IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

HDFNet: A Hybrid Deep Fusion Network for Automated Skin Lesion Cancer Detection in Dermoscopic Images

SYED NEHAL HASSAN SHAH[1], IMRAN TAJ[2], SYED MUHAMMAD USMAN.[3],SYED ABDULLAH SHAH[1],ALI SHARIQ IMRAN[4],SHEHZAD KHALID[5]

[1]Department of Creative Technologies, Faculty of Computing and Artificial Intelligence, Air University, Islamabad, 44000, Pakistan email: 222378@students.au.edu.pk, 222376@students.au.edu.pk

[2]College of Interdisciplinary Studies, Zayed University, Abu Dhabi P.O. Box 144534, United Arab Emirates email: MuhammadImran.Taj@zu.ac.ae

[3]Department of Computer Science, Bahria School of Engineering and Applied Sciences, Bahria University, Islamabad, 44000, Pakistan (email: drsyedmusman@gmail.com

[4]Department of Computer Science, Norwegian University of Science and Technology, 2815, Gjøvik, Norway email: ali.imran@ntnu.no

[5]Department of Computer Engineering, Bahria School of Engineering and Applied Sciences, Bahria University, Islamabad, 44000, Pakistan (email: shehzad@bahria.edu.pk

Corresponding author: Ali Shariq Imran (e-mail: ali.imran@ntnu.no).

ABSTRACT Skin Cancer, identified by abnormal skin cell growth, has constituted a significant challenge for global healthcare systems. Basal Cell Carcinoma (BCC) and Melanoma (MEL) are the more common forms of skin cancer. Among these, BCC is the more common variant though it grows slowly and rarely metastasizes whereas MEL is highly aggressive and invades other parts of the body, if not detected and treated in due time. It is prevalently found in regions with higher sun exposure. Advancements in artificial intelligence (AI) have opened new avenues for improving skin cancer classification. Current AI models often face challenges related to interpretability, generalizability across diverse skin types, and the integration of clinical context. We propose a novel multi-layered model, HDFNet, a Hybrid Dense Fusion Network Model, for accurate classification of dermal lesions. We applied preprocessing techniques, including normalization, high-frequency balancing, and data augmentation, to mitigate class imbalance issues on the acquired after selective merging datasets ISIC 19, ISIC 20, DERMQUEST-DERMIS, to insure improved generalization is attained collection of diverse dataset. After preprocessing, features were extracted using custom architectures: Adaptive Layer Configuration VGG-16 (ALC-VGG16) and Dermatological Feature Enhancement ResNet-50 (DFE-ResNet50). Additionally, our custom Vision Transformer (ViT) model incorporates dynamic patching, which adapts patch sizes based on lesion dimensions to optimize feature extraction. This feature set is filtered through the Multilayer Perceptron (MLP) where categorization of labels in to the affirmed domains is performed. 92% accuracy is demonstrated by our recommended model along with AUC score of 92.15%, showing significant improvements from the existing methods.

INDEX TERMS Skin Cancer, Basal Cell Carcinoma, Melanoma, Artificial Intelligence, Deep Learning, Dermoscopic Images, Attention Mechanisms, Model Evaluation.

I. INTRODUCTION

SKIN cancer is one of those evasive diseases with a variety of identifiable patterns which, due to the variations in the skin type itself, lead to many cases of false positives or false negatives. There are a number of other factors that may contribute to this evasive behavior such as image noise, light intensity, angle, etc. With the advancements in AI, automated diagnostic of dermatology is possible [1]. Its Melanoma variant constitutes a serious concern due to its potential for rapid progression to fatal levels. This type of cancer is a result of excessive Ultraviolet (UV) radiation exposure which leads to melanocytes developing cellular mutations and transforming into malignant tumors. Using AI-enabled automated tools as second opinions is quite useful for detection of skin cancer at earlier stages.

Prior to the rapid enhancements in AI algorithms, diagnosis of skin cancer typically relied on the observation skills and expertise of dermatologists. Accuracy of this visual analysis could be compromised due to any number of factors like brightness, angle of observation, and experience of the observer, causing potential oversights that could lead to problematic diagnoses [2] [3]. A variety of deep learning architectures, including CNN, DNN, RNN, LSTM, ResNet, Inception, Xception, and VGG16, have been used for lesion detection and identification

[3]–[5]. InceptionV3 and VGG16 are known to provide good results of lesion classification when combining human expertise [3]. Dermoscopy is an important tool for acquiring images for automated diagnostics. A dermoscopic analysis is a preliminary step in the assessment of any suspicions regarding the development of cancerous regions [6]. It has been noticed that the detection rate for melanoma remains between 75-84% [7] despite the visual inspection being aided by dermoscopy which provides a magnified and illuminated view of skin so that any irregularities can be easily discovered. In contrast, DL models significantly increased the detection rate through a more accurate assessment of skin condition [8]. The recent developments of new models that can extract low-level features without falling into any bias have been useful in the domain of medical categorization of images due to difficulties in classifying the visual aspects of lesions [9]. Challenges of the visual inspection include skin type variations, presentation method of the affected region, and the inherent subjectivity that ensures a thorough dermatological analysis [10]. ResNet, VGG, Inception, and Xception have shown exceptional adaptation capabilities for dermatological images through the Few-Shot Learning techniques [3], [11]–[13]. The objective of this research is the exploration and attainment of a custom deep learning model that provides generalizability in the automated detection of skin cancer accurately.

KEY CONTRIBUTIONS:

Following are the contribution that are desired to be atttained from this study:

- Evaluated methods to improve model generalization beyond training datasets, ensuring reliable performance on unseen data.
- Tested developed models on diverse datasets to assess performance in practical scenarios.
- Explored ensemble and transformer learning techniques to enhance model accuracy in classifying skin cancer.

The following sections are organized in a way that each section contributes to developments that are discussed in the next. II provides an evaluation of the existing literature in organized structure by cover the applied methodology in sections and define the observed research gaps in the applied studies. III covers the data sets utilized in this research and the selected labels that have been applied in our study from overall merger of the datasets. IV gives an overall layout of our research methodology in detail. V gives the detailed outcomes of the applied methodology as to evaluate the attained result with the desired goals. VI

II. LITERATURE REVIEW

A typical method of skin cancer detection using dermoscopic images involves preprocessing, feature extraction, and classification. Figure 1 presents the preprocessing techniques employed, Figure 2 the model applied and techniques for feature extraction, and Figure 3 presents the applied classifier. Mukadam et al. [9] employed image resizing, color preservation, and sharpening filters along with augmentation techniques like rotation, zooming, height and width shift, and rescaling while applying ESRGAN for enhancement. Rasheed et al. [12] applied resizing along with rotation, translation and flipping augmentation techniques. Resizing, black hat filtering, masking filtering, noise removal and augmentation methods of rotation. Thanka et al. [14] applied flipping and blurring [2]. Jaisakthi et al [11] used image standardization and resizing while Yang et al [15] used resizing, normalization, duplication removal, patch extraction, position embedding and augmented through techniques like brightness/contrast alteration, flipping zooming, rotation and shift. Zhao [16] used median frequency balancing, resizing normalization, and augmentation, while Kahia [17] used class balancing, resizing, and augmentation methods of flipping, shifting, rotating, transformation, and zooming. Ali [18] used resolution scaling, black hat transform, masking and fast marching method, and augmentation techniques like rotation, zooming and horizontal and vertical flipping. Gouda et al. [5] utilized oversampling, resizing, custom contrast method, and augmentation techniques like rotation, reflection, shifting, and brightness adjustment, along with ESRGAN for enhancement and noise reduction. Bassel [3] used resizing while Nakai [10] used lesion segmentation, normalization, resizing, and data augmentation. Figure 1 presents the preprocessing methods proposed by researchers in recent years for skin cancer detection using dermoscopic images.

Akter et al [1] used resizing, normalization and augmentation while Medhat [19] used resizing and augmentation. Li [8] employed augmentation, oversampling, and weighted random sampling, whereas Banasode [20] applied image blurring, color thresholding, image masking, segmenting, and transformation along with RGB to HSV and RGB to Grayscale conversions for their data. Datta et al. [4] used oversampling, undersampling, and normalization, while Jain [6] utilized high-frequency sampling, normalization, and augmentation methods of rotation, shifting, and zooming. Rezaoana [13] employed normalization, resizing, and augmentation through rotation, horizontal flip, zoom and shear. Finally Kassem [7] employed minority class sampling, augmentation, and bootstrapped multiclass SVM aggregation.

In feature extraction, Al-Rasheed [12] utilized transfer learning models ResNet50, ResNet101 and VGG16 pretrained on ImageNet for feature extraction while Thanka [14] utilized the transfer learning VGG16 model. Jaisakthi [11] used ResNet50, DenseNet121, Inception ResNetV2, and EfficientNet variants 0-7 for feature extraction purposes, whereas Yang [15] relied upon patch-based feature representation. Bassel [3] employed ResNet50, VGG16 and Xception (Main)

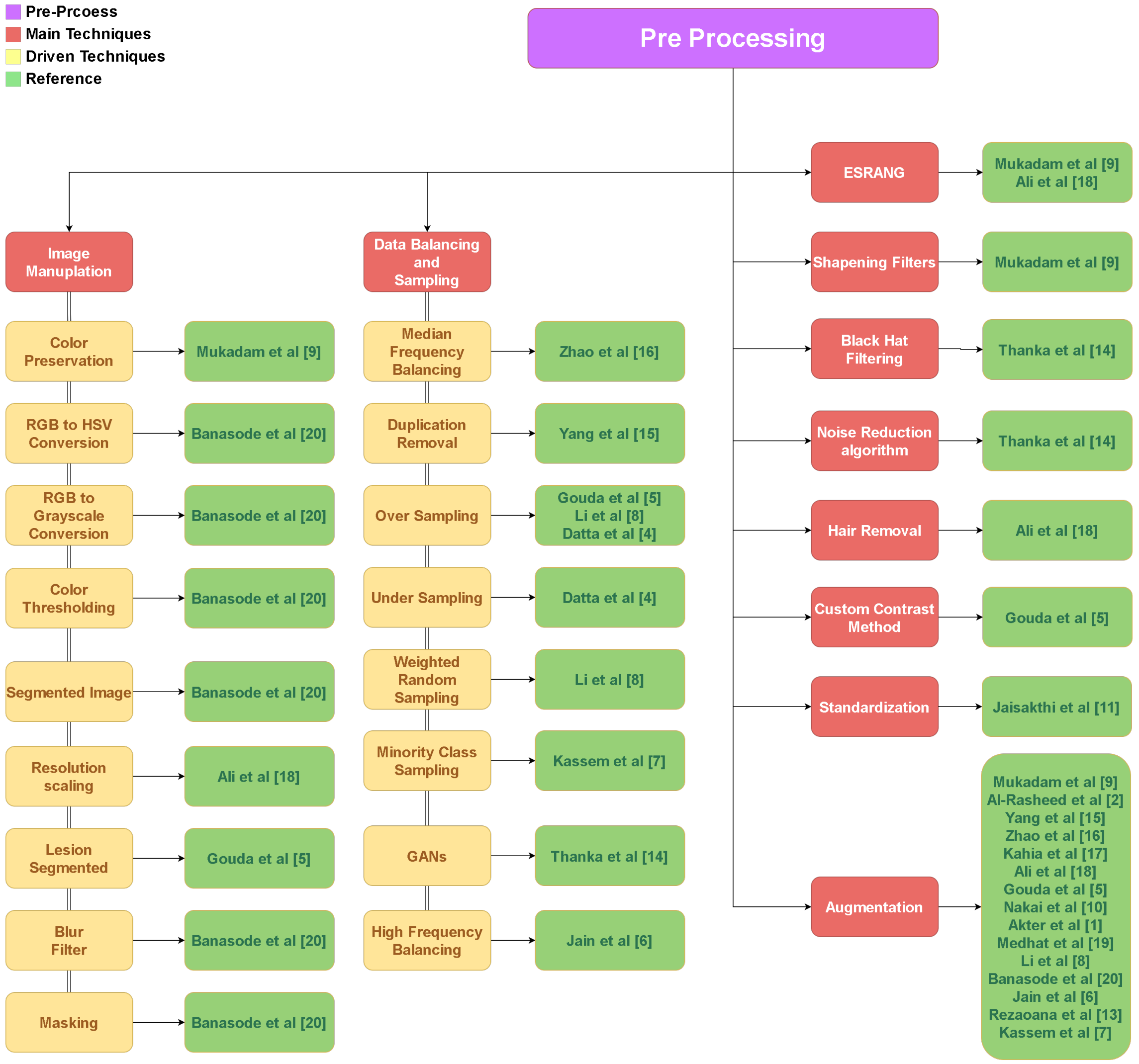

FIGURE 1: Preprocessing techniques applied in the existing methods

for this stage while Akter [1] used ResNet50, VGG16, Inception, Xception and MobileNet. Banasode [20] used a KNN mining/clustering algorithm, whereas Kassem [7] employed GoogleNet for feature extraction purposes. These are all the researches that have utilized feature extractors and so we will move to discuss the variety in the selection of classifiers that were employed by the previous research.

Mukadam et al [9] proposed a custom CNN for seven-category classification, whereas, Al-Rasheed [12] employed VGG16, ResNet50 and ResNet 101 pre-trained on ImageNet along with CGANs and the Ensemble algorithm for the purpose of generating more data and averaging outcomes respectively. Thanka [14] employed a hybrid strategy involving data augmentation, and Conditional Generative Adversarial Networks (CGANs) with VGG16 and XGBoost while Kaya [2] compared the performance of VGGNet11, VGGNet13, VGGNet16, and VGGNet19 and discovered VGG11 to be the best performer. Jaisakthi [11] used the EfficientNetB6 pre-trained on ImageNet for a binary classification (benign or malignant lesions) while Yang [15] used a custom Vision Transformer for a seven-category classification. Zhao [16] employed a number of CNNs, including ResNet18, VGGNet, ViT, and Deep ViT; they found ResNet18 to be the best-performing model among these. Kahia [17] employed VGG16 and InceptionV3 with transfer learning for two-category and three-category classification purposes whereas Ali [18] applied EfficientNet B0-B6 family of models for a comparative seven-category classification and found EfficientNetB4 to be the best performing model. Gouda [5] selected ResNet50, InceptionV3 and InceptionResNet for a binary classification experiment and concluded with InceptionV3 being the best performer while Bassel [3] used DNN, SVM, Random Forest (RF), Regression, Neural Netowrk (NN), k-nearest neighbours (KNN), AdaBoost, Deci-

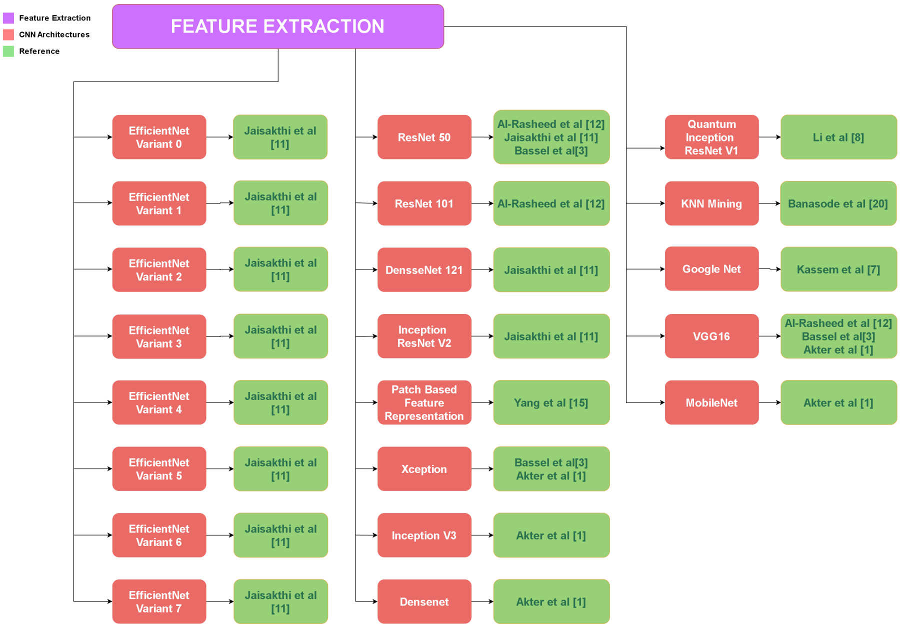

FIGURE 2: Feature Extraction techniques applied in the existing methods

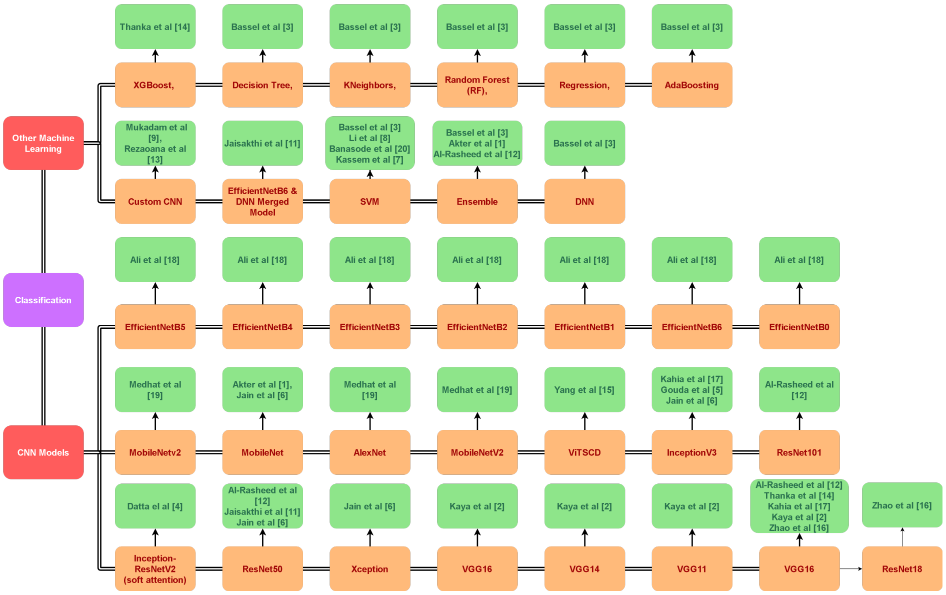

FIGURE 3: Classification models proposed in the existing methods

sion Tree, GaussianN combined with Xception feature extraction using stacking CV algorithm for a binary classification task.

Nakai [10] used an optimized custom bottleneck transformer (EnDBoT) and variants of ResNet and DenseNet architectures, for the task of multi-class categorization. Akter [1] used five distinct stacking models for a seven-category classification whereas Medhat [19] employed MobileNetV2 and AlexNet both with transfer learning and augmentation for a multi-class categorization experiment. Li [8] used quantum computing with InceptionResNetV1 for melanoma classification through an SVM classifier, while Banasode [20] used the SVM for a binary classification. Datta [4] employed soft-attention layers in ResNet34, ResNet50, VGG16, VGG19, InceptionResNetV2 and DenseNet201 [21] for multi-category classification of melanoma while Jain [6] employed a collection of transfer learning models as illustrated in Figure 3 for melanoma classification in multiple categories. Rezaoana [13] utilized a parallel CNN model for a nine-category classification, whereas Kassem [7] used transfer learning GoogleNet and replaced Softmax and classification output layers with SVM for lesion classification in multiple categories. Several gaps have been identified from the existing literature which are listed as follows:

- Many studies rely on biased datasets, lacking comprehensive strategies to address these imbalances and resulting biases.
- Models often demonstrate strong performance on familiar datasets but frequently struggle with new data, indicating a tendency towards overfitting.
- Most models are tested using homogeneous datasets, bypassing the need for evaluation on diverse, real-life skin cancer examples.
- The potential of transformers and ensemble models in skin cancer detection has not been deeply investigated, leaving gaps in understanding their full benefits.

TABLE 1: Comparison Table of Studies in the Literature Review

Reference and Publication Year	**Pre-processing**	**Feature Extraction**	**Methodology**	**Training/Testing**
Mukadam et al. [9] 2023	ESRGAN Augmentation Image Resizing Color Preservation Sharpening Filters		ESRGAN Custom Convolutional Neural Network (CNN) Protocol-III **Binary classification** BKL and MEL	HAM10000: 10,015 images Augmented data: 38,017 training images, 4,224 validation images, 4,694 test images Training=80%, Testing=20%
Al-Rasheed et al. [12] 2023	Re-sizing Augmentation	Transfer Learning Models: VGG16, ResNet50, ResNet101	CGANs generate data Ensemble Algorithm (VGG16, ResNet50, ResNet101) **Multi-class classification** AKIEC, BCC, BKL, DF, MEL, NV, VASC	HAM10000, ISIC 2018: 10,015 images Seven distinct types of skin lesions Training=70%, Testing=10%, Validation=20%
Thanka et al. [14] 2023	Resizing Black hat filtering Noise removal (hair detection and inpainting algorithm) Masking filtering Augmentation GANs	Transfer Learning Model: VGG16	Hybrid method VGG16 and XGBoost Augmentation + GAN **Binary classification** BKL and MEL	ISIC dataset: 1,000 skin lesions 416 images of Malignant Melanoma Training=80%, Testing=20%, Validation=20%
Kaya et al. [2] 2023	Not specified	Not elaborated	VGG-Net (11, 13, 16, 19) **Binary classification** BKL and MEL	Kaggle competition 2021: Total images used=3,297 Training=80%, Testing=20%
Jaisakthi et al. [11] 2023	Image Standardization Image Resizing	Image feature extractor models: DenseNet121, ResNet50, InceptionResNetV2, EfficientNet (variants 0-7)	DCNN Dense Net model ResNet50 Inception ResNet V2 Efficient Net B6 **Binary classification** BKL and MEL	Two datasets: (ISIC 2019 and ISIC 2020) skin cancer classification: ISIC 2020: 33,126 dermoscopic images ISIC 2019: 25,331 images (includes BCN_20000, HAM10000, MSK datasets) Training=33,126, Testing=25,331
Yang et al. [15] 2022	Image Resizing Normalization Augmentation Duplication Removal Patch Extraction Position Embedding	Patch based Feature Representation	Custom ViT for skin cancer detection (ViTfSCD): Variants: VitfSCD-L, VitfSCD-B **Multi-class classification** NV, MEL, BKL, BCC, AKIEC, VASC, DF	ImageNet dataset, HAM10000 (ViTfSCD trained on ImageNet, retrained on HAM10000) Data augmentation for cancer-class rebalancing: 85% training, 15% testing
Zhao et al. [16] 2022	Median frequency balancing Re-sizing Normalization Data augmentation		CNN, VGGNet, ResNet, ViT, Deep ViT **Multi-class classification** MEL, NV, AKIEC, BCC, DF, VASC, BKL	HAM10000: 10,000 + Median frequency balancing, data augmentation Training 80%, Testing 20%
Kahia et al. [17] 2022	Class balancing Augmentation Resizing Images		Three categories: VGG16, Inception v3 With two categories: VGG16, Inception v3 **Multi and Binary classification** MEL, NV, AKIEC	Dataset: - Melanoma images: 374 - Nevus images: 1,372 - Seborrheic keratosis images: 254 (Link available custom dataset) Training=2000, Validation=150, Testing=600
Ali et al. [18] 2022	Resolution scaling Image Augmentation Hair Removal (fast marching method blackhat transform & masking)	Not specified	EfficientNet B0-B6 (custom addition) EfficientNet B4 **Multi-class classification** AKIEC, BCC, BKL, DF, MEL, NV, VASC	HAM10000 dataset: 10,015 images Training (72%), Validation (8%), Testing (20%) + Image Augmentation
Gouda et al. [5] 2022	Oversampling, ESRGAN (Enhancing and reducing noise), Augmentation (rotation, reflection, shifting, brightness adjustment), Image Resizing, Custom Contrast Method	Inception v3 Binary classification (BKL, MEL)	ISIC2018 dataset, 11,527 images Training = 10,015, Testing = 1,512	AUC-ROC = 85.8%

Bassel et al. [3]
2022
	Resizing,
	Normalization
		ResNet50,
		Xception (main),
		VGG16
		Binary classification
		(BKL, MEL)
			ISIC 2019 dataset
			Total = 25,331 images
			Training = 70%, Testing = 30%
				AUC-ROC = 90.9%

Nakai et al. [10]
2022
	Lesion segmentation,
	Normalization,
	Data Augmentation,
	Resizing
		Enhanced Transformer module
		+ benchmark CNN models
		Multi-class classification
		(MEL, NV, AKIEC)
			ISIC2017 dataset, 2,750 images
			Training = 2,000, Validation = 150,
			Testing = 600
				AUC-ROC = 92.1%

Akter et al. [1]
2022
	Image Resizing,
	Normalization,
	Augmentation
		InceptionV3, DenseNet,
		ResNet50, MobileNet, Xception,
		VGG16, Transfer learning
		Multi-class classification
		(AKIEC, BCC, BKL, DF, MEL, NV, VASC)
			HAM10000 dataset, 10,015 images
			Training = 70%, Validation = 10%,
			Testing = 20%
				AUC-ROC = 78%

Medhat et al. [19] 2022	Resizing image, Augmentation	AlexNet (transfer learning) + augmented, MobileNet-V2 (transfer learning) Multi-class classification (BCC, SCC, MEL, NV, ACK, SEK)	PAD-UFES-20 dataset, 2,298 images Training = 80%, Testing = 20%	Accuracy = 94.07%

Li et al. [?]
2022
	Augmentation,
	Oversampling,
	Weighted random sampling
		Quantum Inception ResNet-V1
		Binary classification
		(BKL, MEL)
			ISIC 2019 dataset, 25,331 images
			Training = 80%, Validation = 10%,
			Testing = 10%
				Accuracy = 98%

Banasode et al. [20] 2021	Blur Images, RGB to HSV conversion, RGB to Grayscale conversion, Color thresholding, Masking Image, Segmentation, Image Transformation	Support Vector Machine Binary classification (BKL, MEL)	ISIC dataset, 5,341 images Training = 80%, Testing = 20%	Accuracy = 96.9%

Datta et al. [4]
2021
	Oversampling,
	Under Sampling,
	Normalization
		KNN mining/clustering
		Soft-Attention
		Multi-class classification
		(BCC, SCC, MEL, NV, ACK, SEK)
			HAM10000 dataset = 10,015
			ISIC-2017 = 2,000 images
			ISIC-2017: Training/Validation = 1,400,
			Testing = 600
				Accuracy = 96.9%

Jain et al. [6]
2021
	High-Frequency sampling with Augmentation (rotation, zooming, shifting), Normalization	Transfer learning
		(VGG19, InceptionV3,
		InceptionResNetV2,
		ResNet50, Xception, MobileNet)
		Multi-class classification
		(BCC, SCC, MEL, NV, ACK, SEK)
			HAM10000 dataset, 10,015 images
			After duplication = 5,514
			Training = 70%, Validation = 10%,
			Testing = 20%
				AUC-ROC = 90.48%

	Normalization			

III. DATASETS

We have used four publicly available datasets of dermoscopic images in this research.

A. ISIC-2019

This dataset consists of dermoscopic images collected for the ISIC's 2019 challenge. It consists of 25,331 images comprising various lesion categories that have crucial significance during research regarding medical imaging and skin cancer classification. The 8 categories of the dataset consist of Melanoma (MEL), Melanocytic Nevus (NV), Basal Cellular Carcinoma (BCC), Actinic Keratoses (AKIEC), Benign Keratosis-like Lesions (BKL), Dermatofibroma (DF), Vascular Lesions (VASC) and Squamous Cellular Carcinoma (SCC). Critical metadata has also been included, such as age, sex, and anatomic site, along with precious contextual records for further insights, and even a separate check dataset comprising the outlier class of data points not within the main data is present as well. The dataset also contains images from ISIC-2018 (HAM10000) and ISIC-2017 challenges and can be easily sourced from the ISIC online archive or Kaggle.

B. ISIC-2020

The 2020 version of the ISIC dataset contains 32,542 BKL-category images and 563 MEL images, which, while a mix of crucial skin conditions, may constitute a class imbalance. Each file in this data contains patterns along with the visual photo, basic metadata, and a unique patient identifier; the metadata elaborates on the minute details of the dataset, thereby proving beneficial to dermatologists for research and experimentation. Histopathology has been employed to confirm each malignant diagnosis, while a combination of expert assessment, long-term observation, and histopathology has been employed for every benign case. Due to the existence of class imbalance in this dataset, the standard operation would be to employ augmentation or other specialized sampling methods to tackle the problem before being used as an input to the classifiers.

C. PAD-UFES

Containing a variety of medical images captured through the medium of smartphone devices, this dataset composed of 2298 rotoscope images taken from 1373 patients comprises 6 different lesion categories with 58.4% of the lesions being proven through biopsy, and is a data repository of great significance for skin cancer classification researches.

D. DERMQUEST-DERMIS

Offering a wide range of dermoscopic images accompanied by abundant data for research and analysis, this dataset caters to the needs of skin cancer classification researchers through the inclusion of some of the Melanoma (MEL) and NON-MEL categories of images. This is an excellent dataset that can be used as the foundation for projects that aim to deliver improved machine-learning models with better accuracy and enhanced generalization.

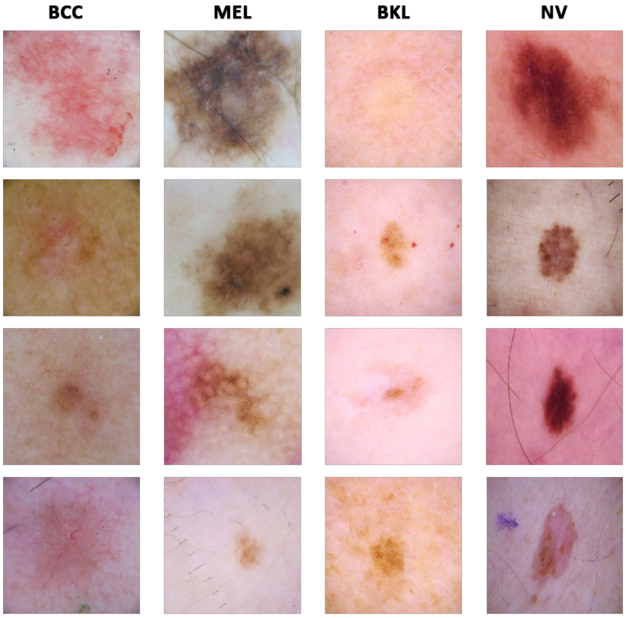

FIGURE 4: The figure presents a short subset of lesion variants, visual sample are arranged in column format such that images of the lesion lie under the Class from which they have been categorized (BCC, MEL, BKL and NV) in the acquired dataset.

E. LABEL DETAILS

In addition to the data sets that we have elaborated upon here, some important labels regarding the categories of lesion images that we have used are also introduced here. There are four lesion categories that we have focused our research on BKL, MEL, NV, and BCC. BKL or Benign Keratosis-like Lesions are non-cancerous though difficult to differentiate from the cancerous lesions due to their resemblance. They look similar to warts and usually form due to long-term sun exposure for adults. MEL or Melanoma is an aggressive, malignant category of lesions that affects melanocytes and is characterized by weird-shaped and colored moles. It forms as a result of ultraviolet exposure, sunburn, or genetic factors, visually represented as A in Fig 4. NV or Nevi are benign melanocyte growths which are typically referred to as moles that may be innate or acquired later in life. They may be flat or raised, and their colors are limited to brown or black, though the irregular borders and the different colors may lead to them being confused with melanoma, visually represented as B in Fig 4. Finally, BCC, or Basal Cell Carcinoma, is the least aggressive and the most common form of skin cancer, which, while unable to metastasize, can still wreak havoc on its locality if untreated. This form of lesion may appear veiny with flesh or brownish colors and usually forms because of exposure to carcinogens, radiation, chronic sun, or even something as simple as older age, visually represented as E and F in Fig 4. Benign Keratosis-like Lesions *BKL* is a

TABLE 2: Summary of the images and classes in the acquired datasets

No.	Dataset	Total Categories	Short Name of Categories	Total Images
1	ISIC 2019 Dataset	8	AKL, BCC, MEL SCC, BKL, NV, DF, VASC	25,331
2	PDF-UFES	6	ACK, BCC, MEL, SCC, NEV, SEK	1,612
3	ISIC 2020 Dataset	2	BKL, MEL	33,126
4	DERM QUEST DERMIS Dataset	2	MEL, NOTMEL	180

common lesion that is not cancerous. Seborrheic keratoses are usually brown, black, or light tan. The growths (lesions) look waxy or scaly and slightly raised as presented by C and D in Figure 4. Table 2 is constructed to present the number of images found in each class label and employed in the study for quantifying purposes.

IV. PROPOSED METHODOLOGY

Proposed method consists of three steps, i.e., preprocessing, feature extraction, and classification of dermoscopic images. Figure 5 shows the flow diagram of the proposed method. The following subsections provide a detailed description of each step.

A. PREPROCESSING

Datasets that have been used in this research include ISIC-2019, ISIC-2020, PAD-UFES and DermQuestDermIS. Figure 6 presents a detailed illustration of preprocessing applied in the proposed methodology. In the first step, images have been resized into 150x150x3. Standardization was used to reduce the different resolutions and formats of the acquired image data to a single resolution (150x150) and format which will be useful for ensuring the effectiveness and the speed of the learning model. Normalization is then applied to reduce the fluctuations between the pixel values into the range of 0-1 so that the learning model does not tend to focus on and develop a bias towards the more prominent features.

Class imbalance problem exists in the dataset; therefore, to balance the classes, we have used High-frequency balancing to reduce the overfitting to some classes that have numbers one-fourth $1/4$ of the class, which have the highest number of representations. We introduce augmentation, dropout layer, and Batch Normalization reversed in a way that the minority class representation will equal to the majority class representation of samples. The High-frequency Balancing and Augmentation method is utilized specifically to introduce transformations into those spaces of the dataset where the minority classes appear with a higher frequency. By ensuring that the minority class is appropriately represented in high-frequency regions without damaging the overall integrity of the data, this technique easily allows for an enhancement to the generalization capabilities and the robustness of the AI model. The dataset is divided into the train, validation, and test sizes of 80%, 10%, and 10%, respectively.

B. DYNAMIC PATCHING VIT MODEL

After preprocessing, we extracted automated features from three different deep learning architectures, including ResNet50 [22], ViT [23], and VGG-16. Features extracted from these three models are concatenated to form a single feature vector, which is then fed into HDFNet for classification [24], [25]. Some important components of the ViT that set it apart from other CNNs are the following: Dynamic

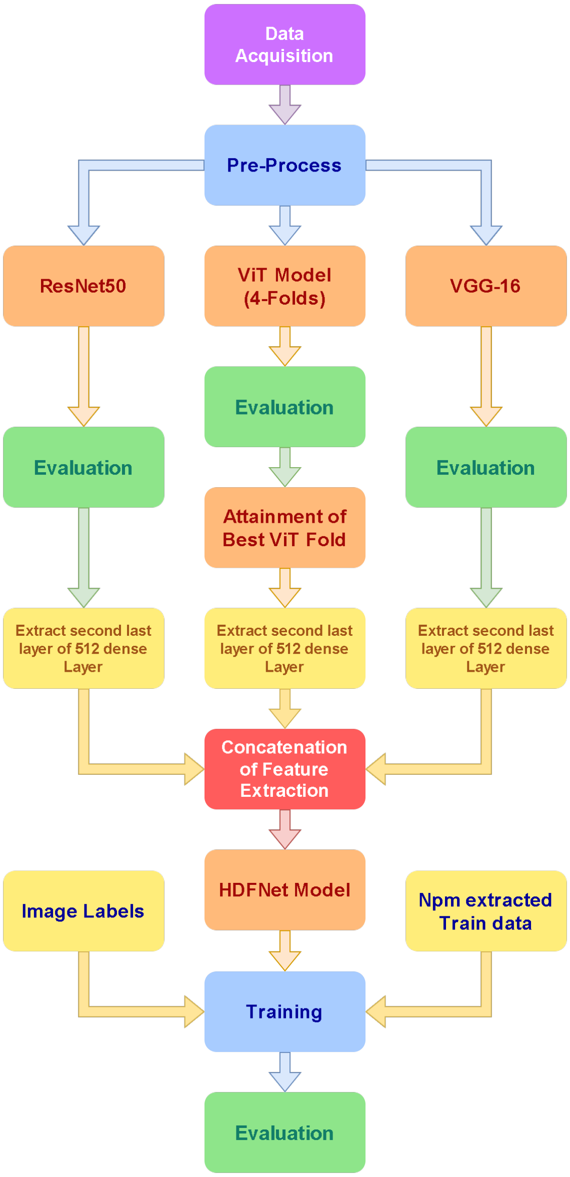
Data Acquisition
Pre-Process
ResNet50
ViT Model (4-Folds)
VGG-16
Evaluation
Evaluation
Evaluation
Attainment of Best ViT Fold
Extract second last layer of 512 dense Layer
Extract second last layer of 512 dense Layer
Extract second last layer of 512 dense Layer
Concatenation of Feature Extraction
HDFNet Model
Image Labels
Npm extracted Train data
Training
Evaluation

FIGURE 5: Flow diagram of the Proposed Method

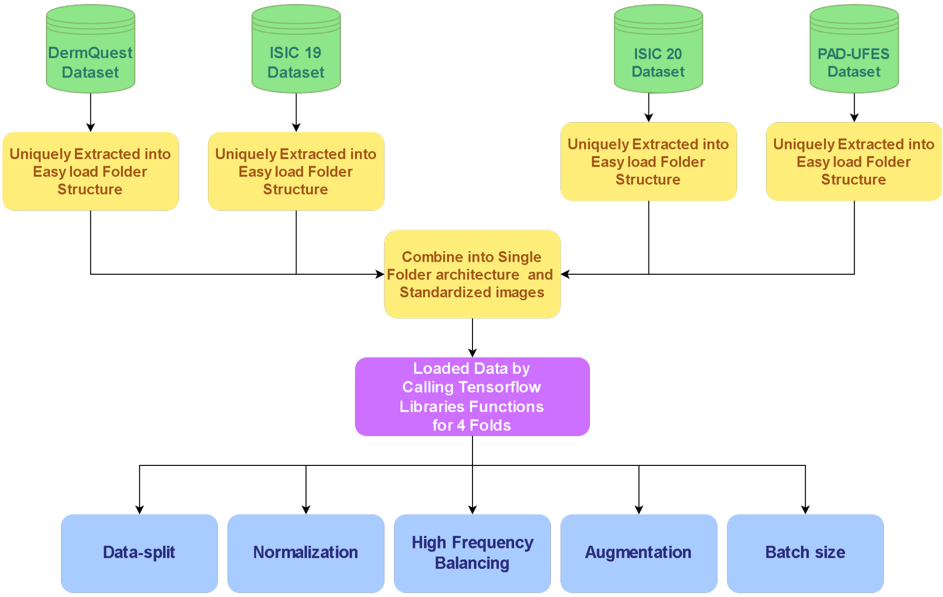
DermQuest Dataset
ISIC 19 Dataset
ISIC 20 Dataset
PAD-UFES Dataset
Uniquely Extracted into Easy load Folder Structure
Uniquely Extracted into Easy load Folder Structure
Uniquely Extracted into Easy load Folder Structure
Uniquely Extracted into Easy load Folder Structure
Combine into Single Folder architecture and Standardized images
Loaded Data by Calling Tensorflow Libraries Functions for 4 Folds
Data-split
Normalization
High Frequency Balancing
Augmentation
Batch size

FIGURE 6: Proposed Preprocessing Method

Patch Extractor, Token Embedder, Transformer Encoder, and MLP. When the data is input to the ViT, the patch extractor divides each image into equal-sized patches upon which linear transformations are applied to transform them to a lower-dimensional vector space [26]. Proposed architecture of the ViT is shown in figure 7. The architecture consists of the following important layers: [27] [28]

- **Dynamic Patch Extractor:** Whereas a normal patch extractor divides images into fixed-size patches, the problem at hand is lesions might be too large or too small in different cases. These variations in lesion sizes mean that fixed-size patch extraction technique might lead to loss of important details or unnecessary fragmentation of lesions. Therefore, this dynamic-patching component works by dividing images into patches of different sizes depending on the size of the lesions. The three different lesion sizes that are considered by us are as follows: if the lesion size in the image is less than 25%, then smaller patch sizes (8x8) are utilized for finer details whereas if the lesion covers 25-50% or greater than 50% of the image, then the appropriate patch size is increased to 16x16 and 32x32 respectively so as to provide more context. A patch extractor works by dividing images of patches of the size $P \times P$. Each image $x \in \mathbb{R}^{H \times W \times C}$ is split into $N = \frac{HW}{P^2}$ patches, where each patch is flattened into a vector $x_p \in \mathbb{R}^{P^2 \cdot C}$.
$$\{x_p^i\}_{i=1}^N \quad \text{where} \quad x_p^i \in \mathbb{R}^{P^2 \cdot C} \tag{1}$$
However for the case of implementation in order to attain patch Figure **??** illustrates the applied methodology to attain the area estimation of lesion from complete images the architecture applied is designed in such a way that the applied images are first convert to gray scales after which multiple threshold-ing limits by manually selection to attain estimated clearer frames after which lesion size is extraction and then simply processed by VIT model
- **Token Embedder:** The flattened patches are then linearly transformed into a lower-dimensional vector space using a learnable matrix $E \in \mathbb{R}^{(P^2 \cdot C) \times D}$, where D is the embedding dimension [29].
$$z_0 = [x_p^1 E; x_p^2 E; \dots; x_p^N E] + E_{pos} \tag{2}$$
Here, E_{pos} is the positional encoding that adds positional information to the patch embeddings.
- **Transformer Encoder:** This component consists of L layers, each containing a Multi-Head Self-Attention (MSA) mechanism and a Feedforward Neural Network (FFN). The input to each layer is processed as follows:
$$z'_l = \text{MSA}(\text{LN}(z_{l-1})) + z_{l-1} \tag{3}$$
$$z_l = \text{FFN}(\text{LN}(z'_l)) + z'_l \tag{4}$$
where $l = 1, \dots, L$, z_0 is the initial embedding, LN is layer normalization, and MSA is computed as:
$$\text{MSA}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) W^O \tag{5}$$
with each attention head i defined as:
$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{6}$$
and the scaled dot-product attention given by:
$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right) V \tag{7}$$

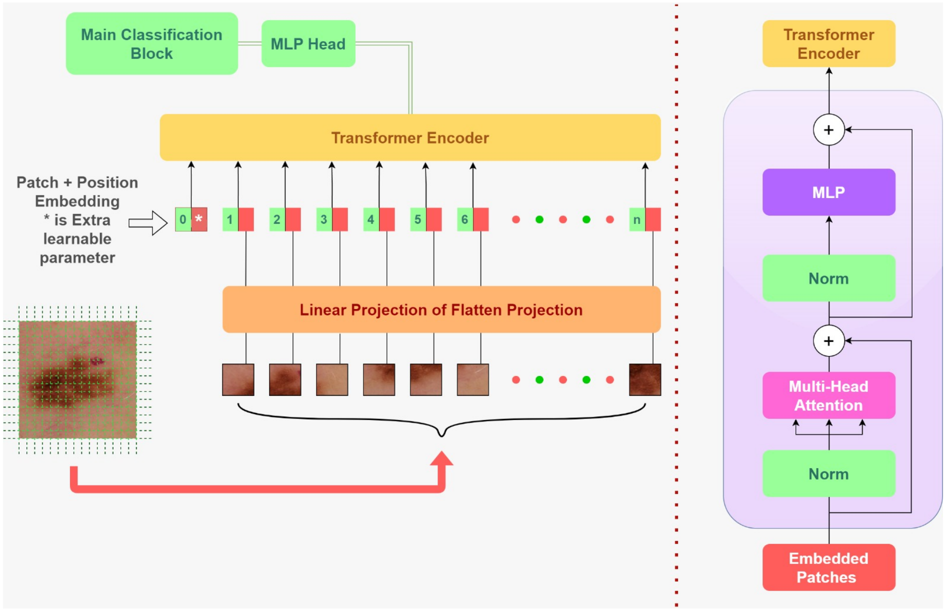
Main Classification Block
MLP Head
Transformer Encoder
Patch + Position Embedding
* is Extra learnable parameter
0
*
1
2
3
4
5
6
n
Linear Projection of Flatten Projection
Transformer Encoder
+
MLP
Norm
+
Multi-Head Attention
Norm
Embedded Patches

FIGURE 7: Proposed Vision Transformer Model

where W_i^Q, W_i^K, W_i^V, and W^O are learnable weight matrices.
The first block includes a skip connection 1 of 2 skip connections in the transformer encoder. This first skip connection will be used later for residual connection. After the skip connection, a layer of normalization is applied to the input. Layer normalization helps normalize the activations across the features and stabilizes the training process. After which multi-head attention is applied to the normalized input. The number of attention heads is specified to 12 in the configuration. The key dimension for each attention head is specified as 512. Multi-head attention allows the model to attend to different positions of the input sequence and capture dependencies. This residual connection helps the model learn the identity function and facilitates the flow of information across layers. after which the output of the previous layer is passed to the second skip connection, skip 2, by assigning the output of the previous addition to it. where normalization is applied to the output of the previous addition. After this, the MLP block is introduced, which provides additional non-linearity and learnable parameters to the model. This residual connection helps the model learn the identity function and facilitates the flow of information across layers. Finally, the function returns the output of the transformer encoder layer.

- **MLP Head:** The final layer is a Multi-Layer Perceptron (MLP) that performs the classification. The MLP consists of multiple fully connected layers with non-linear activation functions such as GeLU or ReLU [30].

$$\mathrm{MLP}(z_L) = W_2 \cdot \mathrm{GeLU}(W_1 \cdot z_L + b_1) + b_2 \tag{8}$$

where z_L is the output from the last Transformer Encoder layer, and W_1, W_2, b_1, and b_2 are learnable parameters.

The first dense layer in the MLP block is set to 3072 in the applied configuration. This layer applies a linear transformation to the input, expanding its dimension to 3072. GELU (Gaussian Error Linear Unit) activation function is used, which introduces non-linearity and helps capture complex patterns in the data. After the first dense layer, a dropout regularization is applied with a rate of 0.1 in the configuration. Dropout randomly sets a fraction of the input units to 0 during training, which helps prevent overfitting and improves generalization. The second dense layer in the MLP block is set to 512 in the configuration. This layer applies another linear transformation, reducing the dimension back to the original hidden dimension of 512. No activation function is used in this layer, allowing the output to have both positive and negative values. Another dropout regularization is applied after the second dense layer, with the same dropout rate of 0.1. The output of the MLP block is returned, which has a shape of (None, 257, 512). This output is then added to the input of the MLP block using a residual connection in the transformer encoder layer. Other models utilized in this research include VGG-16 and ResNet-50.

C. ADAPTIVE LAYER CONFIGURATION VGG16(ALC-VGG16)

The customized VGG-16 utilized by us entails 13 convolutional layers along with 3 fully connected layers. Labeled as ALC-VGG16 (Adaptive Layer Configuration VGG16), it retains the simplicity and uniformity of the original model but includes adaptive layer configuration for enhanced performance on dermoscopic datasets. Key customizations performed on this model include adaptive padding in convolutional layers, LeakyReLU for a better gradient flow, Dynamic pooling instead of max pooling, dropout layers for regularization, and a penultimate dense layer which gives a 512-dimensional feature vectors for fusion. In this document, we describe the customized components of the ALC-VGG16 model, including formulas for each:

1. Adaptive Padding in Convolutional Layers

To maintain consistent spatial dimensions across convolutional layers, adaptive padding is applied. Given an input feature map $X \in \mathbb{R}^{H \times W \times C}$ and a convolutional kernel of size $k \times k$, the padding p is calculated as:

$$p = \left\lceil \frac{k-1}{2} \right\rceil$$

This ensures that the output feature map dimensions match the input dimensions, preserving spatial information.

2. LeakyReLU Activation

For better gradient flow and to avoid inactive neurons, the LeakyReLU activation function replaces the standard ReLU:

$$f(x) = \begin{cases} x, & \text{if } x > 0 \\ \alpha x, & \text{if } x \leq 0 \end{cases}$$

where α is a small constant, typically set to 0.01. This allows a small gradient when $x < 0$, enhancing the learning process by maintaining non-zero gradients.

3. Dynamic Pooling

Dynamic pooling replaces fixed max pooling to capture more nuanced spatial details by selecting pooling behavior based on the input. For a pooling window of features $f_{i,j}$, dynamic pooling can be represented as:

$$P(X) = \text{adaptive\_pool}(f_{i,j}) \quad \forall i, j \in \text{pooling window}$$

This approach adapts the pooling operation based on local feature values, leading to more flexible and expressive spatial representations.

4. Dropout Layer for Regularization

To reduce overfitting, dropout randomly sets a fraction p of the input units to zero during training. Given an input vector z, the dropout operation can be defined as:

$$z' = z \cdot \text{mask}$$

where mask is a binary vector where each element is 1 with probability $(1 - p)$ and 0 otherwise.

5. Penultimate Dense Layer with a 512-Dimensional Feature Vector

The penultimate dense layer generates a 512-dimensional feature vector, providing a compact, high-level representation of the input. This layer can be represented as:

$$\text{Dense}(z) = Wz + b$$

where:

- $W \in \mathbb{R}^{512 \times d}$ is the weight matrix,
- $b \in \mathbb{R}^{512}$ is the bias vector,
- $z \in \mathbb{R}^{d}$ is the input feature vector.

The resulting 512-dimensional vector is useful for feature fusion or further analysis.

Layer	Output Shape	Param #
input_layer	None, 150, 150, 3	0
conv2d	None, 150, 150, 64	1,792
leaky_relu	None, 150, 150, 64	0
conv2d_1	None, 150, 150, 64	36,928
leaky_re_lu_1	None, 150, 150, 64	0
average_pooling2d	None, 75, 75, 64	0
conv2d_2	None, 75, 75, 128	73,856
leaky_re_lu_2	None, 75, 75, 128	0
conv2d_3	None, 75, 75, 128	147,584
leaky_re_lu_3	None, 75, 75, 128	0
average_pooling2d_1	None, 37, 37, 128	0
conv2d_4	None, 37, 37, 256	295,168
leaky_re_lu_4	None, 37, 37, 256	0
conv2d_5	None, 37, 37, 256	590,080
leaky_re_lu_5	None, 37, 37, 256	0
conv2d_6	None, 37, 37, 256	590,080
leaky_re_lu_6	None, 37, 37, 256	0
average_pooling2d_2	None, 18, 18, 256	0
conv2d_7	None, 18, 18, 512	1,180,160
leaky_re_lu_7	None, 18, 18, 512	0
conv2d_8	None, 18, 18, 512	2,359,808
leaky_re_lu_8	None, 18, 18, 512	0
conv2d_9	None, 18, 18, 512	2,359,808
leaky_re_lu_9	None, 18, 18, 512	0
average_pooling2d_3	None, 9, 9, 512	0
conv2d_10	None, 9, 9, 512	2,359,808
leaky_re_lu_10	None, 9, 9, 512	0
conv2d_11	None, 9, 9, 512	2,359,808
leaky_re_lu_11	None, 9, 9, 512	0
conv2d_12	None, 9, 9, 512	2,359,808
leaky_re_lu_12	None, 9, 9, 512	0
average_pooling2d_4	None, 4, 4, 512	0
flatten	None, 8192	0
dense	None, 4096	33,558,528
leaky_re_lu_13	None, 4096	0
dropout	None, 4096	0
dense_1	None, 4096	16,781,312
leaky_re_lu_14	None, 4096	0
dropout_1	None, 4096	0
dense_2	None, 512	2,097,664

TABLE 3: Summary of the ACL VGG16 model architecture

D. RESNET-50 WITH DERMATOLOGICAL FEATURE ENHANCEMENT (DFE-RESNET50)

Our customized ResNet-50 variant consists of 49 convolutional layers and a final fully connected layer, for a total of 50. Retaining the skip connection characteristic of ResNet, it incorporates specific enhancements for dermatological analysis of images through the following customizations: In this context, we describe a feature processing module designed to process input features using two attention mechanisms: **Channel Attention** and **Spatial Attention**.

Input Features

Let the input features be represented as $X \in \mathbb{R}^{C \times H \times W}$, where:

- C is the number of channels,
- H is the height of the input feature map,
- W is the width of the input feature map.

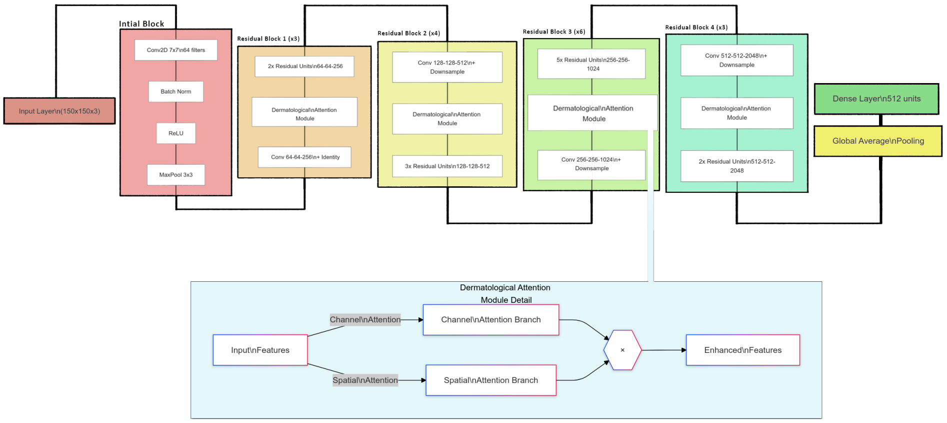
Intial Block
Conv2D 7x7\n64 filters
Batch Norm
ReLU
MaxPool 3x3
Input Layer\n(150x150x3)
Residual Block 1 (x3)
2x Residual Units\n64-64-256
Dermatological\nAttention Module
Conv 64-64-256\n+ Identity
Residual Block 2 (x4)
Conv 128-128-512\n+ Downsample
Dermatological\nAttention Module
3x Residual Units\n128-128-512
Residual Block 3 (x6)
5x Residual Units\n256-256-1024
Dermatological\nAttention Module
Conv 256-256-1024\n+ Downsample
Residual Block 4 (x3)
Conv 512-512-2048\n+ Downsample
Dermatological\nAttention Module
2x Residual Units\n512-512-2048
Dense Layer\n512 units
Global Average\nPooling
Dermatological Attention Module Detail
Input\nFeatures
Channel\nAttention
Channel\nAttention Branch
Spatial\nAttention
Spatial\nAttention Branch
×
Enhanced\nFeatures

FIGURE 8: Customized Architecture of DFE-ResNet 50 Model

Channel Attention Branch

The Channel Attention branch adjusts the importance of each channel by computing a **Channel Attention map**, $M_c \in \mathbb{R}^{C\times 1\times 1}$. This map is applied to the input features as follows:

$$X_c = M_c \odot X \tag{9}$$

where $\odot$ denotes element-wise multiplication, and X_c is the output feature map with enhanced channels. The Channel Attention map M_c is computed by passing X through a series of operations, such as global average pooling and fully connected layers, to highlight important channels.

Spatial Attention Branch

The Spatial Attention branch enhances important spatial regions by computing a **Spatial Attention map**, $M_s \in \mathbb{R}^{1\times H\times W}$. This map is applied to the input features as:

$$X_s = M_s \odot X \tag{10}$$

where X_s is the output feature map with enhanced spatial regions. The Spatial Attention map M_s is derived by applying operations such as convolution and activation functions to focus on significant spatial locations.

Combination (Multiplication) Node

The outputs of the Channel and Spatial Attention branches are combined to produce the final refined features:

$$X_{\text{enhanced}} = X_c \odot X_s = (M_c \odot X) \odot (M_s \odot X) \tag{11}$$

Here, X_{enhanced} represents the **Enhanced Features**, which retain both channel and spatial information, making them more effective for dermatological analysis tasks.

Enhanced Features

The final output, X_{enhanced}, represents features optimized to focus on relevant channels and spatial areas. This refined output can be used for further tasks, such as classification or detection.

E. HDF-NET MODEL

When the different models described above have processed the input and finished with their work, feature vectors are obtained through the extraction of their penultimate layers, which are then fused through the concatenation feature fusion technique, which allows for the dimensionality of the fused feature, to be enhanced. In our case with 3 feature vectors of 512-dimensions each, our final feature vector would have a dimensionality of 1536. [25] [31] This feature fusion technique allows for the strengths of different models to be integrated into a single vector, which, when utilized as input for any model in the future, allows the model to work with greater efficiency. Now our compiled dataset and the fused feature vector are employed as input for a Deep Neural Network (DNN) where the feature vector is passed through a dense layer that reinforces the effect of the combined information for effective learning by a layer containing 1024 units which strengthens the convergence of the variety of feature representations onto a single vector. After which optimal features extracted from 1024 layer is then passed to the classification of a layer of 4 neurons to s the class label attained by model.

V. RESULTS AND DISCUSSION

A. EVALUATION MATRICES

Finished with the description of the datasets and the label details, we now proceed to the description of our selection of evaluation metrics. We chose to employ the AUC-ROC score, Recall, and F1-Score to attain a comprehensive understanding of our models' performance.

1) Receiver Operation Curve (ROC)

ROC is a graphical representation that plots the curve between the true positive rate and the false positive rate. It is derived through the calculation of specificity and sensitivity.

$$\text{ROC} = \{(\text{Specificity}_i, \text{Precision}_i)\}_{i=0}^{n} \tag{12}$$

2) Specificity

Specificity measures the proportion of true negatives that are correctly identified by the neural network architecture. It is calculated through the following formula:

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}} \tag{13}$$

where FP and TN are False Positives and True Negatives, respectively.

3) Precision

Precision refers to the proportion of true positives that are correctly identified by the neural network. Its calculation is carried out through the following process:

$$\text{Precision} = \frac{\text{TP}}{\text{FP} + \text{TP}} \tag{14}$$

where TP and FP are True Positives and False Positives, respectively.

4) AUC

Area Under the ROC Curve or AUC refers to a numerical value existing between 0 and 1 which signifies the efficiency levels of the classifier architectures after calculating the ability of the classifier to correctly identify positive and negative classes.

$$\text{AUC} = \int_0^n \text{Sensitivity}(\text{Specificity})\, d\text{Precision} \tag{15}$$

In the domain of medicine where different categories or progression stages of the disease may vary in terms of their rarity, it is imperative to utilize a metric that has the ability to display varied discriminating thresholds to account for this variation. In contrast, simple accuracy has a simple fixed threshold without any flexibility and is therefore not suitable for usage as our metric, compared to the AUC-ROC. Furthermore, two crucial features of the diagnosis are processed by the AUC-ROC: true positives and true negatives, allowing professionals to easily evade the dangers of false positives or false negatives, which could constitute fatal consequences in real-world settings. In addition, our metric also utilizes the moving threshold to deal with the issue of class imbalance and is, therefore, a more reliable performance evaluator than general accuracy.

5) Recall

Recall is another machine learning evaluation metric used to define how often the employed model calculates and correctly identifies the class positive instances. Following is the formula of the recall function:

$$R = \frac{TP}{TP + FN} \tag{16}$$

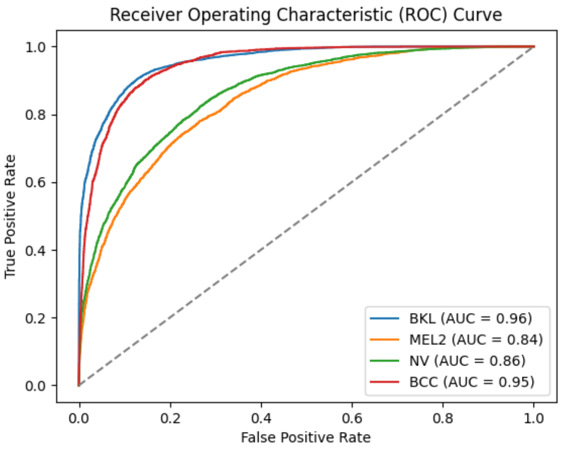

FIGURE 9: AUC-ROC curve of Proposed Custom Hybrid Net Dense Fusion Model.

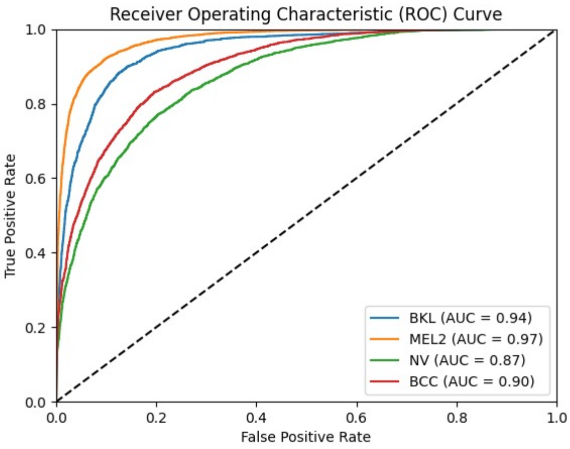

FIGURE 10: AUC-ROC curve achieved using ViT.

6) F1 Score

F1 score is a machine learning evaluation metric that is generally used to evaluate how many times the model has correctly evaluated correct prediction across data sets using precision and recall. The following Formula is used to calculate the F1-score:

$$F1 = 2 \cdot \frac{P \cdot R}{P + R} \tag{17}$$

Figure 9 and 10, shows that the Vision Transformer model's perform betteer than the feature fused DNN. The former's performance scores for the different lesion categories are 0.96 (BKL), 0.84 (MEL2), 0.86 (NV), and 0.95 (BCC), whereas the latter's scores are 0.94 (BKL), 0.97 (MEL2), 0.87 (NV) and 0.90 (BCC). In terms of class discrimination capabilities it is evident that the DNN performs better in terms of MEL and NV compared to BKL and BCCAs is noticeable. Further

TABLE 4: Precision, Recall, F1-Score, AUC-ROC and Overall Accuracy performance on various models with Raw and Standardized Data.

Models	Precision				Recall				F1-Score				AUC-ROC	Overall Accuracy
	BKL	MEL	NV	BCC	BKL	MEL	NV	BCC	BKL	MEL	NV	BCC		
Raw Data + VGG16	0.69	0.68	0.71	0.67	0.64	0.62	0.69	0.65	0.66	0.65	0.70	0.66	0.80	0.68
Raw Data + VGG22	0.67	0.66	0.69	0.65	0.62	0.60	0.67	0.63	0.64	0.63	0.68	0.64	0.78	0.65
Raw Data + ResNet50	0.71	0.70	0.73	0.69	0.68	0.66	0.71	0.67	0.69	0.68	0.72	0.68	0.82	0.70
Raw Data + ResNet101	0.70	0.69	0.72	0.68	0.66	0.65	0.69	0.66	0.68	0.67	0.71	0.67	0.81	0.69
Raw Data + 4 Fold ViT	0.74	0.73	0.76	0.72	0.71	0.69	0.74	0.70	0.72	0.71	0.75	0.71	0.86	0.74
Feature Fused Model	0.77	0.76	0.79	0.75	0.75	0.74	0.77	0.73	0.76	0.75	0.78	0.74	0.88	0.77
Standardized Data + VGG16	0.78	0.77	0.79	0.76	0.76	0.75	0.77	0.74	0.77	0.76	0.78	0.75	0.89	0.75
Standardized Data + VGG22	0.71	0.70	0.72	0.69	0.69	0.68	0.70	0.67	0.70	0.69	0.71	0.68	0.86	0.70
Standardized Data + ResNet50	0.72	0.71	0.73	0.70	0.73	0.72	0.74	0.71	0.72	0.71	0.73	0.70	0.87	0.72
Standardized Data + ResNet101	0.71	0.70	0.72	0.69	0.72	0.71	0.73	0.70	0.71	0.70	0.72	0.69	0.87	0.71
Standardized Data + 4 Fold ViT	0.87	0.86	0.88	0.85	0.86	0.85	0.87	0.84	0.87	0.86	0.88	0.85	0.90	0.87
Standardized Data + Feature Fused Model	0.91	0.90	0.92	0.89	0.89	0.88	0.90	0.87	0.90	0.89	0.91	0.88	0.92	0.90
Standardized Data + ACL-VGG16	0.87	0.62	0.79	0.65	0.77	0.87	0.56	0.70	0.54	0.72	0.66	0.68	0.82	0.73
Standardized Data + ACL-VGG22	0.81	0.75	0.62	0.76	0.75	0.35	0.80	0.81	0.78	0.48	0.63	0.78	0.8975	0.75
Standardized Data + DFE-ResNet50	0.82	0.75	0.63	0.75	0.74	0.35	0.80	0.80	0.79	0.50	0.63	0.76	0.89	0.75
Standardized Data + DFE-ResNet101	0.81	0.74	0.65	0.76	0.74	0.36	0.81	0.80	0.79	0.51	0.66	0.76	0.90	0.74
Standardized Data + 4 Folds Dynamic Patching ViT	0.90	0.91	0.89	0.90	0.90	0.90	0.88	0.91	0.92	0.90	0.88	0.91	0.9025	0.90
Standardized Data + Feature Fused Model	0.95	0.97	0.89	0.91	0.94	0.96	0.88	0.90	0.92	0.95	0.88	0.90	0.9215	0.92

TABLE 5: Comparison of performance of the EfficientNet Series on both AUC-ROC and Accuracy

Model	EfficientNetB1	EfficientNetB2	EfficientNetB3	EfficientNetB4	EfficientNetB5	EfficientNetB6
AUC-ROC	0.86	0.86	0.82	0.85	0.88	0.87
Accuracy	0.77	0.77	0.78	0.77	0.80	0.81

information regarding F1-score, recall, and precision is depicted in Table 4 of the in-depth employed models.

B. EVALUATION

Based on the information above, the precision, recall, and f1-scores for the different classes of lesions are 0.95, 0.94, and 0.95 for BKL, 0.97, 0.96, and 0.96 for MEL, 0.88, 0.89 and 0.88 for NV, and 0.91, 0.90 and 0.90 for BCC, respectively. Table 4 presents more detailed experiments that have been covered in this study, whereas Table 5 presents the attainment of the EfficientNet Model and the presentation of exclusion due to limited achievement in performance on the employed dataset. In the given study a comparison study has been illustrated using Table 6 which presents the attainment of the proposed model with respect to the previous studies exploring the efficient use of AI in the domain of Skin Cancer.

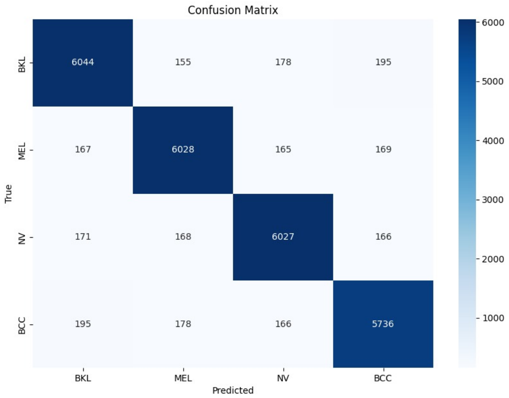

FIGURE 11: Confusion Matrix of Proposed Model.

Furthermore we can easily notice the overall performance differences between the previous models and the feature fused DNN. While ViT, ResNet50, and VGG16 achieved an overall AUC-ROC score of 90.25%, 90.27%, and 89.75%, respectively, the feature fused DNN shows a significant improvement through its score of 92.15%. Figure 11 illustrates the performance of the Hybrid Net Dense Fusion model attained by averaging fusion and illustrates the True Positive, False Positive, True Negative, and False Negative prediction of labels for each label class in the study. The significant improvement in results that we see here results from factors such as the contribution of the strengths of different classifiers through the unified feature map, which allows the DNN to be comprehensively strengthened when dealing with the dataset, thus leading to a more comprehensive analysis of the features which leads to better generalization and greater efficiency. For future work, the feasibility of inclusion of other CNN architectures can be tested by the researchers along with addressing the lack of external validity. In addition, future research could extend by expanding on the four categories of lesion that we have experimented on and include a greater number of categories to work with. Finally, the application of this methodology in real-time settings can also be tested in the future.

VI. CONCLUSION AND FUTURE WORK

We conclude our research with the deep neural network being provided with a unified feature vector obtained from the concatenation of the corresponding feature extractions acquired from custom Adaptive Layer Configuration VGG-16 (ALC-VGG16), Dermatological Feature Enhancement ResNet-50 (DFE-ResNet50), and a custom Vision Transformer (ViT) model. This approach resulted in an impressive AUC-ROC score of 92.15%. The AUC-ROC scores for the specific

TABLE 6: Comparison of Past performed research with the Proposed Models with applied Datasets

Method	**Year**	**Dataset**	**Accuracy / AUC**
Kahia et al [17]	2022	Melanoma images: 374 Nevus images: 1372 Seborrheic keratosis images: 254	Accuracy = 73.33%
Rezaoana et al [13]	2021	Kaggle competition 2019 Images=25,780	Accuracy = 79.45%
Kaya et al. [2]	2023	Kaggle competition 2021 Total images used=3297	Accuracy = 83.33%
Ali et al. [18]	2022	HAM10000 dataset 10015 images	Accuracy = 87.91
Jaisakthi et al. [11]	2022	ISIC 2019 and ISIC 2020 skin cancer classification. ISIC 2020=33,126 ISIC 2019=25331 It includes BCN_20000, HAM10000, and MSK	AUC-ROC = 91%
Gouda et al. [5]	2022	ISIC 2018 dataset Total images = 11,527	AUC-ROC = 85.8%
Bassel et al. [3]	2021	ISIC 2019 Challenge Used 1800 images	AUC-ROC = 90.9%
Nakai et al. [10]	2022	ISIC2017 =2750 Three Classes Dataset	AUC-ROC = 92.1%
Akter et al. [1]	2022	HAM10000 dataset 10015 images	AUC-ROC = 78%
Jain et al. [6]	2021	HAM10000 dataset 10015 images	AUC-ROC = 90.48%
Proposed VIT Model	**2024**	**ISIC 2019, ISIC 2020, PDF-UFES, and DERMQUEST-DERMIS Dataset**	**Acc = 90%, AUC-ROC = 90.25%**
Proposed Hybrid Net Dense Fusion Model	**2024**	**ISIC 2019, ISIC 2020, PDF-UFES, and DERMQUEST-DERMIS Dataset**	**Acc = 92%, AUC-ROC = 92.15%**

categories in this research are BKL (0.94), BCC (0.90), NV (0.87), and MEL (0.97). This level of efficiency is unprecedented among the models considered throughout this research.

In our comparisons, we found that standard CNN architectures such as ResNet-50 and VGG-16 produced AUC-ROC scores around 0.90, highlighting the shortcomings of these conventional models in capturing the complexities of skin cancer classification. The adoption of custom CNN models, specifically ALC-VGG16 and DFE-ResNet50, effectively mitigates these limitations. These architectures are designed to enhance feature extraction through adaptive configurations and dermatological attention mechanisms, respectively, leading to more robust and accurate classifications.

Several models initially considered, such as VGG-16, VGG-22, ResNet-50 and ResNet-101, were ultimately excluded from this study due to issues like the vanishing gradient problem and subpar performance. VGG-22 and VGG-16 faced challenges related to hyperparameter tuning, which could lead to overfitting and biased results. In contrast, the custom architectures implemented in our research successfully address these concerns, offering a viable solution to enhance performance. We did not select vgg-22 it achieve aur-roc and accuracy less then vgg-16. However We did not select resnt-101 because it achieve about same accuracy and auc-roc Secondly we want our system little less complex so we left out vgg-22 and resnet-101

This research opens avenues for further experimentation and investigation. Future researchers should conduct comparative analyses of additional CNN architectures using the same methodology to evaluate the feasibility and versatility of this approach. Additionally, addressing external validity will be crucial for assessing the generalization capabilities of the models with unseen data. While our study focused on four specific categories of skin cancer lesions, expanding this research to encompass a broader range of categories could yield significant insights. Finally, the practical effectiveness of our methodology warrants testing in real-time medical settings, where its potential impact could be substantial.

REFERENCES

[1] M. S. Akter, H. Shahriar, S. Sneha, and A. Cuzzocrea, "Multi-class skin cancer classification architecture based on deep convolutional neural network," in *2022 IEEE International Conference on Big Data (Big Data)*. IEEE, 2022, pp. 5404–5413.

[2] V. Kaya and İ. Akgül, "Classification of skin cancer using vggnet model structures," *Gümüşhane Üniversitesi Fen Bilimleri Dergisi*, vol. 13, no. 1, pp. 190–198, 2022.

[3] A. Bassel, A. B. Abdulkareem, Z. A. A. Alyasseri, N. S. Sani, and H. J. Mohammed, "Automatic malignant and benign skin cancer classification using a hybrid deep learning approach," *Diagnostics*, vol. 12, no. 10, p. 2472, 2022.

[4] S. K. Datta, M. A. Shaikh, S. N. Srihari, and M. Gao, "Soft attention improves skin cancer classification performance," in *Interpretability of Machine Intelligence in Medical Image Computing, and Topological Data Analysis and Its Applications for Medical Data: 4th International Workshop, iMIMIC 2021, and 1st International Workshop, TDA4MedicalData*

2021, Held in Conjunction with MICCAI 2021, Strasbourg, France, September 27, 2021, Proceedings 4. Springer, 2021, pp. 13–23.

[5] W. Gouda, N. U. Sama, G. Al-Waakid, M. Humayun, and N. Z. Jhanjhi, "Detection of skin cancer based on skin lesion images using deep learning," in *Healthcare*, vol. 10, no. 7. MDPI, 2022, p. 1183.

[6] S. Jain, U. Singhania, B. Tripathy, E. A. Nasr, M. K. Aboudaif, and A. K. Kamrani, "Deep learning-based transfer learning for classification of skin cancer," *Sensors*, vol. 21, no. 23, p. 8142, 2021.

[7] M. A. Kassem, K. M. Hosny, and M. M. Fouad, "Skin lesions classification into eight classes for isic 2019 using deep convolutional neural network and transfer learning," *IEEE Access*, vol. 8, pp. 114 822–114 832, 2020.

[8] Z. Li, Z. Chen, X. Che, Y. Wu, D. Huang, H. Ma, and Y. Dong, "A classification method for multi-class skin damage images combining quantum computing and inception-resnet-v1," *Frontiers in Physics*, vol. 10, p. 1046314, 2022.

[9] S. B. Mukadam and H. Y. Patil, "Skin cancer classification framework using enhanced super resolution generative adversarial network and custom convolutional neural network," *Applied Sciences*, vol. 13, no. 2, p. 1210, 2023.

[10] K. Nakai, Y.-W. Chen, and X.-H. Han, "Enhanced deep bottleneck transformer model for skin lesion classification," *Biomedical Signal Processing and Control*, vol. 78, p. 103997, 2022.

[11] J. SM, M. P, C. Aravindan, and R. Appavu, "Classification of skin cancer from dermoscopic images using deep neural network architectures," *Multimedia Tools and Applications*, vol. 82, no. 10, pp. 15 763–15 778, 2023.

[12] A. Al-Rasheed, A. Ksibi, M. Ayadi, A. I. Alzahrani, and M. Mamun Elahi, "An ensemble of transfer learning models for the prediction of skin lesions with conditional generative adversarial networks," *Contrast Media & Molecular Imaging*, vol. 2023, pp. 1–15, 2023.

[13] N. Rezaoana, M. S. Hossain, and K. Andersson, "Detection and classification of skin cancer by using a parallel cnn model," in *2020 IEEE International Women in Engineering (WIE) Conference on Electrical and Computer Engineering (WIECON-ECE)*. IEEE, 2020, pp. 380–386.

[14] M. R. Thanka, E. B. Edwin, V. Ebenezer, K. M. Sagayam, B. J. Reddy, H. Günerhan, and H. Emadifar, "A hybrid approach for melanoma classification using ensemble machine learning techniques with deep transfer learning," *Computer Methods and Programs in Biomedicine Update*, vol. 3, p. 100103, 2023.

[15] G. Yang, S. Luo, and P. Greer, "A novel vision transformer model for skin cancer classification," *Neural Processing Letters*, vol. 55, no. 7, pp. 9335–9351, 2023.

[16] Z. Zhao, "Skin cancer classification based on convolutional neural networks and vision transformers," in *Journal of Physics: Conference Series*, vol. 2405, no. 1. IOP Publishing, 2022, p. 012037.

[17] M. Kahia, A. Echtioui, F. Kallel, and A. B. Hamida, "Skin cancer classification using deep learning models." in *ICAART (1)*, 2022, pp. 554–559.

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

[19] S. Medhat, H. Abdel-Galil, A. E. Aboutabl, and H. Saleh, "Skin cancer diagnosis using convolutional neural networks for smartphone images: A comparative study," *Journal of Radiation Research and Applied Sciences*, vol. 15, no. 1, pp. 262–267, 2022.

[20] P. Banasode, M. Patil, and N. Ammanagi, "A melanoma skin cancer detection using machine learning technique: support vector machine," in *IOP Conference Series: Materials Science and Engineering*, vol. 1065, no. 1. IOP Publishing, 2021, p. 012039.

[21] M. Hashim, A. M. Khattak, and I. Taj, "Efficient detection of skin cancer using deep learning techniques and a comparative analysis study," in *International Conference on Computer Science and its Applications and the International Conference on Ubiquitous Information Technologies and Applications*. Springer, 2022, pp. 203–210.

[22] H. Chen, Y. Wang, T. Guo, C. Xu, Y. Deng, Z. Liu, S. Ma, C. Xu, C. Xu, and W. Gao, "Pre-trained image processing transformer," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 299–12 310.

[23] P. Zhang, X. Dai, J. Yang, B. Xiao, L. Yuan, L. Zhang, and J. Gao, "Multi-scale vision longformer: A new vision transformer for high-resolution image encoding," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 2998–3008.

[24] J. Chen, Y. He, E. C. Frey, Y. Li, and Y. Du, "Vit-v-net: Vision transformer for unsupervised volumetric medical image registration," *arXiv preprint arXiv:2104.06468*, 2021.

[25] I. Ali, M. Muzammil, I. U. Haq, M. Amir, and S. Abdullah, "Deep feature selection and decision level fusion for lungs nodule classification," *IEEE Access*, vol. 9, pp. 18 962–18 973, 2021.

[26] D. Zhou, B. Kang, X. Jin, L. Yang, X. Lian, Z. Jiang, Q. Hou, and J. Feng, "Deepvit: Towards deeper vision transformer," *arXiv preprint arXiv:2103.11886*, 2021.

[27] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.

[28] S. H. Lee, S. Lee, and B. C. Song, "Vision transformer for small-size datasets," *arXiv preprint arXiv:2112.13492*, 2021.

[29] Y. Wang, R. Huang, S. Song, Z. Huang, and G. Huang, "Not all images are worth 16x16 words: Dynamic transformers for efficient image recognition," *Advances in neural information processing systems*, vol. 34, pp. 11 960–11 973, 2021.

[30] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 568–578.

[31] X. Cheng, L. Tan, and F. Ming, "Feature fusion based on convolutional neural network for breast cancer auxiliary diagnosis," *Mathematical Problems in Engineering*, vol. 2021, pp. 1–10, 2021.

SYED NEHAL HASSAN SHAH received the M.S. degree in Artificial Intelligence from Air University, Islamabad, Pakistan, in 2024. Since 2023, he has been working as an AI Developer at Octaloop Technologies. His research interests include deep learning, computer vision, and medical image analysis, with a particular focus on vision transformers.

IMRAN TAJ is currently working as an Assistant Professor at Zayed University. Prior to that he was a Senior Team Lead, Information Systems Branch with BC Public Service, Canada, where he led the digital transformation and automation of an enterprise software application using Artificial Intelligence techniques and Robotic Process Automation tools. He received his Ph.D. degree in computer engineering from University of Paris-Est, France. He has several years of professional experience of applying machine learning operations to craft real world data solutions and has been involved in all phases of Software Development Life Cycle - from inception to implementation - for several systems engineering projects in the fields of Artificial Intelligence, Machine Learning and Data Sciences.

SYED MUHAMMAD USMAN (Member, IEEE) is working as Sr. Assistant Professor in Bahria University, Islamabad, Pakistan. He completed PhD in Computer Engineering from Bahria University, Islamabad and MS in Computer Engineering from NUST, Islamabad, Pakistan. He has teaching and research experience of more than eight years with research interests in biomedical signal processing, medical imaging and precision agriculture. He has more than 20 publications in international journals and peer reviewed conferences.

SYED ABDULLAH SHAH is a dedicated student poised to complete his Master's in Artificial Intelligence from Air University, Islamabad, Pakistan. With a robust academic foundation, he holds a Bachelor's degree from Riphah International University, I-14 Islamabad. He a Pakistani citizen hailing from Punjab born in 25th January 2000. Mr. Abdullah has demonstrated his scholarly aptitude through his involvement in two recent to be be published research papers. Currently focused on his studies, Syed continues to explore the vast potential of artificial intelligence, aiming to make impactful advancements in this rapidly evolving domain.

ALI SHARIQ IMRAN (Member, IEEE) received a master's degree in software engineering and computing from the National University of Science and Technology (NUST), Pakistan, in 2008 and a Ph.D. in computer science from the University of Oslo (UiO), Norway, in 2013. He is associated as an Associate Professor with the Department of Computer Science (IDI), Norwegian University of Science and Technology (NTNU), Norway. With over 15 years of teaching and research experience, he devised innovative ways to design effective multimedia learning objects and integrate the teaching-research nexus frameworks at the graduate level. He served as a commission member of the Ministry of Education of Macedonia in setting up Mother Theresa University in Skjope. He leads a capacity-building project called CONNECT (https://norpart-connect.com) funded by the Higher Education Commission of Norway, DIKU, under the NORPART scheme as a coordinator and three Erasmus+ KA2 projects (PhDICTKES (https://phdictkes.eu), RAPID, and TKAEDiT) as a project manager at NTNU, along with an Excited mini-project funded by NTNU. Dr. Ali also leads a research group on Deep NLP (http://deep-nlp.net) and specializes in applied deep learning research to address various multi-modality media analysis application areas for audio-visual and text processing. He has co-authored over 100 peer-reviewed journals and conference publications and has served as an editor and reviewer for many reputed journals. He is a member of the Intelligent Systems and Analytics research group at NTNU and an IEEE/ACM Member.

SHEHZAD KHALID is a passionate academician, researcher and research management professional with 22+ years of experience and a proven track record of delivering high-quality results in the field of Machine Learning, Computer vision, Signal Processing, Natural Language Processing, Intelligent Medical Diagnostics etc. Possesses a strong background in executive management of Postgraduate Programs, Research Innovation Commercialization (RIC) and Entrepreneurial Ecosystem in academic organization with focus on conducive policy development and execution. Led RIC ecosystem and Incubation center at Bahria university, which has been recognized amongst top ranked ecosystems in Pakistan. Proficient in leveraging advanced research techniques in different applied domains as reflected by track record of around 100 high quality journal publications, 50 conference publications, tens of PhD/MPhil supervisions and successful completion of Rs. 65 million+ worth of research and entrepreneurial projects.