Internship Technical Report

Jessica Klebe

j.s.klebe@gmx.de

April 26, 2021

1 Dataset exploration

Fig. 1 shows the histogram of the frequency of the labels in order to become familiar with the data as a first step.

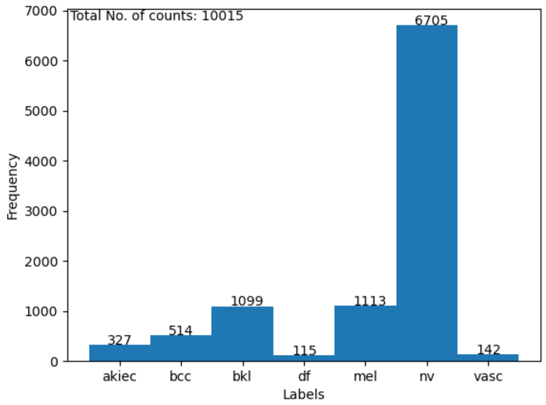


Figure 1: Histogram illustrating the number of images of various skin lesions: Actinic keratoses and intraepithelial carcinoma / Bowen's disease (akiec), basal cell carcinoma (bcc), benign keratosis-like lesions (solar lentigines / seborrheic keratoses and lichen-planus like keratoses, bkl), dermatofibroma (df), melanoma (mel), melanocytic nevi (nv) and vascular lesions (angiomas, angiokeratomas, pyogenic granulomas and hemorrhage, vasc).

While the total number of images in the dataset is 10015, Fig. 1 shows that melanocytic nevi (nv) are represented by 6705 images and, hence, they make 66.95% of the total dataset. Skin lesions with the lowest number of images included in the dataset are dermatofibroma (df) and vascular lesions (vasc), each with about only 100 images which makes about 1% of the dataset.

2 First attempt

In the first attempt, a Resnet34 model has been used. The training set contained 80% of the images of each skin lesion type in order to maintain their frequency. Consequently, the test set consisted of the remaining 20%. No pre-training has been conducted. The images have been resized to a size of 224px each side and normalized according to their mean and standard deviation (mean=[194.6979, 139.2626, 145.4852], std=[22.8551, 30.9032, 33.9032]). The stochastic gradient descent optimizer has been used with a learning rate of 10^{-3} and a momentum of 0.9. 10 epochs each with a batchsize of 100 have been executed. After a few epochs it became clear that the output is always nv which led to an accuracy of 0.6695, a balanced accuracy of 0.1429 and a mean recall of 0.0158.

3 Further attempts

In further attemps the dataset has been split into a training set (70%), a validation set (10%) and a test set (20%).

3.1 Transformation of the training data

In order to avoid the problem of always obtaining nv as an output, the training set has been balanced.

This means all subsets of images of the training set that are not labelled as nv have been added multiple times to the training dataset such that each label is represented roughly equally in the training dataset. Tab. 1 depicts how often a subset of a specific label is included in the training dataset and, consequently, how many images of this particular label are contained by the final training dataset.

Label	No. of subsets of a label	No. of labels
akiec	20	4,578
bcc	13	4,677
bkl	6	4,615
df	58	4,669
mel	6	4,674
nv	1	4,693
vasc	47	4,671

Table 1: In order to obtain a balanced training dataset the subsets of the labels have been added multiple times to the training dataset (middle column). The total number of images of the final training dataset is depicted in the right column.

Having added certain images multiple times to the training set, the neural net could simply learn these images. Therefore, the images need to be altered by every iteration. Perez et al. investigated in [1] data augmentation scenarios for melanoma classification trained, inter alia, on a ResNet. Following this paper, the subsequent transformations have been implemented.

- *transforms.ColorJitter*: Modification of saturation, contrast, and brightness chosen uniformly from [0.6 - 1.4]. (Similar to scenario B from [1].)
- *transforms.RandomCrop*: Image is randomly cropped. The transformer first resizes the image to 256px each side and then randomly crops this image to a size of 224 px x 224 px. (Similar to scenario F from [1].)
- *transforms.RandomHorizontalFlip*: Image randomly flipped horizontally with probability of 50%. (Similar to scenario E from [1].)
- *transforms.RandomVerticalFlip*: Image randomly flipped vertically with probability of 50%. (Similar to scenario E from [1].)
- *transforms.RandomAffine*: Random affine transformation of the image keeping center invariant. Range of degree: (-40°, 40°), scaling factor interval: (0.9, 1.1), shear = 0. (Similar to scenario D from [1].)
- *transforms.RandomPerspective*: Performs a random perspective transformation of the given image with probability 0.5. Degree of distortion: 0.2.

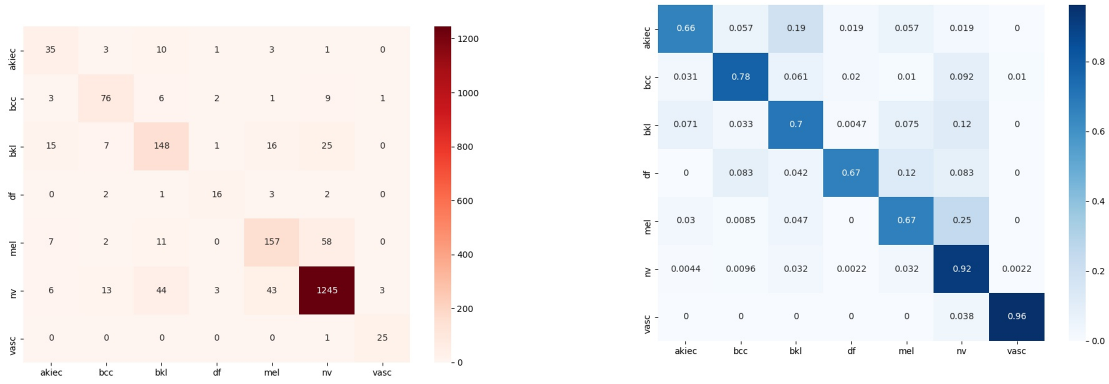
3.2 Further implementation details

Furthermore, the ResNet34 model has been pretrained on ImageNet [2]. As an optimizer the Adam algorithm has been used. Additionally, the learning rate has been adjusted using a cosine annealing schedule ($torch.optim.lr\_scheduler.CosineAnnealingLR$) where the number of update steps is $T_{max} = 1037$ and, hence, corresponds to one epoch. Finally, early stopping method has been used, i.e. if after 10 epochs on the validation set the balanced accuracy has not improved, the training is stopped and the best model is reported. Afterwards, the model is executed on the test set for further evaluation.

3.3 Results

3.3.1 Test set

Fig. 2 illustrates the confusion matrices determined on the test set and Tab. 2 depicts the test set's mean recall, balanced accuracy and accuracy. The diagonal of the normalized confusion matrix contains the recall values of each label. The original confusion matrix containing absolute values (Fig. 2a) has a high value on the diagonal only for nv in comparison to the other labels. This is due to the circumstance that the test set is not balanced and, therefore, it contains a larger number of images containing the label nv. Consequently, the normalized confusion matrix (Fig. 2b) depicts in a more suitable way how successful the model is. Furthermore, the high value on the diagonal for nv in Fig 2a explains why the accuracy is not a suitable evaluation metric.



(a) Original confusion matrix.

(b) Confusion matrix normalized with respect to rows. The values on the diagonal depict the recall values of each label.

Figure 2: Confusion matrices of the test set.

mean recall	balanced accuracy	accuracy
0.7638	0.7330	0.8489

Table 2: Mean recall, balanced accuracy and accuracy of the test set.

3.3.2 Validation set

Fig. 3 illustrates the confusion matrices determined on the validation set and Tab. 3 depicts the validation set's mean recall, balanced accuracy and accuracy. The diagonal of the normalized confusion matrix contains the recall values of each label.

mean recall	balanced accuracy	accuracy
0.8077	0.8342	0.8774

Table 3: Mean recall, balanced accuracy and accuracy of the validation set.

3.3.3 Training set

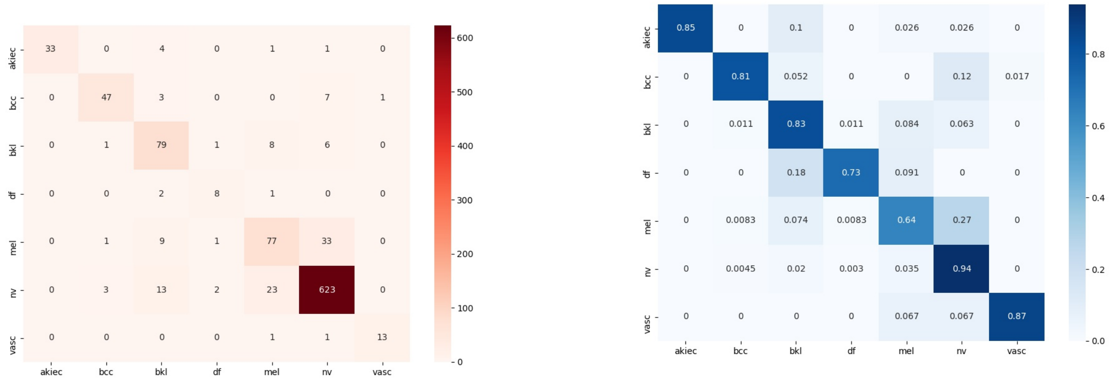
Fig. 4 illustrates the confusion matrices determined on the training set and Tab. 4 depicts the training set's mean recall, balanced accuracy and accuracy. The diagonal of the normalized confusion matrix contains the recall values of each label.

mean recall	balanced accuracy	accuracy
0.9993	0.9993	0.9993

Table 4: Mean recall, balanced accuracy and accuracy of the training set.

4 Conclusion

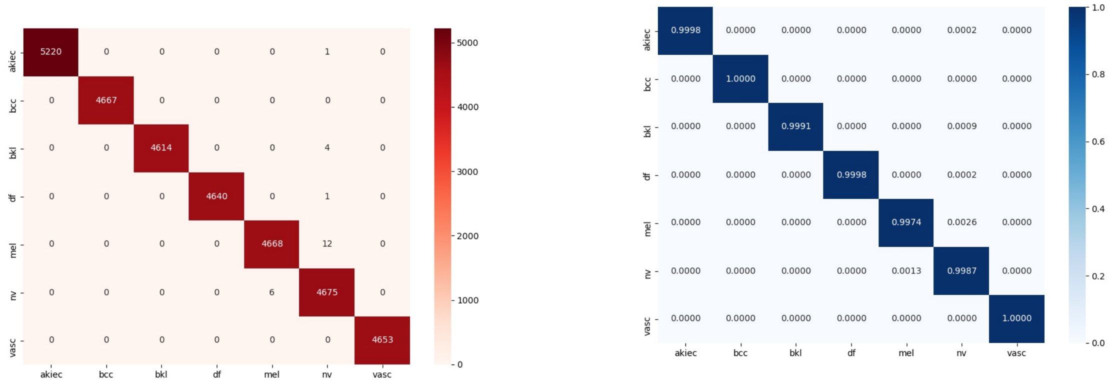
The normalized confusion matrix of the test set - that shows a relatively high number of results on its diagonal - suggests that the model has learned the connection between input and output. This is a progress in comparison to the first model that solely predicted the label nv. Evaluated on the test set, a balanced accuracy of 0.7330 is obtained. The aim was to reach a value comparible to the upper third of the ISIC 2018 challenge. The upper third would be rank 25 or higher. Rank 25 yielded a balanced accuracy 0.718 which is comparible to 0.7330. However, the solution has not been uploaded yet such that the test sets on which the balanced accuracy is evaluated differ. The next steps could be to upload the solution and to



(a) Original confusion matrix.

(b) Confusion matrix normalized with respect to rows. The values on the diagonal depict the recall values of each label.

Figure 3: Confusion matrices of the validation set.



(a) Original confusion matrix.

(b) Confusion matrix normalized with respect to rows. The values on the diagonal depict the recall values of each label.

Figure 4: Confusion matrices of the training set.

train the semantic segmentation model.

References

[1] Fábio Perez, Cristina Vasconcelos, Sandra Avila, and Eduardo Valle. Data augmentation for skin lesion analysis. In *OR 2.0 Context-Aware Operating Theaters, Computer Assisted Robotic Endoscopy, Clinical Image-Based Procedures, and Skin Image Analysis*, pages 303–311. Springer, 2018.

[2] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.