AN IMPROVED METHOD FOR SKIN CANCER PREDICTION USING MACHINE LEARNING TECHNIQUES

MINOR PROJECT II

Submitted by:

Debshishu Ghosh (19103082)

Rishabh Lal Srivastava (19103088)

Roshni Singh (19103034)

Under the supervision of:

Dr. Bharat Gupta

Department of CSE & IT,

Jaypee Institute of Information Technology, Noida

MAY 2022

TABLE OF CONTENTS

List of Tables

List of Figures

Abbreviations

mel : Melanoma

nv : Melanocytic Nevi

Bcc : Basal cell Carcinoma

Akiec : Actinic Keratosis

Bkl : Bening keratosis like lesions

DF : Dermatofibroma

Vasc : Vascular lesions

EDA : Exploratory Data Analysis

CNN : Convolutional Neural Network

ACKNOWLEDGEMENT

I would like to place on record my deep sense of gratitude to **Dr. Bharat Gupta**, Assistant Professor (Senior Grade), Jaypee Institute of Information Technology, India for his/her generous guidance, help and useful suggestions.

I express my sincere gratitude to **Dr. Chakresh Jain**, Dept. of Bioinformatics, Assistant Professor (Senior Grade), Jaypee Institute of Information Technology, India, for his/her stimulating guidance, continuous encouragement and supervision throughout the course of present work.

Signatures:

Name of students:	Roshni Singh (19103034)	Debshishu Ghosh (19103082)	Rishabh Lal Srivastava (19103088)
Date:	20th May 2022		

DECLARATION

We hereby declare that this submission is our own work and that, to the best of our knowledge and beliefs, it contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma from a university or other institute of higher learning, except where due acknowledgment has been made in the text.

Place: JIIT Noida - 62

Date : 20th MAY 2022

Name: Rishabh Lal Srivastava

Enrolment No.: 19103088

Name: Debshishu Ghosh

Enrolment No.: 19103082

Name: Roshni Singh

Enrolment No.: 19103034

CERTIFICATE

This is to certify that the work titled "**An Improved Method for Skin Cancer Prediction Using Machine Learning Techniques**" submitted by Name of Students of B.Tech of Jaypee Institute of Information Technology, Noida has been carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of any other degree or diploma.

Signature of Supervisor :

Name of Supervisor : Dr. Bharat Gupta

Designation : Assistant Professor(Senior Grade)

Date : 20th MAY 2022

Abstract

Among skin diseases the type that causes cancer are the most fatal ones and pose the biggest issues. These issues arise due to the fact that cancers are just much larger quantities of the same cells that are present around the body, which makes diagnosis very difficult until later stages. For skin diseases this is an even bigger issue as most diseases look the same and are generally harmless. Now the onset of artificial intelligence and machine learning techniques, in the field of images, has allowed computers to identify sequences and patterns in images that can never be observed by the naked eye. Hence in order to battle skin cancer in its early stages a system has been proposed to identify and predict skin cancer in its earlier stages. A skin cancer prediction system has hence been created and implemented to predict three major types of skin cancer that affect humans. A dataset of the said skin cancer types and other types of skin diseases have been taken and analysed for this project. Apart from the model a web application has also been constructed for deployment on the web to enable the access of this model to the general masses. This would involve the usage of a camera or a picture file as input and the output would be the type of cancer or just no cancer. This project's main aim is to provide a fast and accurate diagnosis of skin cancer for anyone using the web application. It is essentially free of cost and can provide a great help to people in remote areas using this, who have limited access to proper healthcare.

Chapter 1

1.1 INTRODUCTION

Skin problems are a typical occurrence in the daily lives of humans, which can appear in any part of the body and come in a variety of shapes and sizes. Most of the time, one dismisses them as a common occurrence, unconcerned about the severity of the problem, and in some cases, one is unable to recognize them as a disease [1]. They have a wide range of appearances that even when identified, even trained professionals (doctors or skin specialists) are unsure of the exact disease which has been acquired by the patient at times and thus how to treat them[1]. Hence treatment of such diseases is very difficult and intensive. Among the myriad of skin diseases, skin cancer is the most fatal of them all. If cancers aren't detected in the early stages, they are effectively incurable and result in deaths [2].

Cancers are a category of disease which are caused due to abnormal growth in cells, wherein they do not stop multiplying [2]. The constant growth and appearance of new cells leads to congestions and complications from within[2]. Hence cancers in general can be categorised as some of the most fatal diseases, and if not detected in earlier stages almost certainly leads to death [3]. Detection of cancers is an arduous task even for trained medical professionals, as they are simply the extra addition of the same cells and generally cannot be distinguished from one another unless some new characteristics arise [4].
Although skin diseases can be difficult to detect in general by the naked eye of trained professionals, machine learning is a powerful tool that has come to existence. It has proven to be a very powerful tool in image processing and segmentations. Image segmentation allows the model to identify new characteristics of the images that were previously unknown and helps identify new traits about an image. Recently it has become a very powerful way of detecting and predicting diseases through images (X Rays, MRI scans etc.)[5]. Even among all ML techniques, CNN has proven to be the one of the best and most popular ways to perform image segmentation and analysis [5][7].

A plan was devised by the group to address the problem of skin cancer detection with the help of a machine learning model based on CNN neural analysis. This system would ensure quick detection of the cancer and give out a report on the type of cancer it detects. Since it is a computer run model and can even be run on a mobile device upon full completion, it would be accessible to anyone having a mobile phone. 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

The biggest motivation for this project is to help the people who are suffering from the most prevalent skin cancers and to provide a simple diagnosis for those that are present in areas where healthcare is inaccessible, or just too expensive. In such places this model would prove to be very helpful and save the lives of the people who might be suffering from such a fatal skin disease, all free of cost. All one would need is a stable internet connection and access to a device which can host the internet. Then all they would need to do is get a picture of the area they think is affected and get their diagnosis.

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.

The model being based on a CNN model would also give a very accurate prediction due to its proficiency in image segmentation and classification. This increases the usefulness of the system and would ensure that it helps many more people to get a quick and accurate diagnosis. This report contains the proceedings of how we approached the problem found in research.

Chapter 2

2.1 LITERATURE STUDY

In order to start the preparation for the creation of the system a research is needed in order to identify the current progress in the said field and what can be improved on them. The strengths and weaknesses of the models proposed in the papers were analysed in order to create a model that is in effect a better model and would fulfil our purpose and motivation. The following are the briefings of the major papers for our inspiration and work done and a short analysis of their weaknesses.

The first paper to be analysed was by Mohammed et.al, who prepared a full case study of various Machine Learning models of different countries that have been compared side by side on their key aspects[7]. As it is a survey, it gives a comprehensive analysis of which algorithms to use to get better results and what has been worked on. This gives a full comprehension of what kind of models have been made and their details. It helps in setting a benchmark for future work to be done and gives insight of the most successful models currently in research.

Vidya et.al. who in their research have prepared a skin disease detection model with the help of KNN, SVM and Naive Bayes Classifiers [4]. For their work, skin lesion images were taken from International Skin Imaging Collaboration (ISIC) in which 328 images of benign and 672 images of melanoma. Their classification result obtained is 97.8 % of Accuracy and 0.94 Area under Curve using SVM classifiers [4]. This sets a good benchmark for the levels of accuracy that need to be achieved in order to compete with the ongoing levels of research.

Dai et.al. have a research that is in line with ours and they have proposed a CNN model for tackling the issue while creating a mobile application for its use[5]. However their accuracy stands at 75.2% which is substandard and cannot be used for medical or widespread use[5]. This research focuses solely on the creation of a mobile application for mass usage, which produces a substandard model not capable for real time usage.

ALEnezi et.al in this paper, analysed 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[6]. This is another model which is a very rudimentary model for diagnosis. It achieves an extremely high accuracy which is however unreliable as the dataset used for the purpose of this research is a very small dataset of only 80 images [6].

The research of Kolkur et.al is about a classification system that has been made for the K.E.M. hospital in Parel, Mumbai [8]. 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%[8]. This gives an insight on a model which is currently in use by professionals.

Harper et.al in their research have given us a collection of research on the vulnerabilities of CNN models in skin cancer diagnosis. This is an analysis of the vulnerabilities of CNN networks in the detection of skin cancer. It highlights the issue of skin cancers being very similar in nature and hence causing misdiagnosis in the system [9]. It then explains how the models can be improved for the future [9]. This has a briefing of the issues that might arise in the future due to unforced errors by the model and hence they need to be curbed to provide a much more reliable model. Below is a table of comparison for all the major papers used for inspiration (Table 2.1)..

Title	Year	Author	Main Topic of Concern	Algorithms Used	Accuracy if Prediction Algorithm	Remarks
Skin Disease Classification Based on ML: A Survey	2020	Mohammed et.al	All Skin Diseases	ANN, KNN, Kmeans, SVM, CNN, Naïve Bayes	Ranging from 62% to 100%	This is a survey which helps in highlighting the best model for the skin cancer classification system
Skin Cancer Detection using Machine Learning Techniques	2020	Vidya et.al	Skin Cancer	KNN	97.80%	Uses an older method for detection not optimal for modern skin disease segmentation
Machine Learning on Mobile: An On-device Inference App for Skin Cancer Detection	2019	Dai et.al	Skin Cancer	CNN	75.80%	This is a very low accuracy model not relevant for general public use. The idea of the usage of a mobile device is one that is pivotal for the project idea however.
A Method Of Skin Disease Detection Using Image Processing And Machine Learning	2019	ALEnezi et.al	Skin Lesions	CNN with SVM	100%	The model uses an extremely small dataset of 80 images which is too small for a proper accurate training result.
Machine Learning Approaches to Multi-Class Human Skin Disease Detection	2018	Kolkur et.al	All Skin Diseases	ANN, KNN, SVM, Decision Trees, Random Forest.	98.94%	A non-autonomous model which is solely responsible for assisting a doctor with the results a doctor puts in. It is not an image based analysis.
Clinically Relevant Vulnerabilities of Deep Machine Learning Systems for Skin Cancer Diagnosis	2021	Harpur et.al	Skin Cancer	CNN	NA	Talks about the vulnerabilities and issues of CNN skin prediction system due to the nature of skin diseases being very similar to one another

Table 2.1: Research Comparison

2.2 GAPS IN RESEARCH

1. A lot of the papers have used ANN and KNN models which are not the best algorithms for the analysis of images and consume much more time and computation power to achieve the same results as that of a CNN model [4][8][7].
2. Low accuracy of the model can result in incorrect outcomes and predictions. A wrong diagnosis of the system can be very harmful for the user [5][7]. The same is applicable for models that have not been trained properly with the right amount of data [6].
3. The varied datasets used also create more variables that cause changes in the accuracy due to the inherent faults of some datasets [7]. Hence the dataset needs to be a balanced dataset. The imaging techniques should also be specified to obtain similar results upon a set way of imaging.
4. Not all skin diseases can be classified by the use of one system.[7]
5. In a non image-based classification for the case of Kolkur et.al, it becomes a system which is reliant on dermatologists[8]. Hence a doctor's inference is needed first before the system can analyse which prevents the system from being a widespread and readily available system.
6. In a manual input system, the input method is slow and needs a lot of parameters making it an inefficient system[8].
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[5][6][7][8][9].
8. Only one system has a proper frontend mobile application for the model's use ([5]). The rest of the models cannot be implemented for a free large scale purpose.

The gaps and issues of the research papers have been identified. The major issue for these researches is that they aren't made for the general masses for diagnosis which is hence our primary goal. Commercialisation of this system is very beneficial for our target audience who are present in underdeveloped localities.

Chapter 3

3. REQUIREMENT ANALYSIS AND DETAILED DESIGN

In this chapter we will be detailing our objectives and provide a structured process for the workflow of the system to be made using the inferences from the research above.

3.1 RESEARCH OBJECTIVES

The main objectives for the project would be to solve the gaps identified in the research. The major objectives that have been planned to solve for this project are to implement a high accuracy CNN model for faster output times and better results. This is to solve the first and the second gaps which had been found. The other objectives planned to fulfil are for the model to be fully autonomous and have a web application for the system. This application would allow people in every area to have access to the system.

3.1 PROBLEM STATEMENT

With the shortcomings of the above research in mind and having clear research objectives, the proposal is to make an accurate CNN model for skin cancer detection which would then also be hosted on the internet so that everyone can access it and gain a fast and accurate diagnosis.

3.2 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. A CNN model will be used to analyse the images for the segments and to classify the images into their said categories.

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** 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 [10].

The above algorithms would be used for analysing 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).

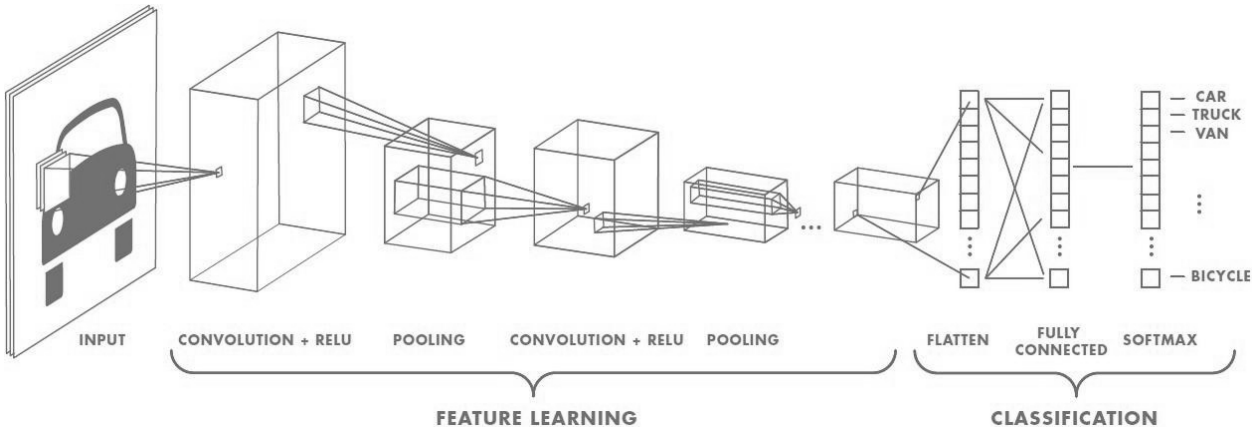

Fig 3.1: CNN algorithm[10]

3.3 PROPOSED ARCHITECTURE

An architecture has been proposed which addresses the gaps identified in the background 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 and the analysis has been done after.

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. We have also identified the duplicates in the images from this dataset and actively worked to reduce the effects of this duplication through oversampling of images.

The segmented images and their pixel RGB values are then used for the training of the model. To obtain the segmented RGB values of the images we have sorted the pixel RGB values into an array and then appended them in the dataset for evaluation. Oversampling was then done to overcome the class

imbalance of the data and the data was reset accordingly to obtain a fully balanced dataset. This new dataset had a total of 46935 images to train with. The new balanced dataset would hence give us better prediction results when compared to the original dataset.

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 analyse 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:

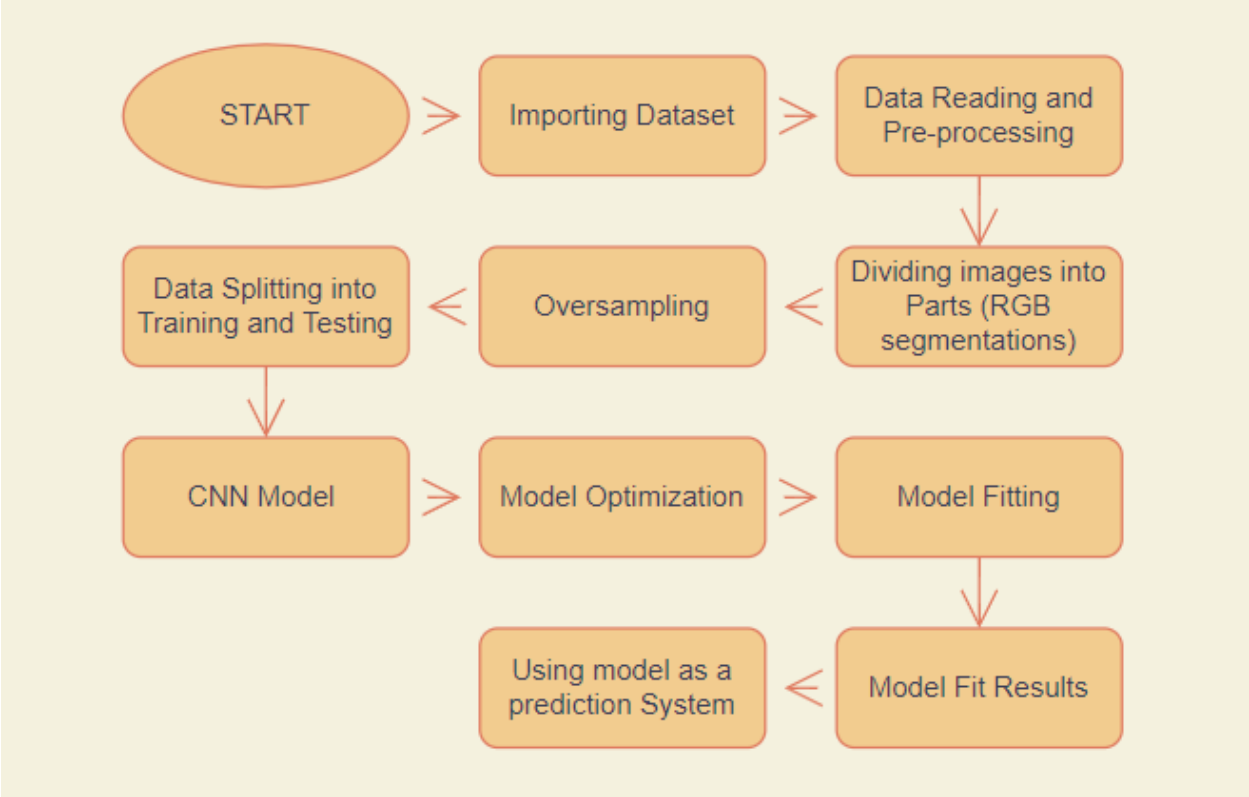

Fig 3.3.1: Proposed Architecture

Chapter 4

4. DATA PROCESSING AND MODEL CREATION

After getting clear criterias and a set architecture for the model to be made, one can now devise a plan for the implementation of the model and start developing it.

4.1 THE DATASET

The HAM10000 dataset [13] 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 like Lesions (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.

Other than the type of cancers present in the dataset, the information on the age of the people, their gender, the location of the affected disease, the method of diagnosis of the disease and the images of the disease are given to us. This forms the details of the entire dataset.

4.1.1 DATASET OPERATIONS

A set of operations were performed for better visibility of the dataset which are fully labelling all the diseases as numbers according to the type of disease, from numbers 0 to 6. The NULL values were found in the dataset and replaced with the mean of the data, hence a total of 57, NULL values were normalised. The rest of the information of the dataset was also obtained and analysed accordingly for any useful information. Another process was done in order to obtain the information on duplicate data. For this comparison between image IDs was done and similar IDs were found within the dataset. This gave us the information that around 4501 images were duplicates, these were then normalised during the process of oversampling. The data was then grouped according to the disease type for further use.

The original images were converted to a Red-Blue-Green (RGB) format after segmentation and this flattened dataset of segmented 28 x 28 images was then used to train the model. The data was readily available for our system however for any further predictions one would have to segment the image and then use them for classification. Hence this new dataset was appended to our original dataset for association and model training along with properly labelled data.

The main interests for our project are the three skin cancer variants from these lesions which are Basal Cell Carcinoma, Melanoma and Actinic Keratosis. These three are the more common types of skin cancer that are known with melanoma being the most fatal type among them [12]. A small data analysis of the dataset was conducted and it was observed that the data was quite unbalanced with some lesion types having vastly more data compared to the other ones.

4.1.2 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.

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.

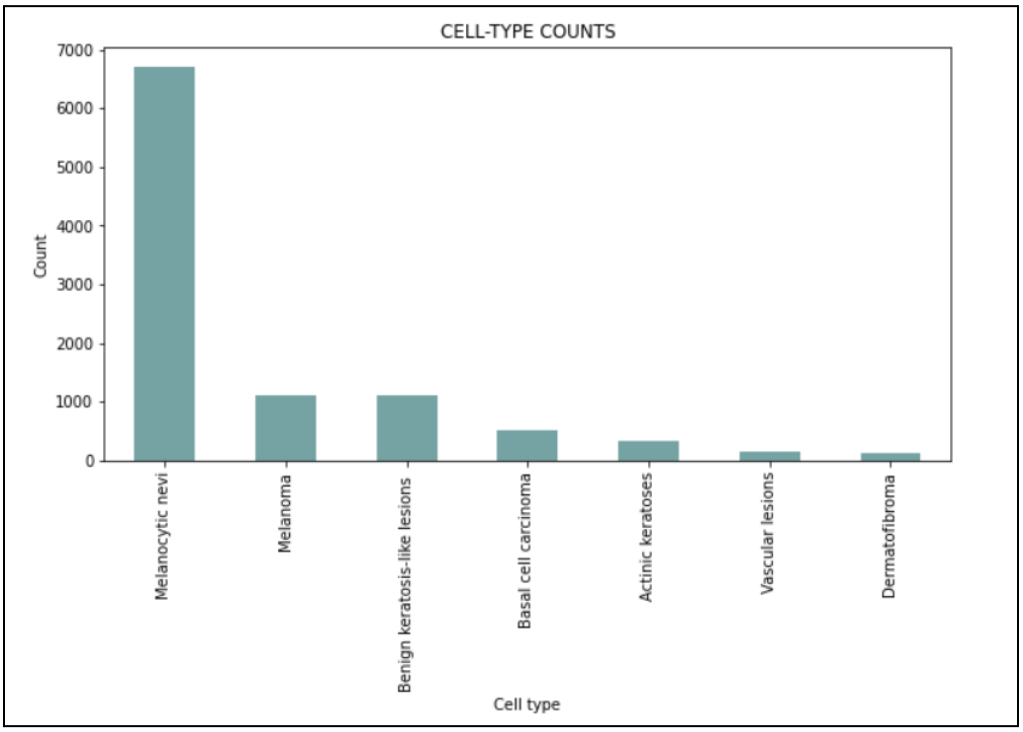

Fig 4.1.1: Cell/Image count of the different cell types

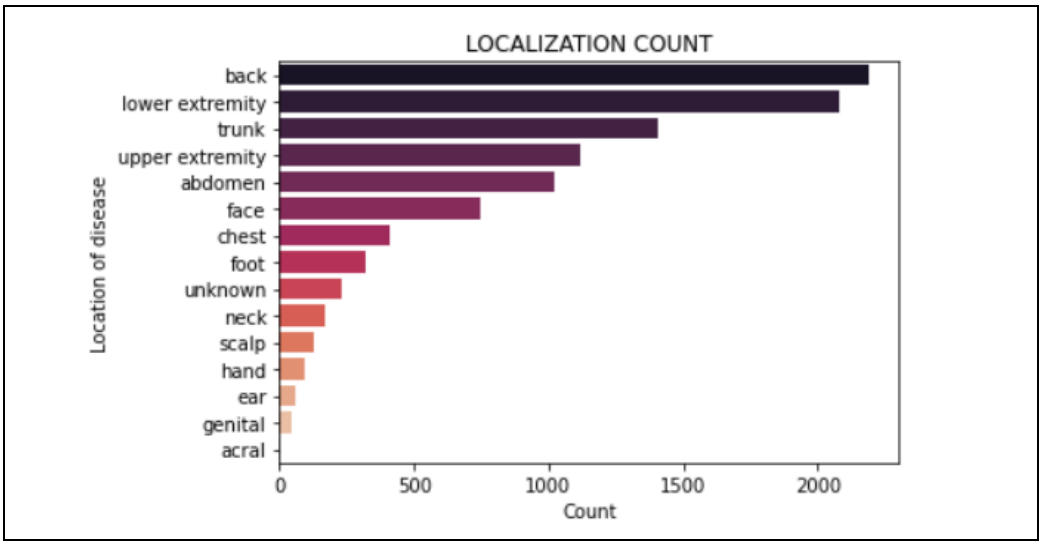

Fig 4.1.2: Localization of the cells in different parts of the body

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. Now we come to the distribution of the images of disease among the various age groups. The following two graphs (fig 4.1.3, 4.1.4) show us the age of the people who have acquired the diseases and their distribution.

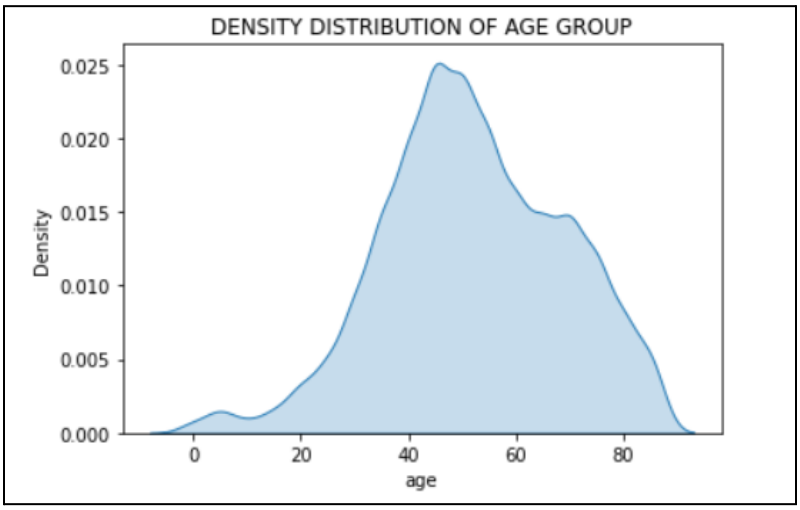

Fig 4.1.3: Density Distribution in Age groups

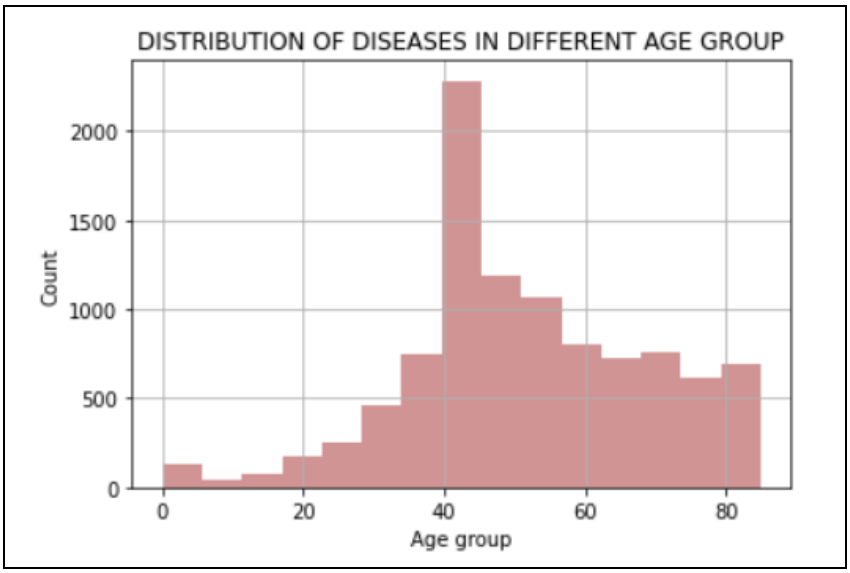

Fig 4.1.4: Count Distribution among Age groups

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. 4.1.5].

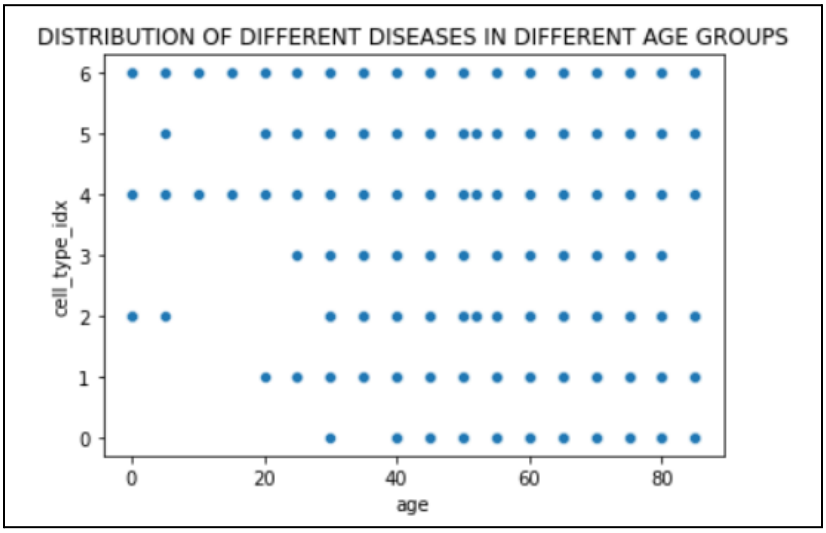

Fig 4.1.5: Overall Disease Distribution

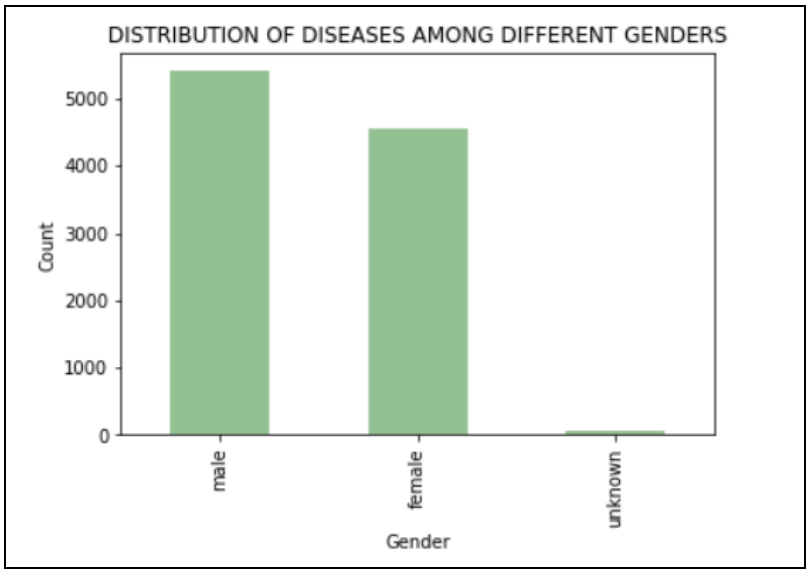

Fig 4.1.6: Gender Distribution

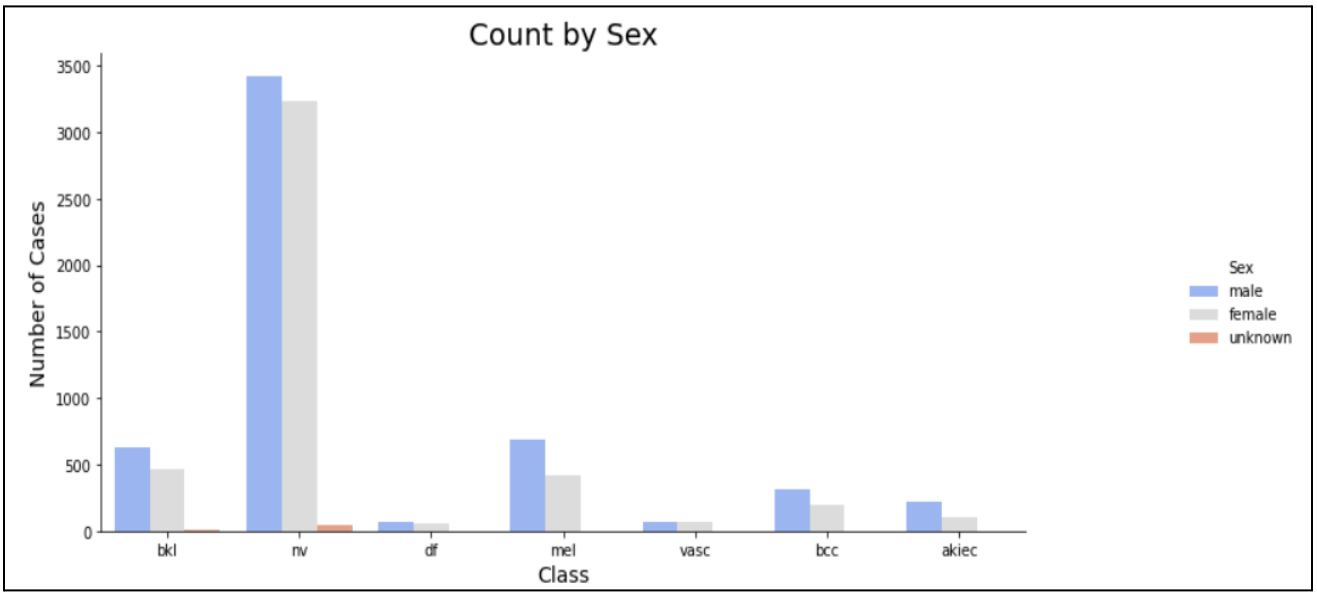

Fig 4.1.7: Individual Diseases among Gender

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 4.1.6, 4.1.7). Here we see no major difference in the graphs and both genders are equally susceptible to all the diseases.

This analysis has helped identify the imbalance in the data and hence balance the data using oversampling with the help of the analysis done. After this an oversampling of the data was done which would help in obtaining a balanced dataset for better classification. The random oversampling produced the final dataset of 46934 images from the original dataset of 10015 images.

4.2 DATA 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.

The first step in the process of pre-processing is to male the data a balanced form of data. This is necessary so that the model can train equally for all types of classes. Hence to ensure that, oversampling is done which randomly selects data and duplicates them and adds them to the dataset. This new training set has a much larger number of images and becomes a dataset of 46937 images.

	pixel0000	pixel0001	pixel0002	pixel0003	pixel0004	pixel0005	\
0	192	153	193	195	155	192	
1	25	14	30	68	48	75	
2	192	138	153	200	145	163	
3	38	19	30	95	59	72	
4	158	113	139	194	144	174	
...	...	...	...	...	...	...	
46930	164	110	113	173	124	125	
46931	179	141	158	183	151	168	
46932	217	179	197	216	175	194	
46933	250	242	252	251	242	252	
46934	165	141	160	168	135	155	

Fig: 4.2.1 Oversampled Images

4.2.1 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.

4.3 THE TRAINING MODEL

The model has a total of 9 layers which are a combination of 2D convolutional layers, MaxPooling and Dense. The CNN architecture used for the model is:

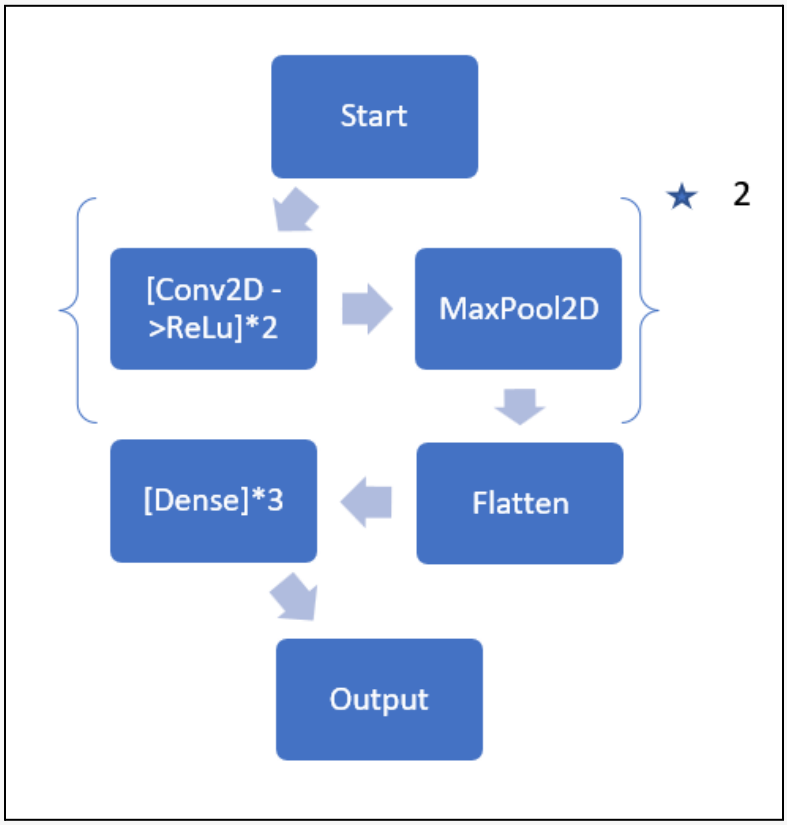

Fig: 4.3.1: CNN Model Architecture

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. This also helps in reducing the computation complexity of the system by reducing the number of variables.

'Relu' stands for rectifier (maximum activation function) (0,x). The rectifier activation function is utilised to give the network non-linearity. It removes all the negative values from the pattern identification and segmentation phase to reduce unwanted variables and get a clear crisp output. To transform the final feature maps into a single 1D vector, utilise 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 three 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")). This uses the softmax activation function which is the best activation function for a CNN model, as it predicts a multinomial probability distribution. It normalises the data the best for further computation for the ANN classification done by the Dense layer.

4.3.1 OPTIMISER

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).

4.3.2 MODEL TESTING AND TRAINING

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 20 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 5

5. EXPERIMENTS AND RESULTS

Using HAM10000 dataset [13], 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 normalisation 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. However, due to further classification of the model into cancerous and non-cancerous systems it will give the output among the three cancerous types in the end. The predictions sometimes vary as the initial stages of a few diseases look like others.

5.1.1 TRAINING OF MODEL

```
Epoch 1/20
232/235 [============================>.] - ETA: 0s - loss: 1.2389 - accuracy: 0.5149
Epoch 1: saving model to /content/drive/MyDrive/Model_Minor/model.h5
235/235 [==============================] - 3s 12ms/step - loss: 1.2352 - accuracy: 0.5164 - val_loss: 0.9132 - val_accuracy: 0.6403
Epoch 2/20
231/235 [============================>.] - ETA: 0s - loss: 0.6949 - accuracy: 0.7467
Epoch 2: saving model to /content/drive/MyDrive/Model_Minor/model.h5
235/235 [==============================] - 2s 8ms/step - loss: 0.6928 - accuracy: 0.7474 - val_loss: 0.5886 - val_accuracy: 0.7714
Epoch 3/20
231/235 [============================>.] - ETA: 0s - loss: 0.4381 - accuracy: 0.8436
Epoch 3: saving model to /content/drive/MyDrive/Model_Minor/model.h5
235/235 [==============================] - 2s 8ms/step - loss: 0.4367 - accuracy: 0.8441 - val_loss: 0.3684 - val_accuracy: 0.8714
Epoch 4/20
232/235 [============================>.] - ETA: 0s - loss: 0.2933 - accuracy: 0.8954
Epoch 4: saving model to /content/drive/MyDrive/Model_Minor/model.h5
235/235 [==============================] - 2s 8ms/step - loss: 0.2925 - accuracy: 0.8954 - val_loss: 0.2671 - val_accuracy: 0.8999
Epoch 5/20
231/235 [============================>.] - ETA: 0s - loss: 0.2154 - accuracy: 0.9232
Epoch 5: saving model to /content/drive/MyDrive/Model_Minor/model.h5
235/235 [==============================] - 2s 8ms/step - loss: 0.2157 - accuracy: 0.9230 - val_loss: 0.2378 - val_accuracy: 0.9193
Epoch 6/20
230/235 [============================>.] - ETA: 0s - loss: 0.1681 - accuracy: 0.9418
Epoch 6: saving model to /content/drive/MyDrive/Model_Minor/model.h5
235/235 [==============================] - 2s 8ms/step - loss: 0.1683 - accuracy: 0.9418 - val_loss: 0.2203 - val_accuracy: 0.9224
Epoch 7/20
234/235 [============================>.] - ETA: 0s - loss: 0.1288 - accuracy: 0.9532
Epoch 7: saving model to /content/drive/MyDrive/Model_Minor/model.h5
235/235 [==============================] - 2s 8ms/step - loss: 0.1289 - accuracy: 0.9532 - val_loss: 0.1877 - val_accuracy: 0.9407
Epoch 8/20
232/235 [============================>.] - ETA: 0s - loss: 0.0983 - accuracy: 0.9654
Epoch 8: saving model to /content/drive/MyDrive/Model_Minor/model.h5
```

Fig 5.1.1: Epoch Training

Upon training the model we find that the accuracy of the model starts from 64% and rises up all the way to 91.9% in just 5 epochs. The learning rate for our model starts at 0.001 and is set to fall to 0.00001 at minimum. This ensures that our model gets the best possible results in the least amount of time.

5.1.2 MODEL EVALUATION:

The model achieved an accuracy score of 97.89% and suffered a net loss of 15.5% during training. The classification report for all the data classes in the system is given in Fig. 5.2 below.

	precision	recall	f1-score	support
akiec	1.00	1.00	1.00	1359
bcc	0.99	1.00	0.99	1318
bkl	0.95	0.99	0.97	1262
df	1.00	1.00	1.00	1351
mel	0.99	0.87	0.92	1374
nv	1.00	1.00	1.00	1358
vasc	0.93	0.99	0.96	1365
accuracy			0.98	9387
macro avg	0.98	0.98	0.98	9387
weighted avg	0.98	0.98	0.98	9387

Fig: 5.1.2: Classification Report

5.1.3 GRAPHICAL ANALYSIS AND COMPARISON

(Model Accuracy and Model Loss**)**

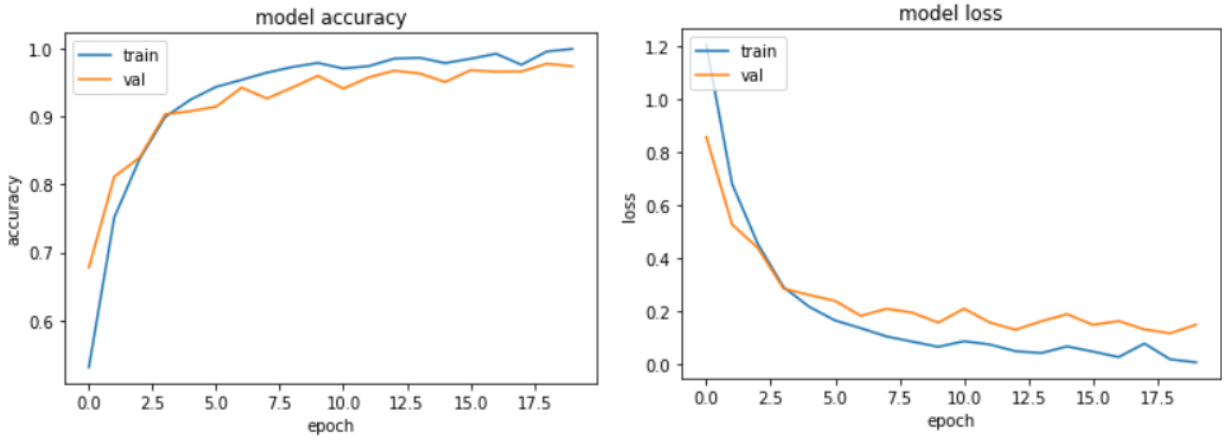

Fig: 5.2 Model Accuracy and Loss Graph

The Fig. 5.2 shows us the graph of the accuracy rate and the loss rate of the system over time during training, in classifying the images respectively. Our model has hence achieved an accuracy of 97.89%. In the paper by Mohammed et.al, the average accuracy rates of all the given models for the survey amount to 89.17% [7]. Hence the model prepared could be said to be better than the average model for skin disease detection worldwide. For skin cancer related CNN models only is around 95.4% which could be said as the real benchmark to be obtained [7]. This model having 97.89% accuracy could be said as an above

average model for skin cancer detection. The best models for skin cancer detection can upto 99% with higher computation models and better algorithms.

5.2 MODEL USAGE

The above model is now loaded onto the streamlit library which is an open source application framework specialised for machine learning setups and analysis. This forms the frontend of the system which will be deployed and can be accessed on multiple devices.

The interface is simple for the system, it has a total of 5 sections to browse into. The first section is a simple directions section where it is shown how to use the webpage. The second section is a live image capturing section where we can capture the image of the affected body part through an integrated camera to the system. The space bar would click the image for it, and then the diagnosis is listed below the affected image.

The third section would be that of an image uploading system wherein an already present image can be uploaded into it and it would give the result of that image. A briefing of how the model has been made and an EDA for it is provided in the fourth section and the final section is just a credits section. The application is very simple to use and with the instructions on how to use it, basically anyone can operate it as long as they have an internet connection.

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

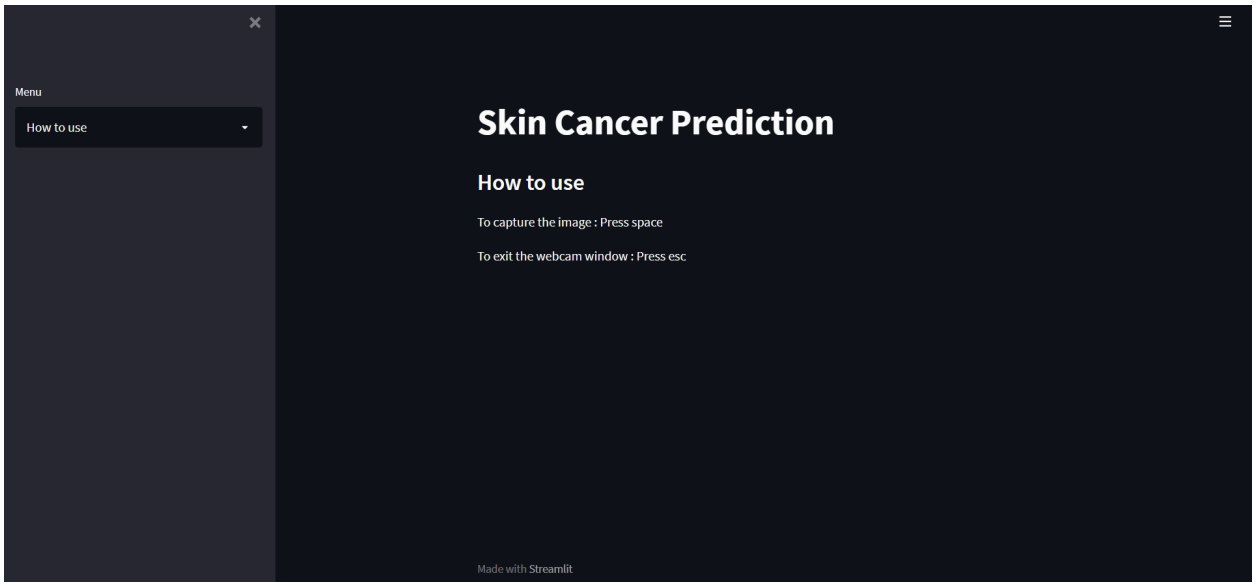

Fig: 5.2.1 How to use

Fig: 5.2.2 Webcam Live Feed

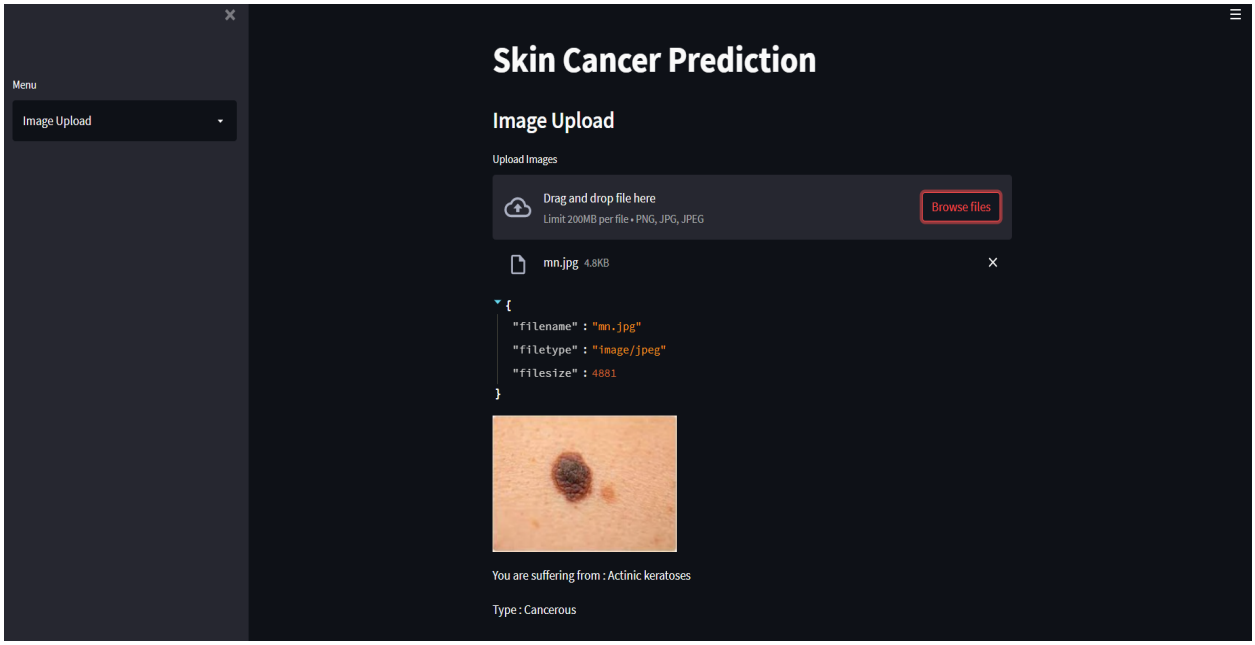

Fig: 5.2.3 Image Upload

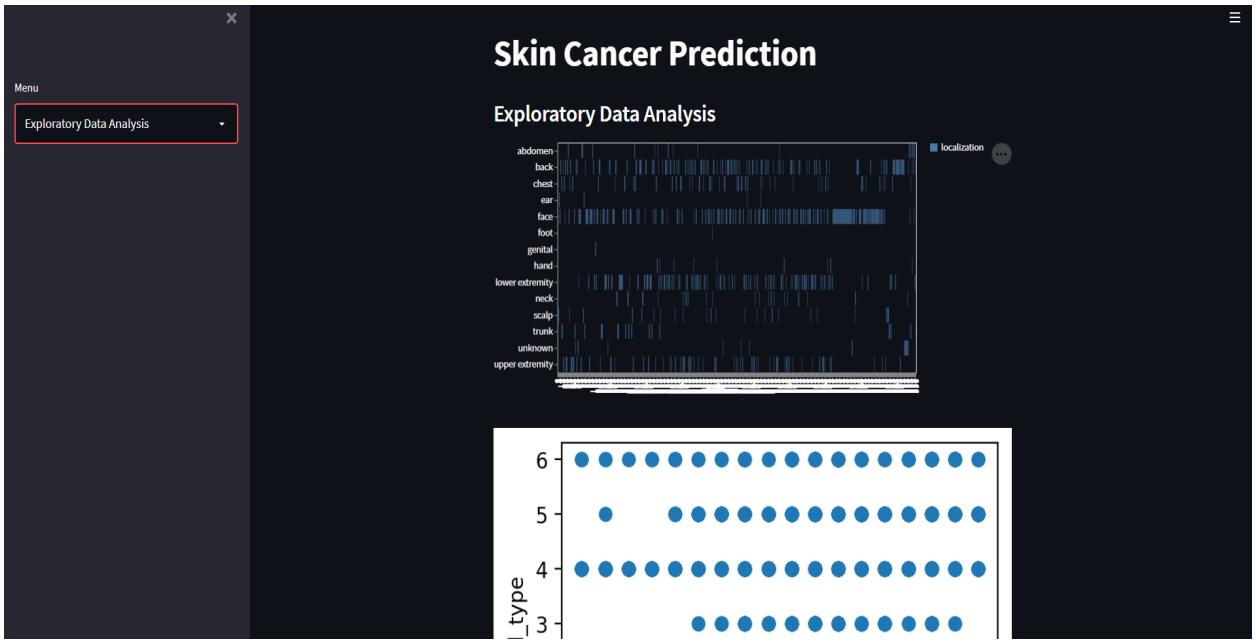

Fig: 5.2.4 Exploratory Data Analysis

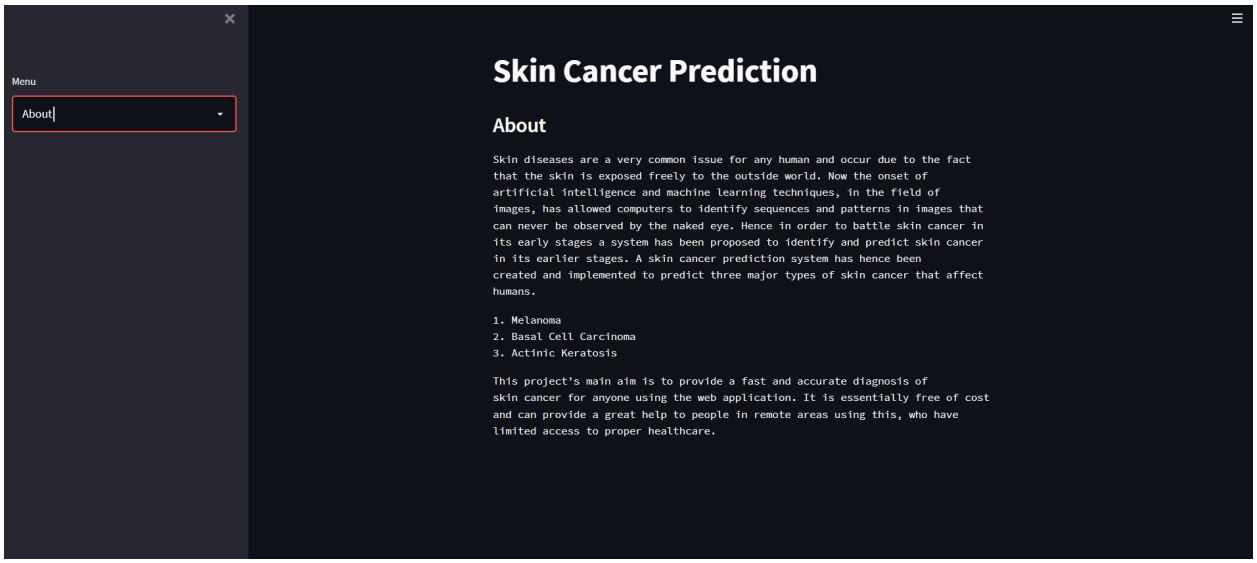

Fig: 5.2.5 About

CHAPTER 6: 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 97.89%. 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 fulfil the main purpose for which this model was built.

6.1 FUTURE WORK

This final model can now be deployed on Heroku or Firebase etc. to gain access to the facilities provided by the model to get a proper diagnosis for themselves. The creation of an android application would also greatly benefit the cause and help in spreading awareness and better healthcare of this serious disease.

REFERENCES:

1. Katherine Brind'Amour, All About Common Skin Disorders , https://www.healthline.com/health/skin-disorders#prevention (accessed on: 09/05/2022)
2. Gavin P. Dunn, Lloyd J. Old, Robert D. Schreiber, The Immunobiology of Cancer Immunosurveillance and Immunoediting, Immunity, Volume 21, Issue 2, 2004, Pages 137-148, ISSN 1074-7613, https://doi.org/10.1016/j.immuni.2004.07.017.
3. Ames, B N et al. "The causes and prevention of cancer." *Proceedings of the National Academy of Sciences of the United States of America* vol. 92,12 (1995): 5258-65. doi:10.1073/pnas.92.12.5258
4. M. Vidya and M. V. Karki, "Skin Cancer Detection using Machine Learning Techniques," *2020 IEEE International Conference on Electronics, Computing and Communication Technologies (CONECCT)*, 2020, pp. 1-5, doi: 10.1109/CONECCT50063.2020.9198489.
5. X. Dai, I. Spasić, B. Meyer, S. Chapman and F. Andres, "Machine Learning on Mobile: An On-device Inference App for Skin Cancer Detection," 2019 Fourth International Conference on Fog and Mobile Edge Computing (FMEC), 2019, pp. 301-305, doi: 10.1109/FMEC.2019.8795362.
6. Nawal Soliman ALKolifi ALEnezi, A Method Of Skin Disease Detection Using Image Processing And Machine Learning, Procedia Computer Science, Volume 163, 2019, Pages 85-92, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2019.12.090.
7. 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
8. 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
9. Du-Harpur, X., Arthurs, C., Ganier, C., Woolf, R., Laftah, Z., Lakhan, M., Salam, A., Wan, B., Watt, F. M., Luscombe, N. M., & Lynch, M. D. (2021). Clinically Relevant Vulnerabilities of Deep Machine Learning Systems for Skin Cancer Diagnosis. *The Journal of investigative dermatology*, *141*(4), 916–920. https://doi.org/10.1016/j.jid.2020.07.034
10. 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**.(**accessed on 09/05/2022)
11. Artificial Neural Network Tutorial, https://www.javatpoint.com/artificial-neural-network**,** (accessed on 09/05/2022)

12. Mayo Clinic, “Skin Cancer”, https://www.mayoclinic.org/diseases-conditions/skin-cancer/symptoms-causes/syc-20377605#:~:text=Skin%20cancer%20%E2%80%94%20the%20abnormal%20growth,squamous%20cell%20carcinoma%20and%20melanoma.,(accessed on 10/05/2022)
13. Tschandl, Philipp, 2018, "The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions", https://doi.org/10.7910/DVN/DBW86T, Harvard Dataverse, V3