Deep Learning Assignment 2023

Laura Gozzo

9 October 2023

1 Exploratory Data Analysis

1.1 Sample Images

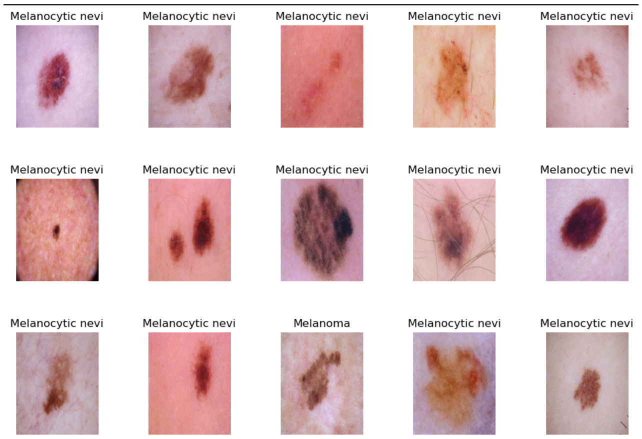


Figure 1: 15 images sampled randomly from the dataset

1.2 Class label distribution

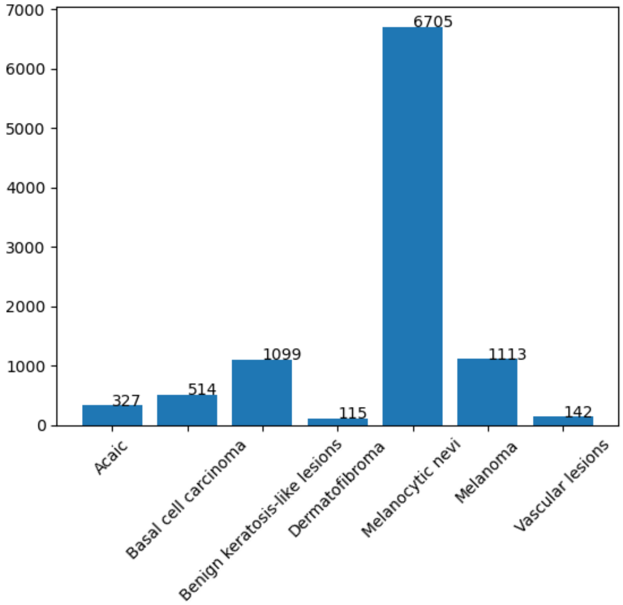


Figure 2: Class label distribution of the dataset

2 Baseline model

2.1 Training and validation losses and accuracies

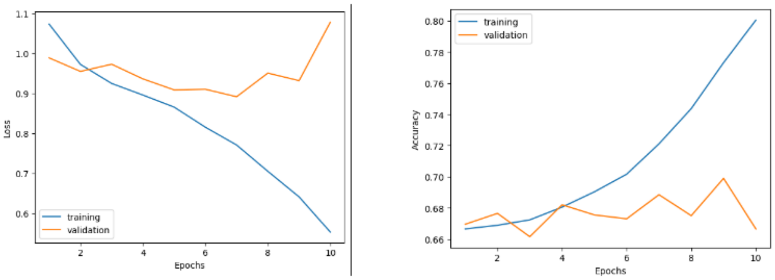


Figure 3: Visualization of history of the training and validation losses and accuracies of the baseline model

2.2 ROC-AUC curves

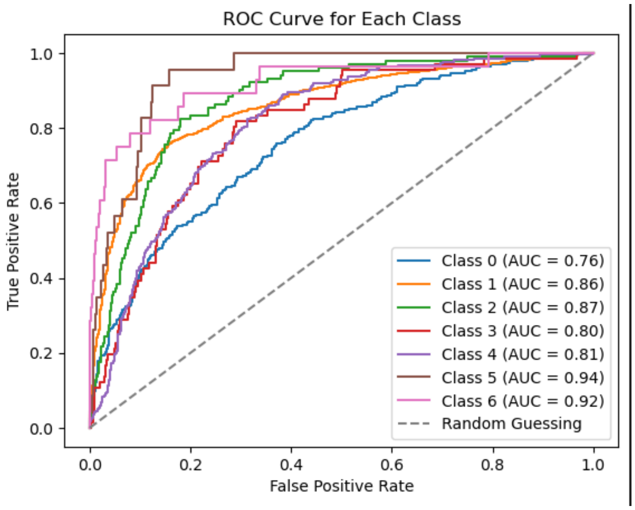


Figure 4: ROC curves and AUC scores of the baseline model (0= Melanoma 1= Melanocytic nevi 2= Basal cell carcinoma 3= Acaic 4= Benign keratosis-like lesions 5= Dermatofibroma 6= Vascular lesions)

2.3 Confusion matrixes

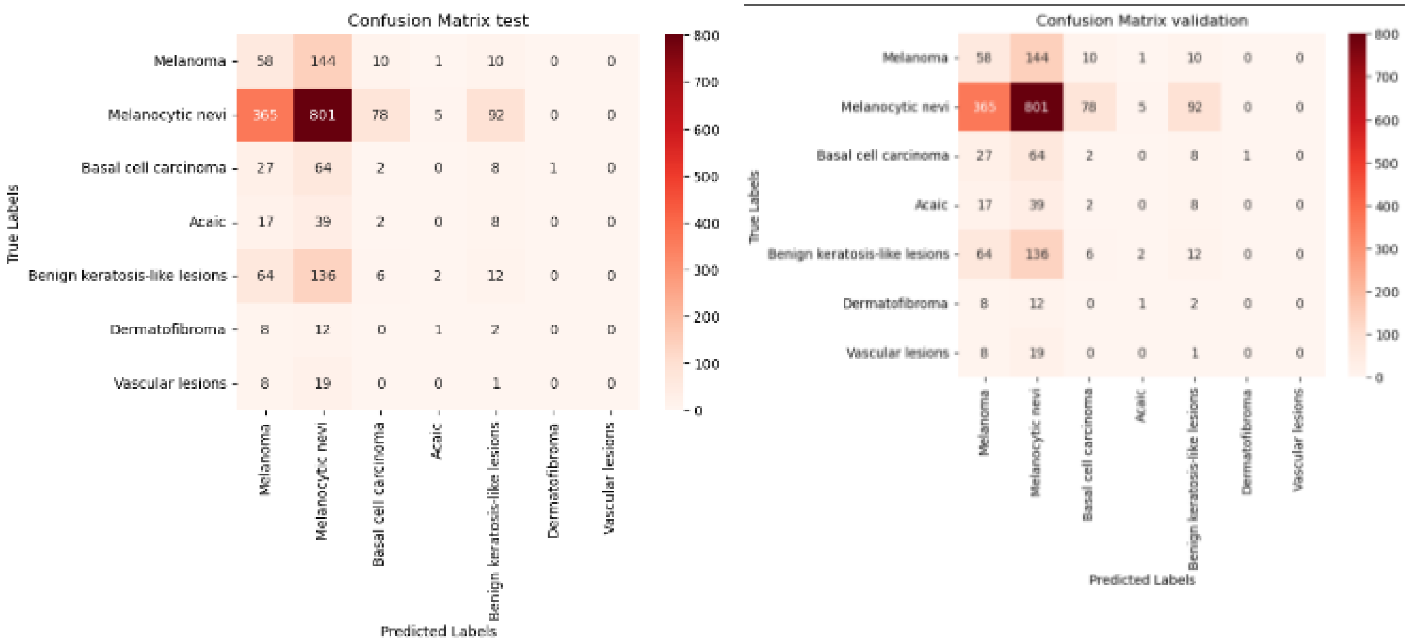
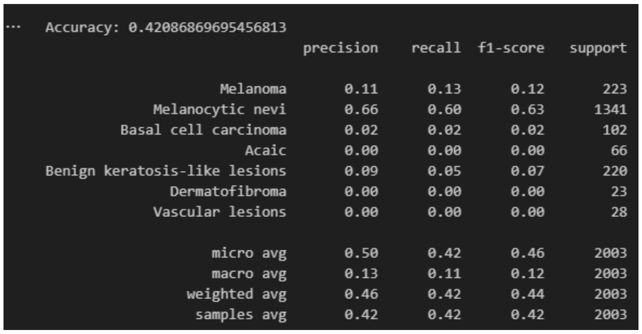


Figure 5: Confusion matrixes of baseline model for validation and test set

2.4 Performance metrics

[?]



```
...  Accuracy: 0.42086869695456813
                               precision    recall  f1-score   support

                     Melanoma       0.11      0.13      0.12       223
             Melanocytic nevi       0.66      0.60      0.63      1341
         Basal cell carcinoma       0.02      0.02      0.02       102
                        Acaic       0.00      0.00      0.00        66
Benign keratosis-like lesions       0.09      0.05      0.07       220
               Dermatofibroma       0.00      0.00      0.00        23
             Vascular lesions       0.00      0.00      0.00        28

                    micro avg       0.50      0.42      0.46      2003
                    macro avg       0.13      0.11      0.12      2003
                 weighted avg       0.46      0.42      0.44      2003
                  samples avg       0.42      0.42      0.42      2003
```

Figure 6: Accuracy (for the test set), Recall, Precision and F1-score for the baseline model

3 Enhanced model

3.1 ROC-AUC curves

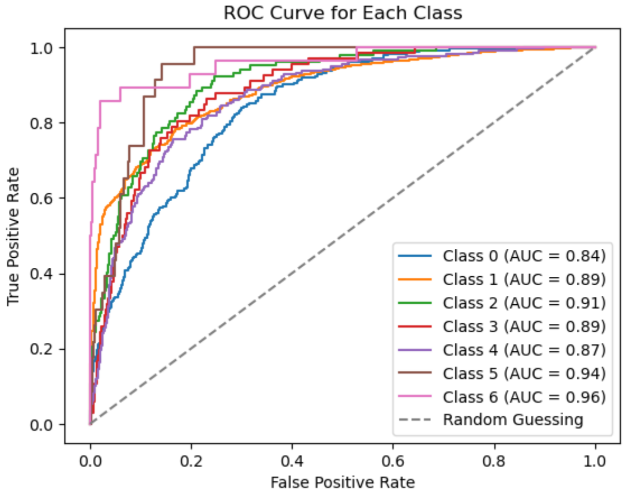
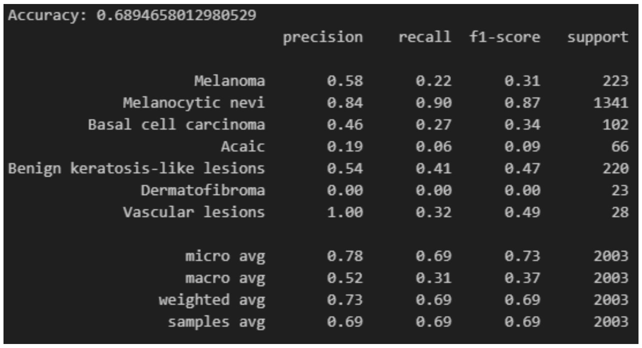


Figure 7: ROC curves and AUC scores of the enhanced model (0= Melanoma 1= Melanocytic nevi 2= Basal cell carcinoma 3= Acaic 4= Benign keratosis-like lesions 5= Dermatofibroma 6= Vascular lesions)

3.2 Performance metrics



```
Accuracy: 0.6894658012980529
                              precision    recall  f1-score   support

                     Melanoma       0.58      0.22      0.31       223
              Melanocytic nevi       0.84      0.90      0.87      1341
         Basal cell carcinoma       0.46      0.27      0.34       102
                        Acaic       0.19      0.06      0.09        66
Benign keratosis-like lesions       0.54      0.41      0.47       220
               Dermatofibroma       0.00      0.00      0.00        23
             Vascular lesions       1.00      0.32      0.49        28

                    micro avg       0.78      0.69      0.73      2003
                    macro avg       0.52      0.31      0.37      2003
                 weighted avg       0.73      0.69      0.69      2003
                  samples avg       0.69      0.69      0.69      2003
```

Figure 8: Accuracy (for the test set), Recall, Precision and F1-score for the enhanced model

4 Discussion about enhanced model

Over all metrics, the enhanced model performed significantly better than the baseline model (Accuracy= 0.69 vs 0.42, F1= 0.37 vs 0.12, Precision= 0.52 vs 0.13, Recall= 0.31 vs 0.11). The enhanced model also got better or equal AUC scores for all classes (0= 0.84 vs 0.76 ,1= 0.89 vs 0.86, 2= 0.91 vs 0.87, 3= 0.89 vs 0.80, 4= 0.87 vs 0.81, 5= 0.94 vs 0.94, 0.96 vs 0.92). One important difference between the models is that in the baseline model, the 3 smallest classes have an F1 score of 0, and all but the 2 biggest classes have F1 scores close to 0 (<1). This means that the model fails to recognise them and tends to classify the vast majority of images as being from the biggest class. The enhanced model partially solves this issue as it has only 1 class with an F1 score of 0 and one other close to 0 (<1); all other classes have a much greater F1 score compared to the baseline model (>30). The following changes have been made to enhance the baseline model.

4.1 Changing filter sizes

Finding the ideal filter size took some trial and error and research to come to the conclusion that for image classification, it is better to increase filter size in the deeper layers. This is because in the deeper layers the representation of the image becomes more complex and bigger filters are more useful to capture this complexity. Filter sizes were changed from (64,32,64,32) to (32,32,64,64).

4.2 Drop out

Dropout increases the performance of the model by training the model with different subsets of data. Drop out trains the model to perform well even when some parts of the network are missing, therefore, it not only increased accuracy but also reduced overfitting (the gap between between training and validation accuracy before introducing drop out was 0.12 vs 0.04). A drop out rate of 0.2 was used after the first pooling layer and a drop out rate of 0.4 was used after the second pooling layer. Because complexity increases in deeper layers they become more prone to overfitting, therefore drop out rate was increased in deeper layers.

4.3 Padding

Setting padding to 'same' improves the model by ensuring that the filter is applied to all elements of the input, it makes sure that no spatial information is lost at the edges of the image.

4.4 Early stopping

The final regularization technique implemented to improve the performance of the model was early stopping. Instead of trying to find the ideal number of epochs through trial and error, early stopping allows you to train the model until a set performance metric (in this case validation accuracy) as been maximised, and would not increase further with more epochs. This made the number of epochs go from 10 to 23.

5 Discussion about possible enhancements

5.1 L2 regularization

A way to improve the performance of the model would have been to implement regularization techniques such as L2 regularization to reduce overfitting. The models discussed above all produce greater accuracy for the training set than for the test and validation sets; this means that they are overfitting. Regularization techniques trade off a higher bias to get a lower variance and therefore, less overfitting. L2 regularization increases bias by introducing a regularization term to the cost function, this causes the weights to decrease towards 0 without dropping to 0 (Van Laarhoven, 2017).

5.2 Fine-tuning hyper-parameters through grid search

Fine-tuning hyper-parameters is a more efficient way to set hyper-parameters than trying out different parameters manually. Grid-search allows to test a range of hyper-parameters for a given layers and outputs the hyper-parameter that produces the best performance for the model (Mvoulana, 2021). This enhances the performance of the model by making sure that the hyper-parameters are as efficient as possible, the reason that I did not implement grid search is that it entails training the model multiple times and therefore requires much more time and computational resources to implement than setting hyper-parameters manually.

5.3 Data augmentation

Data augmentation is a great way to improve a model's performance by 1- providing more data for the model to learn from and 2- decreasing class imbalances (Afzal, 2019). In this case the data is highly unbalanced with 6705 images in the biggest class, 1113 in the second biggest class and only 115 in the smallest class; this causes the models mentioned above to have a strong tendency to classify images as the biggest class and not recognise images from the smallest classes. Therefore, augmenting the data from all classes (except the biggest) could have improved the performance of the models, in particular their ability to correctly identify examples from smaller classes. I did not implement data augmentation because processing enough images to make the dataset balanced, or close to, would have required too much computational resources and time.

6 Transfer learning: using VGG16

6.1 ROC-AUC curves

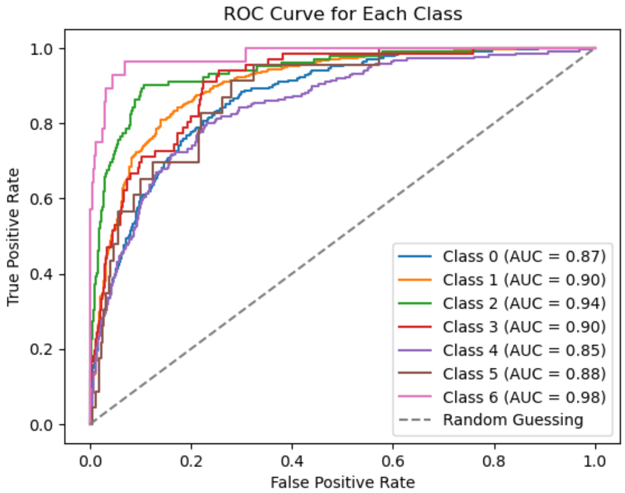


Figure 9: ROC curves and AUC scores of the transfer model (0= Melanoma 1= Melanocytic nevi 2= Basal cell carcinoma 3= Acaic 4= Benign keratosis-like lesions 5= Dermatofibroma 6= Vascular lesions)

6.2 Confusion matrix

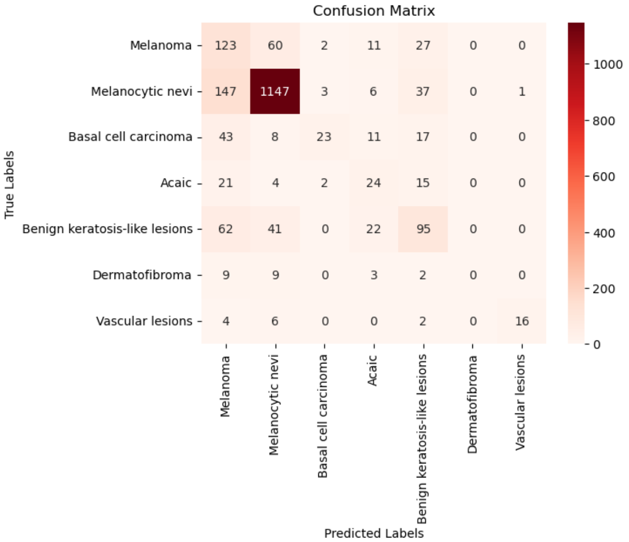
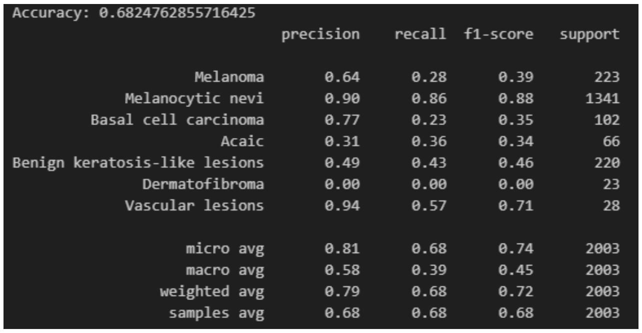


Figure 10: Confusion matrix of transfer model (using test set)

6.3 Performance measures



```
Accuracy: 0.6824762855716425
                               precision    recall  f1-score   support

                     Melanoma       0.64      0.28      0.39       223
             Melanocytic nevi       0.90      0.86      0.88      1341
         Basal cell carcinoma       0.77      0.23      0.35       102
                        Acaic       0.31      0.36      0.34        66
Benign keratosis-like lesions       0.49      0.43      0.46       220
               Dermatofibroma       0.00      0.00      0.00        23
             Vascular lesions       0.94      0.57      0.71        28

                    micro avg       0.81      0.68      0.74      2003
                    macro avg       0.58      0.39      0.45      2003
                 weighted avg       0.79      0.68      0.72      2003
                  samples avg       0.68      0.68      0.68      2003
```

Figure 11: Accuracy (for the test set), Recall, Precision and F1-score for the transfer model

6.4 Discussion on results of transfer learning model

Compared to the baseline model, the model trained using transfer learning produced better results over all metrics (Accuracy = 0.68 vs 0.42, F1 = 0.45 vs 0.12, Recall = 0.39 vs 0.11 and Precision = 0.58 vs 0.13). For all metrics except accuracy it also performed better than the enhanced model (Accuracy = 0.68 vs 0.69, F1 = 0.45 vs 0.37, Recall = 0.39 vs 0.31 and Precision = 0.58 vs 0.52). The transfer model is also better than the enhanced at recognizing smaller classes, only one class has an F1 score of 0 and the next smallest F1 score is 0.34. Transfer models perform particularly well because they are pre-trained and have knowledge from a different task (Tammina, 2019). They disprove the assumption that knowledge about features has to be build from scratch, and they have the advantage of being pre-trained with a big amount of data already (Tammina, 2019).

7 References

Afzal, S., Maqsood, M., Nazir, F., Khan, U., Aadil, F., Awan, K. M., ... Song, O. Y. (2019). A data augmentation-based framework to handle class imbalance problem for Alzheimer's stage detection. IEEE access, 7, 115528-115539.

Mvoulana, A., Kachouri, R., Akil, M. (2021, January). Fine-tuning Convolutional Neural Networks: a comprehensive guide and benchmark analysis for Glaucoma Screening. In 2020 25th International Conference on Pattern Recognition (ICPR) (pp. 6120-6127). IEEE.

Tammina, S. (2019). Transfer learning using VGG-16 with Deep Convolutional Neural Network for Classifying Images. International Journal of Scientific and Research Publications, 9(10), p9420. https://doi.org/10.29322/ijsrp.9.10.2019.p9420

Van Laarhoven, T. (2017). L2 regularization versus batch and weight normalization. arXiv preprint arXiv:1706.05350.