Newcastle University

CSC8635

Machine Learning Extended Project

Name: Shubham Rajesh
Student No.: 210462188

Table of Contents

Introduction:

The skin is composed of three fundamental layers. Skin cancer originates in the outermost layer, which is constituted of squamous cells in the first layer, basal cells in the second layer, and melanocytes in the third layer. Squamous and basal cells are sometimes referred to as non-melanoma tumours. Non-melanoma skin cancer is always curable and seldom metastasizes to other skin tissues. Melanoma is a more deadly kind of skin cancer than the majority of other forms. If it is not diagnosed early on, it rapidly invades surrounding tissues and spreads to other regions of the body. The biopsy is the formal diagnostic approach for skin cancer detection. A biopsy is a procedure that involves the removal of a piece of tissue or a sample of cells from the patient's body for laboratory analysis. It is an inconvenient procedure. The biopsy method is time-intensive for both the patient and the physician due to the length of time required for testing. The biopsy is performed by extracting skin tissue (skin cells) and subjecting the sample to a battery of laboratory tests. There is a risk of sickness spreading to another portion of the body. It is more dangerous. Considering the aforementioned situations, it is recommended to diagnose skin cancer using CNN. This methodology uses a digital image processing tool with CNN for classification. This technology has inspired the early diagnosis of skin cancer and requires that the patient should not apply oil to your skin to get clear sharp images of your moles. In this sense, it's a speedier, cleaner and less painful method. But, most crucially, because of its increased magnification, Skin Cancer Detection Using CNN may avoid the wasteful removal of completely innocuous moles and skin blemishes (Ansari, U.B. and Sarode, T., 2017).

Objective:

The purpose of this research is to cut down the amount of human error involved in biopsy diagnosis by recognising the type of skin cancer that is being investigated in the images. A machine learning (ML) algorithm is used to perform this recognition. Machine learning is an Artificial intelligence (AI) technology that employs statistical models and algorithms to learn from data in order to predict the attributes of new samples and accomplish the job. In this case, the sophisticated algorithms are intended to carry out tasks that would otherwise be difficult for human brains to do by themselves. The goal is to develop a model that can analyse skin images and identify the type of skin cancer that is presently based on the image that has been processed.

Convolution Neural Network:

Convolutional Neural Networks have achieved ground-breaking breakthroughs in a range of domains connected to pattern identification during the last decade, ranging from image processing to speech recognition. The most advantageous feature of CNNs is that they reduce the number of parameters in ANNs. This accomplishment has motivated both academics and developers to consider bigger models in order to perform difficult tasks, which was previously impossible with classical ANNs. The most fundamental premise is that problems handled by CNN should not have spatially dependent aspects. In other words, in an image recognition application, for example, it is not required to be concerned with the location of the features in the photos. The only thing that matters is that they are detected regardless of their location in the provided photos. Another critical component of CNN is the

acquisition of abstract characteristics when data propagates deeper levels (S. Albawi, T. A. Mohammed and S. Al-Zawi, 2017)

How does CNN perform Image classification?

Pixels are the fundamental units of a picture. Each pixel is assigned a value between 0 and 255. So, it is possible that each image is represented digitally, allowing computers to modify them. The CNN architecture includes various building blocks, including convolution layers, pooling layers, and fully connected layers. A typical design is composed of repeated layers of many convolution layers and pooling layers, followed by one or more fully linked layers.

1. Convolution layer:

Convolution is a mathematical method for merging data sets. This research utilizes a convolution filter on the input data to produce a feature map. Convolutional layers are used to detect edges, lines, colour dips, and other visual characteristics in an image. It can distinguish a characteristic in any portion of the image after learning it at a specific location. CNN use a filter also known as kernels or feature detector to detect edges in images. A filter is simply a set of weighted values trained to identify certain qualities. The filter examines the image for the feature it seeks. For each feature, the filter conducts a convolution operation (an element-wise product and sum between two matrices) to generate a result.

2. Activation map:
   It is a technique to retrieve a CNN's categorised image regions used to determine a class in a picture. These feature maps are passed non-linearly. They are then combined with a bias term and ReLu, a nonlinear activation function. Because images are made up of discrete objects that are not linear to one another, the activation function's goal is to create non-linearity in our network.

3. Max pooling:
   It is used to reduce dimensionality. This minimises training time while preventing overfitting. Pooling layers reduces the height and width of each feature map while retaining the depth. The pooling layer of the data is reduced data by using max-pooling or mean-pooling. Max-pooling selects the maximum value, while main pooling finds the average.

4. Flattening:
   Flattening involves converting 2D arrays to 1D arrays. So that a fully connected network can classify the input images. This creates a lengthy feature vector. Multi-input rows are concatenated to generate a longer feature vector.

5. Fully connected layer:
   Fully connected layer - A Fully Connected layer determines which high-level traits are most closely related to a certain class and weights them accordingly.

Then the information is passed through the network to calculate the prediction error. The algorithm then uses error to enhance prediction. Then the model's accuracy is calculated, and the best model is chosen for future use.

Data Description:

There are in total 2 data sets:

1. The “HAM10000” dataset was taken from Kaggle for the Skin Cancer Image Recognition Project. The CSV dataset includes 7 variables connected with each image and patient: lesion_id, image id, dx, dx type, age, sex, and localization.

2. Lesions (injuries) of melanoma are described by 'lesion id' and 'image id'. Below are the lesion type considered:
   - **nv: (Melanocytic nevi)**
     This series covers benign melanocyte neoplasms in all their forms. Dermatologically, the variants may be vastly different.
   - **mel : (Melanoma**)
     Melanoma is a malignant tumour that develops from melanocytes. Early detection allows for straightforward surgical excision. Melanomas are malignancies that can be invasive or non-invasive (in situ). Non-pigmented, subungual, ocular, and mucosal melanoma were excluded.
   - **bkl: (Benign keratosis**)
     Benign keratoses Inflammation and regression of seborrheic keratoses or solar lentigo are all examples of "benign keratoses." Although the three subgroups appear to be distinct dermoscopically, they are physiologically related and frequently reported histopathological under the same general label. They are routinely biopsied or removed for diagnostic purposes and can have morphologic features similar to melanoma.
   - **bcc: (Basal)**
     Basal cell carcinoma is a rare epithelial skin cancer that can be fatal if ignored. It has various morphological forms (flat, nodular, pigmented, cystic, etc).
   - **akiec: (Actinic Keratoses)**
     Actinic keratoses are more common on the face, but Bowen's disease is more common elsewhere. Except with Bowen's disease, which is caused by a papillomavirus infection rather than UV light, the surrounding skin is frequently sun-damaged. Actinic keratoses can be pigmented.
   - **vasc: (Vascular Lesions)**
     The dataset includes cherry angiomas, angiokeratomas, and pyogenic granulomas. This also includes haemorrhage.
   - **df: (Dermatofibroma)**
     Dermatofibroma is a benign development or an inflammatory response to minor trauma. Dermoscopically, it is brown with a fibrotic core.

3. The localization variable tells us where on the body we can discover skin cancer. Below are the variables considered in the data set:

back, ear, neck chest, face, lower extremity, trunk, upper extremity, abdomen, scalp, genital, acral, foot, hand, unknown.

4. "dx type" variable shows a technical validation field type for skin lesion diagnosis. There are four types of technical validation fields considered in the data set:
   - Histopathology
   - Confocal
   - Consensus
   - Follow-up

5. The age and gender of the person whose image is being examined by provided information on the data set person's "age" and "gender".

Data Pre-processing:

1. Start with importing required libraries and raw data in Jupyter Notebook.
2. A more precise, thorough, and concise analysis of new columns were added.
3. Added the image path and cell type details
4. Cell labels were created, a numerical acknowledgement for each type of Lesion.
5. Created a column describing the type of skin cancer, e.g. 'Melanocyctic nevi' for accurate analysis.
6. As there are only 7470 unique lesion ids in the data, duplicate lesion ids are deleted.
7. There are 54 NAs in the Age variable, so this can be handled by replacing the NAs with the mean of the vales of Age.
8. The width and height of images in raw data are reduced by 20% for improving the computational power of the model and gaining greater accuracy.

Exploratory Data Analysis:

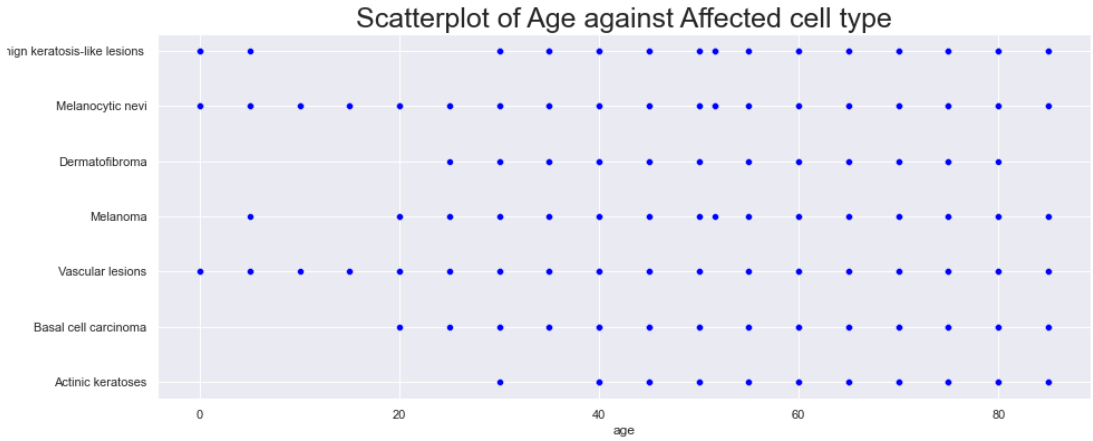
1. There are around 7470 total unique instances in the data set.

2. Count the cell type present in the data present. We can see that there are around 6705 instances where the "Melanocytic nevi" cell type was detected.

```
Out[6]: Melanocytic nevi                6705
        Melanoma                        1113
        Benign keratosis-like lesions   1099
        Basal cell carcinoma             514
        Actinic keratoses                327
        Vascular lesions                 142
        Dermatofibroma                   115
        Name: cell_type, dtype: int64
```

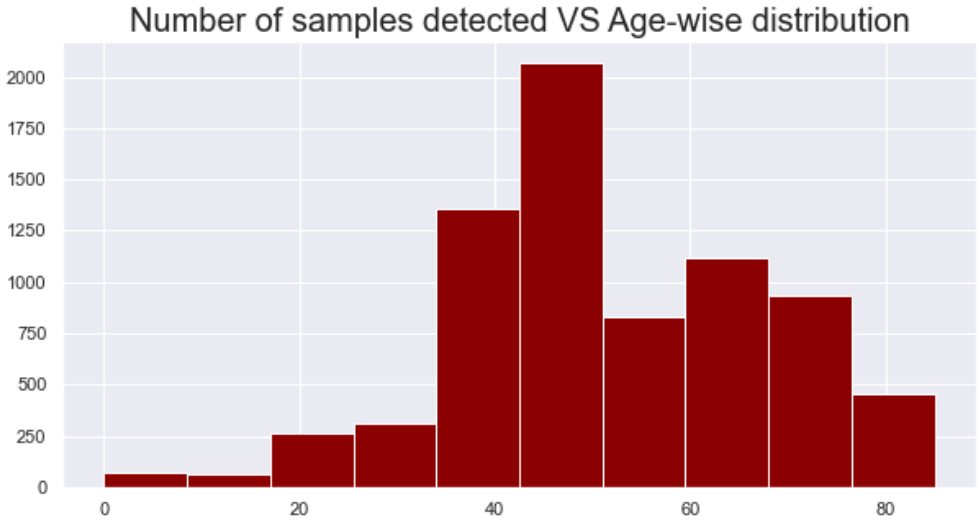
3. This scatter plot shows the distribution of Age against every cell type in the data set.

It can be seen that skin cancer is mostly diagnosed in people above age 35. In younger patients (age 0 to 20), “high keratosis like lesions”, “Melanocytic” and “Vascular lesions” are mostly found. The skin cancer type, “Dermatofibroma”, “Actinic keratoses”, “Basal cell carcinoma” can be found in most people above age 25.



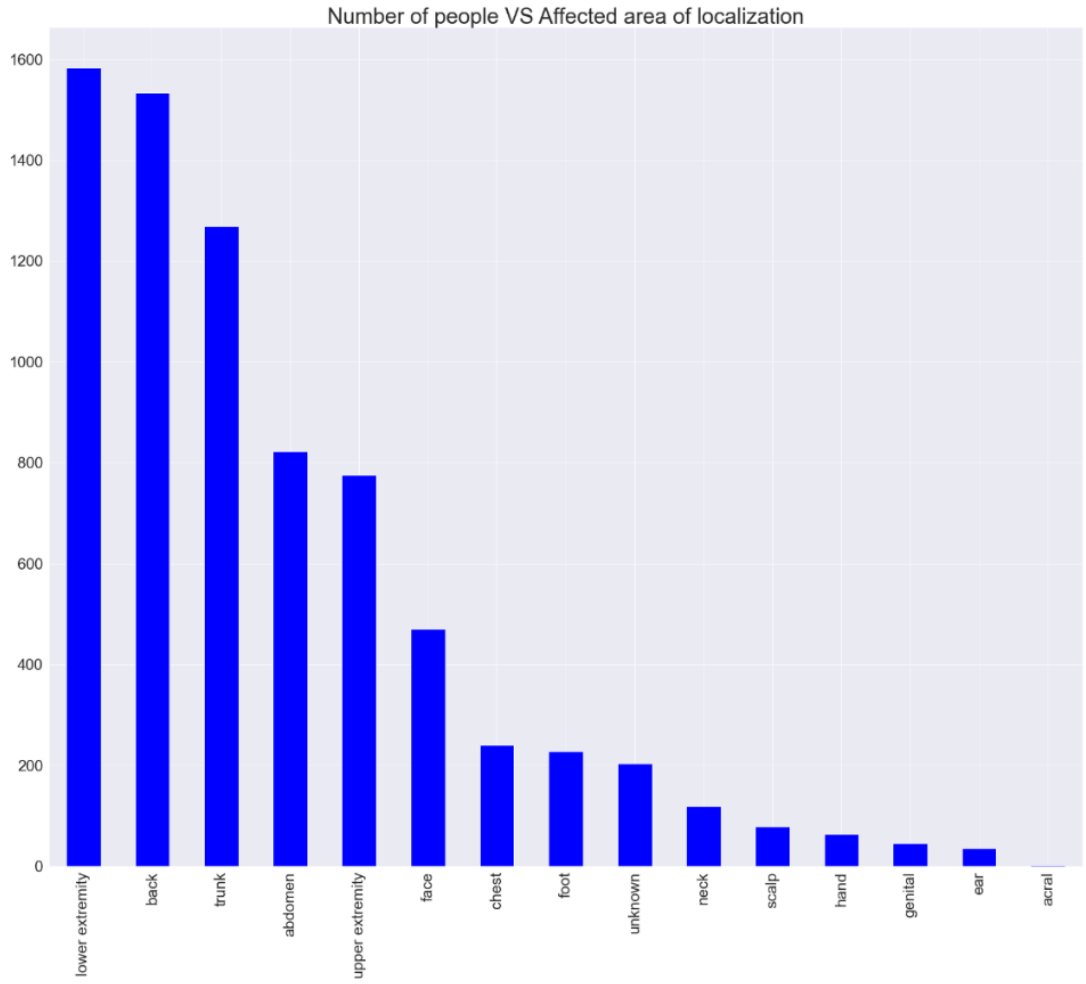
4. Histogram distribution of number of samples with respect to age.

Here, it can be clearly seen that the most people diagnosing of skin cancer are between age of 40 to 50.



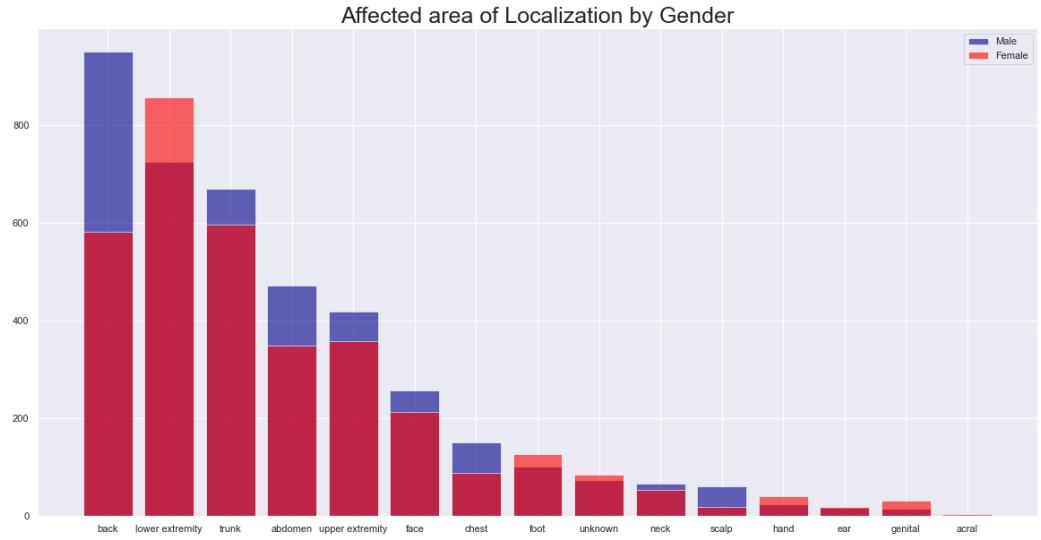
5. Bar plot of analysis between the number of people and area where the skin cancer is detected.

It can be inferred from the graph that most of the detected cases are from “lower extremity” and “back” whereas very rare cases are detected in “acral” and “ear”.



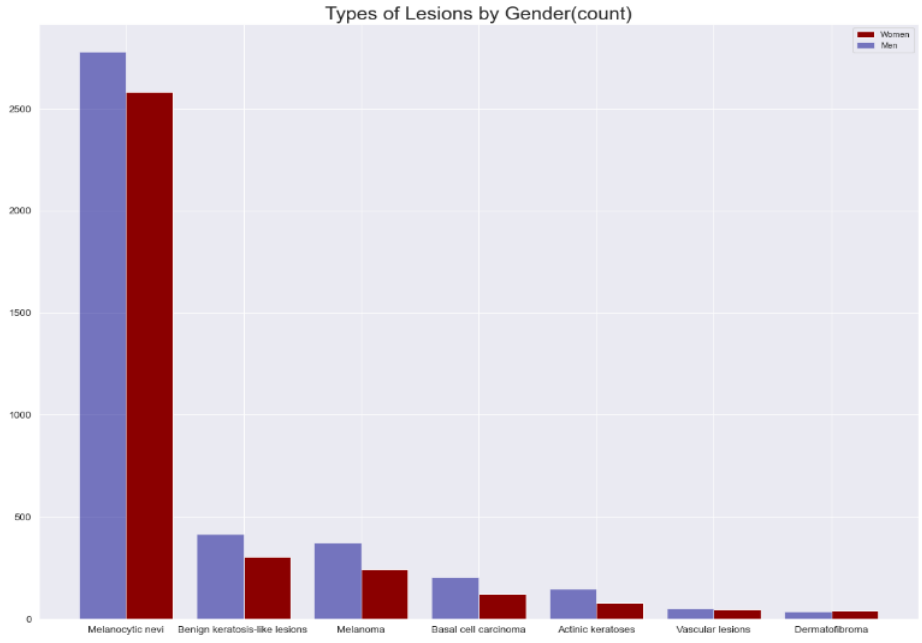
6. Stack bar chart for analysing the number of cases according to the affected area where the colours red and blue are used for females and males respectively.

It can be seen that the “back” is most affected by skin cancer in men whereas the most affected area by the skin in women is “lower extremity”. There are no traces that the affected area “acral” is found for males.



7. The analysis about the lesion type and the gender is visualized using a double bar graph.

It can be clearly seen that “melanocytic nevl” skin cancer is the most common type of skin answer in men and women. It can also be inferred that from the given data set, skin cancer is more common in men than women.



Data Modelling:

Splitting dataset:

Modelling began with splitting the dataset into two sections: training data- 50% and testing data-50%. Then the training data was divided into Validation-25% and training data as 75%. To guarantee that each split contains sufficient data from each class for acceptable modelling, each split will be applied independently to each class.

Normalizing image variables for training, validation, and test data:

To begin, a list of image data is created; this image data consists of the pixel data for each image, which is a critical component for CNN modelling. Now, for each of the three datasets, normalised data is calculated by removing the mean from the actual data and dividing it by the training data's standard deviation.

One hot encoding:

One hot encoding is a technique for transforming categorical data variables into variables that may be used by machine learning algorithms to increase prediction accuracy. One-hot encoding is a critical component of machine learning feature engineering.

Data Augmentation:

This procedure is used in machine learning models to enhance the total amount of data by adding mod ified copies of current data or preprocessed data that was prepared prior to the modelling processes. T his process is carried out because it helps in regulating the data by reducing the variance and also help s in reducing overfitting during the model's learning process, If overfitting is not addressed, the model learns the junk and wrong assumption are considered to the point where the model's performance is impaired.

CNN Model Creation:

1. In this step, CNN techniques like convolution, activation mapping, map pooling, and flattening are utilised to process the fully connected layer to the output layer.

2. Keras Sequential API allows for layer-by-layer model development but is incompatible with models that share layers or have numerous input and outputs; as in functional Keras API) for model creation.

3. The convolution 2D function is applied, followed by the ReLu activation function (the ReLU (Rectified Linear Unit) activation function makes all pixel values be zero when a pixel image has a value less than zero), processing with Max-pooling 2D, the model pixel parameters are flattened creating a 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.

4. In post-processing with the creation stage, all of the parameters are trained, with no untrained parameters being discovered.

CNN Model fitting:

In this step the model is run on the train data set and validation dataset, to check the model's accuracy. The training data used for this purpose is augmented and normalised x_train data, as well as y_train data. The validation data used is normalised x validation and y validation data. Additionally, during model fitting, other variables such as batch size, epoch size, verbose, steps per epoch, and call back_lr are supplied in the function.

Saving the Model:

Here the model built on test and validation data is saved in the file system, which can also be viewed by any other person without actually running the model again.

Loading the Model:

Here the value of Root means square error is loaded and the also the saved model is loaded from the file system.

Checking the accuracy and other metrics for the train and validation model:

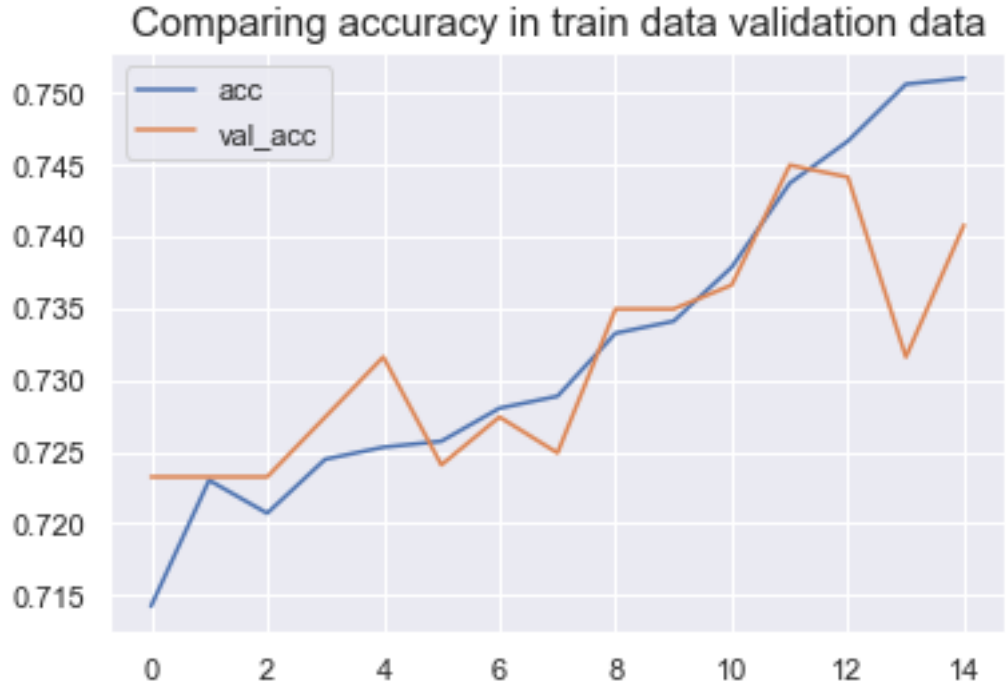
In this step, the metrics like Accuracy, loss, RMSE and MSE for train and validation data is calculated.

Checking the accuracy and other metrics for the test model:

In this step, the metrics like Accuracy, loss, RMSE and MSE for test data is calculated.

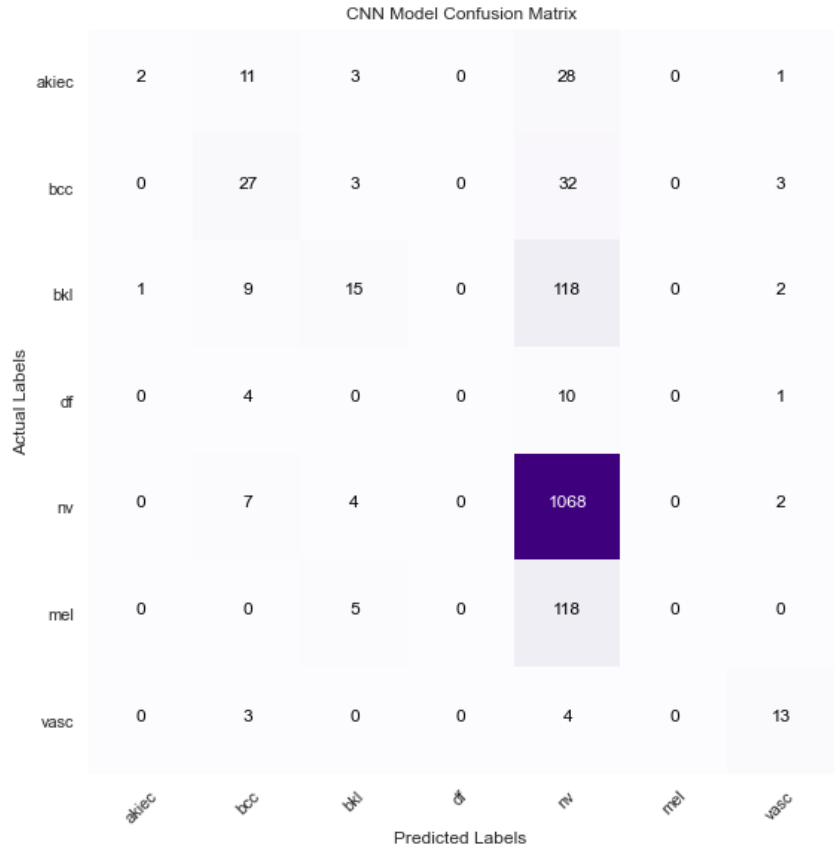
```
Metrics for Train Data
Accuracy: 78.70292663574219 %
Loss: 0.5805390477180481
RMSE: 38596720.0
MSE: 0.04023636505007744

Metrics for Validation Data
Accuracy: 74.16387796401978 %
Loss: 0.7650584578514099
RMSE: 42502608.0
MSE: 0.046362441033124924

Metrics for Test Data
Accuracy: 76.90762877464294 %
Loss: 0.6776171922683716
RMSE: 40009828.0
MSE: 0.04274426028132439
```

Plots for comparison of loss and accuracy between train and validation:



Confusion Matrix:

A confusion matrix is a way of describing a classification algorithm's performance. Classification accuracy alone may be deceptive if the number of observations in each class is uneven or if your dataset has more than two classifications. Calculating a confusion matrix may help you see where your classification model is succeeding and where it is failing.



The highest value in the above confusion matrix is 118, according to analysis it can be seen that the model is confusing between the images of “mel” and “nv”. Also there is confusion between “nv” and “bkl”.

Classification Report:

A classification report is a machine learning performance statistic. It is used to demonstrate your trained classification model's accuracy, recall, F1 Score, and support.

There are four ways to check if the predictions are right or wrong:

TN / True Negative: the instance was negative and predicted negative

TP / True Positive: the instance was positive and predicted positive

FN / False Negative: the instance was positive but predicted negative

FP / False Positive: the instance was negative but predicted positive

Precision — *What percentage of your predictions were correct?*

Precision refers to a classifier's ability to avoid labelling a negative instance as positive. It is defined as the ratio of true positives to the ratio of true positives and false positives for each class.

Precision: *Accuracy of positive predictions.*

Precision = TP/(TP + FP)

	precision	recall	f1-score	support
akiec	0.67	0.04	0.08	45
bcc	0.44	0.42	0.43	65
bkl	0.50	0.10	0.17	145
df	0.00	0.00	0.00	15
nv	0.78	0.99	0.87	1081
mel	0.00	0.00	0.00	123
vasc	0.59	0.65	0.62	20
accuracy			0.75	1494
macro avg	0.43	0.31	0.31	1494
weighted avg	0.66	0.75	0.67	1494

Recall — *What percentage of the positive cases did you catch?*

Recall is a classifier's capacity to discover all positive occurrences. The ratio of true positives to the total of true positives and false negatives is determined for each class.

Recall: Fraction of positives that were correctly identified.

Recall = TP/(TP+FN)

F1 score — *What percentage of positive predictions were correct?*

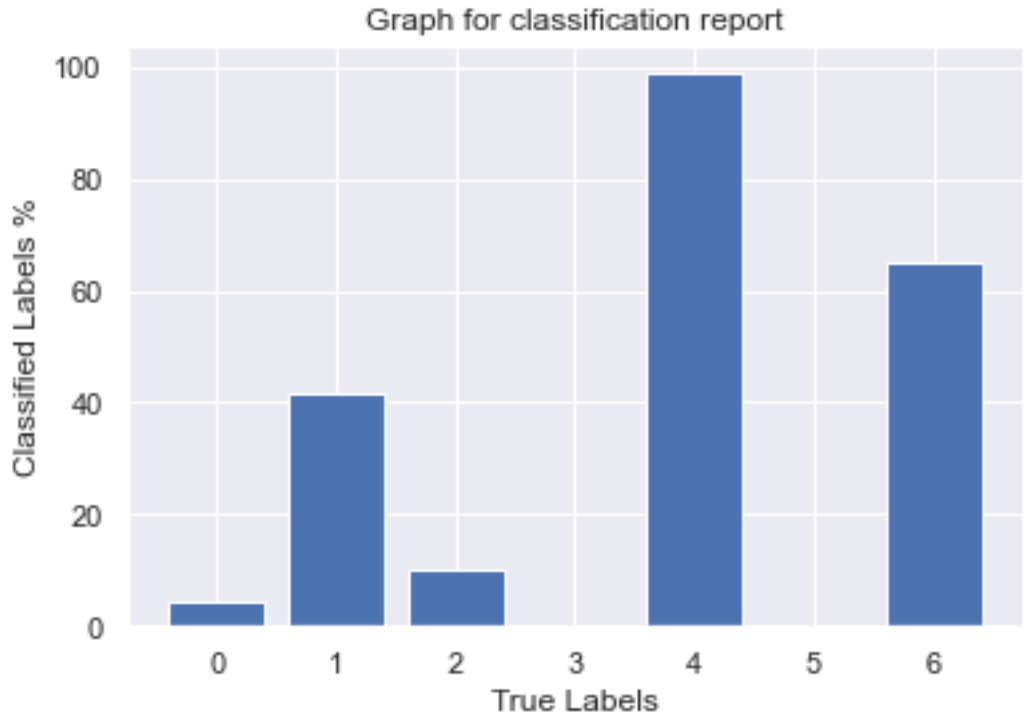
The F1 score is a weighted harmonic mean of precision and recall, with a maximum of 1.0 and a minimum of 0.0. F1 scores are lower than accurate measurements due to the calculation of precision and recall. As a general rule, classifier models should be compared on the basis of their weighted average of F1, not their global accuracy.

F1 Score = 2*(Recall * Precision) / (Recall + Precision)

Support

The number of actual instances of the class in the provided dataset is called support. Poorly balanced support in the training data may suggest fundamental flaws in the classifier's reported scores, indicating the necessity for stratified sampling or rebalancing. Support does not vary across models; rather, it serves as a diagnostic for the evaluation process.

Below is the classification report for skin cancer data analysis:



Conclusion:

CNN Model gave accuracy about 74% for validation data and 78% for training data, whereas 77% for Test data. It can also be concluded that, this accuracy can be increased by increasing the epoch and batch size. For class 3 Basal cell carcinoma and class 5 Vascular lesion there are no images classified.

```
Metrics for Train Data
Accuracy: 74.8744785785675 %
Loss: 0.6693912744522095
RMSE: 42440116.0
MSE: 0.0457882396876812

Metrics for Validation Data
Accuracy: 74.08027052879333 %
Loss: 0.750418484210968
RMSE: 44122864.0
MSE: 0.048736702650785446


Metrics for Test Data
Accuracy: 75.30120611190796 %
Loss: 0.7361882328987122
RMSE: 43168968.0
MSE: 0.0478905625641346
```

To construct this report, I used Jupyter Notebook, which not only offered an interactive data science environment (IDE), but it also provided a means of presenting the analysis and code written in Python in a descriptive manner, which was very helpful in the reporting process. Jupyter notebook made it very easy to combine graphs, images, and code with the findings, and it was also quite fast. In this project, it was possible to create the report in markdown in conjunction with the code using Jupyter Notebook, which increased the readability and repeatability of the report.

References:

1. Ansari, U.B. and Sarode, T., 2017. Skin cancer detection using image processing. *Int Res J Eng Technol*, *4*(4), pp.2875-2881.

2. S. Albawi, T. A. Mohammed and S. Al-Zawi, "Understanding of a convolutional neural network," 2017 International Conference on Engineering and Technology (ICET), 2017, pp. 1-6, doi: 10.1109/ICEngTechnol.2017.8308186.

3. How to One Hot Encode Sequence Data in Python (machinelearningmastery.com)

4. https://medium.com/@kohlishivam5522/understanding-a-classification-report-for-your-machine-learning-model-88815e2ce397