SKIN LESION DETECTION USING DEEP LEARNING

by

Rajit Chandra

A Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Master of Science

Department of Computer Science

Fort Wayne, Indiana

May 2022

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF COMMITTEE APPROVAL

Dr. Mohammadreza Hajiarbabi, Chair

Department of Computer Science3

Dr. Venkata N. Inukollu

Department of Computer Science

Dr. Yihao Deng

Department of Mathematical Sciences

Approved by:

Dr. Venkata N. Inukollu

ACKNOWLEDGMENTS

First and foremost, I'd want to express my gratitude to Professor Dr. Mohammadreza Hajiarbabi for enabling me to collaborate with him. Without his expert guidance and unwavering support, I would not have been able to complete my master’s degree. Also, I'd want to convey my gratitude to Dr. Venkata Inukollu and Dr. Yihao Deng who served on my thesis committee. The Computer Science whole faculty deserves credit for getting me to this point in the program.

I'd want to express my gratitude to my parents for encouraged me to seek this degree. I am indebted to them for their love, blessings, compassion, and sacrifices.

I don't have any meaningful words to convey my gratitude, but my heart is still overflowing with gratitude for the kindness shown to me by everyone. Thank you very much!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

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 explainability 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.

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 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. Because the design of CNNs can understand the complex structure of images.

2. LITERATURE REVIEW

[1], Computer-aided diagnostic (CAD) systems are now a need in today's world for diagnosing and evaluating clinical imaging. In the United States, CAD is used for diagnosing breast cancer (US). Melanoma is a type of skin tumor that begins to increase and expand all across the upper layer of skin before infiltrating the deeper layers, in which it eventually links with the blood arteries and lymph arteries, if left untreated. if tumor is discovered in its initial phases, it can be treated successfully. Researchers proposed machine learning based solution to diagnose skin cancer. The binary classification task was conducted to classify the images into two classes i.e. “malignant” and “benign”. In the machine learning tasks, features of images are extracted to classify images into predefined classes. Regularization is a strategy for handling the complexity of a predicting algorithm. A machine learning model is regularized by employing dropout technique in neural networks or by using the L1 regularizer and L2 regularizer. But researchers suggested a regularization technique which is based on the standard deviation value of the weight matrix of the classifier. They limited the model complexity by penalizing the dispersion of the values of weight matrix. Consequently, the values will be near together. The maximum average accuracy obtained throughout that method was 97.49 percent for 100 epochs. As the dataset was highly imbalanced, weighted accuracy, which is termed as “the mean of the true positive rate achieved for each class”, was also calculated. [2]**,** melanoma causes 75% of all skin tumor fatalities. The researchers used machine learning for classification of three classes skin lesions. The three classes are “normal”, “abnormal”, and “melanoma”. They also constructed a system that would help doctors make better decisions in diagnosis skin lesion. In the PH2 data set, a diagnostic research was conducted using the machine classifiers developed for melanoma diagnosis. Researchers conducted different experiments by using different algorithms to compare the results. They employed “Artificial Neural Network (ANN), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and Decision Tree (DT)” classifiers on the same PH2 data set. For the items in this research, categorical values are coded using "one-of-N coding." The best value for "k" in the k-fold cross-validation procedure was found to be 5 and 10 in experimental investigations. The results showed that ANN outperformed in identifying the classes then the other used algorithms. [3], the researchers offered a “fully automated system” for skin lesion classification in the paper. For training the classifiers, the researchers used the training set, validation set, and test set images

from the dataset provided by ISIC of year 2016 and 2017. There are 2037 colored dermoscopic skin photographs in total, the dataset has total of 411 images belongs to the class malignant melanoma, 254 images were present in the class seborrheic keratosis, and benign nevi class has 1372 images in the dataset. For the feature extraction process, three pre-trained classifiers were used. The three feature extractors that were used were, AlexNet, VGG16 and ResNet-18. These extracted features are then fed into the model to train support vector machine (SVM) models to classify the images into given classes. For every experiment, the researchers trained different classifiers and then average the scores to reach the final score of classification. The researchers used the logistic regression to convert SVM results to probabilities in order to evaluate the classification results. The researchers combine the scores for the two binary classification tasks that were presented in the ISIC 2017 challenge, the one was malignant melanoma versus all images and the other was seborrheic keratosis versus all images. The primary contribution of the paper is a "hybrid strategy" for skin lesion classification found on fusion of deep features, which involves training many SVM models and combining fusion probabilities to obtain good classification performance. [4], dermatologists diagnose melanoma in patients by using "ABCD" parameters.

The ABCD stands for,
Asymmetrical form - melanoma tumors are often asymmetric
Boundaries - melanoma lesions have abnormal boundaries
Color - melanoma lesions typically have more than one color
Diameter - melanoma lesions are typically greater than 6mm in diameter

On the ISIC 2017 dataset, the researchers used two input images at various scales of resolution to fine-tune an ImageNet pre-trained Inception-v3 deep neural network. The two scales correlate to a coarse level that preserves the broader vision and form features of the lesion, whereas the image at the finer scale provides granular details and several low-level lesion characteristics that are useful for discriminating across lesion classes. The size of input images was kept to [224 * 224] at the coarse scale, whereas at the finer level, the input data are first resized to [448* 448] before a center crop of size [224 *224] is produced. The input values are then scaled to the [0,1] range. Prior to training or inference, the images are not pre-processed in any way. The multi-scale network is generated by using the same Inception feature extractor to process both the lower

quality image and the higher quality center-crop image. Each image's feature vectors are concatenated into a single 4096-element vector, that is next processed through a fully-connected layer with 1024 hidden units and ReLU activations. Finally, for each of the three classes: melanoma, seborrheic keratosis, and nevus, a three-way SoftMax is employed to provide probability estimates. The researchers developed numerous models with modest modifications on the original formulation to enhance the variance of the outputs and hence boost the effectiveness of the final version of model. After that, geometrical mean was used to aggregate the final forecasts from each model. The researchers integrated the results of ten distinct models in all. Extending the batch size, raising the amount of unfrozen inception modules, and retraining on various folds of the dataset the were all changes made to the previously mentioned model. The ensemble additionally included a single-scale model trained on images enlarged to [448*448] instead of [224*224]. [5], Deep learning frameworks can assist in avoiding the phase of manually extracting features. This can save time and inform the patient if a questionable signal is detected. There are 200 lesion photos in PH2, the dataset had 80 typical nevi and same number of atypical nevi, and half of melanoma i.e., 40. Due to the small number of samples in the dataset, multiple machine learning techniques and deep learning techniques cannot reliably validate and rely on the findings achieved using these samples. The PH2 dataset's labelled images were enhanced with Augmentor to yield a total of 4000 labelled images. The final dataset after augmentation contained 2000 photos of Nonmelanoma and 2000 images of Melanoma, resulting in a suitable data for developing deep learning models. To equalize every class of skin lesion, the HAM10000 dataset's labelled images were also enhanced using Augmentor to a total of 34415 tagged images. It was utilized to obtain findings for assessments using different models such as MobileNet, VGG-16, and Custom model. The researchers investigated the efficiency of two pre-trained state-of-the-art models, MobileNet and VGG16, on the PH2 dataset and the HAM10000 dataset in two situations, one with data augmentation and the other without data augmentation. The researchers also created their own customized deep learning framework then contrasted its performance to over to the other two models to demonstrate that training a well-designed model from the ground can also be effective. In basis of test accuracy, AUC, and F1 score assessment metrics, MobileNet and the custom model outperformed the VGG-16. [6], a smart healthcare system to detect skin tumor can be a useful method for clinicians. The researchers formulated the binary classification problem: determining if a dermoscopic image of a lesion comprises a melanoma lesion or a benign lesion. The VGG16

architecture was chosen for this project because it has been proved to generalize better to different datasets. The input layer of network requires a [224*224] pixel RGB picture. Size rescaling, horizontal shift and flipping, image zooming, and rotations of 40 degrees were the data augmentation techniques chosen. Furthermore, data augmentation is intended to aid in the prevention of overfitting (a problem that occurs when a model is subjected to very few instances and acquires features that do not apply to new data) and, as a result, improve the model's ability to generalize. Under sampling is used to balance the dataset. The first 173 photos from every group in the training data and the first 75 input images from every class in the test set were chosen alphabetically. The number of input images in training set was 346 while the number of testing images were 150 in the final dataset. The researchers used three methods to train the classifier.

Method 1:

The model was trained from scratch; the framework was rained for epochs after being initialized 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.

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.

Method 3:

Another frequent transfer learning strategy entails not only training the model by sing 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 initialize 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 initialize the fifth and final convolutional block. The evaluation metrics showed that the third method performed better than Method 1 and Method 2. [7], dermoscopic is the state-of-the-art procedure for skin cancer screening, with a diagnosis accuracy that is higher than the naked eye. 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 dermoscopic image, and a histological diagnosis detail. 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 (i.e., eyes, facial landmarks, jewelry, or garment). ResNet-50 was used to extract the features of the images. Three kinds of experiments were conducted,

"Full multimodality classification"

When all mentioned three modes (macroscopic image of lesions, dermoscopic images, and metadata of patients) were provided, the researchers built a network with two image feature extractions, one for dermoscopic input images and the other for macroscopic input images.

"Partial multimodality classification"

The researchers excluded the other two from the complete network when only one image modality (macroscopic images or dermoscopic 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.

"Single image classification"

When there was only one image type for classification and no 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. [8], using image representations produced from Google's Inception-v3 model, the proposed automated approach intends to detect the kind and cause of cancer directly. 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. [9], 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. 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.

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.

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.

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.

The 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. [10], in image recognition nowadays, two basic types of feature sets are routinely used. 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 utilize 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. [11], 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 images can help with diagnosis. 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 preprocessing, 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 (Malignant Melanoma /Sabrohtic Keratosis/ 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. [12], the “attention residual learning convolutional neural network (ARL-CNN)” model for skin lesion categorization is proposed in this research. 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 dermoscopic 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. [3], 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. 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 dermoscopic 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.

3. DATA AND METHODS

3.1 **Skin lesion**

In the United States, a total of more than 9500 people are identified with the skin cancer on the daily basis and the rate of deaths from this disease is also alarming. In a day, more than two persons die of this disease every hour. It is horrific that people are detected with skin cancer more than all the other cancers in United states. It is a predicted that one in five U.S. inhabitants are vulnerable to this cancer by the age of 70. Skin cancer has several types and "actinic keratosis" is the most frequent diagnosed kind in people of America. Almost more than 58 million people suffers from this this type of skin cancer. Moreover, according to the statistics of the treatment cost, $8.1 billion are spend annually in the treatment of this cancer. The nonmelanoma type of cancer has a treatment cost of $4.8 billion while melanoma skin cancers have a cost of $3.3 billion. It shows that nonmelanoma skin cancers are more common in people of America. This type of skin cancers have grown by 77% between the year 1994 and the year 2014. The one of the main reasons for developing nonmelanoma skin cancer is the excessive exposure of skin to sun. The Ultraviolet radiations in the sun are dangerous for the skin and these are the radiations that causes this cancer to develop. Although there can be other reasons but 90% of the time, the reason for developing this cancer is the sun exposure. The nonmelanoma skin cancer can be divided into several forms. One form of nonmelanoma skin cancer is Basal Cell Carcinoma and around 3.6 million people are diagnosed with this form every year. Another form of nonmelanoma skin lesion is Squamous Cell Carcinoma and this form has most cases after Basal Cell Carcinoma. Almost 1.8 million cases of SCC are detected every year in the United States. A total number of 15,000 deaths are reported from Squamous Cell Carcinoma every year. And it is more than twice the deaths that occurs from melanoma skin cancer. The nonmelanoma form of skin cancer takes the life of more than 5400 people around the globe. The Squamous Cell Carcinoma is usually diagnosed in the people who go through the organ transplant surgery. While other people are also vulnerable to this form but organ transplant patients are 100% more likely to develop this form. one very rare form of skin cancer is Merkel Cell Carcinoma which is alarmingly increasing from 2000 to 2013.

The frequency of new instances of melanoma detected in year 2021 is predicted to boost by 5.8%. And for the year 2021, the frequency of people dying from melanoma is predicted to rise by 4.8 percent. In the United States, a projected 207,390 cases of melanoma will be diagnosed in the year 2021. Moreover, the projection was made that there will be 106,110 noninvasive instances restricted to the epidermis (i.e. upper layer of human's skin), and 101,280 invasive cases entering the epidermis into the second layer of skin (the dermis). The projection is also made for the gender, men will account for 62,260 of the invasive cases, while women will account for 43,850. The amount of new aggressive melanoma detected instances each year grew by 44% over the last decade (2011-2021). Melanoma will kill 7,180 individuals in 2021, according to estimates. There will be 4,600 males and 2,580 females among them. As in the case of nonmelanoma the sun is responsible for the great proportion of melanoma patients. According to one stud of United Kingdom the ultraviolet (UV) light from the sun is responsible for roughly 86 percent of melanoma cases. An individual's risk of acquiring melanoma doubles after five sunburns on average,15 but only one severe sunburn during childhood or teenage more than doubles a person's odds of acquiring melanoma eventually in life. It is the deadliest of all skin cancers, and its prevalence has skyrocketed in recent decades. When melanoma is detected when it is still confined to the skin's surface layers, simple excision is often curative, with a 5-year survival rate of roughly 98 percent. Regrettably, despite the fact that melanoma can be detected early with a basic eye examination, many people are nevertheless identified with severe cancer.

Melanoma is usually identified with a straightforward visual examination and is usually brought to the attention of a doctor due to a worrying appearance. Even for professionals, early-stage melanomas can be difficult to detect from benign skin lesions despite being apparent to the naked eye. Despite an increasing pandemic of skin biopsies, this has resulted in numerous undetected melanomas. The frequency of needless biopsies changes depending on the clinical situation, the examiner's ability, and the use of technologies. For example, in youngsters, where melanoma frequencies are low yet changing moles are widespread, nearly 500,000 biopsies are performed every year to identify approximately 400 melanomas in the United States.

It is obvious that there is a need to increase the efficiency, effectiveness, and accuracy of cancer detection. The financial and personal implications of failing to detect cancer early are high. On the

other side, inexperienced melanoma screening can result in several needless biopsies. Another issue that has arisen because of increasing cancer screening is the diagnosis and treatment of indolent lesions that cannot currently be separated from potentially fatal tumors. This latter scenario is known as "overdiagnosis," and it raised questions about democratizing availability to automated mobile testing equipment in populations with a lower risk for death from skin cancer.
So the task in hand was to employ machine learning algorithms in detecting the melanoma. The ability of deep learning to understand the images can help the examiners in decision making. The very powerful algorithm of deep learning is convolutional neural networks.

3.2 **Problem framing**

The skin lesion can be divided into seven classes. The problem statement is to detect the type of skin lesion by examining the image. So it is a multiclass classification task. The 7 classes of skin lesion are:

1. Melanoma
2. Melanocytic Nevus
3. Basal Cell Carcinoma
4. Actinic Keratosis
5. Benign Keratosis
6. Dermatofibroma
7. Vascular Lesion

3.3 **Dataset**

The dataset used for classifying the skin lesion into classes is provided by ISIC. The International Skin Imaging Collaboration (ISIC) is an international effort supported by the International Society for Digital Imaging of the Skin to enhance melanoma diagnosis (ISDIS). The ISIC Archive houses the world's largest archive of properly regulated dermoscopic images of skin lesions. The dataset that was used come under the “task 3” of classification.

3.4 **Convolutional Neural Networks**

The Convolutional Neural Network (CNN) [13] has made incredible progress. In the area of deep learning, it is now one of the most prominent neural networks. Face recognition, driverless automobiles, self-service supermarkets, and intelligent healthcare are just a few examples of what computer vision based on convolutional neural networks has allowed mankind to achieve in the last few decades. Artificial Neural Networks and Convolutional Neural Networks are inextricably linked (ANN). McCulloch and Pitts introduced the MP model, which is the first mathematical model of neurons. It was introduced in 1943. Rosenblatt developed a one-layer perceptron model by incorporating learning capacity to the MP model in the late 1950s and early 1960s. A single-layer perceptron network, on the other hand, is not capable of solving the problems that holds linear inseparability, one example is the XOR problems. Later, the Back Propagation Network was introduced which is in place of single-layer, a "multi-layer feedforward network" which is trained using the backpropagation technique to solve issues that single-layer perceptron was not able to solve. After that "one-dimensional" convolutional neural network called the "Time Delay Neural Network" for the task of speech detection was introduced.

A convolutional neural network is a one of the types of feedforward neural network which utilizes the "convolution structures" to extract the features from data. traditional feature extraction techniques need manual feature extraction but in convolutional neural network this process is automated. There is no manual feature extraction required. The CNNs are framed in a way that is influenced by the visual perception of the input data.

The term "artificial neuron" is inspired from the biological neuron, the biological neuron contains receptors, in CNN the kernels behave like receptors of artificial neuron. The kernels interact with many features in convolutional neural network. In human brain, neurons pass only those signals to next neurons that possess a value greater then threshold, this same process is done in CNN by the activation function. Moreover, the optimizers and the loss functions that we use to teach CNN that what is the expected output from it. The convolutional neural networks have many benefits over traditional ANNs.

3.4.1 Benefits of Convolutional Neural Networks

- Faster Convergence:

  Every neuron in the network is only connected to only few other neurons, unlike traditional network where every single neuron was connected to every other neuron. These less connections reduces the network parameters and the model converges faster.
- Parameter Reduction:

  In CNN, the parameters are reduced further by the making the connections share weights.
- Dimensionality Reduction:

  Convolutional Neural Networks contains many layers and of them is a pooling layer. The pooling layer reduces the size of images while keeping the crucial information.

These factors make the Convolutional Neural Networks very effective and it is one of the most use algorithm for deep learning tasks.

3.4.2 Components of Convolutional Neural Networks

Convolutional Neural Networks contains four components,

1. Convolution
2. Padding
3. Convolutional Kernel
4. Pooling

Convolution is crucial for extracting the features from input data. the output that is generated from these convolutions are termed as “feature maps”. The convolution kernel will lose the information that is stored in the borders. To counter this issue, padding is introduced in the process. The zeros are padded in the borders to increase the size of input. The strides are introduced to manage the density of “convolving”. The higher value of stride makes the density low. When features are convolved, the number of features increases, this increase in features increase the chances of overfitting. To counter this issue, pooling is introduced, by pooling the features are reduced. There are two types of pooling:

1. Average Pooling
2. Max pooling

The process of Convolutional Neural Networks is shown as:

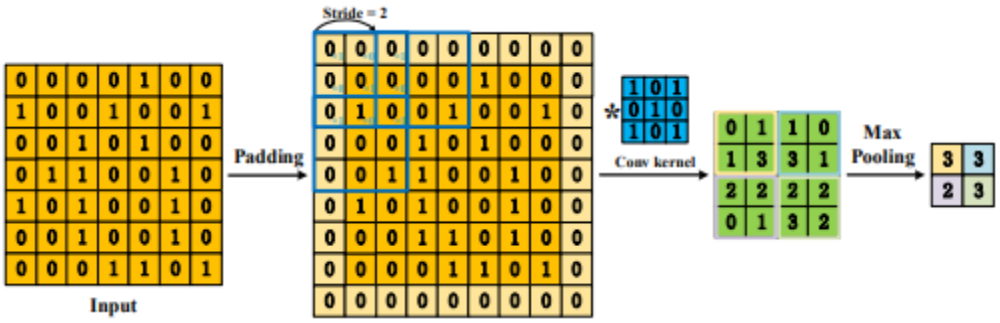


Figure 1:CNN Process

Over the last decade, Convolutional Neural Networks have achieved ground-breaking breakthroughs in a wide range of pattern recognition applications, from image processing to speech recognition. recognition. CNNs' greatest advantageous feature is their ability to reduce the amount of ANN parameters This accomplishment has inspired both research community to consider larger projects models in order to fix complex jobs, that was not before done possible with traditional ANNs. A key component of CNN is the ability to extract abstract features when input spreads to deeper levels. In picture classification, for example, the edges may be recognized during first layers, followed by simpler shapes in the next levels, and then higher level characteristics such as faces as the network goes deeper. The pixel values of images are represented by the height, width and depth.

Moreover, a modification of traditional CNNs was introduced in which the convolution kernels perceive the wider area. This modified CNN architecture is called “Dilated Convolution”.

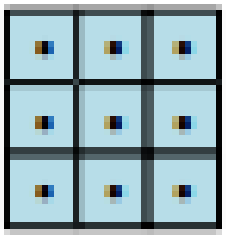


Figure 2: Traditional Convolution Kernel

This is the traditional convolution kernel with the shape of 3*3.

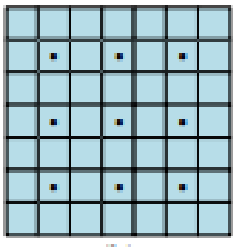


Figure 3: 2-dilated Convolution Kernel

This is modified version called dilated version of CNN. This is 3*3 kernel and 2-dilated. There is also a 4-dilated kernel.

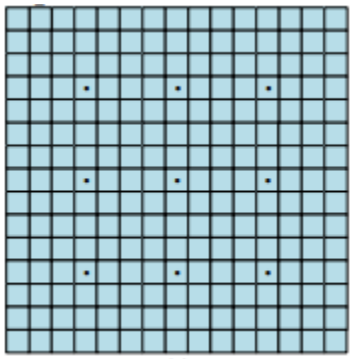


Figure 4: 4-dilated Convolution Kernel

This is 4-dilated kernel with the shape of 3*3.

There is also another convolution kernel called “deformable convolution kernel”. It was introduced to address the object irregularity issue. Because the object shapes are irregular so these convolutions were introduced. These convolutions only focus on the main things. A comparison between traditional convolution and the deformable convolution can be demonstrated as:

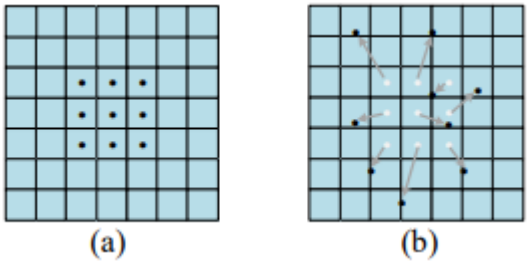


Figure 5: general (left) and deformable convolution (right)

The shape of both the convolutions kernels is 3*3.

3.4.3 Convolution Neural Networks Models

One of the key features of CNN is that the number of parameters are reduced by using different filters, strides and paddings. Since the CNN is very powerful architecture and coding the one form scratch can be tricky, a lot of architectures are built by experts of the field and that are trained on millions of parameters.

Some of these are:

Table 1: CNN Models

CNN Model	Year Introduced
LeNet-5	1998
AlexNet	2012
ZFNet	2013
VGGNets	2014
NiN	2014
Inception v1	2014
ResNet	2015
DCGAN	2016
Inception v2	2016
Inception v3	2016
SqueezeNet	2016
MobileNet v1	2017
Xception	2017
ResNeXt	2017
DenseNet	2017
ShuffleNet v1	2017
SENet	2017
Inception v4	2017
MobileNet v2	2018
ShuffleNet v2	2018
MobileNet v3	2018
GhostNet	2020

3.5 DenseNet:

DenseNet [14] is a novel Convolutional neural network architecture that was able to achieve state of the art performance on the classification problems for the dataset of ImageNet, CIFAR and

SVHN. The main contribution of this model is that it uses less parameters in achieving highly accurate results. On some classification benchmarks, DenseNet is more efficient. Its parameters are more representative, and it has less duplicated layers. This makes it more feasible to perform efficient computation.

3.5.1 Architecture of DenseNet:

Dense Net contains "Dense blocks". These blocks consist of layers that are densely linked to each other. One DenseNet block contains input layer, batch normalization and activation function layers.

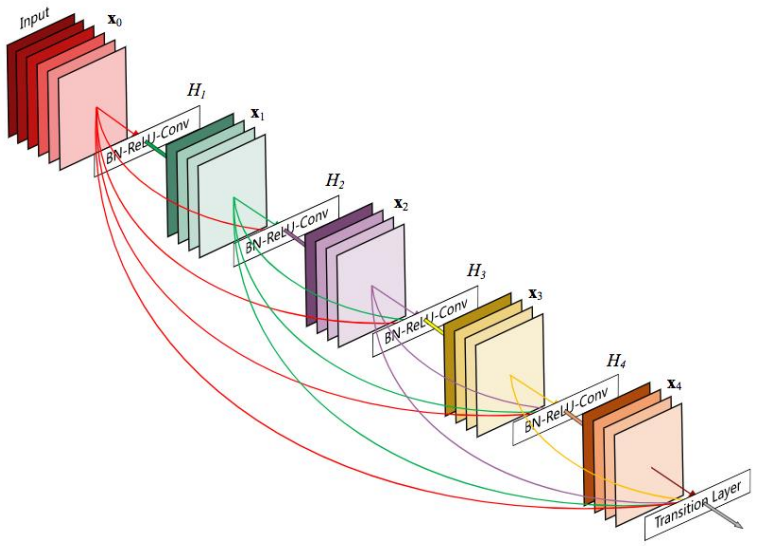


Figure 6: Dense Block

Every layer in the DenseNet block receives all the feature maps generated by previous layers. If we have L layers then the direct connections in the dense block can be numerically presented as:

$$L = (L + 1)/2$$

DenseNet links every single layer in the architecture to each layer. The input feature map for every layer is the concatenation of all the feature maps that are generated from previous layers. DenseNet reduce the vanishing-gradient problem, the usage of number of parameters also decreases and it also benefits from reuse of features. To concatenate all the feature maps, it is important to have the same size of the feature maps to be concatenated. But in convolutional neural network, the

main idea is that as the features propagate, the size of feature map is reduced. This problem is addressed by using multiple dense blocks. The dense blocks use same feature map size and for downsampling the inputs, the pooling layer and the convolutional layer is used in between the blocks.

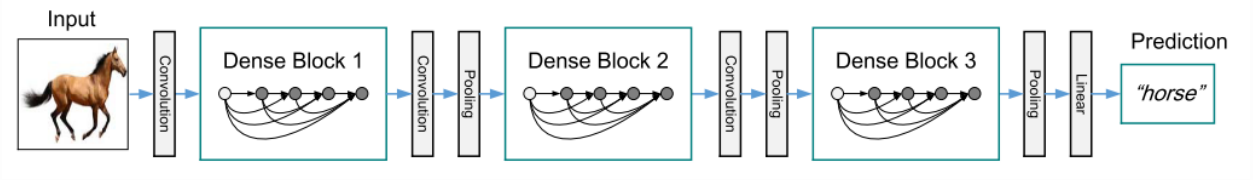


Figure 7: DenseNet Architecture

The convolutional layer and the pooling layers in between the dense block are named as "transition layers" by the authors.

When an image x_0 is feed into CNN, every layer of the network apply a "non-linear transformation" $H_l(.)$. The l here represents the index of the layers L. The transformation function $H_l(.)$ can implement multiple operations like batch normalization, pooling, rectified linear unit or convolution. The output generated by the l^{th} layer is represented by the symbol x_l. The DenseNet architecture allows direct connections from any layer to all the following layers. The l^{th} layer of the DenseNet is fed with all the previous layers $(x_0, \dots \dots, x_{l-1})$ feature maps.

$$x_l = H_l([x_0, x_1, \dots., x_{l-1}])$$

Where H_l is the composite function of three operations.

- Batch Normalization
- Rectified Linear Units
- 3×3 convlutions

The size of the feature maps increases when it passes through every dense layer and the new features and existing features are concatenated. These features are the "global state" of the network and every layer adds k features to the global state. This k is called the "growth rate" of the network. Different versions of DenseNet are introduced on the basis of the number of layers. However, the dense blocks are same in every version introduced.

Table 2: DenseNet Variants

Layers	Output Size	DenseNet-121	DenseNet-169	DenseNet-201	DenseNet-264
Convolution	112 × 112	7 × 7 conv, stride 2			
Pooling	56 × 56	3 × 3 max pool, stride 2			
Dense Block (1)	56 × 56	[1 × 1 conv; 3 × 3 conv] × 6	[1 × 1 conv; 3 × 3 conv] × 6	[1 × 1 conv; 3 × 3 conv] × 6	[1 × 1 conv; 3 × 3 conv] × 6
Transition Layer (1)	56 × 56	1 × 1 conv			
	28 × 28	2 × 2 average pool, stride 2			
Dense Block (2)	28 × 28	[1 × 1 conv; 3 × 3 conv] × 12	[1 × 1 conv; 3 × 3 conv] × 12	[1 × 1 conv; 3 × 3 conv] × 12	[1 × 1 conv; 3 × 3 conv] × 12
Transition Layer (2)	28 × 28	1 × 1 conv			
	14 × 14	2 × 2 average pool, stride 2			
Dense Block (3)	14 × 14	[1 × 1 conv; 3 × 3 conv] × 24	[1 × 1 conv; 3 × 3 conv] × 32	[1 × 1 conv; 3 × 3 conv] × 48	[1 × 1 conv; 3 × 3 conv] × 64
Transition Layer (3)	14 × 14	1 × 1 conv			
	7 × 7	2 × 2 average pool, stride 2			
Dense Block (4)	7 × 7	[1 × 1 conv; 3 × 3 conv] × 16	[1 × 1 conv; 3 × 3 conv] × 32	[1 × 1 conv; 3 × 3 conv] × 32	[1 × 1 conv; 3 × 3 conv] × 48
Classification Layer	1 × 1	7 × 7 global average pool			
		1000D fully-connected, softmax			

Every version contains four dense blocks with different number of layers:

DenseNet-121 (6, 12, 24, 16)

DenseNet-169 (6, 12, 32, 32)

DenseNet-201 (6, 12, 48, 32)

DenseNet-264 (6, 12, 64, 48)

For the classification of skin lesions, the DenseNet-121 architecture is used.

3.6 **GoogLeNet:**

GoogLeNet architecture was able to win the ILSVRC award in 2014 for the classification algorithms. This architecture was built by stacking the inception modules. There are different variations of these inception networks.

- Inception v1
- Inception v2
- Inception v3
- Inception v4

3.6.1 Inception v1

[15], the components in images that we take from cameras have varying distances, if an object in an image is covering the larger area then a large convolution kernels are employed. Or alternately, few small kernels are used. and if the object is small in the image then the opposite design of kernels is used. So the convolution kernels with the larger size have more training parameters. The inception v1 employed 1x1, 3x3 and 5x5 size kernels to build a "wider network". The varying sizes of convolution kernels are useful in extracting feature maps of different scales in the images. These extracted feature maps are combined together to get a final representation of the network. The 1x1 size kernel is utilized to decrease the channel number. It results in the reduction of computational cost.

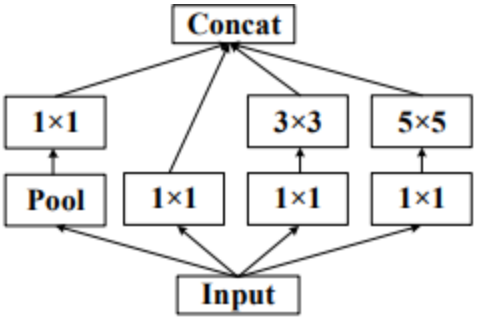


Figure 8: Inception v1

3.6.2 Inception v2

Inception v2 uses the batch normalization to manage the "covariate shift" problem. The output value that is generated from each layer is processed to come under the normal distribution. It helps in making the model more robust. Moreover, the training is done by using a larger value of learning rate. The convolution kernel that was represented as 5x5 size in Inception v1 has a different representation in Inception v2. Two 3x3 kernels are used in the place of one 5x5 kernel. Moreover, one nx1 and another nx1 layer is used in the place of one nxn size layer. The factorization in the initial layers is not very beneficent to use so it was utilized in the feature maps that have a medium size. The filter banks were also expanded to make improvements in the higher dimension. So the factorization was only used in last convolution of size 3x3 of each branch.

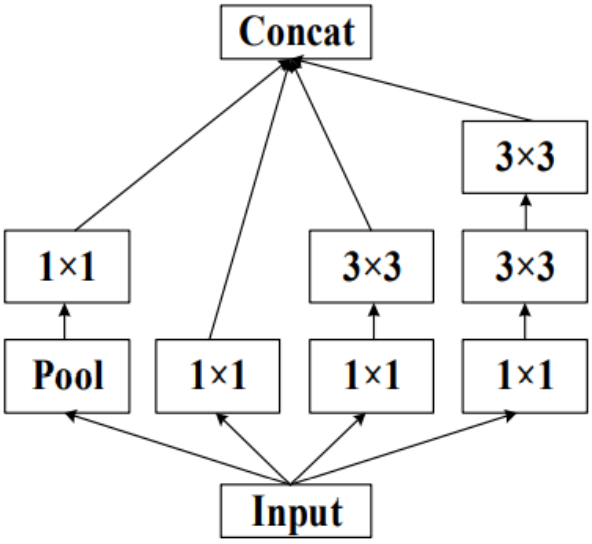


Figure 9: Two 3x3 in place of one 5x5

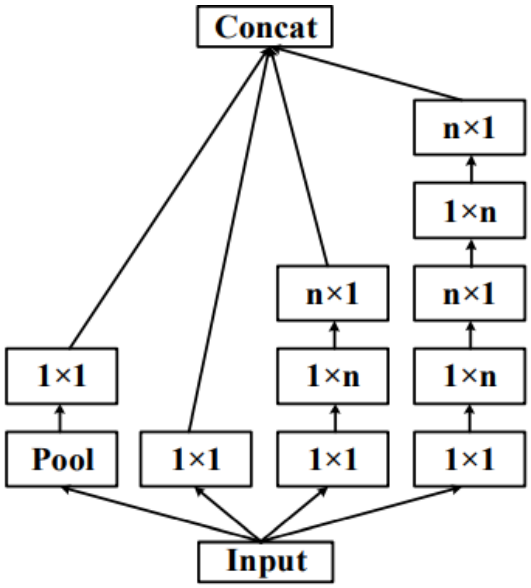


Figure 10: 1xn and nx1 in place of nxn convolution

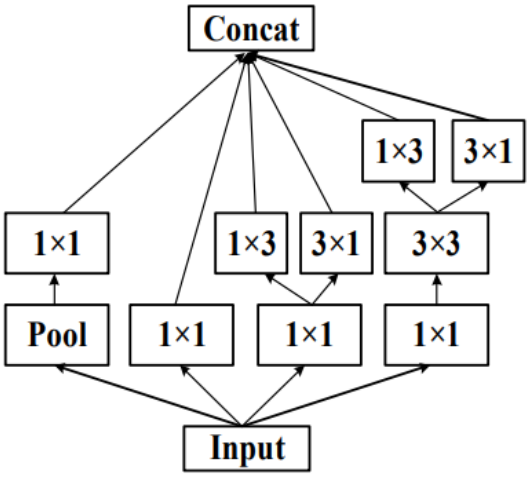


Figure 11: last convolution is factorized

3.6.3 Inception v3

The inception v3 has most of the attributes introduced in the inception v2. The 5x5 kernels and 3x3 size kernels are factorized into two kernels with one dimension, 1x7 and 7x1, 1x3 and 3x1 respectively. This step enhances the speed of training and minimize the time it takes and by doing this the depth of network is also increased. The standard size of input was 224x224 in the previous versions, in inception v3 it is increased to 229x229. The regularization term of the model is calculated by analyzing the outcome of “label-dropout” during training. This technique makes it hard for the classifier to predict a class with higher confidence. It refines the error rate by 0.2%. this version uses the RMSProp optimizer. The auxiliary classifier in the model uses batch normalization in the full connected (FC) layer. For optimal performance, an Inception network should use minimal computational resources and require less computation load. This is achieved using various convolutional filter sizes. For the high-performance output of an Inception network, efficient use of computational resources with little increase in computation load is required. The ability to extract features from input data at different scales by using different convolutional filter sizes. 1x1 conv filters learn cross-channel patterns, enhancing the network's overall feature extraction capabilities.

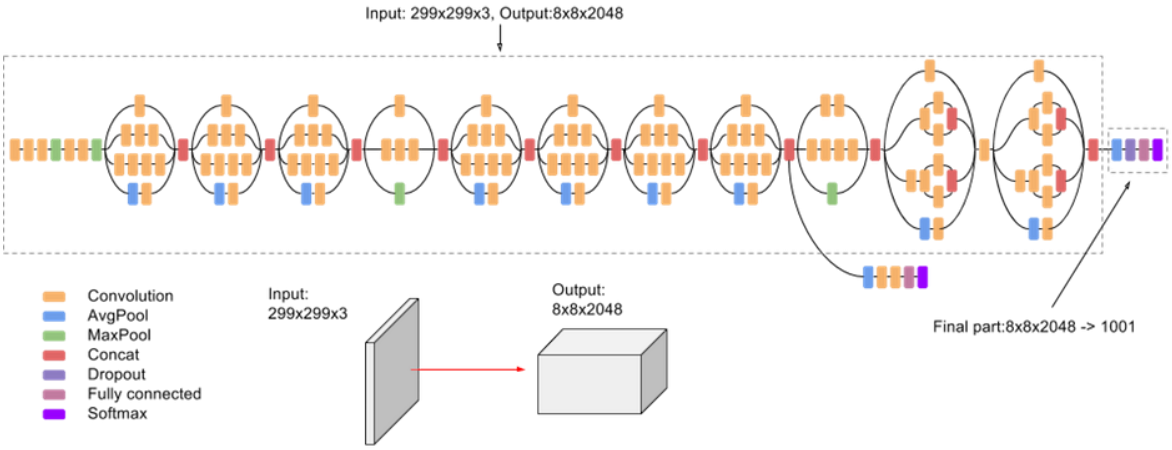


Figure 12: Inception V3 architectures

3.7 **Optimizers**

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

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.

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

η is learning rate.

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:

$$v(w,t) \coloneqq \gamma v(w,t-1) + (1-\gamma)(\nabla Q_i(w))^2$$

The learning parameters are updated as follows:

$$w \coloneqq w - \frac{\eta}{\sqrt{v(w,t)}} \nabla \mathrm{Q}_i\,(w)$$

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).

"Root Mean Square Propagation" (RMSProp), which additionally sustains adaptive per-parameter learning rates on the basis of mean of recent gradient magnitudes for the weights (e.g. how swiftly it is changing). This indicates that the algorithm performs effectively on non-stationary and online issues (e.g. noisy). Adam recognizes the value of both "AdaGrad" and "RMSProp". Rather than modifying the learning rates solely on the basis of the mean first moment (the mean), as RMSProp does, Adam also uses the average of the gradients' second moments (the uncentered variance).To be more precise, the algorithm creates an exponential moving average of the "gradient and the squared gradient", with parameters beta1 and beta2 controlling the decomposition rates of both moving averages. Hyper-parameters are often easy to read and require little adjustment. To calculate the moments, Adam optimizer uses exponentially moving averages, which are calculated on the gradient that is computed on a current mini-batch:

$$m_t = \beta_1 m_{t-1} + (1-\beta_1) g_t$$

$$v_t = \beta_2 v_{t-1} + (1 - \beta_2) g_t^2$$

Where,

$$g_t = \frac{\delta L}{\delta W_t}$$

m and v refer to moving averages,

g refers to the gradient on current mini-batch,

β_1 and β_2 refer to new hyper-parameters.

The default value for β_1 is 0.9 and it is 0.99 for β_2.

The visual representation of how Adam works is given as:

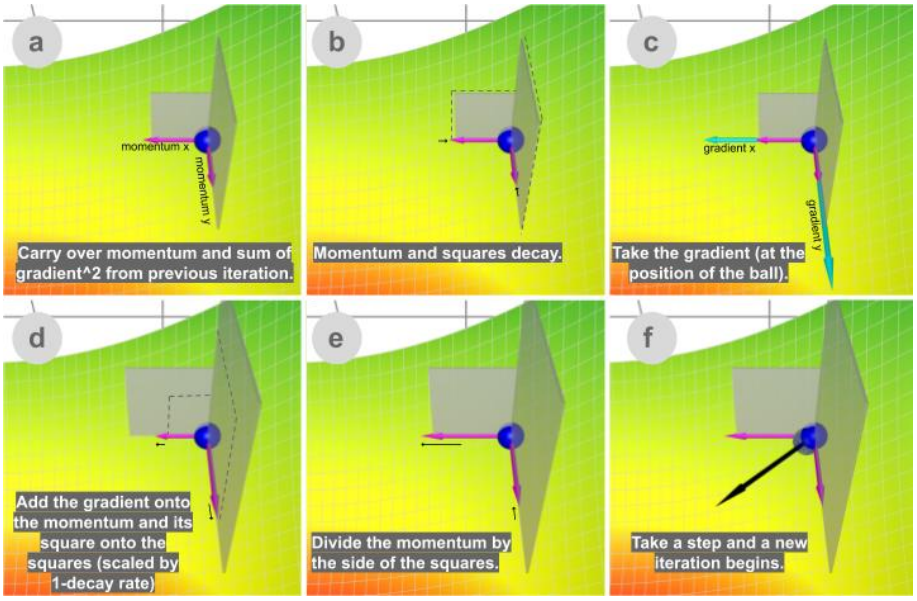


Figure 13: working of adam

3.8 **Loss function**

The loss function in deep learning is used to calculate the distance between actual value and the predicted value. It is utilized to make the model learn. The goal of the model is to minimize this loss value. The model is trained with different hyper parameters to achieve this goal. There are different loss functions that are used according to the nature of problem. the mean absolute error, cross entropy, mean square error are commonly used.

The data was highly imbalance so using categorical cross entropy was not the efficient to use, the other solution was to use weighted categorical cross entropy, where the classes with less examples are given more weightage as compared to the classes with more examples. This technique is also efficient but the "focal loss" [16] that was introduced to address the class imbalance problem was more efficient to use. The focal loss for the binary case can be defined as,

$$FL_{(pt)} = (1 - p_t)^{\gamma} \log p_t$$

Where $\gamma \geq 0$ and is termed as "focusing paramter" and

$$p_t = \begin{cases} p & if\ y = 1 \\ 1 - p & otherwise, \end{cases}$$

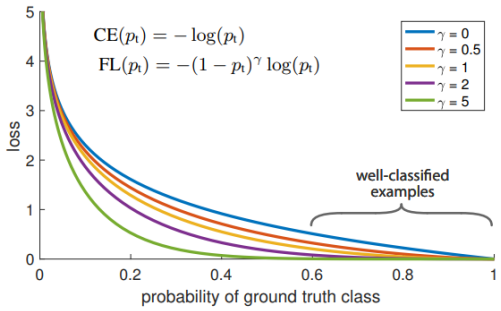


Figure 14: focal loss

1. When any input image is not classified correctly, the term pt move to 0 and the adjusting factor will become 1, due to which the loss function is not affected. And if the input image is classified correctly, pt will be 1 and the adjusting factor will be 0. It will make the loss value near to 0 which will "down-weight" the input image.

2. The focusing parameter γ regulate the rate at which the input images that are easily classified are down-weighted.

3.9 **Medical Aided Visualizations**

Convolutional Neural Networks allowed major breakthroughs in different tasks related to computer vision. Whether it is classifying the images, detecting the objects in image, segmenting the images or captioning them, all these computer vision tasks use Convolutional Neural Networks.

But these models are difficult to decompose into separate components. It makes the model interpretability very difficult. This is a major problem in computer vision tasks. If any system does not give satisfactory performance, there is no way of knowing the reasons behind generating a specific output. This drawback impacts the reliability of system. It is important to make the artificial intelligence based system trust worthy. So it is needed that we build the systems that can explain why a specific output is generated. This explainability of the AI based system can help in two stages of AI evolution.

1. Despite all the advancements in AI, it is still very weak than human beings and thus it is not deployable in many fields such as visual question answering. The explainability of the model can help researchers to focus in the better direction.
2. The explainability of the models will help in trust building in the AI based applications.

There is always a trade-off between explainability and accuracy of the model. The traditional rule-based systems can be interpreted easily but they do not give satisfactory accuracy. If pipelines can be decomposed, and every phase is designed by hand then these kinds of models are considered to be more interpretable. The deep learning models are highly accurate but least interpretable. It is because deep models use more layers as well as they train on end to end training. for example, the deep model ResNet have 200 layers and this architecture turned out very accurate in various computer vision tasks. Due to this complexity, deep models are not interpretable.

Similarly, employing deep learning in healthcare becomes problematic because the complicated design of neural networks usually makes them far more difficult to interpret than typical machine learning techniques (e.g. linear models). Class Activation Maps are one of the most prevalent ways for improving model explainability for the computer vision problems (CAM). Class activation maps are important for determining where the model is "seeing" when predicting the class of images, it might belong to. To know that which region of the image the model to see to make the decision, the “Gradient-weighted Class Activation Mapping” (Grad-CAM) is used. It employs gradients of the predicted class, streaming into the model’s final convolutional layer to generate a localization map highlighting the essential regions in the picture for the prediction.

3.9.1 Grad-CAM

Many researchers in the past have proposed deeper representations in the Convolutional neural network which produced higher-level visualizations. The convolutional layers lock the "spatial information" which is lost in fully connected layers. It shows that the final layers in CNN can have a compromise between "spatial information" and high level semantics. The neuron units in the last layers look for semantic information which is specific to class in the image. Grad-CAM [17] utilizes the gradient information that flows into last convolutional layer of the Convolutional Neural network. It allocates the important values to every neuron for a specific decision of interest. The technique used in Grad-CAM is general because its w.r.t feature activation maps of the convolutional layer. These gradients are "global averaged pooled" over the height and width dimension to get the neuron importance. can be utilized to explain the activations of any layer in the CNN. But we are more concerned with the decision of output layer. To get the localization map of the class, the score gradient for class is calculated.

$$\alpha_k^c = \overbrace{\frac{1}{Z}\sum_i\sum_j}^{\text{global average pooling}} \underbrace{\frac{\partial y^c}{\partial A_{ij}^k}}_{\text{gradients via backprop}}$$

The final equation of the Grad-CAM is as follows,

$$L^c_{\text{Grad-CAM}} = ReLU\underbrace{\left(\sum_k \alpha_k^c A^k\right)}_{\text{linear combination}}$$

The workflow of the Grad-CAM in the convolutional neural network is explained in this diagram,

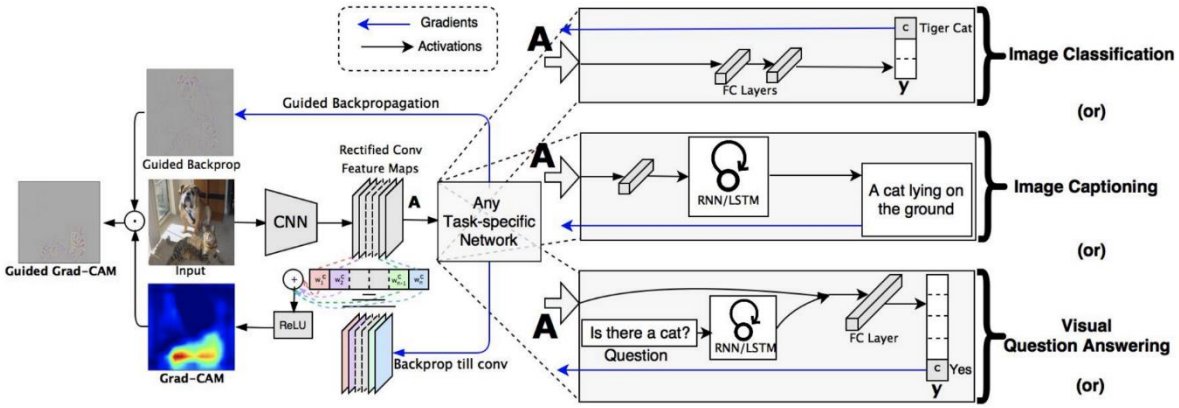


Figure 15: GRAD-CAM

Grad-CAM are useful in localizing the regions but it does not explain the reasons for the prediction the model makes. But it is beneficial in knowing the regions the model looked at to make the final decision.

3.10 **Language and Frameworks**

To experiment the skin lesion classification model, following languages and frameworks were used.

Language:

Python 3.6

Framework:

Tensorflow

Keras

4. 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. Following evaluation metrics were used to evaluate the models.

1. AUC-ROC

   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 formulae of true positive rate and false positive rate are as follows:

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

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

   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.

2. Accuracy

   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 given as:

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

3. Precision

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

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

4. Recall

   Recall indicates fraction of actual positives that were predicted correctly.

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

5. F1 Score

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

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

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 use 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,

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

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

With Augmentations:

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]

Without Augmentations:

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

7. DenseNet [RMSPROP]
8. DenseNet [ADAM]
9. DenseNet [SGD]
10. Inception v3 [RMSPROP]
11. Inception V3 [ADAM]
12. Inception V3 [SGD]

4.1 **DenseNet model**

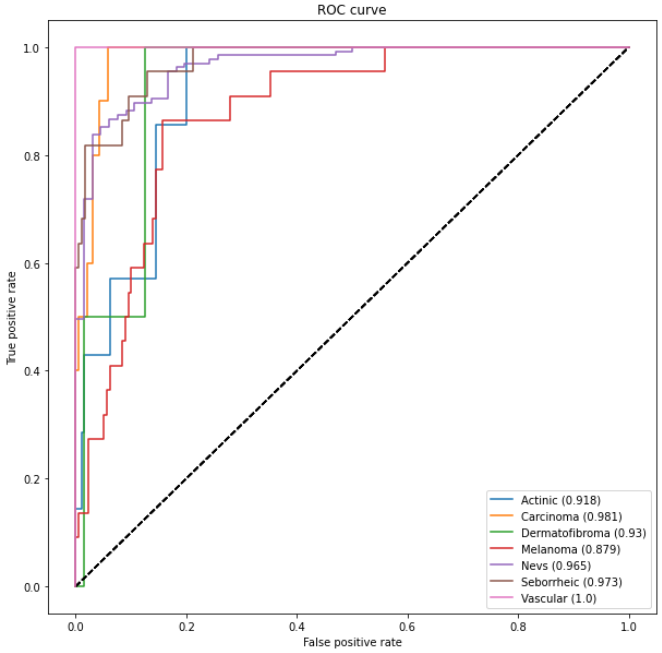


Figure 16: ROC CURVE for RMSprop

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

Table 3: 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	0.79	35
SGD	0.81	0.82	0.81	0.81	34

Per class AUC-ROC [DenseNet, Adam, 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

Per class AUC-ROC [DenseNet, RMS Prop, 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

Per class AUC-ROC [DenseNet, SGD, focal Loss, with Augmentations]

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

Focal Loss – Without Augmentations, [DenseNet]

Table 4: DenseNet without augmentations

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

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

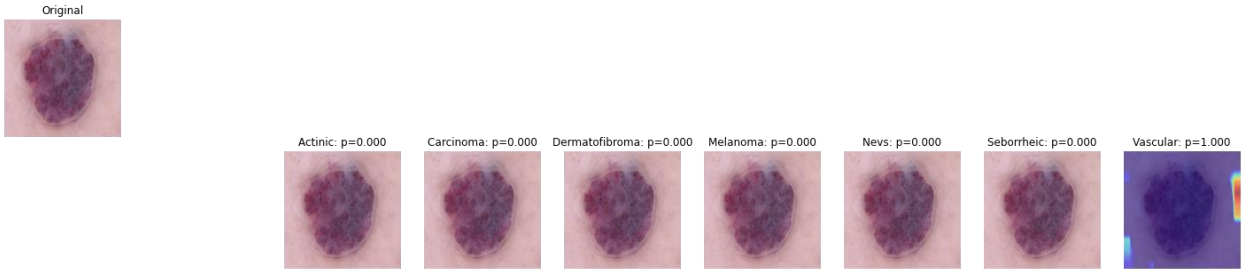
Per class AUC-ROC [DenseNet, RMSProp , focal Loss, without Augmentations]

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

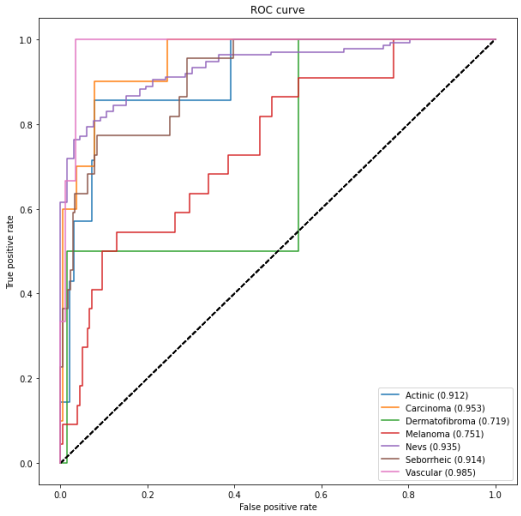
Per class AUC-ROC [DenseNet, SGD , focal Loss, without Augmentations]

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

4.1.1 Grad-CAM of DenseNet model



4.2 Inception V3 Model



Focal Loss – With Augmentations, [Inception v3]

Table 5: 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

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

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

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

Focal Loss – Without Augmentations, [Inception v3]

Table 6: Inception V3 without Augmentations

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

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

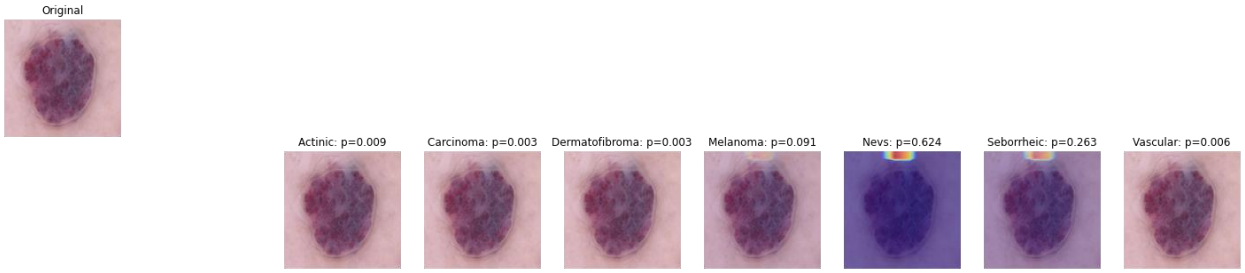
Per class AUC-ROC [Inception, RMSProp, focal Loss, without Augmentations]

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

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

4.2.1 Grad-CAM of Inception V3



This is Grad-CAM for inception V3. Both the grad-CAMs are for vascular class image and it is interesting to notice that DenseNet predicted the image correctly and inception v3 miss predicted the image. The overall performance of DenseNet is better than inception V3 in terms of evaluations metrics.

5. 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 hyperparameter, 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. The 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.

6. 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:

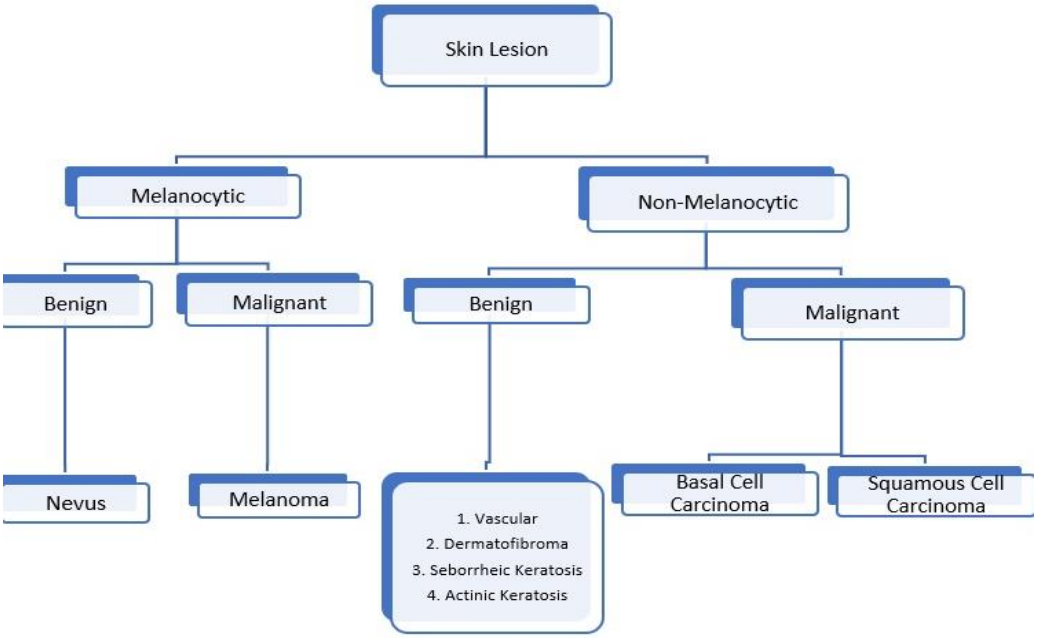


Figure 17:skin lesion hierarchy

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.

REFERENCES

[1] M. A. L. I. Albahar, "Skin Lesion Classification Using Convolutional Neural Network With Novel Regularizer," *IEEE Access*, vol. 7, pp. 38306–38313, 2019, doi: 10.1109/ACCESS.2019.2906241.

[2] I. A. Ozkan and R. Garcia, "Skin Lesion Classification using Machine Learning Algorithms," doi: 10.18201/ijisae.2017534420.

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

[4] T. Devries, "Skin Lesion Classification Using Deep Multi-scale Convolutional Neural Networks," 2017.

[5] A. C. Salian, "Skin Lesion Classification using Deep Learning Architectures," pp. 168–173, 2020.

[6] A. R. Lopez and X. Giro-i-nieto, "SKIN LESION CLASSIFICATION FROM DERMOSCOPIC IMAGES USING DEEP LEARNING TECHNIQUES."

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] S. Albawi, T. A. M. Mohammed, and S. Alzawi, "Understanding of a Convolutional Neural Network," *Ieee*, p. 16, 2017.

[14] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *Proc. - 30th IEEE Conf. Comput. Vis. Pattern Recognition, CVPR 2017*, vol. 2017-Janua, pp. 2261–2269, 2017, doi: 10.1109/CVPR.2017.243.

[15] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception Architecture for Computer Vision," *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, vol. 2016-Decem, pp. 2818–2826, 2016, doi: 10.1109/CVPR.2016.308.

[16] T.-Y. L. P. G. R. G. K. H. P. Dollar´ and F. A. R. (FAIR), "Focal Loss for Dense Object Detection," *13C-NMR Nat. Prod.*, pp. 30–33, 1992, doi: 10.1007/978-1-4615-3288-0_5.

[17] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *Int. J. Comput. Vis.*, vol. 128, no. 2, pp. 336–359, 2020, doi: 10.1007/s11263-019-01228-7.