74	0.77	0.72	0.77	0.81
ViT B16 (LN)*	0.61	0.86	0.83	0.78	0.91	0.57	0.86	0.87
ViT B16 (BN + BCE + DA)*	0.56	0.88	0.82	0.75	0.88	0.8	0.87	0.89
ViT B16 (LN + BCE + DA)*	0.60	0.88	0.84	0.82	0.89	0.86	0.88	0.93
ViT B16 (BN + FL + DA)*	0.51	0.90	0.81	0.81	0.89	0.78	0.88	0.91
ViT B16 (LN + FL + DA)*	**0.77**	0.37	**0.86**	0.83	**0.96**	0.62	**0.91**	0.92

in Table.4.3, where simple training and test partitioning were done and compared the result of current state-of-the-art[17]. Evaluating nevus and Seborrheic Keratosis, we achieved a state-of-the-art recall score, with a performance improvement of 2% weighted recall and 1% AUC score, and also got higher specificity of 84% as well. In the case of Seb vs [Melanoma & Neves], we achieved a 0.96 recall score which is higher than the results obtained in [17].

In the case of setting S2 under data partitioning scheme SS2, the results are given in Table.4.4, where simple training and test partitioning were done and compared the result with [45] ARL-CNN, SEnet, ResNet, RAN14, where they have used the original tasks in ISIC 2017 challenge. We achieved higher sensitivity and accuracy than ARL-CNN in both tasks (i.e., [Melanoma] vs [Nevus & Seb] and [Seb] vs [Melanoma & Nevus. In [Melanoma] vs [Nevus & Seb], we achieved a higher sensitivity of 77% and accuracy of 86%, which is higher than all the other state-of-the-art models presented in the Table. 4.4. In task 2 (i.e., Seb vs [Melanoma & Neves]), we achieved a sensitivity of 96% and an accuracy of 91 % which is higher than ARL-CNN [45] by 8% in terms of sensitivity and 4% in terms of accuracy.

CHAPTER 5

Discussion

Various experiments have been conducted to evaluate the performance of ViTs in the field of skin cancer lesion classification. We have explored various predictive models, such as CNN, CNN with soft attention [17], the basic ViT architecture, ViT integrated with batch normalisation [30], and a version of ViT enhanced with layer normalisation and additional regularisation. The combination of the ViT with layer normalization and regularisation methods pro- reduced the most significant results.

In order to improve the performance of the model, various preprocessing techniques were investigated. To achieve optimal results various preprocessing steps have been experimented such as hair removal, image resizing, and data augmentation techniques. Surprisingly, hair removal didn't produce the expected performance. It performed well in Task 2 but failed to produce better results in Task 2. The algorithm used for hair removal is explained in section 3.5. In order to speed up the training process, images were resized to 224x224 dimensions, using the Lanczos kernel method. Lanczos kernel method helped in maintaining the picture quality. Data augmentation methods have been utilized to augment the number of images within each class. Given focal loss as the chosen loss function, we intentionally preserved class imbalance while applying augmentation, resulting in a fivefold increase in the number of images per class.

A rigorous five-fold cross-validation was performed across all scenarios. In each fold, ViTs with layer normalization and regularization consistently delivered excellent performance. Even though the batch normalization produced good results in 1 or 2 folds, it failed to produce consistent results across each fold leading to poor results while

considering the average results. Also, we have noticed that data augmentation improves the performance of the model compared to the model without it. This underscores the capability of focal loss to manage larger imbalances compared to binary cross-entropy loss.

Some of the challenges in model performance include interclass similarity and intra-class dissimilarity. Melanoma and benign nevi, for example, share visual characteristics, making it difficult to distinguish between them especially when training artificial intelligence using smaller datasets. These similarity issues are also problems in diagnosing results in earlier stages of the above example. Innovative approaches are required to overcome this problem, which may include enhanced data augmentation methods and hybrid models.

Dataset size is another challenge faced in improving the performance of the model. Moving forward the viability of the ViT model on larger and more varied datasets must be investigated which can result in numerous advantages. The broader range of variables that larger datasets provide allows models like ViT to extract more complex patterns and perform better. Larger data repositories offer the possibility of improved feature extraction and performance, especially given the limitations of the ISIC 2017 dataset.

In future, my suggestion is to incorporate all the above-mentioned challenges while implementing the skin cancer classification tasks, so that it will help to improve the performance as well as the usability of the model across diverse skin tones.

Furthermore, there is a noticeable disparity in death rates among people with darker skin tones, which is frequently related to delayed diagnoses [1]. Even experienced dermatologists struggle to identify diseases on darker skin, which causes concern within the medical community. Also, the datasets for these images are not easily available to implement a machine learning model. So there is a high requirement for the implementation of the advanced model to classify the skin cancer lesion which can easily classify skin cancer in diverse skin tones.

CHAPTER 6

Conclusions

In this research, we are introducing a cutting-edge 9-layer ViT model created especially for the task of classifying skin cancer lesions. Our approach is based on the fundamental ViT architecture which is strengthened by the 8 additional customised layers. The imbalance in dataset distribution is one of the major challenges in the classification, which has been solved by integrating focal loss as the loss function. We have blended the basic vision transformer architecture with advanced regularization techniques and dense layers. Our model exhibits a collaborative strength that leads to a reliable and precise classification.

After exhaustive research and analysis with different combinations of task settings as mentioned in section 3.7, we came to the conclusion that our proposed ViT model surpassed the benchmark set by the current state-of-the-art [17] model in terms of sensitivity, weighted recall, accuracy in the binary classification task of ISIC 2017 challenge. The model outperforms the previous state-of-the-art approach [17] with accuracy ratings of 91% and 92% in tasks 1 and 2, respectively, of binary classification. This represents an impressive improvement of 1% and 2%. 5-fold Cross-validation has been implemented in all the scenarios and our ViT with layer normalization and additional regularization shows a consistent performance in all folds, while batch normalization struggled to get consistent results even though got good results in one or two folds.

In conclusion, our novel 9-layer ViT model represents a significant advancement in the field of medical picture classification. Its integration of specialised layers, innovative architecture, and focused loss function not only improves efficiency but also has the abil-

ity to raise the bar for precise medical picture categorization. Our study highlights the revolutionary potential of incorporating cutting-edge deep learning methodologies into medical diagnostics, pointing the way towards more accurate and reliable classification systems for skin cancer lesions.

APPENDIX

Appendix

Download the PDF file: Click here

APPENDIX B

Another Appendix

Bibliography

[1] H. M. Gloster Jr and K. Neal, "Skin cancer in skin of color," *Journal of the American Academy of Dermatology*, vol. 55, no. 5, pp. 741–760, 2006.

[2] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Transactions on knowledge and data engineering*, vol. 22, no. 10, pp. 1345–1359, 2009.

[3] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," *ACM computing surveys (CSUR)*, vol. 54, no. 10s, pp. 1–41, 2022.

[4] G. Yang, S. Luo, and P. Greer, "A novel vision transformer model for skin cancer classification," *Neural Processing Letters*, pp. 1–17, 2023.

[5] D. L. Narayanan, R. N. Saladi, and J. L. Fox, "Ultraviolet radiation and skin cancer," *International journal of dermatology*, vol. 49, no. 9, pp. 978–986, 2010.

[6] P. T. Bradford, "Skin cancer in skin of color," *Dermatology nursing/Dermatology Nurses' Association*, vol. 21, no. 4, p. 170, 2009.

[7] M. Goyal, T. Knackstedt, S. Yan, and S. Hassanpour, "Artificial intelligence-based image classification methods for diagnosis of skin cancer: Challenges and opportunities," *Computers in biology and medicine*, vol. 127, p. 104065, 2020.

[8] K. Ali, Z. A. Shaikh, A. A. Khan, and A. A. Laghari, "Multiclass skin cancer classification using efficientnets–a first step towards preventing skin cancer," *Neuroscience Informatics*, vol. 2, no. 4, p. 100034, 2022.

[9] A. Budhiman, S. Suyanto, and A. Arifianto, "Melanoma cancer classification using resnet with data augmentation," in *2019 International Seminar on Research of Information Technology and Intelligent Systems (ISRITI)*. IEEE, 2019, pp. 17–20.

[10] J.-A. Almaraz-Damian, V. Ponomaryov, S. Sadovnychiy, and H. Castillejos-Fernandez, "Melanoma and nevus skin lesion classification using handcraft and deep learning feature fusion via mutual information measures," *Entropy*, vol. 22, no. 4, p. 484, 2020.

[11] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun, "Dermatologist-level classification of skin cancer with deep neural networks," *nature*, vol. 542, no. 7639, pp. 115–118, 2017.

[12] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[13] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[14] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[15] Y. Gulzar and S. A. Khan, "Skin lesion segmentation based on vision transformers and convolutional neural networksâa comparative study," *Applied Sciences*, vol. 12, no. 12, p. 5990, 2022.

[16] M. Berseth, "Isic 2017-skin lesion analysis towards melanoma detection," *arXiv preprint arXiv:1703.00523*, 2017.

[17] S. K. Datta, M. A. Shaikh, S. N. Srihari, and M. Gao, "Soft attention improves skin cancer classification performance," in *Interpretability of Machine Intelligence in Medical Image Computing, and Topological Data Analysis and Its Applications for Medical Data: 4th International Workshop, iMIMIC 2021, and 1st International Workshop, TDA4MedicalData 2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 4*. Springer, 2021, pp. 13–23.

[18] K. Han, Y. Wang, H. Chen, X. Chen, J. Guo, Z. Liu, Y. Tang, A. Xiao, C. Xu, Y. Xu *et al.*, "A survey on vision transformer," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 1, pp. 87–110, 2022.

[19] E. Craythorne and F. Al-Niami, "Skin cancer," *Medicine*, vol. 45, no. 7, pp. 431–434, 2017.

[20] A. F. Jerant, J. T. Johnson, C. D. Sheridan, and T. J. Caffrey, "Early detection and treatment of skin cancer," *American family physician*, vol. 62, no. 2, pp. 357–368, 2000.

[21] D. N. Dorrell and L. C. Strowd, "Skin cancer detection technology," *Dermatologic clinics*, vol. 37, no. 4, pp. 527–536, 2019.

[22] V. Narayanamurthy, P. Padmapriya, A. Noorasafrin, B. Pooja, K. Hema, K. Nithyakalyani, F. Samsuri *et al.*, "Skin cancer detection using non-invasive techniques," *RSC advances*, vol. 8, no. 49, pp. 28 095–28 130, 2018.

[23] S. Jain, N. Pise *et al.*, "Computer aided melanoma skin cancer detection using image processing," *Procedia Computer Science*, vol. 48, pp. 735–740, 2015.

[24] T. Saba, "Computer vision for microscopic skin cancer diagnosis using handcrafted and non-handcrafted features," *Microscopy Research and Technique*, vol. 84, no. 6, pp. 1272–1283, 2021.

[25] Y. N. Fuâadah, N. C. Pratiwi, M. A. Pramudito, and N. Ibrahim, "Convolutional neural network (cnn) for automatic skin cancer classification system," in *IOP conference series: materials science and engineering*, vol. 982, no. 1. IOP Publishing, 2020, p. 012005.

[26] R. R. Subramanian, D. Achuth, P. S. Kumar, K. N. kumar Reddy, S. Amara, and A. S. Chowdary, "Skin cancer classification using convolutional neural networks," in *2021 11th International Conference on Cloud Computing, Data Science & Engineering (Confluence)*. IEEE, 2021, pp. 13–19.

[27] F. Zhuang, Z. Qi, K. Duan, D. Xi, Y. Zhu, H. Zhu, H. Xiong, and Q. He, "A comprehensive survey on transfer learning," *Proceedings of the IEEE*, vol. 109, no. 1, pp. 43–76, 2020.

[28] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, no. 1, pp. 1–40, 2016.

[29] U.-O. Dorj, K.-K. Lee, J.-Y. Choi, and M. Lee, "The skin cancer classification using deep convolutional neural network," *Multimedia Tools and Applications*, vol. 77, pp. 9909–9924, 2018.

[30] C. Xin, Z. Liu, K. Zhao, L. Miao, Y. Ma, X. Zhu, Q. Zhou, S. Wang, L. Li, F. Yang *et al.*, "An improved transformer network for skin cancer classification," *Computers in Biology and Medicine*, vol. 149, p. 105939, 2022.

[31] N. C. Codella, D. Gutman, M. E. Celebi, B. Helba, M. A. Marchetti, S. W. Dusza, A. Kalloo, K. Liopyris, N. Mishra, H. Kittler *et al.*, "Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic)," in *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*. IEEE, 2018, pp. 168–172.

[32] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data*, vol. 5, no. 1, pp. 1–9, 2018.

[33] M. Combalia, N. C. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, C. Carrera, A. Barreiro, A. C. Halpern, S. Puig *et al.*, "Bcn20000: Dermoscopic lesions in the wild," *arXiv preprint arXiv:1908.02288*, 2019.

[34] I. Giotis, N. Molders, S. Land, M. Biehl, M. F. Jonkman, and N. Petkov, "Med-node: A computer-assisted melanoma diagnosis system using non-dermoscopic images," *Expert systems with applications*, vol. 42, no. 19, pp. 6578–6585, 2015.

[35] S. S. Han, M. S. Kim, W. Lim, G. H. Park, I. Park, and S. E. Chang, "Classification of the clinical images for benign and malignant cutaneous tumors using a deep learning algorithm," *Journal of Investigative Dermatology*, vol. 138, no. 7, pp. 1529–1538, 2018.

[36] J. N. Weinstein, E. A. Collisson, G. B. Mills, K. R. Shaw, B. A. Ozenberger, K. Ellrott, I. Shmulevich, C. Sander, and J. M. Stuart, "The cancer genome atlas pan-cancer analysis project," *Nature genetics*, vol. 45, no. 10, pp. 1113–1120, 2013.

[37] R. Fisher and J. Rees, "Dermofit: A cognitive prosthesis to aid focal skin lesion diagnosis," 2012.

[38] T. Mendonça, P. M. Ferreira, J. S. Marques, A. R. Marcal, and J. Rozeira, "Ph 2-a dermoscopic image database for research and benchmarking," in *2013 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)*. IEEE, 2013, pp. 5437–5440.

[39] J. Kawahara, S. Daneshvar, G. Argenziano, and G. Hamarneh, "Seven-point checklist and skin lesion classification using multitask multimodal neural nets," *IEEE journal of biomedical and health informatics*, vol. 23, no. 2, pp. 538–546, 2018.

[40] B. Vandame, "Fast and efficient resampling for multi-frame super-resolution," in *21st European Signal Processing Conference (EUSIPCO 2013)*. IEEE, 2013, pp. 1–5.

[41] T. Moraes, P. Amorim, J. V. Da Silva, and H. Pedrini, "Medical image interpolation based on 3d lanczos filtering," *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, vol. 8, no. 3, pp. 294–300, 2020.

[42] S. Jadon, "A survey of loss functions for semantic segmentation," in *2020 IEEE conference on computational intelligence in bioinformatics and computational biology (CIBCB)*. IEEE, 2020, pp. 1–7.

[43] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.

[44] G. S. Tran, T. P. Nghiem, V. T. Nguyen, C. M. Luong, J.-C. Burie *et al.*, "Improving accuracy of lung nodule classification using deep learning with focal loss," *Journal of healthcare engineering*, vol. 2019, 2019.

[45] J. Zhang, Y. Xie, Y. Xia, and C. Shen, "Attention residual learning for skin lesion classification," *IEEE transactions on medical imaging*, vol. 38, no. 9, pp. 2092–2103, 2019.