Melanoma segmentation and classification using dermoscopy images via integrated deep neural network

Aqsa Hassan[1*], Akash Harijan[1] and Mohammad Jaber Hossain[1]

[1]Norwegian University of Science and Technology, Gjøvik, Norway.

*Corresponding author(s). E-mail(s): aqsah@stud.ntnu.no;
Contributing authors: akashh@stud.ntnu.no;
mohammjh@stud.ntnu.no;

Abstract

Skin lesion is one of the common contributors to deaths globally. However, early detection of this disease can lead to a very high chance of survival. Deep learning neural networks have been successful in demonstrating promising progress on medical imaging for skin lesion classification. Therefore, in this project, we have evaluated the performance comparison of the recent state-of-the-art deep convolutional neural networks, that have already achieved significant performance on lesion image segmentation and classification, for the task of skin lesion detection using HAM10000 through transfer learning. In order to address the limitation of training data, several data augmentation techniques applied to the training dataset to avoid overfitting for skin lesion classification. Moreover, different optimization techniques are used while training these models, such as learning rate decay and dropout, which made the model generalized better for the skin lesion classification. In addition, we investigated ensemble learning to build a single strong model that has a higher performance accuracy than any of the base models. Intensive experimentation has also been performed to identify the role of segmentation in the skin lesion classification where ISIC 2018 dataset has been used for training the double-UNet architecture. This achieved 0.76 in Dice coefficient, 0.72 in specificity and 0.80 sensitivity. The overall results of experimentation revealed that the ensemble of InceptionV3 and DenseNet201 has outperformed by achieving 0.76 accuracy without segmentation and 0.75 with segmentation.

Keywords: Melanoma Segmentation, Melanoma Classification, skin lesion, deep learning

1 Introduction

Among all the skin lesions, melanoma is a type of lesion which is referred as a dermatological cancer that can be detected by dermatological screening and biopsy tests of the patient's infected skin. These tests are highly expensive and require high time to get the result where some medical experts required to present for analyzing the test results. The numbers of melanoma patients are increased by more than 50 percent in last 10 years due to the increased amount of UAV radiation arriving to the surface of the earth [1]. Computer aided solution depends on applying different image processing techniques are being utilized before whereas for the time being deep learning based solution are giving better performance in less time rather than before. Automatic detection of skin lesion possess different difficult problems as skin having hair and blood vessels in the infested area, skin colour of the infected area including texture, shape and size of the area varied along with un-noticeable infection area edges [2].

Skin lesions detection is a serious concern as it is one of the most common contributors to the cause of death all over the world but early detection of this disease can lead to 99% of 5-year survival rate. Computer-aided diagnostic systems can drastically aid physicians to detect skin cancer in the early stages and avoid unnecessary biopsies, improving patient care and reducing cost. In addition, the recent advancement in deep learning techniques for medical image analysis has enabled the medical imaging-based diagnosis of skin cancer which has proven successful to help dermatologists in giving better treatment to their patients. The emergence of intelligent medical imaging-based diagnosis systems also assisted to avoid unnecessary biopsies, improving patient care, and reducing costs. Early detection of skin cancer can be aided by portable systems and even mobile applications to help patients by providing suggested diagnoses with the supervision of the physicians that can act as a warning sign. The visual inspection by dermatologists and dermatoscopic images when combined together, results in a 75% to 84% of melanoma detection accuracy by dermatologists.

Moreover, the change in viewing angle and illumination conditions during image acquisition has played a crucial role in skin cancer classification[3]. In addition, there are many drawbacks related to errors and the loss of information due to the multiple sizes and shapes of images, noise and artifacts presence, irregular fuzzy boundaries, low contrast, and color illumination[4] as shown in Figure 1.

Training a successful deep learning model highly depends upon the availability of sufficient labeled training data. Due to this limitation, the model overfits the training dataset and does not perform well on unseen input data.

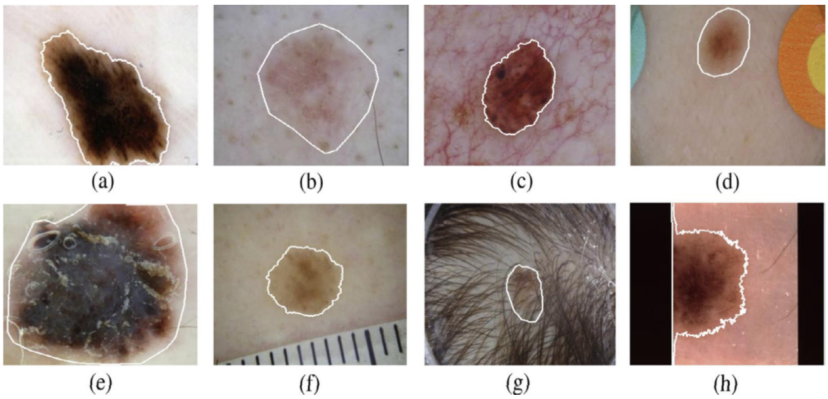


Fig. 1 Examples of some complicated cases of skin lesions including (a) irregular boundary, (b) low contrast, (c) blood vessels, (d) illumination color, (e) bubble, (f) artifact ruler, (g) artifact hair and (h) artifact frame.

Moreover, skin cancer classification is a complicated task because of the variation in image kinds and sources. There is such an enormous variation in the appearance of human skin that it is difficult and challenging to discover the skin lesion. The low inter-class variance and high intra-class variance requires a high level of domain-specific expertise to distinguish between the type of skin cancer. The variations in viewing angle and illumination conditions during image acquisition also plays a crucial role in skin lesions classification. Furthermore, there are many drawbacks related to errors and the loss of information due to the multiple sizes and shapes of images, noise and artifacts presence, irregular fuzzy boundaries, low contrast, and color illumination as shown in Figure 1.

In this study, we have investigated several deep convolutional neural networks using HAM10000[5] dataset to provide a performance evaluation of different deep learning architectures for skin lesions classification with segmentation and classification without segmentation. The main contribution of this work are summarized as follows:

- We evaluated the performance of deep CNNs, that have already achieved significant performance on image classification, for the task of skin lesion detection using HAM10000 [5] dataset.
- Each deep learning model for classification has been trained using transfer learning. We have fine-tuned all the layers of InceptionV3 and DenseNet201 and fine-tuned only the output layers for the other models.
- In order to address the limitation of labeled training data, several data augmentation techniques have been investigated to increase the number of samples in training and to avoid overfitting for skin lesion classification and segmentation.

- Furthermore, certain optimization techniques such as dropout and learning rate decay are used while training the model that has resulted in an adequate generalization of the model for skin lesion classification.
- Intensive experimentation has been performed to identify the role of segmentation in the skin lesion classification where we have investigated double Unet architecture.
- Finally, we investigated ensemble learning to build a single strong model that has a higher performance accuracy than any of the base models. Ensemble learning combines multiple base models to produce one optimal predictive model.

The rest of this paper is organized as follows: In Section 2, the recent promising research work on skin lesion segmentation. The details of the methodology for performance evaluation of these networks is presented in Section 3. Section 4 elaborates the evaluation and results. Finally, in Section 5 the some conclusions of this paper are drawn.

2 Related works

Significant amount of work gone through to solve the lesion related problem easier using automated or semiautomatic systems over the time where segmentation and classification of lesion attribute have considerable contribution which address different solution based on different classical computer vision based image processing technique and deep neural architecture dependent solutions.Significant amount of study went through the data augmentation techniques due to the size of the datasets publicly available are not large enough to efficiently train the deep neural networks as the models are data hungry. The study [6] identified 13 data augmentation methods which can apply on skin lesion related problems including changing Saturation, Contrast, hue and Brightness and applying affine operation such as rotation, shearing and scaling and adding noise or random erasing which will create a large range of variant in the dataset that are going to be used for the deep neural network training and testing. Some image processing technique has a commendable impact on further processing as it is benefited to represent the image to the model in particular characteristics. To be precise, remarkable amount of the work done earlier gone through image resizing techniques to make uniform input images for the training and testing set [7–9]. To train deep learning based data driven model efficiently, existing convolutional neural networks can be trained on large datasets which generate the weights for initializing the network training, rather than assigning random weights which is known as transfer learning and this type of architectures often utilized in classification problems [10–13].In [14], the study investigated Dropout and Dropblock as regularization techniques and applied muli-weighted focal loss by setting C(Mel)=1.8 to achieve a BACC of 0.864 whereas modified RandAugment helped to gain significant performance which is both computationally inexpensive and efficient. Kassem et al.[15] modified the GoogleNet architecture by additional filters to each layer

to enhance features extraction and reduce noise. The model was trained using transfer learning. It helped the proposed model to not overfit even without pre-processing steps. They used precision, specificity, sensitivity and accuracy as a performance measure which is 80.36%, 97%, 79.8% and 94% respectively.

To find an efficient model for Skin Lesion Segmentation (SLS), considering the advantage of the U-net architectures encoder-decoder model and recognizing the high efficiency of MobileNetV3, the study [8] comes up with a hybrid model combining this two architectures whereas encoder portion utilized the MobileNetV3 and decoder portion exercised different types of U-net architectures in particular Separable-UNet, BCDU, LSTM-UNet and UNet where MobileNetV3-UNet outperformed then others. The famous deep learning based architecture that made huge improvement in segmentation problem overall is U-NET [16]. After that, people are trying to imitate this architecture in medical imaging as well. There is a recent work on medical imaging based segmentation, based on U-NET is Double-Unet [17] that gives quite good and promising results.

The availability of large training data and the advancement in both software and hardware resources have made it possible for more deep neural network architectures to make impressive advances and evolvement in Computer Vision in recent years, starting from AlexNet in 2012. The special characteristics of Deep CNN is that the first layers usually learn very general and "low-level" features of images, while the output layers of the network learn the semantics and high-level features.

2.1 Deep learning neural networks for classification

The availability of large training data and the advancement in both software and hardware resources have made it possible for more deep neural network architectures to make impressive advances and evolvement in Computer Vision in recent years, starting from AlexNet in 2012. The special characteristics of Deep CNN is that the first layers usually learn very general and "low-level" features of images, while the output layers of the network learn the semantics and high-level features. Two state-of-the-art networks and one modified deep neural network are briefly introduced in the following subsections. These networks have achieved remarkable performance on image classification. Their performances on skin lesion classification are also compared.

2.1.1 Inception

The Inception network Inception networks is also known as GoogLeNet[18]. It consisted of twenty two convolutional layers and nine inception modules as presented in Figure 2. This architecture won the ImageNet 2014 competition by reducing a top-5 error rate by 6.7%. Inception V3 was the top performers on ImageNet with 0.937 accuracy for top-5 and 0.779 for top-1. The basic idea behind GoogleNet architecture was the implementation of several size filters such as 1x1, 3x3 and 5x5. This has helped the network go deeper than VGG

network and is smaller than ALexNet[19]. However, this architecture faced the issue of vanishing gradient. Another variant of this network is Inception-ResNet that combines both the residual learning and inception block to improve the classification performance.

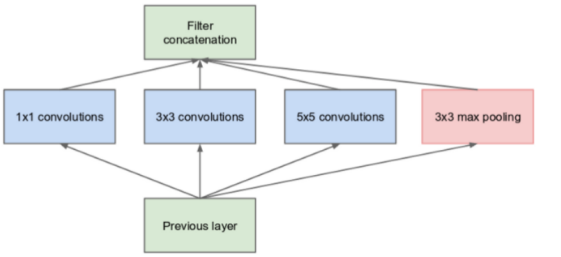


Fig. 2 Inception Module, naive version[18]

2.1.2 DenseNet

DenseNet[20] architecture explicitly differentiates between information that using skip connection is added to the network between layers as shown in Figure 5 and information that is preserved by concatenation of features. This

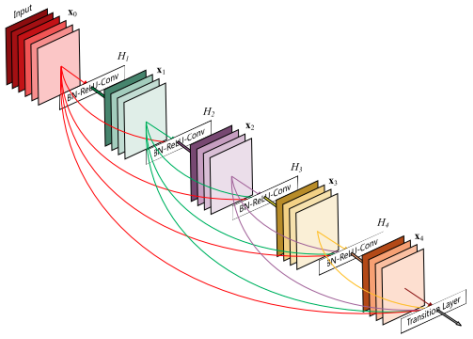


Fig. 3 The DenseNet architecture[20] using skip connections where each layer is linked with all the other layers.

has been beneficial in increasing the variation of the input in following layers and improves efficiency. DenseNet architecture is one of the top performers on ImageNet by achieving 0.936 top-1 and 0.773 top-5 accuracy. The performance of DenseNet is also competitive to InceptionV3 while it has less parameters (i.e. approximately 20 millions compare with approximate 23 millions of Inception V3).

2.1.3 Ensemble learning

The performance of deep convolutional neural networks is highly dependent upon the availability of large training annotation dataset. In addition, these networks are nonlinear and demonstrate a high variance. Ensemble Learning[21] is the method that combines multiple base models to produce one optimal classification model that has been successful to outperformed than any of the base models. It has been proven beneficial to minimize the variance of predictions and generalization error. This method can be implemented in two different ways i.e. we can build ensemble of homogeneous base learners that consisted of using one type of base learning algorithm or heterogeneous learners that use different types of learning algorithms. Bagging, Boosting and a bucket of models are the three main categories of ensemble learning.

2.2 Deep learning neural networks for classification with segmentation

A good number of study done earlier separately for segmentation and classification to achieve high performance indexes whereas segmentation with classification might provide better performance according to the indexes for the diagnosis of skin lesion affecting cancer melanoma. The proposed method [22] addressed the problem and proposed a combined architecture having coarse segmentation network with mask classification network and go around with another enhanced segmentation network. Coarse SN benefited the mask CN with coarsely segmented lesion area to classify the lesion accurately, then the generated mask CN assisted the enhanced SN by providing collected localization information to get segmented area, in this method both segmentation and classification helped each other mutually to get better performance. But classification on segmented images showed performance degradation in [23] where the proposed method segmented the lesions using SegNet and bidirectional convolutional LSTM with U-Net,then the segmented images classified using ResNet50, Xception, VGG16, InceptionV3, where every architecture found lower performance indexes rather than directly fed images.

3 Methodology

In this section, we discussed about the datasets used for the segmentation and classification task and reasons behind choosing the datasets. Besides the data augmentation applied for increasing the dataset size and introduce diversity on data. Alongside, the detail workflow of segmentation task, classification task on direct images and segmented images. These experiments are performed on high-performance machines in order to minimized the time for training and testing. The Nvidia Tesla V100 machine was used, having 640 Tensor Cores, 8 GB GPU memory and 90GB RAM. The operating system used was Linux Ubuntu 18.04 LTS that has Cuda 10.2, cuDNN v7.5.0, Python 3.6 with its deep learning libraries such as Keras, Tensorflow, etc.

3.1 Lesion Classification without Segmentation

3.1.1 Exploratory Data Analysis and Data Transformation

There are a very limited number of benchmark datasets for skin lesion images. Moreover, most of these datasets are highly imbalanced. The ISIC dataset consisted of 33126 training samples which is a huge dataset. However, the number of images in each class is highly imbalanced as shown in Table 1.

Table 1 The number of samples in each class of HAM and ISIC dataset.The ISIC dataset contains a huge number of images in total but the distribution of images is highly imbalanced as shown in the table.

Available Classes	HAM dataset	ISIC 2018 dataset
nv: Melanocytic_nevi	6705	11861
mel: melanoma	1113	1056
bkl: Benign_keratosis-like_leisons	1099	477
bcc: Basal_cell_carcinoma	514	72
akiec: Actinic_keratoses	327	2
vasc: Vascular_lesions	142	0
df: Dermatofibroma	115	7

The models trained on such a dataset will not perform well. Therefore, we have used the HAM10000 dataset for experimentation in this study. The publicly available HAM10000 dataset was first introduced by Tschandl et al [5] that was collected over the period of twenty years and is considered as one of the largest dermoscopic image datasets which later became the part of ISIC 2018 dataset. The HAM10000 dataset contains only 10015 images where majority of the images are not lesions as shown by the graph in present Fig 4. The distribution of images in each class of HAM1000 dataset is illustrated in the graph of Fig 5.

Now let's explore our dataset with the help of a provided metadata csv file. First we can visualize where the lesions are mostly located in our images as shown in Fig 6. and then we can explore the treatment that patients received as shown in Fig 7.

Referred to Fig 7, following are the description for each dx_type:

histo: Histopathologic diagnoses of excised lesions have been performed by specialized dermatopathologists.

follow_up: If nevi monitored by digital dermatoscopy did not show any changes during 3 follow-up visits or 1.5 years we accepted this as evidence of biologic benignity. Only nevi, but no other benign diagnoses were labeled

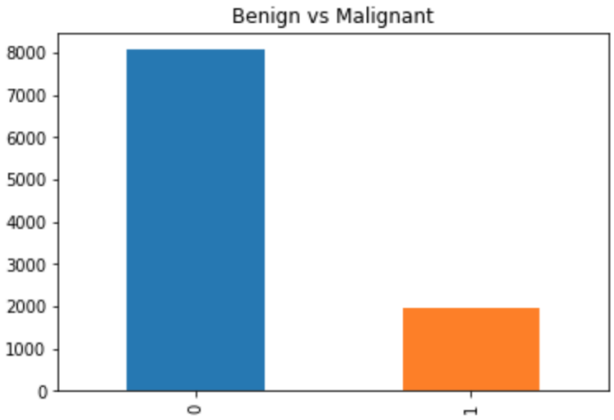


Fig. 4 The HAM10000 dataset consists of 10015 images with seven classes of skin lesions where eight thousand images in the dataset are benign while only two thousand images are malignant.

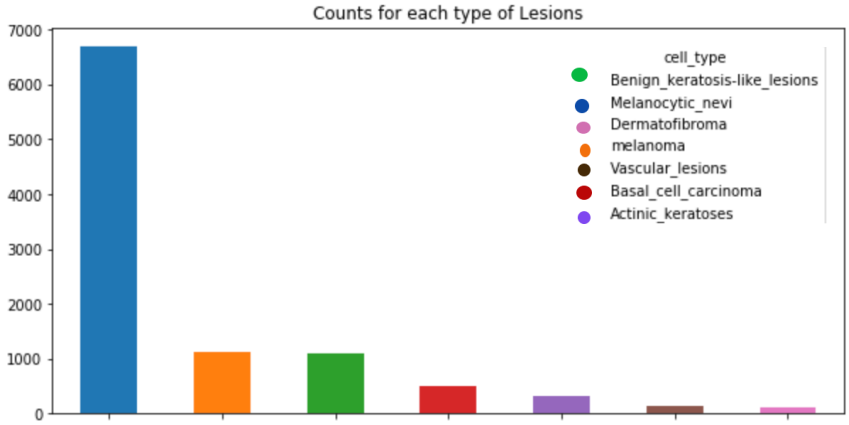


Fig. 5 The dataset contains 1113 images of melanocytic nevi, 327 images of AK, 1099 images of benign keratosis, 514 images of basal cell carcinomas, 115 images of dermatofibromas, 6705 images of melanomas,and 142 images of vascular skin lesions. The dataset is also biased towardMelanocytic nevi class as it has the highest number of images.

with this type of ground-truth because dermatologists usually do not monitor dermatofibromas, seborrheic keratoses, or vascular lesions.

consensus: For typical benign cases without histopathology or follow-up we provide an expert consensus rating of authors PT and HK. We applied the consensus label only if both authors independently gave the same unequivocal benign diagnosis. Lesions with this type of ground-truth were usually photographed for educational reasons and did not need further follow-up or biopsy for confirmation.

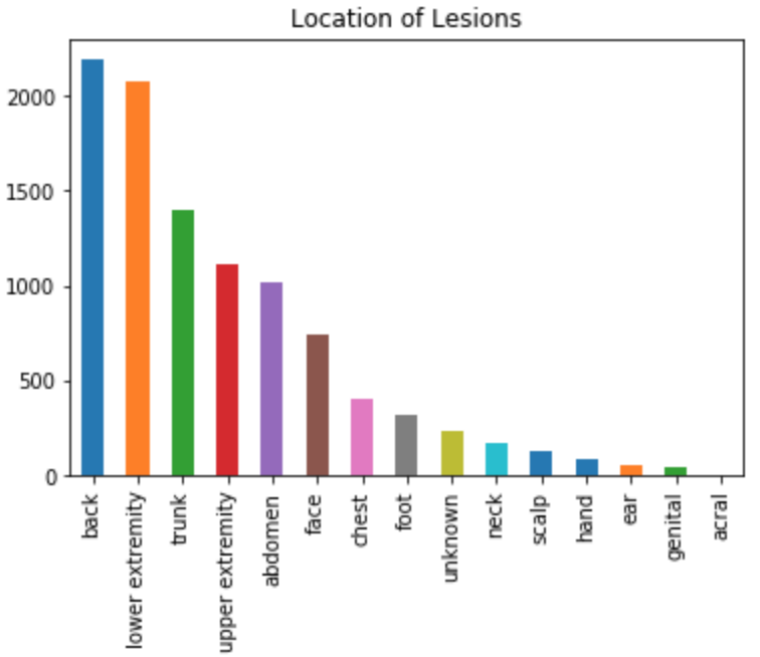


Fig. 6

confocal: Reflectance confocal microscopy is an in-vivo imaging technique with a resolution at near-cellular level, and some facial benign keratosis were verified by this method.

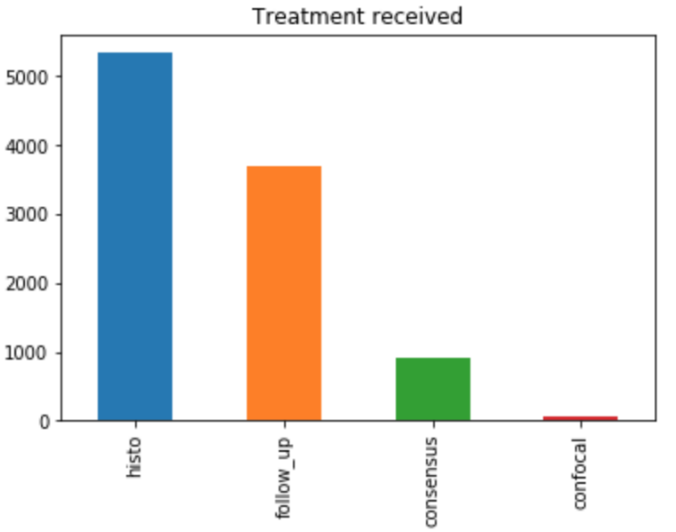


Fig. 7

We can explore some characteristics of our patients as well. As illustrated from the graphs shown in Fig 8. we can clearly observe that most of the patients are above 30. However, for the malignant cases, most patients are 50 and above, and 70s - year - old patients are the most present.

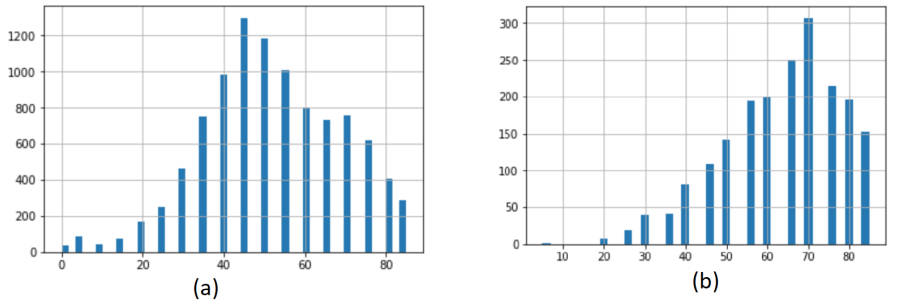


Fig. 8 (a) the age of the patients and (b) Number of malignant case with respect to the ages

We have more male patients than female patients in both the general population and in malignant cases as shown in graphs of Fig 9.

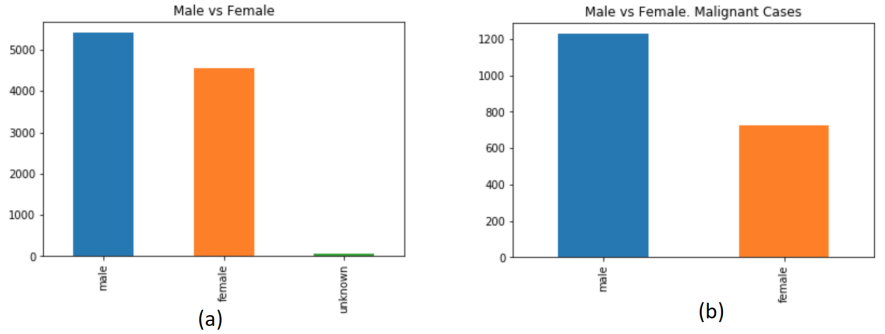


Fig. 9 (a) the gender of the patients and (b) Number of the malignant case with respect to the gender.

3.1.2 Architectures used for Lesion Classification

In order to estimate the difficulty of skin lesion classification task, we commenced our evaluations by building a simple baseline convolutional neutral network that consist of six layers(i.e. three convolutional layers and three max pooling layers). To avoid overfitting, several data augmentation methods has been applied on the training dataset. Before the output layers, a dropout of 0.5 has also been added which randomly inactivate neurons in a network so that it can generalized better on training dataset. Learning rate decay is also

scheduled during the model training such that the learning rate will reduced to after certain number of epochs. In order to fine-tune InceptionV3, the output fully-connected layers are removed, and new fully-connected layers are added at the top. These additional layers consisted of max pooling layers, 512 units of fully connected layer, 0.5 dropout and finally the output layer consisted the softmax activation for the classification of seven types of skin lesions diseases. Initially, all layers in network are frozen to perform feature extraction for the newly added fully connected layers so that the weights for these layers are not random initialed and the gradient would not be too large when we perform fine-tuning on our dataset. Once we trained the upper layers for feature extraction on HAM1000 dataset, the upper convolutional block of InceptionV3 is unfreeze for fine-tuning the network. Later, we have also experimented fine tuning all the layers of InceptionV3. Throughout the training process, learning rate of 0.001 and Adam optimizer are used. The same data augmentation and learning rate decay strategy as in baseline model is used. Another variant of InceptionV3 that is top performer on ImageNet is InceptionResNetV2. This network consisted of the Residual connection, which is proved to be inherently necessary for training very deep convolutional models. InceptionResNetV2 was fine-tuned for only for the top layers. The same training parameters were used for fine-tuning InceptionResNetV2 as used for InceptionV3.Finally DenseNet201 was also trained by both fine tuning the weights of only the output layers and also by fine tuning all the layers of the network. The details of dataset preparation, data augmentation and other optimization methods used during the experiments, are discussed in the following subsections.

3.1.3 Data Pre-processing

The original size of images in the HAM10000 dataset is too big i.e. 450 by 600. Hence, the images must be pre-processed before feeding to the networks for training. For the baseline model, the images are resized to 64 by 64 where as the images are resized to 192x256 for other deep learning models that are to be fine-tuned on the HAM10000 dataset.In addition, the image pixel values are also normalized by dividing the image pixels by 255. Finally, the dataset is divided into 72% (i.e. 7210) images for training, 18% (i.e.1803) for validation and 10% (i.e. 1002) for testing.

3.1.4 Data augmentation

Data augmentation is a technique to increase the variation in a dataset by applying transformations to the original images. It is often used when the training data is limited and also as an effective way of preventing overfitting. Data augmentation does not only help when we have limited data samples but also helps the model to generalize better by increasing variance in the dataset so that the same type of images are not shown to the model again and again during the training. By Varying the original images slightly using data augmentation methods through different types of transformation, the model is prevented from learning the training data so well that it will not overfit the

samples used while training the model. We have used different data augmentation techniques to increase the number of images in the training dataset by applying transformation on images such as shearing, resizing, zooming, left or right flipping, random cropping and random erasing as shown in Fig 10..

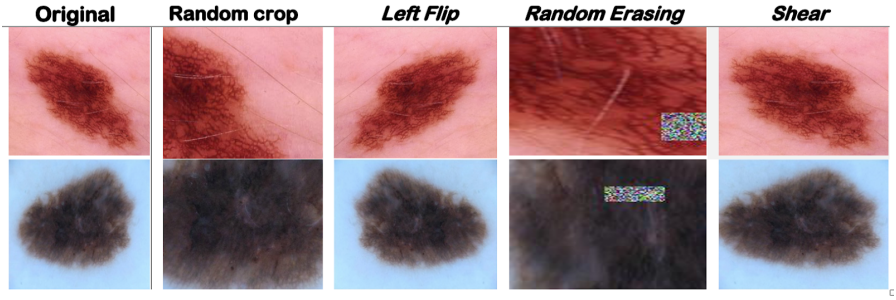
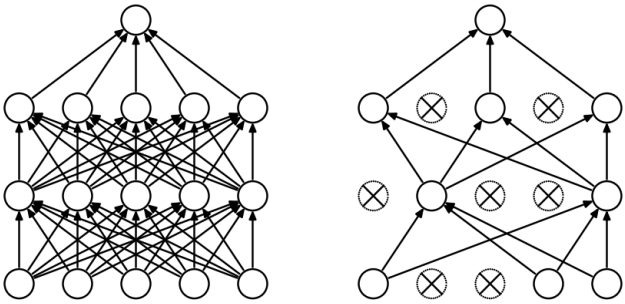


Fig. 10 The illustration of applied transformation on sample images that are used for training the models.

3.1.5 Dropout Regularization:

Usually when the training dataset is small, the model can easily overfit the training images which means that the model performs well on these images but loses its ability to predict unseen data images. There are several methods to address this problem and Dropout regularization is the most common and effective approach to tackle this issue. Hence, a dropout layer has also been introduced after the upper dense layer and before the output layer during our experiments. Dropout can effectively alleviate overfitting that blocks certain activation values of random neurons during training that temporarily blocks all incoming and outgoing connections to avoid overfitting, as shown in Fig 11..



(a) Standard Neural Net (b) After applying dropout.

Fig. 11 The term "dropout" refers to dropping out units (hidden and visible) in a neural network [8].

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         (None, 64, 64, 3)         0
_________________________________________________________________
conv2d (Conv2D)              (None, 64, 64, 16)        448
_________________________________________________________________
max_pooling2d (MaxPooling2D) (None, 32, 32, 16)        0
_________________________________________________________________
conv2d_1 (Conv2D)            (None, 32, 32, 32)        4640
_________________________________________________________________
max_pooling2d_1 (MaxPooling2 (None, 16, 16, 32)        0
_________________________________________________________________
conv2d_2 (Conv2D)            (None, 16, 16, 64)        18496
_________________________________________________________________
max_pooling2d_2 (MaxPooling2 (None, 8, 8, 64)          0
_________________________________________________________________
flatten (Flatten)            (None, 4096)              0
_________________________________________________________________
dense (Dense)                (None, 512)               2097664
_________________________________________________________________
dropout (Dropout)            (None, 512)               0
_________________________________________________________________
dense_1 (Dense)              (None, 7)                 3591
=================================================================
Total params: 2,124,839
Trainable params: 2,124,839
Non-trainable params: 0
_________________________________________________________________
```

Fig. 12 Baseline model architecture

3.1.6 Transfer learning

Deep learning models take days, or even weeks, to train a very large datasets. So a one way forward is to use transfer learning where features learned on the pretrained models that were previously trained on huge datasets, namely ImageNet, are taken and leveraged on a new and similar problem. The Lower level layers of these pretrained models are usually frozen to avoid destroying the generic features of the dataset they contain and add new layers on top of the frozen layers that are used for learning more specific features during the training on new dataset. The special characteristics of Deep CNN is that the initial layers of the network usually learn very generic features and "low-level" features of images, while the top layers of the network contain the semantics and high-level features. Therefore by fine-tuning deep Convolutional Neural Networks trained for image classification tasks on one dataset, can be reused for other image classification tasks with different dataset. Fine-tuning is one method of performing transfer learning where you change the model output to fit the new task and train only the output model. Consequently, fine-tuning has been used widely in computer vision and deep learning.

3.1.7 Baseline Model

In order to estimate the difficulty of the skin lesion classification task, we commenced our evaluations by building a simple baseline convolutional neural network that consist of six layers(i.e. three convolutional layers and three max pool-ing layers). The architecture of the baseline convolutional neural network showed in Fig 12.

To avoid overfitting, several data augmentation methods have been applied on the training dataset as discussed above. Before the output layers, a dropout of 0.5 has also been added which randomly inactivates neurons in a network so that it can generalize better on training dataset. Learning rate decay is also scheduled during the model training such that the learning rate will be reduced after a certain number of epochs.

3.2 Lesion Segmentation

Segmentation is the classification of each pixel in the image and in lesion segmentation we have to classify the pixels of image that whether they belong to a part of image of lesion (foreground) or part of image that is not lesion (background). Recently due to the advancements in deep learning techniques, segmentation has got quite mature. There has been significant work in the medical imaging based segmentation as well, that's why I have opted to go for deep learning based techniques in this project.

3.2.1 Double U-Net Architecture

U-Net is the encoder and decoder based Convolution Neural Network as seen in Fig 13. It takes an image as input, applies CNNs and max-pooling based operations on it and creates a one-dimensional vector in latent space, after that, it up samples the one dimensional vector to an output image, a segmented image. Architecture of Double-Net uses two modules of encoder-decoder based architectures as seen in Fig 14, output of first encoder-decoder module is the input of second encoder-decoder module architecture, and weights of encoder of first module is also shared with the decoder of second module.

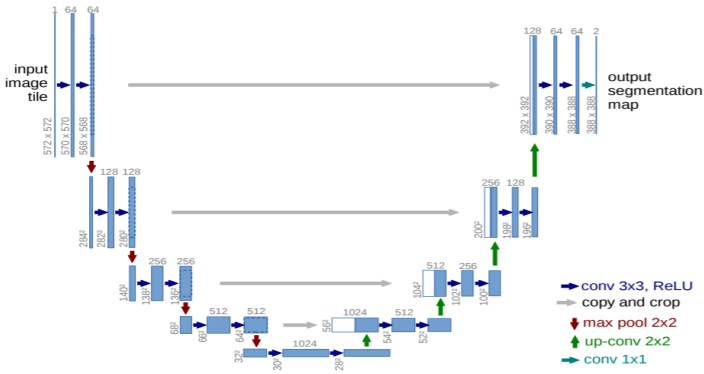


Fig. 13 Architecture of U-Net [16]

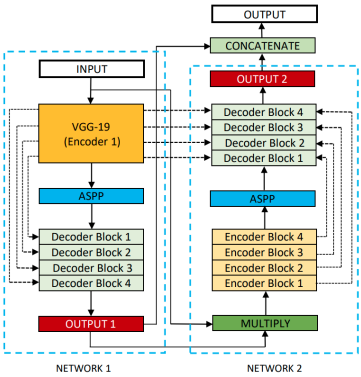


Fig. 14 Architecture of Double-Net [17]

3.2.2 Dataset for Segmentation

Dataset used to train the segmentation module of this project was ISIC 2018 [24] as this dataset includes labeled data for segmentation that is the mask of each lesion image which is not available in 2019 and 2020 datasets of ISIC. As it can be seen in Fig 15 in which I have included three examples from ISIC 2018 dataset, including its ground truth that is mask and its respective output from segmentation module. In Each example first patch is the ground truth mask, second one is input image and third one is output of segmentation module and it can be seen visually about good segmentation module is working, descriptive results of this model will be mentioned in the results part.

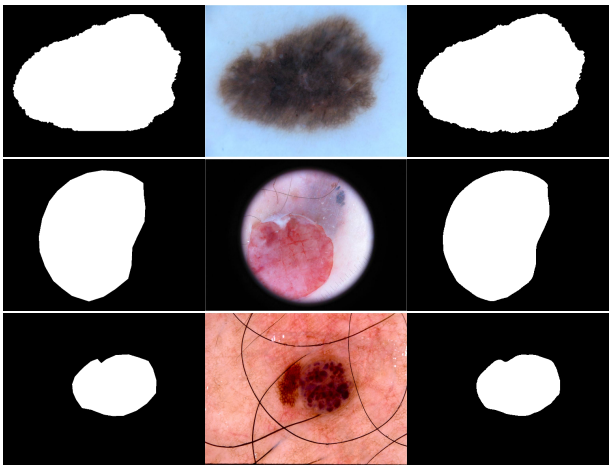


Fig. 15 ISIC 2018 Dataset: First and second images are mask and input image, third image is the output of Segmentation Module

ISIC 2018 dataset has 2595 training images, that were used to train the model but model on these images was not giving good results. As deep learning models are data hungry algorithms so I needed more data to train a better model that's why I opted data augmentation to augment the data. In augmentation either we can do offline augmentation or online augmentation. I opted for offline augmentation to reduce the overall training time. For augmentation I have used different types of techniques as they were already implemented in python library Albumentations [25]. These augmentation techniques are, Center Crop, Crop, Random Rotate, Transpose, Elastic Transpose, Grid Distortion, Optical Distortion, Vertical Flip, Horizontal Flip, Random Brightness Contrast, Random Gamma, Hue Saturation Value, RGB Shift, Random Brightness, Random Contrast, Motion Blur, Median Blur, Guassian Blur, Gauss Noise, Channel Shuffle and Coarse Dropout.

3.2.3 Hyperparameters

In this section, some important parameters used to train the model will be highlighted. First one is that, this model was trained on a virtual machine with V100 gpu. It was trained with input size of image of 192x256 pixels, with Adam optimizer available in tensorflow library. Loss function used to train this model was Dice Loss [26]. This model was trained for 100 epoch on augmented dataset mentioned previously.

3.3 Lesion Classification with Segmentation

Different skin lesion size can have an impact on the test accuracy and validation accuracy, it will happen when dataset size is limited. So, another way to consider to see the performance, investigating the impact of segmentation on classification task.Here the proposed method choose best performing classification architecture DenseNet chosen to apply on the segmented images, where the images are segmented using U-Net based architecture Double-Unet. Additionally, the proposed method applied segmentation with the Inception architecture as well, to get the comparison what can be the result of another model which utilized for classification task only. All the images gone through the pre-processing step before going through the classification pipeline. To see the comparative result, during the training instead of direct images, the proposed method trained the models on segmented images which helps to get more relevant feature from segmented images, but another thing can happen in terms of semantic segmentation, if there is any representative feature segmented as background which can show performance degradation.According to the performance indexes, the DenseNet method outperformed over Inception considering the methods with segmentation. But classification task on direct images have better performance in both of the methods then the method with the segmented images. Fig 16 shows the details workflow of the Classification with Segmentation.

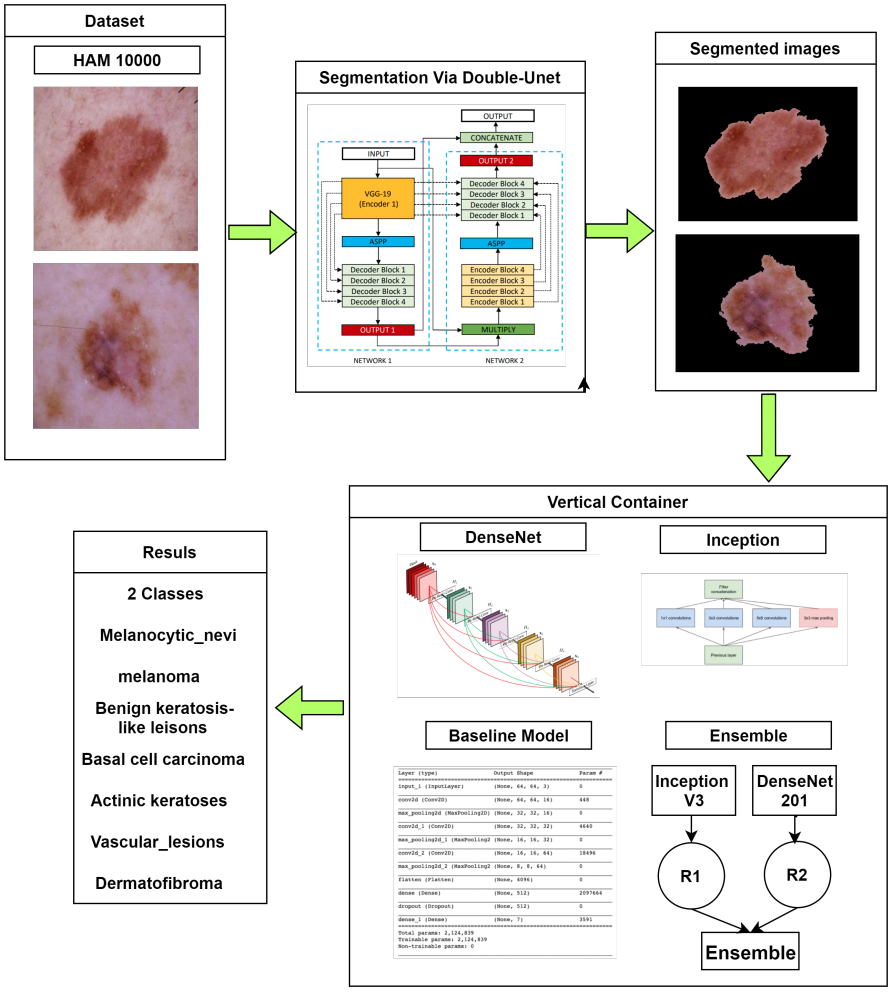


Fig. 16 Workflow diagram of segmentation with classification.

4 Result and Discussion

4.1 Evaluation Metrics

Sensitivity: This index is measured by the ratio of correctly identified patients having problems along with total of correctly identified patients having problems and incorrectly spotted as healthy. This is crucial in the medical image diagnosis to have better sensitivity, the ration leads to decreasing if the number of incorrectly spotted as healthy amount is increased, the patient will not get the treatment on time which is not expected then. This Sensitivity index

also used as precision in classification task.

$$Sensitivity = \frac{TP}{TP + FN} \tag{1}$$

Specificity: This index calculated by the ratio of actual healthy person and total amount of actual healthy person with incorrectly identified as patients. The index give low value when the amount of incorrectly identified as patients amount will increase which may leads to wrong treatment as healthy people having the medicine without having the disease. This Specificity index also used as recall in classification task.

$$Specificity = \frac{TN}{TN + FP} \tag{2}$$

Accuracy: This index determined by the ratio of total number of patients correctly and incorrectly identified verses the total number of correctly and incorrectly identified healthy and patients.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

Dice coefficient: The Dice coefficient which is also known as Dice similarity index measured using following formula.

$$Dice\ similarity\ index = \frac{2 * TP}{2 * TP + FP + FN} \tag{4}$$

4.2 Segmentation Results

Accuracy Metrics used for segmentation module are Dice coefficient (DSC), mean Intersection over Union (mIoU), Precision and Recall. In Table 2, there is a comparison of Double U-Net model with previous best models used for Medical Image segmentation with accuracy metrics.

Table 2 Performance & Comparison of Segmentation Methods with Double U-Net

Method	DSC	mIoU	Specificity	Sensitivity
FCN-VGG [27]	0.7023	0.5420	-	-
Mask R-CNN with Resnet [28]	0.7042	0.6124	-	-
U-Net [16]	0.2920	0.1759	0.5930	0.2021
Double U-Net	0.7649	0.6255	0.7156	0.8007

This model was also tested on ISIC 2020 Dataset as well, that is totally new dataset to model but we couldn't get any descriptive accuracy metrics as ground truth for segmentation is not available in ISIC 2020 Dataset, but

we can see visually how good this model is working on ISIC 2020 Dataset. In Fig 17, there are some images from ISIC 2020 dataset with predicted output of segmentation module.

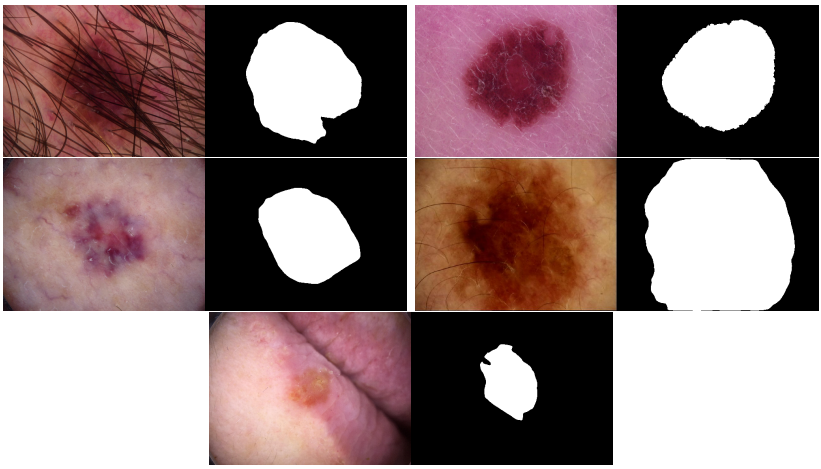


Fig. 17 ISIC 2020 Dataset: In each Image, first patch is the input image and second patch is the output of segmentation module

4.3 Classification Results

Table 3 Results obtained through fine tuning all layers of the models.

Model	*Validation Accuracy*	*Testing Accuracy*	*Testing Loss*	*Number of layers*	*Number of Parameters*
Baseline model	.75	.74	0.69	11 layers	2,124,839
Inception V3	.72	.77	0.7482	315 layers	22, 855, 463
DenseNet 201	***.75***	.69	0.691	711 layers	19, 309, 127
Ensemble	.76	.67	0.61	-	-

Table 4 Performance of different methods with segmentation and without segmentation images.

Method	**Testing accuracy**	**Testing loss**	**Precision**	**Recall**
Inception with segmentation	0.72	0.78	0.83	0.64
DenseNet with segmentation	0.76	0.66	0.84	0.68
Inception without segmentation	0.72	0.77	0.82	0.64
DenseNet without segmentation	0.76	0.68	0.83	0.69
Ensemble model without segmentation	0.76	0.67	0.86	0.67
Ensemble model with segmentation	0.75	0.67	0.85	0.66

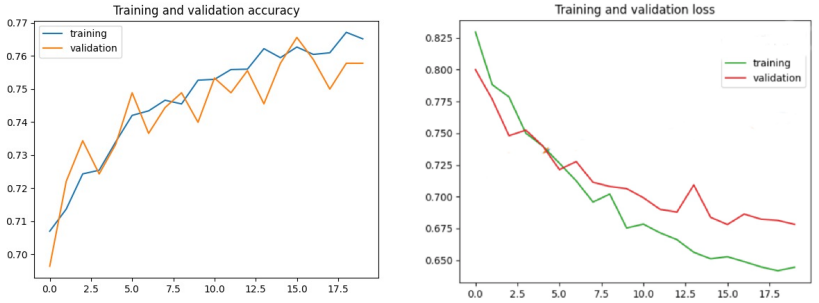


Fig. 18 Performance of DenseNet with segmentation

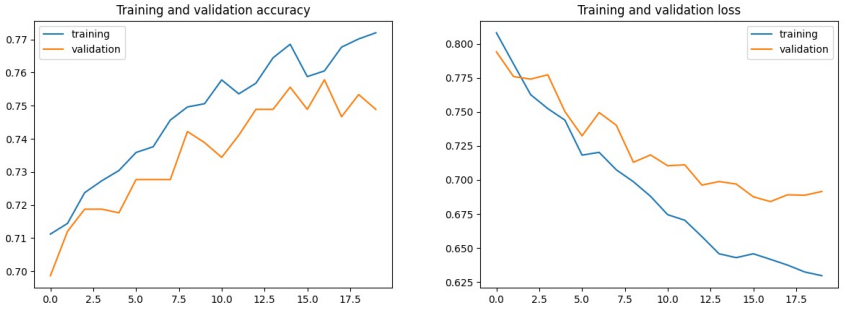


Fig. 19 Performance of DenseNet without segmentation

In this section, the performance of the deep neural networks on HAM10000 dataset has been evaluated. Several deep neural networks has been investigated using HAM1000 dataset and compared with the baseline model. The models involved in the performance evaluation are Baseline model, Inception V3 and DenseNet201. Each of these deep neural networks are fine-tuned such that the lower level layers were frozen to avoid destroying the generic features of the dataset they contain and add new layers on top of the frozen layers. The top layers are further trained on HAM10000 dataset. In addition, the experiment where we investigated the performance by fine-tuning the all layers are performed with Inception V3 and Dense Net 201 only. Finally, in order to build a single strong model, an ensemble of Inception V3 and DenseNet201 was created which has outperformed other deep learning architectures on HAM10000 dataset. The results with the number of layers and parameters are summarized in Table 3.

Fig 18 shows the graph of the traning and validation loss of DenseNet with Segmentation. Simontinously, Fig 19, Fig 20 , Fig 21 presents the graphs of the results of DenseNet without segmentation, Inception with segmentation, Inception without segmentation. And Fig 22, Fig 23 delivers the graph of the

accuracy performances of base model with segmentation , basemodel without segmentation.

Table 4 summarise the performance indexes of all the methods we applied for classification with segmentation and classification without segmentation.Densenet with segmentation and Ensemble model without segmentation have better accuracy then the other methods which archived 0.76 in accuracy. In terms of Precision Ensemble model without segmentation outperformed then other models which gained 0.86. DenseNet without segmentation outperformed in Recall index which achived 0.69 which is better then other methods encounterd. Overall, the methods applied on direct images rather then the segmented images performing better according to the performance indexes.

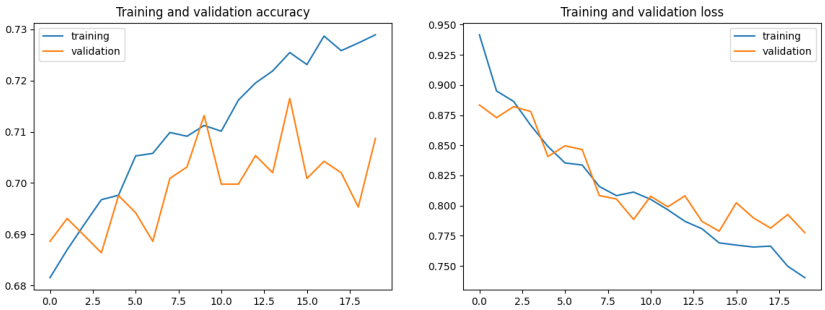


Fig. 20 Results of Inception with segmentation

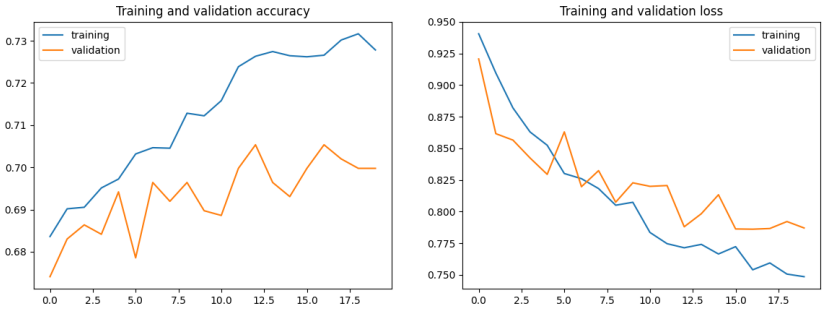


Fig. 21 Results of Inception without segmentation

5 Limitations

The skin lesion classification is a challenging task due the low inter-class and high intra-class variance which makes it hard to distinguish between classes.

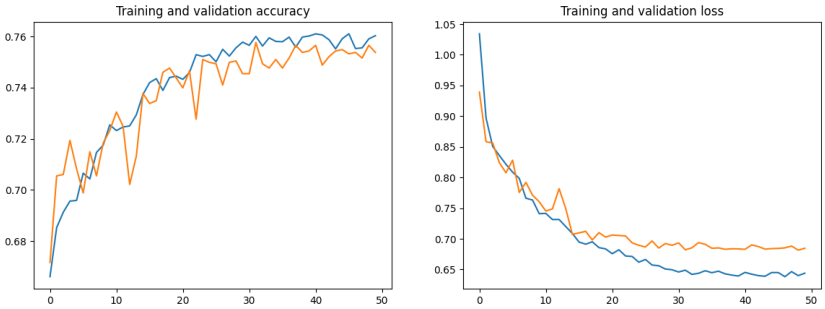


Fig. 22 Results of base model with segmentation

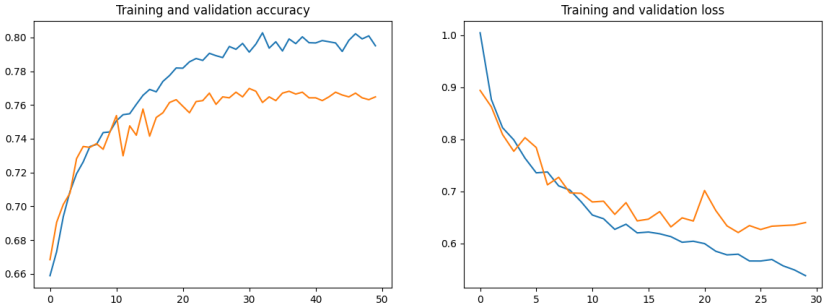


Fig. 23 Results of base model without segmentation

Moreover, the variety of illumination and viewing angles setup while taking the image also plays a vital role. However, these limiting conditions can be improved by collecting larger and more diverse datasets, generating augmented data samples and pre-processing the data to improve the poor lighting conditions.

6 Conclusion

In this work, two state-of-the-art networks and one ensemble models has been encountered that have achieved remarkable performance on image classification, have been investigated for the task of skin lesion classification into seven types of classes using the HAM10000 dataset. Initially, a baseline model has been trained on this dataset to evaluate the difficulty of skin lesion classification task and compare the performance of other deep learning models on the dataset. Transfer learning has been used for training these deep neural networks due to time constraint. In order to avoid overfitting and help the model generalized better, several data augmentation methods have been applied on the training dataset and other optimization techniques namely dropout and learning rate decay have also been investigated that has been successfully in

improving the model performance for skin lesion classification. The ensemble of InceptionV3 and DenseNet201 has outperformed. Intensive experimentation has also been performed to identify the role of segmentation in the skin lesion classification where ISIC 2018 dataset has been used for training the double-UNet architecture. This achieved 0.76 in Dice coefficient, 0.72 in specificity and 0.80 sensitivity. The overall results of experimentation revealed that the ensemble of InceptionV3 and DenseNet201 has outperformed by achieving 0.76 accuracy without segmentation and 0.75 with segmentation.

References

[1] Brinker, T.J., Hekler, A., Utikal, J.S., Grabe, N., Schadendorf, D., Klode, J., Berking, C., Steeb, T., Enk, A.H., Von Kalle, C.: Skin cancer classification using convolutional neural networks: systematic review. Journal of medical Internet research **20**(10), 11936 (2018)

[2] Al-Masni, M.A., Al-Antari, M.A., Choi, M.-T., Han, S.-M., Kim, T.-S.: Skin lesion segmentation in dermoscopy images via deep full resolution convolutional networks. Computer methods and programs in biomedicine **162**, 221–231 (2018)

[3] Barata, C., Celebi, M.E., Marques, J.S.: A Survey of Feature Extraction in Dermoscopy Image Analysis of Skin Cancer. IEEE J. Biomed. Health Inf. **23**(3), 1096–1109 (2019) 29994234. https://doi.org/10.1109/JBHI.2018.2845939

[4] Kareem, O., Mohsin Abdulazeez, A., Zeebaree, D.: Skin Lesions Classification Using Deep Learning Techniques: Review. ResearchGate, 1–22 (2021). https://doi.org/10.9734/AJRCOS/2021/v9i130210

[5] Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions - Scientific Data. Sci. Data **5**(180161), 1–9 (2018). https://doi.org/10.1038/sdata.2018.161

[6] Perez, F., Vasconcelos, C., Avila, S., Valle, E.: Data augmentation for skin lesion analysis. In: OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis, pp. 303–311. Springer, ??? (2018)

[7] Lei, B., Xia, Z., Jiang, F., Jiang, X., Ge, Z., Xu, Y., Qin, J., Chen, S., Wang, T., Wang, S.: Skin lesion segmentation via generative adversarial networks with dual discriminators. Medical Image Analysis **64**, 101716 (2020)

[8] Wibowo, A., Purnama, S.R., Wirawan, P.W., Rasyidi, H.: Lightweight

encoder-decoder model for automatic skin lesion segmentation. Informatics in Medicine Unlocked, 100640 (2021)

[9] Liu, L., Mou, L., Zhu, X.X., Mandal, M.: Skin lesion segmentation based on improved u-net. In: 2019 IEEE Canadian Conference of Electrical and Computer Engineering (CCECE), pp. 1–4 (2019). IEEE

[10] Al Nazi, Z., Abir, T.A.: Automatic skin lesion segmentation and melanoma detection: Transfer learning approach with u-net and dcnn-svm. In: Proceedings of International Joint Conference on Computational Intelligence, pp. 371–381 (2020). Springer

[11] Alqudah, A.M., Alquraan, H., Qasmieh, I.A.: Segmented and non-segmented skin lesions classification using transfer learning and adaptive moment learning rate technique using pretrained convolutional neural network. In: Journal of Biomimetics, Biomaterials and Biomedical Engineering, vol. 42, pp. 67–78 (2019). Trans Tech Publ

[12] Rahman, Z., Ami, A.M.: A transfer learning based approach for skin lesion classification from imbalanced data. In: 2020 11th International Conference on Electrical and Computer Engineering (ICECE), pp. 65–68 (2020). IEEE

[13] Rahman, Z., Hossain, M.S., Islam, M.R., Hasan, M.M., Hridhee, R.A.: An approach for multiclass skin lesion classification based on ensemble learning. Informatics in Medicine Unlocked **25**, 100659 (2021)

[14] Yao, P., Shen, S., Xu, M., Liu, P., Zhang, F., Xing, J., Shao, P., Kaffenberger, B., Xu, R.X.: Single Model Deep Learning on Imbalanced Small Datasets for Skin Lesion Classification. arXiv (2021) 2102.01284

[15] Kassem, M.A., Hosny, K.M., Fouad, M.M.: Skin lesions classification into eight classes for isic 2019 using deep convolutional neural network and transfer learning. IEEE Access **8**, 114822–114832 (2020). https://doi.org/10.1109/ACCESS.2020.3003890

[16] Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In: International Conference on Medical Image Computing and Computer-assisted Intervention, pp. 234–241 (2015). Springer

[17] Jha, D., Riegler, M.A., Johansen, D., Halvorsen, P., Johansen, H.D.: Doubleu-net: A deep convolutional neural network for medical image segmentation. In: 2020 IEEE 33rd International Symposium on Computer-based Medical Systems (CBMS), pp. 558–564 (2020). IEEE

[18] Szegedy, C., Ioffe, S., Vanhoucke, V., Alemi, A.: Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning. arXiv (2016) 1602.07261

[19] Iandola, F.N., Han, S., Moskewicz, M.W., Ashraf, K., Dally, W.J., Keutzer, K.: SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size. arXiv (2016) 1602.07360

[20] Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2261–2269 (2017). https://doi.org/10.1109/CVPR.2017.243

[21] Zhou, Z.-H.: Ensemble Learning. In: Encyclopedia of Biometrics, pp. 270–273. Springer, Boston, MA, USA (2009). https://doi.org/10.1007/978-0-387-73003-5

[22] Xie, Y., Zhang, J., Xia, Y., Shen, C.: A mutual bootstrapping model for automated skin lesion segmentation and classification. IEEE transactions on medical imaging **39**(7), 2482–2493 (2020)

[23] Rasul, M.F., Dey, N.K., Hashem, M.: A comparative study of neural network architectures for lesion segmentation and melanoma detection. In: 2020 IEEE Region 10 Symposium (TENSYMP), pp. 1572–1575 (2020). IEEE

[24] Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., Dusza, S., Gutman, D., Helba, B., Kalloo, A., Liopyris, K., Marchetti, M., et al.: Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic). arXiv preprint arXiv:1902.03368 (2019)

[25] Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A.A.: Albumentations: fast and flexible image augmentations. Information **11**(2), 125 (2020)

[26] Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., Cardoso, M.J.: Generalised dice overlap as a deep learning loss function for highly unbalanced segmentations. In: Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp. 240–248. Springer, ??? (2017)

[27] Brandao, P., Mazomenos, E., Ciuti, G., Caliò, R., Bianchi, F., Menciassi, A., Dario, P., Koulaouzidis, A., Arezzo, A., Stoyanov, D.: Fully convolutional neural networks for polyp segmentation in colonoscopy. In: Medical Imaging 2017: Computer-Aided Diagnosis, vol. 10134, p. 101340 (2017). International Society for Optics and Photonics

[28] Qadir, H.A., Shin, Y., Solhusvik, J., Bergsland, J., Aabakken, L., Balasingham, I.: Polyp detection and segmentation using mask r-cnn: Does a deeper feature extractor cnn always perform better? In: 2019 13th International Symposium on Medical Information and Communication Technology (ISMICT), pp. 1–6 (2019). IEEE