Configuration Manual

MSc Research Project
Data Analytics

Pooja Rajesh Garje
Student ID: 22241001

School of Computing
National College of Ireland

Supervisor: Prof. Furqan Rustam

National College of Ireland
Project Submission Sheet
School of Computing

Student Name:	Pooja Rajesh Garje
Student ID:	22241001
Programme:	Data Analytics
Year:	2024
Module:	MSc Research Project
Supervisor:	Prof. Furqan Rustam
Submission Due Date:	16/09/2024
Project Title:	Configuration Manual
Word Count:	900
Page Count:	13

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.

ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Pooja Rajesh Garje
Date:	13th September 2024

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST:

Attach a completed copy of this sheet to each project (including multiple copies).	☐
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	☐
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	☐

Assignments that are submitted to the Programme Coordinator office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Configuration Manual

Pooja Rajesh Garje
22241001

1 System Configuration

This project is implemented on a system with 12th Gen Intel(R) Core(TM) i5-1235U processor.It has intel iRIS graphics, 16 GB of RAM, and working on 64-bit system.

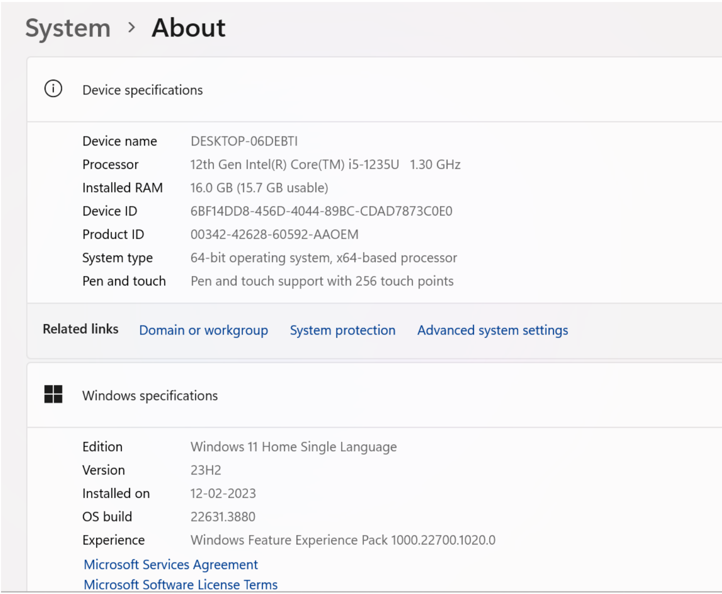


Figure 1: System configuration

2 Software requirement

We used below softwares for the project:

1. Anaconda 3.11.4
2. Python 3.12.1
3. Jupyter Notebook 7.0.8

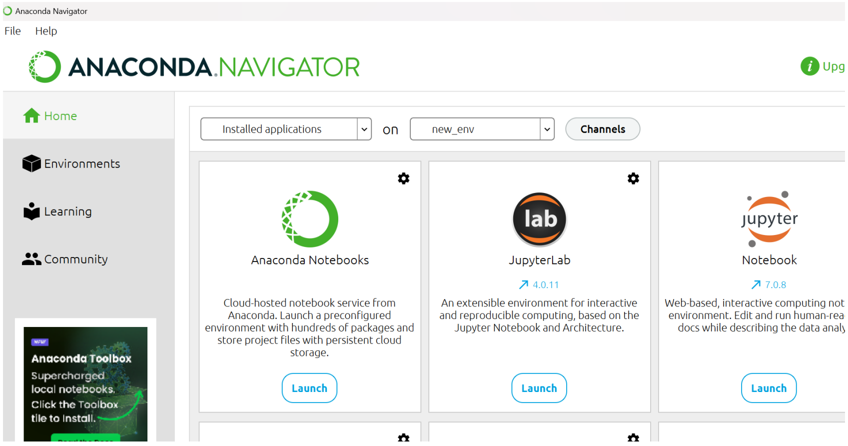


Figure 2: Software Requirement

3 Python Libraries

Here is a list of the main Python packages used in the project:

1. TensorFlow
2. PyTorch
3. Torch Geometric
4. OpenCV
5. Augmentor
6. scikit-learn
7. matplotlib
8. seaborn
9. torchvision
10. scikit-image
11. numpy
12. Pathlib

These packages can be installed using pip in the Jupyter Notebook. This list covers all the major dependencies required for implementing the CNN, GNN, and traditional machine learning models in the project.

4 Dataset Extraction

- The dataset is provided by the ISIC Archive, a well-known repository for dermatological images.It is publicly available on Kaggle.com .
  Original dataset link: https://www.kaggle.com/datasets/nodoubttome/skin-cancer9-classesisic
- There are a total of 2,357 images that include 2,239 pictures for training and 118 ones for testing it offers a vast source in relation to precisely identifying different variations of cancerous tumors affecting the skins.

The dataset was downloaded using the following steps mentioned in Figure 3.The zip file for the dataset is downloaded after executing the below commands. It is stored locally on the system for further model implementation.The path to the folder of the source dataset will change according to the location where the dataset folder is stored.

```
!pip install kaggle

!kaggle datasets download -d nodoubttome/skin-cancer9-classesisic

!tar -xf skin-cancer9-classesisic.zip
```

```
import os
import numpy as np
from skimage import io
from sklearn.model_selection import train_test_split

# paths to train and test folders
train_folder = "C:\\Users\\hp\\Skin cancer ISIC The International Skin Imaging Collaboration\\Train"
test_folder = "C:\\Users\\hp\\Skin cancer ISIC The International Skin Imaging Collaboration\\Test"
```

Figure 3: Dataset extraction

5 Data Pre-processing

The data preprocessing involves resizing the images, normalizing the pixel values, and performing data augmentation.

5.1 Data Augmentation

As shown in Figure 4 augmentation techniques like random horizontal flips and rotations are applied to enhance the dataset.

```
if image_count < 400:
    p = Augmentor.Pipeline(class_dir, output_directory=os.path.join(augmented_dataset_dir, class_name))
    p.rotate(probability=0.7, max_left_rotation=10, max_right_rotation=10)
    p.flip_left_right(probability=0.5)
    p.zoom_random(probability=0.5, percentage_area=0.8)
    p.random_contrast(probability=0.5, min_factor=0.5, max_factor=2.0)
    p.random_brightness(probability=0.5, min_factor=0.5, max_factor=1.5)
    p.random_color(probability=0.5, min_factor=0.5, max_factor=1.5)
    p.random_distortion(probability=0.5, grid_width=4, grid_height=4, magnitude=8)
    num_augmented_images = 400 - image_count
    p.sample(num_augmented_images)
```

Figure 4: Dataset Augmentation

The study was carried out on two separate datasets, the original data (Non-augmented data) and the Augmented data. The distribution of the dataset after augmentation is given in Figure 5.

```
Found 3700 files belonging to 9 classes.
Using 2960 files for training.
Found 3700 files belonging to 9 classes.
Using 740 files for validation.
Image counts in each class after augmentation:
actinic keratosis: 400 images
basal cell carcinoma: 400 images
dermatofibroma: 400 images
melanoma: 438 images
nevus: 400 images
pigmented benign keratosis: 462 images
seborrheic keratosis: 400 images
squamous cell carcinoma: 400 images
vascular lesion: 400 images
Epoch 1/50
```

Figure 5: Data count after augmentation

6 Model Training and Implementation

This study aims at improving the diagnosis of skin cancer through the development of an image segmentation and classification method based on Graph Neural Network (GNN), comparing its results with Convolutional Neural Networks (CNNs) along with traditional machine learning models.The model was first trained on the training and validation data. And only the best model was evaluated on the original test set which comprised of 118 images.The data augmentation step is just performed in the CNN model code and the same augmented data is used by all the models for training. The 22241001.py file in ”code artefacts.zip” folder contains the implementation code for all the six models step-by-step.

Hyperparameter Settings:

- Learning Rate: 0.001
- Batch Size: 32
- Epochs: 25
- Optimizer: Adam
- Loss Function: Cross-Entropy Loss

6.1 Convolutional Neural Network(CNN)

The CNN model is trained using the Adam optimizer and cross-entropy loss. The training loop processes images in batches and updates the weights accordingly.

The model trained on train and validation sets. A detailed classification report and heat map of confusion matrix is printed to analyse the performance of the model for every class.Below figures show the results for CNN model with input images resolution(200,200).

6.1.1 Performance of CNN on Original data and Augmented data

Classification Report (Non-Augmented Dataset):

	precision	recall	f1-score	support
actinic keratosis	0.41	0.32	0.36	22
basal cell carcinoma	0.96	0.89	0.93	84
dermatofibroma	0.87	0.87	0.87	23
melanoma	0.83	0.83	0.83	82
nevus	0.78	0.84	0.81	64
pigmented benign keratosis	0.94	0.92	0.93	83
seborrheic keratosis	0.38	0.53	0.44	17
squamous cell carcinoma	0.88	0.90	0.89	42
vascular lesion	0.97	0.97	0.97	30
accuracy			0.84	447
macro avg	0.78	0.79	0.78	447
weighted avg	0.85	0.84	0.84	447

Figure 6: CNN - Classification report (Original data)

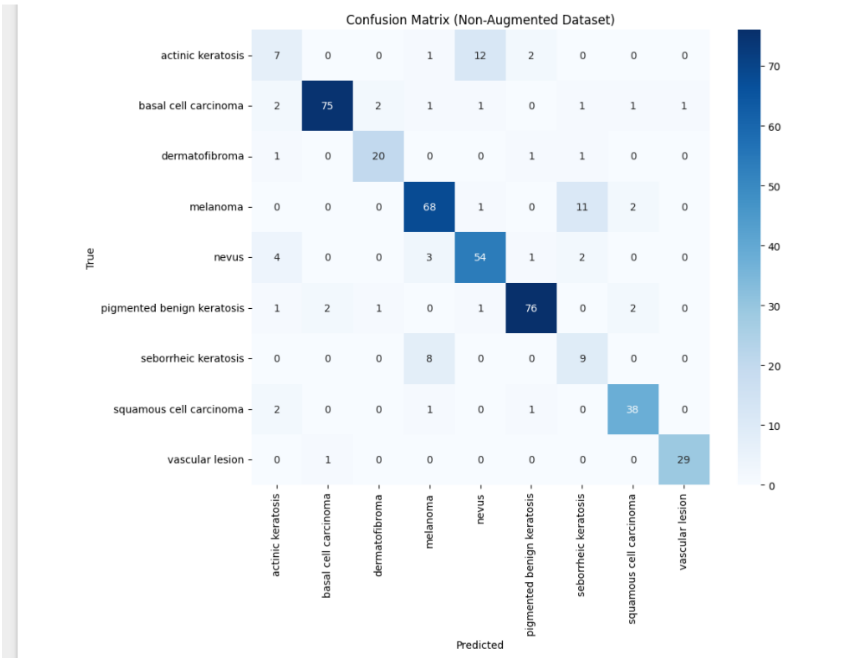


Figure 7: CNN - Confusion matrix (Original data)

Below are the results of CNN model on the augmented data:

6.1.2 Evaluation of CNN

As per the results the best performing model was CNN. The model was finally evaluated on the unseen Test data to validate its reliability.
The Figure 10 is the confusion matrix of the CNN model on test data.

Classification Report (Augmented Dataset):

	precision	recall	f1-score	support
actinic keratosis	0.33	0.36	0.34	75
basal cell carcinoma	0.63	0.55	0.59	87
dermatofibroma	0.70	0.56	0.62	88
melanoma	0.36	0.48	0.41	73
nevus	0.59	0.57	0.58	82
pigmented benign keratosis	0.73	0.66	0.69	101
seborrheic keratosis	0.44	0.59	0.51	75
squamous cell carcinoma	0.46	0.41	0.43	71
vascular lesion	0.89	0.83	0.86	88
accuracy			0.57	740
macro avg	0.57	0.56	0.56	740
weighted avg	0.59	0.57	0.57	740

Figure 8: CNN - Classification report (Augmented data)

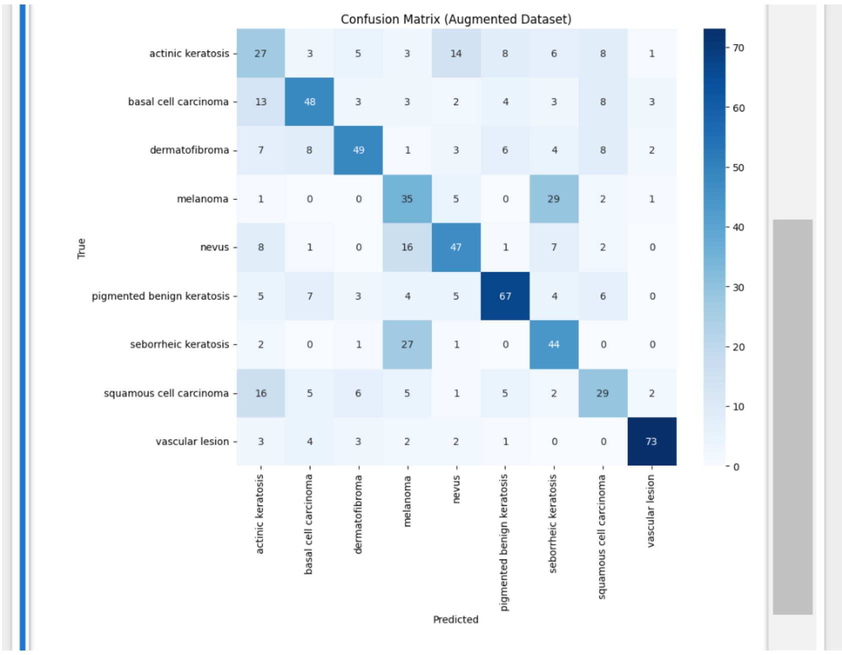


Figure 9: CNN - Confusion matrix (Augmented data)

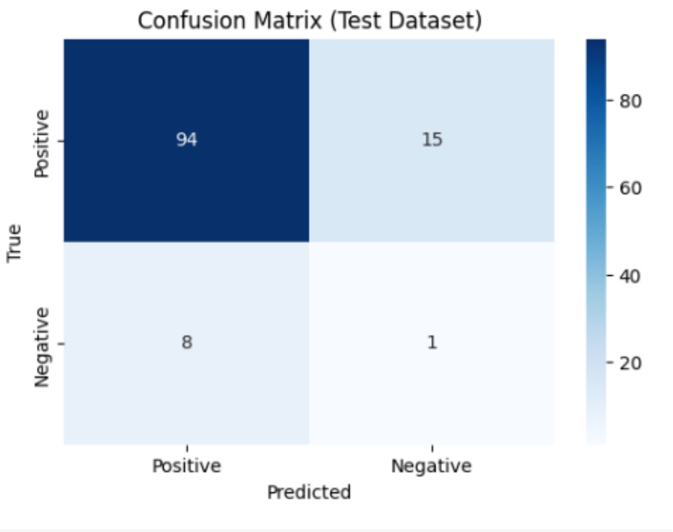


Figure 10: Confusion Matrix

6.2 Graph Neural Network(GNN)

The model trained on train and validation sets. A detailed classification report and heat map of confusion matrix is printed for validation set to analyse the performance of the model for every class. Results for GNN model with input images resolution(200,200).

6.2.1 Performance of GNN on Original and Augmented data

The Figure 11 and Figure 12 show the results of GNN on original data.
The Figure 13 and Figure 14 show the results of GNN on augmented data.

```
Epoch 46/50, Training Accuracy: 0.4846, Training Loss: 1.4158, Validation Accuracy: 0.4710
Epoch 47/50, Training Accuracy: 0.4891, Training Loss: 1.4107, Validation Accuracy: 0.4732
Epoch 48/50, Training Accuracy: 0.4846, Training Loss: 1.4088, Validation Accuracy: 0.4665
Epoch 49/50, Training Accuracy: 0.4863, Training Loss: 1.4081, Validation Accuracy: 0.4754
Epoch 50/50, Training Accuracy: 0.4791, Training Loss: 1.4097, Validation Accuracy: 0.4710

Classification Report:

              precision    recall  f1-score   support

           0       0.33      0.04      0.08        23
           1       0.32      0.79      0.45        75
           2       0.50      0.05      0.10        19
           3       0.59      0.73      0.65        88
           4       0.67      0.34      0.45        71
           5       0.55      0.58      0.57        93
           6       0.00      0.00      0.00        15
           7       0.33      0.03      0.05        36
           8       0.64      0.25      0.36        28

    accuracy                           0.47       448
   macro avg       0.44      0.31      0.30       448
weighted avg       0.49      0.47      0.43       448


Confusion Matrix:
```

Figure 11: GNN - Classification report (Original data)

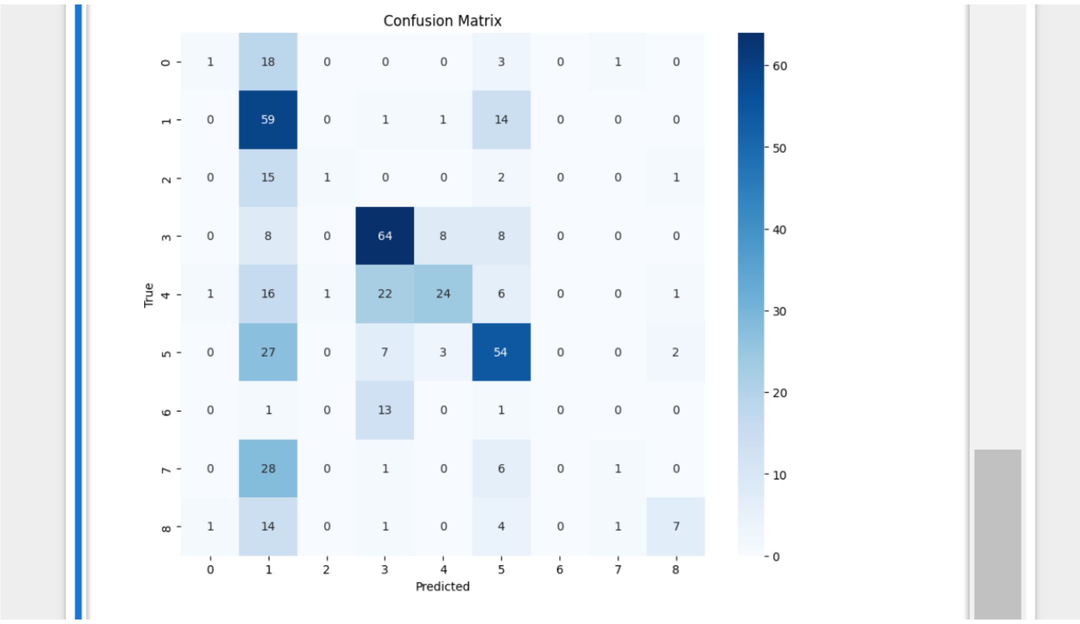
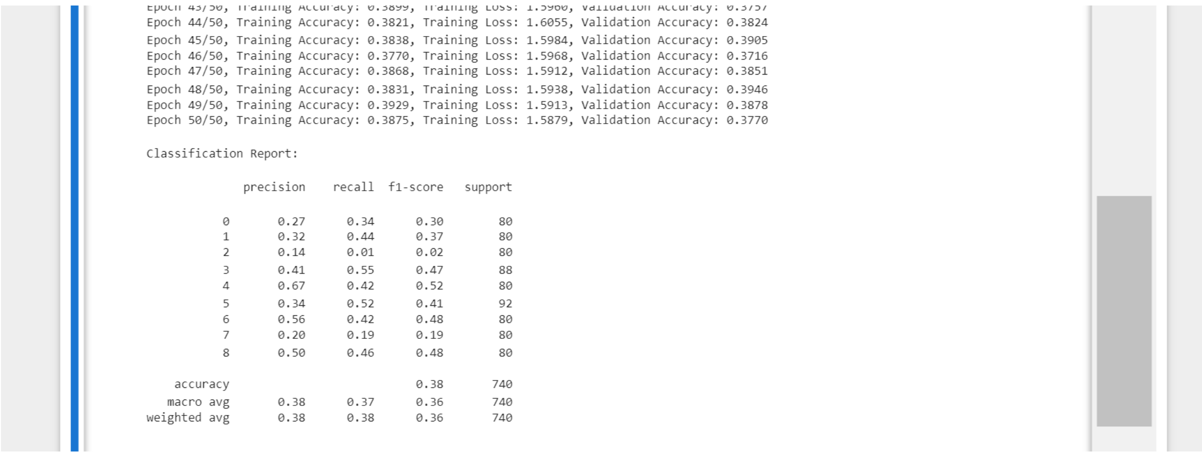


Figure 12: GNN - Confusion matrix (Original data)



```
Epoch 43/50, Training Accuracy: 0.3899, Training Loss: 1.5960, Validation Accuracy: 0.3757
Epoch 44/50, Training Accuracy: 0.3821, Training Loss: 1.6055, Validation Accuracy: 0.3824
Epoch 45/50, Training Accuracy: 0.3838, Training Loss: 1.5984, Validation Accuracy: 0.3905
Epoch 46/50, Training Accuracy: 0.3770, Training Loss: 1.5968, Validation Accuracy: 0.3716
Epoch 47/50, Training Accuracy: 0.3868, Training Loss: 1.5912, Validation Accuracy: 0.3851
Epoch 48/50, Training Accuracy: 0.3831, Training Loss: 1.5938, Validation Accuracy: 0.3946
Epoch 49/50, Training Accuracy: 0.3929, Training Loss: 1.5913, Validation Accuracy: 0.3878
Epoch 50/50, Training Accuracy: 0.3875, Training Loss: 1.5879, Validation Accuracy: 0.3770

Classification Report:

              precision    recall  f1-score   support

           0       0.27      0.34      0.30        80
           1       0.32      0.44      0.37        80
           2       0.14      0.01      0.02        80
           3       0.41      0.55      0.47        88
           4       0.67      0.42      0.52        80
           5       0.34      0.52      0.41        92
           6       0.56      0.42      0.48        80
           7       0.20      0.19      0.19        80
           8       0.50      0.46      0.48        80

    accuracy                           0.38       740
   macro avg       0.38      0.37      0.36       740
weighted avg       0.38      0.38      0.36       740
```

Figure 13: GNN - Classification report (Augmented data)

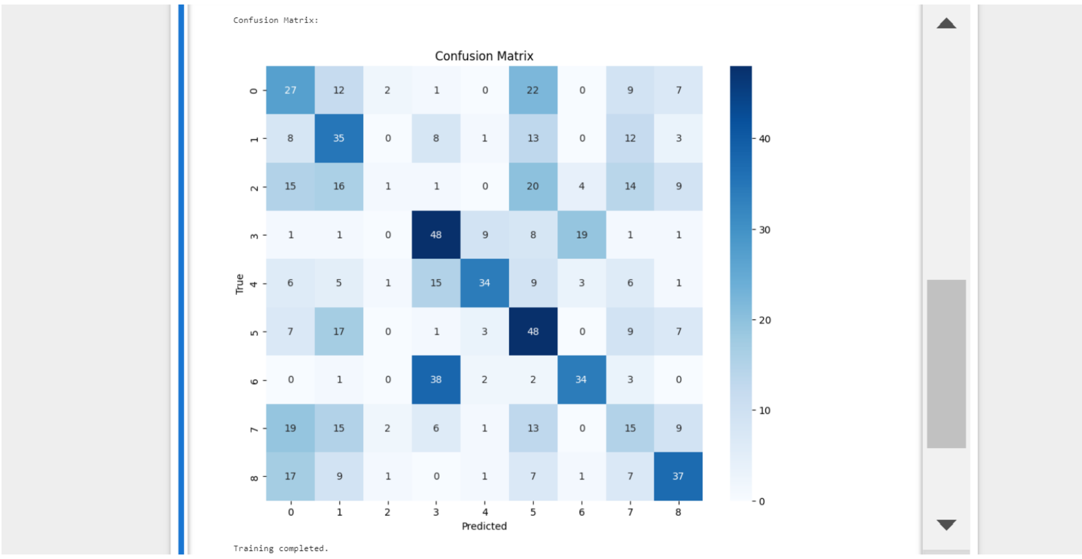


Figure 14: GNN - Confusion matrix (Augmented data)

6.3 Machine Learning Models

The machine learning models were trained on training and validation data. A detailed classification report and heat map of confusion matrix is printed for validation set to analyse the performance of the model for every class.Below figures show the results for machine learning models with input images resolution(200,200).

6.3.1 Logistic Regression

The Figure 15 and Figure 16 show the results for Logistics regression.

Logistic Regression Results:

	precision	recall	f1-score	support
actinic keratosis	0.12	0.15	0.13	20
basal cell carcinoma	0.54	0.48	0.51	93
dermatofibroma	0.15	0.20	0.17	20
melanoma	0.49	0.50	0.49	78
nevus	0.52	0.46	0.49	69
pigmented benign keratosis	0.57	0.51	0.54	96
seborrheic keratosis	0.00	0.00	0.00	18
squamous cell carcinoma	0.24	0.33	0.28	30
vascular lesion	0.55	0.71	0.62	24
accuracy			0.44	448
macro avg	0.35	0.37	0.36	448
weighted avg	0.46	0.44	0.45	448

Figure 15: Logistic Regression (Original data)

Logistic Regression Results:

	precision	recall	f1-score	support
actinic keratosis	0.39	0.24	0.30	96
basal cell carcinoma	0.29	0.29	0.29	82
dermatofibroma	0.44	0.51	0.47	79
melanoma	0.33	0.39	0.36	92
nevus	0.37	0.44	0.41	70
pigmented benign keratosis	0.44	0.36	0.40	90
seborrheic keratosis	0.38	0.38	0.38	78
squamous cell carcinoma	0.20	0.21	0.20	87
vascular lesion	0.53	0.61	0.57	66
accuracy			0.37	740
macro avg	0.38	0.38	0.37	740
weighted avg	0.37	0.37	0.37	740

Figure 16: Logistic Regression(Augmented data)

6.3.2 Random Forest

The Figure 17 and Figure 18 show the results for Random Forest.

Training Random Forest...
Random Forest Results:

	precision	recall	f1-score	support
actinic keratosis	0.00	0.00	0.00	20
basal cell carcinoma	0.52	0.71	0.60	93
dermatofibroma	0.50	0.15	0.23	20
melanoma	0.50	0.53	0.51	78
nevus	0.52	0.54	0.53	69
pigmented benign keratosis	0.55	0.62	0.58	96
seborrheic keratosis	0.00	0.00	0.00	18
squamous cell carcinoma	0.40	0.20	0.27	30
vascular lesion	0.83	0.42	0.56	24
accuracy			0.50	448
macro avg	0.42	0.35	0.36	448
weighted avg	0.48	0.50	0.48	448

Figure 17: Random Forest (Original data)

6.3.3 K-Nearest Neighbor

The Figure 19 and Figure 20 show the results for Random Forest.

Training Random Forest...

Random Forest Results:

	precision	recall	f1-score	support
actinic keratosis	0.41	0.34	0.38	96
basal cell carcinoma	0.37	0.54	0.44	82
dermatofibroma	0.44	0.48	0.46	79
melanoma	0.50	0.43	0.47	92
nevus	0.51	0.56	0.53	70
pigmented benign keratosis	0.41	0.52	0.46	90
seborrheic keratosis	0.46	0.47	0.47	78
squamous cell carcinoma	0.47	0.22	0.30	87
vascular lesion	0.72	0.70	0.71	66
accuracy			0.46	740
macro avg	0.48	0.47	0.47	740
weighted avg	0.47	0.46	0.46	740

Figure 18: Random Forest(Augmented data)

K-Nearest Neighbors Results:

	precision	recall	f1-score	support
actinic keratosis	0.18	0.15	0.16	20
basal cell carcinoma	0.44	0.73	0.55	93
dermatofibroma	0.21	0.20	0.21	20
melanoma	0.53	0.45	0.49	78
nevus	0.52	0.52	0.52	69
pigmented benign keratosis	0.46	0.48	0.47	96
seborrheic keratosis	0.00	0.00	0.00	18
squamous cell carcinoma	0.29	0.07	0.11	30
vascular lesion	0.56	0.42	0.48	24
accuracy			0.46	448
macro avg	0.35	0.33	0.33	448
weighted avg	0.43	0.46	0.43	448

Figure 19: K-Nearest Neighbor (Original data)

Training K-Nearest Neighbors...
K-Nearest Neighbors Results:

	precision	recall	f1-score	support
actinic keratosis	0.32	0.30	0.31	96
basal cell carcinoma	0.27	0.45	0.34	82
dermatofibroma	0.29	0.49	0.37	79
melanoma	0.35	0.21	0.26	92
nevus	0.48	0.50	0.49	70
pigmented benign keratosis	0.38	0.38	0.38	90
seborrheic keratosis	0.34	0.26	0.29	78
squamous cell carcinoma	0.33	0.11	0.17	87
vascular lesion	0.39	0.42	0.41	66
accuracy			0.34	740
macro avg	0.35	0.35	0.33	740
weighted avg	0.35	0.34	0.33	740

Figure 20: K-Nearest Neighbor(Augmented data)

6.3.4 Decision Tree

The Figure 21 and Figure 22 show the results for Random Forest.

Training Decision Tree...
Decision Tree Results:

	precision	recall	f1-score	support
actinic keratosis	0.05	0.10	0.07	20
basal cell carcinoma	0.46	0.34	0.40	93
dermatofibroma	0.08	0.10	0.09	20
melanoma	0.32	0.38	0.35	78
nevus	0.40	0.36	0.38	69
pigmented benign keratosis	0.48	0.44	0.46	96
seborrheic keratosis	0.00	0.00	0.00	18
squamous cell carcinoma	0.18	0.20	0.19	30
vascular lesion	0.38	0.46	0.42	24
accuracy			0.33	448
macro avg	0.26	0.27	0.26	448
weighted avg	0.36	0.33	0.34	448

Figure 21: Decision Tree (Original data)

Training Decision Tree...
Decision Tree Results:

	precision	recall	f1-score	support
actinic keratosis	0.21	0.18	0.19	96
basal cell carcinoma	0.32	0.30	0.31	82
dermatofibroma	0.29	0.34	0.32	79
melanoma	0.26	0.25	0.25	92
nevus	0.35	0.46	0.40	70
pigmented benign keratosis	0.23	0.20	0.21	90
seborrheic keratosis	0.25	0.26	0.25	78
squamous cell carcinoma	0.32	0.28	0.29	87
vascular lesion	0.42	0.45	0.43	66
accuracy			0.29	740
macro avg	0.29	0.30	0.30	740
weighted avg	0.29	0.29	0.29	740

Figure 22: Decision Tree (Augmented data)