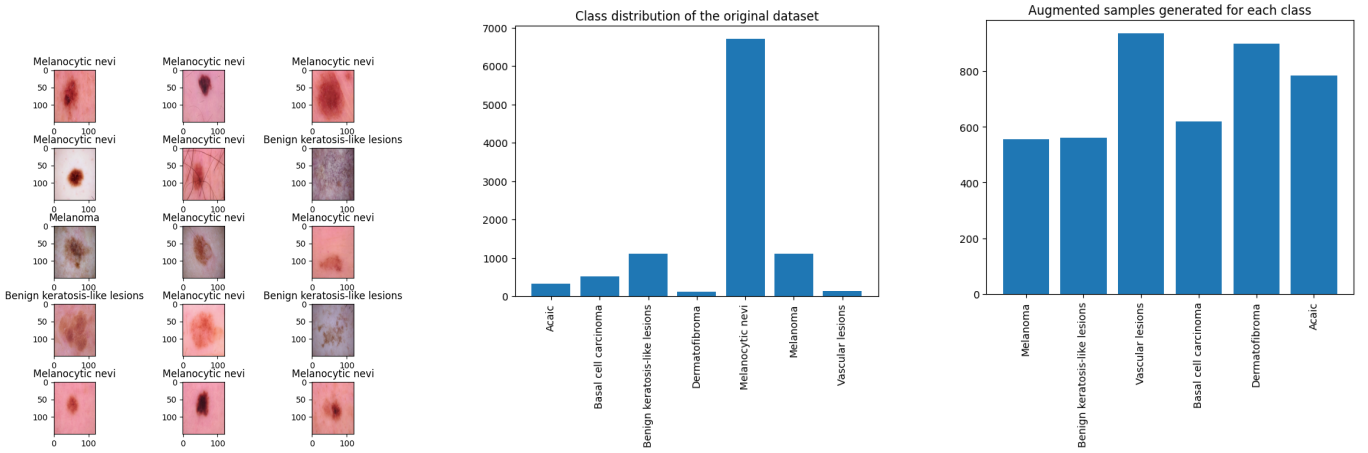
1 LOADING DATASET AND PREPROCESSING

The project that will be described and discussed in the following was about the task of classifying various forms of skin abnormalities and lesions based on their pictures. In order to complete the task, computer vision techniques were used on the Skin Cancer dataset openly provided by the International Skin Imaging Collaboration [1, 3, 9]. The dataset consists of 10015 images of different skin abnormalities and lesions belonging to seven different classes (Actinic keratoses and intraepithelial carcinoma (Acaic), Basal cell carcinoma, Benign keratosis-like lesions, Dermatofibroma, Melanocytic nevi, Melanoma, and Vascular lesions). We started by resizing the pictures from the original dimension of 450x600 size to a smaller size of 120x150 to reduce memory and computational workload. After that, we began with the preprocessing of the data. We normalized the image data and split the data into three sets to get a training (60%), validation (20%) and test (20%) set.

2 VISUALIZING SAMPLE IMAGES AND CLASS LABEL DISTRIBUTION

After the first preprocessing steps, we performed some exploratory data analysis by visualizing a random sample of 15 images of skin lesions from the dataset with their corresponding labels. The sample images can be seen in Figure 1a. Additionally, to get a better overview over the contents of the dataset and the distribution of classes, we created a bar plot. Each bar represents one of the seven classes of skin lesions. As can be observed in Figure 1b, the distribution of classes is highly unbalanced. Nearly 70% of all samples show pictures of the Melanocytic nevi lesion. This imbalanced data distribution is important to keep in mind for the later evaluation and improvement of the model.



(a) 15 Random Images from the Dataset

(b) The Original Class Distribution

(c) The Augmented Distribution of the Minority Classes

Figure 1: Image Samples (a) and Class Distributions of the Original (b) and Augmented (c) Classes

3 BASELINE MODEL

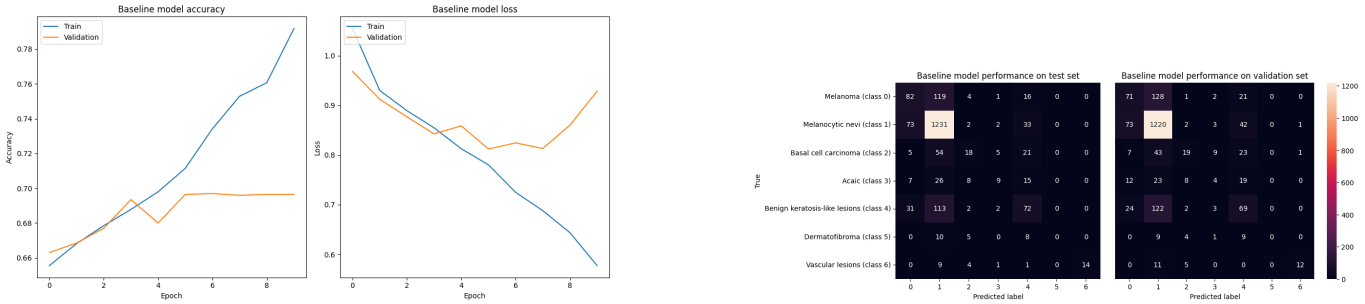
The baseline Convolutional Neural Network (CNN) to perform a first attempt on classifying the images was constructed as specified in the assignment's task description. In order to maintain the shape of the input, padding was set to *same*. Since the task at hand is a multiclass classification problem, a proper activation function for the output layer of the CNN is softmax, as it calculates a probability for all classes and the classification is made according to the highest probability. A fitting loss function for multiclass classification tasks is categorical crossentropy, which we chose when compiling our model.

Class	Precision	Recall	F1-Score
Melanoma (0)	0.45	0.23	0.30
Melanocytic nevi (c1)	0.80	0.93	0.86
Basal cell carcinoma (2)	0.28	0.20	0.24
Acaic (3)	0.24	0.26	0.25
Benign keratosis-like lesions (4)	0.41	0.31	0.35
Dermatofibroma (5)	0.00	0.00	0.00
Vascular lesions (6)	0.43	0.31	0.36
Accuracy			0.70
Macro Avg	0.37	0.32	0.34
Weighted Avg	0.66	0.70	0.67

Figure 2: Classification Report for the Baseline Model

The development of the model's classification accuracy and loss over the epochs is portrayed in Figure 3a. It shows that the model is overfitting to the training data after approximately six epochs. However, accuracy is not a reliable measure of performance for multiclass classification problems based on imbalanced data. Therefore, we also consulted other measures and created a confusion matrix for the classification performance on the test and validation set (Figure 3b) and plotted the Receiver Operating Characteristic (ROC) curve with their corresponding Area Under Curve (AUC) values for all classes (Figure 4a). We additionally got a classification report on the test set with the precision, recall, and f1-score for all classes (Figure 2). Most importantly, it can be seen that the performance of the baseline model is by far best on the majority class

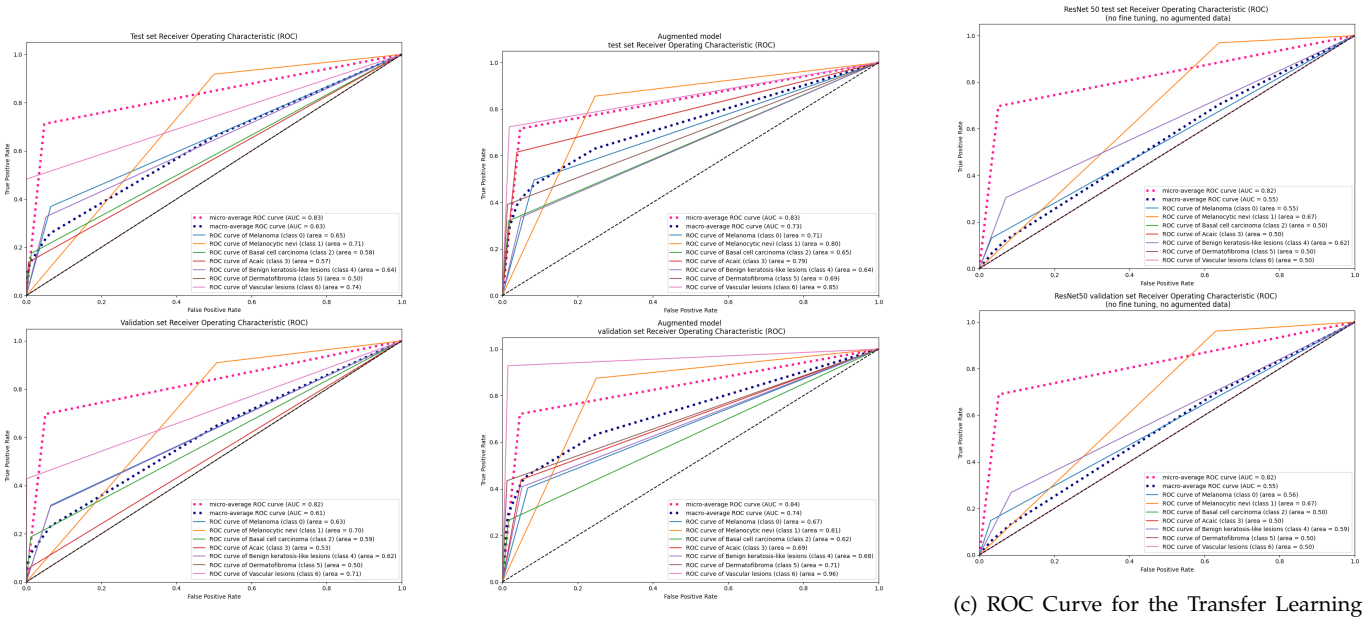
Melanocytic nevi. This is most likely due to the amount of training data that is available for this class but not for the minority ones, and underlines the problem of imbalanced data for (multiclass) classification tasks. The model seems to have learned the characteristic features of the majority class rather well, as many datapoints were available, while its performance is poor for the underrepresented classes.



(a) Accuracy and Loss for the Baseline Model

(b) Confusion Matrix for the Baseline Model

Figure 3: Metrics for the Baseline Model



(a) ROC Curve for the Baseline Model

(b) ROC Curve for the Enhanced Model

(c) ROC Curve for the Transfer Learning Model (no fine-tuning, no augmented data)

Figure 4: The ROC Curves for the Baseline, Enhanced, and Transfer Model

4 ENHANCED MODEL

Accordingly, we had to make some changes to our model to enhance its overall performance.

4.1 *Data Augmentation and Training*

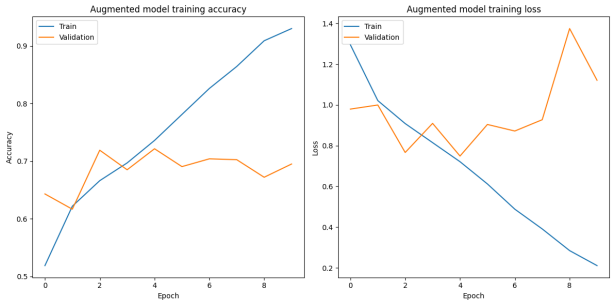
In order to balance the dataset better and get more training examples for the minority classes from the original dataset, we used the keras built in ImageDataGenerator. We augmented all classes despite the majority class Melanocytic nevi, to bring them up to approximately 1000 samples. We thought about upsampling all minority classes to around 2000 examples so that they would at least be half the size of the majority class, however hardware and memory limitations led us to use 1000 samples only and observe tendencies in the model's performance change. Augmentation was done by applying rotation of up to 40 degrees, width and height shifts of +- 20%, shearing by an angle of +-20 degrees, zooming in an out by a factor of +-0.2, and horizontal flipping. Missing pixels caused by any of the previous actions were filled in with the values of the nearest other pixels. A class distribution of the minority samples after augmentation can be seen in Figure 1c. The total number of training images after data augmentation was 10358.

In addition to augmenting the data, we also made some changes to the model itself. A common problem when using the ReLU activation function in CNNs is the dying ReLU effect, where the model outputs zero for all inputs [2]. When testing for dead ReLUs in the baseline model for 20 random samples from the validation set, we found 47-92% of the ReLU activations to be zero. Following the literature [4, 7], we decided to use leaky ReLu instead of the

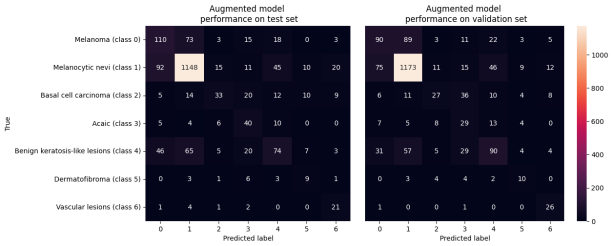
previous ReLu activation function and apply He weight initialization to mitigate the problem. When testing again for dead ReLUs with the enhanced model, no zero activations could be found.

Furthermore, the issue of the internal covariate shift is known to negatively influence the training of CNNs and other deep neural networks. As suggested in the literature [5], we therefore made use of batch normalization to improve our model's performance.

A last change we made to the baseline model was adding more training epochs. To let the model make use of the newly augmented data, we increased the number of epochs to 100. However, the development of the loss function in the baseline model was already suggesting that the model overfits to the training data with only ten epochs of training. To deal with this problem, we fitted our model with early stopping. After five epochs without a performance improvement on the validation set, the training would be stopped and the previously best performing weights would be chosen.



(a) Accuracy and Loss for the Enhanced Model



(b) Confusion Matrix for the Enhanced Model

Figure 5: Metrics for the Enhanced Model

4.2 *Performance Evaluation*

Class	Precision	Recall	F1-Score
Melanoma (0)	0.42↓	0.50↑	0.46↑
Melanocytic nevi (c1)	0.88↑	0.86↓	0.87↑
Basal cell carcinoma (2)	0.52↑	0.32↑	0.40↑
Acaic (3)	0.35↑	0.62↑	0.45↑
Benign keratosis-like lesions (4)	0.46↑	0.34↑	0.39↑
Dermatofibroma (5)	0.25↑	0.39↑	0.31↑
Vascular lesions (6)	0.37↓	0.72↑	0.49↑
Accuracy			0.72↑
Macro Avg	0.46↑	0.53↑	0.48↑
Weighted Avg	0.73↑	0.72↑	0.72↑

Note: ↑ : improved results, ↓ : decreased performance.

Figure 6: Classification Report for the Enhanced Model

For the performance evaluation of the enhanced model, we consulted the same metrics that were seen earlier for the evaluation of the baseline model. The results can be found in Figures 5, 4b, and 6.

Similar to the development in the baseline model, the enhanced model seems to be overfitting to the training data already after six epochs of training, which is indicated by the rise in loss at that point that can be seen in Figure 5a.

Nevertheless, as indicated by the arrows in the classification report shown in Figure 6, the values for almost all metrics of previously underrepresented classes had improved. The performance on the Melanocytic nevi class however was slightly worse than in the baseline model, which could be due to the fact that no augmented data was generated for this class. Less diverse data for the Melanocytic nevi class than for others might have negatively influenced the model's performance. Creating augmented data for the Melanocytic nevi class as well might solve this issue, however, we refrained from trying this due to the memory limitations that were mentioned earlier.

5 TRANSFER LEARNING MODEL

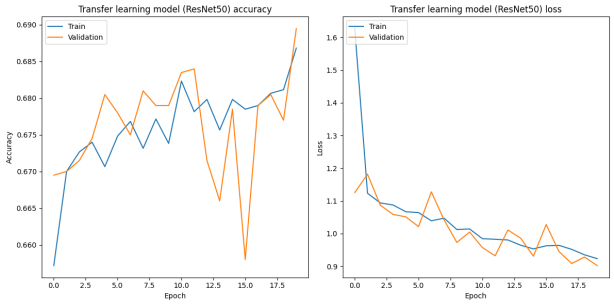
Transfer learning can be a powerful tool for training CNNs in computer vision tasks, especially when relatively few datapoints are available. With this method one can take advantage of models that are trained on large image corpora like ImageNet. For the task at hand we decided to use the ResNet50 network with ImageNet weights on the original (non-augmented) data. First, we tried to add two dense layers with 32 neurons each and an output layer with seven nodes as suggested for the baseline model, but observed very poor performance. Therefore, after searching for inspiration in the literature [8], we decided to add two dense layers with 4096 nodes each and ReLu activation function, plus an output layer with seven neurons and a softmax activation function. We specified 20 learning epochs but included early stopping with a patience of five epochs after what we had observed earlier in our model. With these changes the transfer learning model performed better, however, the training and validation loss and accuracy that are shown in Figure 8a did still not reach satisfactory

Class	Precision	Recall	F1-Score
Melanoma (0)	0.33↓	0.13↓	0.19↓
Melanocytic nevi (0)	0.75↓	0.97↑	0.85↓
Basal cell carcinoma (2)	0.0↓	0.0↓	0.0↓
Acaic (3)	0.0↓	0.0↓	0.0↓
Benign keratosis-like lesions (4)	0.35↓	0.30↓	0.32↓
Dermatofibroma (5)	0.0	0.0	0.0
Vascular lesions (6)	0.0↓	0.0↓	0.0↓
Accuracy			0.70↓
Macro Avg	0.20↓	0.20↓	0.19↓
Weighted Avg	0.58↓	0.70↓	0.62↓

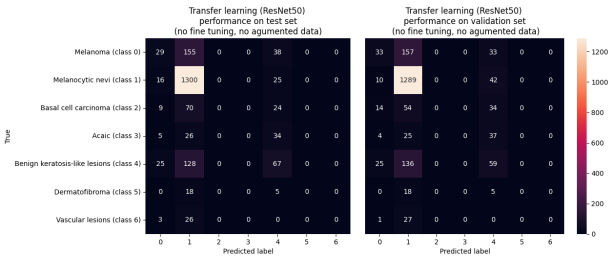
Note: ↑ : decreased performance, ↓ : decreased performance.

Figure 7: Classification Report for the Transfer Learning Model (no fine-tuning, no augmented data)

values. The corresponding confusion matrices, ROC curves, and the classification report metrics (Figures 8b, 4c, and 7) show that this transfer model even performed worse than the baseline model.



(a) Accuracy and Loss for the Transfer Learning Model (no fine-tuning, no augmented data)



(b) Confusion Matrix for the Transfer Learning Model (no fine-tuning, no augmented data)

Figure 8: Metrics for the Transfer Learning Model (no fine-tuning, no augmented data)

After consulting the literature [6, 8, 10], we concluded that these outcomes might be caused by either the limited number and size of the dense layers we constructed on top of the ResNet50 architecture, or by the fact that we did not unfreeze any of the last layers of the original network. While the first layers of the ResNet50 network like any other CNN for computer vision are trained to extract the rather basic image features, like edges, the late layers capture more detail that can be very specific to the classes in the training data [10]. Given the fact that ImageNet mainly contains pictures of animals, food and everyday objects, it is unlikely that the part of ResNet50 extracting the more high level features of these objects would generalize well to the skin images in our task. Therefore, we decided to implement a fine-tuning strategy instead of trying to add an even deeper dense neural network to the complete architecture of ResNet50, following a rationale provided in the literature [6, 8, 10].

Class	Precision	Recall	F1-Score
Melanoma (0)	0.64↑	0.43↓	0.52↑
Melanocytic nevi (0)	0.89↑	0.94↑	0.92↑
Basal cell carcinoma (2)	0.66↑	0.69↑	0.67↑
Acaic (3)	0.49↑	0.52↓	0.51↑
Benign keratosis-like lesions (4)	0.65↑	0.64↑	0.64↑
Dermatofibroma (5)	0.40↑	0.26↓	0.32↑
Vascular lesions (6)	0.96↑	0.79↑	0.87↑
Accuracy			0.82↑
Macro Avg	0.67↑	0.61↑	0.63↑
Weighted Avg	0.81↑	0.82↑	0.81↑

Note: ↑ : decreased performance, ↓ : decreased performance.

Figure 9: Classification Report for the Transfer Learning Model (with fine-tuning and augmented data)

Instead of keeping the whole ResNet50 network as it was, we decided to unfreeze the last three layers of the network and train the whole network on the augmented data instead of on the original. Additionally, we added layers of the same structure as in the previous transfer learning model on top of the ResNet50 architecture. Due to the previously mentioned limitations of Google Colab, we trained this modified version of the model in a different environment, using an Nvida A40 GPU combined with the Intel Xeon Gold 6248R CPU. As to not change the pre-trained weights of the unfrozen layers too drastically, we trained our model with an Adam optimizer with a learning rate of 1e-5 instead of the default rate of 0.001. These modifications changed the transfer model's performance significantly and nearly all the performance metrics that can be seen in Figure 9 show better results than our own enhanced model when trained on the augmented data. In the appendix, Figures 10 and 11 can be found, displaying the accuracy and loss functions, confusion matrices, and ROC curves for this version of the transfer model as well.

6 CONCLUSION AND DISCUSSION

For the described project of classifying various skin abnormalities and lesions into seven different classes, we used a CNN as a computer vision architecture. After creating a baseline model, and analysing its outcomes, we concluded that the model's performance mainly suffered from the great class imbalance in the dataset. In order to build a better model, we therefore conducted data augmentation on the minority classes and additionally made use of He weight initialization, batch normalization, a leaky ReLu activation function, and regularization through early stopping techniques. Performance improved especially on the minority classes and we suggest that in a setting with less memory limitations and more data, the model could perform quite well overall. Moreover, we explored a transfer learning approach using the ResNet50 network and adding three more layers to it. The model performed very poorly on the original data when all original weights were frozen also in the last layers of the architecture. We therefore applied some fine-tuning measures to the model and unfroze the last three layers before our own added layers. When carefully trained on the augmented data, this version of a transfer model outperformed our own enhanced model, and as before, we expect it to perform even better if provided with more data. Additionally, the overall performance of all models discussed here could be further improved if more fine-tuning of the hyperparameters would be done (dropout, L1/L2 regularization, different numbers of neurons and layers).

REFERENCES

[1] "Skin cancer dataset:," https://challenge.isic-archive.com/data/#2018, accessed: 2023-09-17.

[2] A. F. Agarap, "Deep learning using rectified linear units (relu)," *arXiv preprint arXiv:1803.08375*, 2018.

[3] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti *et al.*, "Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (isic)," *arXiv preprint arXiv:1902.03368*, 2019.

[4] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1026–1034.

[5] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *International conference on machine learning*. pmlr, 2015, pp. 448–456.

[6] G. Litjens, T. Kooi, B. E. Bejnordi, A. A. A. Setio, F. Ciompi, M. Ghafoorian, J. A. Van Der Laak, B. Van Ginneken, and C. I. Sánchez, "A survey on deep learning in medical image analysis," *Medical image analysis*, vol. 42, pp. 60–88, 2017.

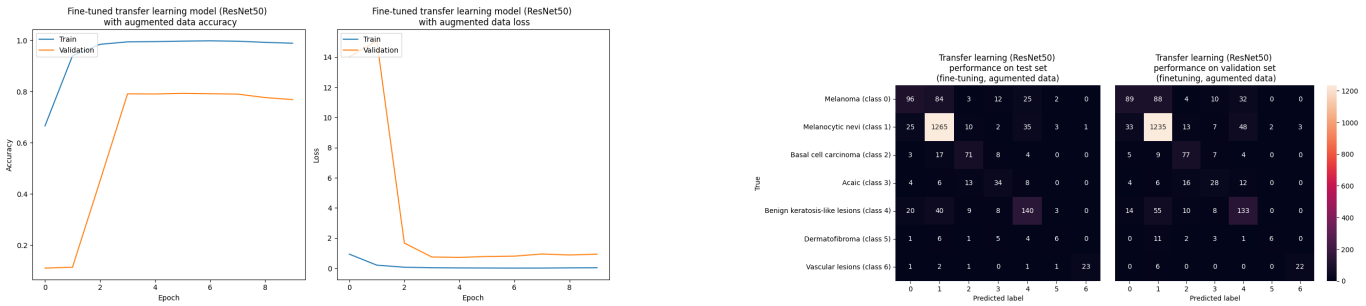
[7] A. L. Maas, A. Y. Hannun, A. Y. Ng *et al.*, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. icml*, vol. 30, no. 1. Atlanta, GA, 2013, p. 3.

[8] N. Tajbakhsh, J. Y. Shin, S. R. Gurudu, R. T. Hurst, C. B. Kendall, M. B. Gotway, and J. Liang, "Convolutional neural networks for medical image analysis: Full training or fine tuning?" *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1299–1312, 2016.

[9] P. Tschandl, C. Rosendahl, and H. Kittler, "The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions," *Scientific data*, vol. 5, no. 1, pp. 1–9, 2018.

[10] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson, "How transferable are features in deep neural networks?" *Advances in neural information processing systems*, vol. 27, 2014.

APPENDIX A



(a) Accuracy and Loss for the Transfer Learning Model (with fine-tuning and augmented data)

(b) Confusion Matrix for the Transfer Learning Model (with fine-tuning and augmented data)

Figure 10: Metrics for the Transfer Learning Model (with fine-tuning and augmented data)

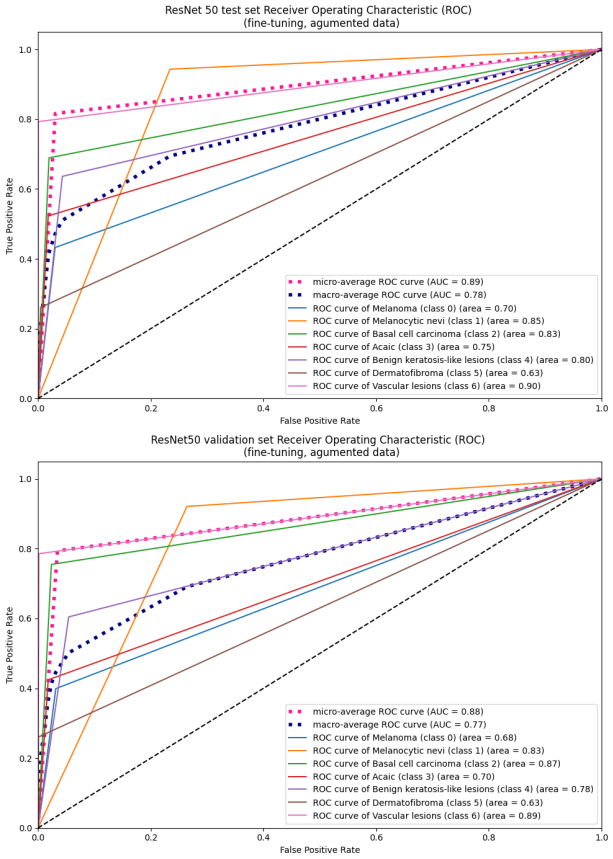


Figure 11: ROC curve for the Transfer Learning Model (with fine-tuning and augmented data)