Newcastle University

REPORT: SKIN CANCER IMAGE RECOGNITION

CSC8635 Machine Learning with Project.

Parth Umeshkumar Shah
STUDENT ID: 210468180

TABLE OF CONTENT:

Introduction

- **Cancer**, it is one of the most common causes of death in the world's population. According to global statistics, about 10.0 million people would have died from cancer in 2020 (9.9 million excluding non-melanoma skin cancer).

- **Skin Cancer**, this term signifies the out-of-control growth of abnormal cells in the epidermis, in the outermost layer of the skin, it is caused by unrepaired DNA damage that triggers mutations. The sun's harmful ultraviolet (UV) radiation and the usage of UV tanning beds are the two main causes of skin cancer. These mutations cause skin cells to grow rapidly, resulting in cancerous tumours. Basal cell carcinoma (BCC), squamous cell carcinoma (SCC), melanoma, and Merkel cell carcinoma are the most common kinds of skin cancer (MCC).

- **Melanoma**, it is a type of cancer that arises from melanocytes, which are skin cells that generate the pigment melanin, which gives skin its colour.
  - Melanoma is the skin cancer with the worst prognosis, Melanomas look a lot like moles and can sometimes develop from them. They can be found on any part of the body, including those not often exposed to the sun.
  - Melanomas is curable when caught and treated early, it is diagnosed based on clinical evaluation and classic features on lesion biopsy. Overall, early detection is key for the effective treatment and better outcomes these cancers.
  - The time between detection and action is critical, yet diagnosis and human error can often drag things down.

- **Nonmelanoma** skin cancer encompasses all skin cancers that aren't melanoma. Nonmelanoma skin cancer encompasses a number of different forms of skin cancer, the most frequent of which are basal cell carcinoma and squamous cell carcinoma.

Aim:

- The goal of this study is to reduce the human error involved in diagnosis for biopsy by accurate recognition of images detecting the class of Skin Cancer.

- This detection is processed by **Machine Learning** (ML), it is an Artificial intelligence (AI) technique that uses statistical models and algorithms to learn from data in order to anticipate the properties of fresh samples and complete a task. Here, the complicated algorithms are designed to execute jobs that would otherwise be difficult for human brains to complete.

- The state-of-the-art network for pattern identification in medical image analysis is the **Convolutional Neural Network (CNN)**, a type of machine learning that models the processing of biological neurons.

- Here using CNN on loaded dataset, we aim to build a model to classify the skin images and recognize the type of skin cancer is possessed with reference to the processed image.

Convolutional Neural Network

- CNNs excels in analysing visual imagery as they are fully connected feed forward neural networks that reduce the number of parameters very efficiently without losing out on the quality of models.
- Convolutional Neural Networks (CNN or ConvNets) are regular neural networks that have picture inputs. They're used to classify and analyse photos, cluster images based on similarity, and recognise objects inside a frame. Convolutional neural networks (ConvNets or CNNs), for example, are used to recognise faces, individuals, street signs, cancers, platypuses, and a variety of other visual data.

Convolutional Neural Networks - how do they learn?

- Pixels are the building blocks of images. A number between 0 and 255 is assigned to each pixel. As a result, each image has a digital representation, which allows computers to manipulate them.
  - In CNN image detection/classification, there are four major procedures.
  - Shown below there are of major operations in CNN image detection/classification:

- 1.Convolution
  - Convolution is a mathematical procedure for combining two sets of data. Convolution is applied to the input data using a convolution filter to create a feature map in our project.
  - Here, the fundamental goal of a convolutional layer is to detect edges, lines, colour dips, and other visual elements in images. This is an intriguing property since it can recognise a characteristic in any region of the image after it has learned it at a specific place in the image.
  - Filters (also known as kernels or feature detectors) are used by CNNs to recognise characteristics such as edges that are present throughout an image. A filter is just a set of weighted values that have been taught to detect specific traits. The filter scans each area of the image to see if the feature it's looking for is there. The filter performs a convolution operation, which is an element-wise product and sum between two matrices, to generate a value expressing how confident it is that a specific feature is there.

- 2.Activation map
  - A method for getting the classified image regions used by a CNN to identify a given class in an image.A non-linear mapping is used to pass these feature maps. The feature maps are added together with a bias term and then run through ReLu, a nonlinear activation function. The activation function's objective is to induce non-linearity into our network because the images are made up of distinct objects that are not linear to one another, resulting in extremely non-linear images.
  - example
    - Activation Rel-U
    - ReLU (Rectified Linear Unit) activation makes all pixel values to be zero when a pixel image has a value of less than zero. [f(x) = x; if x>0 and 0; if x<= 0]
    - Rel-U activation layer in CNN is to increase the training stage on neural networks that have advantages to minimize errors.
- Activation map Another activation functions is sigmoid, tanh, leaky ReLu etc.

- 3.Max pooling
    - It is commonly used to minimise dimensionality. This allows us to limit the number of parameters, which reduces training time while also preventing overfitting. Each feature map is downsampled independently by pooling layers, reducing the height and breadth while maintaining the depth.
    - Using max-pooling or mean-pooling, the Pooling Layer is used to decrease data. The maximum value will be chosen by max-pooling, while the average value will be found via main pooling.

- 4.Flattening
    - Flattening means that anything greater than 1 dimension-array must be convert to 1D array/vector. This is done to feed the output of CNN to fully connected network to classify the input images.
    - All the rows are concatenated to form a long feature vector. If multiple input layers are present, its rows are also concatenated to form an even longer feature vector.

- 5.Fully connected layer
    - The fully connected layer is connected to the output layer is where we get the predicted classes.
    - A Fully Connected layer examines which high-level features are most strongly associated with a specific class and assigns weights to them so that when the products of the weights and the preceding layer are computed, the right probability for the various classes are obtained.

- The information is passed through the network and the error of prediction is calculated. The error is then backpropagated through the system to improve the prediction. Following this the Accuracy of the Machine-Learning model is also calculated and hence post this, working on different train-test dataset model comparison is made and the best model is recognized for future usage.

Brief On Dataset Used

- The HAM10000 dataset: This dataset was downloaded from Kaggle for skin Cancer Image Recognition Project, Dataset Consists of csv files and files containing the images to process in model.

- As we aim for classification, on analysing the csv dataset we observe that the dataset includes 7 attributes associated with each image and patient: lesion_id, image_id, dx, dx_type, age, sex and localization.
- Here 'lesion_id' and 'image_id' informs us about a lesion_id and a unique image_id
- 'dx' explains about the various lesions (injury) of melanoma. Here they are noted as:

    - **nv : (Melanocytic nevi)** Melanocytic nevi are benign melanocyte neoplasms that come in a variety of forms, all of which are covered in this series. From a dermatoscopic standpoint, the variants may differ dramatically.

    - **mel : (melanoma)** Melanoma is a malignant tumour that arises from melanocytes and comes in a variety of forms. If caught early enough, it can be treated with a simple surgical excision. Melanomas are cancerous tumours that can be invasive or non-invasive (in situ). All types of melanoma were considered, including melanoma in situ, although non-pigmented, subungual, ophthalmic, or mucosal melanoma were excluded.
    - **bkl : Benign keratosis)** like lesions Seborrheic keratoses ("senile warts"), solar lentigo (a flat version of seborrheic keratosis), and lichen-planus like keratoses (LPLK), which corresponds to a seborrheic keratosis or a solar lentigo with inflammation and regression, are all examples of "benign keratosis." The three

subgroups may appear to be different dermatoscopically, but we grouped them together since they are physiologically similar and are frequently reported histopathologically under the same generic title. Lichen planus-like keratoses are particularly challenging from a dermatoscopic standpoint since they might have morphologic features that resemble melanoma and are frequently biopsied or excised for diagnostic purposes.

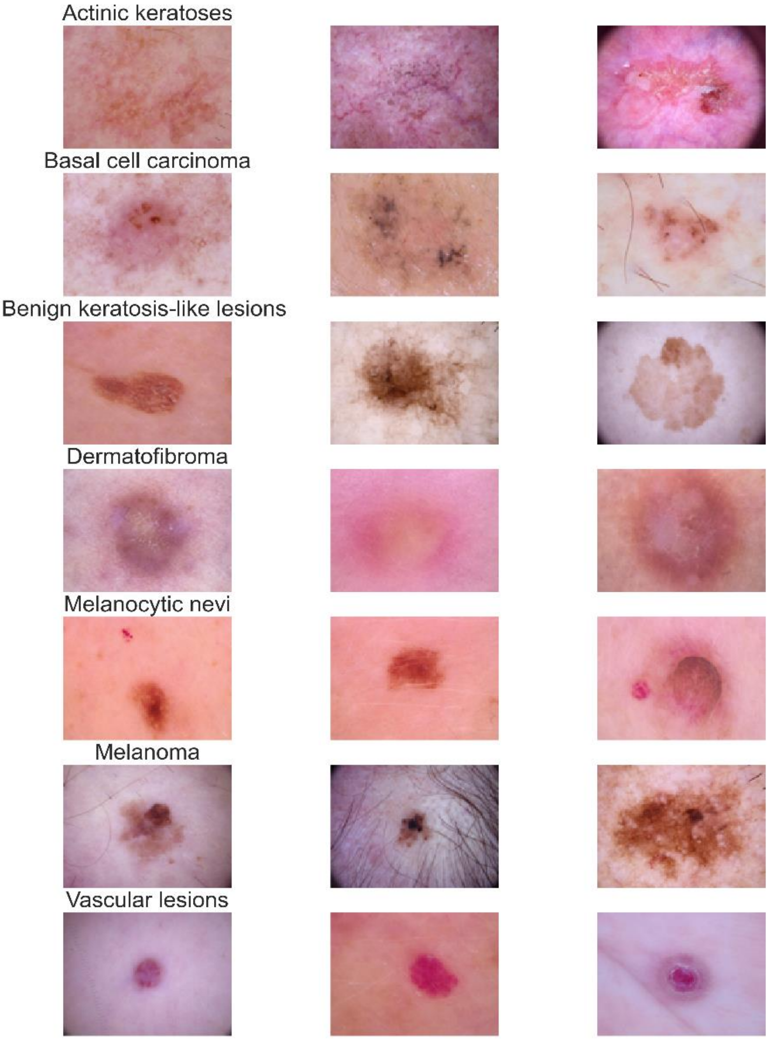


Figure 1/Lesions images

- **(bcc: Basal cell carcinoma)**
  Basal cell carcinoma is a type of epithelial skin cancer that seldom spreads but can be deadly if left untreated. It comes in a variety of morphologic forms (flat, nodular, pigmented, cystic, etc).
- **akiec : Actinic Keratoses** On the face, actinic keratoses are more common, while Bowen's disease is more common elsewhere on the body. Because both

forms are produced by UV light, the surrounding skin is usually sun-damaged, with the exception of Bowen's disease, which is caused by infection with the human papilloma virus rather than UV. Actinic keratoses have pigmented variations.

- **vasc : Vascular Lesions** Cherry angiomas, angiokeratomas, and pyogenic granulomas are among the vascular skin lesions in the dataset. This category also includes haemorrhage.
- **df : Dermatofibroma** Dermatofibroma is a noncancerous skin lesion that can be classified as either a benign growth or an inflammatory response to minor trauma. Dermoscopically, it is brown, with a central zone of fibrosis.

- 'dx_type' depicts about a technical validation field type, which indicates how the skin lesion diagnosis was made. The four types of technical validation fields are as follows:
  - Histopathology [Histo]
  - Confocal
  - Follow-up
  - Consensus
- 'age' and 'sex' provides information about the person of whose image are taken here in analyzation, this gives any idea about which age group has the skin cancer disease and about the gender of the person.
- Localization gives brief classification on what part of body we can find the skin cancer disease as here as per data:
  - back, lower extremity, trunk, upper extremity, abdomen, face, chest, foot, unknown, neck, scalp, hand, ear, genital, acral.

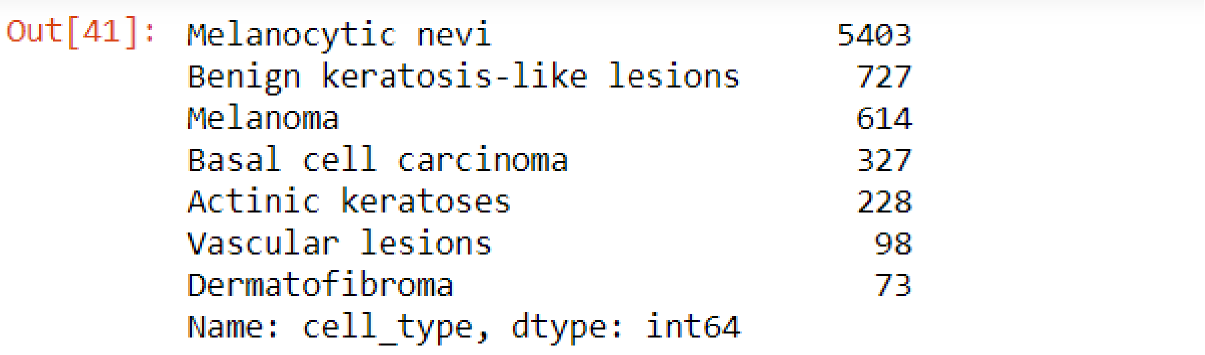
Data Import and Initial Pre-processing:

- Imported libraries of python and Raw data was loaded for essential analysis
- Added few columns for further exact,detailed and brief analysis
  1. Made a section i.e., column describing the type of skin cancer, eg ‘Melanocyctic nevi’ etc for proper outlook in analysis purpose.
  2. Here, first of all as we observe different Lesions, we introduced Cell label, a numerical acknowledgement with respect to each of the type of Lesion.
  3. Also added the image path and cell type details.
  4. It can be observed that there are only 7470 unique lesion id through data, here by, deleting rows with duplicate lesion Id.
  5. Post this, all the columns were checked where there may be null values present, on checking it was found out that, column describing the ‘age’ has 52 null values. Hence, filled in the gaps in the age column, which has 52 NA values, replacing it with the mean of the age column as age is a numeric value.
  6. Post analysing the images height and width, reduced the size of image by decreasing the parameters to 20% of its original loaded raw data parameters for betterment of model computational power, gaining a good accuracy and less losses.

Initial Exploratory Data Analysis:

Following observations:

1. On counting the label counts and cell-type we found out that most of the skin cancer cases in the following data are from Lesion - Melanocytic nevi and least are of Lesion ‘Dermatofibroma’ type.

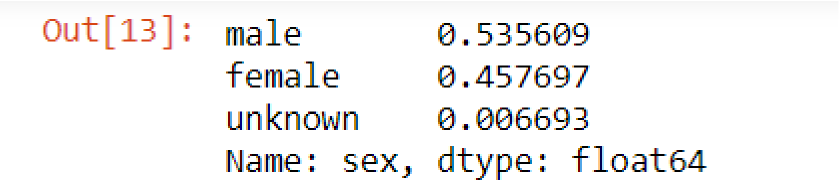


```
Out[41]: Melanocytic nevi                  5403
         Benign keratosis-like lesions      727
         Melanoma                           614
         Basal cell carcinoma               327
         Actinic keratoses                  228
         Vascular lesions                    98
         Dermatofibroma                      73
         Name: cell_type, dtype: int64
```

Figure 2/Count On Lesion images

2. The Unique lesion id present in the dataset is 7470 in total.

3. There are about 54% of male contribution in skin images and rest 46% are of females.



```
Out[13]: male       0.535609
         female     0.457697
         unknown    0.006693
         Name: sex, dtype: float64
```

Figure 3/total count based on Gender

4. Most of the people whose images are present for diagnosing of the disease are of age group “40-70”.

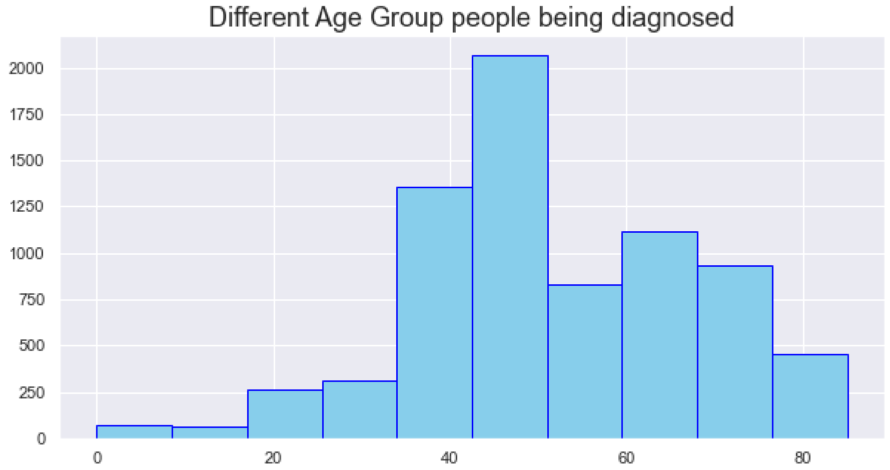
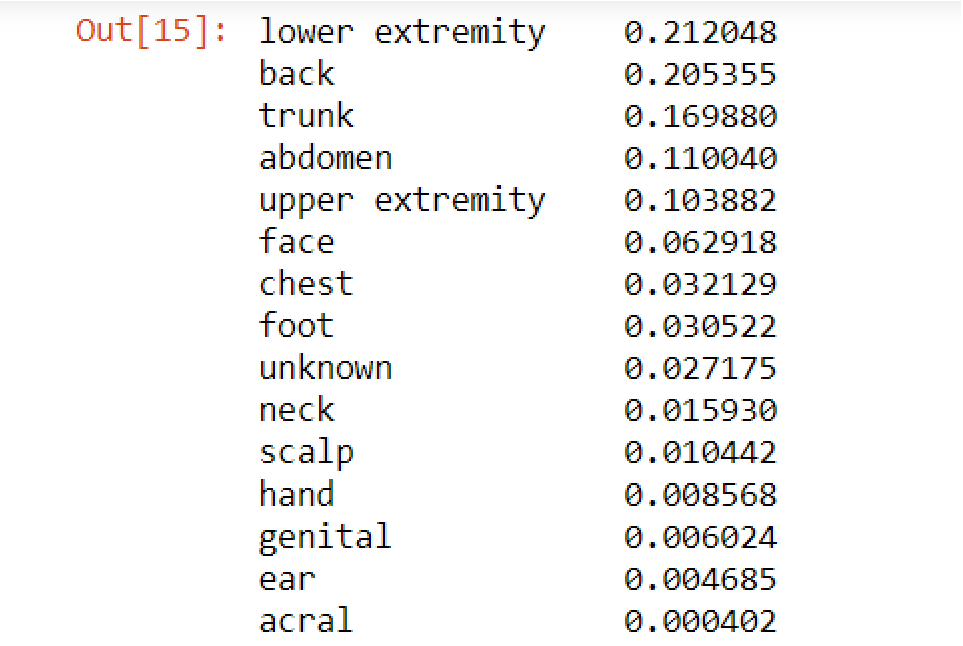


Figure 4/Age-Group Classified

5. About 22% of the skin cancer images are of body part - 'back' and ‘lower-extremity’ from the given dataset and least images were found from the 'acral' and ‘ear’.



```
Out[15]: lower extremity    0.212048
         back               0.205355
         trunk              0.169880
         abdomen            0.110040
         upper extremity    0.103882
         face               0.062918
         chest              0.032129
         foot               0.030522
         unknown            0.027175
         neck               0.015930
         scalp              0.010442
         hand               0.008568
         genital            0.006024
         ear                0.004685
         acral              0.000402
```

Figure 5/Parts of body affected in given data

6. From given data in males, most frequent affected part is ‘back’ whereas in females the most affected part is found to be ’lower extremity’, although in both genders the least is found in ‘acral’ part.

   According to the data, mainly most of the people are affected by back, lower extremity, trunk, abdomen, upper extremity and face skin cancers, visualization can be seen in figure below:

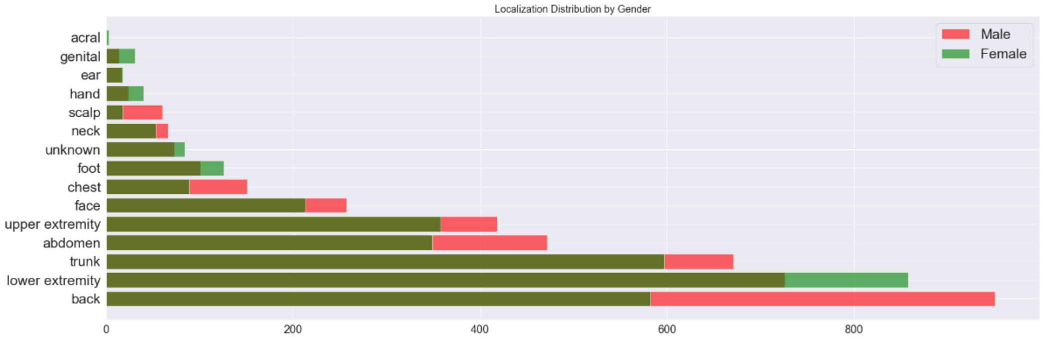


Figure 6/affected body part by gender

7. The Melanocytic nevi type lesions contributes the most in both male and female data used here for diagnosing purpose, followed by Benign keratosis, melanoma, basal cell carcinoma, actinic keratoses, vascular, dermatofibroma in order as mentioned here.

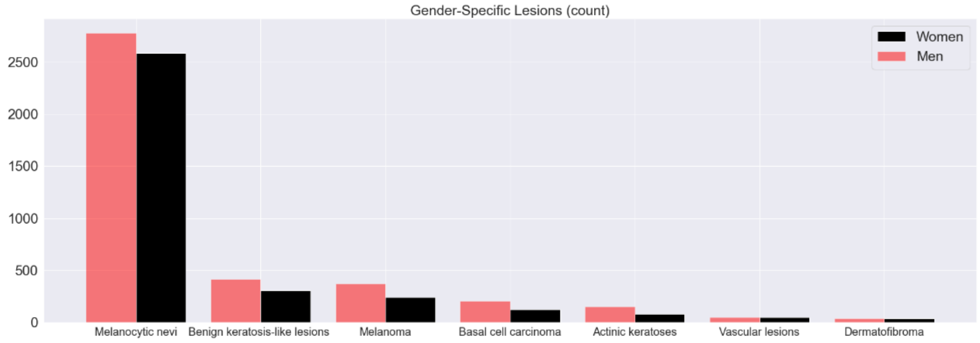


Figure 7/Distribution of lesions according to gender

8. Technical validation field type is majorly of Histopathology (histo) type in whole data. Details of it can be visualised in following graph:

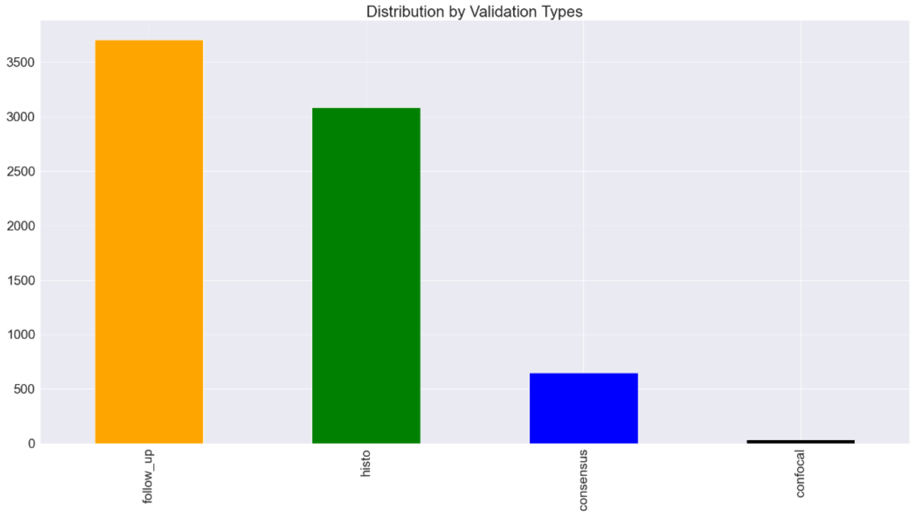


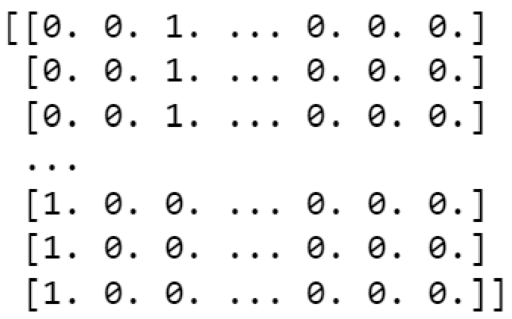
Figure 8/Count and distribution of validation type

CNN-Modelling

Data Segmentation:

- For Convolutional Neural Network (CNN) normal encoding won’t work for categorical data, thus performing one hot encoding for 7 lesions.
- Hot encoding is one method of converting data to prepare it for an algorithm and get a better prediction. Here by, we have first of all splitted the data and created two new data frames one consisting of whole pre-processed data without Labels and other one consisting of only label. Here ‘label’ refers to numerical numbering corresponding to the Lesions as defined in pre-processing.

Here we performed one hot encoding as an example Categorizing the label along with Number of Lesions, result obtained explains for label 1 through matrix plot can be concluded seeing 1 at the 2nd index, signifying that the cell label for the first row is 2 and the cell type is Lesions that resemble benign keratosis.



```
[[0. 0. 1. ... 0. 0. 0.]
 [0. 0. 1. ... 0. 0. 0.]
 [0. 0. 1. ... 0. 0. 0.]
 ...
 [1. 0. 0. ... 0. 0. 0.]
 [1. 0. 0. ... 0. 0. 0.]
 [1. 0. 0. ... 0. 0. 0.]]
```

Figure 9/Resulting Matrix of performed One hot encoding

Splitting dataset:

We started by separating the dataset into two parts: training and testing data. To develop our 75 percent Training - 10% Validation - 15% Testing model, we divide the 85 percent training data into 88.235 percent training data and 11.765 percent validation data. To ensure that each split includes enough data from each class for adequate modelling, the split will be applied to each class independently. As a result, each class will be split 75:10:15. This is performed by using the train test split method to set our goal to 'stratify.'

Data Preparation:

❖ **Applying normalisation to image variable to training, validation and test data:**

➢ Here First of all creation of list of image data is performed, this image data is the pixel data of each image which is a key part to be used in Cnn modelling.

➢ Post listing process of datasets, calculation is performed on training dataset to generate mean and standard deviation of training data, these parameters are then used on testing and validation data as the data sets are now will used in normalized form and hence, the mean and standard deviation parameters similar.

➢ Now the normalized data is calculated for each of the three datasets respectively, this is performed by subtracting the mean data from the actual data and dividing it by the standard deviation data of the training data. Normalized dataset is generated as Normalization is a data preparation technique that is used frequently, it is a process of converting the values of numeric columns in a dataset to a similar scale without distorting the ranges of values or losing information.

❖ **Data Augmentation:**

- In Machine learning models we perform this process to increase the total data amount by adding modified copies of the existing data or the pre-processed data, which was created here before the modelling steps.

- This process is performed as it regulates the data by reducing the variance and also helps in by decreasing the overfitting at the time when we train the model as while learning process of model, if overfitting is not reduced the model learns the noise and negativity increases to an extent that it impacts the model performance.

- Here in the performed CNN model we used augmentation technique to setup input mean and standard deviation, whitening; zooming; shifting and flipping the images on the normalized data which from now on now will be a part of the model characteristics.

Creation of CNN Model:

- As we begin the creation, firstly in this step we will be proceeding using the CNN techniques of convolution, activation map, map pooling, and flattening through to the fully connected layer to process it to the output layer.

- Here by, first using (calling) the Keras Sequential API (allows to create model layer-by-layer, although not suitable for models that share layers or thereby having multiple input and outputs; as in functional Keras API) for the creation of the model.

- Post this step Convolution 2D is applied followed by ReLu activation function (ReLU (Rectified Linear Unit) activation makes all pixel values to be zero when a pixel image has a value of less than zero), processing with Max-pooling 2D the model pixel parameters are flattened creating 1D array/vector which works as an input to the Artificial Neural Networks where the dense layer i.e. fully connected layer (detail description given in one of the previous sections) is added which connects the neural network to the output layer of our model.

- Post processing with the creation stage we got all parameters trained, none of the parameters was found untrained.

- The usage of root mean square error (rmse) was defined which will be useful in calculation of the losses and optimizer was called which defines the models loss functions and model result parameters i.e. accuracy, rmse, mean square error (mse) and mean absolute percentage error (mape).

Compilation and running of CNN Model:

Once the model is created, it is complied with use of optimizer and loss function, post it epoch and batch size of the model are set upped along with defining the call-backs.

a. Epochs: An epoch is how many times the model trains on our whole data set
b. Batch Size: Batch can be explained as taking in small amounts, train and take some more.
c. Each epoch must finish all batch before moving to the next epoch.
d. Call-backs:

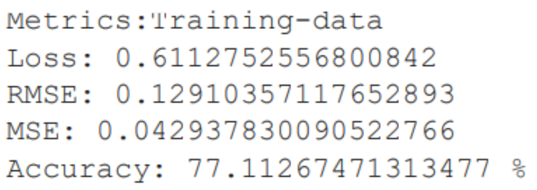
This model is fitted on batches of the normalized train data and normalized validation data after which the model is executed (run) and saved.

Results and Analysis of the CNN Model:

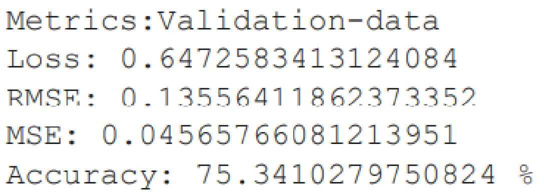
On Loading the model along, we can see the training data, validation data and testing data parameters results as shown below in figure:

Here we got Accuracy of 77.11% on training data, followed by testing data having 76.18% accuracy and Validation data possessing the 75.34% accuracy.

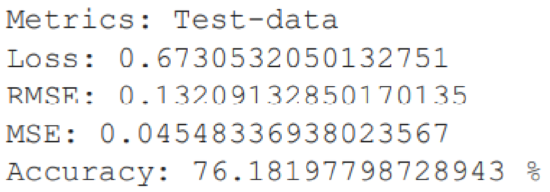
Details are shown below:



```
Metrics:Training-data
Loss: 0.6112752556800842
RMSE: 0.12910357117652893
MSE: 0.042937830090522766
Accuracy: 77.11267471313477 %
```



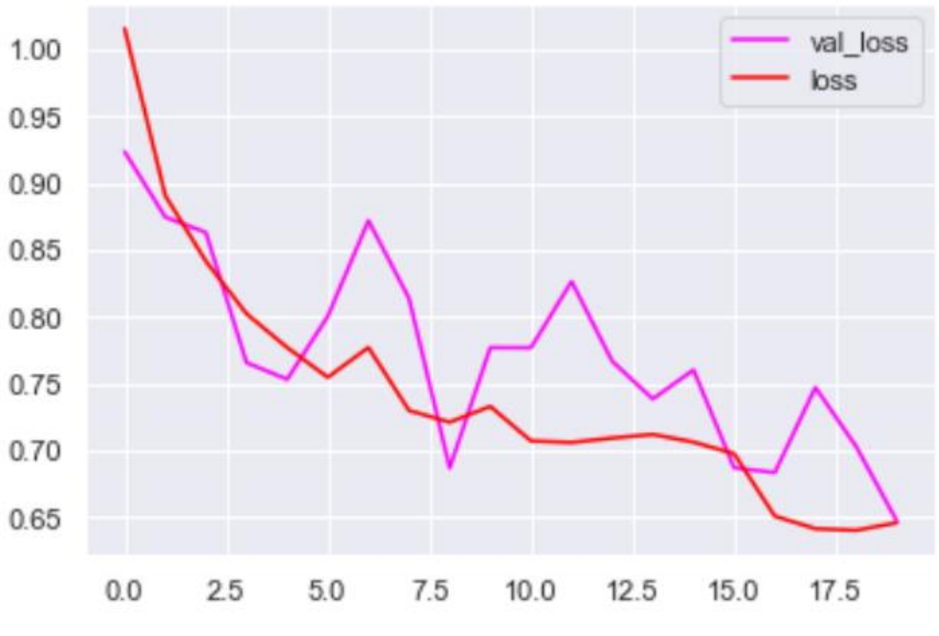
```
Metrics:Validation-data
Loss: 0.6472583413124084
RMSE: 0.13556411862373352
MSE: 0.04565766081213951
Accuracy: 75.3410279750824 %
```



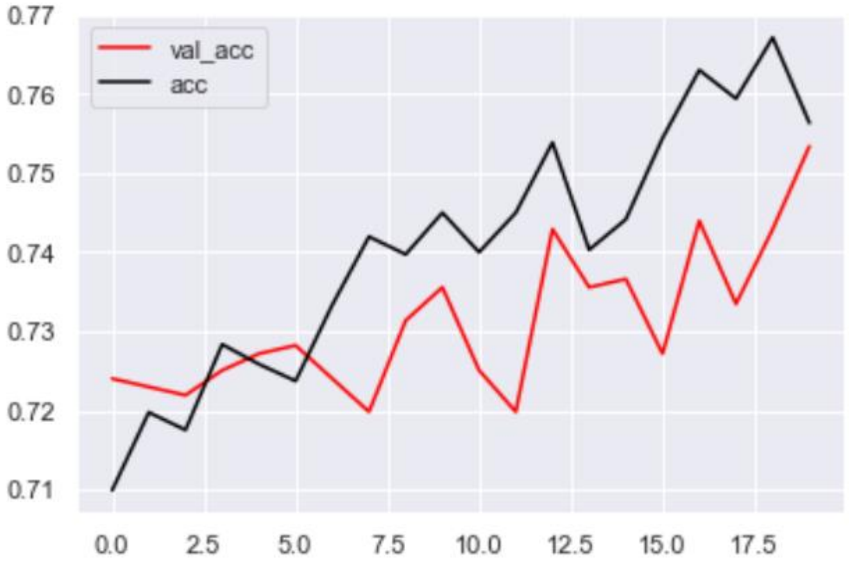
```
Metrics: Test-data
Loss: 0.6730532050132751
RMSE: 0.13209132850170135
MSE: 0.04548336938023567
Accuracy: 76.18197798728943 %
```

The comparison-plots of accuracy and losses are shown below:

- The plots describe the comparison between the testing and validation datasets, here we can see with respect to training dataset the losses are definitely more in validation data and accuracy is also less but comparable to training dataset, this defines that the model is working in right direction with a good amount of efficiency.



Losses comparison graph is shown above:

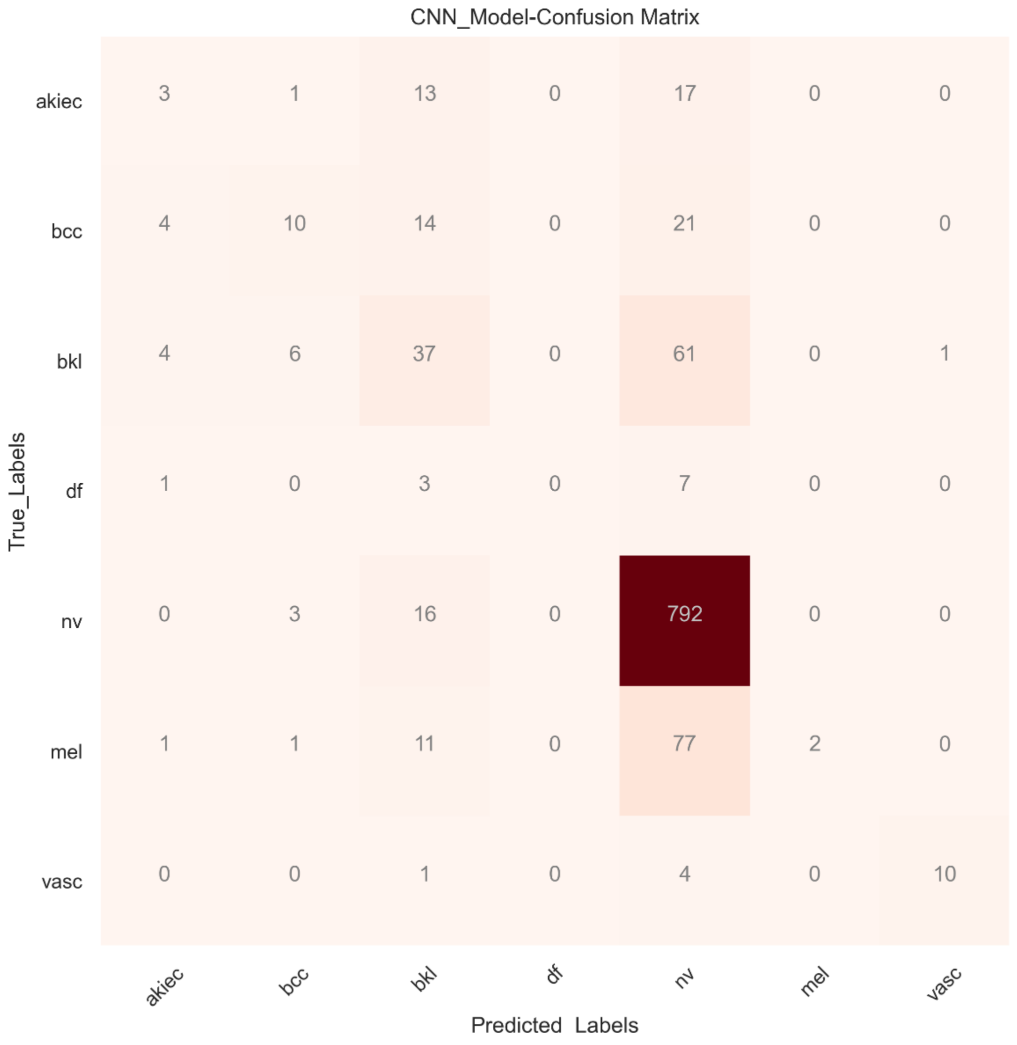


Accuracy plot is shown above:

Plotting the Confusion Matrix:

Confusion Matrix: It describes the summary of the results predicted here, on the skin cancer classification. Here the number of the correct and incorrect predictions are summarized with count values and splitted down by class.

Confusion matrix is shown below:



	akiec	bcc	bkl	df	nv	mel	vasc
akiec	3	1	13	0	17	0	0
bcc	4	10	14	0	21	0	0
bkl	4	6	37	0	61	0	1
df	1	0	3	0	7	0	0
nv	0	3	16	0	792	0	0
mel	1	1	11	0	77	2	0
vasc	0	0	1	0	4	0	10

Analysing the Confusion Matrix, we can say that the highest value in the above confusion matrix is 77, according to analysis it can be seen that the model is confusing between the images of “Melanoma nevi” and “Melanocytic nevi”. Also, there is confusion between “Melanocytic nevi” and “benign keratosis”.

Finally, A Classification Report is generated which shows the precision, recall, f1-score and support values in respect to each label.

There are four techniques to determine if the forecasts are correct or incorrect:

- True Negative (TN): the case was negative and was projected to be negative.
- True Positive (TP): the case was positive and projected to be positive.
- FN / False Negative: the case was positive, but the outcome was projected to be negative.
- FP / False Positive: a case that was projected to be positive but was really negative.

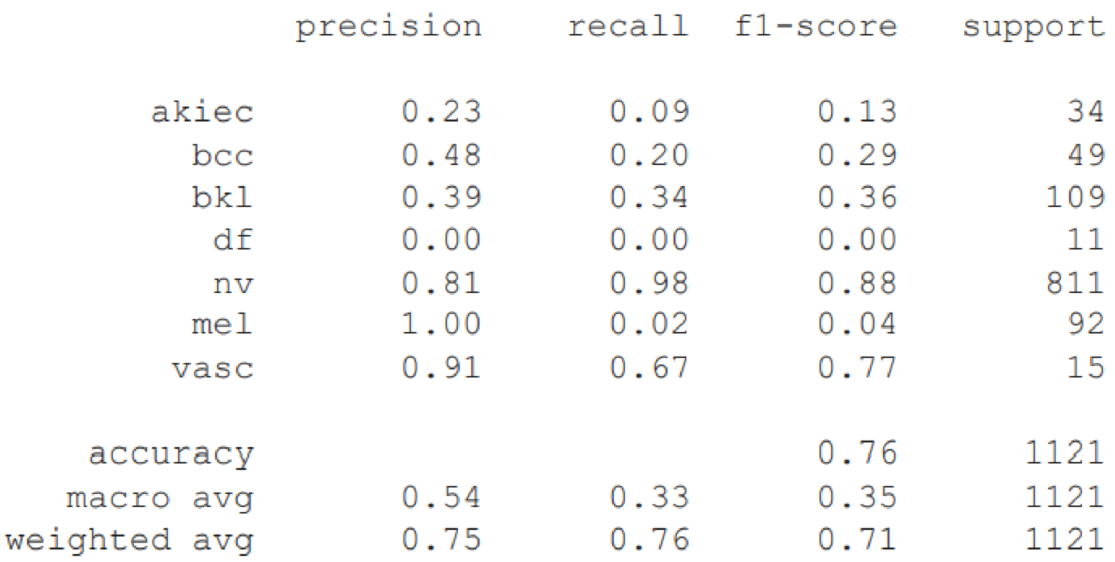
a) Precision: it is the ratio of true positives to the sum of true and false positives.

[Precision:- Accuracy of positive predictions.
Precision = TP/(TP + FP)]

b) Recall: it is the ratio of true positives to the sum of true positives and false negatives.
[Recall:- Fraction of positives that were correctly identified.
Recall = TP/(TP+FN)]

c) F1-score is the weighted harmonic mean of precision and recall, the closer the F1 score is near 1.0, the better the model's projected performance will be.
As a rule of thumb, the weighted average of F1 should be used to compare classifier models, not global accuracy.
F1 Score = 2*(Recall * Precision) / (Recall + Precision)

d) Support: The number of actual instances of the class in the dataset is known as support. It does not differ between models; rather, it diagnoses the process of performance evaluation. Support doesn’t change between models but instead diagnoses the evaluation process.

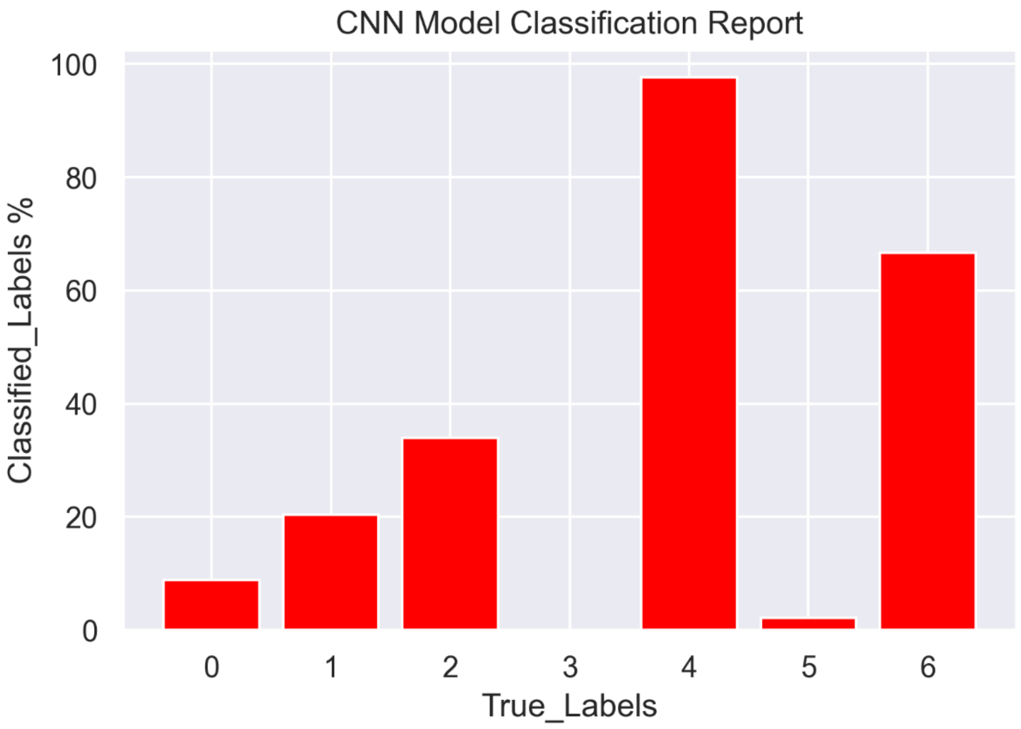


```
              precision    recall  f1-score   support

       akiec       0.23      0.09      0.13        34
         bcc       0.48      0.20      0.29        49
         bkl       0.39      0.34      0.36       109
          df       0.00      0.00      0.00        11
          nv       0.81      0.98      0.88       811
         mel       1.00      0.02      0.04        92
        vasc       0.91      0.67      0.77        15

    accuracy                           0.76      1121
   macro avg       0.54      0.33      0.35      1121
weighted avg       0.75      0.76      0.71      1121
```

Classification Report Can be shown below:



Plotting recall value vs true labels shows how much percentage of the fraction positives of each label i.e., lesions in numeric, that were correctly identified.

Here it can be visualized that label-3 classification report shows nil, which confirms that the lesion assigned with this designated label does not have enough data for modelling predictions.

Conclusion:

- The model returned good accuracy was a pretrained model with convolutional layers and additional fully connected layers. The batch size was 20 and it returned a training accuracy score of 77.11% and a testing accuracy score of 76.18%. The training also included class weights to give more attention to label 3 – basal cell carcinoma- and label 5 – vascular lesions- with the intention of increasing the recall for those specific classes. However, according to the confusion matrices and classification accuracy plot in the model we ran, it could be seen that the more training data is key for better performance of model at predicting them during test time. More images of one having better positive statistics in classification report had to be provided for more accurate results and less losses in classification, as the main aim in the medical field is to reduce the number of patients dying. Overall, this project highlighted the potential of CNNs in the context of disease classification to lead further advances in the medical field with its image recognition abilities.

❖ To improve on these outcomes, the following measures must be taken:

➢ Investigate different skin lesion data sets to get more training data for our models, particularly for underrepresented and malignant skin lesion classifications.

- Improve the basic CNN's design and implementation by fine-tuning hyperparameters such as the number of layers, the kind of layers, and the hyperparameter values of the layers. With the right processing resources, it is conceivable to maximise all of the available hyperparameters in each network.

- Using **Jupyter Notebook to** build the report not only helped to give an interactive data science environment (IDE), but it also gave a means to display the analysis and code written in the Python language in a clear manner. Using Jupyter notebook, it was simple to combine graphs, images, and code with the data. In this project, it was possible to create the report in markdown alongside the code using Jupyter Notebook, which increased readability and reproducibility.

References:

1. https://www.skincancer.org/skin-cancer-information
2. https://towardsdatascience.com/applied-deep-learning-part-4-convolutional-neural-networks-584bc134c1e2
3. https://www.degruyter.com/downloadpdf/j/itms.2017.20.issue-1/itms-2017-0003/itms-2017-0003.pdf
4. https://towardsdatascience.com/best-python-libraries-for-machine-learning-and-deep-learning-b0bd40c7e8c