SKINNET-14: A FINE-TUNED CCT MODEL FOR CLASSIFYING SKIN CANCER ADDRESSING COMPUTATIONAL COMPLEXITY AND TRAINING TIME

BY

ABDULLAH AL MAHMUD
ID: 191-15-2527

AND

INAM ULLAH KHAN
ID: 191-15-2575

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering

Supervised By

Dr. S. M. Aminul Haque
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised By

Dr. Md Zahid Hasan
Associate Professor
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

JANUARY 2023

APPROVAL

This Project/internship titled **"SkinNet-14: A fine-tuned CCT model for classifying skin cancer addressing computational complexity and training time"**, submitted by Abdullah Al Mahmud ID No: 191-15-2527 and Inam Ullah Khan, ID No: 191-15-2575 to the Department of Computer Science and Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfilment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on *30/01/2023*.

BOARD OF EXAMINERS

Chairman

Dr. Touhid Bhuiyan
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

SArefin

Dr. Mohammad Shamsul Arefin
Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Examiner

Ms. Sharmin Akter
Lecturer (Senior Scale)
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Examiner

Dr. Mohammad Shorif Uddin
Professor
Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

We hereby declare that, this study was completed under the direction of **Dr. S. M. Aminul Haque, Associate Professor,** Department of Computer Science and Engineering, Daffodil International University. We further declare that neither the entire project nor any portion of it has been submitted elsewhere for a degree or certificate.

Supervised by:

Dr. S. M. Aminul Haque
Associate Professor
Department of CSE
Daffodil International University

Co-Supervised by:

Dr. Md Zahid Hasan
Associate Professor
Department of CSE
Daffodil International University

Submitted by:

Mahmud

Abdullah Al Mahmud
ID: 191-15-2527
Department of CSE
Daffodil International University

Inam

Inam Ullah Khan
ID: 191-15-2575
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our deepest appreciation and gratitude to the all-powerful God, whose divine favor enables us to successfully finish the senior project/internship.

We really grateful and wish our profound our indebtedness to **Dr. S. M. Aminul Haque, Associate Professor and Associate Head and Dr. Md Zahid Hasan, Associate Professor,** Department of CSE Daffodil International University, Dhaka. Our supervisor's in-depth knowledge and deep interest in the topic of "Computer Vision" made it possible to complete this assignment. This project could not have been completed without his inexhaustible patience, intellectual direction, continuous encouragement, constant and energetic supervision, constructive criticism, invaluable counsel, reviewing several substandard drafts and correcting them at every level.

We would like to express our heartiest gratitude to Dr. S. M. Aminul Haque, Dr. Md Zahid Hasan, and Dr. Touhid Bhuiyan, Head**,** Department of CSE, for his kind help to finish our project and also to other faculty member and the staff of CSE department of Daffodil International University.

We would like to thank every Daffodil International University classmate who participated in this discussion while completing course work.

Finally, we must respectfully thank the unwavering support and tolerance of our parents.

ABSTRACT

In recent years, the occurrence and mortality rate due to skin cancer has increased to a higher extent worldwide. It is crucial to identify such cancers early and accurately to provide proper treatment, and research has shown that deep intelligent learning-based ways to address this issue have been proved successful. The main motivation of this study is to classify skin cancer using deep learning techniques on dermoscopy dataset with optimal performance while training time taken into account. The aim is to develop such an automated framework which can perform optimally across three different dermoscopy datasets having diverse characteristics. We have proposed a model SkinNet-14 by altering compact convolutional transformer (CCT) using 32 × 32 sized input image which results in minimizing time complexity to classify skin cancer into different classes. The SkinNet-14 architecture is developed through ablation study conducted on CCT model using HAM dataset. Prior to that, several data augmentation techniques and preprocessing methods are applied to enhance the image quality and quantity of all the datasets. Afterwards, the proposed model is evaluated with the rest two datasets. Results show that, the model which was proposed, achieved an accuracy of 97.85% on the HAM dataset, 96.0% on the ISIC dataset, and 98.14% on the PAD dataset. Moreover, the proposed model yields better performance in terms of number of parameters, accuracy and training time than six transfer learning model while training with 32 × 32 sized images.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

Introduction

1.1 Introduction

Cancer is currently one of the most severe threats to global health where skin cancer is one of the most prevalent types of cancer in the world. Skin cancer comes in many forms. The most common types of skin cancer include melanoma, basal cell carcinoma (BCC), squamous cell carcinoma (SCC), actinic keratoses and intraepithelial carcinoma (AKIEC), dermatofibroma (DF), melanocytic nevi, etc. [1]. Globally, approximately 115,320 skin cancer cases and nearly 11540 skin cancer deaths were recorded in 2021 [2]. By 2040, 28.4 million new cases of cancer are predicted to have occurred, representing a 47% increase in the global cancer burden [3].

Deep learning has made significant strides in Computer Aided Diagnosis (CAD) systems, and these systems are now frequently used in research on CAD systems for various medical imaging interpretations. Due to their end-to-end feature representation capabilities, convolutional neural networks (CNNs) have made great advancements in skin lesion detection at present. However, precise classification of skin lesions remains challenging due to the following issues: (1) the need for a large number of training images as well as a lengthy and complex training process [4]. (2) Inter-class similarities and intra-class variations, and (3) lack of the ability to focus on discriminative skin lesion parts. [5] Large dataset requirements might be addressed by implementing transfer learning, but other issues like lengthy computational requirements and training times, generalization potential, performance consistency, and robustness of the model still need to be addressed [6].

Vision Transformer (ViT) [7], a model based on self-attention [8] and influenced by natural language processing (NLP), was initially implemented in computer vision tasks. The used architecture was a pure transformer design. In contrast to standard CNN architectures, the self-attention layers of the Transformer architecture may detect long-range dependencies [8], [9]. However, due to the lack of inductive bias in its architecture, ViT is a data-hungry model, according to the findings of this study [8]. This data-hungry approach of ViT has made transformers inapplicable for a variety of essential tasks, as data is scarce in many fields. In order to overcome the massive data limitations of ViTs, Hassani et al. [10] introduce the Compact Convolutional Transformer (CCT) model that implements sequential pooling and replaces patch embedding with convolutional embedding, allowing for more inductive bias.

The aim of this study is to develop a single framework including similar image pre-processing, data augmentation and model architecture that is capable of classifying skin cancer lesions from three different datasets. An ablation study on CCT is used to propose a robust model. Due to the presence of noise, hairs, dark corners, color charts, uneven illumination, and marker ink in dermoscopic images [11], image pre-processing methods are used to improve the performance of our proposed model.

1.2 Problem Statement

For several reasons, estimating the incidence of skin cancer is particularly challenging. Although melanoma only accounts for about 5% of skin cancer cases, it causes 75% of skin cancer deaths [12]. Due to the high mortality rate of melanoma, skin cancer is sometimes divided into melanoma and non-melanoma. Non-melanoma skin cancer is often not tracked by cancer registries [13]. Dermatologists have difficulty identifying skin cancer from a dermoscopy image of a skin lesion [12]. A biopsy and a pathology review may be required in some circumstances to diagnose cancer. Moreover, manual disease monitoring is time-consuming, labor-intensive, and sensitive to observer variability [6]. In addition, a lack of radiologists and an increase in the number of skin cancer patients may result in diagnostic and treatment delays. To address these challenges, it is essential to implement an automated diagnostic strategy for the detection of skin cancer that reduces diagnostic time and increases medical efficiency.

1.3 Research Objectives

The main objectives of the paper can be summarized as follows:

a) To utilize three datasets named HAM10000, PAD-UEFS and ISIC having a maximum of nine classes, different characteristics and imaging protocols, to employ to classify skin lesions.

b) To use three different data augmentation techniques to increase the volume of the datasets. Best data augmentation method will be selected based on the model performance.

c) To remove challenging artifacts and improve the quality of the image using various image pre-processing techniques.

d) To propose a model called SKINNET-14 by modifying the original compact convolutional transformer model for the efficient classification of skin lesions.

e) To solve the problems of lengthy training times and insufficient amount of data with the proposed model.

f) To perform an ablation study is by altering the layer architecture and the hyper-parameters of the base model to propose the robust model to classify skin lesions.

g) To compare the performance of the proposed SKINNET-14 model with six transfer learning CNN based models on all three dataset

1.4 Research Questions

a) How can we investigate research gaps in existing machine vision-based systems for correctly classifying different skin types?

b) How can we develop attention-based model, utilizing lower time complexity and low resolution image, approach for improving the accuracy for classify skin cancer according to their class?

1.5 Report Layout

Chapter 1 presents the research introduction, objectives, and key research questions.

Chapter 2 Brief summaries of the literature review are provided.

Chapter 3 describes the proposed methodology with a detailed description.

Chapter 4 explains paper's experimental results and discussed.

Chapter 5 concludes the present research along with a direction for future work.

CHAPTER 2

Literature Review

2.1 Related works

In recent works, several researchers have proposed various transformer, deep-learning, and machine learning based methods for classifying skin lesions. This section presents some literature on classifying skin diseases.

Mohamed et al. [14] proposed a skin lesion classification technique by modifying the architecture of the GoogleNet. The proposed model achieved an accuracy of 94.92% in multiclass classification. In another study, Jason et al. [15] combined conventional image processing with deep learning by fusing features to achieve greater accuracy in dermoscopy images for melanoma diagnosis. The deep learning component uses knowledge transfer via a modified ResNet-50 network to classify the melanoma from ISIC dataset and achieved 94% accuracy with an AUC of 0.90. In this work, Moloud et al. [16] introduced the Three-Way Decision (TWD) theory and used it for the analysis of images of skin cancer. Two uncertainty quantification (UQ) techniques, named ensemble MC dropout (EMC) and deep ensemble (DE), have been incorporated into the proposed hybrid deep learning model TWDBDL. In the final phase, the model's accuracy was 88.95% and its AUC was 0.92. In their study, Simon et al. [17] classified skin tissue into 12 meaningful dermatological classes using CNN and machine learning. The study also showed that semantic segmentation permits a network to interpretably learn the complete context of skin tissue types. The approach attained accuracy between 93% and 97%. Ameri et al. [18] proposed a skin cancer detection method that utilizes CNN to classify images into benign and malignant categories. No segmentation or feature extraction techniques were applied to lesions. The HAM10000 dataset yielded a classification accuracy of 84%.

Transformer networks are infrequently used to classify skin cancer. Chao et al. [19] proposed a ViT based SkinTrans model to classify skin cancer on HAM10000 and a clinical dataset. This paper uses overlapping multiscale sliding windows to serialize images using multiscale patch embedding. The proposed model achieved 94.3% accuracy on the HAM10000 dataset and 94.1% accuracy on the clinical dataset. Xiaoyu et al. [5] proposed a model named DeMAL-CNN for skin lesion classification in dermoscopy images. In DeMAL-CNN, a TPN consisting of three weight-shared embedding extraction networks and a mixed attention mechanism that takes both spatial-wise and channel-wise attention information into account were developed and implemented. The results of the ablation analysis indicated that DeMAL-CNN obtained a maximum accuracy of 92.7% on the ISIC dataset. In another study, Jingye et al. [20] proposed transformer- UNet based MT-TransUNet. It can segment and classify skin lesions simultaneously by mediating multi-task tokens in Transformers. The model achieved 91.2% accuracy on multiclass classification. In order to enhance the deep convolutional neural network's (DCNN) capacity for discriminative representation, Jianpeng et al. [21] propose the attention residual learning convolutional neural network (ARL-CNN) model for the detection of skin lesions in dermoscopy images. After applying the ARL-CNN model to the ISIC-skin 2017 dataset, the model attained an AUC of 0.905. Work by Nils et al. [22] proposed a unique patch-based attention architecture to successfully classify both the high-class imbalance and high-resolution real-world multi-class skin cancer datasets. The model gives global context between small, high-resolution patches. According to the results, using an attention-based approach increases MC-sensitivity by up to 7%. The maximum sensitivity was 67.8%. To improve skin cancer classification performance Soumyya et al. [23] merged soft attention with DenseNet, VGG, Inception-ResNet v2, and ResNet architectures. The authors observed that Soft-Attention enhances the performance of the original network. Their suggested Inception-ResNet v2 + soft attention (IRv+SA) model achieved the greatest 90.40% accuracy on the ISIC-2017 dataset.

2.2 Scope of the Problem

Several machine learning and deep learning-based models have been employed to classify skin cancer, as is evident from previous studies. In addition, transformer and attention-based models are employed to make progress in the tasks of skin classification. However, there are drawbacks such as high time complexity and the inability to utilize low-quality photos. There is scope for improvement in the classification of skin cancer pictures by addressing the shortcomings. In this study, these obstacles are considered in the context of establishing a single framework with strong interpretive skills.

2.3 Challenges

The focused challenges for this research are:

a) **Data Collection and merge**: Different datasets having a different classes need to be collect, to train and test this model.

b) **Data Augmentation:** Have to perform data augmentation techniques to increase the volume of the datasets. Challenging artifacts are removed and the quality of the image is enhanced using various image pre-processing techniques.

c) **Image Processing:** Images gathered from various sources occasionally had low or high contrast, or they were noisy. The challenge is to create noise-free, contrast-enhanced images that are perfect for classification.

d) **Select Base Model:** To solve the problems of lengthy training times and insufficient amount of data a perfect base model need to be choose for ablation study.

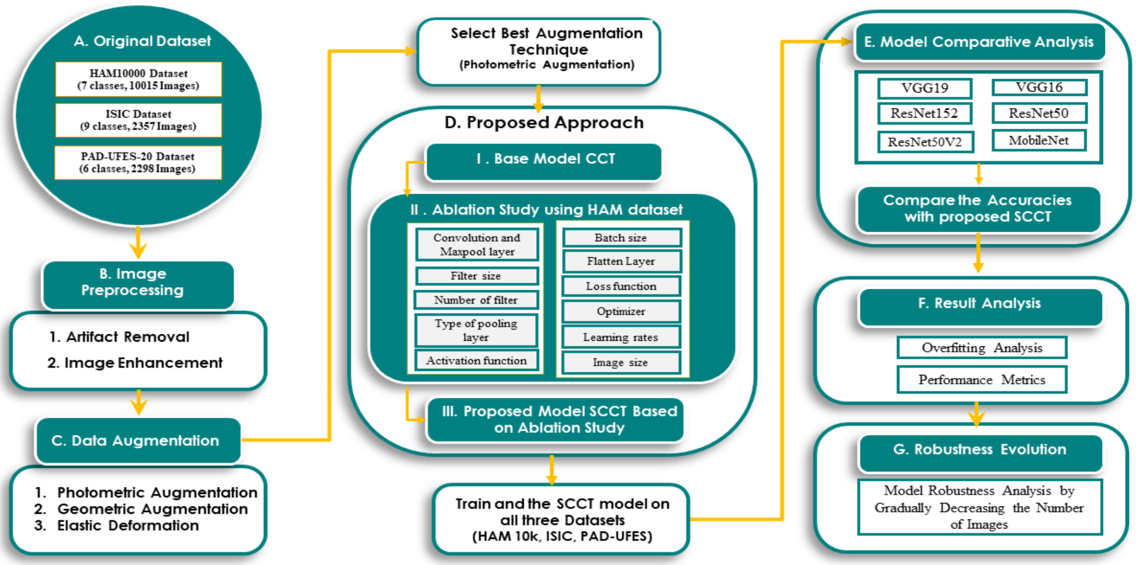
e) **Propose Robust Model and Improve Accuracy:** Many research is used many different model, in our approach we have to utilize an attention and CNN based hybrid approach to classify skin cancer more precisely.

CHAPTER 3

Materials and Methods

3.1 Working Process

To develop an effective transformer-based skin lesion classification model various steps are performed. The entire step-by-step methodology is illustrated in figure 3.1.



A. Original Dataset
HAM10000 Dataset (7 classes, 10015 Images)
ISIC Dataset (9 classes, 2357 Images)
PAD-UFES-20 Dataset (6 classes, 2298 Images)
B. Image Preprocessing
1. Artifact Removal
2. Image Enhancement
C. Data Augmentation
1. Photometric Augmentation
2. Geometric Augmentation
3. Elastic Deformation
Select Best Augmentation Technique (Photometric Augmentation)
D. Proposed Approach
I . Base Model CCT
II . Ablation Study using HAM dataset
Convolution and Maxpool layer
Filter size
Number of filter
Type of pooling layer
Activation function
Batch size
Flatten Layer
Loss function
Optimizer
Learning rates
Image size
III. Proposed Model SCCT Based on Ablation Study
Train and the SCCT model on all three Datasets (HAM 10k, ISIC, PAD-UFES)
E. Model Comparative Analysis
VGG19
VGG16
ResNet152
ResNet50
ResNet50V2
MobileNet
Compare the Accuracies with proposed SCCT
F. Result Analysis
Overfitting Analysis
Performance Metrics
G. Robustness Evolution
Model Robustness Analysis by Gradually Decreasing the Number of Images

Figure 3.1: Overall methodology to classify multiclass skin disease on three dataset

This study utilizes three publicly available skin cancer-related datasets named HAM10000, ISIC and PAD-UFES to conduct all the experiments. On each of the three datasets, multiple and similar image processing methods are used to eliminate artifacts and enhance the photos. Afterwards, a number of data augmentation approaches including photometric, geometric and elastic deformation are implemented to address data imbalance and scarcity issues. Among all the strategies, best data augmentation method is selected based on the model performance. After that, augmented data from the HAM10000 dataset are divided into 75% for training, 10% for validation, and 15% for testing, and fed into the CCT base model where the input image pixels size is 32×32. HAM10000 have the maximum number of data than other two datasets. A ten-case ablation study is conducted to assure optimal performance and to address the time complexity. On the basis of ablation research, the Skin Compact Convolutional Transformer (SKINNET-14) is developed by altering the layer structure and hyperparameters of the original CCT model. In addition, the SKINNET-14 model are trained and evaluated on all the rest two datasets and it is found that the model is robust enough to yield optimal performance over several dermoscopy datasets with image size 32 x 32. With this image size, six state of the art transfer learning models named ResNet50, ResNet152, VGG19, MobileNet, VGG16 and ResNet50V2 are trained and evaluated with all the three datasets and performance is compared with our proposed model in terms of accuracy and time complexity. Then, the outcomes of the SKINNET-14 model on all three datasets are analyzed with several performance metrics, and the likelihood of overfitting is examined. Finally, the resilience of the model is evaluated further by comparing its performance with decreasing numbers of photos gradually. The sections and subsections below provide a brief description of each process.

3.2 Dataset description

In this research, the evaluation of the suggested model is performed on three publically available dataset. The details of the dataset are discussed below.

a) HAM 10000 Dataset

The HAM 10000 [24] dataset is a popular publicly available kaggle dataset. The dataset consists of 10015 skin lesion images. It has seven classes including Actinic Keratosis, , Benign keratosis, Basal Cell Carcinoma, Dermatofibroma, Melanocytic Nevi, Melanoma and Vascular Lesions. The resolution of the images are 644 × 450 pixel.

b) ISIC dataset

The ISIC [25] dataset contains 2357 images which was collected from the International Skin Imaging Collaboration (ISIC) database. The dataset contains nine classes: Actinic Keratosis, Pigmented Benign Keratosis, Basal Cell Carcinoma, Melanoma, Dermatofibroma, Nevus, Squamous Cell Carcinoma, Seborrheic Keratosis and Vascular Lesions. The images of this dataset are 600 × 450 pixels.

c) PAD-UFES-20 Dataset

The PAD-UFES-20 [26] dataset consist of 2298 images with six classes. The classes are: Actinic Keratosis, Basal Cell Carcinoma, Melanoma, Nevus, Seborrheic Keratosis, Squamous Cell Carcinoma and Vascular Lesions. It is a very challenging dataset. The images of this dataset are 1050 × 1050 pixels.

An overview of all three of datasets are given in Table 3.1.

Table 3.1 Dataset description

Name	Description		
HAM10000 Dataset	Total No. Of Image	10015	
	No. of classes	7	
	Dimension	600 × 450	
	Actinic Keratosis	327	Number of images of each Class
	Benign Keratosis-Like Lesions	1099	
	Basal Cell Carcinoma	514	
	Dermatofibroma	115	
	Melanocytic Nevi	6,847	
	Melanoma	1,113	
	Vascular Lesions	142	
	Total no. of image	2357	
	Dimension	600 × 450	
	No. of classes	9	
	Actinic Keratosis	114	Number of images of each Class
	Dermatofibroma	95	
	Basal Cell Carcinoma	376	

ISIC Dataset	Melanoma	438	
	Pigmented Benign Keratosis	462	
	Nevus	357	
	Seborrheic Keratosis	77	
	Vascular Lesions	139	
	Squamous Cell Carcinoma	181	
PAD-UFES-20 Dataset	Total no. of image	2298	
	No. of classes	6	
	Dimension	1050 × 1050	
	Actinic Keratosis	730	Number of images of each Class
	Melanoma	52	
	Basal Cell Carcinoma	845	
	Nevus	244	
	Squamous Cell Carcinoma	235	
	Seborrheic Keratosis	192	

3.2.1 Skin Lesion Description

The dataset includes numerous skin lessons. Melanoma, squamous cell carcinoma, and basal cell carcinoma are the three main kinds of skin cancer [27]. The most prevalent non-melanoma skin cancers are basal cell carcinoma and squamous cell carcinoma. Most common basal cell carcinoma grows slowly and rarely spreads. Squamous cell carcinoma penetrates deeper and spreads more than basal cell carcinoma. Melanocyte-based malignancies, or melanomas, are malignant. Melanomas, the most malignant type of skin cancer, spread to other organs and are hard to cure. Figure 3.2 shows the images of each skin classes marking according to tumor, artifacts, and normal skin.

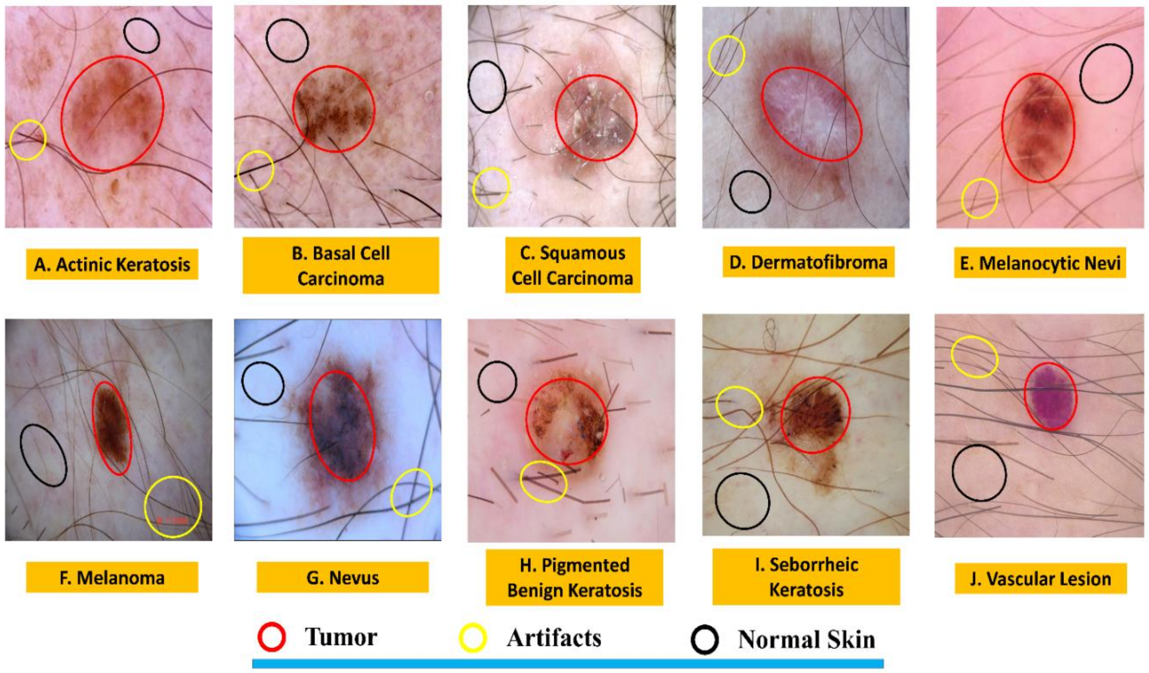


Figure 3.2: Cancer lesion of different skin classes

Actinic keratosis (Fig. 3.2-A) might appear differently like rough, dry, or scaly skin patch, on the top layer of skin, a patch or bump that is flat to slightly elevated. In certain instances, a rough, wart-like surface, bleeding and itching. Basal cell carcinoma (Fig. 3.2-B) causes skin changes like growths or sores that will not heal. Lesions are typically characterized by a shiny, transparent, skin-colored lump, a brown, black, or blue lesion, or a flat, scaly patch with a raised border or a whitish, waxy, scar-like lesion lacking a distinct boundary. Skeletal Cell Carcinoma (Fig. 3.2-C) can manifest as elevated growths with a central depression, open sores, scaly red patches, rough, thickened, or wart-like skin. It can occasionally itch, bleed or crust over. As visible in Fig. 3.2-D, sizes of dermatofibromas range from 0.5 to 1.5 cm in diameter. Color of dermatofibroma can range from pink to light brown on people with white skin to dark brown to black on people with dark skin; certain colors appear paler in the middle. Although dermatofibromas rarely exhibit symptoms, they can occasionally be tender, painful, or irritating. Melanocytic nevi (Fig. 3.2-E) typically grow to a maximum size of 40 cm. Tan to black is the color spectrum, and it can get lighter or darker with time. A nevi's surface can be smooth, uneven, elevated, thickened, or bumpy; it can differ in different areas of the nevus and it can alter with time. The nevus's skin is frequently dry, prone to irritation, and itchy. Melanomas (Fig. 3.2-F) usually shaped asymmetric with irregular border. The diameter of the melanoma mole is larger than 6 mm and has uneven color. The mole size and color changes over time and can face bleeding or itching. Nevus (Fig. 3.2-G) normally has round smooth mole, with a single color. Common nevi can seem tan, brown or pink and might have a flat or dome-shaped appearance. Typical nevi are unharmful clusters of colored cells. Pigmented benign keratosis (Fig. 3.2-I) and seborrheic keratosis (Fig. 3.2-J) are similar type and may appear as an oval growth with a minor rise or as a flat growth. The average size of a mole is 2.5 centimeters, and it can have a single or many growths that range in color from tan to black to brown. Vascular Lesions are dark or brilliant red in color and can cause the breakdown of the skin's surface, bleeding, and/or infection. It typically expands outward on the surface of the skin, whereas deeper lesions resemble bruises on the skin with a mass of soft tissue underneath.

3.3 Image preprocessing

Pre-processing images before putting them into a neural network optimizes the model's performance and computation time. This study uses different widely-used techniques to remove artifacts and enhance image quality by adjusting brightness and contrast. First, morphological opening is applied removes artifacts. Non Local Means Denoising (NLMD) is introduced to minimize noise and CLAHE to improve brightness and contrast. Finally, Gaussian blur algorithm smooths pixels while preserving ROI edges. In this regard, all the image pre-processing techniques are applied to all of our three datasets.

3.3.1 Artifact Removal

Morphological opening is a technique that eliminates all single-pixel artifacts, such as noisy spikes and tiny spurs, and blackens small objects [28]. To apply morphological opening, the image is first turned into binary format. Thus, small noises become more visible after the conversion to binary format. Using a kernel, morphological opening is applied to the binary image. The shape and size of this kernel are determined by the characteristics of the artifacts to be removed. After experimenting with a variety of kernel shapes and sizes, a rectangular kernel of size 5×5 is applied since, for this kernel, artifacts are successfully removed while essential information is preserved. Thus, a noise-free binary mask is produced, which is subsequently combined with the original picture via a bitwise AND operation.

3.3.2 Image Enhancement

Complex dermoscopy details and concealed information make it difficult for a model to accurately classify classes. To attain best performance, appropriate image enhancement techniques may assist in enhancing the visual contrast between Regions of Interest (ROIs) and backgrounds.

a) Non-Local Means Denoise (NLMD)

NLMD algorithm [29] is based on a basic principle: replacing a pixel's color with the average of the colors of neighbouring pixels. This leads to significantly improved post-filtering clarity and less loss of image detail than local mean methods. NLMD is implemented to reduce the noise of the images. The denoising of an image $z = (z1;\ z2;\ z3)$ in channel i to the pixel j is executed as follows [nlmd]:

$$\widehat{z_i}(x) = \frac{1}{C(x)} \sum_{k \in B(x,r)} z_i(x) \omega(x,k), \tag{1}$$

$$C(x) = \sum_{k \in B(x,r)} \omega(x,k) \tag{2}$$

here, $B\,(x,r)$ denotes the area of pixels x inside a radius of r. The weight $\omega(x,k)$ is determined by the squared Frobenius norm distance between color patches with centers at x and k that degrade under a Gaussian kernel.

b) Contrast limited adaptive histogram equalization (CLAHE)

CLAHE [30] is performed to rectify excessive contrast amplification and restore overall contrast balance. CLAHE is a variation of adaptive histogram equalization in which contrast amplification is limited in order to reduce this noise amplification issue. In CLAHE, the contrast enhancement close to a particular pixel value is determined by the slope of the transformation function.

c) Gaussian blur

Gaussian blurring [31] is used in image processing to minimize noise and eliminate speckles from an image. It is essential to remove the extremely high frequency components that surpass those connected with the gradient filter, as these can lead to the detection of erroneous edges. A two-dimensional Gaussian function's formula is:

$$G(i,j) = \frac{1}{2\pi\sigma^2} e^{\frac{i^2}{2\sigma^2}} \tag{3}$$

Here, i represents the horizontal axis' distance from the origin, j represents the vertical axis' distance, and σ represents the Gaussian distribution's standard deviation. The origin of these axes is (0, 0). The formula creates a surface in two dimensions with concentric circles that have a Gaussian distribution away from the center point. Fig. 3.3 shows the whole image pre-process steps.

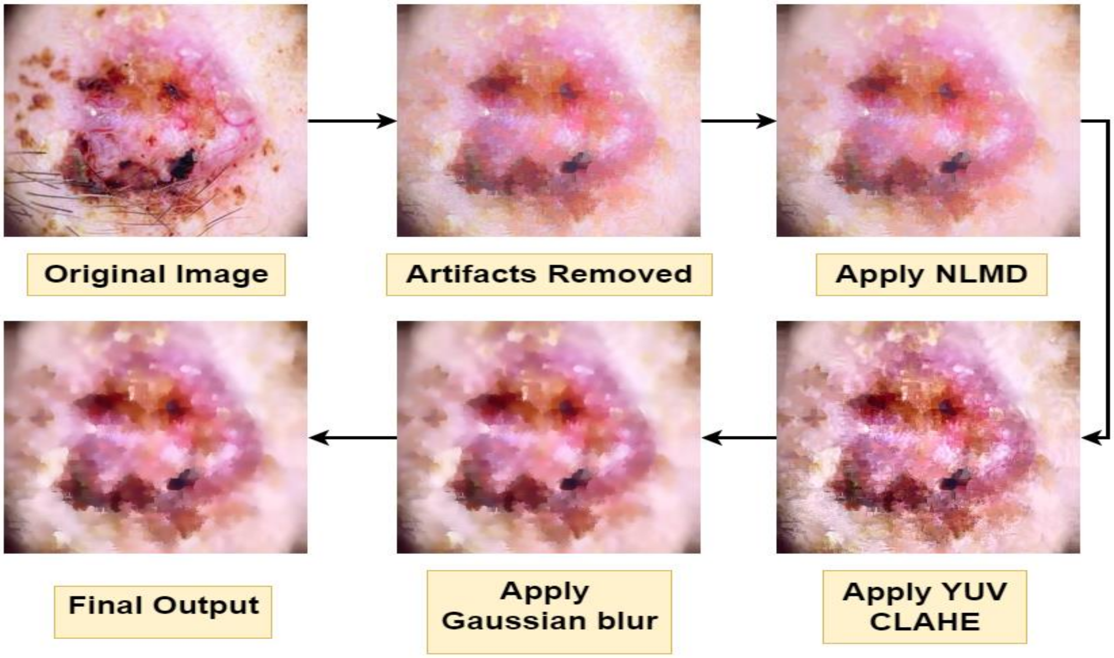


Figure 3.3: Image after each pre-process stage

3.4 Data Augmentation

The technique of artificially generating new training dataset samples from existing data is known as data augmentation. Data augmentation is vital for AI applications in medical imaging asannotated data is expensive and sparse. Data augmentation is essential since it increases labelled data. In this study, three different data augmentation techniques named photometric augmentation, geometric augmentation and elastic deformation were applied.

a) Geometric Data Augmentation

One of the most used augmentation methods to increase the amount of data is geometric transformation [32]. Geometric augmentation is the process of modifying an image's geometric shape by changing the values to their matching new values. It is a successful image enhancement method that just changes the image's shape without affecting the image's quality. Among several geometric augmentation methods in medical imaging research our study used, vertical flipping which can be used on matrices to flip their rows and columns vertically, horizontal flipping which allows the image to be flipped either to the left or to the right, vertical and horizontal flipping maintaining the image's natural horizontal-vertical column structure and rotation by rotate the images to any degree.

b) Photometric Data Augmentation

Photometric augmentation involves modifying pixel values such as brightness, sharpness, blurriness, color and contrast. Photometric augmentation transforms RGB channels by shifting each pixel's (r, g, and b) value to a new pixel's (r′, g′, and b′) value. It mainly alters visual color and lighting, not geometry [33]. It includes color jittering, grayscaling, filtering, brightness perturbation, noise addition, contrast adjustment, random erasing, etc [33]. However, this process must be carried out in such a way that critical pixel information is not lost. Among a number of different photometric methods, altering the brightness by maintaining level of lightness or darkness, contrast by making the light regions get lighter and the dark regions darker, color by changing the color balance of an image, and sharpness by sharping the details of an image yielded the best results as a photometric augmentation in this paper.

c) **Elastic Deformation**

Elastic deformations [34]as data augmentation stretches and changes the shape of images differently according on skin location and compression strength.

There are two steps involved in obtaining a distortion of a skin cancer image. The first step is to create a random stress field for the Δ_a and Δ_b directions, respectively. A random number between ν × [0.5, 0.5] is selected consistently for each pixel in each direction. The obtained horizontal and vertical pictures are applied a Gaussian filter independently (eq. (4) and (5)) to ensure that nearby pixels have equal displacement. The transformations contain the maximum value for the initial random displacement (ν) and the degree of smoothing, which is determined by the Gaussian filter's standard deviation (σ). Based on the overall look of the patches, we decided on a value of = 300 and a value of = 20 for deformation. Then, the image, skin segmentation mask, and mass annotations are stressed. In order to achieve this, each pixel is moved to a new location (eq. (6)), and intensities at integer coordinates are obtained using order one spline interpolation [34].

$$\Delta_a = G(\sigma) * \left(\nu \times Rand(w, z)\right) \tag{4}$$

$$\Delta_b = G(\sigma) * \left(\nu \times Rand(w, z)\right) \tag{5}$$

$$I_{transform}\left(j + \Delta_x(j,k), k + \Delta_y(j,k)\right) = I(j,k) \tag{6}$$

Here, I and $I_{transform}$ are the original and transformation images, respectively; w and z are the dimensions of the retinal fundus image.

3.5 Proposed model

CCT is a hybrid compact ViT with convolution [35]. With a local receptive field that preserves the image's local information, CCT models use CNN blocks as patching blocks. The self-attention mechanism detects relationships between image components and combines all relevant data. In this study, the original CCT model serves as the foundation for an ablation study in which different components are modified to achieve the optimal performance configuration.

3.5.1 Compact Convolutional Transformer (CCT)

CCT architecture consists of two main blocks. One is Convolutional Tokenization and another is Transformer with sequential pooling. The CCT methodology is shown in Figure 3.4.

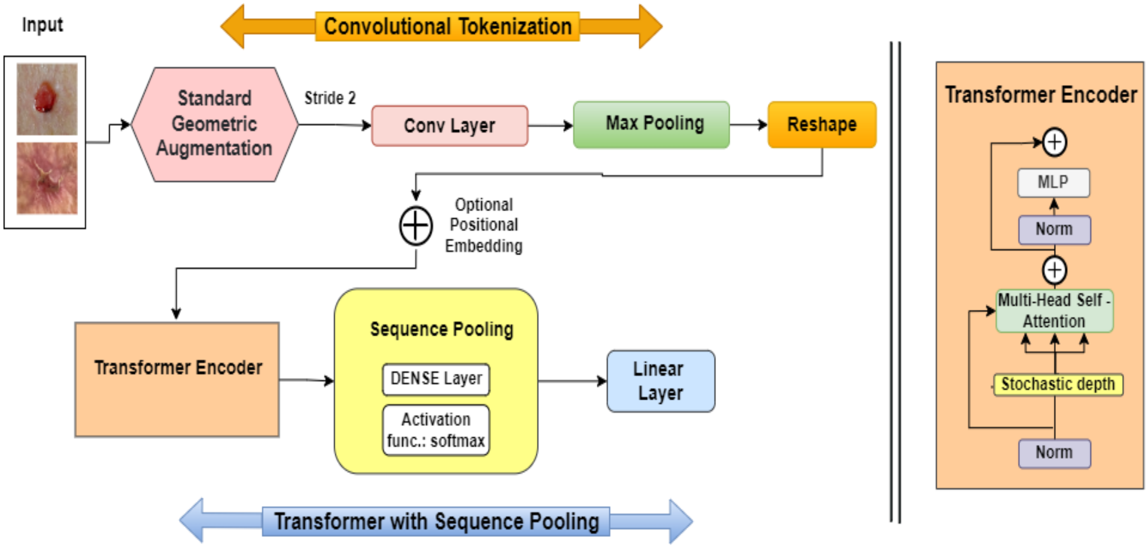

Input
Convolutional Tokenization
Standard Geometric Augmentation
Stride 2
Conv Layer
Max Pooling
Reshape
Optional Positional Embedding
Transformer Encoder
Sequence Pooling
DENSE Layer
Activation func.: softmax
Linear Layer
Transformer with Sequence Pooling
Transformer Encoder
MLP
Norm
Multi-Head Self - Attention
Stochastic depth
Norm

Figure 3.4: Architecture of base model

Convolutional Tokenization generates image patches [35]. Convolutional Tokenization processes for image z using the following formula:

$$z_0 = \text{MaxPool}(\text{ReLU}(\text{Conv2D}(z))) \qquad (7)$$

Here, the convolutional layer (Conv2d) contains 64 filters with strides 2 coupled with the ReLU activation function. The generated Conv2D feature maps are then downscaled by the maxpool layer. The convolutional tokenization block can accept images of any size as input. As a result, CCT models do not need that all image patches be the same size. Because of these convolutional patches, the CNN layers assist the model in retaining local spatial information.

Following that, the first block's image patches are sent to the transformer-based backbone, where an encoder block is composed of a Multihead self-attention (MSA) layer and a Multilayer perceptron (MLP) head. The transformer encoder employs layer normalization (LN), GELU activation, and dropout. In CCT models where the positional embeddings are learnable, layer normalization is applied after positional embedding.

The sequence pooling layer pools the output of the transformer backbone, rather than employing a class token to convert sequential outputs to a single class. The network may evaluate the sequential embeddings of latent space created by the transformer encoder and improve data correlation for the input data using this sequence pooling. Because it comprises relevant data from many input image regions, the sequence pooling layer pools the entire sequence of data. This is referred to as mapping transformation, and it is defined by equation 8:

$$x_L = f(x_0) \in \mathbb{R}^{(b\times n\times d)} \tag{8}$$

where x_0 is a layer transformer encoder and x_L is its output. Furthermore, b denotes a mini-batch size, d the embedding dimension, and n the sequence length. The output is routed through equation 9 where a linear layer and the softmax activation function is used.

$$x'_L = softmax(g(x_L)^T) \in \mathbb{R}^{(b\times 1\times n)} \tag{9}$$

The final output can be computed as:

$$z = x'_L x_L = softmax(g(x_L)^T) \times x_L \in \mathbb{R}^{(b\times 1\times d)} \tag{10}$$

As a result of pooling the second dimension, z is generated as an output. The images are then categorized after passing through a linear classification layer.

3.5.2 Base Model Architecture

This article presents a SKINNET-14 model, which is achieved by doing ablation studies on a CCT model as its foundation. Figure 3.5 depicts the architecture of CCT's Base Model.

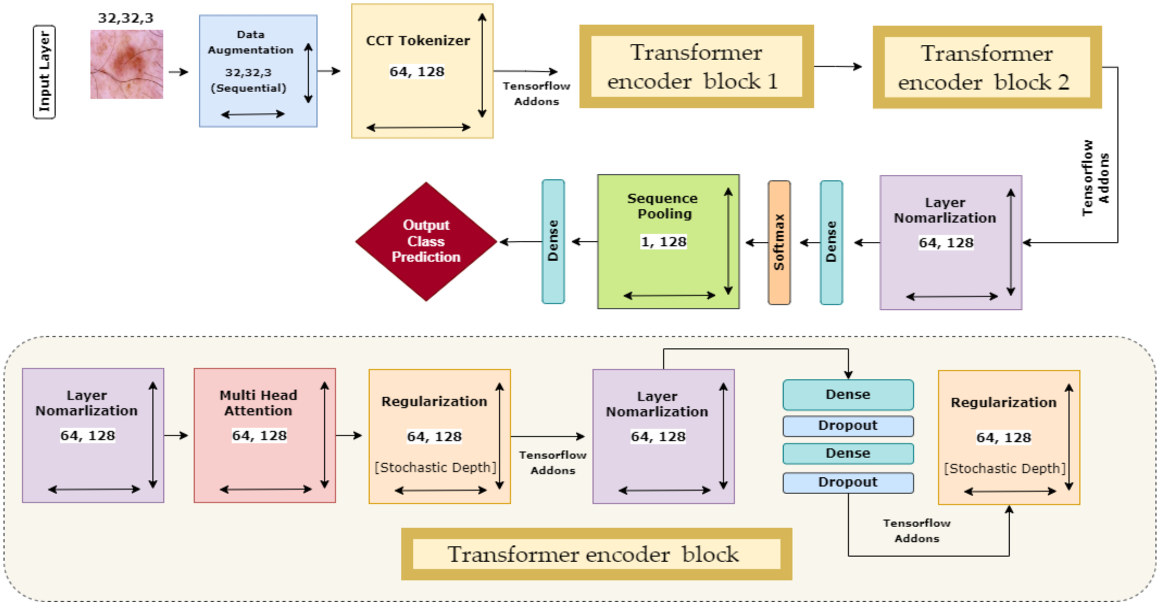


Figure 3.5: Detailed architecture of base model

The base CCT architecture includes the input layer, the CCT Tokenizer, the data augmentation layer, a regularization (stochastic depth) layer, multi-head attention layers, pooling layers, dense layers, dropout layers and output dense layers with softmax activation. The data augmentation layer augments 32×32×3 input images geometrically. The CCT Tokenizer block resizes the augmented pictures to 64×128. The convolutional layer of the CCT Tokenizer block initially consists of strides of size 2 and kernels of size 4, as well as a pooling layer kernel size of 4. After tokenization, data goes through tensorflow addons and the transformer encoder block. This block has numerous layers in a specified order: normalization (1), multi-head attention, regularization, normalization (2), then two pairs of dense and dropout layers with a 0.1 dropout factor. The transformer encoder block ends with an additional regularization layer. This layer's 64×128 output is regularized again with the Regularization layer, followed by another transformer encoder block. The output of the second transformer encoder block is passed through two layers of regularization and normalization. Normalized output travels through dense and softmax layers, producing 64×1 output data. This is passed to a sequence pooling layer, which produces 1×128 data. Finally, a linear classification layer classifies the images of the skin into the different classes of skin cancer.

3.5.3 Ablation Study

As previously mentioned, we conducted ablation research by changing the fine-tuning the hyper parameters and layer design and in order to maximize the performance of this CCT model. There are ten ablation studies, which vary in stride size, pooling layer kernel size, kernel size, batch size, loss function, optimizer, learning rate, and input layer image size. The number of transformer encoder blocks may also be increased or decreased. The activation functions and the type of pooling layers may also be changed. The proposed SKINNET-14 model is reached having a more robust design, improved accuracy of classification, and faster processing speeds after all ablation studies were finished. The findings of the research on ablation are described in Chapter 4.

3.5.4 Proposed SKINNET-14 architecture

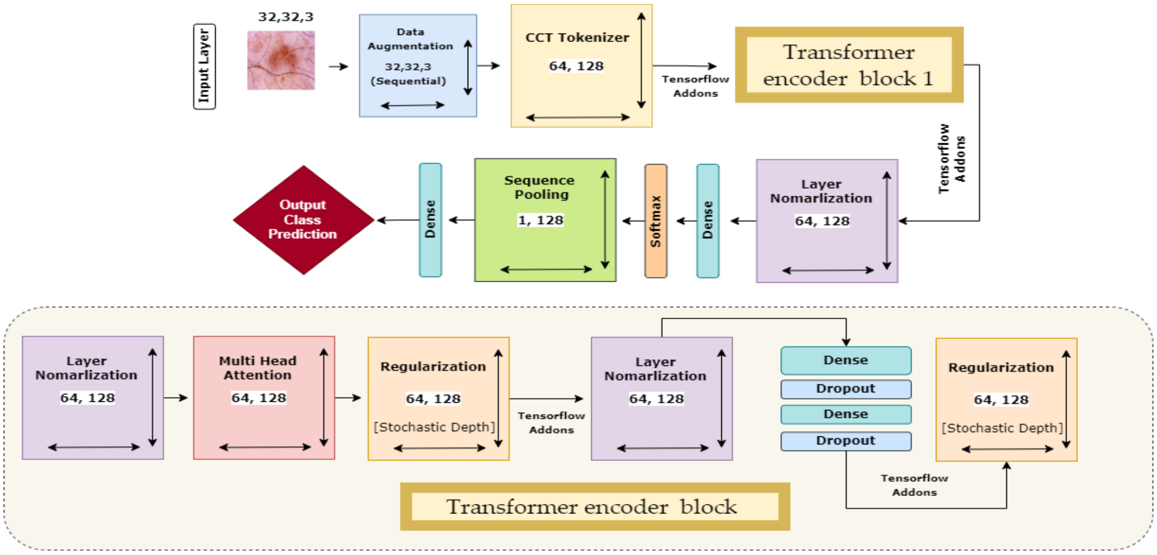


Figure 3.6: Proposed SKINNET-14 model's architecture

The optimized SKINNET-14 design reduces training time, maximizes performance, and limits time complexity. The final SKINNET-14 design features fewer transformer encoder blocks than the original CCT variant. Figure 3.6 shows that the SKINNET-14 model has one transformer encoder block, while the CCT architecture has two. This enables faster training and a smaller model. Except for a few model hyper parameters like stride size and kernel size, the architecture remains the same.

This model does not require positional encoding, reducing processing costs. The computational complexity of self-attention is $O(n^2.d)$, where n is the length of the input sequence and d is the number of vector dimensions. The addition of positional encoding $O(n^2.d + n.d^2))$ increases the computational complexity [8]. Since positional encoding is not required in the SKINNET-14 model and the transformer backbone only uses self-attention, the training and testing phases of the proposed model are shorter and require fewer resources. Therefore, the model is significantly more efficient.

3.5.5 Training Strategy

In order to train the base CCT model the batch size was set at 128, learning rate as 0.001 with optimizer Adam. For the PAD dataset, there will be 400 epochs; for the HAMM and

ISIC datasets. Categorical cross-entropy is employed initially as this is the standard loss function in multiclass instances [36]. The same configuration is considered while training the transfer learning models. = However, while conducting ablation study the model is experimented with different hyperparameters. In order to test different models and configurations, we used three PCs, each of which has an Intel Core i5-8400 processor, 16 GB of RAM, an NVidia GeForce GTX 1660 GPU, and a 256 GB DDR4 SSD for storage.

3.5.6 Transfer Learning Models

We compare the performance of multiple Transfer Learning models trained with the same datasets, taking training time into account, in order to assess the performance of our proposed technique. In total, 128 batches are executed over 400 epochs for the PAD dataset, 200 for HAMM, and 200 for ISIC.

a) VGG Architecture

The VGG networks [37] with 16 (VGG16) and 19 (VGG19) layers served as the foundation for the Visual Geometry Group (VGG) entry to the ImageNet Challenge 2014.

VGG16 having 16 weighted layers is a cutting-edge transfer learning algorithm that achieved 92.7% accuracy for the top five test results in the ImageNet dataset. Because the VGG model has more depth, it can assist the kernel in learning more complicated features.

VGG19 is a VGG model version with 19 weighted layers. In addition to the VGG16 model, there are three additional FC layers of 4096, 4096, and 1000 neurons, respectively. There are also five maxpool layers and a Softmax classification layer. In the convolutional layers, the ReLU activation function is utilized.

b) ResNet Architecture

Residual Networks (ResNets) [38] skip blocks of convolutional layers to create residual blocks. Stacking residual blocks improves training and reduces network deterioration.

ResNet50 employs different-sized convolution filters to reduce CNN model deterioration and training time. 48 convolutional layers, a maxpool, and an average pool layer make up this architecture. Model has 23 million trainable parameters.

ResNet152 has 152 layers. ResNet152 allowed training of neural networks with more than 150 layers. ResNet is an easy-to-optimize and effective deep learning architecture. As the network design contains several layers, it is time-consuming.

ResNet50V2 [38] is a modified version of the original ResNet50. ResNet50V2 outperforms ResNet50 and ResNet101 on ImageNet. ResNet50V2 modified how block connections propagate.

c) MobileNet

MobileNet [39] models are designed to replace costly convolutional layers with depth-separable convolutional blocks. MobileNet is a faster, smaller CNN that uses Depth-wise Separable Convolution. MobileNet models are beneficial for mobile and embedded devices due to their small size.

CHAPTER 4

Experimental Results and Discussion

4.1 Result and discussion

In this section, the results of this research are explained, including the outcomes of numerous ablation experiments and model validation metrics. This part also includes a description of the accuracy loss curves and confusion matrix to further examine the efficacy of the proposed SKINNET-14 model.

Evaluation metrics

Several metrics are investigated to determine how well the suggested classification model performs. A true positive (TP) is a finding where the model correctly classifies the positive category. A result is considered to be true negative (TN) if the model correctly identified the negative class. False positive (FP) and false negative (FN) outcomes are those in which the model wrongly predicts the positive class and the negative class, respectively. The percentage of accurate predictions is known as accuracy. Equations of the performance metrices used in this study are given below.

$$\mathrm{Accuracy} = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{11}$$

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{12}$$

$$\mathrm{Precision} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{13}$$

$$\mathrm{F_1} = 2\ \frac{\mathrm{precision} * \mathrm{recall}}{\mathrm{precision} + \mathrm{recall}} \tag{14}$$

4.2 Ablation study Results

This section contains the details of all the ablation studies undertaken to achieve the optimal model architecture. First, in order to discover the optimal augmentation method, a total of three augmentation methods are investigated by training the base model, and test accuracy results are shown in Table 4.1.

Table 4.1: Ablation study on various augmentation techniques on HAMM dataset

Technique	**Parameters**	**Size of Image**	**Training time × epoch**	**Test accuracy (%)**	**Outcomes**
Geometric	0.41M	32×32	16s × 100	87.68	Poor accuracy
Photometric	**0.41M**	**32×32**	**16s × 100**	**89.85**	Best **accuracy**
Elastic Deformation	0.41M	32×32	15s × 100	83.83	Poor accuracy

With a test accuracy of 89.85%, the photometric augmentation methodology clearly exceeds the other data augmentation methods. Consequently, additional ablation studies have been conducted employing photometric augmented images.

By adjusting the model's features various experiments are conducted to evaluating the model's performance. The performance of a classification model can be enhanced by changing a few of its features. Ten separate studies are conducted for this research. The outcomes of these ablation experiments are given in Tables 4.2, Tables 4.3, and Tables 4.4.

Table 4.2: Ablation study on modifying transformer layer, activation function, pooling layer, stride size.

Modification 1: Transformer layer changes						
No.	Transformer encoder block count	Parameters	Training time × epoch	Overall time	Test accuracy (%)	Outcomes
1	3	0.57M	200 x 26s	41-44 minutes	89.63	Good accuracy with High time
2	2	0.41M	200 x 16s	33-35 minutes	89.85	Good accuracy with medium time
3	**1**	**0.24M**	**200 x 7s**	**21-24 minutes**	**89.55**	**Almost good accuracy with lower time**

Modification 2: Activation function changes					
No.	Activation function	Parameters	Training time × epoch	Test accuracy (%)	Outcomes
1	softplus	0.24M	10s × 200	88.97	Poor accuracy
2	softsign	0.24M	10s × 200	90.88	Almost good accuracy
3	elu	0.24M	11s × 200	90.38	Almost good accuracy
4	**relu**	**0.24M**	**10s × 200**	**91.24**	**Best accuracy**

5	Tanh	0.24M	10s × 200	89.55	Poor accuracy
Modification 3: Pooling layer changes					
No.	Pooling layer types	Parameters	Training time × epoch	Test accuracy (%)	Outcomes
1	Average	0.24M	7s × 200	91.24	Good accuracy
2	**Max**	**0.24M**	7s × 200	**92.37**	**Best accuracy**
Modification 4: Stride size changes					
No.	Strides numbers	Parameters	Training time × epoch	Test accuracy (%)	Outcomes
1	**1**	**0.24M**	7s × 200	**93.57**	**Best accuracy**
2	2	0.24M	4s × 200	91.14	Almost good accuracy
3	3	0.24M	4s × 200	91.37	Almost good accuracy
4	4	0.24M	4s × 200	89.63	Poor accuracy

- **Modification 1: Transformer layer changes**

In this research, the transformer layer is changed by varying the number of encoder blocks. In table 4.2, it is visible that increasing the number of blocks increases the number of parameters and the duration of time, yet the accuracy is nearly identical. A single transformer block with 0.24M parameters, 21–24 minutes, and 89.55% accuracy achieves the maximum performance. The configuration has the smallest number of trainable parameters and the least training time per epoch. Therefore, configuration 3 is selected for additional ablation experiments.

- **Modification 2: Activation function changes**

Different activation functions influence the classification model, and the optimal activation function improves model performance. Six different activation functions named Tanh, ELU, ReLU, SoftSign, and SoftPlus are applied to the model (Table 4.2). ReLU scored the highest accuracy among six activation functions, 91.24 %, with 10 seconds each epoch. Therefore, ReLU activation is selected for additional ablation experiments.

- **Modification 3: Type of pooling layer changes**

Pooling layers downsample feature maps by summarizing feature presence in patches. Average pooling and maxpooling layers are applied for this experiment (Table 4.2). The test accuracy went from 91.24 % to 92.37 % after using the max pooling layer. As a result, maxpooling layer is selected for additional ablation experiments.

- **Modification 4: Stride size changes**

The stride selection impacts the network's matrix structure after convolution. Various stride sizes like 4, 3, 2 and 1 are applied in the transformer layers. Table 4.2 shows that using a single stride improved the accuracy to 93.57% with 7 seconds per epoch. So, further ablation experiments continued with stride size 1.

Table 4.3: Ablation study on modifying kernel size, pooling layer kernel size, loss function, batch size

Modification 5: Kernel size changes					
No.	Kernel size count	Parameters	Training time × epoch	Test accuracy (%)	Outcomes
1	4	0.3M	8s × 200	93.83	Good accuracy
2	**3**	**0.24M**	**7s** × 200	**94.77**	**Best accuracy**
3	2	0.2M	9s × 200	93.57	Good accuracy

4	1	0.17M	10s × 200	88.33	Poor accuracy
Modification 6: Loss function changes					
No.	Loss Function	Parameters	Training time × epoch	Test accuracy (%)	Outcomes
1	Binary Crossentropy	0.24M	7s × 200	94.88	Good accuracy
2	**Categorical Crossentropy**	**0.24M**	7s × 200	**95.80**	**Best accuracy**
3	Mean Squared Error	0.24M	7s × 200	94.81	Good accuracy
4	Mean absolute error	0.24M	7s × 200	94.63	Good accuracy
5	Mean squared logarithmic error	0.24M	7s × 200	28.76	Poor accuracy
Modification 7: Batch size changes					
No.	Batch size	Parameters	Training time × epoch	Test accuracy (%)	Outcomes
1	256	0.24M	6s × 200	94.09	Good accuracy
2	**128**	**0.24M**	**7s × 200**	**96.68**	**Best accuracy**
3	64	0.24M	11s × 200	95.56	Good accuracy

4	32	0.24M	16s × 200	95.30	Good accuracy

- **Modification 5: Kernel size changes**

The kernel size impacts transition speed and can be optimized through calculation of kernel density. Various kernel sizes including 4, 3, 2, and 1 are utilized, and Table 4..3 demonstrates that kernel size 3 yielded the highest accuracy of 94.77% and the shortest time per epoch of 7 seconds. Consequently, the kernel size of 3 is maintained for future ablation studies.

- **Modification 6: Loss function changes**

Loss functions are used to assess how effectively a model predicts the outcome. In the experiment, five distinct loss functions are implemented. They are Categorical Crossentropy, Binary Crossentropy, Mean Squared Logarithmic Error, Mean Absolute Error, and Mean Squared Error. Categorical Crossentropy's 95.80% result was the highest of all five loss functions (Table 4.3). Categorical Crossentropy is therefore adjusted for subsequent ablation experiments.

- **Modification 7: Batch size changes**

Different batch sizes affect a classification model's performance. For the modification, 256, 128, 64, and 32-batch sizes are evaluated (Table 4.3). Training the model with 128 batches results in a maximum accuracy of 96.68% with 10 seconds for each epoch. Whereas, other batch sizes reduce accuracy (Table 4.3). Further ablation studies use batch size 128.

Table 4.4: Ablation study on modifying optimizer, learning rate, image size

Modification 8: Optimizer changes					
No.	Optimizer	Parameters	Training time × epoch	Test accuracy (%)	Outcomes
1	**Adam**	**0.24M**	**7s × 200**	**96.68**	**Best accuracy**
2	Nadam	0.24M	7s × 200	88.62	Poor accuracy
3	SGD	0.24M	7s × 200	94.46	Good accuracy
4	Adamax	0.24M	7s × 200	95.23	Good accuracy
5	RMSprop	0.24M	7s × 200	94.48	Good accuracy
Modification 9: Learning rate changes					
No.	Learning rate	Parameters	Training time × epoch	Test accuracy (%)	Outcomes
1	0.01	0.24M	7s × 200	92.23	Poor accuracy
2	0.006	0.24M	7s × 200	95.47	Good accuracy
3	**0.001**	**0.24M**	**7s × 200**	**97.85**	**Best accuracy**
4	0.0008	0.24M	7s × 200	96.68	Good accuracy
Modification 10: Image size changes					
No.	Image size	Parameters	Training time × epoch	Test accuracy (%)	Outcomes

1	64	0.24M	24s × 200	96.17	Near best accuracy
2	**32**	**0.24M**	**7s × 200**	**97.85**	**Best accuracy**
3	28	0.24M	6s × 200	95.17	Good accuracy
4	16	0.24M	5s × 200	94.88	Good accuracy

- **Modification 8: Optimizer changes**

An optimizer for neural networks modifies weights and learning rate. It decreases loss and increases accuracy. In this study, five optimizers named Nadam, Adam, Adamax, RMSprop, and SGD were tested with a learning rate of 0.001. The best accuracy of 96.68%, is attained with the Adam optimizer (Table 4.4). Therefore, Adam optimizer is retained for the remainder of the ablation research.

- **Modification 9: Learning rate changes**

Learning rate affects loss gradient weights in neural networks. With the Adam optimizer, learning rates of 0.0008, 0.01, 0.001 and 0.006 are utilized. The Adam optimizer still obtains the best result of 97.85% with a learning rate of 0.001 (Table 4.4). Hence, a learning rate of 0.001 was established for the following ablation studies.

- **Modification 10: changing the image size**

The final study involves doing experimentation with the input layer picture dimensions (image height and width). We test 64×64, 32×32, 28×28, and 16×16 pixel sized images. The study's findings are presented in Table 4.4. The model was able to be trained in just 10 seconds per epoch, while still achieving the best testing accuracy of 97.85%, with an image size of 32×32 on HAMM dataset. However, the image size of 64×64 also achieved a very good test accuracy of 96.17%, but the training time was 24 seconds per epoch.

The input image dimension is 32 × 32 pixels since it takes minimal training time while keeping high performance. This is essential because our objective is to design a model with high performance that also takes time complexity into account. Figure 4.1 depicts how test accuracy grew gradually during all ablation studies conducted on the base model.

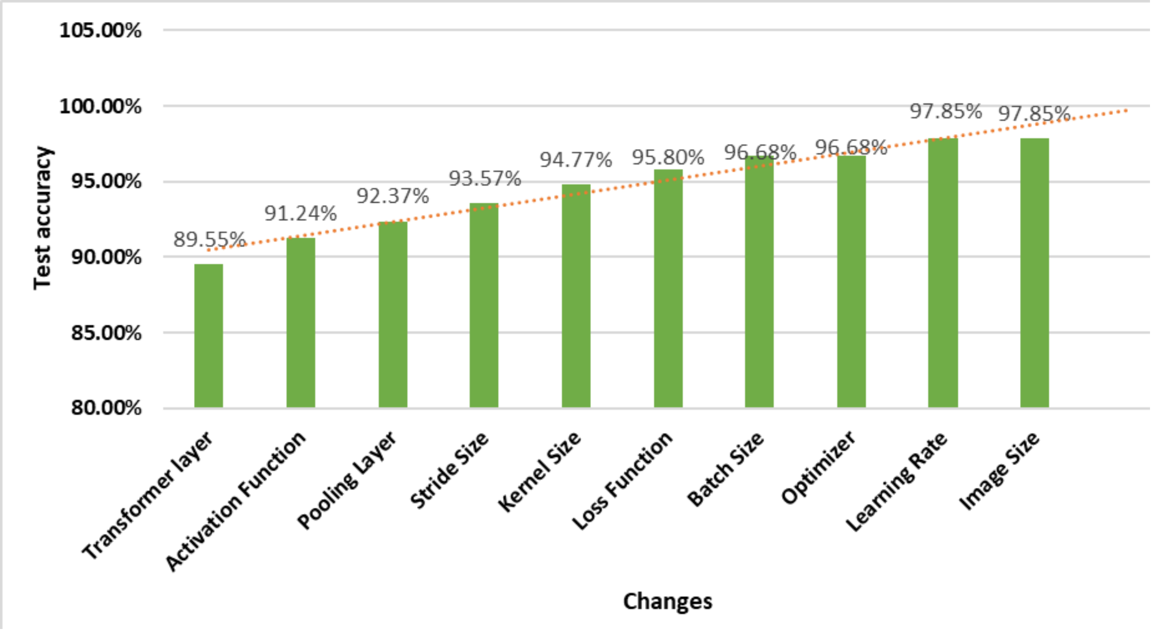


Figure 4.1: Test Accuracy increasing over 10 ablation study

After ablation study, the configuration of the proposed SKINNET-14 are, 32 × 32 image size, Adam optimizer with learning rate of 0.001, batch size of 128 and kernel size 3. The activation function of SKINNET-14 is relu, loss function is Categorical Crossentropy and pooling layer is Max Pooling. Pooling layer kernel size is 3 and stride size is 1. Finally, model's projection_dim is 128, stochastic_depth_rate is 0.1 and weight_decay is 0.0001.

4.2.1 Performance evaluation of the proposed model

By completing ablation experiments on the base model, the final SKINNET-14 model has been created whose classification performance is significantly enhanced. This is accomplished by modifying and configuring the model in various ways. Table 4.5 shows some statistical analysis for the proposed SKINNET-14 model, such as precision, recall, f1-score, and number of images tested on each class for three dataset.

Table 4.5: Different matrices calculated for SKINNET-14 model performance evaluation

	Skin Class	Precision	Recall	F1-score	Support	
HAM1000 Dataset	Actinic Keratosis	0.99	0.98	0.99	245	**Test Accuracy: 97.85%**
	Basal Cell Carcinoma	1.00	0.99	0.99	386	
	Benign Keratosis	1.00	0.99	0.99	824	
	Dermatofibroma	1.00	1.00	1.00	86	
	Melanoma	0.99	0.98	0.98	835	
	Melanocytic Nevi	0.98	0.99	0.99	5135	
	Vascular Lesions	0.18	0.04	0.06	107	
ISIC Dataset	Actinic Keratosis	1.00	1.00	1.00	85	**Test Accuracy: 96.01%**
	Basal Cell Carcinoma	1.00	1.00	1.00	282	
	Dermatofibroma	1.00	0.99	0.99	71	
	Melanoma	0.91	0.90	0.90	329	
	Nevus	0.98	0.99	0.99	268	
	Pigmented Benign Keratosis	1.00	1.00	1.00	347	
	Seborrheic Keratosis	0.50	0.52	0.51	58	
	Squamous Cell Carcinoma	1.00	0.99	1.00	136	
	Vascular Lesions	1.00	1.00	1.00	104	
PAD-UFES-20 Dataset	Actinic Keratosis	0.98	0.98	0.98	548	
	Basal Cell Carcinoma	0.99	0.98	0.98	634	
	Melanoma	0.93	1.00	0.96	39	

	Nevus	0.99	0.99	0.99	183	**Test Accuracy: 98.14%**
	Seborrheic Keratosis	0.99	0.99	0.99	176	
	Squamous Cell Carcinoma	0.93	0.97	0.95	144	

The resultsof Table 6 clearly shows that the proposed model performed exceptionall y well on all three datasets. In the Ham dataset, the model achieved a good statistical score on six classes of the dataset except vasculer lesion. Though the average accuracy obtained on the HAM dataset is 97.85%. In the ISIC dataset, the model gained a good statistical score on all eight datasets, except Seborrheic Keratosis. The average accuracy on the dataset is 96.01%. Finally, on the most challenging PAD-UFES dataset, the proposed model achieved the highest average accuracy of 98.14%. It is visible that the model obtained a good statistical score of precision, recall, and f1-score on all six classes of this dataset.

The confusion matrix on three datasets produced by the SKINNET-14 model is shown in Figure 4.2. The true labelling of the test photos are indicated by row values. Column values are used to represent the labels that the model predicted for the test set photos. The confusion matrix (Figure 4.2)'s diagonal values show how many test images the model successfully predicted. It is clear that the model is not biased toward any one class or classes, nor does it predict any class significantly more accurately than the others. In fact, the model provides about equal numbers of accurate predictions for each class, further demonstrating the model's robustness.

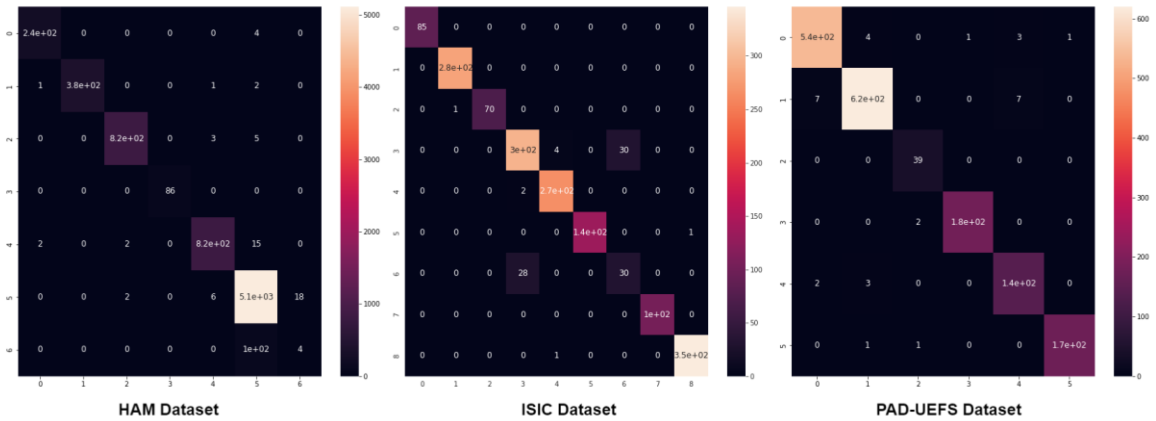
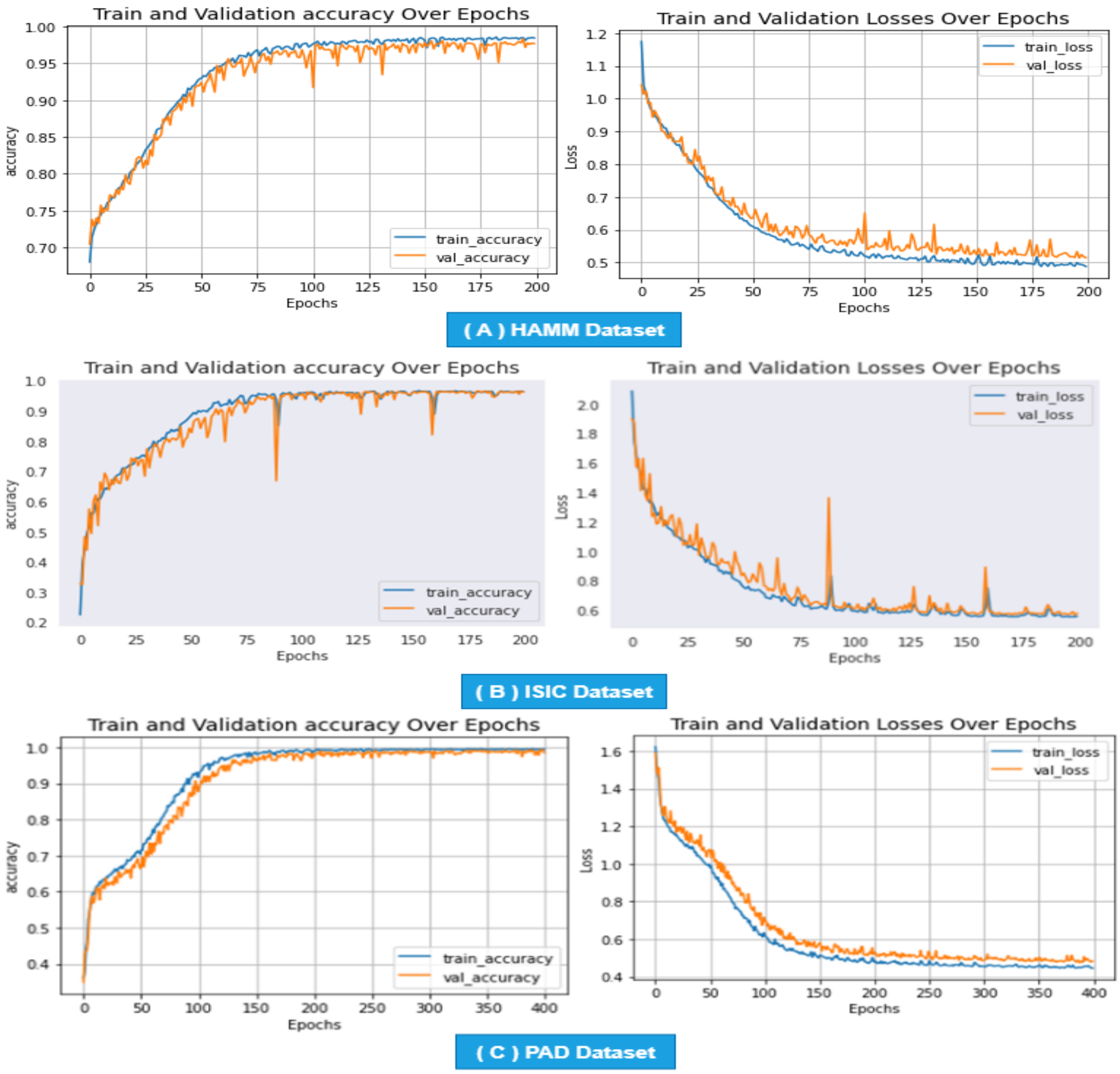


Figure 4.2: Confusion matrix for the proposed SKINNET-14 model on all three datasets following the ablation

Figure 4.3 depicts the SKINNET-14 model's accuracy and loss curves on HAMM, ISIC and PAD dataset. From the figures of all three dataset, it is visible that the model's training and validation curves converging without substantial gaps, indicating no overfitting. Similarly, Loss curves (Figure 4.3) converging steadily from start to ending epoch. It can be said that neither overfitting nor under-fitting occurred during the model's training phase.



(A) HAMM Dataset

(B) ISIC Dataset

(C) PAD Dataset

Figure 4.3: Accuracy curve and loss curve of SKINNET-14 model on (A) HAMM dataset (B) ISIC Dataset (C) PAD dataset.

4.2.2 Comparison with CNN based transfer learning models

Six state-of-the-art transfer learning CNN models are used to evaluate the proposed approach. All six models are trained and tested on the three dataset as the proposed model, with 32×32 pixel input images. The optimizer is Adam, the batch size is 128, and the learning rate for each model in the table is 0.001. Table 4.6 shows experiment results.

Table 4.6: Comparison of performance with six state-of-the-art transfer learning models

Model	Parameters	**HAM DATASET**			**ISIC DATASET**			**PAD DATASET**		
		epochs	Per epoch time	Accuracy	epochs	Per epoch time	Accuracy	epochs	Per epoch time	Accuracy
VGG19	20026436	200	65-67s	80.47 %	200	30-34s	70.87 %	400	28-30s	**82.97%**
VGG16	14716740	200	65-67s	**81.21 %**	200	30-34s	**71.21 %**	400	28-30s	81.38%
ResNet152	58379140	200	65-67s	65.79 %	200	30-34s	75.79 %	400	28-30s	78.79%
ResNet50	23595908	200	65-67s	69.27 %	200	30-34s	68.57 %	400	28-30s	72.97%
ResNet50 V2	23572996	200	65-67s	66.25 %	200	30-34s	63.21 %	400	28-30s	77.15%
MobileNet	3232964	200	65-67s	43.42 %	200	30-34s	49.12 %	400	28-30s	55.48%
SKINNET -14	241861	200	7-8s	97.85 %	200	2-3s	96.01 %	400	2-3s	98.14%

VGG16 achieved the highest test accuracy of 81.21 % on the HAM dataset and 71.21 % on the ISIC dataset, outperforming all other transfer learning models. On the PAD dataset, VGG19 achieved the highest score of the six CNN-based pre trained models with 82.97 %. On all three datasets containing 32-32 pixel pictures, the accuracy of the remaining transfer learning models varied between 40% and 80%. It is also evident that the parameters of all transfer learning models were high, which raised the temporal complexity and per epoch time, which ranged between 65 and 67 seconds for the HAM dataset, 30 and 34 seconds for the ISIC dataset, and 28 and 30 seconds for the PAD dataset. In contrast, our suggested model achieves the highest accuracy of 97.85 % on the HAM dataset, 96.0 % on the ISIC dataset, and 98.14 % on the PAD dataset. In terms of accuracy, the SKINNET-14 model outperformed all six transfer learning methods. In addition, the suggested model's parameter size is 241861, resulting in a reduced temporal complexity of 7 to 8 seconds per epoch on the HAM dataset, 2 to 3 seconds on the PAD dataset, and 1 to 2 seconds on the ISIC dataset. With our methodology, training takes about 6–24 minutes as opposed to approximately two hours for transfer learning methods. This is a substantial improvement in terms of time-intensiveness. Additionally, achieving near-optimal performance with smaller image sizes takes less memory and storage space, making the model less resource-hungry and contributing to a reduction in space complexity.

CHAPTER 5

Impact on Society, Environment and Sustainability

5.1 Impact on Society

The use of skin cancer prediction can have a significant impact on society by improving the early detection and treatment of skin cancer. Early detection and treatment of skin cancer can significantly increase the chances of successful treatment and can potentially save lives.

One potential benefit of skin cancer prediction is that it can help to reduce the burden on healthcare systems and providers by increasing the efficiency and accuracy of the diagnostic process. This can potentially reduce the need for multiple diagnostic tests or visits to the doctor, which can help to reduce healthcare costs and increase access to care.

Another potential benefit of skin cancer prediction is that it can help to reduce the risk of misdiagnosis or delayed diagnosis, which can have serious consequences for patients. By identifying potential skin cancers early on, skin cancer prediction can help to ensure that patients receive prompt and appropriate treatment, which can improve outcomes and reduce the overall burden of the disease on society.

Overall, the use of skin cancer prediction can have a significant positive impact on society by improving the early detection and treatment of skin cancer, reducing the burden on healthcare systems, and reducing the risk of misdiagnosis or delayed diagnosis.

5.2 Impact on Environment

The use of computer vision for the early detection of skin cancer can potentially have both positive and negative impacts on the environment.

On the positive side, using computer vision for early detection can increase the efficiency and accuracy of the diagnostic process, potentially reducing the need for multiple visits to the doctor or specialist and the use of certain diagnostic tools. This can in turn reduce the environmental impact of transportation and the production and disposal of certain diagnostic materials.

However, it is also important to consider the energy and resources required to power and maintain the computer systems and other equipment used for the diagnostic process. These systems may rely on non-renewable energy sources and the production and disposal of electronic devices can also have environmental impacts.

Overall, the environmental impact of using computer vision for early detection of skin cancer will depend on the balance between the potential positive and negative impacts of the technology. It is important to carefully consider these potential impacts and take steps to minimize any negative effects.

5.4 Sustainability Plan

Here are some potential components of a sustainability plan for using computer vision for the early detection of skin cancer:

a) **Energy efficiency:** Ensuring that the computer systems and other equipment used for the diagnostic process are energy efficient can help to reduce the environmental impact of the process.

b) **Use of renewable energy**: Using renewable energy sources, such as solar or wind power, to power the computer systems and equipment used for the diagnostic process can further reduce the environmental impact.

c) **Responsible disposal of equipment:** Properly disposing of equipment, such as computers and other electronic devices, at the end of their useful life can help to reduce waste and prevent harmful substances from entering the environment.

d) **Collaboration with healthcare providers:** Partnering with healthcare providers to ensure that the diagnostic process is integrated into regular care can help to reduce the overall environmental impact by reducing the need for additional transportation and other resources.

e) **Education and outreach:** Providing education and outreach to patients and healthcare providers about the benefits of using computer vision for early detection of skin cancer can help to increase the adoption of this diagnostic approach and contribute to its sustainability.

f) **Continuous improvement:** Regularly evaluating and improving the sustainability of the diagnostic process can help to ensure that it is as environmentally friendly as possible over the long term.

CHAPTER 6

Conclusion and Future Work

5.1 Conclusion

This research proposes a robust model for detecting skin cancer. First, three distinct renowned publically accessible datasets are gathered to evaluate the model's performance. Several image preparation approaches are then employed to eliminate artifacts, improve the quality of skin lesions, and prevent the overfitting problem. On the foundational model CCT, an ablation investigation is conducted to propose the resilient model SkinNet-14. The suggested model achieved an accuracy of 97.85 % on the HAM dataset, 96.0 % on the ISIC dataset, and 98.14 % on the PAD dataset. The proposed model utilized a low-quality, 32-by-32-pixel picture and a small number of parameters with a time-intensive addressing process. The results of the suggested model are compared to the outcomes of six transfer learning models, with the new model performing better. The accuracy is significantly higher than comparable works, demonstrating the effectiveness of the suggested system.

5.2 Limitation and Future Work

Despite of a good result, the dataset is insufficient for the model. Lack of Real data. Future proposals may include a model that is more precise and robust. Some different image preprocess methods and augmentation techniques can be applied to help the model under the ROI. Different types of segmentation work can be carried out. Other illnesses can be studied using the model. Later on, a similar smartphone application will be developed to detect skin cancer and display skin data to users.

REFERENCES

[1] Y. G. Xu *et al.*, “Nonmelanoma Skin Cancers: Basal Cell and Squamous Cell Carcinomas,” *Abeloff’s Clinical Oncology*, pp. 1052-1073.e8, Jan. 2020, doi: 10.1016/B978-0-323-47674-4.00067-0.

[2] R. L. Siegel, K. D. Miller, H. E. Fuchs, and A. Jemal, “Cancer Statistics, 2021,” *CA Cancer J Clin*, vol. 71, no. 1, pp. 7–33, Jan. 2021, doi: 10.3322/CAAC.21654.

[3] H. Sung *et al.*, “Global Cancer Statistics 2020: GLOBOCAN Estimates of Incidence and Mortality Worldwide for 36 Cancers in 185 Countries,” *CA Cancer J Clin*, vol. 71, no. 3, pp. 209–249, May 2021, doi: 10.3322/caac.21660.

[4] A. Mohan, A. K. Singh, B. Kumar, and R. Dwivedi, “Review on remote sensing methods for landslide detection using machine and deep learning,” *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 7, p. e3998, Jul. 2021, doi: 10.1002/ETT.3998.

[5] X. He, Y. Wang, S. Zhao, and C. Yao, “Deep metric attention learning for skin lesion classification in dermoscopy images,” *Complex and Intelligent Systems*, vol. 8, no. 2, pp. 1487–1504, Apr. 2022, doi: 10.1007/s40747-021-00587-4.

[6] M. R. H. Mondal, S. Bharati, P. Podder, and P. Podder, “Data analytics for novel coronavirus disease,” *Inform Med Unlocked*, vol. 20, p. 100374, Jan. 2020, doi: 10.1016/J.IMU.2020.100374.

[7] A. Dosovitskiy *et al.*, “An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale,” *CoRR*, Oct. 2020, doi: 10.48550/arxiv.2010.11929.

[8] A. Vaswani *et al.*, “Attention is all you need,” 2017.

[9] J. Chen *et al.*, “TransUNet: Transformers Make Strong Encoders for Medical Image Segmentation,” *CoRR*, Feb. 2021, doi: 10.48550/arxiv.2102.04306.

[10] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi, “Escaping the big data paradigm with compact transformers,” 2021.

[11] S. Montaha, S. Azam, A. K. M. Rakibul Haque Rafid, S. Islam, P. Ghosh, and M. Jonkman, “A shallow deep learning approach to classify skin cancer using down-scaling method to minimize time and space complexity,” *PLoS One*, vol. 17, no. 8 August, Aug. 2022, doi: 10.1371/journal.pone.0269826.

[12] A. Esteva *et al.*, “Dermatologist-level classification of skin cancer with deep neural networks,” *Nature 2017 542:7639*, vol. 542, no. 7639, pp. 115–118, Jan. 2017, doi: 10.1038/nature21056.

[13] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, “Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries,” *CA Cancer J Clin*, vol. 68, no. 6, pp. 394–424, Nov. 2018, doi: 10.3322/CAAC.21492.

[14] M. A. Kassem, K. M. Hosny, and M. M. Fouad, “Skin Lesions Classification into Eight Classes for ISIC 2019 Using Deep Convolutional Neural Network and Transfer Learning,” *IEEE Access*, vol. 8, pp. 114822–114832, 2020, doi: 10.1109/ACCESS.2020.3003890.

[15] J. R. Hagerty *et al.*, “Deep Learning and Handcrafted Method Fusion: Higher Diagnostic Accuracy for Melanoma Dermoscopy Images,” *IEEE J Biomed Health Inform*, vol. 23, no. 4, pp. 1385–1391, Jul. 2019, doi: 10.1109/JBHI.2019.2891049.

[16] M. Abdar *et al.*, “Uncertainty quantification in skin cancer classification using three-way decision-based Bayesian deep learning,” *Comput Biol Med*, vol. 135, Aug. 2021, doi: 10.1016/j.compbiomed.2021.104418.

[17] S. M. Thomas, J. G. Lefevre, G. Baxter, and N. A. Hamilton, “Interpretable deep learning systems for multi-class segmentation and classification of non-melanoma skin cancer,” *Med Image Anal*, vol. 68, Feb. 2021, doi: 10.1016/j.media.2020.101915.

[18] A. Ameri, “A deep learning approach to skin cancer detection in dermoscopy images,” *J Biomed Phys Eng*, vol. 10, no. 6, pp. 801–806, 2020, doi: 10.31661/jbpe.v0i0.2004-1107.

[19] C. Xin *et al.*, “An improved transformer network for skin cancer classification,” *Comput Biol Med*, vol. 149, Oct. 2022, doi: 10.1016/j.compbiomed.2022.105939.

[20] J. Chen, J. Chen, Z. Zhou, B. Li, A. Yuille, and Y. Lu, “MT-TransUNet: Mediating Multi-Task Tokens in Transformers for Skin Lesion Segmentation and Classification,” Dec. 2021, [Online]. Available: http://arxiv.org/abs/2112.01767

[21] J. Zhang, Y. Xie, Y. Xia, and C. Shen, “Attention Residual Learning for Skin Lesion Classification,” *IEEE Trans Med Imaging*, vol. 38, no. 9, pp. 2092–2103, Sep. 2019, doi: 10.1109/TMI.2019.2893944.

[22] N. Gessert *et al.*, “Skin Lesion Classification Using CNNs with Patch-Based Attention and Diagnosis-Guided Loss Weighting,” *IEEE Trans Biomed Eng*, vol. 67, no. 2, pp. 495–503, Feb. 2020, doi: 10.1109/TBME.2019.2915839.

[23] S. K. Datta, M. A. Shaikh, S. N. Srihari, and M. Gao, “Soft-Attention Improves Skin Cancer Classification Performance,” May 2021, [Online]. Available: http://arxiv.org/abs/2105.03358

[24] P. Tschandl, C. Rosendahl, and H. Kittler, “Data descriptor: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions,” *Sci Data*, vol. 5, Aug. 2018, doi: 10.1038/SDATA.2018.161.

[25] “Skin Cancer ISIC | Kaggle.” https://www.kaggle.com/datasets/nodoubttome/skin-cancer9-classesisic (accessed Dec. 25, 2022).

[26] A. G. C. Pacheco *et al.*, “PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones,” *Data Brief*, vol. 32, p. 106221, Oct. 2020, doi: 10.1016/J.DIB.2020.106221.

[27] M. Nawaz *et al.*, “Skin cancer detection from dermoscopic images using deep learning and fuzzy k-means clustering,” *Microsc Res Tech*, vol. 85, no. 1, pp. 339–351, Jan. 2022, doi: 10.1002/JEMT.23908.

[28] J. A. Salido and C. Ruiz, “Hair artifact removal and skin lesion segmentation of dermoscopy images,” *Asian Journal of Pharmaceutical and Clinical Research*, vol. 11, no. Special Issue 3, pp. 36–39, 2018, doi: 10.22159/ajpcr.2018.v11s3.30025.

[29] B. D. Reddy, D. Bhattacharyya, N. T. Rao, and T. Kim, “Medical Image Denoising Using Non-Local Means Filtering,” pp. 123–127, 2022, doi: 10.1007/978-981-16-8364-0_15.

[30] S. Tripathy and T. Swarnkar, “Unified Preprocessing and Enhancement Technique for Mammogram Images,” *Procedia Comput Sci*, vol. 167, pp. 285–292, Jan. 2020, doi: 10.1016/J.PROCS.2020.03.223.

[31] Y. Zhang *et al.*, “A Poisson-Gaussian Denoising Dataset With Real Fluorescence Microscopy Images.” pp. 11710–11718, 2019.

[32] C. Shorten and T. M. Khoshgoftaar, “A survey on Image Data Augmentation for Deep Learning,” *J Big Data*, vol. 6, no. 1, pp. 1–48, Dec. 2019, doi: 10.1186/S40537-019-0197-0/FIGURES/33.

[33] L. Taylor and G. Nitschke, “Improving Deep Learning with Generic Data Augmentation,” *Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence, SSCI 2018*, pp. 1542–1547, Jan. 2019, doi: 10.1109/SSCI.2018.8628742.

[34] E. Castro, J. S. Cardoso, and J. C. Pereira, “Elastic deformations for data augmentation in breast cancer mass detection,” *2018 IEEE EMBS International Conference on Biomedical and Health Informatics, BHI 2018*, vol. 2018-January, pp. 230–234, Apr. 2018, doi: 10.1109/BHI.2018.8333411.

[35] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. v. Le, “AutoAugment: Learning Augmentation Policies from Data,” *Cvpr 2019*, no. Section 3, pp. 113–123, May 2018, doi: 10.48550/arxiv.1805.09501.

[36] I. Lorencin *et al.*, “On Urinary Bladder Cancer Diagnosis: Utilization of Deep Convolutional Generative Adversarial Networks for Data Augmentation,” *Biology 2021, Vol. 10, Page 175*, vol. 10, no. 3, p. 175, Feb. 2021, doi: 10.3390/BIOLOGY10030175.

[37] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, Sep. 2014, doi: 10.48550/arxiv.1409.1556.

[38] K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 770–778, Dec. 2015, doi: 10.48550/arxiv.1512.03385.

[39] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted Residuals and Linear Bottlenecks.” pp. 4510–4520, 2018.

SkinNet-14: A fine-tuned CCT model for classifying skin cancer addressing computational complexity and training time

ORIGINALITY REPORT

19%	10%	17%	3%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Inam Ullah Khan, Sami Azam, Sidratul Montaha, Abdullah Al Mahmud, A.K.M. Rakibul Haque Rafid, Md. Zahid Hasan, Mirjam Jonkman. "An Effective Approach to Address Processing Time and Computational Complexity Employing Modified CCT for Lung Disease Classification", Intelligent Systems with Applications, 2022 Publication	4%
2	mdpi-res.com Internet Source	1%
3	www.mdpi.com Internet Source	1%
4	www.researchgate.net Internet Source	1%
5	dokumen.pub Internet Source	1%
6	doaj.org Internet Source	1%