Convolutional Neural Network Based Classification of Melanoma Images

David Krieger, Manuel Plank

August 2020

1 Introduction

Malignant melanoma is an aggressive skin cancer that stems from an out-of-control growth of pigment-producing cells known as melanocytes. Even though it is the least common skin tumor, it has a high mortality rate and accounts for the majority of skin cancer-related deaths across the globe (Schadendorf et al. 2018). Early detection significantly increases the chances of long-term tumor containment or even complete removal via surgery before the cancer cells are capable of spreading. In diagnosis, dermatologists usually inspect a patient's skin and apply the "ABCDE" rule that evaluates lesions on the parameters Asymmetry, Border irregularity, Color variation, Diameter, and Evolving over time. Towards that end, dermoscopy is used which is a non-invasive imaging technique that delivers a zoomed representation of the skin's surface and its deeper layers. The ability of machine learning to assist in this kind of biomedical imaging task and the potential for a classifier to serve as a computer-aided detection (CAD) system has been subject to research and controversial debate alike. While experienced dermatologists are still the gold standard when it comes to diagnosing melanoma, Brinker et al. 2019 report for the first time that a classifier performs on par with dermatologists from German universities in terms of sensitivity. In this landmark study, they also report on the high variance of physicians on performance metrics due to their wide range in practical experience. In contrast to that, an algorithm is not subjective and produces robust results demonstrating that CADs could support dermatological work and lead to more consistent diagnosis.

The interest in applying computer vision for medical imaging has grown in recent years because of the advances in data availability, not only in the domain of medical imaging but also for object recognition tasks in general. Large-scale annotated datasets such as ImageNet (Deng et al. 2009), which contains over 14 million images with objects from over 20,000 categories were essential in training very deep neural networks that pushed the performance on object recognition benchmark datasets. In this project, we exploit a convolutional neural network (CNN) that has been pre-trained on the ImageNet synset and adapt it's architecture to the task of melanoma classification. The remainder of this technical report is outlined as follows: In section 2, we will give a brief overview over state-of-the-art approaches in medical image recognition. Section 3 describes the data used in this project and the preprocessing

steps that are part of our input pipeline. Here, re-sampling strategies and image augmentation techniques are discussed as well. Section 4 deals with the transfer learning approach to the classification problem at hand, whereas section 5 discusses the results. The last section concludes.

2 Related Work

Most of the previous work on computer-based melanoma classification deals primarily with methods to extract features from dermoscopy images. In that fashion, Kusumoputro and Ariyanto 1998, inspired by the "ABCDE" rule, extract 18 cancer shape and color features from dermoscopy images already in the late 90's. They then separate malignant melanoma from benign moles using neural networks and obtain 91.8% accuracy. More recently, Xie et al. 2016 compute a bigger set of color, texture and border features, including statistics on RGB values and the convex hull of moles resulting in a total of 57 features. After dimensionality reduction by means of a principal components analysis (PCA), they feed the features into an ensemble of neural networks and report 95% sensitivity, 93.75% specificity and 94.17% accuracy.[1]

The ever more refined way to extract features to improve performance that can be observed over time in the literature is challenged by another approach. As algorithms that are able to process images as inputs and automatize feature extraction began to gain more attention, researchers began to apply that approach to the domain of medical imaging. Now, models with millions of parameters could be trained and re-used for other tasks at a cheap cost as large-scale image datasets became available. Shin et al. 2016 evaluated CNN performance on two medical image tasks, namely thoraco-abdominal lymph node detection and interstitial lung disease classification, to infer the potential for models trained on ImageNet to serve as CAD in the medical domain. Interestingly, although the ImageNet dataset does not contain any data remotely similar to lymph nodes or lung images, they achieve state-of-the-art performance in those tasks. Because of their extensive empirical evaluation of different architectures, e.g. AlexNet and GoogleLeNet, and different training strategies, their work is most relevant to our project.

3 Datasets and Preprocessing Steps

The Society for Imaging Informatics in Medicine (SIIM) and the International Skin Imaging Collaboration (ISIC) hosted the data used in this project on the data science platform Kaggle.com as part of a competition. It consists of 33,126 images of benign lesions and malignant melanoma from over 2,000 patients of the following institutions: Hospital Clínic de Barcelona, Medical University of Vienna, Memorial Sloan Kettering Cancer Center, Melanoma Institute Australia, University of Queensland, and the University of Athens Medical School. At the time of writing, a publication describing the data generation process is forthcoming. But

[1]Reported results have to be read with caution as they are not evaluated on the same dataset and are often based on questionable sample sizes, especially for positive observations.

the ISIC guarantees that all diagnoses have been confirmed either by expert agreement or histopathology.

The data consists of 32,542 negative cases and 584 positive ones. Thus, we are dealing with a fairly unbalanced dataset. Highly imbalanced data adds difficulty to the task of accurate diagnosis since learning algorithms tend to reveal bias towards the majority class while underweighting the minority class. Consequently, samples belonging to the minority group tend to be misclassified more often than those belonging to the majority group. Furthermore, class imbalance can mislead the researcher when interpreting evaluation metrics. Considering a binary dataset with a positive to negative class ratio of 95% to 5%, the most straightforward naïve learning approach would yield 95% accuracy. Especially in medical domains, misclassifications of the minority group can have drastic consequences to an individuals' health and life expectancy. The minority of samples mostly represents the disease that is to be detected by the learner. False-negative classifications of such instances can prevent medical treatments for subjects that suffer from the disease but are not positively diagnosed (Johnson and T. Khoshgoftaar 2019).

In this project, the class imbalance problem is addressed on the algorithm level. Algorithm level methods for handling class imbalance intervene in the learning process by assigning more importance to the minority class. Approaches such as class-weighted learning implement this concept by allocating different penalties to different prediction errors during the model fitting process. In our case, the prediction errors occurring in the samples are weighted with the inverse of the class distribution and then backpropagated to adjust the weights of the neural net. This concept can consequently be considered a form of resampling, penalizing false-positive and false-negative classifications differently (Ling and Sheng 2008).

Apart from handling class imbalance, an image classification tool must deal with dermoscopic variances, including different viewpoints and lighting. To make an image classifier more robust and generalizable, the training data can be transformed and upsampled so that it encompasses slight variations of the original data taking dermoscopic variances into account (Shorten and T. M. Khoshgoftaar 2019). We integrated geometric and color space augmentation techniques such as random flipping and random change of hue, saturation, contrast, and brightness, on model level. As a result, the augmentations were performed in real time, implementing new image variations for each fold of the cross validated model training. Data augmentation must be implemented carefully, taking into account the primary goal of producing "new" slightly varying images that resemble the test images. Applying unrealistic transformations can decrease the model's performance drastically (Shorten and T. M. Khoshgoftaar 2019).

In the datasets, each patient and each image is uniquely identified with an ID. Besides the images themselves, there is meta-information on the image level available. A patient's age and sex and the location of the imaged site on the patient's body are likely to be relevant features.[2] Thus, in the input pipeline we merge image data stored in the jpg-format with metadata stored in CSV files as visualized in figure 1.

Before merging, we apply the preprocessing steps. We re-scale images to a 256x256 pixel resolution using inter area interpolation. We do so to guarantee fixed image dimensions and

[2]At least in the USA, melanoma occurs 1.6 times more often in males than in females. https://gis.cdc.gov/Cancer/USCS/DataViz.html

because models based on ImageNet are designed on these exact dimensions. We leave RGB channels untouched to be aligned with ImageNet pre-trained models resulting in images shaped 256x256x3. For 1.6% of the training images we observe missing values in the age feature, the other features exhibit a lower percentage of missing values. To solve this issue, we impute the age mean and the sex mode. Missing values in the categorical variable location of the imaged site are labeled as such. To prepare the data for machine learning algorithms we further standardize the age feature using first and second moments of the training sample. Additionally, we one-hot encode sex and location of the imaged site.

Overall, the data provided by ISIC is larger than 100GB, which poses practical problems of data loading as the data size exceeds RAM capacities. In our implementation, we serialize the data and store it in a set of files where each file is small enough to fit into memory. The resulting TFRecords allow for efficient streaming over the data with the advantage that when data needs to be deserialized and transformed, it can be distributed across multiple cores to do so as the images are independent of each other. Distribution is also possible when it comes to the training of machine learning models. One common technique is to use synchronous distributed training where different cores train over different sets of data in parallel and aggregate gradients at each step. For that, we connect to Google-owned Tensor Processing Units (TPUs) that implement synchronous distributed training on hardware specifically designed for deep learning.

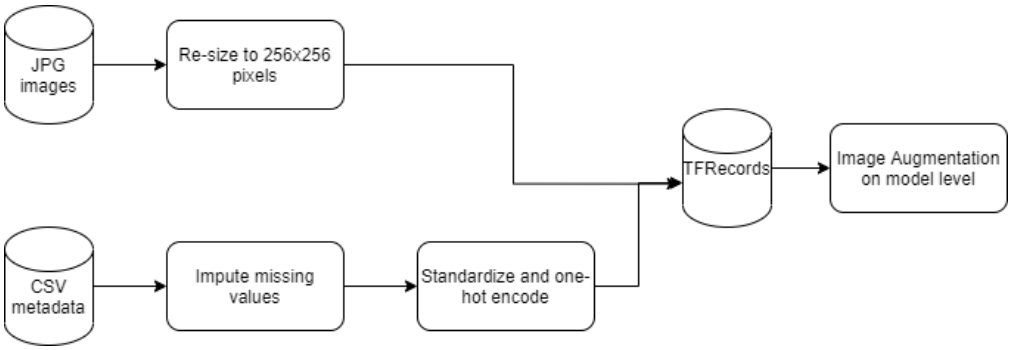

Figure 1: Input pipeline

4 Convolutional Neural Network and Transfer Learning

We employ a novel CNN from the *EfficientNet* model family (Tan and Le 2019) for melanoma detection. Often, model architectures choices are made arbitrarily, resulting in sub-optimal performance. However, experiments by Tan and Le 2019 showed that when a baseline model is scaled up in terms of hyperparameters width, depth, and resolution with a constant ratio, accurate and efficient results can be obtained. The most complex model presented in their paper, *EfficientNet-B7*, achieves the best performance on the benchmark dataset ImageNet with 84.4% accuracy when only the most probable class serves as prediction and 97.1% accuracy when top-5 classes are considered.

Given that the dataset does not contain enough data to train those kinds of models from scratch and the risk of overfitting is high, we opt for a transfer learning approach.[3] Towards that end, we chose the EfficientNet-B5 model guided by experimentation with all models of that family. The weights of the pre-trained models are publicly available and we make use of them to start the optimization problem close to the optimal solution or at least closer than with random initialization of parameters. Shin et al. 2016 report state-of-the-art performance using transfer learning of ImageNet trained models on two datasets in the medical domain. They further discuss two learning strategies and find that fine-tuning of the entire network leads to consistent improvement in performance while using a pre-trained model in an off-the-shelf fashion without learning is less performing.

The model architecture is visualized in figure 2. It consists of the EfficientNet-B5 model architecture described in detail in Tan and Le 2019. The fully connected layer at the top of the network is removed. Instead, we add a global max pooling 2D layer before the feature maps of the last convolution are vectorized. Fully connected layers often lead to overfitting which is tackled by the last pooling operation. The output is then concatenated with the features of the metadata, such as sex and age. Before finally feeding the input into the last fully connected layer with random parameters, a dropout layer is inserted. The dropout layer randomly sets a certain number of input neurons to zero (in our case 20% of the input units) to additionally avoid overfitting.

The described model is trained in mini-batches of 64 images and with a binary cross entropy loss function to penalize those cases having predicted probability and actual class label diverging strongly. At the beginning of model creation, we observed a performance convergence after several epochs. Consequently, we reduced the number of epochs from 15 to 8. As an optimizer, we implemented Adaptive Moment Estimation (Adam). According to Ruder 2016's extensive overview of gradient descent algorithms, Adam slightly outperforms other optimizers in similar visual recognition tasks. The last hyperparameter is the learning rate, which we defined by a base learning rate and decay factor. The base learning rate is set at 0.00003 already relatively low because we are careful not to lose previous knowledge in the network. A decay that reduces the learning rate once learning stagnates was chosen to prevent the overshoot of areas of low loss. All hyperparameter tuning and model selection choices were based on results from 5-fold cross validation. Generally, we found that tuning these parameters is challenging, especially with deep networks such as modern CNNs.

5 Discussion of Results

For the Kaggle submission, the predictions resulting from each of the five cross validation steps were averaged and then handed in. This Kaggle competition used the area under the curve (AUC) as evaluation metric on two different parts of the test set. The *public* AUC score is available to monitor the model performance during the ongoing competition. In contrast, the *private* AUC score is calculated after the submission deadline, serving as final performance score to determine the rank in the competition leaderboard. **The public AUC score of our submitted model was 0.8173, and the private AUC score was**

[3] In fact, the EfficientNet-B5 model that we are using has 28,515,689 parameters.

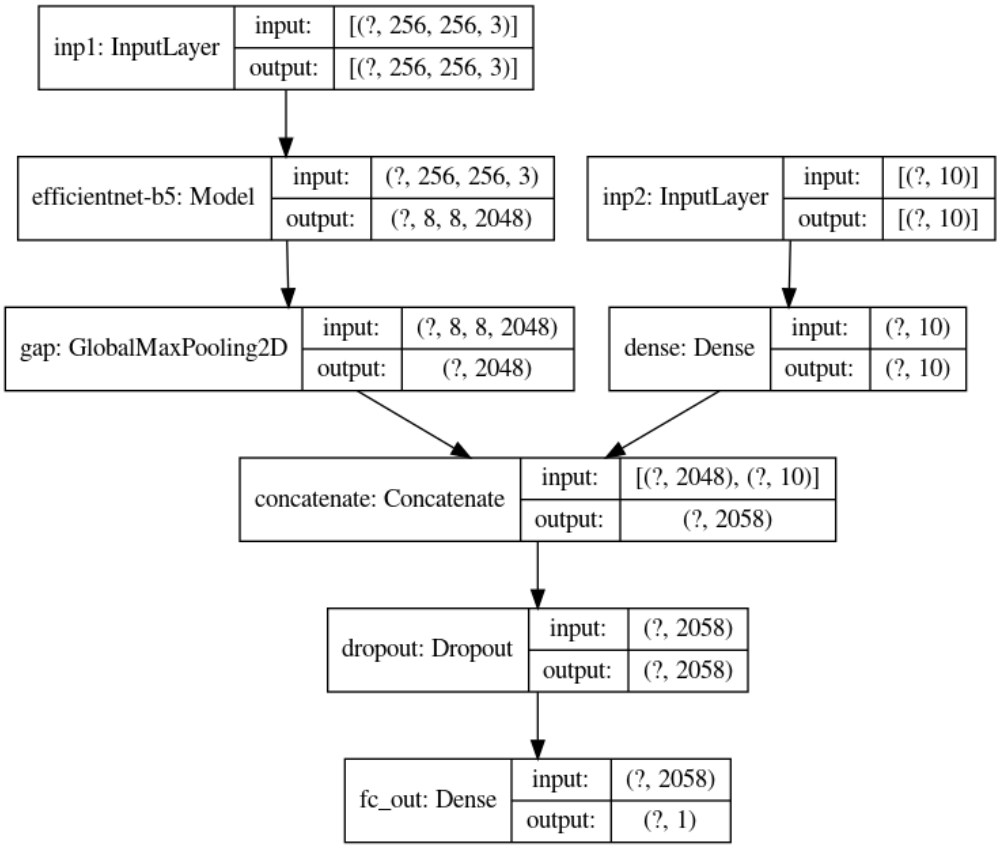

Figure 2: Model architecture

0.8518.[4] Figure 3 represents the training and validation loss as well as the training and validation accuracy and AUC score as a function of the model epochs.[5] It must be pointed out that the model history is based on the validation data since Kaggle did not provide any class labels for the test data to the users.

The AUC score describes the ability of the model to rank a random positive case more highly than a random negative case. Thus, it summarises how well the model discriminates images that represent malign melanomas from those showing benign melanomas. An AUC = 0.5 means that the model performs on random guess level, and an AUC = 1 shows ideal performance (perfect discrimination between positive and negative cases). Considering our final private AUC score of 0.8518, we can infer that given a randomly chosen observation, a positive instance will receive a higher probability than a negative instance with a probability of 85.18% (Hanley and McNeil 1982).

Although this seems to be an acceptable performance value, the left panel in figure 3 shows that the training AUC increases continually over the epochs, whereas the validation AUC does only improve slightly. The training and validation loss curves visualized in the right panel of figure 3 show relative convergence over the epochs. Thus, overfitting does not seem to be the primary reason for the divergence of the training and validation AUC. Although the cause for

[4]The public AUC score is based on 30% of the test data and the private AUC score is based on 70%.

[5]Note that the illustrated curves show the course of performance metrics averaged over the 5 folds.

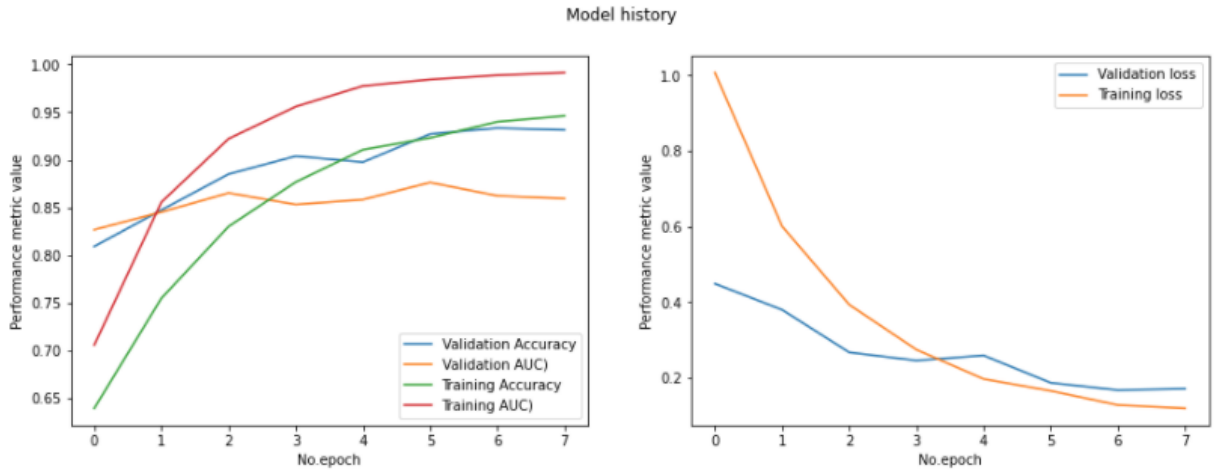

Figure 3: Accuracy, AUC, and loss as a function of epochs

the divergence of the training and validation AUC is not clear at all, the class imbalance could be an influencing factor. Due to the random assignment of samples to the TFRecords and cross validation folds, unequal distribution of the few positively labeled cases could represent a drawback of our data input pipeline in this regard. An approach to analyzing the effect of this sampling bias could be performed with stratified sampling, allocating the same number of positive samples to each TFRecord file. Even though we implemented techniques such as image augmentation and class weighting to overcome the issue of data imbalance, the crucial problem could be based in the data input pipeline. The resulting bias on TFRecords level might be too strong to be equalized by imbalance approaches on model level.

Beyond that, a few model architecture and preprocessing aspects must be examined critically for future machine learning approaches dealing with the detection of melanoma: we applied transfer learning based on an EfficientNet architecture that learns on the image database "ImageNet". This database encompasses a vast number of images that are conceptually allocated to different nodes of the WordNet hierarchy. Consequently, this image corpus is a representative ground-truth for multiple domains rather than a database for a specific application area. The fact that we rely on pretrained weights that were calculated on images that do not resemble the images that are to be classified can thus be examined critically. We cannot say but what margin a dedicated, larger dermoscopic dataset would improve performance.

Apart from the transfer learning implementation, different model architectures could be applied and analyzed in terms of performance. The EfficientNet approach includes multiple model architectures that differ in terms of complexity. An Ensemble Learning approach averaging various model architectures could increase the accuracy of melanoma detection. Considering the data preprocessing, the model performance could benefit from the removal of hairs that partially shade main regions of interests and thus contaminate the images. Borys et al. 2015 introduce an algorithm that removes hairs from dermoscopic images and consequently increases the image quality.

On data level, the input that is provided by Kaggle shows strong substantial bias since it exclusively comprises dermoscopic images from white people. This fact should be consid-

ered critically since it excludes black people from medical improvements based on technical developments in the machine learning domain. Although the skin cancer prevalence is significantly higher in white people (35-45% of all tumors) than in black people (1-2% of all tumors), black people tend to get the skin cancer diagnosis in a later disease stage, worsening the disease prognosis (Bradford 2009). From an ethical point of view, it would consequently be necessary to also include images from black people as data input as we don't know how our model performs when confronted with dark skin.

6 Conclusion

All in all, our model showed reasonable performance metrics. Nevertheless, there are several approaches that could additionally be implemented on data input and preprocessing level, as well as on model level, to further increase the potential to improve a classifiers' diagnostic abilities. Given the vast number of possibilities to intervene in the learning process of CNNs and the complexity of such models, it is inherently difficult to get the optimal model architecture with an ideal tuning of parameters.

Appendix

Manuel was responsible for all data handling related tasks including preprocessing and serialization of the data. David focused on augmentation and re-sampling techniques to combat the class-imbalance problem in the dataset. The choice for a CNN architecture paired with a transfer learning approach for the classification problem at hand is the result of experiments conducted by both in equal parts.

References

Hanley, James A and Barbara J McNeil (1982). "The meaning and use of the area under a receiver operating characteristic (ROC) curve." In: *Radiology* 143.1, pp. 29–36.

Kusumoputro, Benjamin and Aripin Ariyanto (1998). "Neural network diagnosis of malignant skin cancers using principal component analysis as a preprocessor". In: *1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No. 98CH36227)*. Vol. 1. IEEE, pp. 310–315.

Ling, Charles X and Victor S Sheng (2008). *Cost-sensitive learning and the class imbalance problem*.

Bradford, Porcia T (2009). "Skin cancer in skin of color". In: *Dermatology nursing/Dermatology Nurses' Association* 21.4, p. 170.

Deng, Jia et al. (2009). "Imagenet: A large-scale hierarchical image database". In: *2009 IEEE conference on computer vision and pattern recognition*. Ieee, pp. 248–255.

Borys, Damian et al. (2015). "A simple hair removal algorithm from dermoscopic images". In: *International Conference on Bioinformatics and Biomedical Engineering*. Springer, pp. 262–273.

Ruder, Sebastian (2016). "An overview of gradient descent optimization algorithms". In: *arXiv preprint arXiv:1609.04747*.

Shin, Hoo-Chang et al. (2016). "Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning". In: *IEEE transactions on medical imaging* 35.5, pp. 1285–1298.

Xie, Fengying et al. (2016). "Melanoma classification on dermoscopy images using a neural network ensemble model". In: *IEEE transactions on medical imaging* 36.3, pp. 849–858.

Schadendorf, Dirk et al. (2018). "Melanoma". In: *The Lancet* 392.10151, pp. 971–984. ISSN: 0140-6736. DOI: https://doi.org/10.1016/S0140-6736(18)31559-9. URL: http://www.sciencedirect.com/science/article/pii/S0140673618315599.

Brinker, Titus J et al. (2019). "A convolutional neural network trained with dermoscopic images performed on par with 145 dermatologists in a clinical melanoma image classification task". In: *European Journal of Cancer* 111, pp. 148–154.

Johnson, Justin and Taghi Khoshgoftaar (Mar. 2019). "Survey on deep learning with class imbalance". In: *Journal of Big Data* 6, p. 27. DOI: 10.1186/s40537-019-0192-5.

Shorten, Connor and Taghi M Khoshgoftaar (2019). "A survey on image data augmentation for deep learning". In: *Journal of Big Data* 6.1, p. 60.

Tan, Mingxing and Quoc V Le (2019). "Efficientnet: Rethinking model scaling for convolutional neural networks". In: *arXiv preprint arXiv:1905.11946*.