Supplementary Material

## 1. Algorithm, model, modelling, training and evaluation

**1.1 Model architecture**

The *analyze* network is an image classifier based on an inception\_v3 model. The image classification assesses the whole image and accepts images of single skin lesions as input. Images are therefore decoded and resized to 299 (x-axis) x 299 (y-axis) x 3 (RGB) pixels to fit the input requirements of the inception model. Based on the input, the algorithm returns probability scores for each of the possible 47 lesions. Each score lies between 0% and 100% and reflects the probability of being the correct diagnosis. All scores of a lesion add up to 100%. The assessment is performed in significantly less than a second and can also be used in real-time like for video input.

The *detect* network is based on the faster\_rcnn\_inception\_v2\_coco model. It accepts images containing one or several skin lesions as input and returns a list of bounding boxes, a label assigned to each box and a corresponding probability. It consists of three components: convolution layer, region proposal network (RPN), and “classes and bounding boxes prediction”. The convolution layer serves as an initial filter for relevant features of the image. The RPN is a small network, that localizes objects in the image and generates boxes around these objects. The “classes and bounding boxes prediction” calculates the label probabilities for each box and returns the most likely diagnosis, together with its probability. Finally, overlapping boxes are removed. The decision, which box to keep, is based on 1) higher risk and 2) higher probability score. The cumulative probability of all labels can exceed 100% as each box is evaluated separately.

**1.2 Data sources**

An image data set of 19,576 anonymized images, was used. It was split (the training script splits the data by random) into a training and validation data set (18,384 dermatologists labeled images and a test data set (n=1,192) in order to ensure that test data set images are not used in both training and validation of the model the filenames of the test data set images start with tds\*.jpg.
The image data set (n=19,576) contains anonymized representative (age of participants: 18 to 86 years, Fitzpatrick skin type 1 to 4, a clear contrast to surrounding skin, not covered by hair or opaque/glittering substances, not previously traumatized except for label „skin injury") images.
The test data set (n=1,192), which was not used for training, contains anonymized representative images (age of participants: 18 to 86 years, Fitzpatrick skin type 1 to 4, a clear contrast to surrounding skin, not covered by hair or opaque/glittering substances, not previously traumatized).

**1.3 Model training**

The models (both, analyze and detect) were retrained using the training data set for CENSORED iterations. The training process runs through the training image data set in batches and iterations, with validation run on each iteration.

The following settings were used for analyze network to train:

* learning\_rate: CENSORED
* testing\_percentage: 10
* validation\_percentage: 10
* eval\_step\_interval: 10
* train\_batch\_size: CENSORED
* test\_batch\_size: CENSORED
* validation\_batch\_size: 100

The following settings were used to train the detect network:

* truncated\_normal\_initializer stddev: 0.01
* first\_stage\_nms\_score\_threshold: 0.0
* first\_stage\_nms\_iou\_threshold: 0.7
* first\_stage\_max\_proposals: 300
* first\_stage\_localization\_loss\_weight: 2.0
* first\_stage\_objectness\_loss\_weight: 1.0
* initial\_crop\_size: 14
* maxpool\_kernel\_size: 2
* maxpool\_stride: 2
* dropout: false
* dropout\_keep\_probability: 1.0
* fc\_hyperparams l2\_regularizer weight: 0.0
* variance\_scaling\_initializer factor: 1.0 uniform: true mode: FAN\_AVG
* second\_stage\_post\_processing batch\_non\_max\_suppression score\_threshold: 0.0
* iou\_threshold: 0.6
* max\_detections\_per\_class: 100
* max\_total\_detections: 300
* score\_converter: SOFTMAX
* second\_stage\_localization\_loss\_weight: 2.0
* second\_stage\_classification\_loss\_weight: 1.0
* batch\_size: CENSORED
* optimizer momentum\_optimizer
* learning\_rate: manual\_step\_learning\_rate
* initial\_learning\_rate: CENSORED
* momentum\_optimizer\_value: 0.9
* use\_moving\_average: false
* gradient\_clipping\_by\_norm: 10.0
* eval\_config: num\_examples: 381
* max\_evals: 10

**1.4 Model evaluation**

The models were tested with test dataset (n=1,192) to measure sensitivity and specificity for detecting the risk classes. None of the images in the test dataset was used to train the model.

|  |  |  |
| --- | --- | --- |
| network type | sensitivity | specificity |
| ANALYZE | 91,2% | 94,9% |
| DETECT | 93,8% | 96,8% |

2. Case number planning

Assumptions for case number planning:

We assume 85% sensitivity for melanoma and squamous cell carcinoma, respectively, and 70% sensitivity for basal cell carcinoma.

No possible drop-outs are considered in case planning.

3. Sample size planning

With a sample size of 196, the sensitivity of 85% (70%) can be estimated with an accuracy of 5% (6.4%).

With a sample size of 323, the sensitivity of 85% (70%) can be estimated with an accuracy of 3.9% (5%).

Further, it should be considered whether to stratify by different skin types.

For the case number calculation, nQuery Advisor + nTerim 4.0 was used (two-sided 95% confidence interval for a proportion with normal distribution approximation).

4. Calculation “High and medium risk versus low risk”

a) Sensitivity:

"Analyze" mean: 97.9% (CI 96,.7-99.2)

"Detect" mean: 96.1% (CI 93.6-98.6)

b) Specificity:

"Analyze" mean: 98.6% (CI 97.6-99.6)

"Detect" mean: 97.7% (CI 95.9-99.5)