Dermatology Practical & Conceptual

Algorithm

The CNN, consisting of two neuronal networks, was trained with the help of a region-based CNN (faster RCNN [19]) using a dataset that consisted of various skin lesions as well as normal structures that could mimic pathologic lesions. The machine learning algorithms are used through the Classifier and Object Detection components by the user interface in the assessment scan and result screens. The machine learning algorithm in the application receives an input image through the user interface and device camera. There are two separate machine learning algorithms used. The first is an image classifier algorithm based on the inception_v3 model; the second is an object detection algorithm based on the faster_rcnn_inception_v2_coco model. Both models were retrained on the skin lesion training dataset. The image classifier algorithm is used in real-time during the assessment scan screen and the camera test screen. The algorithm processes the live image from the camera using TensorFlow to identify skin and only then enables the taking of an image. If the real-time classifier does not identify skin, it is not possible to take a picture. When any lesion is detected by the algorithm, the scan indicator turns green and allows the user to start the scan. The image classifier algorithm is not used in risk assessment, only in the scan screen and camera test screen. The object detection algorithm returns a set of lesion objects detected in that image. The lesion objects contain the lesion category label, the lesion box coordinates in the image, and the score that represents the probability of the algorithm's having detected the lesion. The score is a float value between 0.0 to 1.0, where a score of 0.0 means the algorithm detected no probability of the lesion object, and a score of 1.0 means the algorithm detected a 100% probability of the lesion object. The assessment algorithm uses the object detection algorithm results to present the user with a risk classification (high, medium, low). The assessment algorithm analyzes the object detection algorithm results through several additional steps to determine the risk classification.

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 one second and can also be used in real-time 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 of 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 dataset of 19,576 anonymized images was used. It was split (the training script splits the data randomly) into a training and validation dataset (18,384 dermatologists labeled images and a test dataset (n=1,192) in order to ensure that test dataset images are not used in both training and validation of the model the filenames of the test dataset images start with tds*.jpg.

The image dataset (n=19,576) contains anonymized representative (age of participants: 18 to 86 years, Fitzpatrick skin type 1–4, a clear contrast with surrounding skin, not covered by hair or opaque/glittering substances, not previously traumatized except for the label "skin injury") images.

The test dataset (n=1,192), which was not used for training, contains anonymized representative images (age of participants: 18 to 86 years, Fitzpatrick skin type 1–4, a clear contrast with 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 dataset for CENSORED iterations. The training process runs through the training image dataset 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 detection of the risk classes. None of the images in the test dataset was used to train the model.

1.5 Network Type

Analyze: Sensitivity: 91.2%; Specificity: 94.9%
Detect: Sensitivity: 93.8%; Specificity: 96.8%

2. Case Number Planning

Assumptions for case number planning:

We assume 85% sensitivity for melanoma and squamous cell carcinoma 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*
&quot;Analyze&quot; mean: 97.9% (95% CI: 96.7–99.2)
&quot;Detect&quot; mean: 96.1% (95% CI: 93.6–98.6)

b. *Specificity*
&quot;Analyze&quot; mean: 98.6% (95% CI: 97.6–99.6)
&quot;Detect&quot; mean: 97.7% (95% CI: 95.9–99.5)

Sample Size Estimation

To ensure the best possible accuracy and precision of the results, a weighted number of cases is used, with a statistically required number of lesions from each of the three risk categories. The sample size was estimated using the software nQuery. To verify the non-inferiority of the app in terms of risk assessment of lesions (sensitivity), the non-inferiority margin was set at 90% based on a requirement of the US

Food and Drug Administration (FDA). With a sample size of 399 lesions using a one-sided exact test (alpha=2.5%) and assuming that a sensitivity of 93.94% is observed, a power of >80% can be achieved. In the previous clinical study to test the diagnostic accuracy of the app before market placement, a sensitivity of 96.4% (95% CI: 93.94%–98.85%) was observed. The lower limit of the confidence interval (93.94%) was used for case number planning. The resulting sample size (399 lesions) represents the weighted number of cases assigned to the "therapy" category (medium risk and high risk) by the expert panel. If this applies to approximately 27% of the images/lesions, a total sample size of 1428 images/lesions is required, which also includes the category "no therapy".