sciendo

VOLUME 16, N° 3 2022
Journal of Automation, Mobile Robotics and Intelligent Systems

# Skin Lesion Detection Using Deep Learning

Submitted: 12th May 2022; accepted: 28th July 2022

*Rajit Chandra, Mohammadreza Hajiarbabi*

DOI: 10.14313/JAMRIS/3-2022/24

**Abstract:**
*Skin lesion can be deadliest if not detected early. Early detection of skin lesion can save many lives. Artificial Intelligence and Machine learning is helping healthcare in many ways and so in the diagnosis of skin lesion. Computer aided diagnosis help clinicians in detecting the cancer. The study was conducted to classify the seven classes of skin lesion using very powerful convolutional neural networks. The two pre trained models i.e DenseNet and Incepton-v3 were employed to train the model and accuracy, precision, recall, f1score and ROC-AUC was calculated for every class prediction. Moreover, gradient class activation maps were also used to aid the clinicians in determining what are the regions of image that influence model to make a certain decision. These visualizations are used for explain ability of the model. Experiments showed that DenseNet performed better then Inception V3. Also it was noted that gradient class activation maps highlighted different regions for predicting same class. The main contribution was to introduce medical aided visualizations in lesion classification model that will help clinicians in understanding the decisions of the model. It will enhance the reliability of the model. Also, different optimizers were employed with both models to compare the accuracies.*

**Keywords:** *Skin lesion, DenseNet, Inception V3*

## 1. Introduction

Dermatologists use technological approaches for detecting skin cancer to facilitate in the early detection of skin cancer. Such lesions are produced by aberrant melanocyte cell formation and it usually happens when skin is exposed to sun more than necessary. Melanocytes cells generates "melanin". Melanin is the substance that is responsible for producing pigmentation in the skin. Moreover, the amount of skin cancer cases has risen dramatically, resulting in a growth in the mortality rate from the condition, notably from melanoma instances. That is why the skin lesion is a big concern in all over the world. Skin lesion has many different kinds, and some kinds if not detected early can become skin cancer and so it is important to detect this disease in the early stage. Like every other field, technology is also used in this area to facilitate clinicians and to contribute to human health. Machine learning is sub field of artificial intelligence and it is proved to outperform in various fields. With the enhancement in the computational power and the huge data availability, it became possible to use deep learning models. Deep learning models have the power to take in the complex structure of images and to learn the pattern out of it. The process in making the deep learning model includes collecting the data, pre-processing it, the image data is then segmented and features are extracted. These features are then fed into the model and probabilities are calculated. The class label having the highest probability is predicted. Data is the most important factor for machine learning algorithms. Experts uses various strategies to collect the data. The two types of images are used in medical AI, i.e. dermoscopic images and macroscopic images. For the study, the dataset provided by the International Skin Imaging Collaboration is used. The ISIC has provided various versions of the dataset. The ISIC-2018 dataset is used for the making the model. The 2018 archive contains seven different classes of skin lesion. So it was a multiclass classification problem. The images that are provided by ISIC are the dermoscopic images of the lesion. Convolutional Neural Networks are neural networks that are primarily used for the computer vision tasks. The reason is that CNNs are able to understand the complex structure of images.

Dermoscopy is the state-of-the-art procedure for skin cancer screening, with a diagnosis accuracy that is higher than the naked eye [2]. In this paper, the researchers offered a method for improving the accuracy of automated skin lesion identification by combining different imaging modalities with the metadata of patients. Only those cases were kept that had metadata of patients, a macroscopic image, a dermatoscopic image, and a histological diagnosis details. Moreover, only instances where input images are of adequate quality and untainted by any identifying traits the were picked by repeated hand scanning of all images (ie, eyes, facial landmarks, jewellery or garment). ResNet-50 was used to extract the features of the images. Three kinds of experiments were conducted.

### 1.1. Full Multimodality Classification

When all mentioned three modes (macroscopic image of lesions, dermatoscopic images, and metadata of patients) were provided, the researchers built a network

with two image feature extractions, one for dermatoscopic input images and the other for macroscopic input images.

### 1.2. Partial multimodality classification

The researchers excluded the other two from the complete network when only one image modality (macroscopic images or dermatoscopic images) and information were supplied for classifying the images. Before passing it through the embedding network, the researchers generated only one feature vector of image and combined it with the feature vector of metadata.

### 1.3. Single image classification

When there was only one image type for classification and there was no metadata, the image was sent through the image feature extraction network, and the extracted features were then transmitted via the network. In the testing phase, it came out that the metadata variables of patients like age, sex and location did not enhance precision for pigmented skin lesions appreciably. As a result, it was concluded that available models rely substantially on tight image criteria and may be unstable in clinical practice. Furthermore, selecting datasets may contain unintended biases for specific input patterns.

Using image representations produced from Google's Inception-v3 model, the proposed automated approach intends to detect the kind and cause of cancer directly [3]. The researchers used a feed forward neural network having two layers with softmax activation function in the output layer to perform two-phase classification based on the representation vector. Two separate neural networks with the same representation vector were used to perform the two-phase classification. In phase one, the researchers determined the type of cancer, whether it was malignant or benign, and in phase two, the researchers determined whether the cancer was caused by melanocytic or nonmelanocytic cells. The training dataset includes 2000 JPEG dermoscopic images of skin lesions, as well as ground truth values. The validation set had 150 photos, whereas the testing set contained 600. The method identifies the images automatically using Google's inspection model and the image representation produced from the dermoscopic images.

This paper had two major contributions: first, the researchers offered a classification model that used Deep Convolutional Neural Network and Augmentation of data to evaluate the classification of skin lesion images [4]. Second, the researchers showed how data augmentation could be used to overcome data scarcity, and the researchers looked at how varying numbers of augmented data samples affect the performance of different models. The researchers used three methods of data augmentation in melanoma classification.

### 1.4. Geometric augmentation

The semantic interpretation of the skin lesion is preserved by the position and scale of lesion mark within the image; therefore, its ultimate classification is unaffected. As a result, input images were randomly cropped and horizontal and vertical flips were used to produce new samples under the same label as the original.

### 1.5. Color augmentation

The images of skin lesions were gathered from various sources and made using various devices. As a result, while using photographs for training and testing any system, it is critical to scale the colors of the images to increase the classification system's performance.

### 1.6. Data warping based on the knowledge of specialist

The clinicians diagnose the melanoma by seeing the patterns that surrounds the lesion. So, affine transformations including distorting, shearing and scaling the data can be helpful in classifying the images. As a result, warping is an excellent way to supplement data in order to improve performance and reduce overfitting in melanoma classification.

In [5] three classifiers named SVM, Random forests and Neural Networks were used to classify the image dataset. The results showed that different augmentations performed differently in this case. The neural networks performed best for classification task.

In image recognition nowadays, two basic types of feature sets are routinely used [5]. The traditional kind is based on what are known as "hand-crafted features", which are created by academics with the goal of capturing visual aspects of a picture, such as texture or color. A new sort of feature set was just presented that was motivated by how brain decode images and derived from powerful Convolutional Neural Networks. These new features beat "hand-crafted" features when combined with deep learning, and as a result, they are increasingly popular in computer vision. The researchers proposed in this study to utilise a mix of both sorts of features to classify skin lesions. "RSurf features" was extracted by the researchers for image description. This feature set's concept is to divide the input image into "parallel sequences of intensity values from the upper-left corner to the bottom-right corner". The concept behind such extraction technique is based on the texture unit model, in which an input image's texture spectrum is defined. The support vector machine with Gaussian kernel and standardized models was used in the first categorization. It estimated the class for a given input image using RSurf features and LBPR=1,3,5. CNN characteristics were used in the second SVM classifier, which had a Gaussian kernel and standardized predictors. The researchers used the AlexNet to extract the features. The researchers chose the label with the greatest absolute score value for each image that was tested. As a result, the final classifier incorporated both approaches, including hand-crafted characteristics as well as features acquired from the deep learning method.

It's critical to distinguish malignant form of skin lesions from benign form of lesions like "seborrheic

keratosis" or "benign nevi", and good computerized classification of skin lesion imagess can help with diagnosis [6]accurate discrimination of malignant skin lesions from benign lesions such as seborrheic keratoses or benign nevi is crucial, while accurate computerised classification of skin lesion images is of great interest to support diagnosis. In this paper, we propose a fully automatic computerised method to classify skin lesions from dermoscopic images. Our approach is based on a novel ensemble scheme for convolutional neural networks (CNNs. The researchers offer a completely automated method for classifying skin lesions from dermoscopic pictures in this study. For tasks like object detection and natural picture categorization, deep neural network algorithm, particularly convolutional neural networks, outperformed alternative methods. The well-established CNN architectures were used to attain great accuracy. Transfer learning had been applied in medical field for other tasks too. The pipeline of the model includes the data pre-processing, fine-tuning of neural networks and then the features were extracted, these features were fed into the SVM model. Then the outputs of the model were assembled together. To facilitate improved generalization ability when tested on additional datasets, the researchers kept the data pre-processing minimum in suggested pipeline. Only one task-specific pre-processing step (related to skin lesion categorization) was included in the technique, while the rest were typical pre-processing stages to prepare the pictures before fed them to model. Normalization, resizing, and color standardization were employed. VGG16, which included 16 weight layers, the number of convolutional layers were 13, and 3 FC layers were employed. In addition to vgg16, the powerful ResNet-18 and ResNet-101, which have varying depths, were used for extracting the features. To solve the three class classification (Malignat Melanoma /Sabrohtic Kerosis/ benign nevi) classification, the 190 final fully connected layers and the last layer which was output layer of all pre-trained networks were eliminated and replaced by two new fully connected layers of 64 nodes and 3 nodes. The new fully connected layers' weights were chosen at random using a normal distribution with average value of zero and a standard deviation of [195 0.01]. The researchers froze the weight values of the earliest layers of the deep models. By freezing the weights, the issue of overfitting was addressed. Also freezing the weights can be helpful in decreasing the training time. The researchers froze the early layers up to the $4^{th}$ layers and up to the $10^{th}$ layers for AlexNet and VGG16, respectively, and up to the $4^{th}$ residual block and $30^{th}$ residual blocks for ResNet-18 and ResNet-101 respectively. To avoid overfitting of the little training dataset, the researchers used data augmentation to boost the training size artificially. As key data augmentation approaches, the researchers used rotation of 90 degrees, 180 degrees and 270 degrees and they also employed horizontal flipping. A ternary SVM classifier was trained using the collected deep features and the related labels defining the lesion kinds. The researchers examined linear kernel as well as radial basis function (RBF) kernels and found that the RBF kernel performed marginally better. In the final models, the researchers used 265 one-vs-all multiclass SVM classifiers with radial basis function kernels. The major participation of the method is that it proposed a hybrid deep neural network method for classifying the skin lesion that extracted deep features from data images using multiple DNNs 395 and assembles features in a support vector machine classifier that produced very accurate results without needing exhaustive pre-processing or lesion area segmentation. The results demonstrated that combining information in this way improves discrimination and is complimentary to the 525 individual networks.

The "attention residual learning convolutional neural network (ARL-CNN)" model for skin lesion categorization is proposed in this research[7]. The researchers combined a residual learning framework for training a deep convolutional neural network with a small number of data images with an attention learning mechanism to improve the DCNN's particular representation capacity by allowing it to object more on "semantically" important regions of dermoscopy images (i.e. lesions). The suggested attention learning mechanism made full usage classification--trained DCNNs' innate and impressive self-attention capacity, and it could work under any deep convolutional neural network framework without appending any additional "attention" layers, which was important for the learning problems having small dataset as in the problem in hand for classifying the images. In terms of implementing this technique, each so-called ARL block might include both "residual learning" and "attention learning". By stacking numerous ARL blocks and training the model end-to-end, an ARLCNN model with any depth could be created. The researchers tested the suggested ARLCNN model using the ISIC-skin 2017 dataset, and it outperformed the competition. The research contributed in many aspects. The researchers proposed a novel ARLCNN model for accurate skin lesion categorization, which incorporates both residual learning and attention learning methods. The researchers created an effective attention framework that took full advantage of DCNNs' inherent "self-attention" ability, i.e., instead of learning the attention mask with extra layers, the researchers used the feature maps acquired by upper layer as the attention mask of a lower level layer; and the researchers achieved "state-of-the-art" lesion classification accuracy on the ISIC-skin 2017 dataset by using only one model with 50 layers, which was foremost for CAD of skin cancer.

Researchers addressed two problems in the paper. The first task entailed classifying skin lesions using dermoscopic pictures. "Dermoscopic" images and the metadata of patients were used for the second task [1]. For the first job, the researchers use a variety of CNNs to classify dermoscopic images. The deep learning models for task 2 are divided into two sections:

a convolutional neural network for dermoscopy images and a "dense neural network" for processing the patients' metadata. In the beginning, the researchers just trained the convolutional neural network on image data (task 1). The weight values of CNN are then frozen, and the metadata neural network is attached. Only the weights of the metadata neural network and the classification layer are trained in the second step. The researchers rely heavily on EfficientNets (EN), which were pre-trained on a very large dataset called ImageNet. These models consist of eight separate models that are architecturally similar and follow particular principles for adjusting the image size if it is larger. The version B0 which is also smallest of all, uses [224 *224] as the input size. In bigger versions, up to B7, the input size is raised while the network breadth and network depth are scaled up. The researchers use efficient net versions of B0 to B6. The researchers also trained SENet154 and the two versions of powerful ResNet for the training.

In developing the model, three optimizers were used to compare the results. The following optimizers were used

1. Stochastic gradient descent
2. RMSprop
3. Adam

#### 1.6.1. Stochastic gradient descent

It is an 'iterative method' that optimizes the loss function with differentiable properties. The goal of machine learning is to optimize the loss function or objective function. Mathematically,

$$Q(w)=\frac{1}{n}\sum_{i-n}^{n} Q_i(w)$$

Here "w" is estimated which minimizes Q. Because it is the iterative method so it performs following iterations to minimize the objective function.

$$w := w - \eta \nabla Q(w) = w - \frac{\eta}{n}\sum_{i-n}^{n} \nabla Q_i(w)$$

$\eta$ is learning rate.

#### 1.6.2. RMSProp

Root mean square propagation is also an optimization algorithm in which learning rate is adjusted for parameters. The 'running average' is calculated as follows:

$$\text{ÐÐ}(\text{ÐÐ}, ) := \text{gn}( \ , -1) + (1-\text{g})(\nabla \ _i( \ ))^2$$

The learning parameters are updated as follows:

$$w := w - \frac{\eta}{\sqrt{n(w,t)}} \nabla Q_i(w)$$

#### 1.6.3. Adam

It is an optimization algorithm that is used in place of the standard stochastic gradient descent process to iteratively update weights in neural network using training data. Diederik Kingma of "OpenAI" and Jimmy Ba of the "University of Toronto" presented Adam in their 2015 ICLR paper (poster) titled "Adam: A Stochastic Optimization Method."

Adam, the authors explain, integrates the benefits of two stochastic gradient descent enhancements. More precisely, an "Adaptive Gradient Algorithm" (AdaGrad) is responsible for managing the per-parameter learning rate and hence increases the efficiency on issues with sparse gradients (e.g. computer vision problems and natural language processing problems).

To experiment the skin lesion classification model, Python 3.6 were used as progemming languaue. Tensorflow and Keras were used for frameworks.

## 2. Methods

### 2.1. Method 1

The model was trained from scratch; the framework was rained for epochs after being initialised with random weights. The algorithm learnt attributes from input and calculates weights by backpropagation after every epoch. If the dataset is not very large, this strategy is unlikely to yield the most accurate results. However, it can still be used as a comparison point for the two other methods.

### 2.2. Method 2

For the second experiment, ConvNet were used as a feature extractor because most dermatological datasets have a small number of photos of skin lesions, this method used the weights from the available pre trained model VGG16 which was trained on a bigger dataset (i.e. ImageNet), this practice is titled as "transfer learning". This pre-trained model had previously learnt features that could be relevant for the classifying the skin lesion images, it is the core idea underpinning transfer learning.

### 2.3. Method 3

Another frequent transfer learning strategy entails not only training the model by assigning pre-trained weights, but also fine-tuning the model by solely training the upper layers of the convolutional network and using the backpropagation. The researchers recommended freezing the lower layers of the network in this paper since they contain more generic dataset properties. Because of their ability to extract more particular features, they were mainly interested in training the model's top layers. The parameters from the ImageNet dataset were used to initialise the first four layers of convolution neural network in the final framework in this method. The model weights that were saved was loaded from the matching convolutional layer in Method 1 were used to initialise the fifth and final convolutional block. The evaluation metrics showed that the third method performed better than Method 1 and Method 2.

## 3. Results

The data was divided into train, validation and test split.

| Train set images | Validation set images | Test set images |
|---|---|---|
| 9714 | 100 | 201 |

The training set was augmented with the images generated by introducing the changes into original dataset. The images were horizontally flipped, the rotation range was 90 degrees and the zoom range was kept 0.2. the images were also rescaled before feeding into the model.

### 3.1. Evaluation Metrics

Following evaluation metrics were used to evaluate the models.

The Receiver Operator Characteristic (ROC) curve is metric that is used to evaluate the classification models of machine learning. It presents a probability curve that plots the true positive rate against false positive rates at many threshold values. It basically distinct the 'signal' from the 'noise'. The formula of true positive rate and false positive rate are as follows:

$$\text{True positive rate} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}}$$

$$\text{ÐÐÐÐÐÐÐÐ} = \frac{\text{false positive}}{\text{false positive} + \text{true negative}}$$

The Area Under the Curve (AUC) measures the performance of the classifier by evaluating its ability to differentiate between classes. It is utilized as the summary of Receiver Operator Characteristic (ROC) curve. The higher value of AUC means that the classification model is performing accurately in differentiating the negative and positive classes.

Accuracy is also an evaluation metric that is used for evaluation of classification models. The accuracy value represents the fraction of predictions that model predicts correctly. The formula of accuracy is:

$$\text{Accuracy} = \frac{\text{total number of correct predictions}}{\text{total predictions}}$$

Precision indicates the fraction of positive predictions that were actually correct. The formula of precision is

$$\text{Precision} = \frac{\text{true positive}}{\text{true positive} + \text{false positive}}$$

Recall indicates fraction of actual positives that were predicted correctly.

$$\text{Recall} = \frac{\text{true positive}}{\text{true positive} + \text{false negative}}$$

It shows the balance between recall and precision. The formula of F1 Score is as follows:

$$\text{F1 Score} = \frac{2 * (\text{precision} * \text{recall})}{\text{precision} + \text{recall}}$$

### 3.2. L2 Regularization

L2 regularization is applied to models to combat overfitting. Overfitting is a term used to describe a situation where training loss decreases but the validation loss increases. In other words, the model is well fitted on training data but it is not predicting accurately for validation data. The model is not able to generalize. This is serious because If model is not generalizing then it will not produce accurate results when it will be implemented in real world scenario. There are different techniques that can be used to control overfitting. Regularization is used to control the complexity of model. When regularization is added, the model not only minimize the loss, but it also minimizes the complexity of model. So, the goal of machine learning model after adding regularization is,

$$\text{minimize}(\text{Loss}(\text{Data}|\text{Model})) + \text{complexity}(\text{Model}))$$

The complexity of the models used in paper was minimized by using L2 regularization. The formula of L2 regularization is the sum of square of all the weights,

$$L_2 \text{ regularization term} = \|w\|_2^2 = w_1^2 + w_2^2 + \ldots + w_n^2$$

In the models, two layers of L2 regularization was used before the final softmax layer.

A total of 12 experiments were conducted by using different optimizers. The three optimizers Adam, RMSprop, Stochastic Gradient Descent were used in DenseNet and inception v3. Moreover, experiments were conducted with augmentations and without augmentations to see whether the augmentations are useful in our case or not. The details of the experiments are given below

#### 3.2.1. With Augmentation

Different augmentations were applied to the dataset to increase the image data to avoid overfitting. If the model is trained on less data, it will learn the pattern but will not generalize it. In other words, the training accuracy is more than testing accuracy. The model does not generalize for unseen data. Different augmentations i.e. rotation range, horizontal flip and zoom range was applied on the dataset. Six experiments were performed with augmentations.

1. DenseNet [RMSPROP]
2. DenseNet [ADAM]
3. DenseNet [SGD]
4. Inception v3 [RMSPROP]
5. Inception V3 [ADAM]
6. Inception V3 [SGD]

### 3.2.2. Without Augmentation

These experiments were also conducted without augmentations to see if the model can generalize well without augmentations.

1. DenseNet [RMSPROP]
2. DenseNet [ADAM]
3. DenseNet [SGD]
4. Inception v3 [RMSPROP]
5. Inception V3 [ADAM]
6. Inception V3 [SGD]

## 4. Discussion

Early detection of skin lesion can save many lives and Artificial Intelligence is helping the medical science in serving this purpose. Convolutional Neural Networks are useful in medical imaging. The two state of the art architectures of convolutional neural network were experimented in this paper and they both showed good results overall. It turned out that DenseNet performed better then Inception V3 in classifying the images into different classes. In order to evaluate the model performance, AUC-ROC curves, precision, recall, F1 score and accuracy were employed. The reason of choosing multiple metrics was that the data was highly imbalance. So, accuracy metric alone might be a deceiving metric. The data imbalance issue was resolved by using focal loss. The per class ROC curves of classes in the DenseNet model are better than the Inception V3 model. Also the overall accuracy, precision, recall and F1 Score figures are better in DenseNet model. The models were run for 60 epochs and early stopping criteria was applied. The reason of applying early stopping was to ensure that model does not overfit. If the model is trained on too many epochs, there are chances that model will overlearn the pattern. And if the model is run for few epochs, the model can underfit i.e. it won't learn the pattern completely. Since number of epochs is a hyperparamter, so it has to be tuned. Normally, the model is run with huge number of epochs and when it stops learning, it is stopped. In keras, the early stopping callback is provided and that was used in experiments. In the result tables, termination epoch is also provided. The purpose of mentioning termination epoch was to see which optimizer converge on what epoch. The idea was to see that which optimizer converge relatively fast. In Dense Net model, Adam converged on $39^{th}$ epoch and gave accuracy of 79% but stochastic gradient descent converged on $35^{th}$ epoch and was 81% accurate. It means that stochastic gradient descent performed better in both perspectives. It gave higher accuracy with less epochs. In the experiments where augmentations were not applied, the accuracies were comparatively better than experiments with augmentations. But the experiments without augmentations faced overfitting problem. this is because the data was very less and the model learnt the training data but did not generalize well on testing data. The purpose of applying augmentations in deep learning is to increase the data because deep learning models requires huge data to learn. The training accuracies of experiments without augmentations were more than 90%. Although L2 regularization were also applied to overcome the issue of overfitting. In case of Inception V3, very interesting figures were produced. Adam optimizer achieved 75% test accuracy in 22 epochs while stochastic gradient descent produced same accuracy in 60 epochs. Moreover, the RMSprop optimizer produced 76% accuracy in 30 epochs. So for the given problem, stochastic gradient descent optimizer with inception V3 is not a suitable choice. The experiments without augmentations showed that RMSprop is a better choice. It gave 81% accuracy in 38 epochs. While Adam and SGD run for same number of epochs and gave 80% and 79% accuracies respectively. Another interesting thing was to see the per class AUC-ROC of Dermatofibroma class. It showed AUC-ROC around 60% in experiments without augmentations. And in experiments with augmentations, it showed AUC-ROC scores around 70%. While this was not the pattern in DenseNet experiments. All the AUC-ROC scores are around 90%. It shows that Inception V3 architecture did not learn the pattern of Dermatofibroma class very efficiently.

The loss function that was used for experiments was focal loss which performed well. It was used to overcome the class imbalance issue. In deep learning, it is important to have equal distribution of the classes. If data entries of one class are more than others, the model will learn efficiently the class with more examples. And when the model is deployed, it predicts every image belong to that class. The data was highly imbalance. There are multiple ways to solve this issue. One method is to use weighted loss. But recently, another loss function as introduced called focal loss. it focuses the class with few examples more than the class with more number of examples. It showed good performance overall. In the given problem, the Vascular class had very few examples in training dataset. focal loss focused on this class and on test dataset almost all experiments accurately classified the Vascular class.

The accuracies are better in DenseNet then Inception V3. Moreover, the grad activation maps show that the two models have seen different places to classify the same image. The focus region of inception V3 is different from the focus region of DenseNet. Inception V3 model misclassified Vascular class as it is shown in figure. While we cannot know from grad activation maps the reason of focusing the certain region, this is the black box to understand. But these visualizations can help medical staff in knowing that why the model is predicting the certain image to belong to certain class. Because the explainability of the machine learning models is important especially in the sensitive area of medical science. It will help medical staff to understand the model prediction without knowing much about artificial intelligence, machine learning and convolutional neural networks.

## 5. Future Work

In future the focus would be to improve the model accuracy by experimenting other models like AlexNet and vgg-16. The accuracy of the models will be compared and the best accurate model will be chosen. Also, the skin lesion follows a certain hierarchy that can be incorporated in future research. The hierarchy of skin lesion goes like:

In this paper, the seven classes from the third level are incorporated. Total of eight classes belongs to the third level but in the dataset of skin lesion 2018, the seven classes are given. In future the focus would be to consider the complete hierarchy. In the first stage, the first level will be classified, in second phase, the second level will be classified and in the third level all the seven classes will be classified by the model.

## 6. Figures and Tables

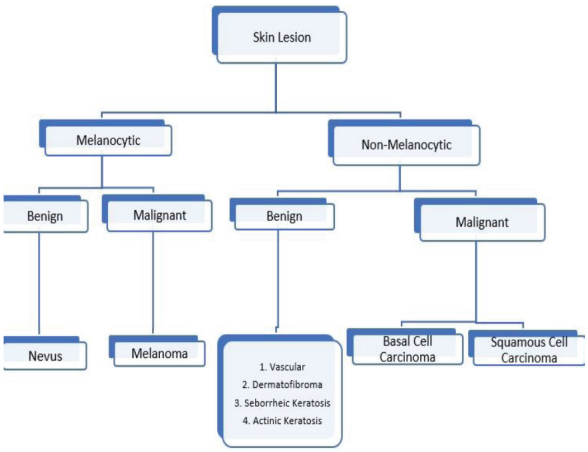


**Fig. 1.** Skin Lesion Hierarchy

### DenseNet model:

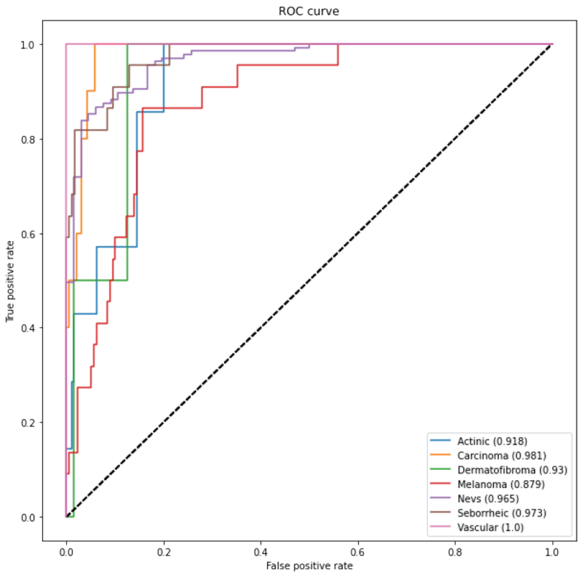


**Fig. 2.** ROC Curve for RMSProp

The per class AUC-ROC is highly accurate. The results of other experiments are following,

**Tab. 1.** DenseNet Comparison Table

| Optimizer | Accuracy | Precision | Recall | F1-SCORE | Termination epoch # |
|---|---|---|---|---|---|
| Adam | 0.79 | 0.82 | 0.79 | 0.79 | 39 |
| RMSProp | 0.80 | 0.80 | 0.80 | 79 | 35 |
| SGD | 0.81 | 0.82 | 0.81 | 0.81 | 34 |

**Tab. 2.** Per class AUC-ROC [DenseNet, RMS Prop, focal Loss, with Augmentations]

| Class | AUC-ROC |
|---|---|
| Actinic | 0.957 |
| Carcinoma | 0.98 |
| Dermatofibroma | 0.985 |
| Melanoma | 0.921 |
| Nevs | 0.962 |
| Seborrheic | 0.958 |
| Vascular | 1.0 |

**Tab. 3.** Per class AUC-ROC [DenseNet, SGD, focal Loss, with Augmentations]

| Class | AUC-ROC |
|---|---|
| Actinic | 0.918 |
| Carcinoma | 0.981 |
| Dermatofibroma | 0.93 |
| Melanoma | 0.879 |
| Nevs | 0.965 |
| Seborrheic | 0.973 |
| Vascular | 1.0 |

**Tab. 4.** Focal Loss – Without Augmentation, [DenseNet]

| Class | AUC-ROC |
|---|---|
| Actinic | 0.971 |
| Carcinoma | 0.977 |
| Dermatofibroma | 0.915 |
| Melanoma | 0.864 |
| Nevs | 0.959 |
| Seborrheic | 0.945 |
| Vascular | 1.0 |

**Tab. 5.** DenseNet without Augmentation

| Optimizer | Accuracy | Precision | Recall | F1-Score | Termination epoch # |
|---|---|---|---|---|---|
| Adam | 0.81 | 0.79 | 0.81 | 0.80 | 38 |
| RMSprop | 0.82 | 0.82 | 0.82 | 0.82 | 38 |
| SGD | 0.81 | 0.80 | 0.81 | 0.80 | 29 |

**Tab. 6.** Per class AUC-ROC [DeseNet, Adam, focal Loss, without Augmentations]

| Class | AUC-ROC |
|---|---|
| Actinic | 0.965 |
| Carcinoma | 0.979 |
| Dermatofibroma | 0.869 |
| Melanoma | 0.924 |
| Nevs | 0.94 |
| Seborrheic | 0.957 |
| Vascular | 1.0 |

**Tab. 7.** Per class AUC-ROC [DenseNet, RMSProp , focal Loss, without Augmentations]

| Class | AUC-ROC |
|---|---|
| Actinic | 0.946 |
| Carcinoma | 0.986 |
| Dermatofibroma | 0.982 |
| Melanoma | 0.905 |
| Nevs | 0.96 |
| Seborrheic | 0.956 |
| Vascular | 1.0 |

**Tab. 8.** Per class AUC-ROC [DenseNet, SGD , focal Loss, without Augmentations]

| Class | AUC-ROC |
|---|---|
| Actinic | 0.944 |
| Carcinoma | 0.975 |
| Dermatofibroma | 0.975 |
| Melanoma | 0.928 |
| Nevs | 0.958 |
| Seborrheic | 0.964 |
| Vascular | 1.0 |

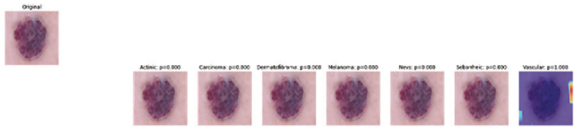

**Fig. 3.** Grad-CAM of DenseNet model

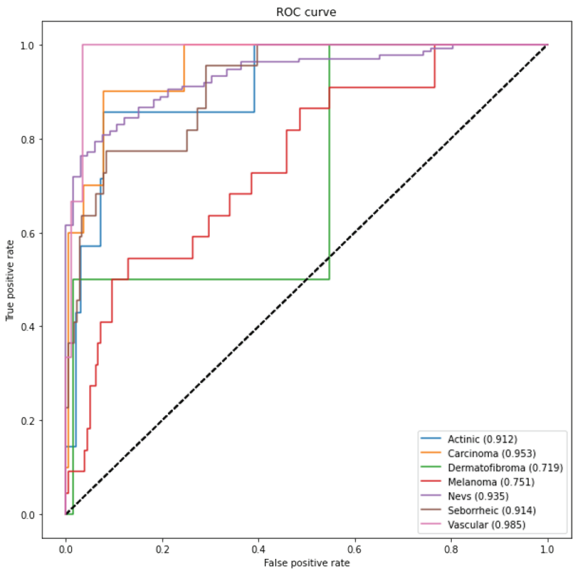


**Fig. 4.** Focal Loss – With Augmentations, [Inception v3]

**Tab. 9.** Inception V3 comparison table

| Optimizer | Accuracy | Precision | Recall | F1-Score | Termination epoch # |
|---|---|---|---|---|---|
| Adam | 0.75 | 0.78 | 0.75 | 0.75 | 22 |
| RMSprop | 0.76 | 0.71 | 0.76 | 0.73 | 30 |
| SGD | 0.75 | 0.74 | 0.75 | 0.74 | 60 |

**Tab. 10.** Per class AUC-ROC [Inception Adam, focal Loss, with Augmentations]

| Class | AUC-ROC |
|---|---|
| Actinic | 0.887 |
| Carcinoma | 0.959 |
| Dermatofibroma | 0.859 |
| Melanoma | 0.791 |
| Nevs | 0.92 |
| Seborrheic | 0.911 |
| Vascular | 0.99 |

**Tab. 11.** Per class AUC-ROC [Inception RMSprop, focal Loss, with Augmentations]

| Class | AUC-ROC |
|---|---|
| Actinic | 0.912 |
| Carcinoma | 0.953 |
| Dermatofibroma | 0.719 |
| Melanoma | 0.751 |
| Nevs | 0.935 |
| Seborrheic | 0.914 |
| Vascular | 0.985 |

**Tab. 12.** Per class AUC-ROC [Inception, SGD, focal Loss, with Augmentations]

| Class | AUC-ROC |
|---|---|
| Actinic | 0.929 |
| Carcinoma | 0.953 |
| Dermatofibroma | 0.786 |
| Melanoma | 0.826 |
| Nevs | 0.94 |
| Seborrheic | 0.905 |
| Vascular | 0.998 |

**Tab. 13.** Focal Loss – Without Augmentations, [Inception v3]

| Optimizer | Accuracy | Precision | Recall | F1-Score | Termination epoch # |
|---|---|---|---|---|---|
| Adam | 0.80 | 0.80 | 0.80 | 0.80 | 43 |
| RMSprop | 0.81 | 0.81 | 0.81 | 0.80 | 38 |
| SGD | 0.79 | 0.79 | 0.79 | 0.79 | 43 |

**Tab. 14.** Per class AUC-ROC [Inception, Adam, focal Loss, without Augmentations]

| Class | AUC-ROC |
|---|---|
| Actinic | 0.921 |
| Carcinoma | 0.937 |
| Dermatofibroma | 0.613 |
| Melanoma | 0.868 |
| Nevs | 0.947 |
| Seborrheic | 0.928 |
| Vascular | 0.998 |

**Tab. 15.** Per class AUC-ROC [Inception, RMSProp, focal Loss, without Augmentations]

| Class | AUC-ROC |
|---|---|
| Actinic | 0.903 |
| Carcinoma | 0.933 |
| Dermatofibroma | 0.673 |
| Melanoma | 0.864 |
| Nevs | 0.946 |
| Seborrheic | 0.906 |
| Vascular | 0.997 |

**Tab. 16.** Per class AUC-ROC [Inception, SGD, focal Loss, without Augmentations]

| Class | AUC-ROC |
|---|---|
| Actinic | 0.909 |
| Carcinoma | 0.946 |
| Dermatofibroma | 0.671 |
| Melanoma | 0.863 |
| Nevs | 0.954 |
| Seborrheic | 0.932 |
| Vascular | 0.997 |

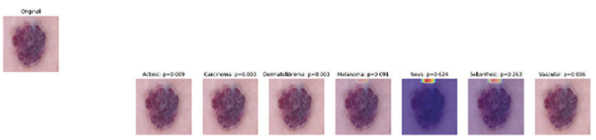

**Fig. 5.** Grad-CAM of Inception V3

## AUTHORS

**Rajit Chandra** – Computer Science Department, Purdue Fort Wayne, Fort Wayne, 46805, USA, E-mail: chanr02@pfw.edu.

**Mohammadreza Hajiarbabi*** – Computer Science Department, Purdue Fort Wayne, Fort Wayne, 46805, USA, E-mail: hajiarbm@pfw.edu.

*Corresponding author

## References

[1] N. Gessert, M. Nielsen, M. Shaikh, R. Werner, and A. Schlaefer, "Skin sion classification using ensembles of multi-resolution EfficientNets with meta data," *MethodsX*, vol. 7, p. 100864, 2020, DOI: 10.1016/j.mex.2020.100864.

[2] J. Yap, W. Yolland, and P. Tschandl, "Multimodal skin lesion classification using deep learning," *Exp. Dermatol.*, vol. 27, no. 11, pp. 1261–1267, 2018, DOI: 10.1111/exd. 13777.

[3] P. Mirunalini, A. Chandrabose, V. Gokul, and S. M. Jaisakthi, "Deep Learning for Skin Lesion Classification," 2017, [Online]. Available: http://arxiv.org/abs/1703.04364.

[4] T. C. Pham, C. M. Luong, M. Visani, and V. D. Hoang, "Deep CNN and Data Augmentation for Skin Lesion Classification," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 10752 LNAI, no. June, pp. 573–582, 2018, DOI: 10.1007/978-3-319-75420-8_54.

[5] T. Majtner, S. Yildirim-Yayilgan, and J. Y. Hardeberg, "Combining deep learning and hand-crafted features for skin lesion classification," *2016 6th Int. Conf. Image Process. Theory, Tools Appl. IPTA 2016*, no. December, 2017, DOI: 10.1109/IPTA.2016.7821017.

[6] A. Mahbod, G. Schaefer, I. Ellinger, R. Ecker, A. Pitiot, and C. Wang, "Fusing fine-tuned deep features for skin lesion classification," *Comput. Med. Imaging Graph.*, vol. 71, pp. 19–29, 2019, DOI: 10.1016/j.compmedimag.2018.10.007.

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