ORIGINAL ARTICLE

SkinNet-14: a deep learning framework for accurate skin cancer classification using low-resolution dermoscopy images with optimized training time

Abdullah Al Mahmud[1] · Sami Azam[2] · Inam Ullah Khan[1] · Sidratul Montaha[3] · Asif Karim[2] · Aminul Haque[1] · Md. Zahid Hasan[1] · Mark Brady[4] · Ritu Biswas[1] · Mirjam Jonkman[2]

Abstract

The increasing incidence of skin cancer necessitates advancements in early detection methods, where deep learning can be beneficial. This study introduces SkinNet-14, a novel deep learning model designed to classify skin cancer types using low-resolution dermoscopy images. Unlike existing models that require high-resolution images and extensive training times, SkinNet-14 leverages a modified compact convolutional transformer (CCT) architecture to effectively process 32×32 pixel images, significantly reducing the computational load and training duration. The framework employs several image preprocessing and augmentation strategies to enhance input image quality and balance the dataset to address class imbalances in medical datasets. The model was tested on three distinct datasets—HAM10000, ISIC and PAD—demonstrating high performance with accuracies of 97.85%, 96.00% and 98.14%, respectively, while significantly reducing the training time to 2–8 s per epoch. Compared to traditional transfer learning models, SkinNet-14 not only improves accuracy but also ensures stability even with smaller training sets. This research addresses a critical gap in automated skin cancer detection, specifically in contexts with limited resources, and highlights the capabilities of transformer-based models that are efficient in medical image analysis.

Keywords Skin cancer · Geometric augmentation · Photometric augmentation · Image preprocessing

1 Introduction

Skin cancer represents a severe threat to global health and is one of the most widespread forms of cancer around the world. Skin cancer appears in many forms. The most common types of skin cancer include melanoma, basal cell carcinoma (BCC), actinic keratoses and intra-epithelial carcinoma (AKIEC), squamous cell carcinoma (SCC), dermatofibroma (DF), melanocytic nevi and more [1]. Approximately 325,000 new cases of melanoma were identified worldwide in 2020, leading to an estimated 57,000 deaths [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]. For several reasons, estimating the incidence of skin cancer is particularly challenging. Although about 5% of all cases of skin cancer are melanomas, it is responsible for 75% of all deaths from skin cancer [4]. Due to the high death rate of melanoma, skin cancer is occasionally divided into melanoma and non-melanoma subtypes. Cancer registries frequently do not keep track of non-melanoma skin cancer [5]. Dermatologists also have difficulty identifying skin cancer from dermoscopy images of skin lesions [4]. In some circumstances, to effectively diagnose cancer a biopsy and pathology review may be required. Moreover, manual disease monitoring is time-consuming, labor-intensive and sensitive to observer variability [6]. In addition, a lack of

✉ Asif Karim
asif.karim@cdu.edu.au; sami.azam@cdu.edu.au

[1] Department of Computer Science and Engineering, Health Informatics Research Laboratory (HIRL), Daffodil International University, Dhaka 1341, Bangladesh

[2] Faculty of Science and Technology, Charles Darwin University, Casuarina, NT 0909, Australia

[3] Department of Computer Science, University of Calgary, Calgary, Canada

[4] Faculty of Arts and Society, Charles Darwin University, Casuarina, NT 0909, Australia

radiologists and an increase in the number of skin cancer patients may result in delays in diagnosis and treatment. To address these challenges, it is essential to implement an effective diagnostic strategy for the detection of skin cancer that reduces diagnostic time and increases medical efficiency.

Deep learning has made significant strides in computer-based medical diagnostic systems, and these systems are now frequently used in research to interpret medical images. Due to their end-to-end feature representation capabilities, deep convolutional neural networks (CNNs) have made notable advancements in skin lesion detection. However, the 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 [7], (2) inter-class similarities and intra-class variations, and (3) lack of the ability to focus on discriminative skin lesion parts [8]. Transfer learning may be used to solve the need for large datasets, but other difficulties such as lengthy training times, computational needs, generalizability, robustness and performance stability of the model, should also be addressed [6].

Vision Transformer (ViT) [9], a model based on self-attention [10] and influenced by natural language processing (NLP), was initially implemented in computer vision tasks. In contrast to standard CNN architectures, the self-attention layers of the Transformer architecture may detect long-range dependencies [10, 11]. However, due to the lack of inductive bias in its architecture, ViT is a data-hungry model [10]. This data-hungry approach of ViT has made transformers unsuitable for a variety of essential tasks, as training datasets remain scarce in many fields. In order to overcome the massive data limitations of ViTs, Hassani et al. [12] introduced the compact convolutional transformer (CCT) model that implements sequential pooling and replaces patch embedding with convolutional embedding, allowing for more inductive bias. Due to the presence of noise, hairs, dark corners, color charts, uneven illumination and marker ink in dermoscopic images [13], various image processing techniques are used to improve the performance of the proposed model.

This study addresses important gaps in skin cancer detection by developing SkinNet-14, a deep learning model capable of classifying low-resolution dermoscopy images while overcoming the challenges faced by traditional models that rely on high-resolution data. By enhancing the CCT architecture, SkinNet-14 not only reduces computational demands but also maintains high diagnostic accuracy, making it particularly ideal for resource-constrained clinical settings. The model further addresses the issue of data imbalance through enhanced preprocessing and augmentation, promising improved performance and being more applicable in real-world diagnostic scenarios. The primary contributions of the manuscript can be outlined as follows:

1. All the datasets of this study feature an uneven distribution of images across the classes, which could hamper model performance. To solve this problem, three different data augmentation techniques are experimented to increase the volume of the datasets. The highest performing data augmentation method is selected based on the model performance.

2. We propose SkinNet-14, which is a modified CCT model, for skin cancer classification. The model addresses problems associated with high computational complexity. Convolutional blocks are used to tokenize the vision transformer, lowering model training time and ensuring good performance even with low-resolution images.

3. We further optimize the performance and efficiency of the proposed model by conducting an ablation study that involves modifying the layer design and hyperparameters. The intention is to reduce the parameter number and computation complexity while improving overall performance.

4. In order to evaluate the effectiveness of the suggested SkinNet-14 model in terms of training time and accuracy with 32×32 sized images, a number of transfer learning models, including ResNet50, ResNet152, MobileNet [47], ResNet50V2 [48], VGG16 [49] and VGG19 [50], are applied to all three datasets and compared with the proposed model.

5. The classifier is trained by progressively reducing the number of images available in order to assess the sustainability and generalization potential of our model relative to the training dataset size. Results indicate that the model produces good performance even with fewer images, demonstrating the stability of the SkinNet-14 model in skin lesion detection.

The remainder of this paper is divided into the following sections. A summary of the literature review is provided in Sect. 2. Information on the proposed methodology is provided in Sect. 3. The adopted datasets are described in Sect. 4. Details on image preprocessing methods are provided in Sect. 5. An overview of data augmentation strategies is given in Sect. 6. Section 7 includes a description of the models with experimental configuration. The ablation experiment and findings are covered in Sect. 8. Section 9 concludes the paper.

2 Literature review

In recent literature, researchers propose several transformer, deep learning and machine learning-based methods for classifying skin lesions. This section presents the summary of existing literature on classifying skin diseases.

2.1 Deep learning

Mohamed et al. [14] proposed a skin lesion classification technique by modifying the architecture of 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 ResNet50 network, to classify the melanoma from the ISIC-2019 dataset and achieved 94% accuracy with an AUC of 90%. The article by, Moloud et al. [16] introduces the three-way decision (TWD) theory and applies it to analyze skin cancer images. Two uncertainties presents an analysis of skin cancer images using TWD theory. The proposed hybrid deep learning model TWDBDL has integrated two uncertainty quantification (UQ) techniques, namely deep ensemble (DE) and ensemble MC dropout (EMC). The study concludes that the model achieved an accuracy of 88.95% and an AUC of 92% in its final phase. In their study, Simon et al. [17] classify skin tissue into 12 meaningful dermatological categories 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 an accuracy of between 93 and 97%. The utilization of a convolutional neural network (CNN) for image classification into benign and malignant categories was proposed by Ameri et al. [18] in their study on skin cancer detection. The images were not subjected to any segmentation or feature extraction techniques. An accuracy rate of 84% was achieved through the utilization of the HAM10000 dataset. Jitendra et al. [19] proposed an ensemble-based machine learning (ML) and deep learning (DL) model. After feature extraction and classification, the model achieved 93% accuracy on ISIC Kaggle dataset. To help physicians in diagnosis, Reza et al. [20] represented a viable deep learning-based approach to detect skin cancer using 57,536 images. The model performed 89.3% ± 1.1% accuracy on multiclass classification. In their work, Abdelhafeez et al. [21] proposed a novel hybrid method for skin cancer classification that combines deep learning, neutrosophic techniques and feature fusion. Using ISIC-2019 data, the model attained 85.74%. To classify skin cancer lesions based on features extracted from preprocessed images Khater et al. [22] used explainable artificial intelligence techniques to interpret the model results. The model achieved an accuracy of 94% on PH2 dataset. The studies by Surono et al. [23, 24] explore the use of convolutional neural networks (CNNs) combined with various machine learning (ML) algorithms and the use of the U-Net architecture for semantic segmentation of CT images with varying resolutions. In another study, Sornsuwit et al. [25] propose a new ensemble learning algorithm called least error boosting (LEBoosting) to improve the classification of cardiovascular disease. The survey studies on skin cancer explore various ML–DL models and show how they assist dermatologists with complex and composite preprocessing. The articles also state that processing skin-infected images by ML–DL faces challenges such as low contrast between infected and normal skin, and artifacts like hair, bubbles and ink. Additionally, algorithm design must consider issues like extensive training needs, prevalence of light-skinned individuals in datasets, minimal variation between classes, and the diverse sizes and shapes of lesions in unbalanced datasets [26–29].

2.2 Attention and transformer

Transformer networks are infrequently used as a classifier of skin cancer. The SkinTrans model proposed by Chao et al. [30] utilizes a ViT approach to accurately classifying skin cancer on both HAM10000 and a clinical dataset. The aim of this paper was to utilize multiscale patch embedding to serialize images by implementing overlapping multiscale sliding windows. The study demonstrates that the proposed model attained high accuracy rates of 94.3% and 94.1% on the HAM10000 dataset and clinical dataset, respectively. Xiaoyu et al. [8] proposed a model named DeMAL-CNN for skin lesion classification from dermoscopy images. In DeMAL-CNN, a three-part network (TPN) consisting of three weight-shared embedding extraction networks, and a mixed attention mechanism, which 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-2016–2017 datasets. In another study, Jingye et al. [31] proposed transformerUNet-based MT-TransUNet. It was able to segment and classify skin lesions simultaneously by mediating multitask tokens in Transformers. The model achieved 91.2% accuracy for multiclass classification. To enhance the deep convolutional neural network (DCNN) capacity for discriminative representation, Jianpeng et al. [32] 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. [32] proposed a unique patch-based attention architecture to successfully classify both the high-class imbalance and high-resolution real-world multiclass 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 achieved was 67.8%. To improve skin cancer classification performance Soumyya et al. [33] merged soft attention with DenseNet, VGG, ResNet and Inception-ResNet v2 architectures. The authors found that soft attention enhances the performance of the original network. Their suggested Inception-ResNet v2 + soft attention (IRv + SA) model achieved the greatest accuracy of 90.40% on the ISIC-2017 dataset. On the HAM10000 dataset, Chao et al. [34] suggested an improved ViT network for classifying skin cancer. The proposed approach had a 94.3% accuracy. In this study, Suliman et al. [30] created a two-tier system for accurate skin cancer classification. They used the Medical Vision Transformer (MVT) on the HAM10000 and got a 96% accuracy. Guang et al. [35] introduce a unique approach to classifying skin cancer in HAM10000 to increase classification performance even further. Experiment results suggest that the approach has a classification accuracy of 94.1%. In their study, Arshed et al. [36] proposed a different method for diagnosing various skin cancer disorders based on off-the-shelf ViT. The suggested method was compared to 11 CNN-based transfer learning algorithms to assess its performance. On the HAM10000 dataset, the model got an accuracy of 92.14%.

As shown in previous studies, several machine learning and deep learning-based models have been employed to classify skin cancer. In addition, transformer and attention-based models can be employed to improve the accuracy of skin cancer classification. However, there are drawbacks, such as high time complexity and the inability to utilize low-quality images. There is scope for improvement in the classification of skin cancer images by addressing the noted shortcomings. The limitations of the previous works and a comparison with our work are described in Sect. 8.6. In this study, these challenges are considered in the context of establishing a single framework to achieve robust interpretive capability.

3 Proposed methodology

To develop an effective transformer-based skin lesion classification model various steps are performed. The entire step-by-step methodology is illustrated in Fig. 1.

The purpose of this research was to develop a deep learning model that can accurately classify different types of skin cancer based on low-resolution dermoscopy images. The architectural design of SkinNet-14 is developed in order to address the particular challenges of efficiently processing images with 32×32 pixels. In order to address the common challenges that arise from medical imaging datasets, such as data scarcity, imbalances and artifacts, the existing CCT were modified. An essential element of the design is the ablation study, which aims to identify and optimize the most impactful model parameters for computational efficiency and diagnostic accuracy.

Three publicly accessible skin cancer datasets—HAM10000, ISIC and PAD-UFES—were used in the study. Every dataset went through a standardized preprocessing phase during which image quality was enhanced and artifacts were eliminated. Following this, we applied and compared three different data augmentation methods to address the class imbalance, selecting the best-performing approach based on experimental results. The data from HAM10000 were split into training, validation and testing sets, adhering to a 75:10:15 distribution, respectively. This dataset was chosen for initial model development due to its comprehensive image volume. To validate the model, the optimized model was subsequently implemented on the remaining ISIC and PAD-UFES datasets.

To measure performance, we evaluated SkinNet-14 against six cutting-edge transfer learning models. Each model was evaluated for accuracy, F1 Score and training efficiency across all datasets. To assess model robustness, we conducted a series of experiments where we systematically reduced the number of training images, observing the impact on the model's performance. Performance metrics such as precision, recall and F1-score were calculated for each skin class, and the model's stability was assessed from these quantitative measures. This methodological approach helped us to obtain important findings regarding the capabilities of SkinNet-14 and resulted in the model's excellent classification accuracy, reaffirming its suitability for clinical use where high-resolution imaging may not be available.

4 Dataset description

The experiments conducted in this study utilized three publicly available skin cancer dermoscopy datasets. The HAM10000 dataset is well known for its massive number of images, making it one of the most comprehensive publicly available datasets for skin disease classification and diagnosis. Its vastness and variety help to improve the accuracy of machine learning models. Meanwhile, the ISIC dataset is a comprehensive resource that includes clinical
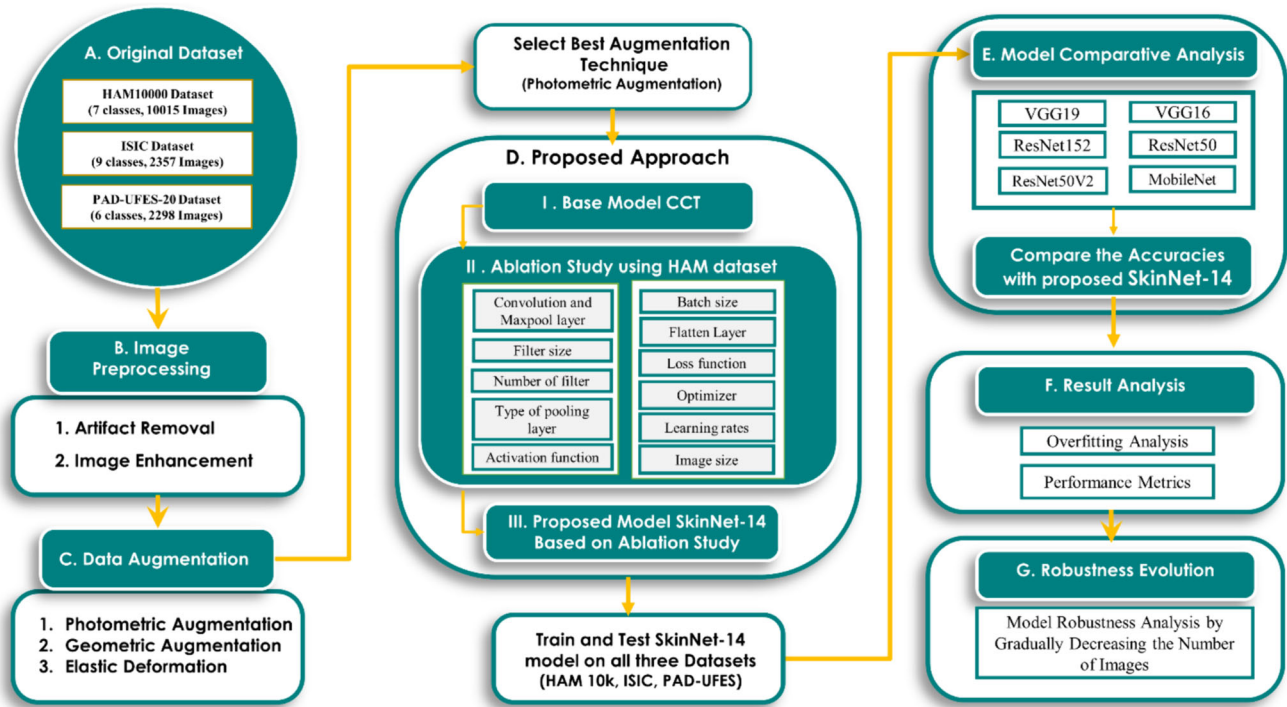
Fig. 1 Overall methodology to classify multiclass skin disease on three datasets

and dermoscopic pictures for a wide spectrum of skin disorders and conditions. This dataset is popular because of its wide range and depth of coverage. Finally, while being smaller than the HAM10000 and ISIC datasets, the PAD-UFES dataset is nevertheless commonly employed in model development because to its high quality and accuracy. The dataset is popular because it includes images with different levels of difficulty, which can help improve the robustness of machine learning models. Below are the descriptions of the datasets.

4.1 HAM10000 dataset

The HAM10000 (Human against machine with 10,000 training images) [37] dataset is a popular publicly available Kaggle dataset. The dataset consists of 10,015 skin lesion images. It has seven classes including Basal Cell Carcinoma (514 images), Benign keratosis (1,099 images), Actinic Keratosis (327 images), Dermatofibroma (115 images), Melanoma (1,113 images), Melanocytic Nevi (6,847 images) and Vascular Lesions (142 images). The image resolution of this dataset is 644×450 pixel.

4.2 ISIC dataset

The ISIC-2019 (International Skin Imaging Collaboration) [38] dataset contains 2357 images which were collected from the Kaggle database. The dataset contains nine

classes: Actinic Keratosis (114 images), Basal Cell Carcinoma (376 images), Pigmented Benign Keratosis (462 images), Dermatofibroma (95 images), Melanoma (438 images), Nevus (357 images), Seborrheic Keratosis (77 images), Squamous Cell Carcinoma (181 images) and Vascular Lesions (139 images). The image resolution of this dataset is 600×450 pixels.

4.3 PAD-UFES-20 dataset

The PAD-UFES-20 (Dermatology image database from Federal University of Espirito Santo) [39] is comprised of 2,298 images that are classified into six distinct categories. This dataset comprises six classes of skin lesions, namely Basal Cell Carcinoma (845 images), Melanoma (52 images), Nevus (244 images), Seborrheic Keratosis (192 images), Actinic Keratosis (730 images) and Squamous Cell Carcinoma (235 images). The image resolution of this dataset is 1050×1050 pixels.

4.4 Skin lesion description

The dataset includes images of numerous skin lesions. Three primary forms of skin cancer are squamous cell carcinoma, melanoma and basal cell carcinoma [40]. Squamous cell carcinoma and Basal cell carcinoma are the most common types of non-melanoma skin cancers. Basal cell carcinoma, the most prevalent form of skin cancer,

exhibits a slow growth rate and rarely metastasizes. Compared to basal cell carcinoma, squamous cell carcinoma spreads more quickly and deeply. Melanomas, which are malignancies based on melanocytes, are inherently malignant. The most aggressive form of skin cancer, melanoma, can spread to other organs and is extremely difficult to treat. Figure 2 shows the images of each skin class marked according to where an image contains tumor, artifacts or normal skin.

Some details on different skin cancers are Actinic keratosis (Fig. 2A) might appear differently as a rough, dry or scaly skin patch, on the top layer of skin, or a patch or bump that is flat to slightly elevated. In certain instances, it presents a rough, wart-like surface, accompanied by bleeding and itching. Basal cell carcinoma (Fig. 2B) causes skin changes such as growths or sores that will not heal. Lesions are typically characterized by a shiny, transparent, skin-colored lump, a brown, blue or black 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. 2C) 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. In (Fig. 2D), the size of dermatofibromas is shown to range from 0.5 to 1.5 cm in diameter. The appearance of dermatofibroma varies from pink to light brown on people with fair skin to dark brown or black on people with darker skin, while certain lesions appear paler

in the middle. Although dermatofibromas rarely exhibit symptoms, they can occasionally be tender, painful or irritating. Melanocytic nevi (Fig. 2E) typically grow to a maximum size of 40 cm. They present as tan to black in color and can become lighter or darker with time. The surface of a nevus can be smooth, uneven, elevated, thickened or bumpy; and it can differ across the nevus and alter with time. Skin around a nevus is frequently dry, prone to irritation and itching. Melanomas (Fig. 2F) are usually asymmetric with an uneven or irregular border. The diameter of a melanoma mole is larger than 6 mm and usually presents with an uneven color. The mole size and color change over time and can evidence bleeding or itching. Nevus (Fig. 2G) normally has a round smooth mole, with a single color. Common nevi can appear tan, brown or pink, and might be flat or dome shaped. Typical nevi manifest as benign clusters of colored cells. Pigmented benign keratosis (Fig. 2H) and seborrheic keratosis (Fig. 2I) are similar and may appear as an oval growth with a minor raised section, or as a flat growth. The average size of a nevi mole is 2.5 cm in diameter, and it may have a single or many growths ranging from tan to brown or black. Vascular lesions (Fig. 2J) appear dark to brilliant red in color and can cause the breakdown of the skin surface, leading to bleeding and/or infection [42, 43]. 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.
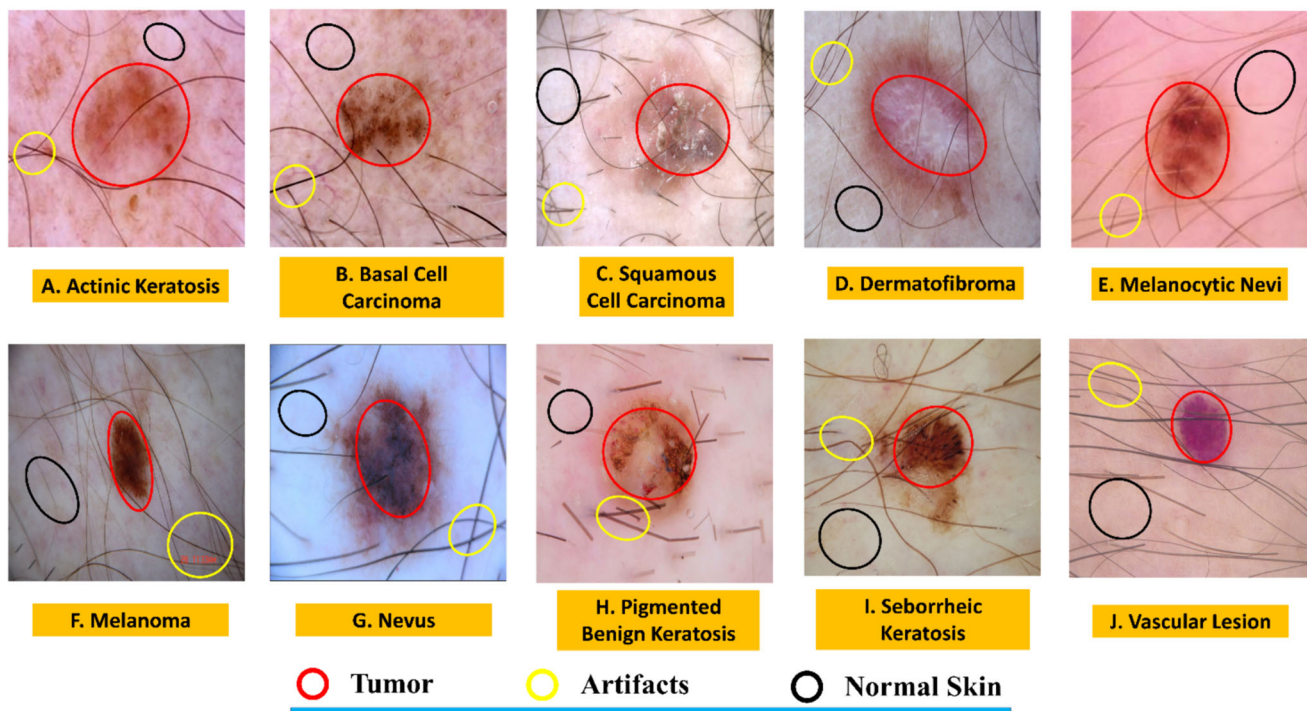

Fig. 2 Cancer lesions of different skin classes [41]

5 Image preprocessing

Preprocessing images before putting them into a neural network optimizes model performance. Morphological opening is used to remove artifacts and focus on the region of interest (ROI). This technique removes artifacts by eroding the boundaries of objects and then dilating them, effectively eliminating small unwanted elements while preserving the overall structure of the image. Following that, NLMD is used to efficiently minimize noise while keeping crucial image information. It compares similar patches in the image to estimate the noise level before applying filtering. CLAHE then increases the image's brightness and contrast by equalizing the histogram in specific locations. It improves the visibility of critical details while reducing noise over amplification. Furthermore, the Gaussian Blur method smooths pixels in an image while maintaining the edges, which assists to minimize noise and improve overall visual appeal. Finally, the processed image is downsized to a standard 224×224 resolution. We experimented with some more strategies and the described processing combinations increased the dermoscopy image quality, allowing for more accurate analysis and subsequent classification tasks. In this step, all the image preprocessing techniques are applied to each of the datasets. Figure 3 shows the complete image preprocessing steps.

5.1 Removal of artifact

Morphological opening is a technique that eliminates all single-pixel artifacts, such as noisy spikes and tiny spurs, and blackens small objects [44]. The process of applying morphological opening to an image involves first converting it into binary format. The conversion to binary format amplifies the visibility of small noises. The application of morphological opening to a binary image is achieved through the use of a kernel. The characteristics of the artifact to be removed determine the shape and size of the kernel. After conducting experiments with various kernel sizes, a 10×10 kernel size is chosen. The successful erasure of unwanted objects while preserving essential image information is achieved through the use of a specific kernel size. The noise-free binary mask is then merged with the original picture using the "bit-wise AND" function. In the context of binary images, "bit-wise AND" and "logical AND" perform the same function. Figure 3 shows the image after performing artifact removal on the original image.

5.2 Image enhancement

The precise classification of various dermoscopy images is challenging due to the complexity of their details and the presence of hidden information. For the best performance, appropriate image enhancement methods assist in adjusting
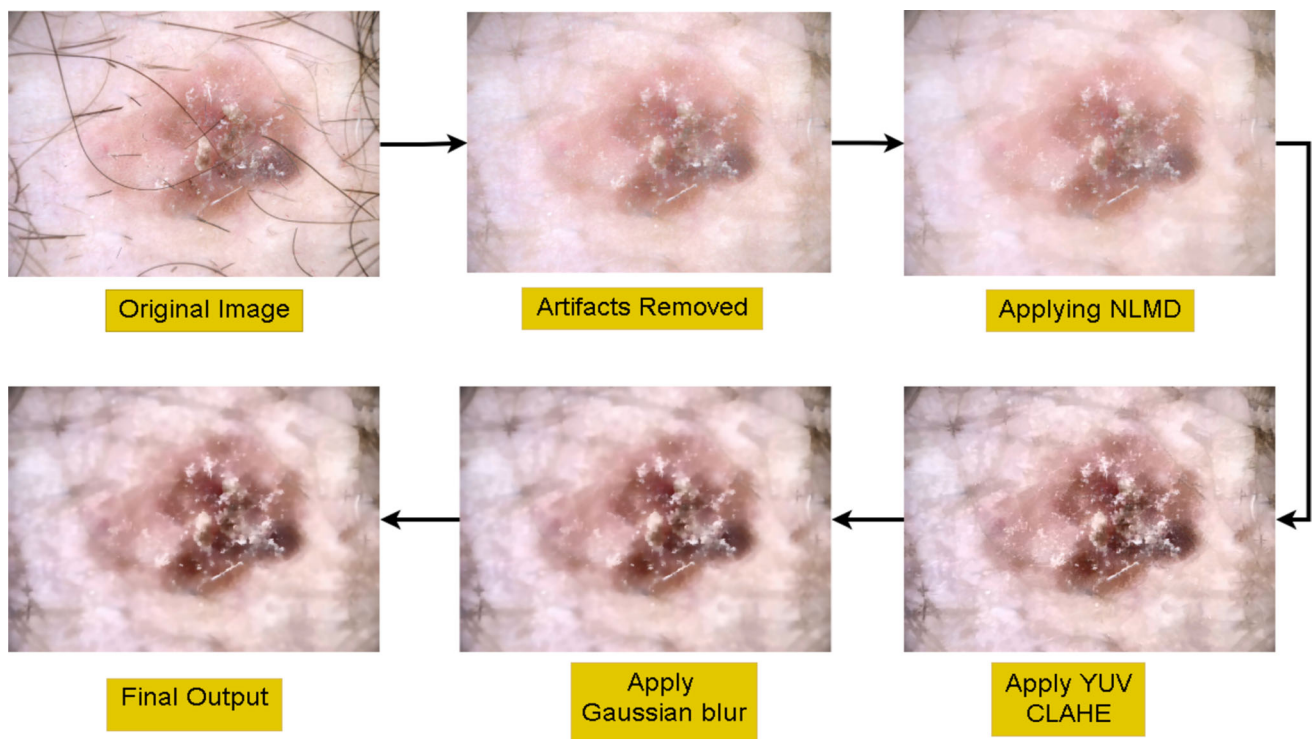

Fig. 3 Output image after each preprocess stage

Table 1 Counts of original versus augmented images

Dataset	Data count without augmentation	Data count after augmentation		
		Geometric	Photometric	Elastic Deformation
HAM10000	10,015	50,075	50,075	50,075
ISIC	2,357	11,785	11,785	11,785
PAD-UFES-20	2,298	11,490	11,490	11,490

Table 2 Ablation study on various augmentation techniques on HAM10000 dataset

Technique	Parameters	Size of Image	Training time × epoch	Test accuracy (%)	Findings
Without Data Augmentation	0.41 M	32 × 32	09s × 100	77.23%	Poor accuracy
Geometric	0.41 M	32 × 32	16s × 100	87.68	Poor accuracy
Photometric	0.41 M	32 × 32	16s × 100	89.85	Best accuracy
Elastic Deformation	0.41 M	32 × 32	15s × 100	83.83	Poor accuracy

the visual contrast between regions of interest (ROIs) and backgrounds.

5.2.1 Non-local means denoise (NLMD)

NLMD is implemented to reduce the noise of the images. The NLMD algorithm [45] is based on the principle of replacing pixel color with the average of the colors of neighboring pixels. This significantly improves post-filtering clarity with less loss of image detail than local mean methods. The denoising of an image $z = (z1; z2; z3)$ in channel i to the pixel j is executed as follows [45]:

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

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

The notation $B(x, r)$ denotes the area of pixels x within a given radius r. The determination of weight $\omega(x, k)$ is performed by the squared Frobenius norm distance within color patches with centers at x and k that degrade under a Gaussian kernel. The OpenCV cv2.fastNlMeansDenoisingColored() function is utilized to execute the NLMD. As the source image, we use the image that resulted following morphological opening. The options available for filter strength tuning in the luminance and color components are h and h Color, respectively. Our proposed method reduced h value to accurately preserve detail. The utilization of the recommended values for the parameters template Window Size and search Window Size, specifically 7 and 21, is implemented. Figure 3 presents the image after applying NLMD from artifact removed images.

5.2.2 Contrast limited adaptive histogram equalization (CLAHE)

CLAHE [46] is performed to rectify excessive contrast amplification and restore overall contrast balance. The determination of contrast enhancement in CLAHE near a specific pixel value is based on the slope of the transformation function. The kernel size for applying CLAHE is 10 × 10, the clip limit is 2.0, and tile grid size is 8 × 8. The color space used for this preprocess is YUV. Figure 3 shows the image after applying CLAHE on NLMD applied images.

5.2.3 Gaussian blur

Gaussian blurring [47] is used in image processing to minimize noise and eliminate speckles from an image. It is essential to remove 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 formula is:

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

The variables i, j and σ are utilized to, respectively, denote the horizontal axis distance from the origin, the vertical axis distance and the standard deviation of the Gaussian distribution. The point (0, 0) serves as the origin for these axes. The formula generates a two-dimensional surface consisting of concentric circles exhibiting a Gaussian distribution as they move away from the central point. The application of Gaussian blur can be achieved through the utilization of the cv2.bilateralFilter() function in OpenCV. The diameter of each pixel neighborhood is set to 9, and sigmaColor and sigmaSpace values are set to 75. In
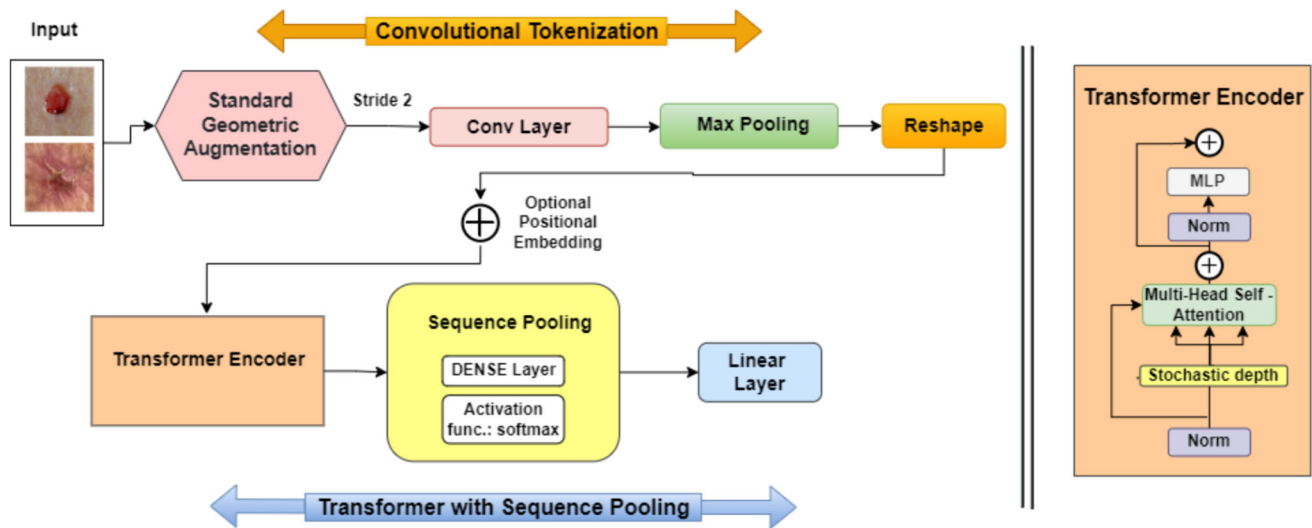
Fig. 4 Basic structure of Compact Convolutional Transformer (Base Model)

the blurring process, Sigma Color represents the range filter and Sigma Space represents the spatial filter. These parameters provide more control over the blurring process, especially in retaining edges and features while minimizing noise in the image. Figure 3 shows the image after applying Gaussian blur on CLAHE applied images.

The preprocessing processes for the HAM10000 dataset were carried out on images with an original size of 644×450 pixels. Similarly, for the ISIC dataset, the original image size was 600×450 pixels. As for the PAD-UFES dataset, the images had an original size of 1050×1050 pixels. After applying the Gaussian blur algorithm, all images were resized to a standard size of 224×224 pixels. This resizing phase ensured consistency among datasets and assisted in subsequent analysis and classification operations. The image size is downsized to 32×32 pixel during insertion in the model and its discussion can be found in Sect. 7.2. Several segmentation approaches, such as thresholding techniques, edge detection and region-based segmentation, are experimented to separate skin lesions from the surrounding skin and the background. However, the segmentation procedures did not increase our model's performance. One reason could be that our model, SkinNet-14, is built to handle a wide range of image qualities and backgrounds without requiring a completely clean input. The architecture, involving convolutional and transformer layers, has shown the ability to focus on relevant features by effectively learning from the entire image context, including subtle signals in the background that could potentially be significant for accurate classification.

6 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 as annotated data is both expensive and sparse. Data augmentation is essential as it increases the quantity of labeled data. In this study, three different data augmentation techniques, photometric augmentation, geometric augmentation and elastic deformation, were applied with the optimal technique selected based on highest model performance.

6.1 Geometric data augmentation

One of the most common augmentation methods employed to increase the quantity of data is geometric transformation [48]. Geometric augmentation is the process of modifying the geometric shape of an image by changing the values to their matching new values. It is a successful image enhancement method that changes the shape of an image without affecting image quality. With several geometric augmentation methods available for medical imaging, this study applied vertical flipping, which is used on matrices to flip the rows and columns vertically, and horizontal flipping, which allows the image to be flipped either to the left or to the right. Vertical and horizontal flipping maintain the natural horizontal–vertical column structure and rotation of an image while rotating the image to any degree. Our geometric augmentation study used four different techniques named vertical flipping, horizontal flipping and rotation (clockwise and anticlockwise 90). No additional parameters are required in geometric techniques; they flip
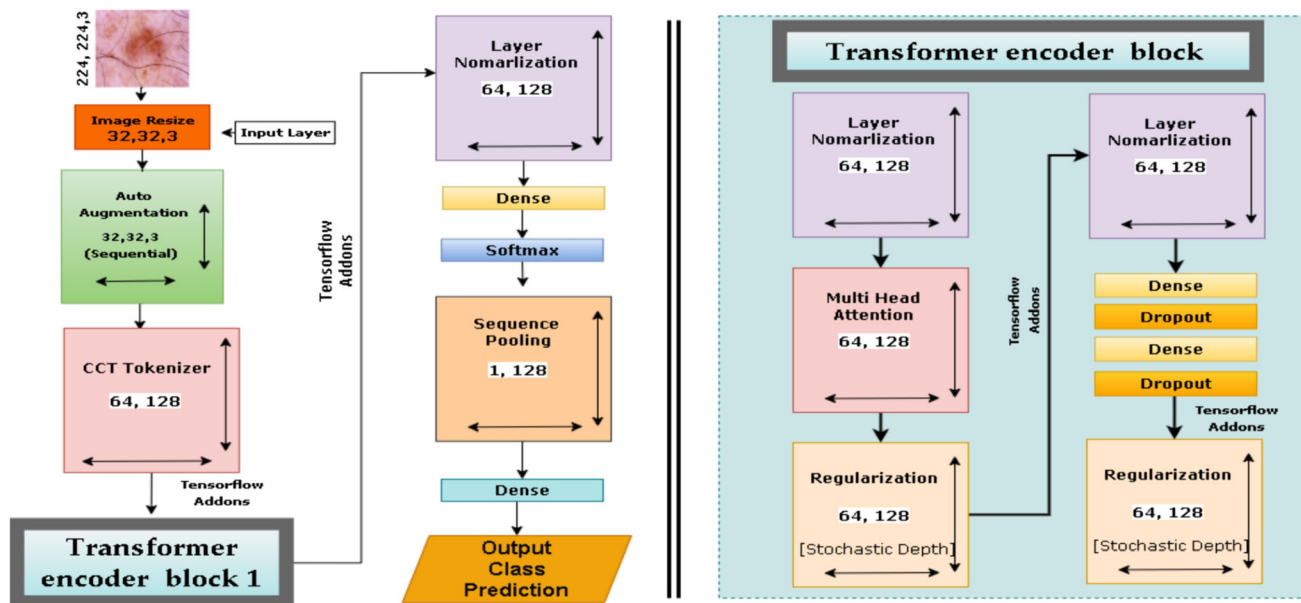
Fig. 5 Proposed SkinNet-14 model architecture

the image horizontally and vertically, and rotate according to the given angles.

6.2 Photometric data augmentation

Photometric augmentation involves modifying pixel values such as brightness, sharpness, blurriness, color and contrast. Photometric augmentation transforms red, green and blue (RGB) channels by shifting the (r, g and b) value of each pixel to a new pixel (r', g' and b') value. It mainly alters visual color and lighting, not geometry [49]. This process must be carried out in such a way that critical pixel information is not lost. In this study, four different types of photometric techniques are used: Changing the brightness, by maintaining the level of lightness or darkness, the contrast, by making the light regions lighter and the dark regions darker, the color, by changing the color balance of an image, and sharpness, by sharpening the details of an image, produced the best results as a photometric augmentation in this paper. This study used several factor values to increase image count without compromising pixel information. The 1.5 factor values generated optimum augmentation of the images.

6.3 Elastic deformation

Elastic deformation [50] for data augmentation stretches and changes the shape of images differently according to skin location and compression strength.

There are two steps involved in obtaining 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 $v \times [0.5, 0.5]$ is selected consistently for each pixel in each direction. A Gaussian filter is applied to the resulting horizontal and vertical images independently (Eq. (4) and (5)) to ensure that nearby pixels have equal displacement. The transformations contain the maximum value for the initial random displacement (v) and the degree of smoothing, which is determined by the Gaussian filter standard deviation (σ). Based on the overall appearance of the patches, a value of = 300 and a value of 20 for deformation were selected. Then, the image segmentation mask is 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 [50].

$$\Delta_a = G(\sigma) * (v \times \text{Rand}(w, z)) \tag{4}$$

$$\Delta_b = G(\sigma) * (v \times \text{Rand}(w, z)) \tag{5}$$

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

Here, I and $I_{transform}$ are the original and transformation images, respectively; w and z are the dimensions of the skin cancer image. Using this technique, the same image of the skin lesion will be visible, but it will be deformed and appear stretched, without losing any important details.

6.3.1 Augmented dataset and augmentation result

From dataset description, it is clearly visible that the data from each class is grossly unbalanced. So, geometric, photometric and elastic deformations are applied to

Table 3 Ablation studies on modifying transformer layer, pooling layer, stride size, activation function

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

No	Activation function	Parameters	Training time × epoch	Test accuracy (%)	Findings
Modification 2: Activation function changes					
1	softplus	0.24 M	10 s × 200	88.97	Poor accuracy
2	softsign	0.24 M	10 s × 200	90.88	Almost good accuracy
3	elu	0.24 M	11 s × 200	90.38	Almost good accuracy
4	relu	0.24 M	10 s × 200	91.24	Best accuracy
5	Tanh	0.24 M	10 s × 200	89.55	Poor accuracy

No	Pooling layer types	Parameters	Training time × epoch	Test accuracy (%)	Findings
Modification 3: Pooling layer changes					
1	Average	0.24 M	7 s × 200	91.24	Good accuracy
2	Max	0.24 M	7 s × 200	92.37	Best accuracy

No	Strides numbers	Parameters	Training time × epoch	Test accuracy (%)	Findings
Modification 4: Stride size changes					
1	1	0.24 M	7 s × 200	93.57	Best accuracy
2	2	0.24 M	4 s × 200	91.14	Almost good accuracy
3	3	0.24 M	4 s × 200	91.37	Almost good accuracy
4	4	0.24 M	4 s × 200	89.63	Poor accuracy

increase the dataset. Table 1 shows the augmented image counts for each class.

The generated images are investigated by training the base CCT model, and test accuracy results are shown in Table 2.

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.

The result without data augmentation is not satisfactory, which might be because of the class imbalance problem. Without augmentation, the model may develop biases, performing well on overrepresented classes and poorly on underrepresented ones.

7 Proposed model

ViT is recognized in computer vision studies for outperforming CNN models in computing efficiency and training time. The ViT encoder–decoder blocks process multiple consecutive datasets faster. Self-attention allows finding long-distance linkages between items. This results in improved image categorization [51, 52]. Since ViTs require significant quantities of data for training, most medical datasets are not adequate for training purposes. CCT, a ViT–convolution hybrid, addresses this issue [12]. CNN blocks patch the CCT local receptive field, which maintains image data. Self-attention identifies visual portions and merges similar data.

7.1 Compact convolutional transformer (CCT)

CCT architecture consists of two main blocks. One is transformer along sequence pooling, and the other is con-

Table 4 Ablation studies on modifying kernel size, batch size, loss function

No	Kernel size count	Training time × epoch	Parameters	Test accuracy (%)	Findings
Modification 5: Kernel size changes					
1	4	8 s × 200	0.3 M	93.83	Good accuracy
2	3	7 s × 200	0.24 M	94.77	Best accuracy
3	2	9 s × 200	0.2 M	93.57	Good accuracy
4	1	10 s × 200	0.17 M	88.33	Poor accuracy

No	Loss Function	Training time × epoch	Parameters	Test accuracy (%)	Findings
Modification 6: Loss function changes					
1	Binary cross-entropy	7 s × 200	0.24 M	94.88	Good accuracy
2	Categorical cross-entropy	7 s × 200	0.24 M	95.80	Best accuracy
3	Mean squared error	7 s × 200	0.24 M	94.81	Good accuracy
4	Mean absolute error	7 s × 200	0.24 M	94.63	Good accuracy
5	Mean squared logarithmic error	7 s × 200	0.24 M	28.76	Poor accuracy

No	Batch size	Training time × epoch	Parameters	Test accuracy (%)	Findings
Modification 7: Batch size changes					
1	256	6 s × 200	0.24 M	94.09	Good accuracy
2	128	7 s × 200	0.24 M	96.68	Best accuracy
3	64	11 s × 200	0.24 M	95.56	Good accuracy
4	32	16 s × 200	0.24 M	95.30	Good accuracy

volutional tokenization. The CCT methodology is shown in Fig. 4.

Convolutional Tokenization generates image patches [12]. Convolutional Tokenization processes for image z using the following formula:

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

Here, Conv2D is convolutional layer which includes 64 filters with 2 strides and the ReLU. Maxpool then downscales Conv2D feature maps. Images of any size can be processed by convolutional tokenization. The use of convolutional patches in CNN layers assists in preserving regional spatial Information.

Following this, the first block image patches are sent to the transformer encoder block. The encoder block contains multilayer perceptron (MLP) and multihead self-attention (MSA) head. GELU activation, dropout and layer normalization are utilized in the transformer encoder.

The sequence pooling layer utilizes sequence pooling to gather the output of the transformer backbone. This thesis explores the use of sequence pooling to enhance data correspondence for input by assessing the sequential embeddings of latent space generated by the encoder. The

sequence pooling layer is responsible for collecting every bit of data, as it effectively captures necessary details from multiple regions within the input image. The term "mapping transformation" is used to describe this particular method.

Finally, the images from the second dimension are then categorized after passing through a linear classification layer.

7.2 Architecture of the base model

This part presents a skin classification model SkinNet-14. The model is developed by doing ablation studies on the architecture of the CCT model.

The CCT architecture is made up of an input layer, a data augmentation layer using different geometric augmentation techniques, a CCT tokenizer, regularization layers, multihead attention layers, dense layers, pooling layers, dropout layers and output dense layers. The $224 \times 224 \times 3$ sized image passes through a resizing process to convert into $32 \times 32 \times 3$ sized images. Then, the auto-augmentation operates on images with $32 \times 32 \times 3$ input dimensions. Auto-augmentation

Table 5 Ablation studies on modifying optimizer, image size, learning rate

No	Optimizer	Training time × epoch	Parameters	Test accuracy (%)	Findings
Modification 8: Optimizer changes					
1	Adam	7 s × 200	0.24 M	96.68	Best accuracy
2	Nadam	7 s × 200	0.24 M	88.62	Poor accuracy
3	SGD	7 s × 200	0.24 M	94.46	Good accuracy
4	Adamax	7 s × 200	0.24 M	95.23	Good accuracy
5	RMSprop	7 s × 200	0.24 M	94.48	Good accuracy

No	Learning rate	Parameters	Training time × epoch	Test accuracy (%)	Findings
Modification 9: Learning rate changes					
1	0.01	0.24 M	7 s × 200	92.23	Poor accuracy
2	0.006	0.24 M	7 s × 200	95.47	Good accuracy
3	0.001	0.24 M	7 s × 200	97.85	Best accuracy
4	0.0008	0.24 M	7 s × 200	96.68	Good accuracy

No	Image size	Parameters	Training time × epoch	Test accuracy (%)	Findings
Modification 10: Image size changes					
1	64	0.24 M	24 s × 200	96.17	Near best accuracy
2	32	0.24 M	7 s × 200	97.85	Best accuracy
3	28	0.24 M	6 s × 200	95.17	Good accuracy
4	16	0.24 M	5 s × 200	94.88	Good accuracy

generates augmented representations of images during training in order to boost the performance of the CCT model. It generates additional training samples with various augmentations, enhancing the model's ability to recognize objects from different perspectives and lighting conditions. This leads to better generalization performance. After passing through auto-augmentation, the CCT Tokenizer block receives enhanced images as input, which are subsequently downsized to 64×128 to produce the output image. The stride and kernel sizes for the convolutional layer tokenizer block are set to 2 and 4 correspondingly, coupled with a kernel size of 4 for the pooling layer. Before the data is delivered to the transformer encoder block, there are tokenization and tensorflow additions. Two sets of dense and dropout layers consisting a ratio of 0.1, followed by the second layer normalization, multihead attention, regularization and the first layer normalization, comprise the layers in the sequence stated. The final layer of the transformer encoder block is connected to another regularization layer. The output, with a size of 64×128, is regularized using the regularization layer. The application of a second transformer encoded block that is comparable to the first, follows next. Then, two additional layers are applied: a regularization layer and a normalizing layer. A dense layer, using the softmax function, produces outputs with a dimension of 64×1, which is then normalized. The sequence pooling layer then receives this and generates output data with a dimension of 1×128. Finally, different groups of skin cancer images are classified by employing a linear classification layer.

In the proposed model, one transformer encoder block is removed. Figure 5 depicts the proposed model architecture that has been generated after the ablation study on the base model.

7.3 Ablation study

As discussed earlier, to optimize the performance of the CCT network, we modify the layer architecture and adjusting the hyper parameter values through an ablation study. This process involved the conduction of ten ablation studies. After completing all ablation investigations, the proposed SkinNet-14 network is established with a more reliable design, improved functionality and shorter processing time.

7.4 *Proposed SkinNet-14 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 (see Sect. 7.2). Figure 5 shows that the SkinNet-14 model has one transformer encoder block, while the CCT architecture has two. This

enables faster training and smaller model size. The architecture remains unchanged, with the exception of some model hyperparameters such as kernel size and stride size. Apart from the data augmentation and softmax functions in the dense layer, the proposed model has 14 layers in its architecture, which inspired the model's name: SKinNet-14.

Positional encoding is not necessary for model function, which reduces processing costs. The computational complexity of self-attention is $O(n^2.d)$, where n is the length of

Fig. 6 Test accuracy increasing over 10 ablation studies
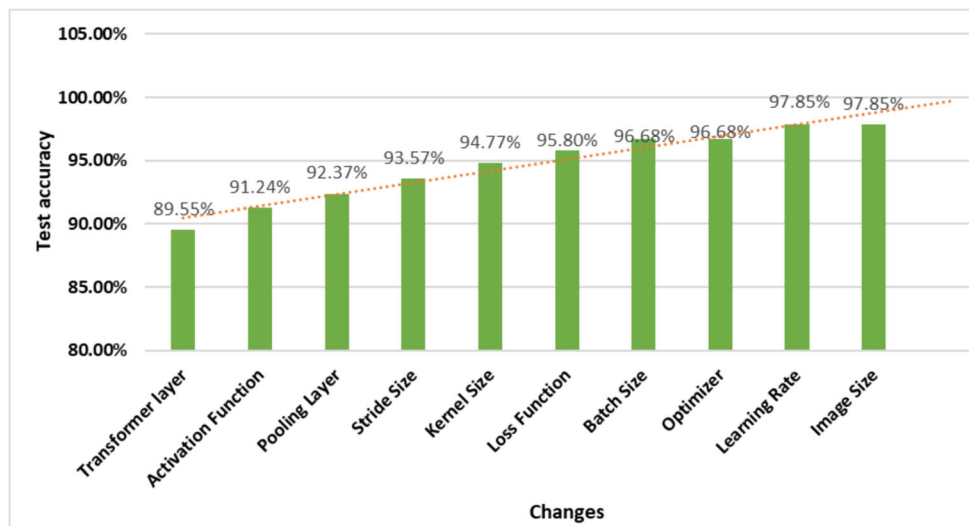
Table 6 Different matrices calculated for SkinNet-14 model performance evaluation

	Skin Class	Precision	Recall	F1-score	Test Accuracy (%)	
HAM10000 dataset	Actinic Keratosis	0.99	0.98	0.99	99.91	Test Accuracy:
	Basal Cell Carcinoma	1.00	0.99	0.99	99.95	97.85%
	Benign Keratosis	1.00	0.99	0.99	99.84	
	Dermatofibroma	1.00	1.00	1.00	100.00	
	Melanoma	0.99	0.98	0.98	99.62	
	Melanocytic Nevi	0.98	0.99	0.99	98	
	Vascular Lesions	0.18	0.04	0.06	98.45	
ISIC dataset	Actinic Keratosis	1.00	1.00	1.00	100.00	Test Accuracy:
	Basal Cell Carcinoma	1.00	1.00	1.00	99.94	96.01%
	Dermatofibroma	1.00	0.99	0.99	99.94	
	Melanoma	0.91	0.90	0.90	96.19	
	Nevus	0.98	0.99	0.99	99.58	
	Pigmented Benign Keratosis	1.00	1.00	1.00	100	
	Seborrheic Keratosis	0.50	0.52	0.51	96.55	
	Squamous Cell Carcinoma	1.00	0.99	1.00	100	
	Vascular Lesions	1.00	1.00	1.00	99.94	
PAD-UFES-20 dataset	Actinic Keratosis	0.98	0.98	0.98	99.02	Test Accuracy:
	Basal Cell Carcinoma	0.99	0.98	0.98	98.73	98.14%
	Melanoma	0.93	1.00	0.96	99.83	
	Nevus	0.99	0.99	0.99	99.83	
	Seborrheic Keratosis	0.99	0.99	0.99	99.13	
	Squamous Cell Carcinoma	0.93	0.97	0.95	99.88	

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 [10]. Because the SkinNet-14 model does not require positional encoding, the training and testing phases use fewer resources. Additionally, the transformer backbone only uses self-attention mechanism. Consequently, the model is significantly more efficient.

Transformer encoder blocks are computationally intensive, therefore reducing a block decreases the complexity, which might lead to faster training and inference times [10]. Additionally, reducing model complexity might be counterbalanced by incorporating architectural features that preserve important information and improve the accuracy.

7.5 Training strategy

The parameters of the base CCT models are: transformer layer = 3, kernel size = 2, learning rate = 0.0008, optimizer = adam, batch size = 64, loss function = mean squared error, pooling layer = average and activation function = tanh. For the PAD dataset, 400 epochs, and 200 epochs for the HAM10000 and ISIC datasets, are used. Several experiments are conducted before deciding on the number of epochs. The split ratio of each skin cancer dataset is 75%, 10% and 15% for training, validation and testing sequentially. During an ablation study, these are the initial parameters that are gradually adjusted by multiple experiments, as shown in Sect. 8.2. Categorical cross-entropy is employed initially as this is the standard loss
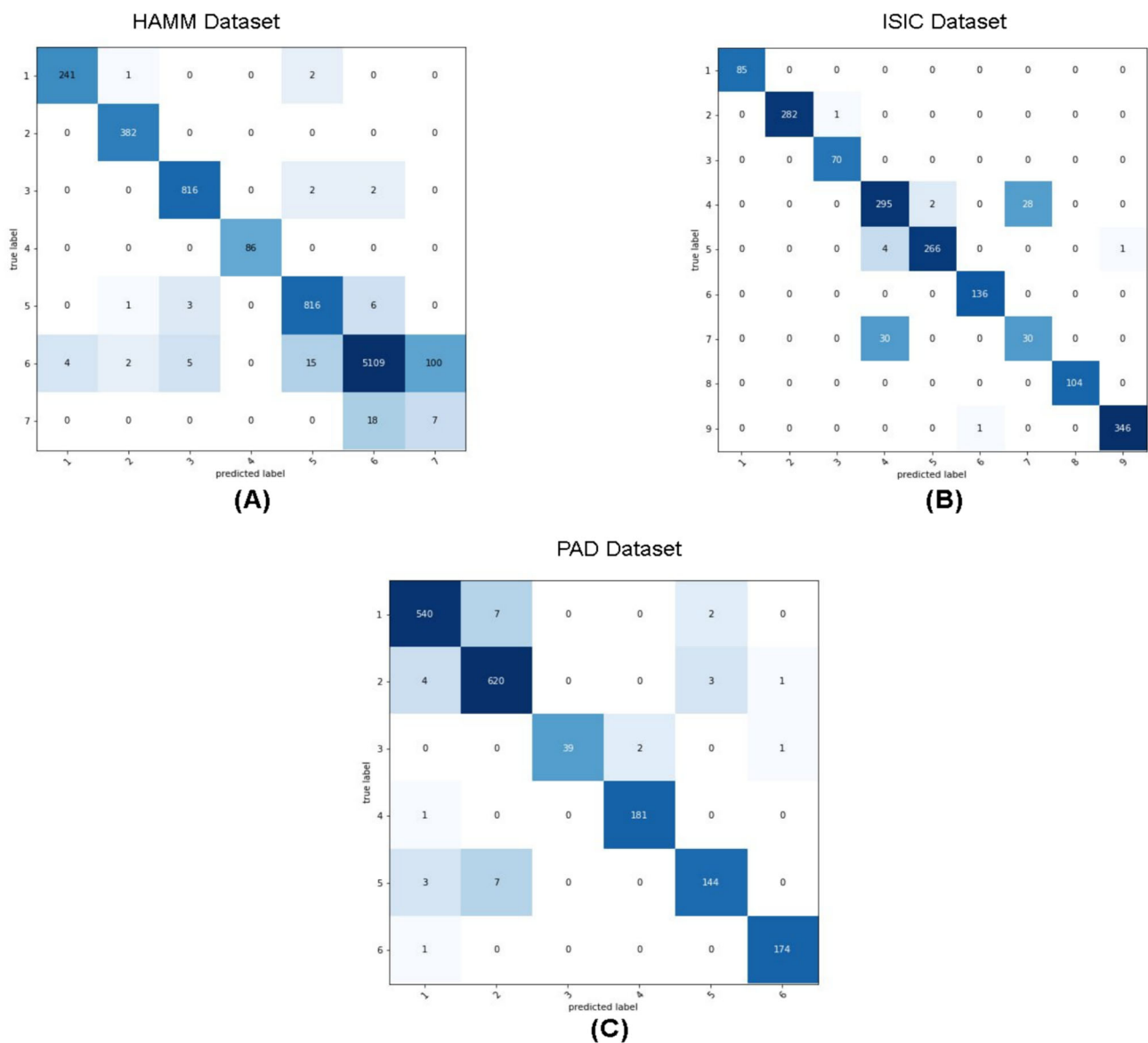

Fig. 7 Confusion matrix for the proposed SkinNet-14 model on: **A** HAM10000 dataset, **B** ISIC dataset, **C** PAD dataset
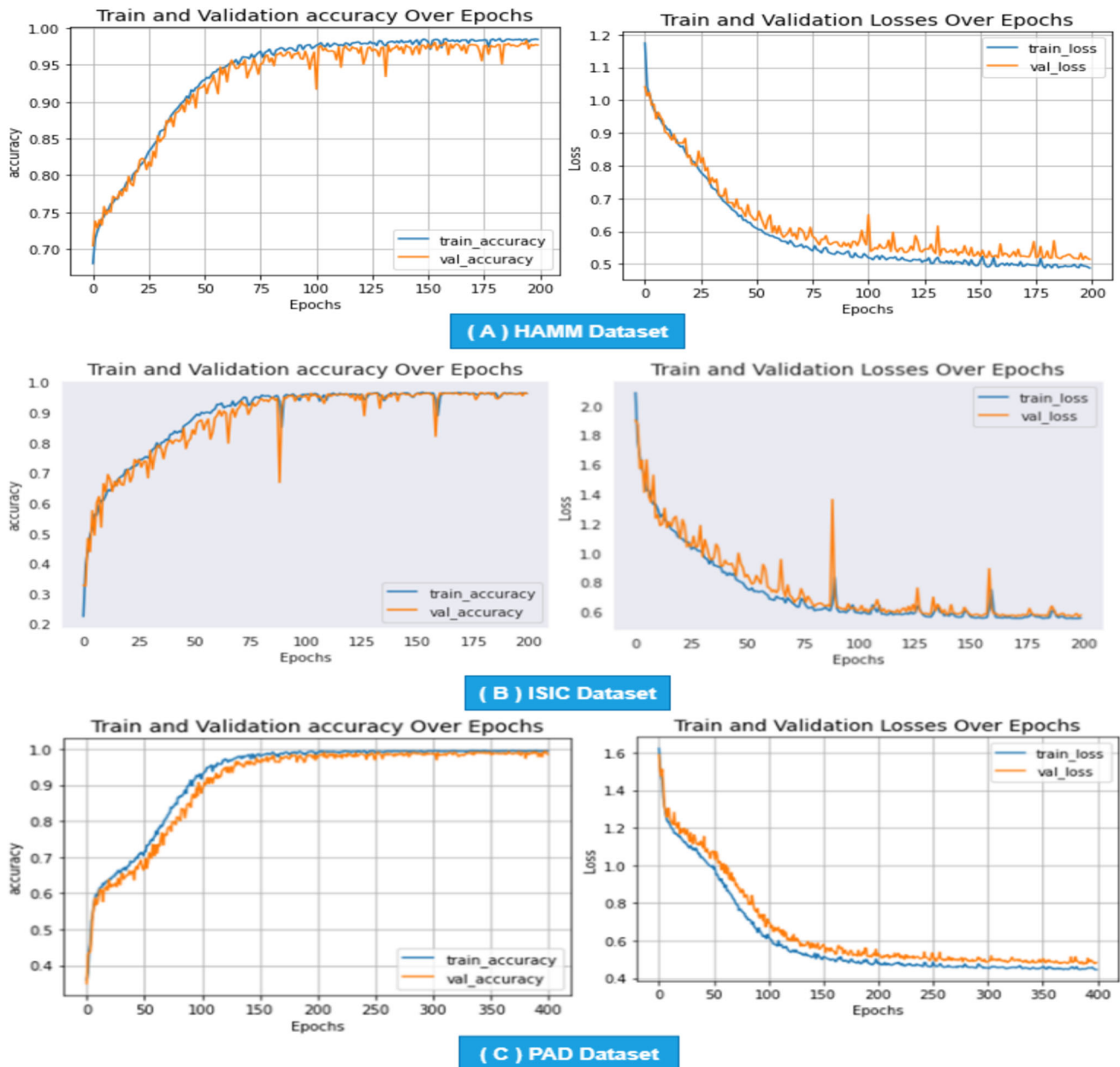
Fig. 8 Accuracy curve and loss curve of SkinNet-14 model on: **A** HAM10000 dataset, **B** ISIC dataset, **C** PAD dataset

function in multiclass instances [53]. The same configuration is considered while training the transfer learning models. In order to test different models and configurations, we engaged three computers, each with an Intel Core i5-8400 processor, 16 GB of RAM, an NVidia GeForce GTX 1660 GPU and a 256 GB DDR4 SSD for storage.

7.6 Transfer learning models

We compare the performance of multiple transfer learning models which 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 HAM10000 and 200 for ISIC. The epoch numbers are chosen following several experiments.

7.6.1 VGG architecture

This study use the Visual Geometry Group (VGG) networks [54], specifically VGG16 and VGG19. VGG16, having 16 weighted layers, is a cutting-edge transfer learning algorithm that achieves an accuracy of 92.7% on the ImageNet dataset. Because the VGG model has more
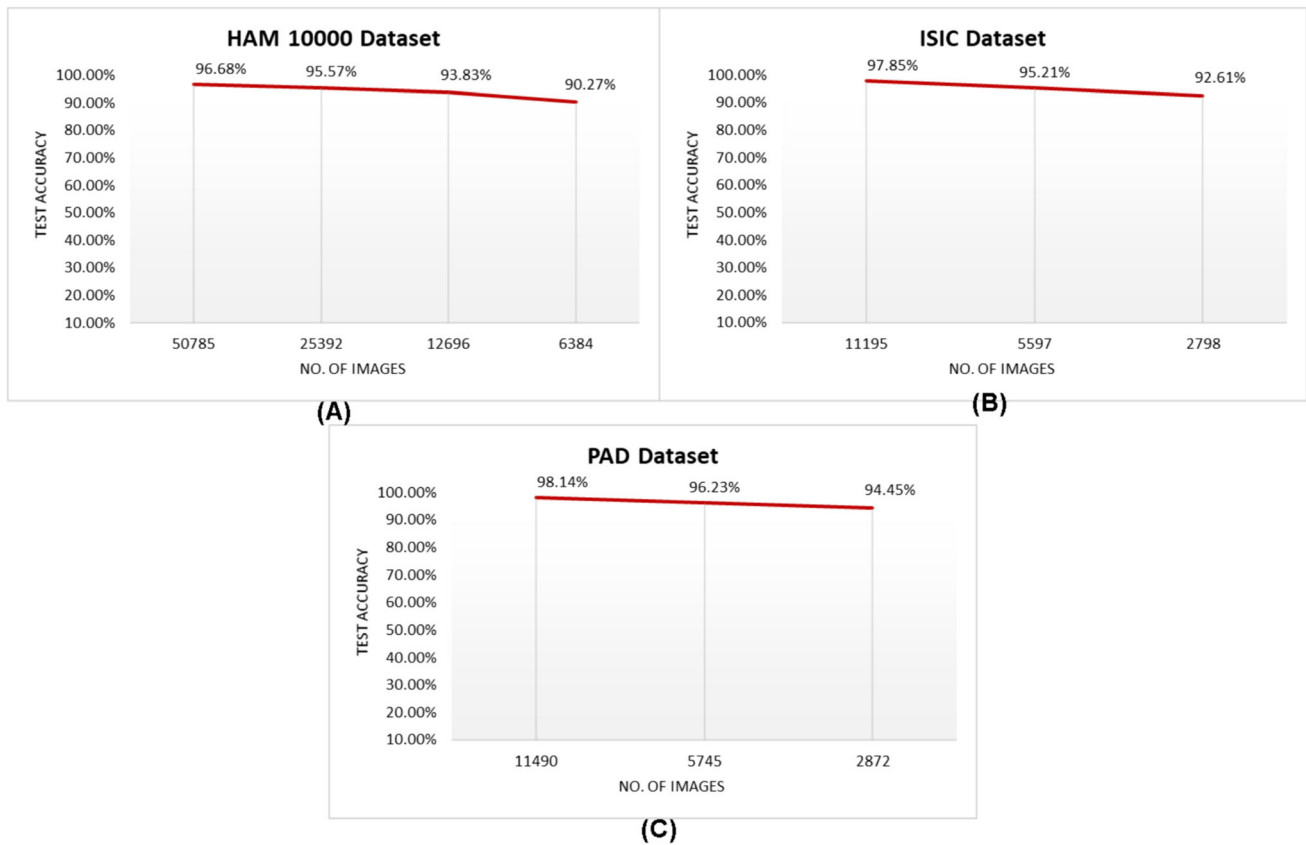
Fig. 9 Testing the performance of the SkinNet-14 using reduced images on: **A** HAM10000 dataset, **B** ISIC dataset, **C** PAD dataset

depth, it can assist the kernel in learning more complicated features. There are five max-pool layers, thirteen conv layer and three dense layer in VGG16.

VGG19 is another VGG model version consisting 19 weighted layers. In the convolutional layers, the ReLU activation function is applied.

7.6.2 ResNet architecture

Residual networks (ResNets) [55] 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 max-pool and an average pool layer make up this architecture. The selected model has 23 million trainable parameters.

ResNet152 has 152 layers. ResNet is an easy-to-optimize and effective deep learning architecture. As the network design contains multiple layers, it is time-consuming.

ResNet50V2 [55] is an altered version of the original ResNet50. ResNet50V2 outperforms ResNet50 and ResNet101 on ImageNet. ResNet50V2 modified the way in which block connections propagate.

7.6.3 MobileNet

The purpose of MobileNet [56] is to provide an effective deep neural network architecture that can perform image classification on embedded and mobile devices with high efficiency. The model's lightweight and high-performance characteristics make it suitable for applications that prioritize power and memory constraints.

MobileNet is a convolutional neural network architecture that utilizes depth-wise separable convolutions to achieve a reduction in the number of parameters and computation required, while still maintaining comparable accuracy to conventional convolutional layers. The feasibility of executing the model on devices with restricted resources is enabled.

8 Result and discussion

In this section, the results of the research are explained, including the findings 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.

Table 7 Comparison of performance with six state-of-the-art transfer learning models

Model	Parameters	HAM10000 dataset				ISIC dataset				PAD dataset			
		epochs	Per epoch time	Accuracy	F1 Score	epochs	Per epoch time	Accuracy	F1 Score	epochs	Per epoch time	Accuracy	F1 Score
VGG19	20,026,436	200	65-67 s	80.47%	81.23%	200	30-34 s	70.87%	71.43%	400	28-30 s	82.97%	83.48%
VGG16	14,716,740	200	65-67 s	81.21%	82.10%	200	30-34 s	71.21%	72.09%	400	28-30 s	81.38%	82.18%
ResNet152	58,379,140	200	65-67 s	65.79%	66.38%	200	30-34 s	75.79%	76.39%	400	28-30 s	78.79%	79.39%
ResNet50	23,595,908	200	65-67 s	69.27%	69.73%	200	30-34 s	68.57%	69.28%	400	28-30 s	72.97%	73.47%
ResNet50V2	23,572,996	200	65-67 s	66.25%	67.11%	200	30-34 s	63.21%	64.09%	400	28-30 s	77.15%	78.06%
MobileNet	3,232,964	200	65-67 s	43.42%	44.19%	200	30-34 s	49.12%	50.04%	400	28-30 s	55.48%	57.22%
SkinNet-14	241,861	200	7-8 s	97.85%	97.92%	200	2-3 s	96.01%	96.50%	400	2-3 s	98.14%	98.57%

8.1 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 identifies the negative class. False positive (FP) and false negative (FN) findings 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 metrics used in this study are given below [57].

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{8}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{10}$$

$$F_1 = 2\frac{\text{precision*recall}}{\text{precision} + \text{recall}} \tag{11}$$

8.2 Ablation study results

The optimal model architecture was achieved through a series of ablation studies, which are detailed in this section. The performance of the skin classification model is optimized through an ablation study conducted on the images of the HAM10000 dataset. Each of the parameters was chosen based on preliminary experiments and literature that suggested the impact of these parameter on deep learning models performance and we experimented with a range of values for these parameters, carefully observing changes in model performance.

Changing a small number of features can enhance the performance of a classification model. The proposed robust model is optimized through a series of experiments that involve adjusting the base model features to determine the optimal configuration. This research comprises ten distinct studies. Tables 4, 5 and 6 present the results of the ablation experiments conducted in this study.

8.2.1 Modification 1: transformer layer changes

In this research, the transformer layer is changed by varying the number of encoder blocks. Table 3 shows 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 0.24 M parameters, 21–24 min and 89.55% accuracy achieves optimum performance. This configuration has the smallest

Table 8 Comparison between the previous studies and proposed research

Author	Model	Dataset	Image resolution	Training time	Used GPU for computation	Accuracy	Limitations
Muhammad et al. [58]	Improved moth flame optimization (IMFO)	HAM10000: 10,015 images	224 × 224 × 3	2.40 h	–	90.67%	1. High computational time. 2. Images are not preprocessed
Jingye et al. [31]	MT-TransUNet	ISIC-2017: 2000 images	224 × 224	8 h	GTX1080Ti GPU with 11GB memory	91.20%	1. High computational time. 2. Low-resolution image bring difficulties for the model. 3. Images are not preprocessed
Jianpeng et al. [32]	ARL-CNN50	ISIC-2017: 2750 images	224 × 224	30 h	NVIDIA GTX Titan XP	91.7%	1. High computational time. 2. Images are not preprocessed
Xiaoyu et al. [8]	DeMAL-CNN	ISIC-2016: 1279 images, ISIC-2017: 2000 images	224 × 224	10 h, 20 h	NVIDIA GTX Titan XP GPU	87.20%, 92.10%	1. High computational time. 2. Images are not preprocessed
Ameri et al. [18]	Proposed DCNN	HAM10000: 10,015 images	227 × 227	0.27 h	Nvidia GTX 1050 GPU	84%	1. Poor Accuracy 2. Images are not preprocessed 3. Low-resolution image bring difficulties for the model
Nazia et al. [59]	Multiclass multilevel (MCML)	ISIC-2016: 1279 images	277 × 277 × 3	3 h	–	96.03%	1. High computational time 2. Limited amount of data
Parvathaneni et al. [60]	MobileNet V2-LSTM	HAM10000: 10,015 images	224 × 224 × 3	0.29 h	RADEON (TM) 530 Graphics 8 GB memory	85.34%	1. Low-resolution image bring difficulties for the model 2. Poor Accuracy 3. Images are not preprocessed
Tembhurne1 et al. [19]	Proposed Ensemble Model	ISIC: 2637 images	224 × 224	–	Nvidia K80 GPU	93%	1. Low-resolution image bring difficulties for the model 2. Mentioned about high computational complexity 3. Limited amount of data
Amal et at. [61]	Proposed work (augmentation + ensemble model)	HAM10000: 10,015 images	256 × 256	–	–	93.5%	1. Images are not preprocessed 2. Low-resolution image bring difficulties for the model
Abdelhafeez et al. [21]	Deep learning-based single-valued neutrosophic sets	ISIC-2017: 25,331 images	256 × 256	–	NVIDIA GeForce GTX 1060 Ti	85.74%	1. Poor Accuracy 2. Images are not preprocessed 3. Low-resolution image bring difficulties for the model
Chao et al. [34]	Proposed transformer network	HAM10000: 10,015 images	224 × 224 × 3	–	2 GPUs (NVIDIA RTX 3060, 12 GB)	94.3%	1. Low-resolution image bring difficulties for the model 2. Mentioned about high computational complexity

Table 8 (continued)

Author	Model	Dataset	Image resolution	Training time	Used GPU for computation	Accuracy	Limitations
MuHAM10000ad et al. [36]	Proposed vision transformer	Skin Cancer MNIST: 6860 images	224 × 224 × 3	–	–	92.1%	1. Low-resolution image bring difficulties for the model 2. Mentioned about high computational complexity 3. Poor Accuracy
Proposed Research	SkinNet-14	HAM10000: 10,015 images, ISIC: 2357 images, PAD: 2298 images	32 × 32	0.38- 0.44 h [23 – 27min], 0.11 – 0.16 h [6-10min], 0.22 – 0.33 h [13-20min]	16 GB of RAM, an NVidia GeForce GTX 1660 GPU	97.85%, 96.00%, 98.14%	Our research addressed these drawbacks: 1. A huge volume of data 2. A well-balanced dataset 3. Image preparation techniques are employed 4. Reduced processing time 5. Uphold model efficiency in a low-quality and noisy dataset

number of trainable parameters and the least training time per epoch. Therefore, configuration 3 is selected for additional ablation experiments.

8.2.2 Modification 2: activation function changes

Different activation functions influence the classification model, and the optimal activation function improves model performance. Six different activation functions, entitled Tanh, ELU, ReLU, SoftSign and SoftPlus are applied to the model (Table 3). ReLU scored the highest accuracy among the six activation functions, 91.24%, with a time duration of 10 s per epoch. Therefore, ReLU activation is selected for additional ablation experiments.

8.2.3 Modification 3: type of pooling layer changes

The pooling layers downsample feature maps by summarizing feature presence in patches. Average pooling and max-pooling layers are applied for this experiment (Table 3). The test accuracy increased from 91.24% to 92.37% after using the max-pooling layer. As a result, the max-pooling layer is selected for additional ablation experiments.

8.2.4 Modification 4: stride size changes

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

8.2.5 Modification 5: kernel size changes

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 demonstrates that kernel size 3 yields the highest accuracy at 94.77% and the shortest time per epoch of 7 s. Consequently, a kernel size of 3 is maintained for future ablation studies.

8.2.6 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 cross-entropy, binary cross-entropy, mean squared logarithmic error, mean absolute error and mean squared error. Categorical cross-entropy, at 95.80% achieved the highest accuracy of all loss functions tested (Table 4). Categorical

cross-entropy is therefore adjusted for subsequent ablation experiments.

8.2.7 Modification 7: batch size changes

Different batch sizes affect classification model performance. For the modification, 256, 128, 64 and 32-batch sizes are evaluated (Table 4). Training the model with 128 batches results in a maximum accuracy of 96.68% with 10 s per epoch, whereas other batch sizes reduce accuracy (Table 4). Accordingly, further ablation studies use batch size 128.

8.2.8 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 known as Nadam, Adam, Adamax, RMSprop and SGD were tested with a learning rate of 0.0008. The best accuracy of 96.68%, is attained with the Adam optimizer (Table 5). Therefore, Adam optimizer is retained for the remainder of the ablation research.

8.2.9 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 tested. The Adam optimizer achieves a best result of 97.85% with a learning rate of 0.001 (Table 5). Hence, a learning rate of 0.001 is applied for subsequent ablation studies.

8.2.10 Modification 10: image size changes

The final study involves doing experimentation with the input layer picture dimensions (image height and width). We tested 64×64, 32×32, 28×28 and 16×16 pixel sized images. The findings are presented in Table 5. The model was able to be trained in just 10 s per epoch, while still achieving the best testing accuracy of 97.85%, with an image size of 32×32 on HAM10000 dataset. However, the image size of 64×64 also achieved a very good test accuracy of 96.17%, but the training time was 24 s per epoch.

The input image dimension selected is 32×32 pixels as it requires minimal training time while retaining high performance. This is essential because the objective of the study is to design a model with high performance that also takes time complexity into account. Figure 6 depicts how test accuracy gradually improved throughout the ablation studies conducted on the base model.

After the ablation study, the configuration of the proposed SkinNet-14 is: 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 cross-entropy and pooling layer is max-pooling. Pooling layer kernel size is 3 and stride size is 1.

8.3 Performance evaluation of the proposed model

After completing ablation experiments on the base model, the final SkinNet-14 model has been created with significantly enhanced classification performance. This is accomplished by modifying and configuring the model in various ways. Table 6 shows a statistical analysis for the proposed SkinNet-14 model, such as precision, f1-score, recall and test accuracy on each class for the three datasets.

The results of Table 6 clearly show that the proposed model performed exceptionally well on all three datasets. In the HAM10000 dataset, the model achieved good performance metrics on six classes of the dataset, except for vascular lesions. The average accuracy obtained on the HAM10000 dataset is 97.85%. In the ISIC dataset, the model achieved good performance metrics for all eight classes, except for Seborrheic Keratosis. The average accuracy of 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 achieved good performance metrics for precision, recall and f1-score on all six classes of this dataset. The test accuracy of the different classes in each dataset ranges from 98 to 100%. The table depicts that the performance of vascular lesions in the HAM10000 dataset and seborrheic keratosis in the ISIC dataset is poorer compared to other classes. This could be due to the fact that the image quality of these classes was subpar, and as a result, reducing their size resulted in a loss of information, making it more difficult for the model to accurately distinguish these lesions from others. However, it is clear that the proposed model performs well in multiclass classification.

Figure 7 displays the confusion matrix of the SkinNet-14 model on three datasets. Row values accurately indicate the correct labeling of test images. The utilization of column values serves as a means to depict the predicted labels of the model for the images in the test set. The successful prediction of test images by the model is indicated by the diagonal values in the confusion matrix (Fig. 7). However, in the ISIC dataset confusion matrix (Fig. 8B), among the 58 images, 30 misclassifications occurred for Seborrheic Keratosis (Class 7) and in the HAM10000 dataset confusion matrix (Fig. 8A), among the 107 images, 100 misclassifications occurred for Vascular Lesions (Class 7). Despite that, the model is not biased toward any particular

class or classes, nor does it predict any class significantly better than others. The robustness of the model is demonstrated by providing of approximately equal numbers of accurate predictions for each class.

Figure 8 depicts the SkinNet-14 model accuracy and loss curves on HAM10000, ISIC and PAD dataset. From the figures of all three dataset, it is visible that the model training and validation curves converge without substantial gaps, indicating minimal overfitting. Similarly, loss curves (Fig. 8) converge steadily from the start to ending epoch. It can be said that neither overfitting nor underfitting occurred during the model training phase.

8.4 Examining the performance stability of proposed SkinNet-14 model

This section evaluates the proposed SkinNet-14 model performance consistency by gradually reducing the amount of input photos at various phases. The dataset image number is decreased by approximately half for each phase. The results of decreasing the number of images are shown in Fig. 9.

After data augmentation, the total number of images in each dataset is 50,785 for the HAM10000 dataset, 11,195 for the ISIC dataset and 11,490 for the PAD dataset. Figure 9 illustrates the performance of the proposed SkinNet-14 model using a reduced number of images. In Fig. 9A for the HAM10000 dataset, we can see that the model attained a 96.68% accuracy with 50,785 images. After that, the images are decreased by half in each phase, as follows: 25,392, 12,696 and 6,384, with a respective accuracy of 95.57%, 93.83% and 90.27%. In Fig. 9B, ISIC dataset with image number 11,195, the suggested model attained an accuracy of 97.85%. After lowering the image to 5,597, the accuracy is 95.21%, and after decreasing it by half to 2,798, the accuracy is 92.61%. In Fig. 9C for the PAD dataset, the accuracy is 98.14% with 11,490 images, 96.23% with a 50% reduction to 5,745 images and 94.45% with 2,872 images. After analyzing the proposed model on each of these three datasets, despite reducing the number of images, the model maintains performance consistency and accuracy. We conclude that utilizing a modest number of images, the suggested SkinNet-14 model may produce optimal results while keeping a low training time, demonstrating the consistency of the model performance. In addition, the model can use fewer images without a significant reduction in test accuracy.

8.5 Comparison with CNN-based transfer learning models

Six state-of-the-art transfer learning models are used to evaluate the proposed approach. All models are trained and tested on the three skin cancer datasets with 32×32 size images. The images are preprocessed and augmented. Table 7 shows the results. The optimizer is Adam, the batch size is 128, and the learning rate for each model in the table is 0.001. Table 7 shows the experiment results.

VGG16 achieved the highest test accuracy of 81.21% and F1-score of 82.10% on the HAM10000 dataset and 71.21% accuracy with 72.09% F1-score, on the ISIC dataset, outperforming all other transfer learning models. On the PAD dataset, VGG19 achieved the highest score of the six CNN-based pretrained models with 82.97% accuracy and 83.48% F1-score. On all three datasets containing 32×32 pixel pictures, the accuracy of the remaining transfer learning models varied between 40 and 80% for both accuracy and F1-score. The parameters of all transfer learning models were high, which raised the temporal complexity and time per epoch, which ranged between 65 and 67 s for the HAM10000 dataset, 30 and 34 s for the ISIC dataset, and 28 and 30 s for the PAD dataset. In contrast, our suggested model achieves the highest accuracy of 97.85% on the HAM10000 dataset, 96.0% on the ISIC dataset and 98.14% on the PAD dataset with 97.92%, 96.50% and 98.57% F1-score, respectively. In terms of accuracy and F1-score, the SkinNet-14 model outperformed all six transfer learning methods. In addition, the suggested model parameter size is 241,861, resulting in a reduced temporal complexity of 7 to 8 s per epoch on the HAM10000 dataset, 2 to 3 s on the PAD dataset and 1 to 2 s on the ISIC dataset. With our methodology, training takes about 6–24 min contrasted with approximately two hours for the transfer learning methods. This represents a substantial improvement in terms of time-intensiveness. Additionally, achieving near-optimal performance with smaller image size uses less memory and storage space, making the model less resource-hungry and contributing to a reduction in space complexity.

8.6 Comparison with other studies

Table 8 represents a comparison table of model, dataset, image resolution, training time, accuracy and limitations between the previous studies and proposed research.

Table 8 compares the literature of other models to our proposed model. The table shows that all studies classified skin cancer using different models. A pixel size of 224×224 is the commonly used image resolution for the models. However, each study shows a number of limitations, such as high computational time, limited amounts of data, poor accuracy and images without preprocessing.

Our research addressed these limitations using a large and well-balanced dataset and applying image processing techniques. In order to construct generic classification models for medical imaging, image resolution is a key

consideration [62]. In order to learn about the effect of the image resolution in the modeling, Sabottke et al. [63] experimented with image resolutions ranging from 32×32 pixel to 600×600 pixel using some well-known deep learning and transfer learning models. The study demonstrated that when the pixel size of the image decreases, the information required for CNNs for classification decreases and as a result, the model suffers in terms of accuracy. This limitation is overcome in our work by recommending a model that utilizes images with low resolution (32×32 pixels) and achieves good accuracies.

9 Conclusion

To summarize the theoretical advancements achieved with SkinNet-14, a model that demonstrates remarkable ability to analyze low-resolution dermoscopy images for the identification of skin cancer. A significant development in the field is demonstrated by this work, which is supported by the model's capacity to overcome the typical challenges of high-resolution demand and high computational cost. The deployment of a modified CCT architecture by the model, which strategically enhances data to address and reduce class imbalances holistically and hence results in better performance, makes significant theoretical advances.

Key results validate the accuracy of SkinNet-14, which reaches up to 97.85% on the HAM10000 dataset, 96.01% on the ISIC dataset and 98.14% on the PAD dataset with lower training time. These results demonstrate the model's dependability and efficiency in a variety of diagnostic contexts by exceeding current benchmarks and confirming its stability despite reduced data volume. Realizing the shortcomings of our method, future work should focus on extracting different features from a larger set of raw photos, including real-time data collecting for a more complete picture of skin cancer types. Moreover, investigating how SkinNet-14 might be included into clinical processes is a promising direction that might bring in a new era of easily available, quick and precise diagnostic methods. Our findings contribute a novel perspective to the existing knowledge pool, challenging and inspiring the academic community to embrace and expand upon our methodologies..

Data availability The HAM10000, ISIC and PAD [37] datasets are publicly available.

Declarations

References

1. Xu YG, Aylward JL, Swanson AM, Spiegelman VS, Vanness ER, Teng JMC et al (2020) Nonmelanoma skin cancers: basal cell and squamous cell carcinomas. Abeloff's Clin Oncol. https://doi.org/10.1016/B978-0-323-47674-4.00067-0
2. Siegel RL, Miller KD, Fuchs HE, Jemal A (2021) Cancer statistics, 2021. CA Cancer J Clin 71:7–33. https://doi.org/10.3322/CAAC.21654
3. Sung H, Ferlay J, Siegel RL, Laversanne M, Soerjomataram I, Jemal A et al (2021) Global cancer statistics 2020: globocan estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 71:209–249. https://doi.org/10.3322/caac.21660
4. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM et al (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542:115–118. https://doi.org/10.1038/nature21056
5. Bray F, Ferlay J, Soerjomataram I, Siegel RL, Torre LA, Jemal A (2018) Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. CA Cancer J Clin 68:394–424. https://doi.org/10.3322/CAAC.21492
6. Mondal MRH, Bharati S, Podder P, Podder P (2020) Data analytics for novel coronavirus disease. Inf Med Unlocked 20:100374. https://doi.org/10.1016/J.IMU.2020.100374
7. Mohan A, Singh AK, Kumar B, Dwivedi R (2021) Review on remote sensing methods for landslide detection using machine and deep learning. Trans Emerg Telecommun Technol 32:e3998. https://doi.org/10.1002/ETT.3998
8. He X, Wang Y, Zhao S, Yao C (2022) Deep metric attention learning for skin lesion classification in dermoscopy images. Complex and Intell Syst 8:1487–1504. https://doi.org/10.1007/s40747-021-00587-4
9. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. (2022) An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. CoRR, https://doi.org/10.48550/arxiv.2010.11929
10. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is All you Need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), Advances in neural information processing systems (Vol. 30). Retrieved from https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

11. Chen J, Lu Y, Yu Q, Luo X, Adeli E, Wang Y, et al. TransUNet: Transformers make strong encoders for medical image segmentation. CoRR. 2021 [cited 25 Dec 2022]. https://doi.org/10.48550/arxiv.2102.04306

12. Hassani, A., Walton, S., Shah, N., Abuduweili, A., Li, J., & Shi, H. (2021). Escaping the Big Data Paradigm with Compact Transformers. CoRR, abs/2104.05704. Retrieved from https://arxiv.org/abs/2104.05704.

13. Montaha S, Azam S, RakibulHaqueRafid AKM, Islam S, Ghosh P, Jonkman M (2022) A shallow deep learning approach to classify skin cancer using down-scaling method to minimize time and space complexity. PLoS ONE. https://doi.org/10.1371/journal.pone.0269826

14. Kassem MA, Hosny KM, Fouad MM (2020) Skin lesions classification into eight classes for ISIC 2019 using deep convolutional neural network and transfer learning. IEEE Access 8:114822–114832. https://doi.org/10.1109/ACCESS.2020.3003890

15. Hagerty JR, Stanley RJ, Almubarak HA, Lama N, Kasmi R, Guo P et al (2019) Deep learning and handcrafted method fusion: higher diagnostic accuracy for melanoma dermoscopy images. IEEE J Biomed Health Inform 23:1385–1391. https://doi.org/10.1109/JBHI.2019.2891049

16. Abdar M, Samami M, DehghaniMahmoodabad S, Doan T, Mazoure B, Hashemifesharaki R et al (2021) Uncertainty quantification in skin cancer classification using three-way decision-based Bayesian deep learning. Comput Biol Med. https://doi.org/10.1016/j.compbiomed.2021.104418

17. Thomas SM, Lefevre JG, Baxter G, Hamilton NA (2021) Interpretable deep learning systems for multi-class segmentation and classification of non-melanoma skin cancer. Med Image Anal. https://doi.org/10.1016/j.media.2020.101915

18. Ameri A (2020) A deep learning approach to skin cancer detection in dermoscopy images. J Biomed Phys Eng 10:801–806. https://doi.org/10.31661/jbpe.v0i0.2004-1107

19. Tembhurne JV, Hebbar N, Patil HY, Diwan T (2023) Skin cancer detection using ensemble of machine learning and deep learning techniques. Multimed Tools Appl 82:27501–27524. https://doi.org/10.1007/s11042-023-14697-3

20. Mehr RA, Ameri A (2022) Skin cancer detection based on deep learning. J Biomed Phys Eng 12:559–568. https://doi.org/10.31661/jbpe.v0i0.2207-1517

21. Abdelhafeez A, Mohamed HK, Maher A, Khalil NA (2023) A novel approach toward skin cancer classification through fused deep features and neutrosophic environment. Front Public Health. https://doi.org/10.3389/fpubh.2023.1123581

22. Khater T, Ansari S, Mahmoud S, Hussain A, Tawfik H (2023) Skin cancer classification using explainable artificial intelligence on pre-extracted image features. Intell Syst Appl. https://doi.org/10.1016/j.iswa.2023.200275

23. Surono S, YahyaFirzaAfitian M, Setyawan A, Arofah DKE, Thobirin A (2023) Comparison of CNN classification model using machine learning with bayesian optimizer. HighTech Innov J 4:531–542. https://doi.org/10.28991/HIJ-2023-04-03-05

24. Surono S, Rivaldi M, Dewi DA, Irsalinda N (2023) New approach to image segmentation: U-Net convolutional network for multiresolution CT image lung segmentation. Emerg Sci J 7:498–506. https://doi.org/10.28991/ESJ-2023-07-02-014

25. Sornsuwit P, Jundahuadong P, Pongsakornrungsilp S (2022) A new efficiency improvement of ensemble learning for heart failure classification by least error boosting. Emerg Sci J 7:135–146. https://doi.org/10.28991/ESJ-2023-07-01-010

26. Bhatt H, Shah V, Shah K, Shah R, Shah M (2023) State-of-the-art machine learning techniques for melanoma skin cancer detection and classification: a comprehensive review. Intell Med 3:180–190. https://doi.org/10.1016/J.IMED.2022.08.004/ASSET/

66A3886B-4387-4695-928D-CC1E2D818DF1/ASSETS/GRAPHIC/2096-9376-03-03-004-F003.PNG

27. Mirikharaji Z, Abhishek K, Bissoto A, Barata C, Avila S, Valle E et al (2023) A survey on deep learning for skin lesion segmentation. Med Image Anal 88:102863. https://doi.org/10.1016/J.MEDIA.2023.102863

28. Zafar M, Sharif MI, Sharif MI, Kadry S, Bukhari SAC, Rauf HT (2023) Skin lesion analysis and cancer detection based on machine/deep learning techniques: a comprehensive survey. Life 13:146. https://doi.org/10.3390/LIFE13010146

29. Mazhar T, Haq I, Ditta A, Mohsan SAH, Rehman F, Zafar I et al (2023) The role of machine learning and deep learning approaches for the detection of skin cancer. Healthcare 11:415. https://doi.org/10.3390/HEALTHCARE11030415

30. Xin C, Liu Z, Zhao K, Miao L, Ma Y, Zhu X et al (2022) An improved transformer network for skin cancer classification. Comput Biol Med. https://doi.org/10.1016/j.compbiomed.2022.105939

31. Chen J, Chen J, Zhou Z, Li B, Yuille A, Lu Y. MT-TransUNet: Mediating Multi-Task Tokens in Transformers for Skin Lesion Segmentation and Classification. 2021. Available: http://arxiv.org/abs/2112.01767, https://doi.org/10.48550/arXiv.2112.01767

32. Zhang J, Xie Y, Xia Y, Shen C (2019) Attention residual learning for skin lesion classification. IEEE Trans Med Imaging 38:2092–2103. https://doi.org/10.1109/TMI.2019.2893944

33. Gessert N, Sentker T, Madesta F, Schmitz R, Kniep H, Baltruschat I et al (2020) Skin lesion classification using CNNs with patch-based attention and diagnosis-guided loss weighting. IEEE Trans Biomed Eng 67:495–503. https://doi.org/10.1109/TBME.2019.2915839

34. Datta SK, Shaikh MA, Srihari SN, Gao M. Soft-attention improves skin cancer classification performance. 2021. Available: http://arxiv.org/abs/2105.03358. https://doi.org/10.48550/arXiv.2105.03358

35. Aladhadh S, Alsanea M, Aloraini M, Khan T, Habib S, Islam M (2022) An effective skin cancer classification mechanism via medical vision transformer. Sensors. https://doi.org/10.3390/s22114008

36. Yang G, Luo S, Greer P (2023) A novel vision transformer model for skin cancer classification. Neural Process Lett. https://doi.org/10.1007/s11063-023-11204-5

37. Arshed MA, Mumtaz S, Ibrahim M, Ahmed S, Tahir M, Shafi M (2023) Multi-class skin cancer classification using vision transformer networks and convolutional neural network-based pretrained models. Information (Switzerland). https://doi.org/10.3390/info14070415

38. Tschandl P, Rosendahl C, Kittler H (2018) Data descriptor: the HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci Data. https://doi.org/10.1038/SDATA.2018.161

39. Skin Cancer ISIC | Kaggle. [cited 25 Dec 2022]. Available: https://www.kaggle.com/datasets/nodoubttome/skin-cancer9-classesisic

40. Pacheco AGC, Lima GR, Salomão AS, Krohling B, Biral IP, de Angelo GG et al (2020) PAD-UFES-20: a skin lesion dataset composed of patient data and clinical images collected from smartphones. Data Brief 32:106221. https://doi.org/10.1016/J.DIB.2020.106221

41. Nawaz M, Mehmood Z, Nazir T, Naqvi RA, Rehman A, Iqbal M et al (2022) Skin cancer detection from dermoscopic images using deep learning and fuzzy k-means clustering. Microsc Res Tech 85:339–351. https://doi.org/10.1002/JEMT.23908

42. Medical Diseases & Conditions – Mayo Clinic. [cited 6 Nov 2023]. Available: https://www.mayoclinic.org/diseases-conditions

43. Fuchs A, Marmur E (2007) The kinetics of skin cancer: progression of actinic keratosis to squamous cell carcinoma. Dermatol Surg 33:1099–1101. https://doi.org/10.1111/J.1524-4725.2007.33224.X

44. Goyal N, Thatai P, Sapra B (2017) Skin cancer: symptoms, mechanistic pathways and treatment rationale for therapeutic delivery. Ther Deliv 8:265–287. https://doi.org/10.4155/TDE-2016-0093

45. Salido JA, Ruiz C (2018) Hair artifact removal and skin lesion segmentation of dermoscopy images. Asian J Pharm Clin Res 11:36–39. https://doi.org/10.22159/ajpcr.2018.v11s3.30025

46. Reddy BD, Bhattacharyya D, Rao NT, Kim T. (2022) Medical Image Denoising Using Non-Local Means Filtering. 123–127. https://doi.org/10.1007/978-981-16-8364-0_15

47. Tripathy S, Swarnkar T (2020) Unified preprocessing and enhancement technique for mammogram images. Procedia Comput Sci 167:285–292. https://doi.org/10.1016/J.PROCS.2020.03.223

48. Zhang Y, Zhu Y, Nichols E, Wang Q, Zhang S, Smith C, et al. (2019) A Poisson-Gaussian Denoising Dataset With Real Fluorescence Microscopy Images. 11710–11718.

49. Shorten C, Khoshgoftaar TM (2019) A survey on image data augmentation for deep learning. J Big Data 6:1–48. https://doi.org/10.1186/S40537-019-0197-0/FIGURES/33

50. Taylor L, Nitschke G. Improving Deep Learning with Generic Data Augmentation. Proceedings of the 2018 IEEE Symposium Series on Computational Intelligence, SSCI 2018. 2019; 1542–1547. https://doi.org/10.1109/SSCI.2018.8628742

51. Castro E, Cardoso JS, Pereira JC. Elastic deformations for data augmentation in breast cancer mass detection. 2018 IEEE EMBS International Conference on Biomedical and Health Informatics, BHI 2018. 2018;2018-January: 230–234. https://doi.org/10.1109/BHI.2018.8333411

52. Huang X, Bi N, Tan J. Visual Transformer-Based Models: A Survey. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2022;13364 LNCS: 295–305. https://doi.org/10.1007/978-3-031-09282-4_25/COVER

53. Khan IU, Azam S, Montaha S, Mahmud Al A, Rafid A, Hasan M et al (2022) An effective approach to address processing time and computational complexity employing modified CCT for lung disease classification. Intell Syst Appl 16:200147. https://doi.org/10.1016/J.ISWA.2022.200147

54. Lorencin I, Šegota SB, Anđelić N, Mrzljak V, Ćabov T, Španjol J, Car Z (2021) On urinary bladder cancer diagnosis: utilization of deep convolutional generative adversarial networks for data augmentation. Biology 10(3):175. https://doi.org/10.3390/biology10030175

55. Simonyan K, Zisserman A. (2014) Very Deep Convolutional Networks for Large-Scale Image Recognition. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. [cited 26 Dec 2022]. https://doi.org/10.48550/arxiv.1409.1556

56. He K, Zhang X, Ren S, Sun J. (2015) Deep Residual Learning for Image Recognition. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2016-December: 770–778. https://doi.org/10.48550/arxiv.1512.03385

57. M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L. -C. Chen, (2018) "MobileNetV2: inverted residuals and linear bottlenecks," 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 4510–4520, https://doi.org/10.1109/CVPR.2018.00474.

58. Al Mahmud A, Karim A, Ullah Khan I, Ghosh P, Azam S, Haque E. (2022) A robust deep learning based framework for high-precision detection of liver disease. The 10th International Conference on Computer and Communications Management. 9–18. https://doi.org/10.1145/3556223.3556225

59. Khan MA, Sharif M, Akram T, Damaševičius R, Maskeliūnas R (2021) Skin lesion segmentation and multiclass classification using deep learning features and improved moth flame optimization. Diagnostics. https://doi.org/10.3390/diagnostics11050811

60. Hameed N, Shabut AM, Ghosh MK, Hossain MA (2020) Multi-class multi-level classification algorithm for skin lesions classification using machine learning techniques. Expert Syst Appl. https://doi.org/10.1016/j.eswa.2019.112961

61. Srinivasu PN, Sivasai JG, Ijaz MF, Bhoi AK, Kim W, Kang JJ (2021) Classification of skin disease using deep learning neural networks with mobilenet v2 and lstm. Sensors. https://doi.org/10.3390/s21082852

62. Al-Rasheed A, Ksibi A, Ayadi M, Alzahrani AIA, Mamun EM (2023) An ensemble of transfer learning models for the prediction of skin lesions with conditional generative adversarial networks. Contrast Media Mol Imaging 2023:1–15. https://doi.org/10.1155/2023/5869513

63. Lakhani P (2020) The importance of image resolution in building deep learning models for medical imaging. Radiol Artif Intell. https://doi.org/10.1148/RYAI.2019190177

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.