Skin lesion classification using classical feature extraction methods and convolutional neural networks

Lennart Maack, Sandipan Mukherjee, Aditya Tandon, Paulina Vennemann, Arpan Dholakiya

Abstract—**In this paper we describe two methods to classify nine different types of skin lesions including malignant melanoma. We were provided with a dataset of 25,331 images for training and validation. This dataset contains images of eight different classes. The final test dataset includes an additional unknown class. The first method was applied in phase one where we used a decision tree classifier after preprocessing the images and extract features based on the ABCD method. Preprocessing was needed because of how differently the images were taken as well as the different resolutions of the images. Our approach included cropping out the region of interest by applying Otsu's method. The final model of phase one was only able to achieve an overall sensitivity of 13.09% on the validation data. In phase two we used a convolutional neural network, namely SEResNext50, for our model. Additionally, we applied different augmentations on the training images, e.g. flipping and color jittering. During the training of the model we made use of a learning rate scheduler, which reduces the learning rate on plateau based on the area under curve of the receiver operating characteristic curve. Our final model in phase two achieved an overall sensitvity of 84.32%.**

I. Introduction

In this paper, we describe our approach towards classifying dermoscopic images across eight known skin lesion classes and one unknown class. The techniques focus primarily on the classification of melanoma skin lesions. In phase 1 skin lesions are classified using traditional feature extraction and machine learning methods. Images were first preprocessed to allow for effective feature extraction. Features were extracted in accordance with the ABCD technique, evaluating the **a**symmetry, **b**orders, **c**olors and **d**iameter of the skin lesion [1]. These features are then input into a classifier and results are evaluated. Multiple classifiers like the Linear SCV, SVM Classifier, Decision Tree, Gaussian NB, K-Neighbors Classifier and Random Forest Classifier were tested and it was observed that the Decision Tree Classifier produced the best results among the tested methods.

In phase 2 of our approach, skin lesions are classified using a Deep Convolutional Neural Network (CNN). Since training a neural network from scratch is both, time and resource intensive, getting a robust and reliable CNN with limited computation and time resources is difficult. To overcome this, a transfer learning approach was applied with a very small learning rate to adapt the pre-trained network to the specified application. Different CNN architectures using SEResNext 50 (SER-50), EfficientNet b7 (EF-b7), ResNet-101 and InceptionNet-V3 were tested and it was observed that SER-50 produced the best results.

II. Understanding the dataset

To develop these learning techniques, the ISIC - 2019 : Training dataset published by the International Skin Imaging Collaboration (ISIC) is used. The ISIC Archive contains the largest publicly available collection of quality-controlled dermoscopic images of skin lesions [2]. The ISIC - 2019 : Training dataset contains a total of 25,331 dermoscopic images, consisting of 8 different types of skin lesions, namely melanoma (MEL), melanocytic nevus (NV), basal cell carcinoma (BCC), actinic keratosis (AK), benign keratosis (BKL), dermatofibroma (DF), vascular lesion (VASC) and squamous cell carcinoma (SCC). This dataset is based on the aggregation of 3 different datasets - the HAM10000 dataset, the BCN20000 dataset and the MSK dataset, cumulatively containing images of varied quality and size. For instance, images from the HAM10000 dataset are of size 800 x 600px at 72DPI, centered and cropped around the lesion, and have already been subjected to manual histogram corrections to enhance visual contrast and color reproduction [3]. The BCN20000 dataset contains high resolution images of size 1024 x 1024 px, however, these images are uncropped and have lesions positioned in difficult and uncommon locations [5]. The MSK dataset contains images of varying sizes.

A. Splitting Training and Validation Data

To facilitate performance estimation of the developed techniques, the ISIC - 2019 : training dataset is further split into a training and validation subset in a 90:10 ratio respectively.

Table I
DISTRIBUTION OF IMAGE CLASSES IN **ISIC 2019: TRAINING** [3] [4] [5] DATASET AFTER A 90:10 TRAINING-VALIDATION DATA SPLIT

Diagnostic Classes	Training	Validation
Melanoma	4060	462
Melanocytic Nevus	11592	1283
Basal Cell Carcinoma	2982	341
Actinic Keratosis	787	80
Benign Keratosis	2368	256
Dermatofibroma	218	21
Vascular Lesion	228	25
Squamous Cell Carcinoma	562	66
Unknown	0	0
Total	**22797**	**2534**

The classifiers are trained using only the training subset of the complete training dataset, and their performance is evaluated by comparing the predictions of the images in the validation subset with the available ground truth values. A breakdown of the distribution of images in the training dataset after performing the training-validation split is as indicated in Table 1. As can be seen from Table 1, the distribution of images belonging to different classes is extremely varied. The dataset contains a large number of images concentrated in a few classes and often does not contain enough images belonging to other classes. This is one of the primary factors affecting models during training, as models tend to become heavily biased towards the classes with higher number of samples. The ISIC - 2019 : Training dataset also contains additional metadata like the age group and gender of the patient as well as the site of the skin lesion in one of eight locations on the body. This metadata is available for a majority of the skin lesion images, however there are missing values in the dataset. In the scope of this paper, however, this metadata was not utilized and the class predictions are made without context of this metadata.

III. Phase 1

A. *Image Preprocessing*

The objective behind preprocessing images is to improve the quality of the image to yield better results in subsequent segmentation and feature extraction steps. Preprocessing operations performed were the removal of artifacts, like hair, and color normalization, to neutralize color variations due to varied lighting conditions or variations in the devices used to capture lesion images. Additional preprocessing operations that are performed are the resizing/cropping of the images around the region of interest, i.e the lesion. Morphological operations were also performed during this step to remove noise and close gaps in the image, thus improving the overall performance of the segmentation operation.

1) Hair Removal: A large majority of images in the dataset contain lesion images, where the lesion is obstructed by the presence of artifacts like hair. The image is converted to grayscale and the necessary morphological operations as well as thresholding is performed. Since the hair in the images are extremely dark regions running across the image, they can be enhanced using a Black Top-Hat morphological operation. The image is then thresholded using a low threshold value and this mask is further fed to the inpainting operation to remove the hair-like features from the image [13]. The final image obtained shows the hair removed to a great extent and this image is then used as the input for the segmentation operation for Region of Interest (ROI) extraction. Although the hair removal operation was implemented, this operation was not applied to the complete dataset.

2) Image Resizing, ROI Extraction and Cropping: Images in the dataset are of different sizes and quality. A large majority of images have black borders present as well as lesions occurring at different areas in the image. Before the features of the lesion can be extracted, standardization in the images is important. Additionally, the presence of black borders can hinder the output of intensity based feature extraction methods. The initial step in the ROI extraction operation was to perform thresholding of the image using Otsu's method. Once thresholded, the output image was operated on using morphological operations such as Dilation, Erosion and Closing to retrieve a well segmented lesion region devoid of inconsistencies and holes. This binarized image is then used to retrieve the lesion without its dark background. The last step in this operation was to center the ROI in the image and crop it to a fixed size around the lesion. Although the cropping step was implemented, the complete dataset was not subjected to this cropping/resizing operation since the efficacy of the implemented algorithm was found to be unsatisfactory, i.e approximately 15% of the images were cropped incorrectly.

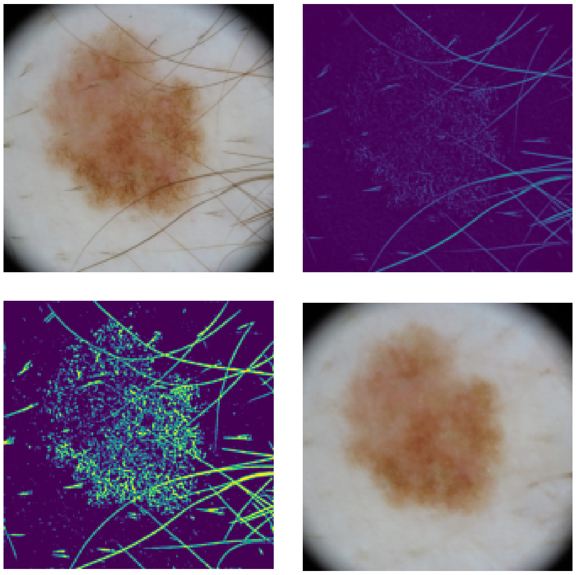

Figure 1. Hair Removal Steps

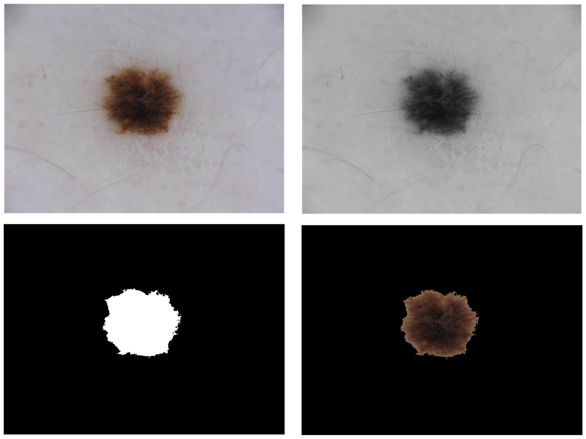

Figure 2. ROI Extraction

B. Feature Extraction

The automated diagnosis of skin lesions was based on the so-called ABCD-rule of dermatoscopy. The choice of following feature extraction methods rests upon this empirical approach in order to characterize the Symmetry, Border irregularity, Colour and Diameter properties. [1]

1) Hu Moments: The first step is the evaluation of the symmetry of the lesions, since asymmetric pattern can be a sign of melanoma. This aspect is implemented by calculating Hu Moments of the examined image. In general image moments indicate the weighted average of pixel intensities. In the specific case of Hu Moments, the result is invariant to translation, scale and rotation. The implemented function provides seven different attributes, which relate to the geometrical properties, computed by central image moments. [10]

2) Color Statistics: The colour of a lesion is in important aspect regarding the recognition of skin cancer. Melanoma can be identified due to formation of colour variations. As descriptors for this appearance different statistical parameters from different colour channels were calculated, including the standard deviation and the average in the HSV and all RGB channels. [10]

3) GLC Matrix: The last feature taken in to consideration was the gray-level co-occurrence matrix (GLCM) [8] which describes texture pattern obtained from the statistical distribution of intensities at particular position relative to each other in an image. The obtained features contain the contrast, dissimilarity, homogeneity, energy, correlation and ASM. [10]

C. Classification

In the next step the obtained feature vector was passed through a classification method to predict the melanoma class to which the image belongs. The classification part of the project was implemented using the Decision Tree Classifier. This choice resulted on the comparison of different classifiers (Linear SCV, SVM Classifier, Decision Tree, Gaussian NB, K-Neighbors Classifier, Random Forest Classifier), regarding accuracy, sensitivity, specificity and the F1 score. All methods were implemented using the functions provided by the scikit-Learn library.

1) Decision Tree: The Decision Tree Classifier is a supervised Machine Learning model, which derives its conclusion based on the structure of a decision tree. The conjunctions of features are represented in the branches in order to derive at a conclusion/class label which is denoted in the leaves. Decision trees are widely used, because of their tangibility and simplicity, but also include some limitations since they can be very non-robust.

D. Results Phase 1

The table below shows the final result of the first phase. It can be observed that the sensitivity and F1 score is significant lower compared to the accuracy and specificity. Reason for this could be the unbalanced spreading of the training data since 85% of the images belong to the second diagnostic class: Melanocytic Nevus.

Table II
SUBMITTED RESULTS [%]

accuracy	sensitivity	specificity	F1 score
85.034	13.09	87.694	11.851

This relation could be an explanation of the high accuracy. Furthermore sustained the project a higher complexity compared to some binary classification from other research papers due to the classification of multiple labels. [11] In addition one could argue that the conducted implementation of ABCD rule is not specific and sufficient enough to predict all the different classes and the choice of the feature extraction methods is the reason for this outcome. One last possible point of failure could be some imprecision in the cropping method, which leads to wrongly classified images.

Future measures in order to improve this outcome could be:

- optimized cropping method to avoid non classifiable images due to cropping errors
- different method of normalization
- different choice of extracted features
- optimized tuning of the hyper-parameters of the decision tree classifier

IV. Phase 2

A. Data normalization

Normalization was performed on both the training and validation data. It was done to eliminate anomalies that could have made the dataset more complicated and to ensure each parameter has a similar distribution. This results in a faster convergence while training the network. Normalization was executed by subtracting the mean pixel value from each channel and dividing the result with their standard deviation.

$$output[channel] = \frac{(input[channel] - mean[channel])}{std[channel]}$$

The sequence of mean values for the input was taken as (0.485, 0.456, 0.406) and the sequence of standard deviation values was taken as (0.229, 0.224, 0.225).

B. Data Augmentation

Many different data augmentations were applied with the help of various transformations in order to diversify the data for the training of the neural network. They were applied with functions from the Pytorch package named TorchVision. This helped to increase the size of the dataset without having to collect new data. The augmentations used were Random Horizontal Flip, Random Vertical Flip, Random Rotation and Random Color Jitter. Data augmentation was only applied on training images.

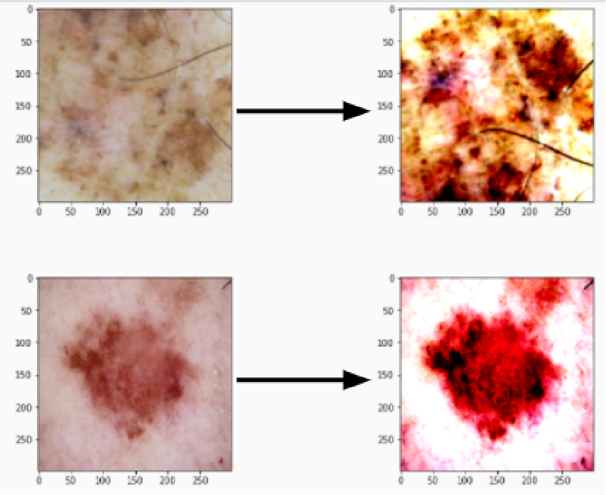

Figure 3. Original images and their corresponding augmented image

C. Neural Network Architectures

To build the architecture of our CNN model we made use of the concept of transfer learning for image classification and tried out the following pre-trained architectures: Efficient-Netb7, ResNet-101, SEResnext50, InceptionNet-V3.

1) Transfer Learning: Transfer learning is a popular method in computer vision that allows fast building of accurate models. With transfer learning, there is no need to start the learning process from scratch, instead the used model contains patterns that already have been trained by solving different problems. Advantages of using transfer learning:

- simple incorporation,
- achieving the same or even better (depending on the dataset) model performance quickly
- less requirement of labeled data

2) SE ResNext 50: When a CNN creates an output feature map from a convolutional layer, it gives equal weighting to each channel. Squeeze and Excitation (SE) Network is a method that adaptively re-calibrates channel-wise feature responses by explicitly modelling inter-dependencies between the channels [6]. A SE block gives an output in the shape of (1 x 1 x channels), which specifies the weighting for each channel. The neural network can learn this weighting by itself like other parameters.

After trying out and comparing the results of the different architectures for our model, we decided to continue with SEResnext50 architecture since it gave us the best results regarding overall sensitivity. The decision was made by training each individual model for 5-10 epochs. As from literature is known the SEResnext50 has good performance in terms of generalising large amounts of data by fine tuning the last few layers of the network [6].

D. Training

For the training of our model we used the training images of size 512x512. Training was performed on a Tesla P100-PCIE-16GB GPU which enabled a maximum possible batch size of eight. As an optimizer we used Adam to converge our model [7]. We started with an initial learning rate of 0.001. Additionally, we implemented a learning rate scheduler which reduces the learning rate on plateau depending on a specific metric. If the metric stops improving over three epochs, the learning rate lowers by the factor 0.1. Since the data is skewed towards the second class, we had an imbalanced classification problem and therefore decided to use the Area Under Curve (AUC) of the receiver operating characteristic curve (ROC curve) as a metric [12]. The AUC is typically used for binary classifications, so we calculated the AUC of each class versus all other classes. Due to the unbalanced training data we used CrossEntropyLoss as the criterion for our loss function. We also made use of an early stopping algorithm which saves the model after each epoch with increasing AUC and stops the training after five epochs without increasing AUC. This reduces the chances overfitting [8]. The evolution of the AUC, accuracy, overall sensitivity and loss can be seen in figure 6. The optimal AUC is at epoch 20. The final model was saved at epoch 20 and used for prediction.

E. Results on the validation data

In this section we present the results of the predictions on the validation data by plotting the ROC curve of each class versus all other classes. We were able to achieve an overall test sensitivity of 84.32% as well as a sensitivity of 72.9% for the Melanoma class. The corresponding ROC curve as well as the AUC value of each class can be found in figure 5.

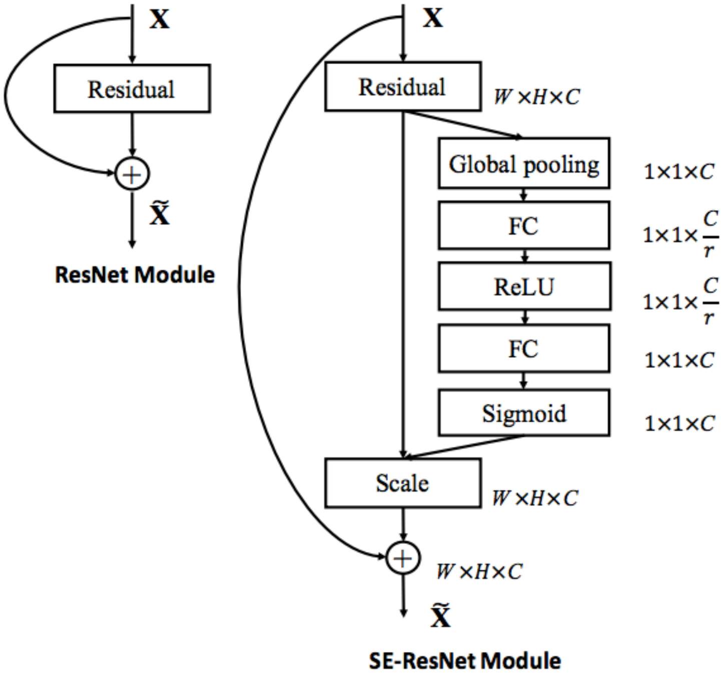

Figure 4. The schema of the original Residual module (left) and the SEResNet module (right).

F. Conclusion and further improvement possibilities

Drawing conclusions from the results of phase 2, the classifier and the methods chosen worked well regarding the limited time and resources we had. Nevertheless further improvements are still needed to achieve better results. Improvements can be made by extending the dataset. For example adding images from previous competitions or from other sources as well as using more augmentations. Further improvements could be made by using ensemble learning. Ensemble learning is especially useful when dealing with imbalanced classification problems because specific models can get "lucky" or "unlucky" for specific classes. If you combine multiple models this effect can get canceled out [8]. Additionally, it is possible to use a better validation strategy, e.g. stratified k-fold cross validation which is useful for imbalanced classification problems as well.

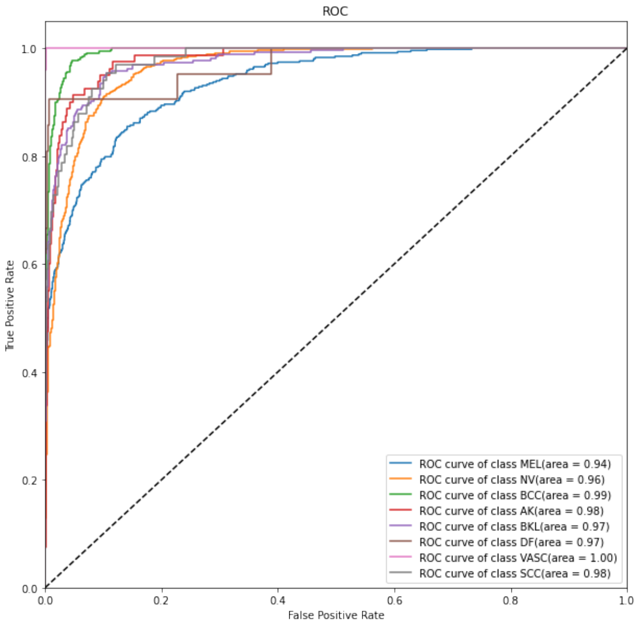

Figure 5. ROC Curve

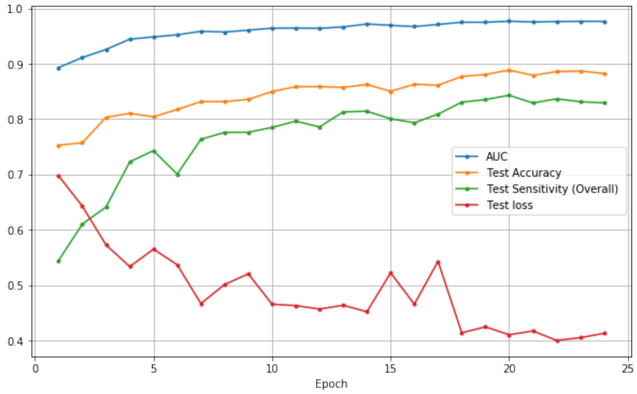

Figure 6. AUC, accuracy, sensitivity, loss per epoch

References

[1] W. Stolz, A. Riemann, A. Cognetta, and O. Braun-Falco, “Abcd rule of dermatoscopy : a new practical method for early recognition of malignant melanoma,” Eur. J. Dermatol., vol. 4, pp. 521–527, 1994.

[2] Overview of "SIIM-ISIC Melanoma Classification", https://www.kaggle.com/c/siim-isic-melanoma-classification/overview.

[3] Tschandl P., Rosendahl C., Kittler H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Sci. Data 5, 180161 doi.10.1038/sdata.2018.161 (2018)

[4] Noel C. F. Codella, David Gutman, M. Emre Celebi, Brian Helba, Michael A. Marchetti, Stephen W. Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, Allan Halpern: “Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC)”, 2017; arXiv:1710.05006.

[5] Marc Combalia, Noel C. F. Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Allan C. Halpern, Susana Puig, Josep Malvehy: “BCN20000: Dermoscopic Lesions in the Wild”, 2019; arXiv:1908.02288.

[6] Hu, Jie, L. Shen and G. Sun. “Squeeze-and-Excitation Networks.” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (2018): 7132-7141.

[7] Diederik P. Kingma and Jimmy Ba: "Adam: A Method for Stochastic Optimization", 2017, arXiv:1412.6980.

[8] R. M. Haralick, K. Shanmugam and I. Dinstein, "Textural Features for Image Classification," in IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-3, no. 6, pp. 610-621, Nov. 1973, doi: 10.1109/TSMC.1973.4309314.

[9] Prechelt L. (2012) Early Stopping — But When?. In: Montavon G., Orr G.B., Müller KR. (eds) Neural Networks: Tricks of the Trade. Lecture Notes in Computer Science, vol 7700. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-35289-8_5.

[10] C. Barata, M. E. Celebi, J. S.Marques, "A Survey of Feature Extraction in Dermoscopy Image Analysis of Skin Cancer" IEEE Journal of Biomedical and Health Informatics, Vol 23, No.3 (2019)

[11] S. Mustafa, A. B. Dauda, M. Dauda, "Image processing and SVM classification for melanoma detection", 2017 International Conference on Computing Networking and Informatics (2017)

[12] Thomas C.W. Landgrebe, Robert P.W. Duin "Approximating the multiclass ROC by pairwise analysis", 2007 Pattern Recognition Letters 28, Electrical Engineering, Mathematics and Computer Science, Delft University of Technology

[13] Telea, Alexandru. “An image inpainting technique based on the fast marching method.” Journal of graphics tools 9.1 (2004): 23-34.

[14] Rokach, L. (2010). "Ensemble-based classifiers". Artificial Intelligence Review. 33 (1–2): 1–39