SKIN CANCER CLASSIFICATION

Project Guide:

Prof. Balaji Padmanabhan

Submission Date:

Nov 15, 2021

Submitted by:

Amulya Geereddy

Vivek Kotha

Jeremy Maury

Olanrewaju Olasupo

Table of Contents

ACKNOWLEDGMENT………………………………………………………….1

INTRODUCTION………………………………………………………………...3

GOALS AND MOTIVATION…………………………………………………....4

## DATASET OVERVIEW………………………………………………………….5

## DATA ANALYSIS AND VISUALISATIONS…………………………………..6

## DATA AUGMENTATION……………………………………………………….8

## DATA PRE-PROCESSING………………………………………………………8

## MODEL BUILDING……………………………………………………………...9

# FUTURE ENHANCEMENTS…………………………………………………...12

# REFERENCES…………………………………………………………………..12

# 

# ACKNOWLEDGEMENT

In this competitive world there is a race of existence where people are having a strong will to come and succeed. Project is a proof of our theoretical learning and practical working. We would like to express our sincere gratitude to **Prof. Balaji Padmanabhan** for his invaluable guidance and support throughout the semester. We would also thank our seniors for their suggestions and references.

Sincerely Team

Targaryen’s

# 

# INTRODUCTION

As a part of the coursework, we wanted to represent a project which replicates the most excruciating concepts that we have learned during this course. We wanted to focus on image recognition as it is an important problem many companies are trying to solve and get command on. We then narrowed our focus to the healthcare field from there based off of the group member’s interests. As upcoming data scientists, we know that the most difficult problem in a data science project is collecting a complete and informative dataset. We then found a dataset on Harvard Data verse containing a large collection of multi-source dermatoscopic images of common pigmented skin lesions.

According to the Mayo Clinic, skin cancer — the abnormal growth of skin cells — most often develops on skin exposed to the sun. But this common form of cancer can also occur on areas of your skin not ordinarily exposed to sunlight.There are three major types of skin cancer — basal cell carcinoma, squamous cell carcinoma and melanoma.

# GOALS AND MOTIVATION

The main goal of our project is to find the type of skin cancer based on the lesion formed on the skin. As we wanted to predict images on an ongoing Skin-cancer diseases with data captured in the format of images. Artificial intelligence that has been trained to classify photos of skin lesions as Melanocytic nevi, Melanoma, Benign keratosis-like lesions, Basal cell carcinoma, Actinic keratoses, Vascular lesions, Dermatofibroma skin malignancies these images are classified by dermatologist Which are ground truths for our scenario. We show how to classify skin lesions with a 3 layered CNN that was trained end-to -end from photos utilizing only pixels and disease labels.

## DATASET OVERVIEW

We sourced data from Harvard Data verse which consists of 10015 clinical pictures to train a CNN. Due to upload size limitations, images are

stored in two files i.e HAM10000\_images\_part1.zip and HAM10000\_images\_part2.zip. And the data is known as HAM10000(“Human Against Machine with 10000 training images”) dataset.

The data mainly consists of images which are collective representation of all the important diagnostic categories in the pigmented lesions. The dataset includes lesions with multiple images, which can be tracked by the lesion-id column. There is total 7 categories which represents all the types of skin cancer lesions.

Below is a table containing the variables and several entries of data.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **lesion\_id** | **image\_id** | **dx** | **dx\_type** | **age** | **sex** | **localization** |
| HAM\_0000118 | ISIC\_0027419 | bkl | histo | 80 | male | scalp |
| HAM\_0000118 | ISIC\_0025030 | bkl | histo | 80 | male | scalp |
| HAM\_0002730 | ISIC\_0026769 | bkl | histo | 80 | male | scalp |
| HAM\_0002730 | ISIC\_0025661 | bkl | histo | 80 | male | scalp |
| HAM\_0001466 | ISIC\_0031633 | bkl | histo | 75 | male | ear |

Data description of the columns:

1. Lesion\_id : ID of the lesion
2. Image\_id : ID of the image
3. dx : Type of the lession
4. dx\_type : dx\_type is the type of method followed by dermatologist to lable the lesion. More than 50% of lesions are confirmed through histopathology (histo), the ground truth for the rest of the cases is either follow-up examination (follow\_up), expert consensus (consensus), or confirmation by in-vivo confocal microscopy (confocal).
5. age : Age of the patient
6. sex. : Gender of the patient
7. localization: Area where the lesion is spotted.

Sample images for the seven types of cancer:

A picture containing background pattern

Description automatically generatedA picture containing text

Description automatically generatedA picture containing text

Description automatically generatedA picture containing background pattern

Description automatically generatedA close up of a person's face

Description automatically generated with medium confidenceBackground pattern

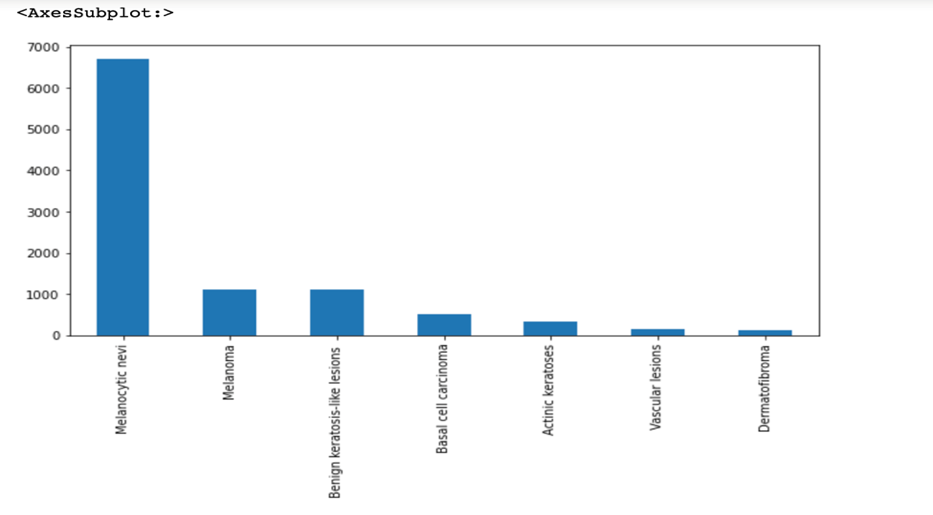
Description automatically generatedA close-up of a person's foot

Description automatically generated with low confidence

## DATA ANALYSIS AND VISUALISATIONS

We will now look at the distributions of our final variables.

To start with, below is the distribution of the 7 classes of the skin cancer in our data set. From the visualization we can see that the dataset is imbalanced dataset and is skewed towards melanocytic nevi. We will address this issue further down in the report with sampling techniques.



A picture containing icon

Description automatically generated

Looking at the distribution for male to female we can see that it skews male slightly but for the most part this variable is balanced.

Chart, histogram

Description automatically generated

We can see from the age distribution that ages 40 to 60 are at the highest risk of skin cancer. This distribution for this variable is fairly normal but skews older.

Chart, histogram

Description automatically generated

From the bar chart above we can see that the most common areas are of the upper body but below the head. These areas include the back, lower & upper extremities, abdomen, and trunk.

Chart

Description automatically generated

Finally, looking at the distribution of where these skin lesions were found and diagnosed, we see that almost all of the cases come from histopathology and follow-up examination visits.

## DATA AUGMENTATION

Since we were able to collect limited number of images, we wanted to used augmentation technique to multiply the data we had. We have used Keras Image Data Generator to augment the data and multiply the size of the dataset using the transformations like Flipping, zoom in, zoom out, Cropping, Rotating. Since we have imbalanced dataset

## DATA PRE-PROCESSING

* We converted the image size of all the images 450x600 x3 into 100x75 for uniformity using PIL Library.
* We labelled the images with the folder names by splitting the name for each level of classification. Labels are 7 different classes of skin cancer types from 0 to 6. We need to encode these labels to one hot vectors.
* We have used Stratified split technique (80:20) to split the data for training and testing in a uniform distributed manner. This will avoid biasing of the model.
* We’ve transformed our test set into 4D Tensor for modelling. Using Label encoding and Keras categorical we transformed the Y-label into multiclass classification data.

## MODEL BUILDING

We used CNN model for building our model as CNN helps to learn about the position and the scale. CNN casts multiple layers (on image and uses filtration to analyze images and understand patterns using filters and NAN images for feature maps. CNN takes tensor as input so CNN can understand spatial relation between pixels of images better thus for complicated images CNN will perform best as CNN is designed to work for images or video classification. We have built a Convolutional neural network to classify images into 7 different classes of skin cancer.

**Architecture**:

We built a CNN model with series of convolution, max pooling layers, fully connected layers followed by a SoftMax layer to classify the labels. We used Adam optimizer and tuned hyperparameters learning rate, batch size, epoch etc. to improve the model accuracy and reduce the model loss during training. We added few dropout layers to our architecture when we realized there’s model overfitting.

Diagram

Description automatically generated

**Steps Followed:**

1. Labelled the dependent variable bases on the class of the image.
2. Split the dataset into 80:20 where 80% is for training and 20% is for testing using stratified split technique.
3. Used Convolutional Neural Networks of three layers using Keras libraries to predict the outcome.

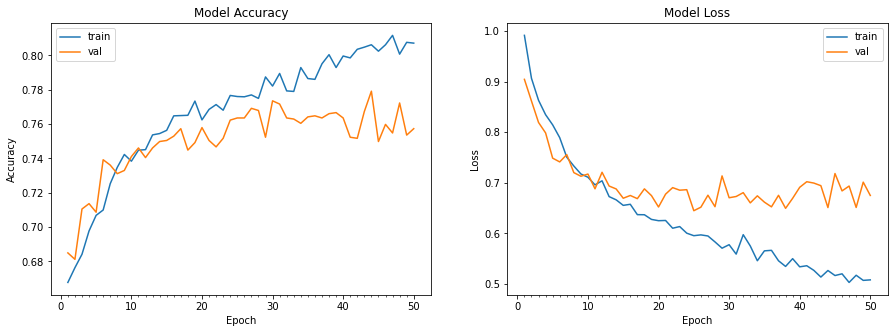
As the dataset is an imbalanced dataset the validation accuracy of the model was 75 % and to make the data balanced, we tried up sampling and down sampling techniques on the data and ran the model.

When we used down sampling of the data i.e., by setting the sample size of lowest class and taking same number of samples from all the other classes, we got an accuracy of 49%, which is relatively very low when compared to the normal this might be because of low number of data points available for training the model.

When we used up sampling of the data i.e., by downsizing the class representing the most and upscaling the remaining data to match that, we got an accuracy of 80% on train, 78 on validation and 76 on test, when we downsized the highest class to next highest class and upscaling the remaining data, we got an accuracy of 83.9% on train and 79% on test and validation. which is better than what we got without any sampling techniques. This model was the best among the three models that we built.

## Learning Curve and the Loss curve for the Up-sampling model

A picture containing text, screenshot

Description automatically generated Learning Curve and loss curve for model without data sampling  Learning Curve and loss curve for model with down sampling

Text

Description automatically generated

From the confusion matrix and the prediction we can observe that the prediction rate for our model appears to have the most inaccurate predictions for Basal cell carcinoma (code 3), followed by Actinic Keratoses (code 4), followed by Melanoma(code 1) and Melanocytic nevi (code 0), and Dermatofibroma(code 6) with the least incorrect predictions.

# Future Enhancements

We used Up sampling and down sampling methods for resampling of the data as the data is imbalanced data. We would like to use resampling techniques besides data augmentation like SMOTE techniques and see the results. Hyper parameter tuning to improve the F1 score of the model. We are planning to use Streamlet or flask packages to create front-end application for predicting the Input Image.

# References

<https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/DBW86T>

<https://www.kaggle.com/sid321axn/step-wise-approach-cnn-model-77-0344-accuracy>

<https://www.nature.com/articles/sdata2018161>

<https://www.jaad.org/article/S0190-9622(10)00394-4/fulltext>

[Noel C. F. Codella](https://arxiv.org/search/cs?searchtype=author&query=Codella%2C+N+C+F), [David Gutman](https://arxiv.org/search/cs?searchtype=author&query=Gutman%2C+D), [M. Emre Celebi](https://arxiv.org/search/cs?searchtype=author&query=Celebi%2C+M+E), [Brian Helba](https://arxiv.org/search/cs?searchtype=author&query=Helba%2C+B), [Michael A. Marchetti](https://arxiv.org/search/cs?searchtype=author&query=Marchetti%2C+M+A), [Stephen W. Dusza](https://arxiv.org/search/cs?searchtype=author&query=Dusza%2C+S+W), [Aadi Kalloo](https://arxiv.org/search/cs?searchtype=author&query=Kalloo%2C+A), [Konstantinos Liopyris](https://arxiv.org/search/cs?searchtype=author&query=Liopyris%2C+K), [Nabin Mishra](https://arxiv.org/search/cs?searchtype=author&query=Mishra%2C+N), [Harald Kittler](https://arxiv.org/search/cs?searchtype=author&query=Kittler%2C+H), [Allan Halpern](https://arxiv.org/search/cs?searchtype=author&query=Halpern%2C+A)

<https://journals.lww.com/melanomaresearch/Fulltext/2009/06000/Computer_versus_human_diagnosis_of_melanoma_.8.aspx>

<https://www.mayoclinic.org/diseases-conditions/skin-cancer/symptoms-causes/syc-20377605>