A PROPOSED MODEL FOR DETECTION AND PREDICTION OF SKIN DISEASE USING MACHINE LEARNING TECHNIQUES

MINOR PROJECT REPORT

By:

Roshni Singh (19103034)

Debshishu Ghosh (19103082)

Rishabh Lal Srivastava (19103088)

Supervised by: Dr. Bharat Gupta

CSE & IT,

JAYPEE INSTITUTE OF INFORMATION TECHNOLOGY, NOIDA

TABLE OF CONTENTS

Chapters	Topics	Page no.
Chapter - 1	**Introduction**	**6**
	1. General Introduction 2. Motivation	
Chapter - 2	**Literature Study**	**7**
	1. Research Papers 2. Gaps in research	
Chapter - 3	**Problem statement and propositions**	**8 - 10**
	1. Problem Statement 2. Proposed Solution 3. Proposed Architecture	
Chapter - 4	**Dataset analysis**	**11 - 19**
	1. The Dataset 1.1. Sample Images 2. Exploratory Data Analysis	
Chapter - 5	**The Code**	**20 - 23**
	1. Data Refining 1.1. Data Pre-processing 1.2. Train - Test Data Splitting 2. The Training model 2.1. Optimiser and Annealer 2.2. Data Augmentation 2.3. Data Fitting	
Chapter - 6	**Results**	**24 - 28**
	1. Experiments and Results 1.1. Model Summary 1.2. Model Evaluation 1.3. Graphs 2. Model Usage 3. Conclusion 4. Future Work	
References		**29**

Students' Self Declaration for Open Source libraries and other source code usage in Minor Project

We Roshni Singh (19103034), Debshishu Ghosh (19103082), Rishabh Lal Srivastava (19103088), hereby declare the following usage of the open-source code and prebuilt libraries in our minor project in the 5th Semester with the consent of our supervisor. We also measure the similarity percentage of pre-written source code and our source code and the same is mentioned below. This measurement is true to the best of our knowledge and abilities.

1. List of pre-build libraries

Numpy, Pandas, Seaborn, Matplotlib, Keras, Tensorflow

2. List of pre-built features in libraries or source code.

Sequential Model, Graph plots, Mathematical Operations, Image preprocessing, Dataset import, Model Import/export

3. Percentage of pre-written source code and source written by us.

Self-Written Source code - 75%	Pre-written source code - 25%

Student ID	**Student Name**	**Student signature**
19103034	Roshni Singh	Roshni Singh
19103082	Debshishu Ghosh	Debshishu Ghosh
19103088	Rishabh Lal Srivastava	Rishabh Lal Srivastava

DECLARATION BY SUPERVISOR

I, ...(Name of Supervisor) declares that I above submitted project with Titled

..... was conducted under my supervision. The project is original and neither the project was copied from External sources nor it was submitted earlier in JIIT. I authenticate this project.

(Any Remarks by Supervisor)

Signature (Supervisor)

ACKNOWLEDGEMENT

We convey our heartfelt thanks to Dr. Bharat Gupta, our mentor for making every session interactive and interesting. We also thank him for being patient and guiding us all through the project. We thank our institution for providing us with an opportunity to develop this minor project, which is sure to play a significant part in our career and any future interviews that we are to face.

Name of students:	Roshni Singh (19103034)	Debshishu Ghosh (19103082)	Rishabh Lal Srivastava (19103088)
Date:	December 2021		

ABSTRACT

A skin illness classification and identification system have been created as a part of this project. The proposed system takes photographs from a camera as input and determines which skin condition the individual has acquired as an output. In this situation, we focused on skin lesions as our major goal, and we employed a variety of algorithms to classify and identify the condition that was presented to the system. This project was created to provide a faster, less expensive, and more accurate diagnosis of major skin illnesses that, if not detected on time, can be fatal to humans. This is also aimed at assisting the poorer members of our society in surviving these diseases.

Chapter 1

1.1 INTRODUCTION

In our daily lives, skin problems are a typical occurrence. They can appear in any part of the body and come in a variety of shapes and sizes. Most of the time, we dismiss them as a common occurrence, unconcerned about the severity of the problem, and in some cases, we are unable to recognize them as a disease. They have such a wide range of appearances that even when they are identified, we are unsure of which disease they are and thus how to treat them.

This can be problematic because some diseases are extremely serious and, if not treated properly and promptly, can be fatal. We don't consider their concern because they happen so frequently, but we should still take care of our skin so that we don't get any further difficulties and stay healthy. In the case of carcinogenic disorders, if they are not detected early on, they can develop into a much more serious condition over time, leading to full-fledged cancer that is then incurable.

As a result, we devised a plan to address the problem of disease detection and attempted to do so. As a result, this initiative is focused on disorders that may be quickly discovered and diagnosed using an image or picture of the abnormality, and thus treated as soon as feasible. This will aid in the saving of many lives from fatal diseases as well as faster and more accurate treatment of diseases without spending time on diagnosis.

1.2 MOTIVATION

Our biggest motivation for this project is to help the people who are suffering from such diseases and are present in areas where healthcare is inaccessible, or just too expensive. There this model would prove to be very helpful and save the lives of the people who require even the most basic diagnosis of the skin disease all free of cost. All they would need is a phone and a stable internet connection.

This system can also be used by doctors to make their work easier and for providing a more accurate analysis for them while they focus on the treatment of the disease. This would effectively lower the cost of their diagnosis and help them in saving lives in time without delays or confusion.

This method can be used by anyone in any area as long as they have access to the internet. It would enable rapid and accurate identification of the disease, allowing for quicker treatment. Because it is free software, it may be used by the poor so that they can receive good treatment and not have to pay merely for the disease's diagnosis.

Chapter 2

2.1 LITERATURE STUDY

A small research was conducted in order to find out the developments made in this field and we drew the following conclusions from the analysis of three research papers.

2.1.1 RESEARCH PAPERS:

1. Nawal Soliman AL Kolifi ALEnezi, in this paper, analyzed how Machine Learning could help humans with diagnosing various diseases and talks about a CNN model with SVM style architecture devised for a small dataset that is trained and used for that project.[1]

2. Saja Salim Mohammed and Jamal Mustafa Al-Tuwaijari, have prepared a full case study of various Machine Learning models of different countries that have been compared according to their respective algorithms, accuracies, the number of images processed, the types of diseases classified and the dataset used. It gives a comprehensive analysis of which algorithms to use to get better results and what has been worked on.[2]

3. This project is about a classification system that has been made for the K.E.M. hospital in Parel, Mumbai. This classification system is made using 5 main algorithms, namely, ANN, KNN, SVM, Decision Trees, Random Forest. The latter two algorithms were used because this is a system that doesn't use images for classifying the disease. The average accuracy of the system is 98.94%.[3]

2.1.2 GAPS:

1. The dataset used for the first project [1] is a very small dataset of just 80 images and thus boasts an accuracy of 94.01% at the second stage.
2. It has used only one algorithm system for classifying the image. [1]
3. The varied datasets used also create more variables that cause changes in the accuracy due to the inherent faults of some datasets.[2]
4. Not all skin diseases can be classified by the use of one system.[2]
5. The third system doesn't use image-based classification, hence a doctor's inference is needed first before the system can analyse.[3]
6. Since it is a manual input system, the input method is slow and needs a lot of parameters making it an inefficient system. [3]
7. It also deals with many diseases however it is still an incomplete model and needs improvement in the number of diseases and the classification.
8. None of the systems designed in them are capable of a simple direct analysis with just the input of a photo.
9. Systems aren't capable of classifying all types of skin diseases due to the similarity between the looks of the diseases.

The gaps and issues of the research have been identified. These researches aren't made for the general masses for diagnosis which is hence our primary goal.

Chapter 3

3.1 PROBLEM STATEMENT

With the shortcomings of the above research in mind, our primary problems are to rectify the gaps found in the systems above and improve them.

Many people may avoid visiting the doctor's office for a variety of personal reasons. They are also too expensive for many people, and just a few people have access to competent healthcare. Hence this harms the economically backward sections of society and causes many fatalities as a result.

The characteristics of the images are diversified so that it is a challenging job to devise an efficient and robust algorithm for automatic detection of the disease and its severity. Skin tone and skin color also play an important role in skin disease detection. Most diseases cannot be classified with just one method.

3.3 PROPOSED SOLUTION

To overcome the above problem the project aims at building a model which is used for the prevention and early detection of skin diseases in a fast and cost-efficient way. An application is built where a person can upload an image from the UI, then the image will be sent to the trained model. The model analyses the image and detects the skin disease that person had. Our system will use Convolution Neural Networks (CNN) to train the images of skin diseases.

This system will allow any users with internet access to upload an image into the database and get results, identifying the disease for them. This result can then be used by the user to take the necessary precautions needed to cure the acquired disease.

This system could also be used by doctors to identify a disease correctly and then verify their diagnosis. New doctors can also learn from this as it'd be more accurate in general and would be a very helpful tool in assisting doctors of the current age, reducing their stress, and allowing them to work more efficiently.

ALGORITHMS AND TECHNIQUES USED:

The algorithms and techniques used for the comparisons are:

1. **CNN (Convolutional Neural Networks):** A **CNN** [6]is a Deep Learning algorithm that can take in an input image, assign importance (learnable weights and biases) to various aspects/objects in the image, and be able to differentiate one from the other.
2. **ANN (Artificial Neural Networks):** An Artificial neural network [7] is usually a computational network based on biological neural networks that construct the structure of the human brain.

The above algorithms would be used for analyzing and predicting the disease acquired by the user. For the sake of a higher prediction rate and more accurate analysis, more algorithms are being used to identify one disease. Then after comparing the accuracy rates of all the algorithms, the most accurate result will be displayed to the user along with accuracy percentages (if possible).

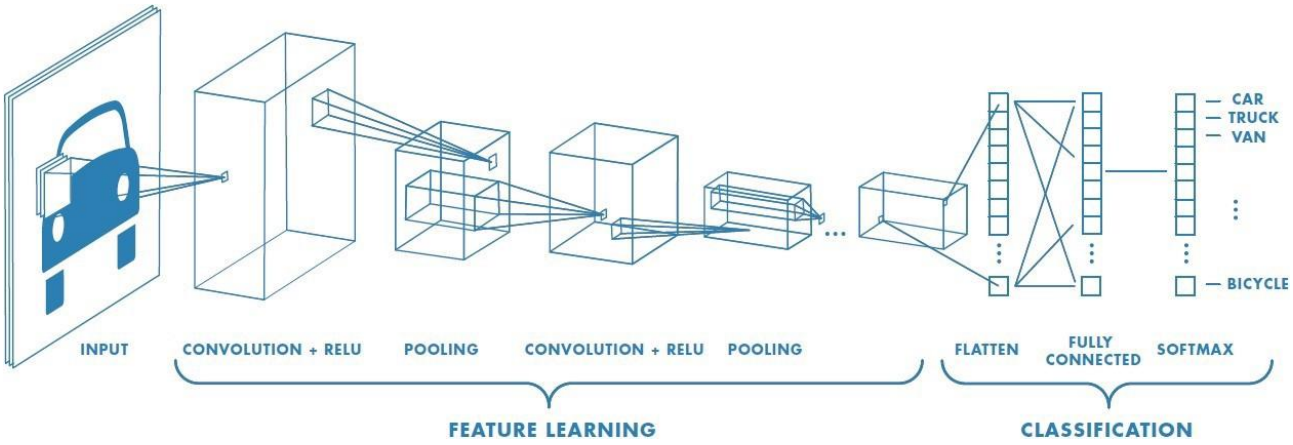


Fig 3.2: *CNN algorithm[6]*

3.2 PROPOSED ARCHITECTURE

An architecture has been proposed which addresses the gaps identified in the literature study. The first step to rectifying the solution is importing the dataset This dataset contains the images for analysis and their metadata in CSV format. The details of the dataset have been given below in the dataset section.

The next step is preprocessing the data then Exploratory data analysis (EDA) which has been obtained. Data preprocessing is wherein we make the raw data into a bit more useful information by setting new variables, cleaning the errors, and filling blank places which can cause errors. EDA is done so that any useful information on the data can be obtained and used for building the CNN model.

Then we resize the data and load the images into the system. The data splitting is done right after, which splits the dataset into the training section and the testing section in ratio 80:20. The testing section is where 20% of the data is kept and will be used for evaluating the model and checking its accuracy.

After this, we normalise the data and images so that the resizing doesn't cause too much noise in the system and the images aren't too distorted. Then one-hot encoding is done on dependent training and testing set which helps categorical variables such as images in this case, are converted into Binary form and is much better understood by the machine on which our model is being trained. It converts the given data into a much lower language than what we as humans would prefer.

The CNN model is made soon after which contains the processes for the classification system and how the image classification will happen in the randomised training set. The data is then augmented and fitted into the model for the classification which basically classifies the randomised data by analysing the images and processing them through the model. It is trained in several steps in epochs and the data is slowly and accurately classified.

After this step we analyse the correctness of our model by comparing it to the test set and validate it. This step gives us the accuracy of the training model and lets us analyse it for further improvements which can be made to the model.

After this we have used this saved model to put it in a system where we can analyze new images by using it. Here we have saved the model and imported it to a different colaboratory and inserted new images in it for proper analysis of the disease. A proper GUI will be created afterward to streamline this process and enable it for all users across the internet. The flow chart of this basic architecture is given below and describes how this project was done in steps. [fig (3.1)]
The proposed architecture for the model is given below:

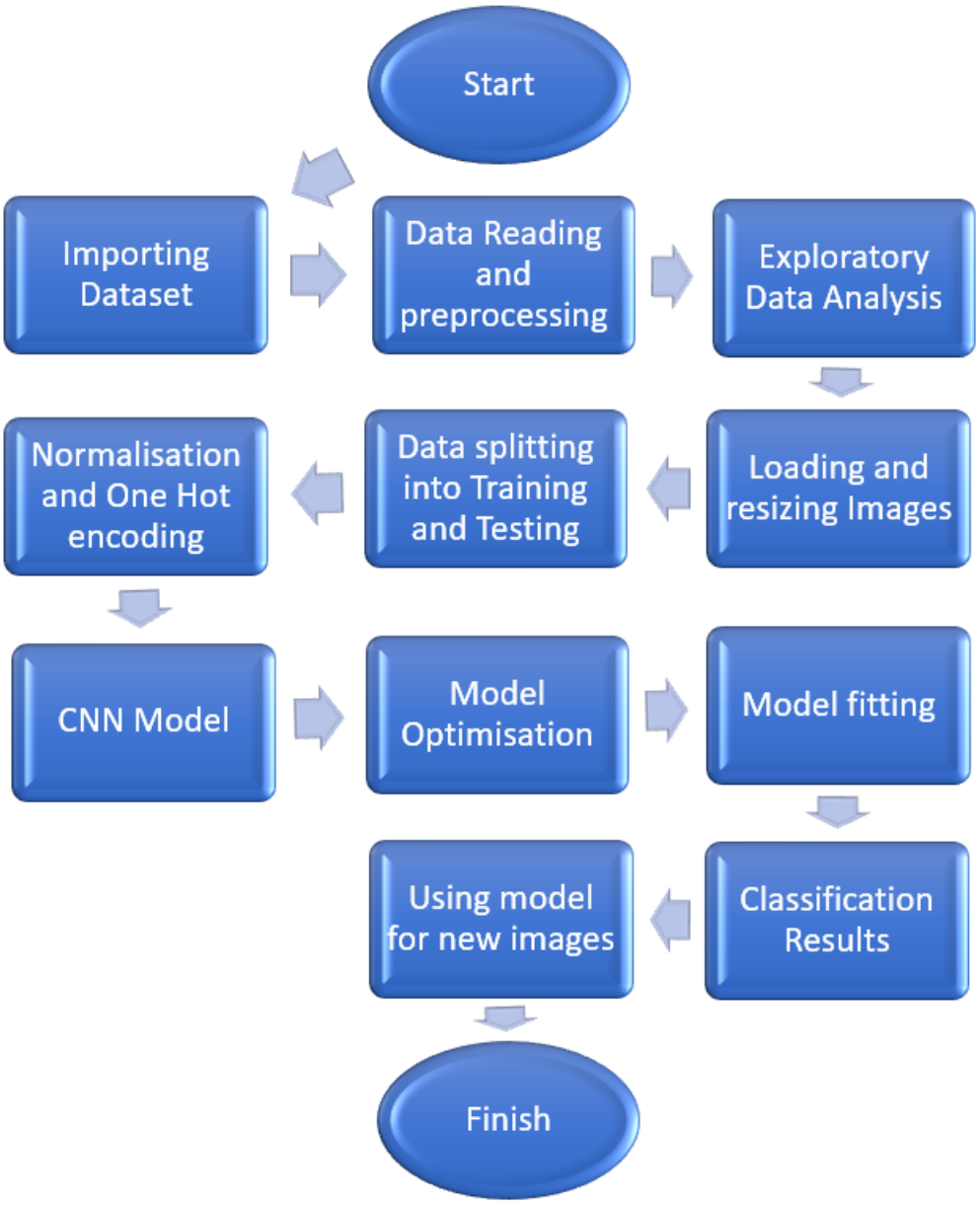


Fig 3.1: Basic Architecture

Chapter 4

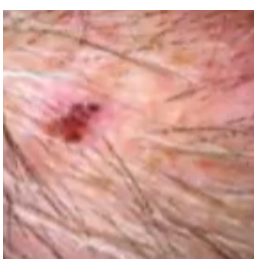
4.1 The Dataset

The HAM10000 dataset [7] consists of 10015 dermatoscopic images of a size of 450 × 600. Cases include a representative collection of all important diagnostic categories in the realm of pigmented lesions: It consists of seven diagnostic classes as follows:

1. **Melanoma(MEL):** Melanoma is a disorder in which melanocytes develop malignant (cancer) cells (cells that colour the skin). There are a variety of cancers that begin in the skin. Melanoma can develop anywhere on the body's surface. Melanoma risk is influenced by unusual moles, sun exposure, and medical history.
2. **Melanocytic Nevi(NV)**: This mole is usually big and is caused by a disease involving melanocytes, or pigment-producing cells (melanin). Melanocytic nevi may be rough, flat, or elevated in appearance. They can be present at birth or develop later in life. The majority of cases do not necessitate treatment, however some do necessitate mole removal.
3. **Basal Cell Carcinoma(BCC):** Basal cell carcinoma is a kind of skin cancer. Basal cell carcinoma starts in the basal cells, which are a type of skin cell that creates new skin cells when the old ones die. Basal cell carcinoma usually shows as a small, translucent lump on the skin, but it can also occur in different ways.
4. **Actinic Keratosis, and Intra-Epithelial Carcinoma(AKIEC):** A rough, scaly area on the skin is called Actinic Keratosis develops after years of sun exposure. It commonly appears on the cheeks, lips, ears, scalp, neck, and backs of hands. An actinic keratosis, also known as a sun keratosis, develops slowly and commonly appears in adults over the age of 40.
5. **Benign Keratosis(BKL):** A seborrheic keratosis is a benign (noncancerous) skin development. It might be white, tan, brown, or black in hue. The majority of them are elevated and appear to be adhered to the skin. They may resemble warts. Seborrheic keratoses can occur on the chest, arms, back, or other parts of the body.
6. **Dermatofibroma(DF):** Dermatofibroma (superficial benign fibrous histiocytoma) is a common cutaneous lesion with an unknown cause that affects women more frequently. Dermatofibroma mainly affects the extremities (particularly the lower legs) and is asymptomatic, though it can cause itching and pain.
7. **Vascular lesions(VASC).:**Birthmarks are vascular lesions, which are very common anomalies of the skin and underlying tissues. Hemangiomas, Vascular Malformations, and Pyogenic Granulomas are the three main types of vascular lesions.

4.1.1 SAMPLE IMAGES





4.2 EXPLORATORY DATA ANALYSIS

In this project, we have used 11 main libraries for the entire project which are, NumPy, pandas, os, seaborn, matplotlib, keras, sklearn, TensorFlow, glob, Image and the drive library. The NumPy and pandas libraries have been used to import, extract, read and refine the data available in the datasets. The seaborn and matplotlib libraries are used for EDA and to show meaningful analysis of the used dataset. Keras, sklearn and TensorFlow are the most important libraries used which are needed for the machine learning model building, testing and training for final use. The drive and os libraries are simply for connections and data transfer.

For the exploratory data analysis using seaborn and matplotlib, we have analysed a few things from the dataset given. Below is a figure of the metadata of the dataset used by us for the project. Here we can see the various columns like the image ID, the diagnosis type, the age of the participant, gender, the localisation, the image path location, the cell type and the cell type index. These are the primary parameters used for the classification of the images.

```
skin_df.head()
```

	lesion_id	image_id	dx	dx_type	age	sex	localization	Path	cell_type	cell_type_idx
0	HAM_0000118	ISIC_0027419	bkl	histo	80.0	male	scalp	/content/drive/MyDrive/HAM10000_images_part_1/...	Benign keratosis-like lesions	2
1	HAM_0000118	ISIC_0025030	bkl	histo	80.0	male	scalp	/content/drive/MyDrive/HAM10000_images_part_1/...	Benign keratosis-like lesions	2
2	HAM_0002730	ISIC_0026769	bkl	histo	80.0	male	scalp	/content/drive/MyDrive/HAM10000_images_part_1/...	Benign keratosis-like lesions	2
3	HAM_0002730	ISIC_0025661	bkl	histo	80.0	male	scalp	/content/drive/MyDrive/HAM10000_images_part_1/...	Benign keratosis-like lesions	2
4	HAM_0001466	ISIC_0031633	bkl	histo	75.0	male	ear	/content/drive/MyDrive/HAM10000_images_part_2/...	Benign keratosis-like lesions	2

Upon analysis of the dataset, we obtained the first graph (fig 4.1.1) that gives us the count of the cell types that we have classified. Here we have 7 types of lesions that we are going to classify for the project namely: Melanocytic Nevi, Melanoma, Benign Keratosis-like lesions, Basal cell carcinoma, Actinic Keratosis, Vascular lesions and Dermatofibroma.

The first graph (fig 4.1.1) shows us that Melanocytic nevi has the highest amount of cell count and hence the most amount of images consist of it. This is hence also the most commonly acquired lesion by inferring this data and from a general study. The next is melanoma which is the most serious condition of skin cancer that exists and is far less common. The rest of the diseases, although common, are not very serious conditions and hence have fewer samples in the dataset. Except for melanoma and basal cell carcinoma, the rest are non-cancerous types of lesions and aren't serious conditions.

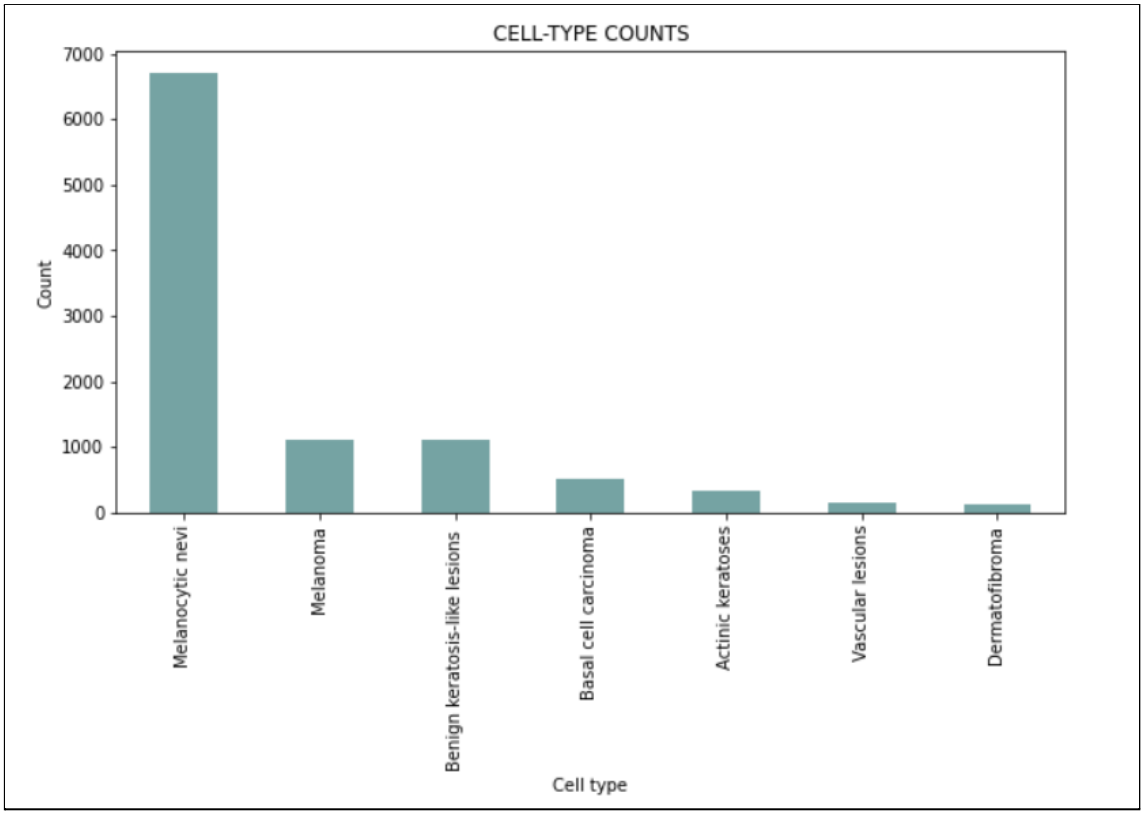


Fig 4.1.1: Cell/Image count of the different cell types

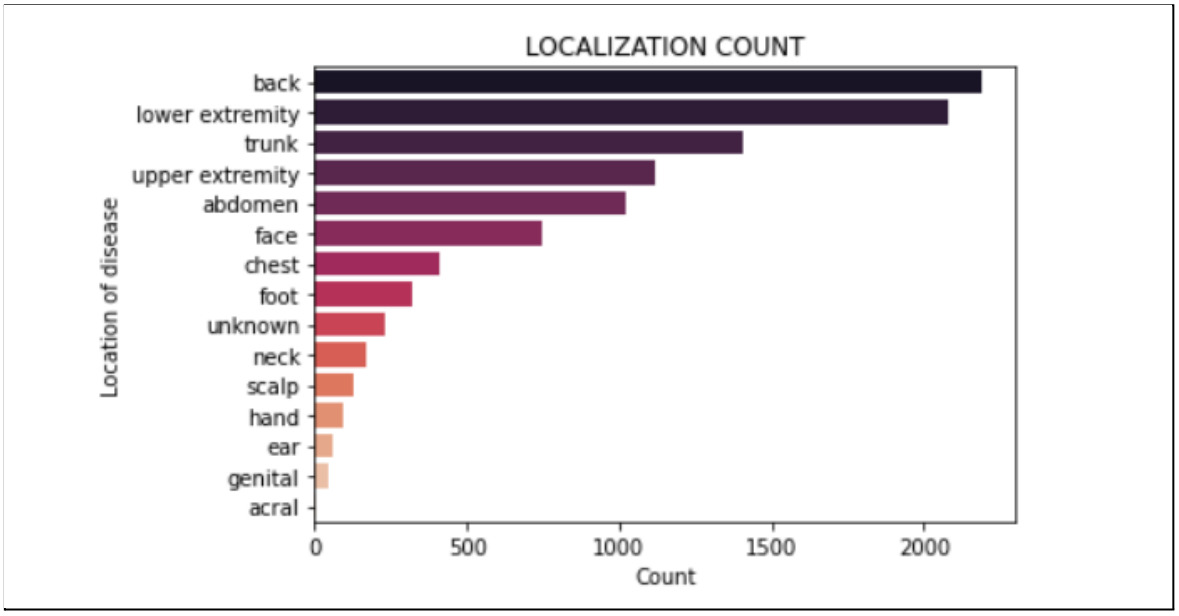


Fig 4.1.2: Localization of the cells in different parts of the body

The above graph (fig 4.1.2) shows us the areas which were affected the most by these diseases. Here we observe that the images taken originate mostly from the back, the lower extremity and the trunk areas, indicating that the disease may have started there due to the areas not receiving much care from being less visible. The other common areas like the face, the foot and the chest are less affected than these due to them being visible locations and receiving more care.

After this, from figures 4.2.1 to 4.2.4 we see that the most common type of lesion for this dataset (melanocytic nevi) has the back, the trunk and the lower extremity as the most commonplace of occurrence. The face is the common location for Benign - keratosis like lesions, dermatofibroma mainly occurs in the lower extremity, and melanoma occurs in the lower extremity, upper extremity, and the back. The rest of the diseases also have images that originate from the lower extremity, upper extremity, and back commonly. In the following few graphs, we see the areas that are affected the most in accordance with the seven different types of diseases that we have analysed.

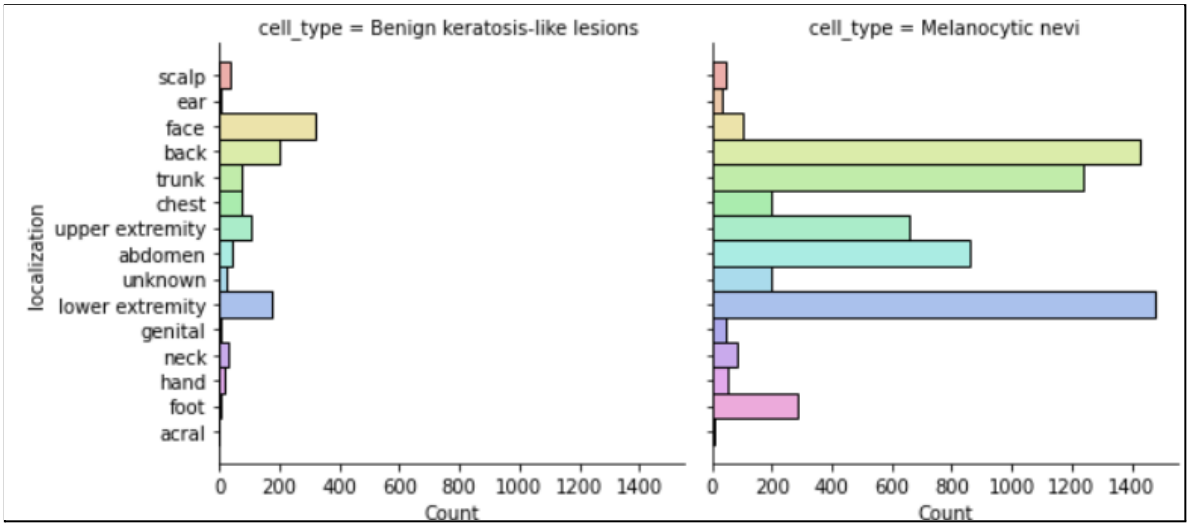


Fig 4.2.1: Benign keratosis and melanocytic nevi

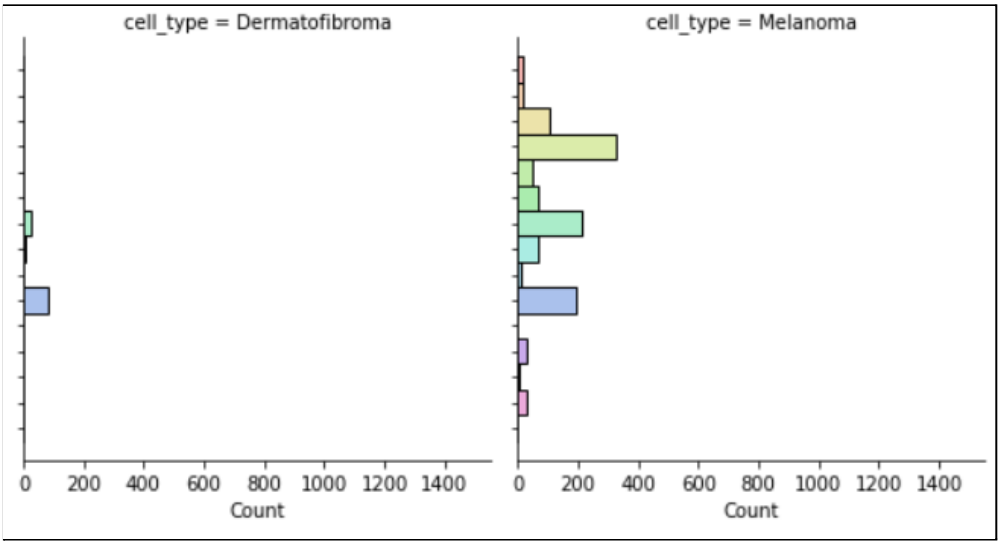


Fig 4.2.2: Melanoma and Dermatofibroma

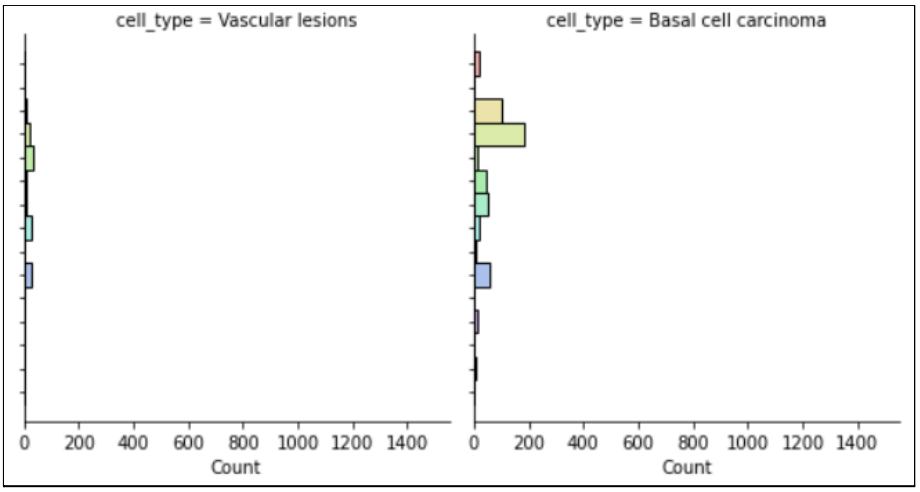


Fig 4.2.3: Vascular Lesions and Basal Cell Carcinoma

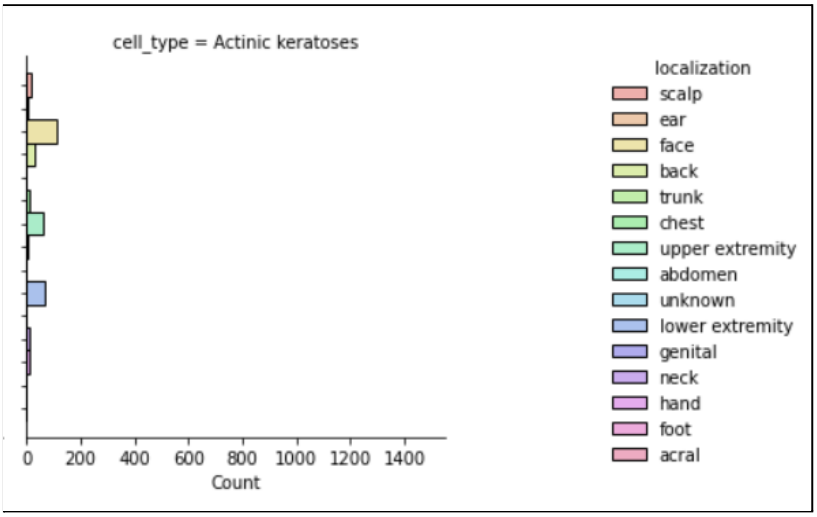


Fig 4.2.4: Actinic Keratosis and the color scheme

Now we come to the distribution of the images of disease among the various age groups. The following two graphs (fig 4.3.1, 4.3.2) show us the age of the people who have acquired the diseases and their distribution.

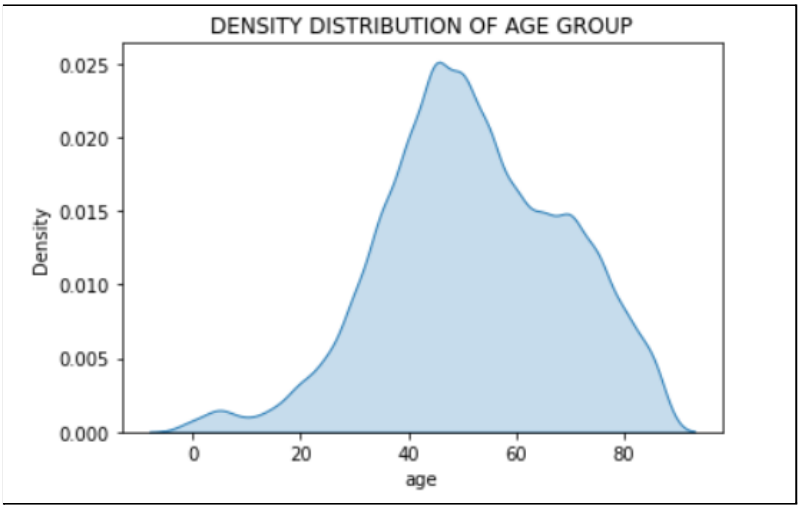


Fig 4.3.1

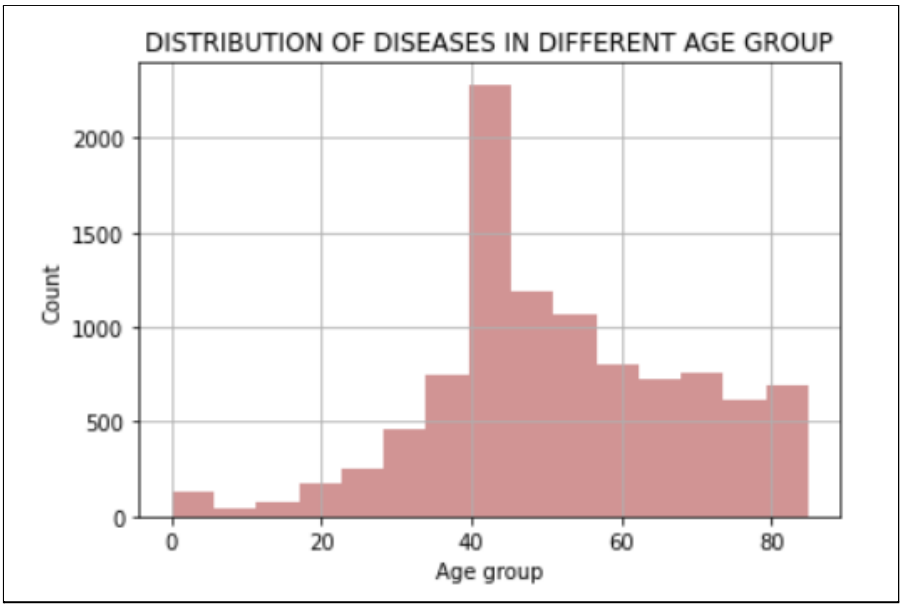


Fig 4.3.2

The scatterplot below shows us which disease is acquired among the various age groups from the list. Here each dot corresponds to a certain age and a cell type.

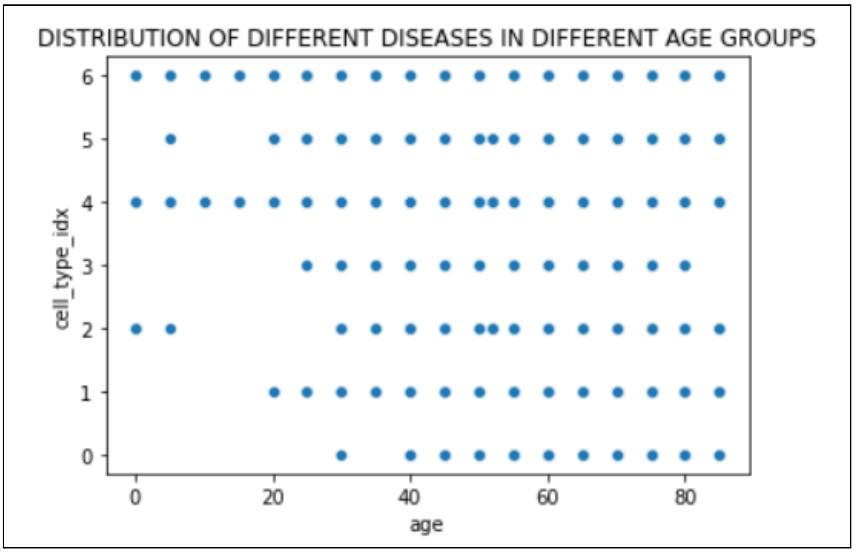


Fig 4.3.3

Now about the gender statistics for the dataset, we first see that the number of males in the images taken is slightly higher than that of females, with some data being unknown. The following graphs show us the distribution of the diseases acquired by males and females (fig 4.4.2). Here we see no major difference in the graphs and both genders are equally susceptible to all the diseases.

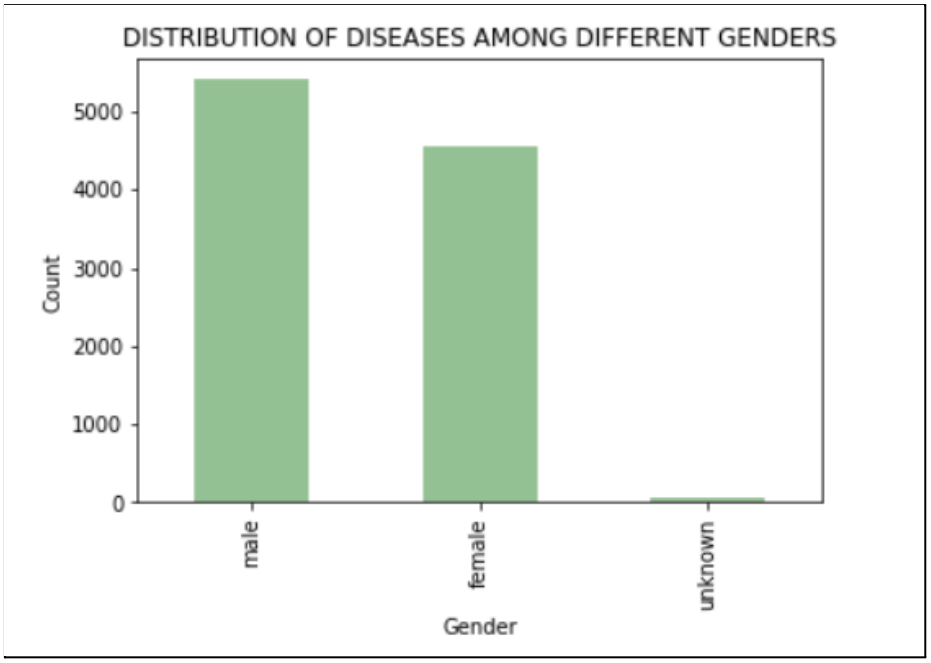


Fig 4.4.1

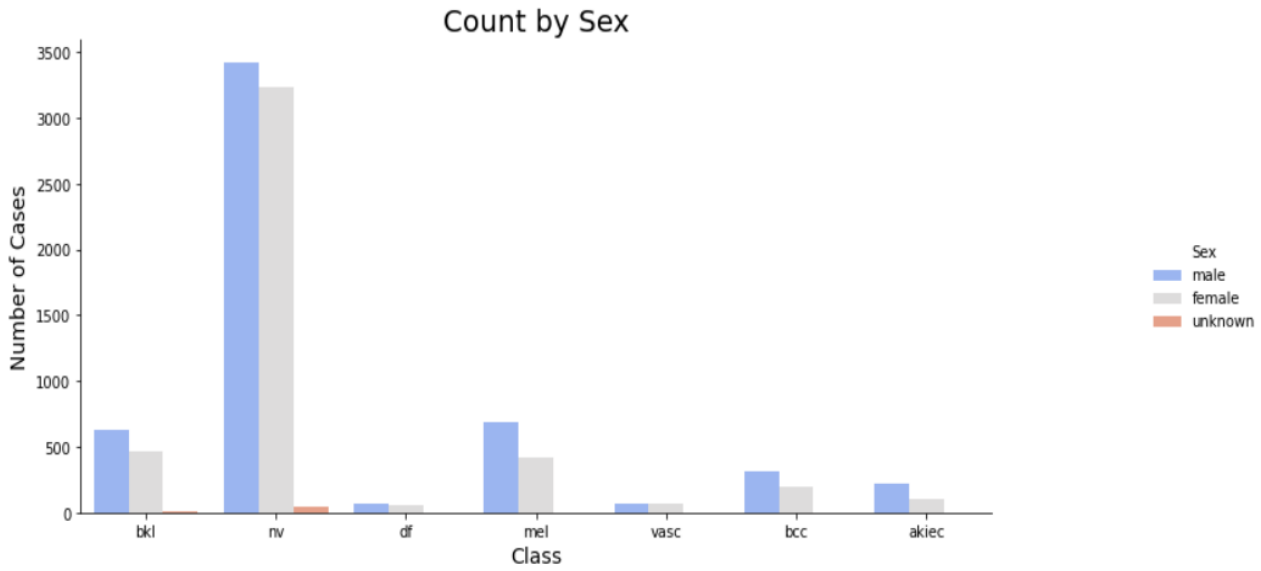


Fig 4.4.2

Next, are graphs of the treatment types used commonly first (fig 4.5.1), then for each individual disease (fig 4.5.2 to 4.5.3). The treatment types are the following:

1. Histopathology(Histo): Histopathologic diagnoses of excised lesions have been performed by specialized dermatopathologists.
2. Confocal: Reflectance confocal microscopy is an in-vivo imaging technique with a resolution at near-cellular level, and some facial benign with a grey-world assumption of all training-set images in Lab-color space before and after manual histogram changes.
3. Follow-up: If nevi monitored by digital dermoscopy did not show any changes during 3 follow-up visits or 1.5 years biologists accepted this as evidence of biologic benignity. Only nevi, but no other benign diagnoses were labeled with this type of ground-truth because dermatologists usually do not monitor dermatofibromas, seborrheic keratoses, or vascular lesions.
4. Consensus: For typical benign cases without histopathology or follow-up biologists provide an expert-consensus rating of authors PT and HK. They applied the consensus label only if both authors independently gave the same unequivocal benign diagnosis. Lesions with this type of ground truth were usually photographed for educational reasons and did not need further follow-up or biopsy for confirmation.

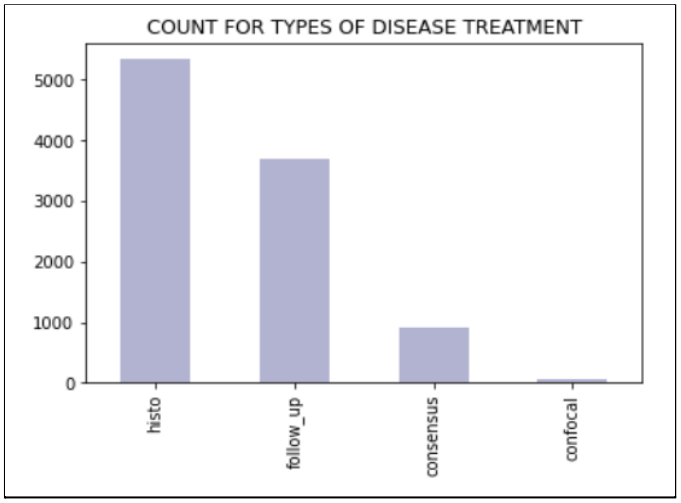


Fig 4.5.1

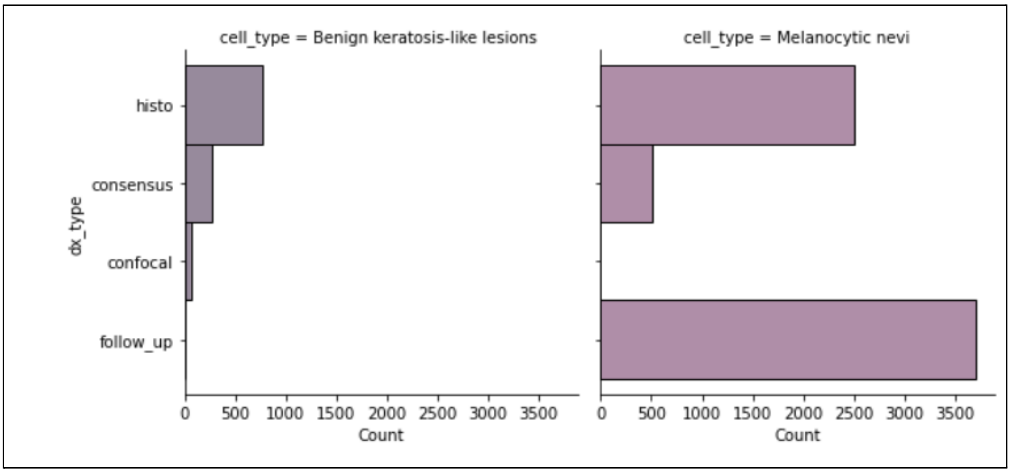


Fig: 4.5.2

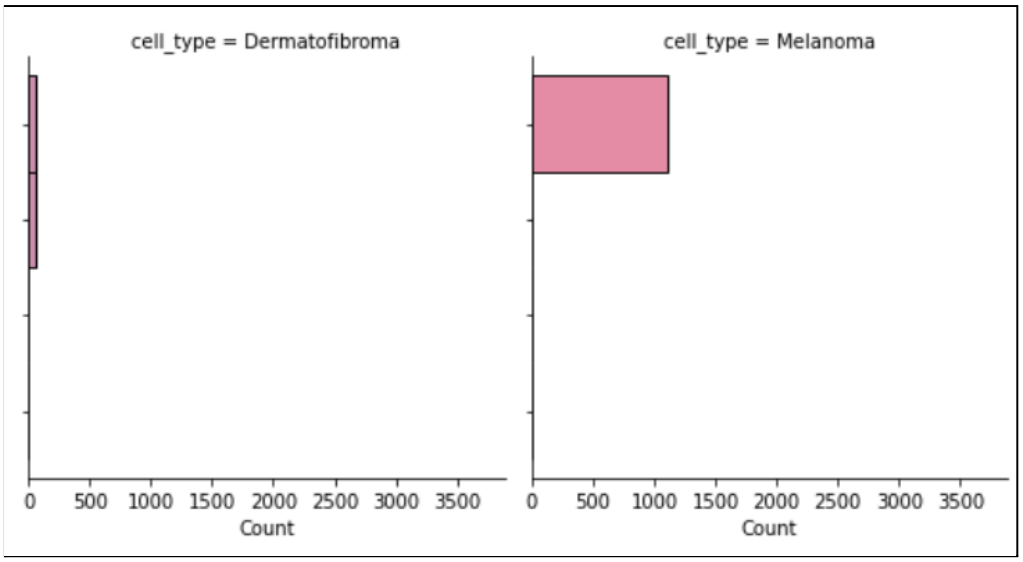


Fig: 4.5.3

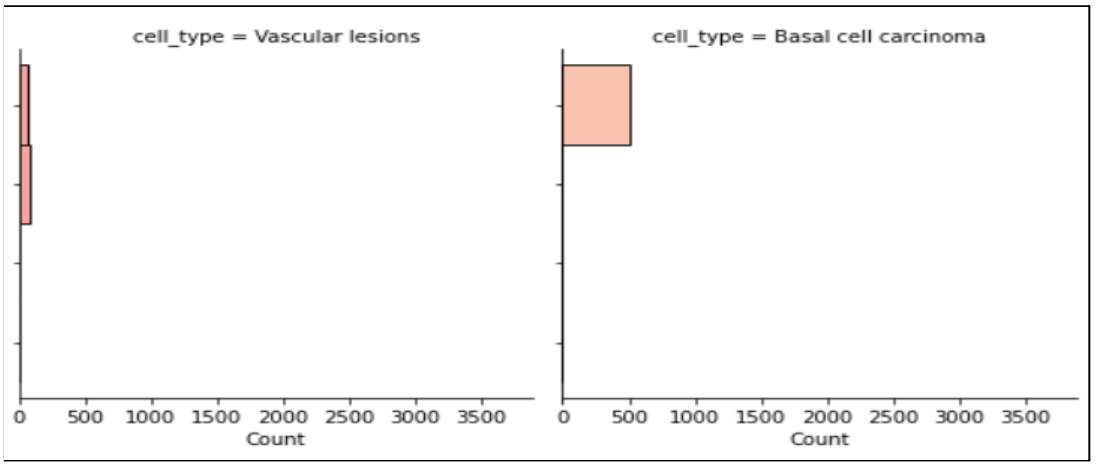


Fig: 4.5.4

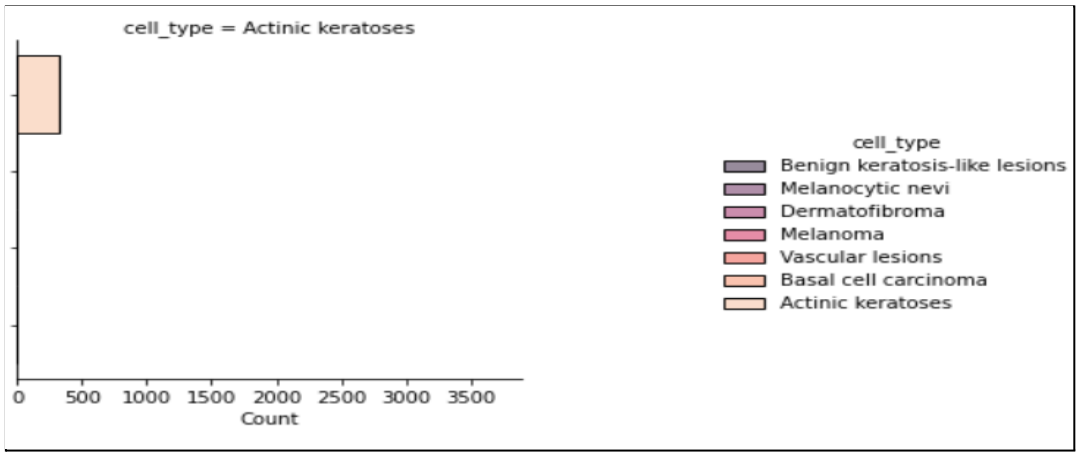


Fig: 4.5.5

With this, we have finished the analysis of the data and are now ready to create the main CNN model to finally classify the dataset and create the working model.

Chapter 5

5.1 DATA REFINING

After finding out all necessary and useful information from the EDA section, we now proceed onto the creation of the main CNN model which will be responsible for the classification and for the final output after taking an image.

5.1.1 THE PRE-PROCESSING:

The first thing we do before starting our main machine learning model is pre-processing the data. Here we make it so that the data becomes much simpler for the main model to read and use. This makes the learning process faster and produces desirable results for us.

```
skin_df['image'] = skin_df['Path'].map(lambda x: np.asarray(Image.open(x).resize((100,75))))
```

We resize the images as the original dimension of images are 450 x 600 x3 which TensorFlow can't handle, so that's why we resize it into 100 x 75.In this step images will be loaded into the column named image from the image path from the image folder.

This is an output of the result:

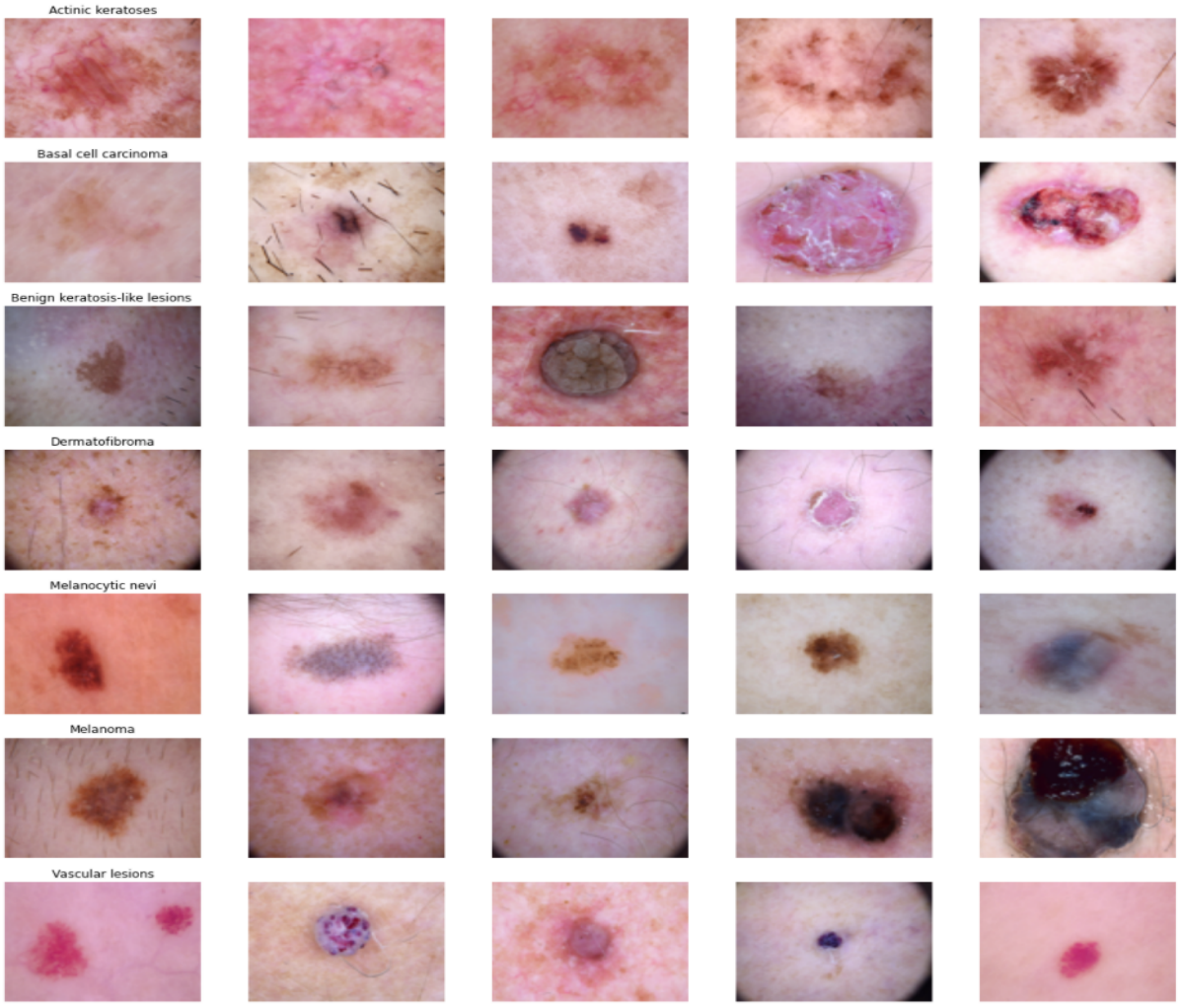


Fig: 5.1.1

5.1.2 TRAIN-TEST DATA SPLITTING:

```
X_train, X_test, Y_train, Y_test = train_test_split(X, y, test_size=0.20,random_state=100)
```

We then split the entire dataset into a train and a test dataset. We have split the one here into an 80:20 ratio of train:test. This allows us to check the data from the 20% left and the rest is randomised to train the classifier model so that it becomes capable of analysing individual images on its own. We have then normalised the dataset accordingly and performed one-hot encoding to reshape and label the data accordingly which makes this dataset finally ready to be analysed.

5.2 THE TRAINING MODEL:

The CNN architecture used for the model is:

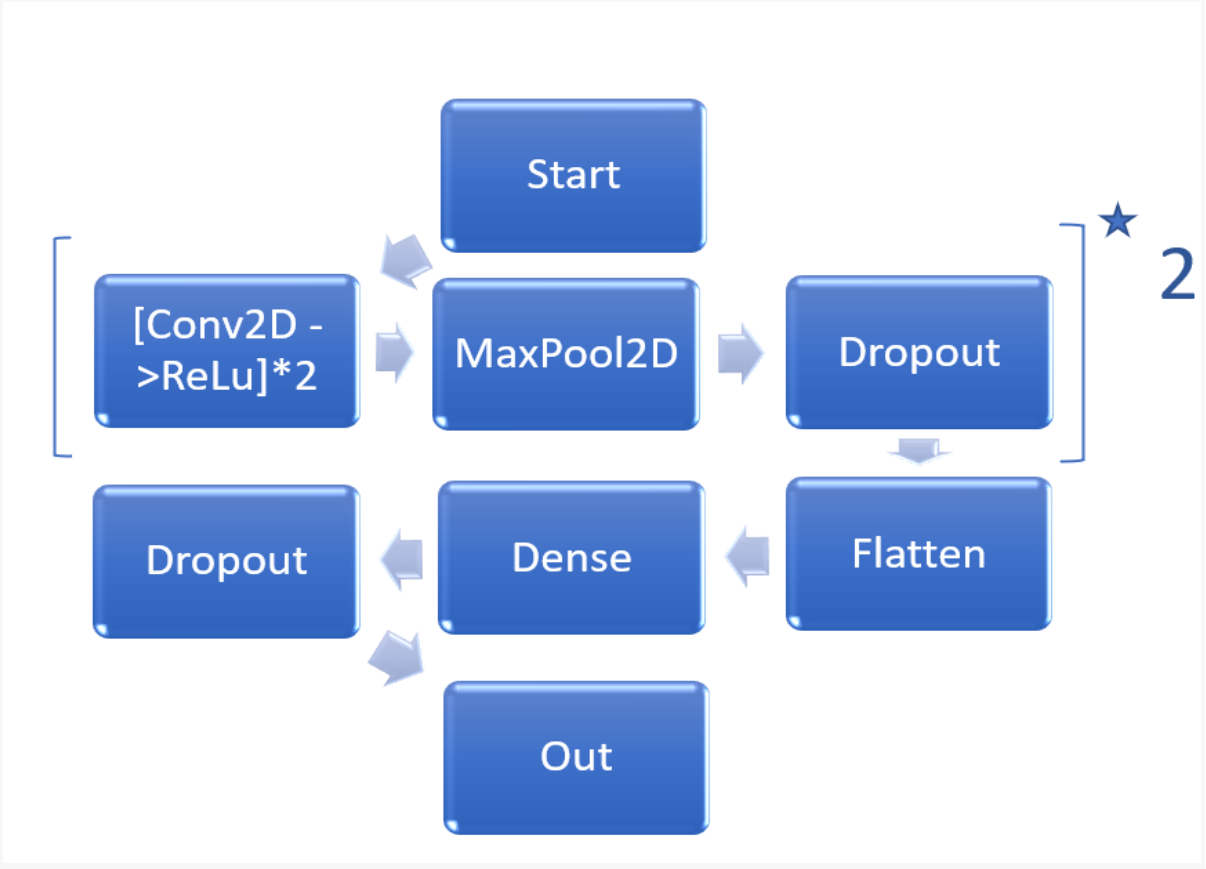


Fig: 5.2.1

We utilised the Keras Sequential API, which allows you to start from the input and add one layer at a time. The convolutional (Conv2D) layer is the first. It's similar to a series of programmable filters. For the first two conv2D layers, we chose 32 filters, and for the latter two, 64 filters. Using the kernel filter, each filter transforms a portion of the image (specified by the kernel size). On the entire image, the kernel filter matrix is applied. Filters can be thought of as image transformations. From these modified images, the CNN can extract features that are useful elsewhere (feature maps).

The pooling (MaxPool2D) layer is the second most essential layer in CNN. Simply said, this layer is a downsampling filter. It compares the values of two adjacent pixels and chooses the one with the highest value. These are used to cut down on computing costs and, to a degree, overfitting. The pooling size (i.e. the area size pooled each time) must be chosen; the larger the pooling dimension, the more essential downsampling is. CNN can aggregate local features and learn more global properties of an image by combining convolutional and pooling layers.

Dropout is a regularisation strategy in which a percentage of nodes in a layer are disregarded (their weights are assigned to zero) at random for each training sample. This forces the network to learn features in a distributed manner by dropping a proportion of the network at random. This method also enhances generalization and decreases overfitting. 'relu' stands for rectifier (maximum activation function) (0,x). The rectifier activation function is utilized to give the network non-linearity. To transform the final feature maps into a single 1D vector, utilize the Flatten layer. This flattening phase is required so that completely linked layers may be used following convolutional/max pool layers. It incorporates all of the previously discovered local characteristics from the convolutional layers.

Finally, we used the features in two dense (completely connected) layers to create an artificial neural network (ANN) classifier. The net produces the probability distribution of each class in the last layer (Dense(10,activation="softmax")).

5.2.1 OPTIMISER AND ANNEALER:

After the model building step we use an optimiser to iteratively improve the parameters to reduce the loss. The Adam optimiser is used here because it combines the advantages of two other extensions of stochastic gradient descent. Specifically:

1. Adaptive Gradient Algorithm (AdaGrad) increases performance on problems with sparse gradients by maintaining a per-parameter learning rate (e.g. natural language and computer vision problems).
2. Root Mean Square Propagation (RMSProp) also preserves per-parameter learning rates that are adjusted based on the average of recent gradient magnitudes for the weights (e.g. how quickly it is changing). This indicates that the technique is effective for both online and non-stationary issues (e.g. noisy).

Adam understands the value of AdaGrad and RMSProp. Adam is a popular deep learning method since it produces good results quickly. Our model's performance is assessed using the metric function "accuracy." This metric function is similar to the loss function, with the exception that the metric evaluation results are not used for training the model (only for evaluation).

5.2.2 DATA AUGMENTATION:

After this the final step before fitting and running the model is data augmentation which helps us fix the problem of overfitting data. Overfitting is an issue which happens due to a model being trained with too much data as here it cannot completely analyse all the perfectly while also removing the noise in the system.

Data augmentation strategies are methods for changing the array representation while keeping the label the same while altering the training data. Grayscales, horizontal and vertical flips, random crops, colour jitters, translations, rotations, and many other augmentations are popular.

For the data augmentation here it was decided to rotate some training photos by 10 degrees at random. Randomly zoom some training images by 10%. Shift photos horizontally by 10% of their width at random. Shift photos vertically by 10% of their height at random. We fit the training dataset once our model is complete.

5.2.3 DATA FITTING:

In this final step the x_train and y_train have been fitted into the model. A batch size of 10 and 50 epochs were chosen for this fitting process. This ensures that we have sufficient epochs to train and no overfitting happens. The model is fitted in steps in 50 different epochs and each epoch helps in increasing the accuracy of the model. Here the learning rate changes sequentially after certain conditions have been made and changes are made accordingly.

Increasing the epochs would essentially increase the accuracy upto a certain point after which the accuracy would remain constant for the system and barely increase.

Chapter 6

6.1 EXPERIMENTS AND RESULTS

Using HAM10000 dataset [7], exploratory data analysis has been done, then resizing of images has been performed so as to fit as a column in the dataframe. Afterwards the dataset has been split into train and test sets and preprocessing is performed such as normalization and one hot encoding. After the preprocessing and splitting up the data into train and test sets, CNN model is built. The number of convolutional layers used is 2.

The model predicts the entered image to be that of a person suffering from Melanoma, Melanocytic Nevi, Basal Cell Carcinoma, Benign keratosis-like lesions, Actinic keratoses, Vascular lesions and Dermatofibroma. The predictions sometimes vary as the initial stages of a few diseases look like others. Also skin disease prediction largely depends upon the skin tone of the patient.

6.1.1 Model Summary: -

```
Model: "sequential"
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d (Conv2D)             (None, 75, 100, 32)       896

 conv2d_1 (Conv2D)           (None, 75, 100, 32)       9248

 max_pooling2d (MaxPooling2D  (None, 37, 50, 32)       0
 )

 dropout (Dropout)           (None, 37, 50, 32)        0

 conv2d_2 (Conv2D)           (None, 37, 50, 64)        18496

 conv2d_3 (Conv2D)           (None, 37, 50, 64)        36928

 max_pooling2d_1 (MaxPooling  (None, 18, 25, 64)       0
 2D)

 dropout_1 (Dropout)         (None, 18, 25, 64)        0

 flatten (Flatten)           (None, 28800)             0

 dense (Dense)               (None, 128)               3686528

 dropout_2 (Dropout)         (None, 128)               0

 dense_1 (Dense)             (None, 7)                 903

=================================================================
```

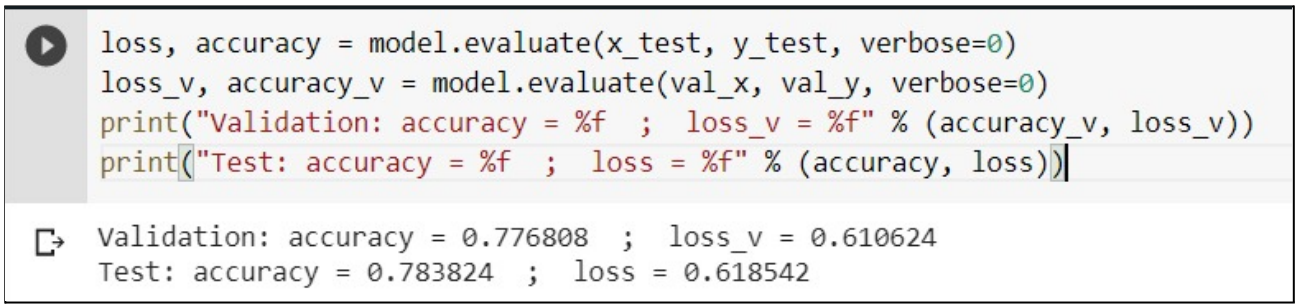
This gives us the statistics of the usage of the CNN techniques used in the system such as Conv2D, max pooling and flattening.

6.1.2 TRAINING OF MODEL: -

```
Epoch 1/50
721/721 [==============================] - 36s 33ms/step - loss: 1.0075 - accuracy: 0.6652 - val_loss: 0.8782 - val_accuracy: 0.6758 - lr: 0.0010
Epoch 2/50
721/721 [==============================] - 23s 32ms/step - loss: 0.9297 - accuracy: 0.6724 - val_loss: 0.8291 - val_accuracy: 0.6845 - lr: 0.0010
Epoch 3/50
721/721 [==============================] - 23s 32ms/step - loss: 0.8774 - accuracy: 0.6766 - val_loss: 0.8285 - val_accuracy: 0.6983 - lr: 0.0010
Epoch 4/50
721/721 [==============================] - 23s 32ms/step - loss: 0.8639 - accuracy: 0.6896 - val_loss: 0.7800 - val_accuracy: 0.6958 - lr: 0.0010
Epoch 5/50
721/721 [==============================] - 22s 31ms/step - loss: 0.8386 - accuracy: 0.6899 - val_loss: 0.7353 - val_accuracy: 0.7120 - lr: 0.0010
Epoch 6/50
721/721 [==============================] - 23s 31ms/step - loss: 0.8293 - accuracy: 0.7028 - val_loss: 0.7409 - val_accuracy: 0.7344 - lr: 0.0010
Epoch 7/50
721/721 [==============================] - 23s 32ms/step - loss: 0.8026 - accuracy: 0.7049 - val_loss: 0.7285 - val_accuracy: 0.7207 - lr: 0.0010
Epoch 8/50
721/721 [==============================] - 23s 31ms/step - loss: 0.7931 - accuracy: 0.7082 - val_loss: 0.6927 - val_accuracy: 0.7431 - lr: 0.0010
Epoch 9/50
721/721 [==============================] - 22s 31ms/step - loss: 0.7882 - accuracy: 0.7108 - val_loss: 0.6941 - val_accuracy: 0.7394 - lr: 0.0010
Epoch 10/50
721/721 [==============================] - 23s 32ms/step - loss: 0.7673 - accuracy: 0.7172 - val_loss: 0.6987 - val_accuracy: 0.7456 - lr: 0.0010
Epoch 11/50
721/721 [==============================] - 23s 31ms/step - loss: 0.7543 - accuracy: 0.7215 - val_loss: 0.6886 - val_accuracy: 0.7506 - lr: 0.0010
Epoch 12/50
721/721 [==============================] - 23s 32ms/step - loss: 0.7468 - accuracy: 0.7230 - val_loss: 0.6766 - val_accuracy: 0.7332 - lr: 0.0010
Epoch 13/50
721/721 [==============================] - 23s 32ms/step - loss: 0.7354 - accuracy: 0.7330 - val_loss: 0.7312 - val_accuracy: 0.7394 - lr: 0.0010
Epoch 14/50
721/721 [==============================] - ETA: 0s - loss: 0.7310 - accuracy: 0.7311
Epoch 00014: ReduceLROnPlateau reducing learning rate to 0.0005000000237487257.
721/721 [==============================] - 23s 32ms/step - loss: 0.7310 - accuracy: 0.7311 - val_loss: 0.6660 - val_accuracy: 0.7444 - lr: 0.0010
Epoch 15/50
721/721 [==============================] - 23s 32ms/step - loss: 0.6933 - accuracy: 0.7426 - val_loss: 0.6453 - val_accuracy: 0.7544 - lr: 5.0000e-04
```

Here Learning Rate has been reduced 7 times from assigned value i.e. 0.0010 to 0.00050; to 0.00025; to 0.000125; to 6.25000029685907e-05; to 3.125000148429535e-05; to 1.5625000742147677e-03; to 1e-05.

6.1.3 MODEL EVALUATION: -



```
loss, accuracy = model.evaluate(x_test, y_test, verbose=0)
loss_v, accuracy_v = model.evaluate(val_x, val_y, verbose=0)
print("Validation: accuracy = %f  ;  loss_v = %f" % (accuracy_v, loss_v))
print("Test: accuracy = %f  ;  loss = %f" % (accuracy, loss))
```

```
Validation: accuracy = 0.776808  ;  loss_v = 0.610624
Test: accuracy = 0.783824  ;  loss = 0.618542
```

We can also further tune our model to easily achieve the accuracy above 80% and it could be said that this model is efficient in comparison to detection with naked eye having 78.38% accuracy.

6.1.4 GRAPHICAL ANALYSIS : -

(Model Accuracy and Model Loss)

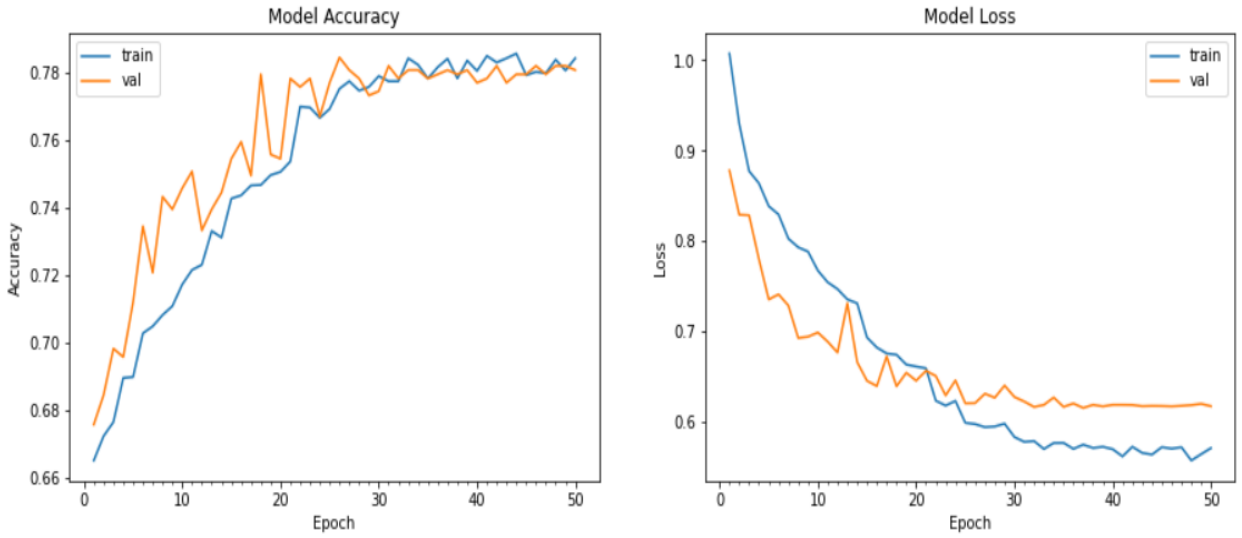


Fig: 6.1.1

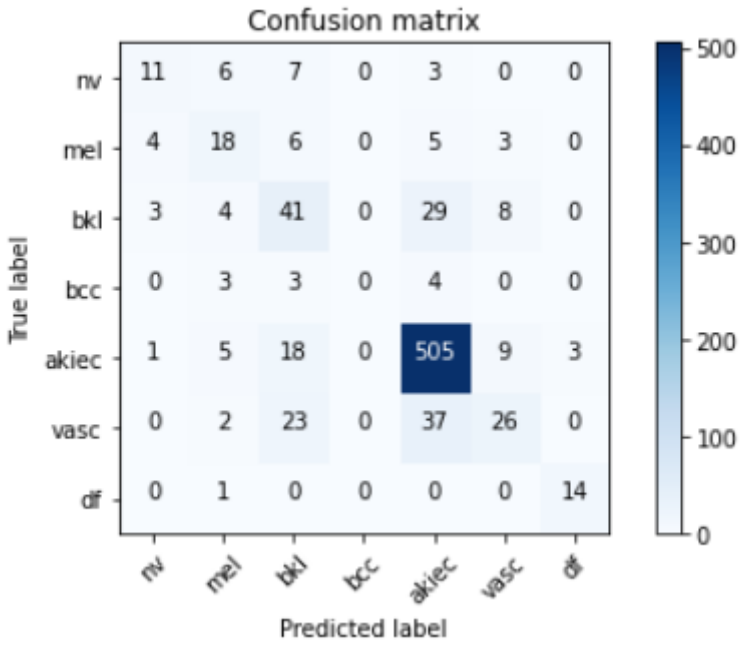


Fig: 6.1.2

The Fig. 6.1.1 shows us the graph of the accuracy rate and the loss rate of the system in classifying the images respectively. From the Confusion Matrix (fig 6.1.2), the main interest was to evaluate how the Actinic Keratosis samples were being classified. An expressive number of samples were mistakenly classified as Actinic Keratosis, almost the same number as the ones that were properly classified. Because of this accuracy in this model is a major issue but still it is approximately equal or better than an average person looking at the disease with naked eye.

6.2 MODEL USAGE

The above model is now loaded onto another code section which is used for accessing the model and using it to detect the diseases in any image given to it. The image uploaded will then be pre-processed and fed into the loaded model then classified into any of the seven types listed in the dictionary below.

```
[ ]  lesion_dict = {
        'nv': 'Melanocytic nevi',
        'mel': 'Melanoma',
        'bkl': 'Benign keratosis-like lesions ',
        'bcc': 'Basal cell carcinoma',
        'akiec': 'Actinic keratoses',
        'vasc': 'Vascular lesions',
        'df': 'Dermatofibroma'
    }
```

This has been further modified to add camera capture integrated into the system which can then allow images to be captured from a camera and be used for the model.

```
from IPython.display import Image
try:
  filename = take_photo()
  print('Saved to {}'.format(filename))

  # Show the image which was just taken.
  display(Image(filename))
except Exception as err:
  # Errors will be thrown if the user does not have a webcam or if they do not
  # grant the page permission to access it.
  print(str(err))
```

```
Saved to photo.jpg
```

This image is then analysed by the model and the following result is obtained.

```
[ ]  print(lesion_dict[result])

     Actinic keratoses
```

6.3 CONCLUSION

This project demonstrates a method that uses techniques related to computer vision to distinguish different kinds of skin lesions for now. Deep learning algorithms have been used for learning algorithms for training and testing purposes. The accuracy attained is 78.382%. The feasibility of building a skin disease classification system has been investigated using CNN model. Better accuracy can be obtained by providing a training set with more variance and also by increasing its size.

This model has then been used to insert more images and obtain results from it which proves its usability and purposes which have been stated above in the report. We can now use this model to obtain images from other people and use those images to diagnose the people and train the model even more. This will allow us to simulate the model and also fulfill the main purpose for which this model was built.

6.4 FUTURE WORK

The present model would be improvised by integrating more data and training it. It can be used further to identify skin problems at an early stage and help the patient seek the right treatment and get cured. The idea can then be commercialized and promoted to a much larger populace for all to use and obtain a free system of diagnosis. This would help people in poorer regions obtain a much faster diagnosis which would be free of cost and available for anyone with internet access. This would all be done after the addition of a proper GUI and by making a proper application for this purpose. Hence our aim to provide a free of cost diagnosis for everyone would be fulfilled.

REFERENCES

1. Nawal Soliman AL Kolifi ALEnezi , "A Method Of Skin Disease Detection Using Image Processing And Machine Learning", presented at the 16th International Learning & Technology Conference 2019, https://www.sciencedirect.com/science/article/pii/S1877050919321295
2. Saja Salim Mohammed and Jamal Mustafa Al-Tuwaijari, "Skin Disease Classification System Based on Machine", 2nd International Scientific Conference of Engineering Sciences (ISCES 2020), https://iopscience.iop.org/article/10.1088/1757-899X/1076/1/012045/pdf
3. Ms Seema Kolkur , Dr D.R. Kalbande , Dr Vidya Kharkar , **"**Machine Learning Approaches to Multi-Class Human Skin Disease Detection"**,** International Journal of Computational Intelligence Research (2018), http://ripublication.com/ijcir18/ijcirv14n1_03.pdf
4. Roderick Hay, Sandra E. Bendeck, Suephy Chen, Roberto Estrada, Anne Haddix, Tonya McLeod, and Antone Mahé, "Skin Diseases", Disease Control Priorities in Developing Countries. 2nd edition, Chapter 37, https://www.ncbi.nlm.nih.gov/books/NBK11733/
5. All About Common Skin Disorders **,**https://www.healthline.com/health/skin-disorders#pictures
6. Sumit Saha, A Comprehensive Guide to Convolutional Neural Networks- the ELI5 way, 2018, https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53
7. Artificial Neural Network Tutorial, https://www.javatpoint.com/artificial-neural-network
8. Tschandl, P., Rosendahl, C. & Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data 5, 180161 (2018), https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T