Skin Lesion Image Classification for Melanoma Detection

Zurich University of Applied Sciences
School of Engineering

Authors	Gabriel Gillmann, Nathan Hess
Supervisors	Norman Juchler, Stefan Glüge
Study Program	DSHEAL
Submitted on	10.04.2025

1 Introduction

Skin cancer is among the most prevalent cancers worldwide, and melanoma represents its most aggressive form, accounting for a majority of skin cancer-related deaths despite being less common than non-melanoma types. The prognosis for melanoma patients heavily depends on early and accurate detection, which makes rapid and reliable diagnostic tools essential in clinical practice [1].

In recent years, deep learning techniques, particularly convolutional neural networks (CNNs), have revolutionized image classification tasks, including medical diagnostics. CNN architectures such as ResNet [2], VGG [3], and EfficientNet [4] have demonstrated high performance in various image recognition benchmarks like ImageNet. Their ability to automatically learn and extract hierarchical image features has made them especially effective in classifying dermoscopic images for skin lesion analysis.

Transfer learning has further amplified the utility of deep CNNs in medical image classification. By fine-tuning pre-trained models on domain-specific datasets, researchers can leverage general image features while adapting to the particular nuances of medical imagery, even when labeled data is limited [5]. This has enabled models like ResNet-101 to achieve impressive results in tasks such as melanoma classification.

The goal of this project is to explore the efficacy of various pre-trained CNN models—including ResNet-50, ResNet-101, VGG16, and EfficientNet-B0—in classifying dermoscopic images from the ISIC 2019 dataset into eight categories of skin lesions. The dataset presents several challenges, notably class imbalance and subtle visual distinctions between lesion types. We ultimately selected ResNet-101 for further fine-tuning based on its superior initial performance in terms of accuracy, precision, and F1-score. The dermoscopic images used in this study are sourced from the ISIC 2019: Skin Lesion Analysis Towards Melanoma Detection dataset, available on Kaggle. The dataset contains a total of 25,331 images labeled into nine diagnostic categories: melanoma, melanocytic nevus, basal cell carcinoma, actinic keratosis, benign keratosis (including solar lentigo, seborrheic keratosis, and lichen planus-like keratosis), dermatofibroma, vascular lesion, squamous cell carcinoma, and “none of the above.” It combines data from previous ISIC challenges (2017 and 2018) and serves as a comprehensive benchmark for automated skin lesion classification[6].

This report presents the methodology, training process, and experimental results for multi-class skin lesion classification, and concludes with a discussion of the outcomes and limitations of the approach.

2 Methods

We implemented all of our models with pytorch in a jupyter notebook [7]. We experimented with four different pre-trained convolutional models during this project. Resnet-50 [8] , Resnet-101 [8], VGG16 [9] and EfficentNet-b0 [10]. We had access to an RTX 4070, a powerful NVIDIA GPU [11], which meant we could use CUDA [12], and could train deep and computationally demanding models such as Resnet-101 quickly. To evaluate which model was best suited for our use-case, we trained each model on our data using the following hyperparameters:

From there we evaluated which model had the best initial performance before fine-tuning

Table 2.1: Initial Hyperparameters

Learning Rate	Batch Size	Optimizer	Epochs
10^{-4}	32	Adam	10

it more. In our case, Resnet-101 had the highest initial accuracy, precision and F1-score. Therefore, all other models were abandoned, and we only continued fine-tuning the Resnet-101 model.

2.1 Resnet-101

The ResNet-101 model has 101 layers. ResNet-101 uses residual connections, which allow the model to learn identity mappings and pass information directly across layers. This helps gradients flow more easily during backpropagation, making deep networks easier to train. The architecture of ResNet-101 consists of an initial convolutional layer and max pooling, followed by four groups of residual blocks. These blocks use a bottleneck design with three layers each: a 1x1 convolution to reduce dimensionality, a 3x3 convolution, and another 1x1 convolution to restore dimensionality. The network includes 3, 4, 23, and 3 of these bottleneck blocks in its four stages, respectively. Despite its depth, it remains computationally efficient due to this bottleneck structure. ResNet-101 has around 44.5 million parameters and achieves high accuracy on image classification tasks such as ImageNet. Its success lies in the residual connections, which make it possible to train very deep networks without the typical problems of vanishing gradients or overfitting [8]. The architecture of Resnet-101 is shown in figure 2.1 [13]

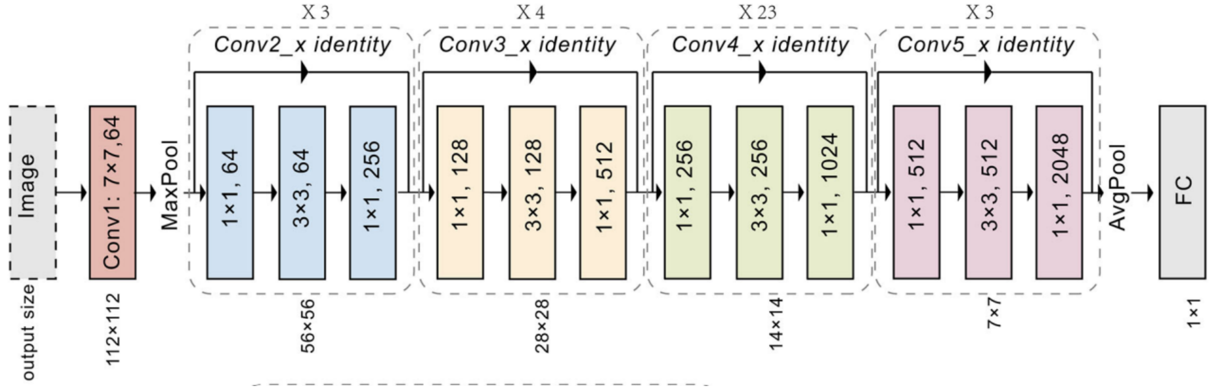

Figure 2.1: Resnet-101 Architecture

2.2 Preprocessing

As mentioned in the Introduction, the dataset used in this project was medium-sized with 25'000 images that included metadata. First, we removed the class "UNK", or "Unknown" and all images of this class from the dataset. The total amount of images was very low, and this group represented several different types of skin cancer, which made it hard for the model to find common features. As we used only pre-trained models in this project, we resized the images to fit the input specifications of our models: Resnet-50, Resnet-101, VGG16 and EfficientNet-b0 all require an input size of 224x224 pixels [8]. In addition, the images were normalised to ensure the images matched the pre-trained expectations of the model. We tested several techniques of image augmentation: Adding noise, subtracting parts of the images and even rotation, but we found that none of these techniques improved model performance, and thus no such techniques were used. Furthermore, we changed the encoding from one-hot encoding to label encoding.

2.3 Model Training

We used 80% of our data to train the model and we used the remaining 20% to evaluate the results. As described in the Model selection chapter, we left the model completely pre-trained at first, only letting it update the final, fully connected layer during training. WeHowever we found that model accuracy could be improved significantly by letting the model update more parameters during training, specifically the weights in the convolutional layers 3 and 4. These layers are responsible for capturing high-level features of the image, and we wanted this part of the model to be trained specifically for our task. We did not implement early stopping, as the model still performed best when training for 10 epochs. We also reduced the learning rate to 10^{-5}

3 Results

The performance of the ResNet-101 model was evaluated on the test set using standard classification metrics: precision, recall, F1-score, and overall accuracy. The classification report, summarized in Table 3.1, indicates that the model achieved a test accuracy of 0.798. The weighted average F1-score was 0.80.

Table 3.1: Overall Performance Metrics

Metric	**Value**
Accuracy	0.7983
Macro Avg F1	0.68
Weighted Avg F1	0.80

Table 3.2: Classification Report for ResNet-101

Class (Label)	**Precision**	**Recall**	**F1-Score**	**Support**
0, Melanoma	0.76	0.65	0.70	915
1, Melanocytic nevus	0.88	0.91	0.89	2567
2, Basal cell carcinoma	0.83	0.81	0.82	658
3, Actinic keratosis	0.54	0.28	0.37	177
4, Benign keratosis	0.65	0.70	0.67	536
5, Dermatofibroma	0.61	0.73	0.67	48
6, Vascular lesion	0.92	0.79	0.85	62
7, Squamous cell carcinoma	0.34	0.59	0.43	104

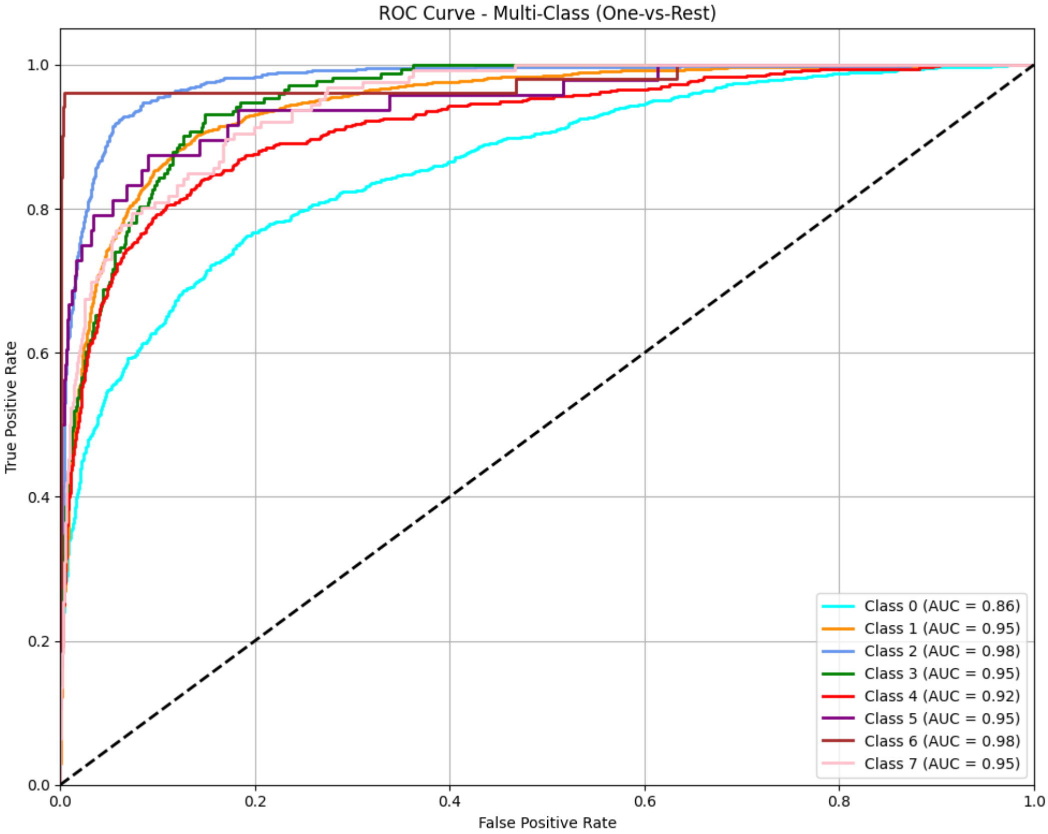

Figure 3.1: ROC Curve for Multi-Class Classification (One-vs-Rest). Most classes show strong AUC values, indicating effective class separability despite imbalance.

As shown in Table 3.2, the results reveal strong classification performance in the most represented classes, particularly class 1, which had the highest support and achieved an F1-score of 0.89. Similarly, class 2 also performed well with an F1-score of 0.82.

In contrast, underrepresented classes such as class 3 and class 7 exhibited significantly lower F1-scores (0.37 and 0.43 respectively), indicating the model's difficulty in accurately classifying these categories. While class 6 had a small support size (62 images), it still achieved a high F1-score of 0.85, suggesting that its visual features may be more distinct or easier to learn.

Despite using a high-performing model and fine-tuning strategies, class imbalance remained a persistent challenge. Attempts to apply class weighting or image augmentation during training did not significantly improve performance. In particular, weighting the classes inversely proportional to their frequency in the dataset led to slight improvements in some of the poorly performing classes, but overall it caused a drop in the model's general performance, suggesting a trade-off between optimizing for minority classes and maintaining strong results across the board.

The ROC curves in Figure 3.1 provide a complementary view of the model's performance across all classes. Most classes achieved high area under the curve (AUC) values, with class 2 and class 6 reaching an AUC of 0.98, indicating excellent separability. Even lower-performing classes in terms of F1-score, such as class 3 and class 7, showed strong ROC results with AUCs of 0.95, suggesting that while the model struggles with precise predictions on these classes, it still maintains good discriminatory capability. This highlights the importance of using multiple evaluation metrics to fully understand model behavior in imbalanced multi-class settings.

4 Conclusion

The results of our classification project might not be perfect. Classification tasks such as this one are challenging, as all images show lesions which are in some way abnormal, and cancerous. Models such as Resnet-101 can easily reach accuracies of 0.95 or upwards when dealing with binary cancer classfication [1].

The accuracy we achieved with this task was lower than what other researchers achieved using the ISIC 2019 dataset for the same task. For example Kassem, Hosny and Fouad reached an accuracy of 94.92% and a F1 score of 80.07%. The true problem with our model is the fact that we did not find a way to deal with class imbalance. Adding weights did not improve model performance. We did not apply oversampling techniques, where we would copy images of the underrepresented classes and augment them using image augmentation techniques. This might be worth looking into to deal with the class imbalance inherent in the dataset.

Bibliography

[1] W. Gouda, N. U. Sama, G. Al-Waakid, M. Humayun, and N. Z. Jhanjhi, "Detection of skin cancer based on skin lesion images using deep learning," *Healthcare*, vol. 10, no. 7, 2022, ISSN: 2227-9032. DOI: 10.3390/healthcare10071183. [Online]. Available: https://www.mdpi.com/2227-9032/10/7/1183.

[2] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *arXiv preprint arXiv:1512.03385*, 2015. [Online]. Available: https://arxiv.org/abs/1512.03385.

[3] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014. [Online]. Available: https://arxiv.org/abs/1409.1556.

[4] M. Tan and Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks," *arXiv preprint arXiv:1905.11946*, 2020. [Online]. Available: https://arxiv.org/abs/1905.11946.

[5] M. A. Kassem, K. M. Hosny, and M. M. Fouad, "Skin lesions classification into eight classes for isic 2019 using deep convolutional neural network and transfer learning," *IEEE Access*, vol. 8, pp. 114 822–114 832, 2020. DOI: 10.1109/ACCESS.2020.3003890. [Online]. Available: https://doi.org/10.1109/ACCESS.2020.3003890.

[6] Larxel, *Skin lesion images for melanoma classification*, https://www.kaggle.com/datasets/andrewmvd/isic-2019, Accessed: 2025-04-10, 2019.

[7] G. G. Nathan Hess, *Dsheal skinmodel*, https://github.com/DeHess/DSHEAL-SkinModel, 2025.

[8] K. He, X. Zhang, S. Ren, and J. Sun, *Deep residual learning for image recognition*, 2015. arXiv: 1512.03385 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1512.03385.

[9] K. Simonyan and A. Zisserman, *Very deep convolutional networks for large-scale image recognition*, 2015. arXiv: 1409.1556 [cs.CV]. [Online]. Available: https://arxiv.org/abs/1409.1556.

[10] M. Tan and Q. V. Le, *Efficientnet: Rethinking model scaling for convolutional neural networks*, 2020. arXiv: 1905.11946 [cs.LG]. [Online]. Available: https://arxiv.org/abs/1905.11946.

[11] NVIDIA Corporation, *Nvidia geforce rtx 4070 family*, Accessed: 2025-04-10, 2025. [Online]. Available: https://www.nvidia.com/de-de/geforce/graphics-cards/40-series/rtx-4070-family/.

[12] NVIDIA Corporation, *Cuda toolkit*, Accessed: 2025-04-10, 2025. [Online]. Available: https://developer.nvidia.com/cuda-toolkit.

[13] Y. Tong, W. Lu, Q.-q. Deng, C. Chen, and Y. Shen, “Automated identification of retinopathy of prematurity by image-based deep learning,” *Eye and Vision*, vol. 7, p. 40, Aug. 2020. DOI: 10.1186/s40662-020-00206-2.