






A Novel Patch-Based Ensembling Approach with Perceptual Attention for Skin Lesion Classification

Tapan Kumar Nayak¹, Annavarapu Chandra Sekhara Rao^{1*}, Soumya Ranjan Nayak²

¹ Department of CSE, Indian Institute of Technology (ISM), Sardar Patel Nagar, Dhanbad 826004, India

² School of Computer Engineering, KIIT Deemed to be University, Bhubaneswar 751024, India

Corresponding Author Email: nayak.soumya17@gmail.com

Copyright: ©2024 The authors. This article is published by IIETA and is licensed under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>).

<https://doi.org/10.18280/ts.410502>

ABSTRACT

Received: 28 October 2023

Revised: 2 April 2024

Accepted: 28 August 2024

Available online: 31 October 2024

Keywords:

image patch, feature ensembling, attention, melanoma, nevus

Most of the time biopsy has been the gold standard for skin lesion evaluation. However, specialists evaluate signs and symptoms for the final decision. Shortage of specialist definitely adds the adverse effect on effective and early detection. Recently, CNN has extended the helping hand for the specialist during the final decision. Also, many pre-trained CNN models have been designed to be used as transfer learning. But, a common approach of random resizing of input images are required before training to get fit to the input layers. This is because the approximate size of most of the available skin lesion images and pre-trained models are of 1000×1000 and 224×224 respectively. Hence the required resizing though solves one problem of size mismatch, it may eliminate principal feature for classification leading to poor accuracy. Hence, in this work, we propose a novel patch-based ensembling approach for the early diagnosis of melanoma and nevus skin lesions. Here the effect of applying patches over classification has been studied on an incremental basis. In the ensembling approach, the resultant features from different patches have been combined for further processing with perceptual attention to maintain the spatial relationship. The proposed model was evaluated on a set of 748 dermoscopy images collected from the ISIC 2017 data set (374 melanoma and 374 nevus images). Our result demonstrates that using image patches as input improves accuracy instead of image scaling. The proposed model performed well enough to serve as a baseline for further studies of skin lesion classification.

1. INTRODUCTION

Nowadays various lifestyle factors, including smoking, drinking alcohol, odd eating habits, physical activity levels, changes in the environment, sun exposure, radiation exposure, viral infections etc. are putting everyone at a potential risk of cancer. Among those, the escalation in skin cancer cases with swelling of cells is a deep concern [1]. Here in this paper, the two cases of skin cancer such as melanoma and nevus are taken for the experimentation. Because of the similarities in structure and symptoms these two types are still challenging task for classification. The early diagnosis is curable and can save the life. The faulty and late detection leads the cancer spreads to adjacent organs and making it more fatal. Melanoma is the most prevalent type of skin condition which affects the cells on skin surface called melanocytes. The proportion of melanoma makes the skin color ranges from fair to black [2]. Melanoma usually found in dark or darker colors, though uncommonly it is developed in skin colors like red, purple, pink, white, or blue [3]. This type of cancer is especially worrisome due to its proclivity for metastasis, or the potential to spread. Melanoma can occur in any part of the human body, while it most commonly develops on the back of legs [4]. In accordance with the research, if skin cancer is discovered at an early stage, the fatality rate can be decreased

by up to 90%. As a result, early detection and classifications are crucial [3, 4]. Many papers have been written about identifying, segmenting, and categorizing skin cancers using various computer vision, image processing, machine learning, and convolutional neural network (CNN). Esteva et al. [5] used a CNN to classify skin cancer, where it has the potential to perform overall and varying tasks across a wide range of fine-grained categories. Iyatomi et al. [6] introduced a semi-automatic system for classifying melanomas, where they used dermoscopic structures like parallel ridges, parallel furrows, and fibrillar patterns as pattern detectors. While Anas et al. [7] created the melanoma classification using four types of categorizations, Almansour et al. [8] illustrated a melanoma classification method using k-means clustering and Support Vector Machine (SVM). Capdehourat et al. [9] used AdaBoost MC to create separate image classification methods for melanomas, cancers, and dermoscopy. Giotis et al. [10] and Ruiz et al. [11] created decision support systems for melanoma using pre-processing technique and neural network algorithm based on lesion texture, color, visual diagnostic features, surrounding tissue, and extent of the damage. CNN as a sub-field of deep neural networks has lately achieved significant progress in computer vision. Since AlexNet's victory in the ImageNet Challenge: ILSVRC 2012 [12], CNN has become a widely used technique in computer vision. In order to attain

greater accuracy, the current sentiment has been to build profound and more complicated networks [13, 14]. As the feature extraction backbone, MobileNetV2 [15] introduces an improved module with a reversed residual structure and non-linearities in narrow layers to achieve state-of-the-art performance for feature extraction and semantic segmentation. Its low parameters in comparison to other pre-trained models like VGG, AlexNet, ResNet, and EfficientNet make it a preferred choice to use it as a feature extractor.

Ensembling is preferred when assembling models with similar architectures because of its reliability and stability. Among different techniques of ensembling, bootstrapping or bagging has been the right choice with different input data or patches of the image. Recent work [16] demonstrates that ensembles can be effectively constructed without a sequential rise in the computational work. While examining different pre-trained models like AlexNet, VGG16, ResNet50, and MobileNet, we found that most of them accepts input image size approximately 224×224 , whereas medical images are available in a variety of sizes, including 1024×1024 . It leads to resizing as a common technique in various pre-processing approach. This eventually ends up eliminating some of the pixels which might have a significant contribution. CNN finds it more difficult to learn the features required for classification or detection as the number of pixels with prominent factors gets significantly reduced when the large image is down-scaled. In a multi-patch-based ensemble model, the spatial relationship among patches must be preserved. Here comes the attention mechanism. Attention modeling has been adopted in computer vision because of the successful implementation in Natural Language Processing (NLP), where different versions of transformer attention are adapted to recognition tasks including object detection and semantic segmentation [17, 18]. Deep learning with attention has changed our approach to designing deep models. It has been a feature retention mechanism that impacts a lot in the field of image classification [19].

In this paper, we illustrate how to build multi-patch ensemble CNN architectures using MobileNetV2 as the base classifier and spatial attention to reconstructing the intermediate feature matrix for the classification of Melanoma and Nevus skin disorder. Here we have taken the skin images as a whole, two patches and four patches, and evaluated the model accuracy. MobileNetV2 is used as a feature extractor from different patches and the spatial attention module maintains the spatial relationship among the patches. The specific contribution of this paper is as follows:

- In order to generate the intermediate feature matrix, we proposed a modified ensembling algorithm.
- To reinstate the spatial relationship, Image Spatial Attention has also been introduced.
- We designed a deep CNN model taking into account the bag of image patches, ensemble MobileNetV2 with spatial attention for classification of melanoma and nevus.
- Our model demonstrates that incremental patches with intact spatial relationships result in higher classification accuracy.

2. RELATED WORK

Extensive research, exploration and development of fresh methods in the relevant field results the precise diagnosis of

skin cancer. Here the associated tasks may be divided based on the method used to classify melanoma and nevus skin cancer. These are multiple patches of an image, modified bagging or bootstrapping, attention mechanism, and classification.

In earlier studies [20] to train deep networks with fine-grained details, images are randomly cropped or resized with small patches, while all the patches are labelled same as of original image. Such an approach results in ambiguity in training examples because of aesthetic, tone, and feature attributes might not accurately depict the details in the entire image. Recent studies have focused on this issue, leading to improved classification precision. One of the most well-known techniques for building ensembles that take into account data variations is bagging [21], while another is AdaBoost [22]. A large number of models are generated by splitting the training data into small subsets, where each subset is used to train a model. Using ensembles, Khatami et al. [23] demonstrated how to significantly improve and achieve cutting-edge outcomes in the field of medical image retrieval, which is plagued by severely unbalanced datasets. A customized bagging algorithm on MobileNetV2 [15] was used in the proposed architecture to produce an intermediate feature matrix. Huang et al. [24] demonstrated that by converging at multiple local minima and preserving the model parameters along its optimization path results in decreased computational cost for ensembling. Hara et al. [25] proposed that regularization techniques such as dropout can be categorized as ensembling methods. They demonstrated that the accuracy of a model can be enhanced by averaging across a network. Huang et al. [26] suggested employing stochastic depth as an ensembling strategy by using average findings over networks. Singh et al. [27] proposed ensembling technique swap out by combining dropout and stochastic depth methods. All these approaches require extensive domain knowledge to construct the initial base models to be ensemble. Recent research has addressed this need by modifying the ensembling to improve feature selection.

The attention of CNN model has been used as a working solution for a diverse range of computer vision problems [28]. The focus of CNNs for the task of classifying images reflects the relevant features which rely on as proof [29]. Due to the multi-patch ensembling of the feature matrix, the spatial relation between patches might be compromised. One simple solution to this issue is to learn precise spatial attention that is comparable to saliency detection [30] and semantic segmentation [31], but it is a bit noisy and different because the label-relevant regions are not well defined. The attention map might also be improved [32] by making it focus on a single compact area instead of several fragmented regions. However, this doesn't work for multiple patches with labels. Hence, by enacting feature matrix-based entire image attention, we suggest an indirect method in this paper to concentrate on spatial attention of patches.

3. PROPOSED METHODOLOGY

In this section, we have covered the elements of our suggested methodology in Figure 1. The methodology shown is a combination of all three models i.e. model for the whole patch, double patches, and four patches. The methodology includes preparing the data which is multi-patch RGB skin images, discussed in subsection 3.1, a modified ensembling for

the intermediate feature matrix, discussed in subsection 3.2, perceptual attention to the images to keep the spatial relationship intact, discussed in subsection 3.3 and the training strategy in subsection 3.4.

3.1 Multi-patch RGB skin images

Medical skin images like melanoma and nevus are high-resolution images of size approx. 1000×1000 or above. This size is much more than the average size used by the pre-trained models. With random resizing, we may lose valuable

information, so instead of random resizing, these images are divided vertically and horizontally as in Figure 2 to create smaller image patches. The first model uses the whole image for training. The second model uses two patches for the intermediate feature extraction and an attention module for keeping the spatial information intact. The third model uses four patches for the intermediate feature extraction and the attention module. Each patch of images is input to the MobileNetV2 to generate the feature matrix, those get combined to generate the final intermediate feature matrix.

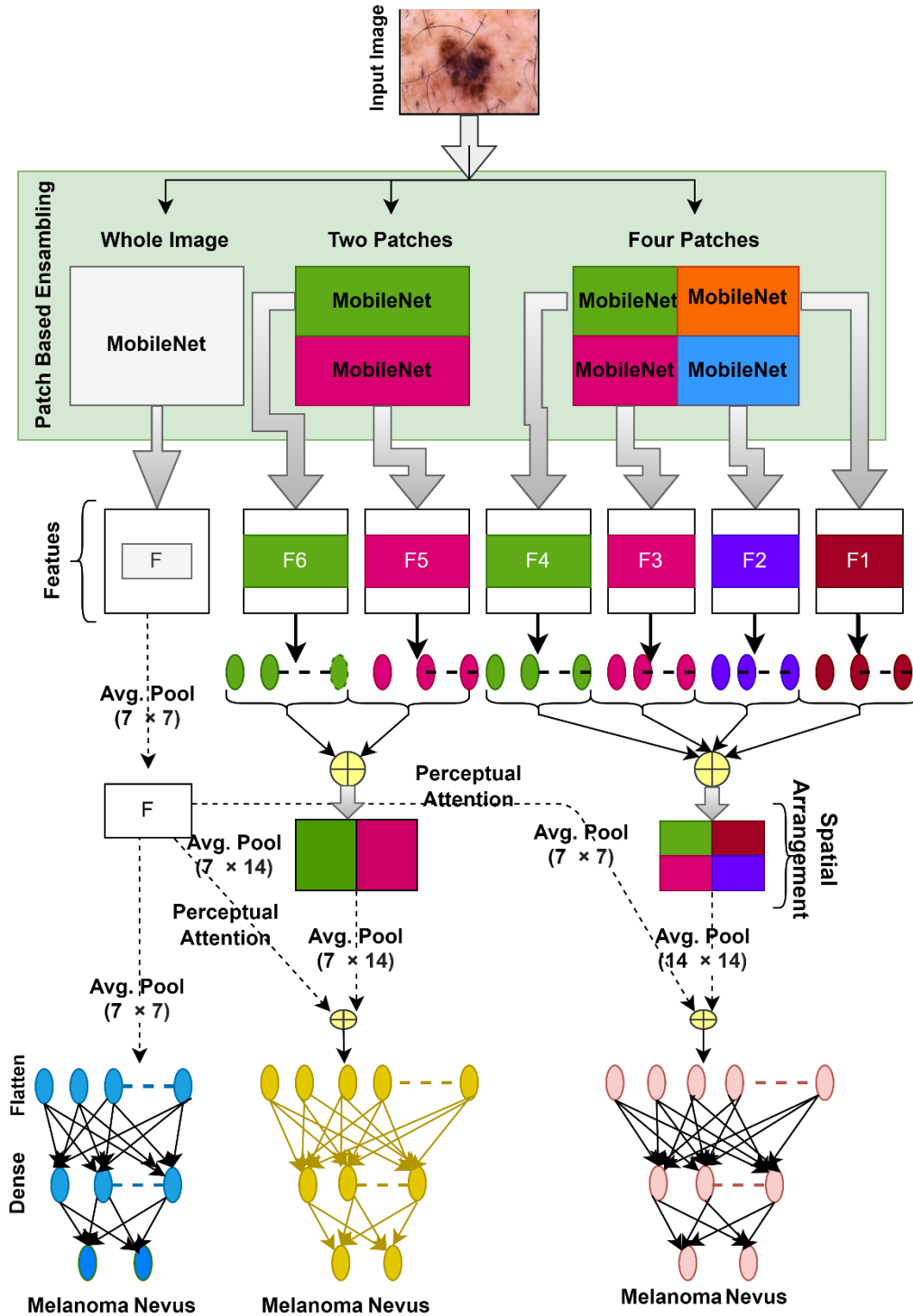


Figure 1. Proposed methodology

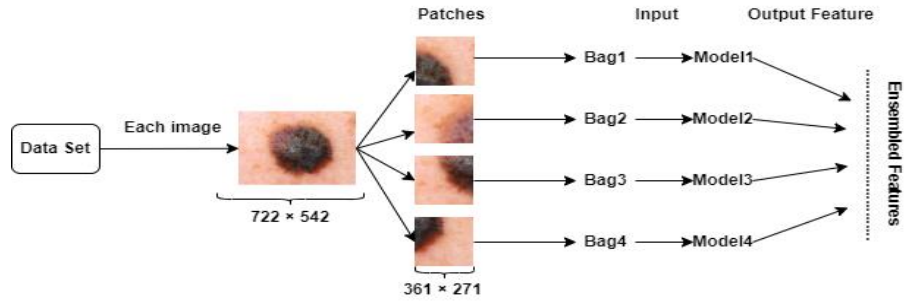


Figure 2. Multi-patches of each image from the data set; each patch moves to each bag of the ensemble model

Algorithm 1: Modified Ensemble Technique

```

1. Input: No. of patches ( $N$ ), MobileNetV2
2. Output: Intermediate Feature Matrix
3. Data: Melanoma and Nevus skin Images ( $img$ )
4.  $n = \lfloor \sqrt[2]{N} \rfloor$  where  $N$  takes values 2, 4
5. Function EnhancedEnsembling ( $img, N, n, MobileNetV2$ ):
6.    $labels, data \leftarrow PrepareData(img, N)$ 
7.    $train, test \leftarrow PrepareInput(N, data_i, labels)$ 
8.   while  $i \leq N$  do
9.      $baseModel_i \leftarrow MobileNetV2()$ 
10.    For  $layer$  in  $baseModel_i.layers$  do
11.       $layer.name \leftarrow layer.name + str(layer)$ 
12.       $baseModel_i \leftarrow baseModel_i.layers[-3].layers$ 
13.    end
14.  end
15.   $IntermediateFeatureMatrix = []$ 
16.   $j = 1$ 
17.  for  $I = 1$  to  $n$  do
18.     $featureMatrix, j \leftarrow Combine(baseModel_j, train_j, n, j)$ 
19.     $j = j + 1$ 
20.     $IntermediateFeatureMatrix \leftarrow VerticalCombine(IntermediateFeatureMatrix, featureMatrix)$ 
21.  end
22.  return  $IntermediateFeatureMatrix$ 
23. Function Combine ( $baseModel_j, train_j, n, k$ ):
24.    $featureH = []$ 
25.    $j = k$ 
26.   for  $i = 1$  to  $n$  do
27.      $feature_i \leftarrow baseModel_j(train_j)$ 
27.      $j = j + 1$ 
27.      $featureH = HorizontalCombine(featureH, feature_i)$ 
28.   end
29.   return  $feature.j$ 
30. Function PrepareData ( $img, N$ ):
31.    $data = []$ 
32.   while  $i \leq N$  do
33.      $data_i = []$ 
34.   end
34.   while  $i \leq N$  do
35.      $image_i \leftarrow verticalSplit(img, n)$ 
36.   end
37.   while  $j \leq N$  do
38.      $data_i \leftarrow data_i + horizontalSplit(image_i, N)$ 
39.   end
40.    $labels = labels$ 
41.   return  $labels, data_i$ 
42. Function PrepareInput ( $N, data_i, labels$ ):
43.    $trainX, testX, trainY, testY \leftarrow trainTestSplit(data_i, labels, testSize)$ 
44.   while  $i \leq N$  do
45.      $trainX_i, testX_i, trainY_i, testY_i \leftarrow trainTestSplit(data_i, labels, testSize, shuffle = false)$ 
46.   end
47.   return  $trainX_i, testX_i, trainY_i, testY_i$ 

```

3.2 Enhanced ensembling

The main goal of the ensemble method is to integrate the output from various model's predictions to improve classification accuracy [33, 34]. The network model predictions, variance and bias may be decreased by the ensembling technique. Unlike traditional ensembling techniques which are based on statistical aggregation of prediction results from different models, our approach focuses on aggregation of intermediate features from different patched input resulting single prediction. This technique helps to get rid of possibilities of biased voting during traditional ensembling. In order to accurately classify skin images, the feature matrix resulting from MobileNetV2 after freezing the top three layers has been assembled according to the suggested approach. The proposed "Modified Ensemble" approach is shown in Algorithm 1. Here instead of dividing the entire data set into no. of bags, the bags contain patches of different images. The pre-trained model MobileNetV2 is chosen as it has the lowest ever parameters than other pre-trained models as per the Table 1. As we are reconstructing the intermediate feature matrix, we have kept all the models the same as MobileNetV2 to get a similar size matrix. During the model execution, the total N+1 number of MobileNetV2(freeze top

three layers) has been saved, where N is the total number of patches. Here N models take N sets of image patches to generate the feature matrix. All the feature matrices are added together as per the algorithm to generate the intermediate feature matrix as shown in Figure 1.

3.3 Perceptual attention

Attention is a technique that attempts to emphasize significant information while diminishing the irrelevant details. As per our proposed approach, we have made multiple patches of a single image to extract more relevant features, and then we ensemble all the intermediate features to generate the intermediate feature matrix. Here there might be a chance that the edge pixels may lose the spatial relationship with the edge pixel of other feature matrices during the ensembling. So, to preserve it we have introduced perceptual attention as in Figure 3, by mapping the feature matrix generated from the complete image with the intermediate feature matrix. The perceptual attention is an area specific attention mechanism inspired from Natural Language Processing (NLP) where it focuses on different words or phrases in the input text by enabling more accurate and relevant output.

Table 1. Network specifications for various pre-trained deep neural networks

Networks	Size (in MB)	Depth of Layers	Image Size	Parameters (in Millions)
Xception [35]	88	81	299×299×3	22.9
VGG16 [13]	528	16	224×224×3	138.4
ResNet101 [36]	171	101	224×224×3	44.7
InceptionV3 [14]	92	189	299×299×3	23.9
DenseNet121 [37]	33	242	224×224×3	8.1
EfficientNetB0 [38]	29	132	224×224×3	5.3
AlexNet [39]	53	8	227×227×3	61
GoogleNet [40]	96	22	224×224×3	7
MobileNetV2 [15]	16	53	224×224×3	3.5

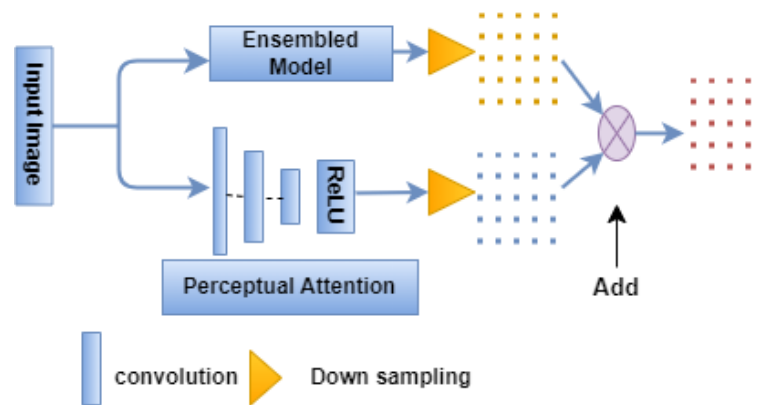


Figure 3. Perceptual attention to ensure the spatial relationship among the pixels

Table 2. Intermediate feature matrix (four patch model) generated by combining the feature matrix generated by base models with different patches where P^1, P^2, P^3, P^4 are patches from ensembled model

$P^1_{1,1}$	$P^1_{1,2}$	$P^1_{1,N/2}$	$P^2_{1,N/2+1}$	$P^2_{1,N/2+2}$	$P^2_{1,N}$
$P^1_{2,1}$	$P^1_{2,2}$	$P^1_{2,N/2}$	$P^2_{2,N/2+1}$	$P^2_{2,N/2+2}$	$P^2_{2,N}$
.
$P^1_{M/2,1}$	$P^1_{M/2,2}$	$P^1_{M/2,N/2}$	$P^2_{M/2,N/2+1}$	$P^2_{M/2,N/2+2}$	$P^2_{M/2,N}$
$P^3_{M/2+1,1}$	$P^3_{M/2+1,2}$	$P^3_{M/2+1,N/2}$	$P^4_{M/2+1,N/2+1}$	$P^4_{M/2+1,N/2+2}$	$P^4_{M/2+1,N}$
$P^3_{M/2+2,1}$	$P^3_{M/2+2,2}$	$P^3_{M/2+2,N/2}$	$P^4_{M/2+2,N/2+1}$	$P^4_{M/2+2,N/2+2}$	$P^4_{M/2+2,N}$
.
$P^3_{M,1}$	$P^3_{M,2}$	$P^3_{M,N/2}$	$P^4_{M,N/2+1}$	$P^4_{M,N/2+2}$	$P^4_{M,N}$

In image processing perceptual attention can be used to enhance spatial relationship among patches. The attention module results feature matrix $F_{i,c}(x)$ is:

$$F_{i,c}(x) = A_{i,c}(x) + I_{i,c}(x) \quad (1)$$

where, i ranges over all spatial positions and $c \in 1, \dots, C$ is the index of the channel. $A_{i,c}(x)$: feature matrix of attention module and $I_{i,c}(x)$: intermediate feature matrix generated from ensembled model. $I_{i,c}(x)$ given in Table 2 is resulted from the enhanced ensembled model.

3.4 Training strategy

The data set contains the skin lesion images for melanoma and nevus, where each of 374 and 374 respectively. Here the model has experimented with a single-image, double-patch, and quad-patch image. The proposed methodology in Figure 1 represents three different models, combined together showing the classification of melanoma and nevus individually. Before designing the ensembled model the base model is chosen out of different pre-trained models like DenseNet121, EfficientNetB0, MobileNetV2, ResNet101, VGG16, VGG19, and Xception. Among all these models as per Table 3 MobileNetV2 shows better accuracy with lesser parameters. After deciding on the base model, the ensembled model has been designed for two patches and four patches with a spatial attention module. The train and validation ratio of image data used is 90:10.

Table 3. Classification accuracy of different pre-trained model with the data set

Networks	Size (in MB)	Parameters (in Million)	Accuracy
DenseNet121	33	8.1	84.85
EfficientNetB0	29	5.3	49.71
Resnet101	171	44.7	54.57
VGG16	528	138.4	77.99
VGG19	549	143.7	74.28
Xception	88	22.9	86.41
MobileNet	16	4.3	82.57
MobileNetV2	14	3.5	88.28

4. DATA DESCRIPTION AND EXPERIMENTAL SETUP

In this section, we present the experimental data description and experimental set up that has been followed to verify the effectiveness of the proposed model. All the experimentations were performed using skin image datasets collected from ISIC2017 data set [41]. The publicly available ISIC 2017 data set has offered three parts of challenges, these are Part.1) Lesion Segmentation Task, Part.2) Dermoscopic Feature classification Task and Part.3) Disease Classification Task. Out of these three tasks in disease classification task participants are offered to classify images belongs to three classes named as “melanoma”, “seborrheic keratosis” and “benign nevi” with 374, 254 and 1372 numbers of training images. In this paper we have considered the data set belongs to the disease classification task. Out of these three classes we have considered only two classes i.e. “melanoma (374 images)” and “nevus (1372 images)”. To make the data set balanced we have considered 374 images from each of the classes for our

experimental purpose. All CNN models were developed using the PyTorch toolbox and all experiments were conducted on the Amazon Web Service platform with NVIDIA Tesla k80 GPU architecture, x86664 CPU architecture, Intel Xenon E5-2686 v4 physical processor, 2.3 GHz clock speed, 4 vCPUs, 15.25(GiB) memory per vCPU and 61 (GiB) memory. Here we have performed the experiment in three different ways. Firstly, we have performed the experiment using the pre-trained models like VGG-16, VGG19, DenseNet121, EfficientNetB0, ResNet101 and MobileNetV2 with the data set. From the experiment we have chosen MobileNetV2 as our base classifier for the next two experiments because of its light weight, fewer parameter and better accuracy. In second and third experiment we have used two patches and four patches of images with an ensembled model with perceptual attention. Performance metrics such as accuracy, loss, sensitivity, specificity, precision and ROC-AUC were considered to evaluate each model.

4.1 Experimental parameters

This subsection provides the parameter set up in Table 4 that has been used for the experiment.

Table 4. Parameter setting for the experiment

Parameters	Settings
pretrained models	DenseNet121, EffcientNetB0, MobilenetV2, ResNet121, VGG16, VGG19, Xception
chosen pretrained model	MobileNetV2
average pool stride	7×7 and 14×14 1
activation	relu and softmax
learning rate	0.0001
drop out	0.5
loss	binary-cross-entropy
epochs	50
images	748
patches	2(dual), 4(quad)
split	90:10

4.2 Performance metrics

The performance of the proposed model has been evaluated using the following metrics.

Accuracy: It measures the proportion of correctly classified instances.

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN} \quad (2)$$

where, TP for true positive, TN for true negative, FP for false positive and FN for false negative instances.

Specificity: It the proportion of true negative instances out of actual negative instances.

$$\text{Specificity} = \frac{TN}{TN+FP} \quad (3)$$

Sensitivity: It is the ratio between predicted true positive instances out of actual positive instances.

$$\text{Sensitivity} = \frac{TP}{TP+FN} \quad (4)$$

Precision: It is the ratio between true positive predictions out of total positive predictions.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (5)$$

ROC (Receiver Operating Characteristic curve): It is a graphical plot of true positive rate (sensitivity) against false positive rate (1-specificity). Area Under the ROC curve (AUC-ROC) is shown to evaluate the performance.

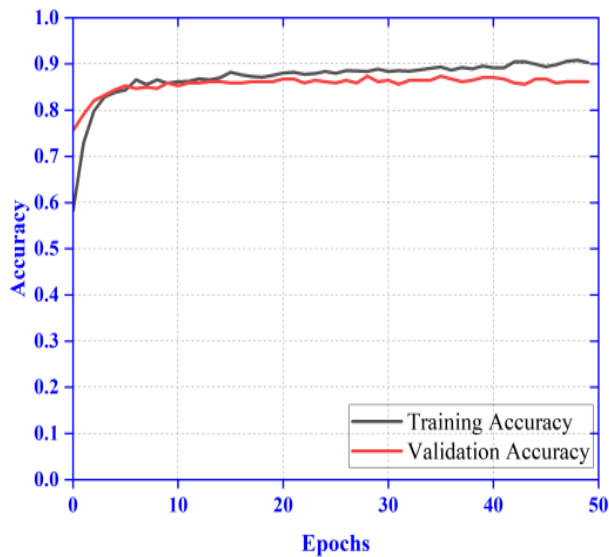
5. RESULTS

In this section, the results are shown in Figures 4-9 and Table 5. Out of these Figure 4, Figure 5 shows the accuracy and loss plot of different pre-trained models. Figures 6-8 shows the accuracy and loss plot of the proposed model with three different image type i.e. whole image as input, two equal halves of image as input (dual patch) and four equal patches of image as input (quad patch). Figure 9 shows the accuracy and loss plot of proposed models. Figure 10 shows the AUC-ROC plot of dual patch model and quad patch model for class:

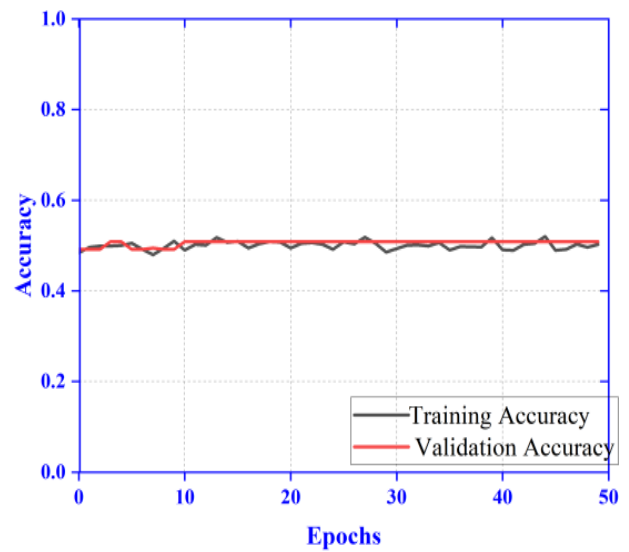
melanoma and class: nevus. Then we set the model to run for two patches showing improved accuracy in Figure 7, which further increases when choosing four patches in Figure 8. The same data set has been used in proposed models shows that instead of random resizing, a large image can be divided to match the input image size at par with the pre-trained model's input size, resulting in better classification accuracy.

5.1 Results of different pretrained models

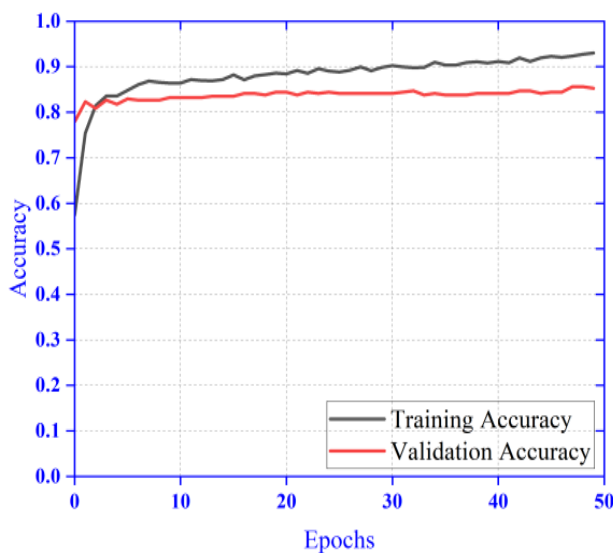
Accuracy and loss with different pre-trained models have been provided here. Figure 4 (a), Figure 5 (a), Figure 4 (b), Figure 5 (b), Figure 4 (c), Figure 5 (c), Figure 4 (d), Figure 5 (d), Figure 4 (e), Figure 5 (e), Figure 4 (f), Figure 5 (f) and Figure 4 (g), Figure 5 (g) describes the accuracy and loss function for different pre-trained models like DenseNet121, EfficientNetB0, MobileNetv2, ResNet101, VGG16, VGG19, and Xception respectively. Considering all these results and the information in Table 1, MobileNetV2 shows promising result as it is light weight with less parameters. Hence it has been chosen as the base pre-trained model for the proposed model.



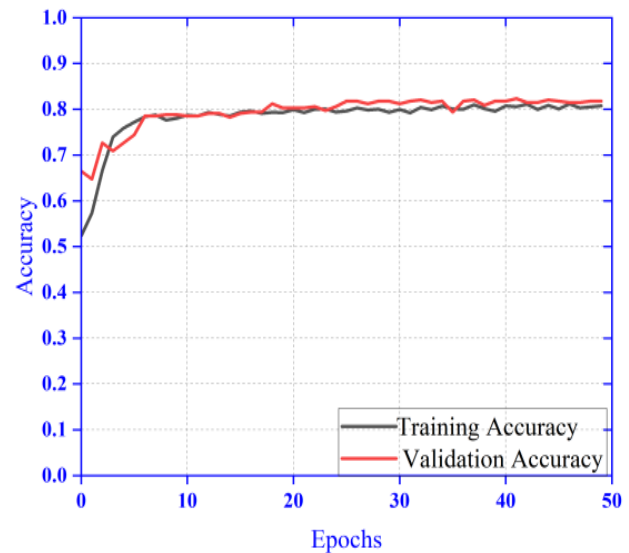
(a) Accuracy with DenseNet121



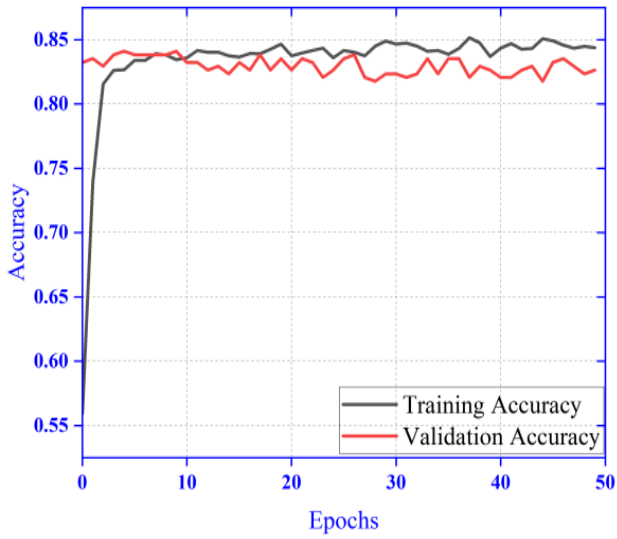
(b) Accuracy with EfficientnetB0



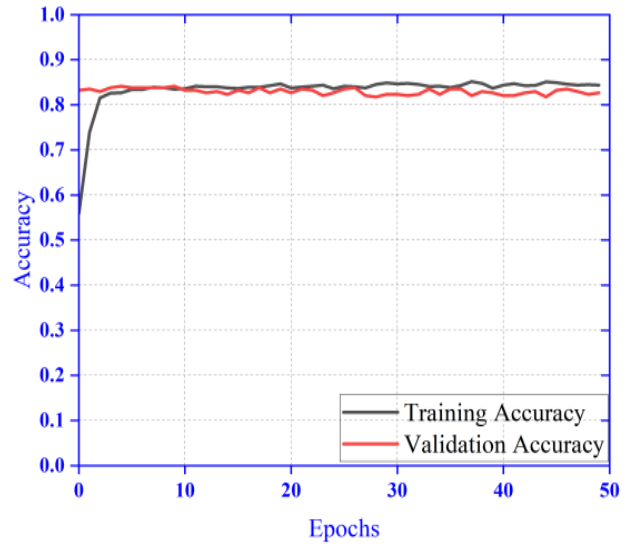
(c) Accuracy with MobileNetV2



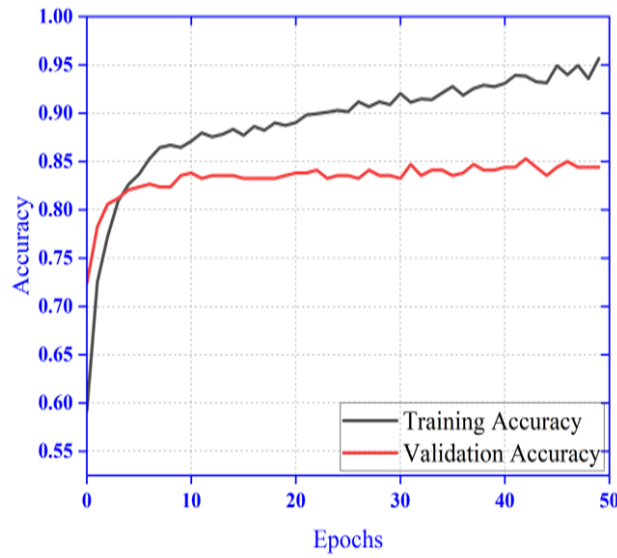
(d) Accuracy with ResNet101



(e) Accuracy with VGG16

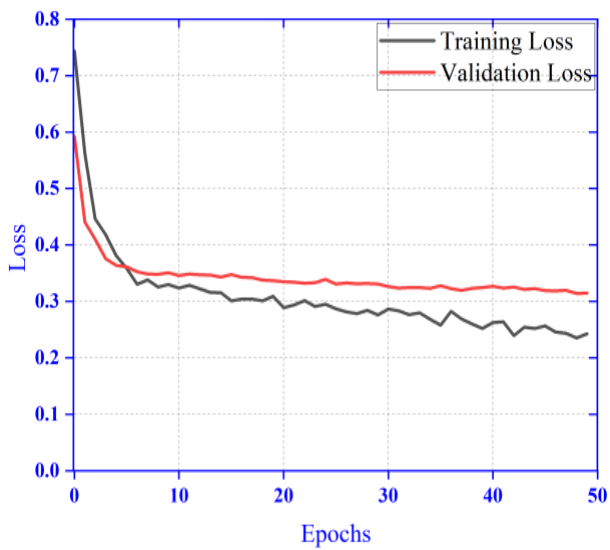


(f) Accuracy with VGG19

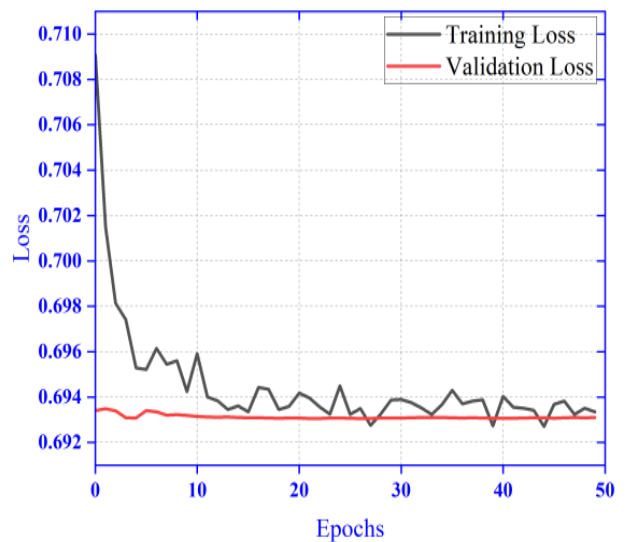


(g) Accuracy with Xception

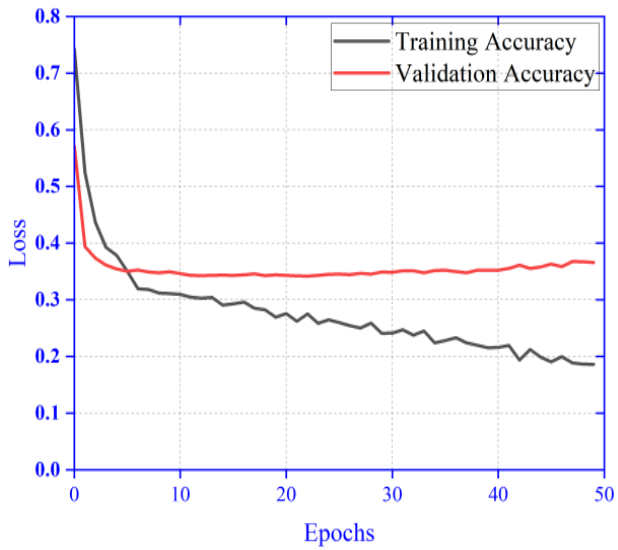
Figure 4. Classification accuracy of different pre-trained models with the data set



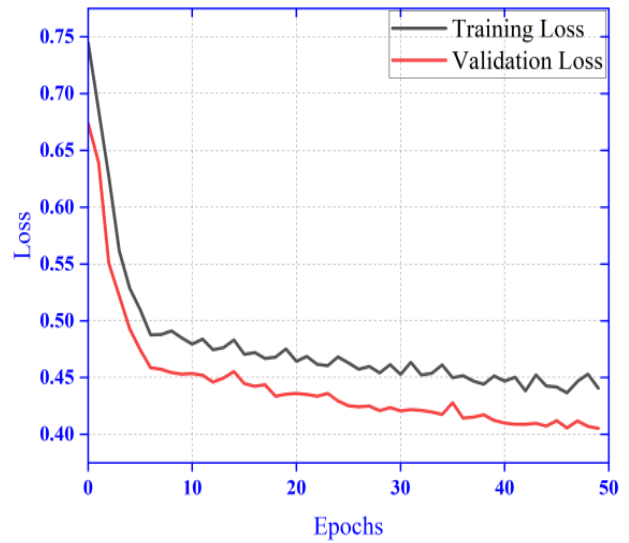
(a) Loss with DenseNet 121



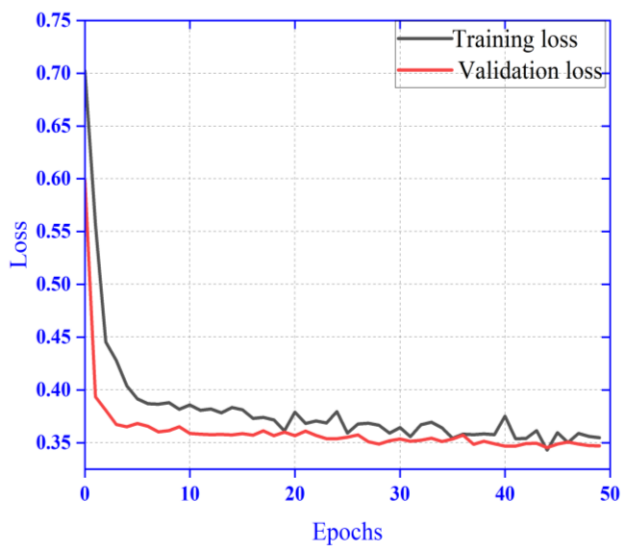
(b) Loss with EfficientNetB0



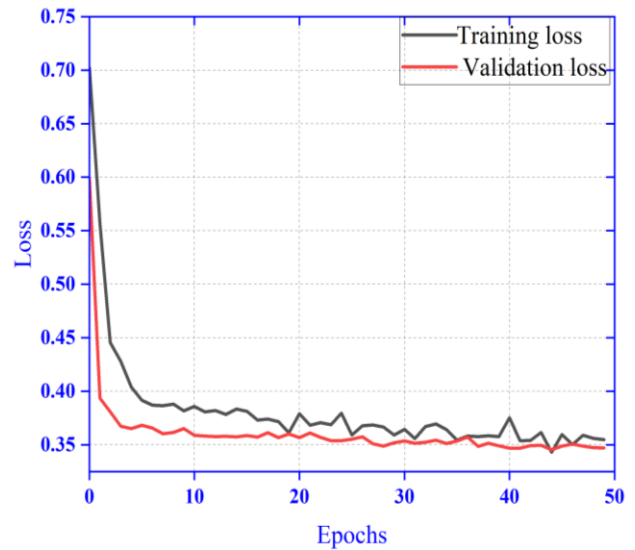
(c) Loss with MobileNetV2



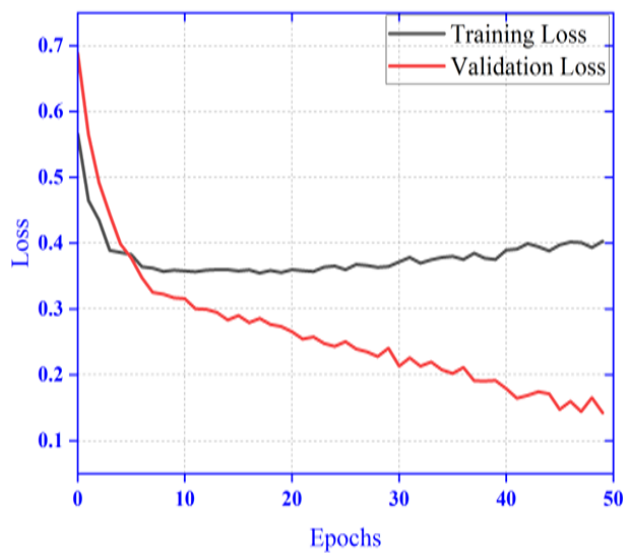
(d) Loss with ResNet101



(e) Loss with VGG16

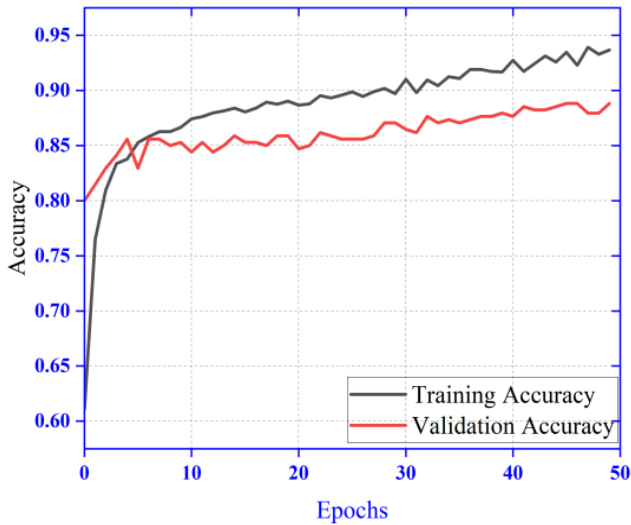


(f) Loss with VGG19

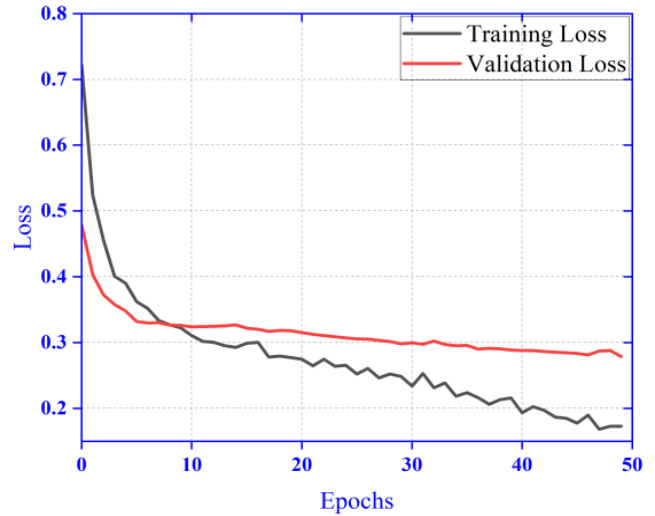


(g) Loss with Xception

Figure 5. Loss with different pre-trained models with the data set

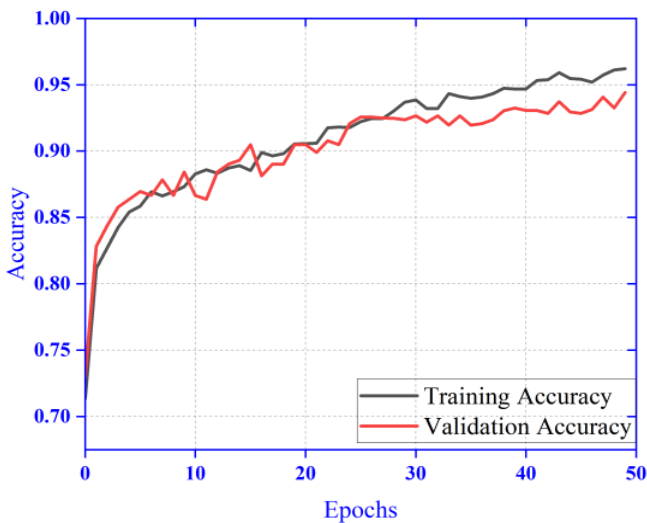


(a) Accuracy with the whole image as input

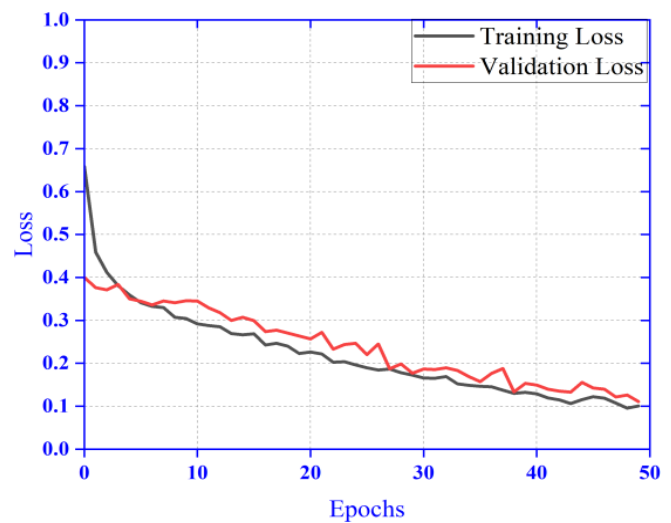


(b) Loss with the whole image as input

Figure 6. Accuracy and loss of the proposed model with the whole image



(a) Accuracy with the dual patch of images



(b) Loss with the dual patch of images

Figure 7. Accuracy and loss of proposed dual patch model

5.2 Results of proposed model taking whole image as input

Here the whole image is used as input with good amount of resizing. The result of accuracy and loss is shown in Figure 6 (a) and Figure 6 (b). The accuracy and loss plot with diverging nature of the graph shows poor performance than the subsequent proposed model.

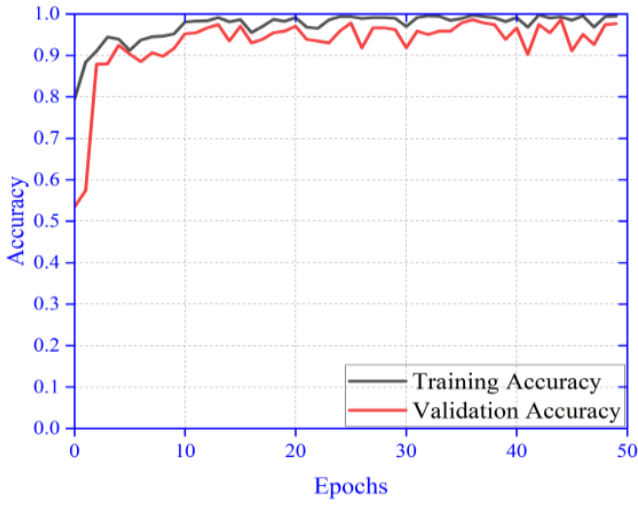
5.3 Results of dual patch model

Here the performance result has been shown in Figure 7 for the proposed ensemble model with dual patch of image as input. Here in this model the amount of input image resizing done is less. The Figure 7 (a) and Figure 7 (b) shows improved result in comparison to the previous experimentation with whole image as input shown in Figure 6 (a) and Figure 6 (b). The AUC-ROC curve for class: melanoma (0.925) and class: nevus (0.889) shown in Figure 10 (a) and Figure 10 (b) is also promising. The results of different performance metrics like

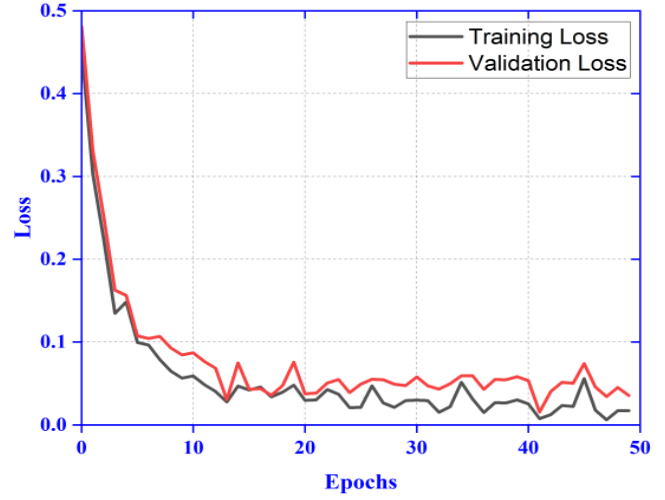
sensitivity, specificity, precision and accuracy are 96.7%, 96.1%, 95.6% and 92.0% respectively.

5.4 Results of quad patch model

Here the performance result has been shown in Figure 8 for the proposed ensemble model with quad patch of image as input. Here in this model very less resizing of input image is done in compared to the previous model. The Figure 8 (a) and Figure 8 (b) shows improved result in comparison to the previous experimentation with dual image patch as input shown in Figure 7 (a) and Figure 7 (b). The AUC-ROC curve for class: melanoma (0.967) and class: nevus (0.927) shown in Figure 10 (a) and Figure 10 (b) is further improved than the previous dual patch model. The results of different performance metrics like sensitivity, specificity, precision and accuracy are 0.981, 0.973, 0.982 and 0.953 respectively. Hence this experiment herewith encourages the division of input images with ensembling of features rather than random resizing.

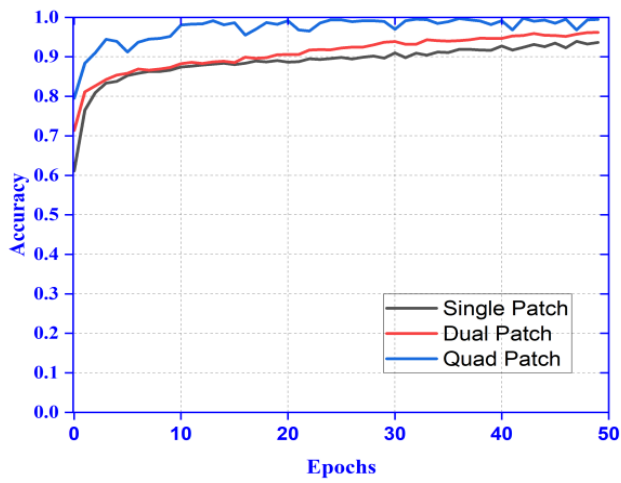


(a) Accuracy with quad patch of image

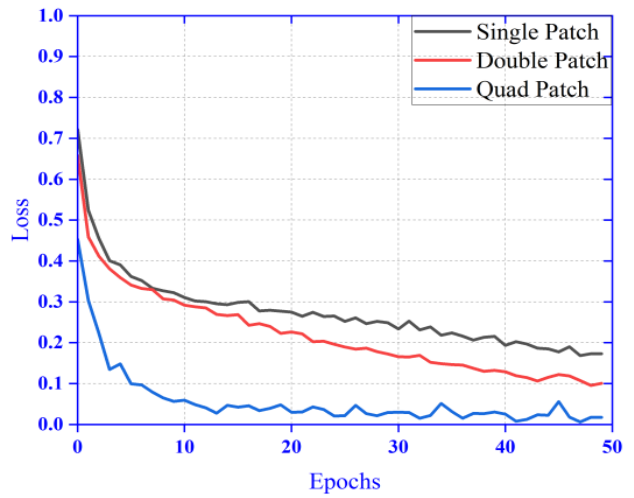


(b) Accuracy with quad patch of image

Figure 8. Accuracy and loss of the proposed model with the quad patch of image

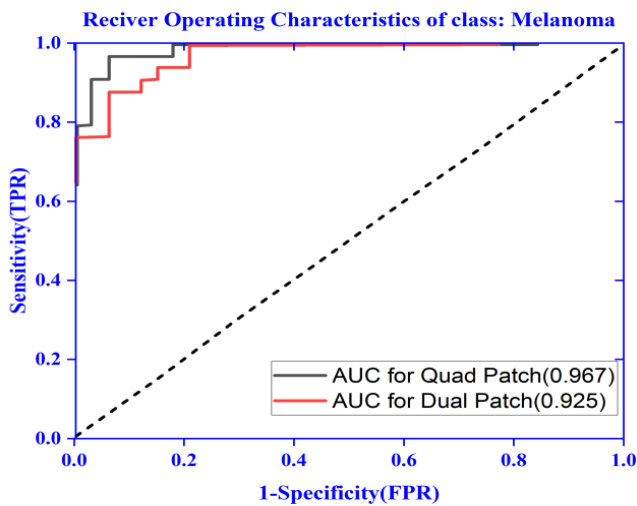


(a) Accuracy plot for single, dual and quad patch

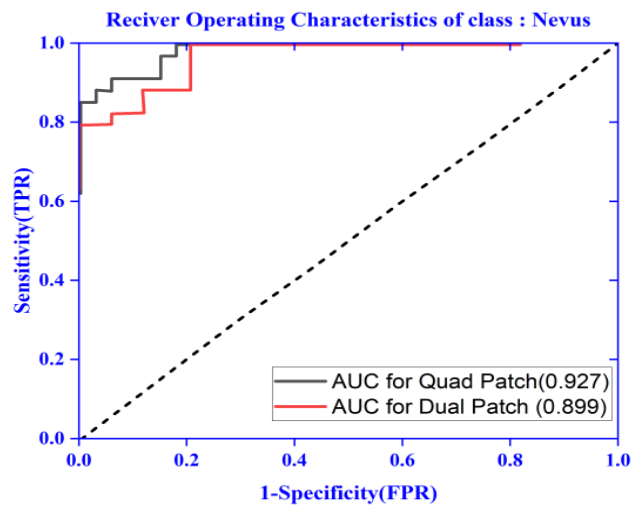


(b) Loss plot for single, dual and quad patch

Figure 9. Combined accuracy and loss plot for single, dual and quad patch of image



(a) AUC-ROC plot for class: Melanoma



(b) AUC-ROC plot for class: nevus

Figure 10. ROC-AUC plot for class: Melanoma and class: nevus

5.5 State of art comparison

The comparison of suggested proposed model was done with the existing deep learning-based schemes. Table 5 shows a summary of the results. It can be seen that the presented method performed better than other approaches when the input image is divided for dual or quad patches to fit the input pre-trained layer rather than random resizing. Although the data set looks small, in the quad patch model it becomes five times (four patches of each image to four different bags for

ensembling and one for perceptual attention), which is used to evaluate the proposed model than in several recent studies [5, 8, 10, 42-50]. In the work proposed by Bisla et al. [51] though the accuracy is 91.5%, the unbalanced data set between nevus and melanoma is a concern. Also, the model proposed by Pomponiu et al. [52] shows an accuracy of 93.64% but it seems region interest has been taken manually. The proposed model also compared with the ensembled methodologies [53-55] showing a better evaluated values for the metrics considered for the classification.

Table 5. Experimental results of the proposed methodology compared with the state-of-the-art methods

Authors	Methods	Sensitivity	Specificity	Precision	Accuracy
Giotis et al. [10]	Texture descriptor	0.62	0.85	0.74	0.76
Giotis et al. [10]	Color descriptor	0.74	0.72	0.64	0.73
Esteva et al. [5]	Inception V3	0.96	N.A.	N.A.	0.72
Almansour et al. [8]	Hybrid texture feature	0.94	0.86	N.A.	0.90
Blum et al. [42]	Diagnostic algorithm	0.93	0.87	N.A.	0.87
Ramlakhan and Shang [43]	kNN classifier	0.61	0.80	N.A.	0.67
Dorj et al. [44]	ECOC SVM classifier	0.97	0.90	N.A.	0.94
Nasr-Esfahani et al. [45]	Illumination correction	0.81	0.80	0.75	0.81
Mukherjee et al. [46]	Optimized NN using PSO	0.86	0.86	N.A.	0.86
Jianu et al. [47]	Neural Network	0.72	0.89	0.87	0.805
Fraiwani and Faouri [48]	ResNet 101	N.A.	N.A.	N.A.	0.829
Bisla et al. [51]	Deep Convolutional GAN	N.A.	N.A.	N.A.	0.915
Pomponiu et al. [52]	Alex-Net + kNN	0.921	0.951	N.A.	0.936
Codella et al. [49]	Alex-Net + UNet with SVM	0.949	0.928	N.A.	0.931
Jojoa et al. [50]	ResNet 152	0.82	0.925	0.75	0.904
Xie et al. [53]	Ensemble of neural networks	0.833	0.95	N.A.	91.1
Taşar [54]	Ensembled pretrained models	0.867	0.970	N.A.	0.831
Moghimi et al. [55]	Boosted CNN (with skin image)	0.782	0.859	0.884	0.862
Proposed Model	Dual Patch	0.967	0.961	0.956	0.920
Proposed Model	Quad Patch	0.981	0.973	0.982	0.953

6. CONCLUSION

We discuss two crucial aspects of classifying skin lesions using dermoscopic image data. First, we make use of the image patches in an incremental manner (image as a whole, dual patch, and quad patch) as the input without random resizing and used a perceptual attention module to maintain the spatial attention. We demonstrate how a novel patch-based technique can enhance feature selection thereby increasing classification precision. The second problem is the intermediate feature assembling with a modified ensembling algorithm followed by classification. We compare different pre-trained models like Xception, VGG16, ResNet101, InceptionV3, DenseNet121, EfficientNetB0, AlexNet, GoogleNet, and MobileNetV2 used as transfer learning for classification and choose MobileNetV2 as the base model because of its smaller size and fewer parameters. Also, we examined our model for three distinct input types using full images, dual patches, and quad patches in incremental techniques, resulting in a considerable improvement in accuracy. Also, AUC-ROC plot is significant for the proposed model.

6.1 Limitations and future scope

Our patch-based ensembling does not encourage the limitless division of images, further research can be carried out in this regard. In this paper we have carried out only binary class classification, in future experiments can be carried out for multiclass classification. The model can be trained with a

large amount of data from different sources with a better resource framework. Training of similar ensembling model can be carried out beyond the quad patch to observe the behaviour of the model.

DECLARATIONS

- Conflict of interest/Competing interests: The authors have no affiliation with any organization with a direct or indirect financial interest in the subject matter discussed in the manuscript.
- Ethics approval: This manuscript has not been submitted to, nor is under review at, another journal or other publishing venue.
- Availability of data and materials: In this paper, we have used ISIC 2017 data set which is available in this link <https://www.isic-archive.com/>.
- Authors' contributions: Model designing, analysis and interpretation of the data is done by the first author; Then first author with other have equal contribution for drafting the article or revising it critically for important intellectual content; and approval of the final version.

REFERENCES

- [1] Taufiq, M.A., Hameed, N., Anjum, A., Hameed, F. (2017). M-Skin doctor: A mobile enabled system for early melanoma skin cancer detection using support

- vector machine. In *eHealth 360°: International Summit on eHealth*, Budapest, Hungary, Springer International Publishing. Springer, Cham, pp. 468-475. https://doi.org/10.1007/978-3-319-49655-9_57
- [2] Li, Q., Chang, L., Liu, H., Zhou, M., Wang, Y., Guo, F. (2015). Skin cells segmentation algorithm based on spectral angle and distance score. *Optics & Laser Technology*, 74: 79-86. <https://doi.org/10.1016/j.optlastec.2015.05.017>
- [3] Kasmi, R., Mokrani, K. (2016). Classification of malignant melanoma and benign skin lesions: Implementation of automatic ABCD rule. *IET Image Processing*, 10(6): 448-455. <https://doi.org/10.1049/iet-ipr.2015.0385>
- [4] Friedman, R.J., Rigel, D.S., Kopf, A.W. (1985). Early detection of malignant melanoma: the role of physician examination and self-examination of the skin. *CA: A Cancer Journal for Clinicians*, 35(3): 130-151. <https://doi.org/10.3322/canjclin.35.3.130>
- [5] Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., Blau, H.M., Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639): 115-118.
- [6] Iyatomi, H., Oka, H., Celebi, M.E., Ogawa, K., Argenziano, G., Soyer, H.P., Koga, H., Saida, T., Ohara, K., Tanaka, M. (2008). Computer-based classification of dermoscopy images of melanocytic lesions on acral volar skin. *Journal of Investigative Dermatology*, 128(8): 2049-2054. <https://doi.org/10.1038/jid.2008.28>
- [7] Anas, M., Gupta, K., Ahmad, S. (2017). Skin cancer classification using K-means clustering. *International Journal of Technical Research and Applications*, 5(1): 62-65.
- [8] Almansour, E., Jaffar, M.A. (2016). Classification of Dermoscopic skin cancer images using color and hybrid texture features. *IJCSNS International Journal of Computer Science and Network Security*, 16(4): 135-139.
- [9] Capdehourat, G., Corez, A., Bazzano, A., Alonso, R., Musé, P. (2011). Toward a combined tool to assist dermatologists in melanoma detection from dermoscopic images of pigmented skin lesions. *Pattern Recognition Letters*, 32(16): 2187-2196. <https://doi.org/10.1016/j.patrec.2011.06.015>
- [10] Giotis, I., Molders, N., Land, S., Biehl, M., Jonkman, M.F., Petkov, N. (2015). MED-NODE: A computer-assisted melanoma diagnosis system using non-dermoscopic images. *Expert Systems with Applications*, 42(19): 6578-6585. <https://doi.org/10.1016/j.eswa.2015.04.034>
- [11] Ruiz, D., Berenguer, V., Soriano, A., Sánchez, B. (2011). A decision support system for the diagnosis of melanoma: A comparative approach. *Expert Systems with Applications*, 38(12): 15217-15223. <https://doi.org/10.1016/j.eswa.2011.05.079>
- [12] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115: 211-252. <https://doi.org/10.1007/s11263-015-0816-y>
- [13] Simonyan, K., Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv Preprint arXiv: 1409.1556*. <https://doi.org/10.48550/arXiv.1409.1556>
- [14] Szegeedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z. (2016). Rethinking the inception architecture for computer vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 2818-2826. <https://doi.org/10.1109/CVPR.2016.308>
- [15] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C. (2018). Mobilenetv2: Inverted residuals and linear bottlenecks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 4510-4520. <https://doi.org/10.1109/CVPR.2018.00474>
- [16] Liu, Q., Yu, L., Luo, L., Dou, Q., Heng, P.A. (2020). Semi-Supervised medical image classification with relation-driven self-ensembling model. *IEEE Transactions on Medical Imaging*, 39(11): 3429-3440. <https://doi.org/10.1109/TMI.2020.2995518>
- [17] Gu, J., Hu, H., Wang, L., Wei, Y., Dai, J. (2018). Learning region features for object detection. *arXiv Preprint arXiv:1803.07066*. <https://doi.org/10.48550/arXiv.1803.07066>
- [18] Hu, H., Gu, J., Zhang, Z., Dai, J., Wei, Y. (2018). Relation networks for object detection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 3588-3597. <https://doi.org/10.1109/CVPR.2018.00378>
- [19] Zhang, X., Shang, S., Tang, X., Feng, J., Jiao, L. (2021). Spectral partitioning residual network with spatial attention mechanism for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 60: 1-14. <https://doi.org/10.1109/TGRS.2021.3074196>
- [20] Lu, X., Lin, Z., Jin, H., Yang, J., Wang, J.Z. (2014). Rapid: Rating pictorial aesthetics using deep learning. In *Proceedings of the 22nd ACM International Conference on Multimedia*, pp. 457-466. <https://doi.org/10.1145/2647868.2654927>
- [21] Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24: 123-140. <https://doi.org/10.1007/BF00058655>
- [22] Freund, Y., Schapire, R.E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1): 119-139. <https://doi.org/10.1006/jcss.1997.1504>
- [23] Khatami, A., Babaie, M., Khosravi, A., Tizhoosh, H.R., Nahavandi, S. (2018). Parallel deep solutions for image retrieval from imbalanced medical imaging archives. *Applied Soft Computing*, 63: 197-205. <https://doi.org/10.1016/j.asoc.2017.11.024>
- [24] Huang, G., Li, Y., Pleiss, G., Liu, Z., Hopcroft, J.E., Weinberger, K.Q. (2017). Snapshot ensembles: Train 1, get m for free. *arXiv Preprint arXiv: 1704.00109*. <https://doi.org/10.48550/arXiv.1704.00109>
- [25] Hara, K., Saitoh, D., Shouno, H. (2016). Analysis of dropout learning regarded as ensemble learning. In *Artificial Neural Networks and Machine Learning-ICANN 2016: 25th International Conference on Artificial Neural Networks*, Barcelona, Spain, September 6-9, 2016, Proceedings, Part II. Springer International Publishing. Springer, Cham, 25: 72-79. https://doi.org/10.1007/978-3-319-44781-0_9
- [26] Huang, G., Sun, Y., Liu, Z., Sedra, D., Weinberger, K.Q. (2016). Deep networks with stochastic depth. *arXiv Preprint arXiv:1603.09382*.

- <https://doi.org/10.48550/arXiv.1603.09382>
- [27] Singh, S., Hoiem, D., Forsyth, D. (2016). Swapout: Learning an ensemble of deep architectures. *Advances in Neural Information Processing Systems*, 29.
- [28] Diba, A., Sharma, V., Pazandeh, A., Pirsiavash, H., Van Gool, L. (2017). Weakly supervised cascaded convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 5131-5139. <https://doi.org/10.1109/CVPR.2017.545>
- [29] Zhang, J., Bargal, S.A., Lin, Z., Brandt, J., Shen, X., Sclaroff, S. (2018). Top-down neural attention by excitation backprop. *International Journal of Computer Vision*, 126(10): 1084-1102. <https://doi.org/10.1007/s11263-017-1059-x>
- [30] Zhao, R., Ouyang, W., Li, H., Wang, X. (2015). Saliency detection by multi-context deep learning. *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, pp. 1265-1274. <https://doi.org/10.1109/CVPR.2015.7298731>
- [31] Long, J., Shelhamer, E., Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, pp. 3431-3440. <https://doi.org/10.1109/CVPR.2015.7298965>
- [32] Guo, H., Fan, X., Wang, S. (2017). Human attribute recognition by refining attention heat map. *Pattern Recognition Letters*, 94: 38-45. <https://doi.org/10.1016/j.patrec.2017.05.012>
- [33] Kumar, A., Kim, J., Lyndon, D., Fulham, M., Feng, D. (2016). An ensemble of fine-tuned convolutional neural networks for medical image classification. *IEEE Journal of Biomedical and Health Informatics*, 21(1): 31-40. <https://doi.org/10.1109/JBHI.2016.2635663>
- [34] Xie, X., Xing, J., Kong, N., Li, C., Li, J., Zhang, S. (2017). Improving colorectal polyp classification based on physical examination data-an ensemble learning approach. *IEEE Robotics and Automation Letters*, 3(1): 434-441. <https://doi.org/10.1109/LRA.2017.2746918>
- [35] Chollet, F. (2017). Xception: Deep learning with depthwise separable convolutions. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 1251-1258. <https://doi.org/10.1109/CVPR.2017.195>
- [36] He, K., Zhang, X., Ren, S., Sun, J. (2016). Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, pp. 770-778. <https://doi.org/10.1109/CVPR.2016.90>
- [37] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q. (2017). Densely connected convolutional networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, HI, USA, pp. 4700-4708. <https://doi.org/10.1109/CVPR.2017.243>
- [38] Tan, M. (2019). Efficientnet: Rethinking model scaling for convolutional neural networks. *arXiv preprint arXiv:1905.11946*. <https://doi.org/10.48550/arXiv.1905.11946>
- [39] Krizhevsky, A., Sutskever, I., Hinton, G.E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6): 84-90. <https://doi.org/10.1145/3065386>
- [40] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A. (2015). Going deeper with convolutions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, MA, USA, pp. 1-9. <https://doi.org/10.1109/CVPR.2015.7298594>
- [41] Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kallou, A., Liopyris, K., Mishra, N., Kittler, H., Halpern, A. (2018). Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, Washington, DC, USA, pp. 168-172. <https://doi.org/10.1109/ISBI.2018.8363547>
- [42] Blum, A., Luedtke, H., Ellwanger, U., Schwabe, R., Rassner, G., Garbe, C. (2004). Digital image analysis for diagnosis of cutaneous melanoma. Development of a highly effective computer algorithm based on analysis of 837 melanocytic lesions. *British Journal of Dermatology*, 151(5): 1029-1038. <https://doi.org/10.1111/j.1365-2133.2004.06210.x>
- [43] Ramlakhan, K., Shang, Y. (2011). A mobile automated skin lesion classification system. In *2011 IEEE 23rd International Conference on Tools with Artificial Intelligence*, Boca Raton, FL, USA, pp. 138-141. <https://doi.org/10.1109/ICTAI.2011.29>
- [44] Dorj, U.O., Lee, K.K., Choi, J.Y., Lee, M. (2018). The skin cancer classification using deep convolutional neural network. *Multimedia Tools and Applications*, 77: 9909-9924. <https://doi.org/10.1007/s11042-018-5714-1>
- [45] Nasr-Esfahani, E., Samavi, S., Karimi, N., Soroushmehr, S.M.R., Jafari, M.H., Ward, K., Najarian, K. (2016). Melanoma detection by analysis of clinical images using convolutional neural network. In *2016 38th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, Orlando, FL, USA, pp. 1373-1376. <https://doi.org/10.1109/EMBC.2016.7590963>
- [46] Mukherjee, S., Adhikari, A., Roy, M. (2019). Malignant melanoma detection using multi layer perceptron with optimized network parameter selection by PSO. In *Contemporary Advances in Innovative and Applicable Information Technology: Proceedings of ICCAIAIT 2018*. Springer Singapore, pp. 101-109. https://doi.org/10.1007/978-981-13-1540-4_11
- [47] Jianu, S.R.S., Ichim, L., Popescu, D. (2019). Automatic diagnosis of skin cancer using neural networks. In *2019 11th International Symposium on Advanced Topics in Electrical Engineering (ATEE)*, Bucharest, Romania, pp. 1-4. <https://doi.org/10.1109/ATEE.2019.8724938>
- [48] Fraiwan, M., Faouri, E. (2022). On the automatic detection and classification of skin cancer using deep transfer learning. *Sensors*, 22(13): 4963. <https://doi.org/10.3390/s22134963>
- [49] Codella, N., Cai, J., Abedini, M., Garnavi, R., Halpern, A., Smith, J.R. (2015). Deep learning, sparse coding, and SVM for melanoma recognition in dermoscopy images. In *International Workshop on Machine Learning in Medical Imaging*. Cham: Springer International Publishing. Springer, Cham, pp. 118-126. https://doi.org/10.1007/978-3-319-24888-2_15
- [50] Jojoa Acosta, M.F., Caballero Tovar, L.Y., Garcia-Zapirain, M.B., Percybrooks, W.S. (2021). Melanoma diagnosis using deep learning techniques on

dermatoscopic images. *BMC Medical Imaging*, 21: 1-11. <https://doi.org/10.1186/s12880-020-00534-8>

[51] Bisla, D., Choromanska, A., Berman, R.S., Stein, J.A., Polsky, D. (2019). Towards automated melanoma detection with deep learning: Data purification and augmentation. In 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), Long Beach, CA, USA, pp. 2720-2728. <https://doi.org/10.1109/CVPRW.2019.00330>

[52] Pomponiu, V., Nejati, H., Cheung, N.M. (2016). Deepmole: Deep neural networks for skin mole lesion classification. In 2016 IEEE International Conference on Image Processing (ICIP), Phoenix, AZ, USA, pp. 2623-2627. <https://doi.org/10.1109/ICIP.2016.7532834>

[53] Xie, F., Fan, H., Li, Y., Jiang, Z., Meng, R., Bovik, A. (2016). Melanoma classification on dermoscopy images using a neural network ensemble model. *IEEE Transactions on Medical Imaging*, 36(3): 849-858. <https://doi.org/10.1109/TMI.2016.2633551>

[54] Taşar, B. (2023). SkinCancerNet: Automated classification of skin lesion using deep transfer learning method. *Traitement du Signal*, 40(1): 285.

<https://doi.org/10.18280/ts.400128>

[55] Moghimi, M., Belongie, S.J., Saberian, M.J., Yang, J., Vasconcelos, N., Li, L.J. (2016). Boosted convolutional neural networks. In *BMVC*, 5: 6.

NOMENCLATURE

CNN	Convolutional Neural Network
ROC	Receiver Operating Characteristic
AUC	Area Under the Curve
<i>TP</i>	True Positive
<i>FP</i>	False Positive
<i>TN</i>	True Negative
<i>FN</i>	False Negative
<i>N</i>	Number of image patches

Subscripts

<i>i</i>	Number of base model/data
set	pixels in the image patch
<i>j</i>	Intermediate feature matrix