1 Description

Dataset name	Skin Lesion Images for Melanoma Classification (ISIC)
Link	Kaggle
License	CC BY-NC 4.0
Medical discipline	Dermatology
Medical procedure	Dermatoscopy
Multi-class problem	✓
Multi-label problem	✗

This dataset contains the training data for the ISIC 2019 challenge, note that it already includes data from previous years (2018 and 2017). The dataset for ISIC 2019 consists of 25,331 images available for the classification of dermoscopic images among nine different diagnostic categories:

- Melanoma
- Melanocytic nevus
- Basal cell carcinoma
- Actinic keratosis
- Benign keratosis (solar lentigo / seborrheic keratosis / lichen planus-like keratosis)
- Dermatofibroma
- Vascular lesion
- Squamous cell carcinoma
- None of the above

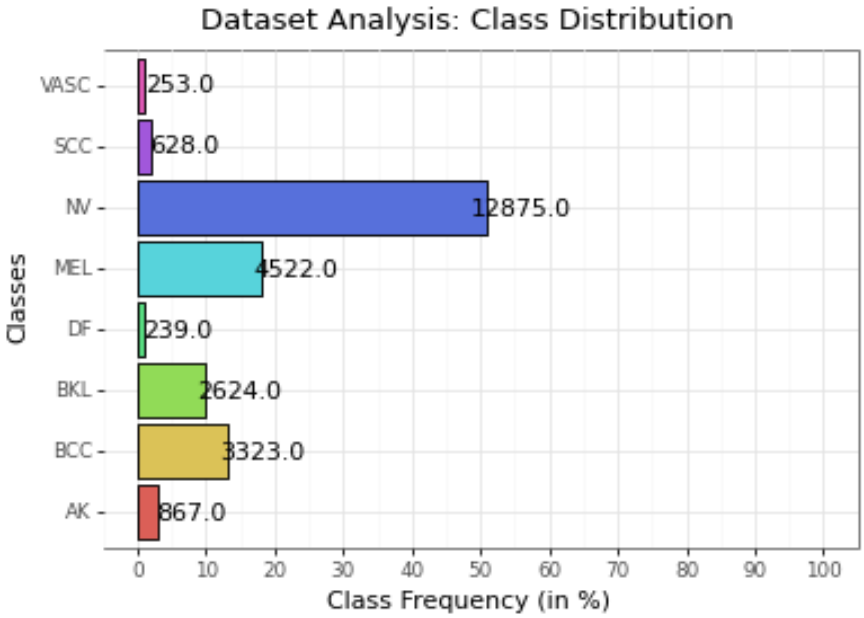


Figure 1: Class distribution: Skin Lesion Images for Melanoma Classification (ISIC)

2 Pre-processing

Since the *unknown* class does not have any samples in this dataset, it is removed from the csv file. No other pre-processing was done.

3 Training

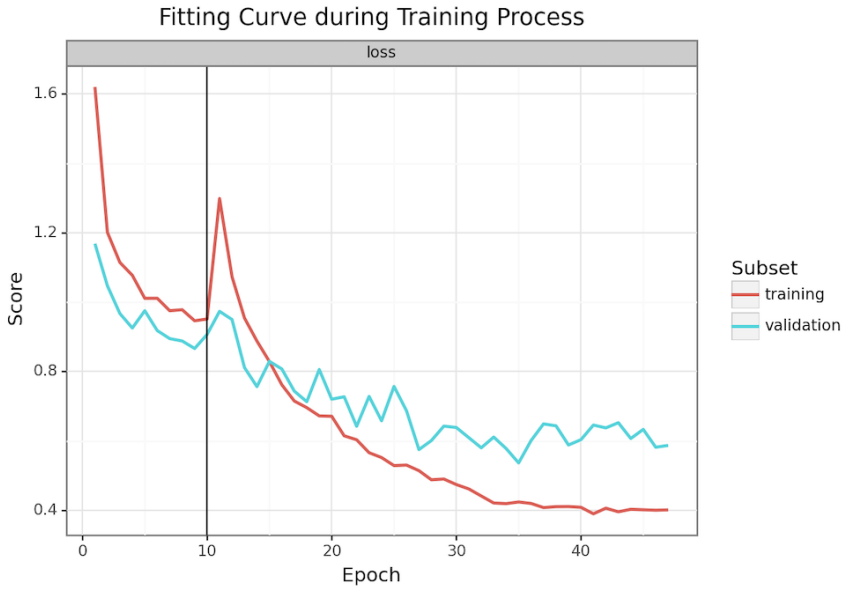


Figure 2: Training: Skin Lesion Images for Melanoma Classification (ISIC)

Due to transfer learning, the model's weights were not adjusted during the first ten episodes; merely its classification head was trained. From episode ten on, the model's weights were unfrozen and the model was trained as a whole. The unfreezing of the weights result in a sudden peak in training and validation loss. That reflects, that the model slightly re-adjusts all of its weights. Very soon validation loss is higher than training loss. In itself, that does not indicate overfitting, but merely means that generalization is challenging for the model. However, the slight increase in validation loss from episode 35 on suggests a tendency to overfit.

4 Results

Figure 3: Metric overview: Skin Lesion Images for Melanoma Classification (ISIC)

Figure 3 shows an overall solid performance with acceptable F1 scores. Due to discrepancies in precision and sensitivity, we will discuss these two metrics in detail.

Precision, formally called *positive predictive value* $PPV = \frac{TP}{TP+FP}$, indicates how many of the classified samples actually belong to this class. Sensitivity, also *true positive rate*, which is defined as $TPR = \frac{TP}{TP+FN}$, expresses how many of a class' samples were correctly identified as such.

Both metrics show high values for classes VASC, NV and BCC. The most significant discrepancy between the two metrics occurs for class DF: Precision only reaches approximately 41 %, whereas its sensitivity lies close to 80 %; specificity reaches almost 100 %.

This discrepancy can easily be explained, if we consider the class distribution, using figure 1. It clearly shows, that DF and VASC classes have less than 300 samples, whereas NV contains more than 12,000 samples. That affects the scores, because precision sets *true positives* in relation to *false positives*. Naturally, there are more samples in absolute numbers for the NV class, than for DF, which results in a low precision score for DF. The high sensitivity score however illustrates, that this class can be identified reliably. Specificity with almost 100 % confirms that.

Despite their low sample size, VASC class seems to be easily identifyable for our model: It shows a good F1 score and an impressive ROC curve (see violet curve in figure 4).

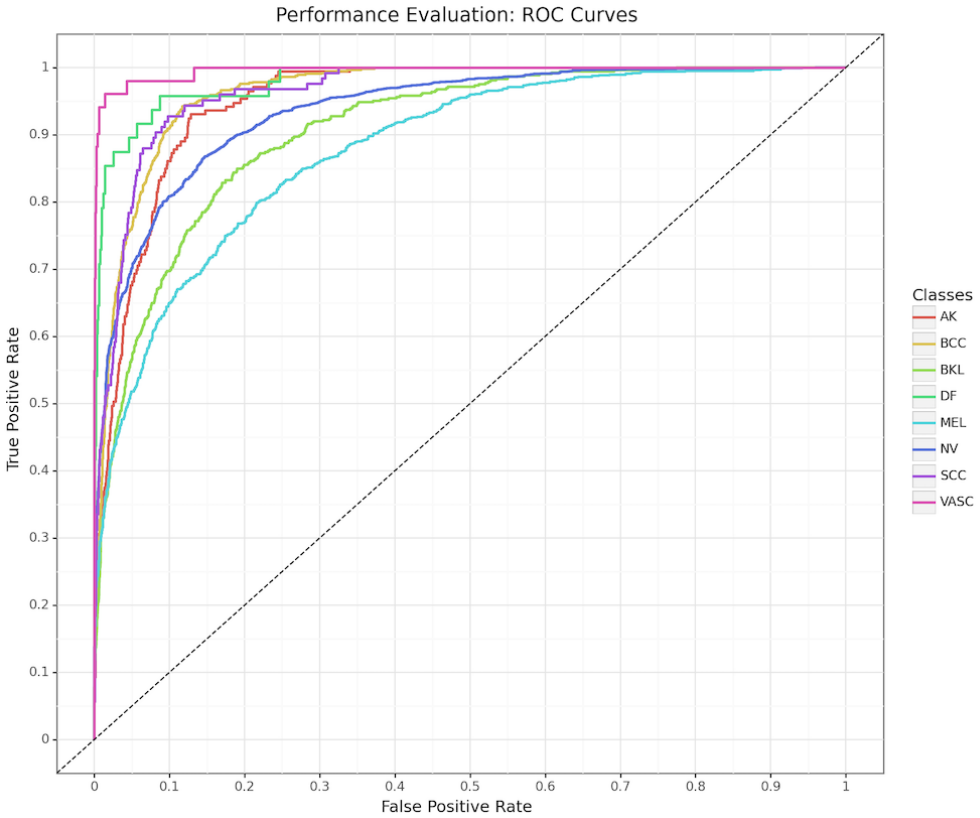


Figure 4: ROC curve: Skin Lesion Images for Melanoma Classification (ISIC)

Figure 5: Confusion matrix: Skin Lesion Images for Melanoma Classification (ISIC)

5 XAI

To gain insights into our model's decision making process, we generated Grad-CAM images.

One thing, that is fairly obvious, if we take a look at the image row displaying image **ISIC_0033808**, is the high confidence in correctly predicting the DF class. The Grad-CAM image shows, that the to our model highly relevant segment lies slightly above the mole. It seems like the light patches within the mole are explicitly ignored. We notice a similar phenomenon for images **ISIC_0026804**, **ISIC_0056787** and **ISIC_0069544**, all displaying correctly classified images. Our model decides to assign the correct classes based on the edges of the suspicious looking skin patches, instead of the whole patches themselves.

As a good example for a wrong classification with high confidence, we should inspect image **ISIC_0032604**. If we consult the confusion matrix, we see that more than 9 % of all MEL images are actually classified as BKL. Based on a single occurrence, it is hard to understand the relation between the two classes. Further steps could include generation of more Grad-CAM images for classes MEL and BKL.

Image: ISIC_0025526

Class: BKL

Classified as: BKL (33.7 %)

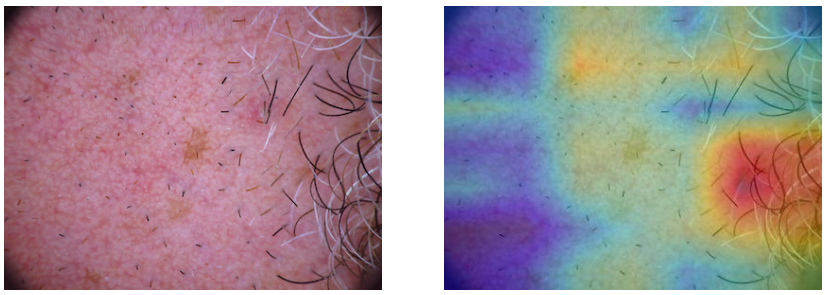


Image: ISIC_0026804

Class: NV

Classified as: NV (53.3 %)

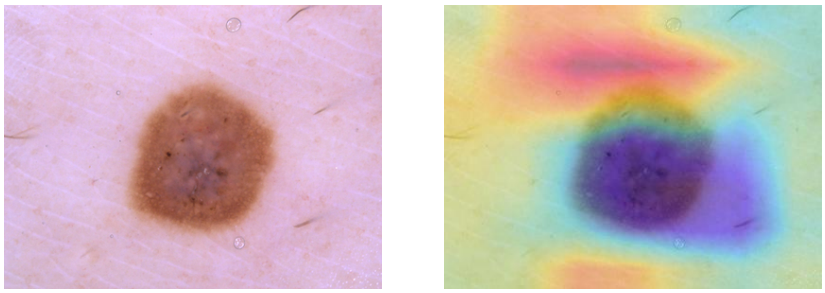
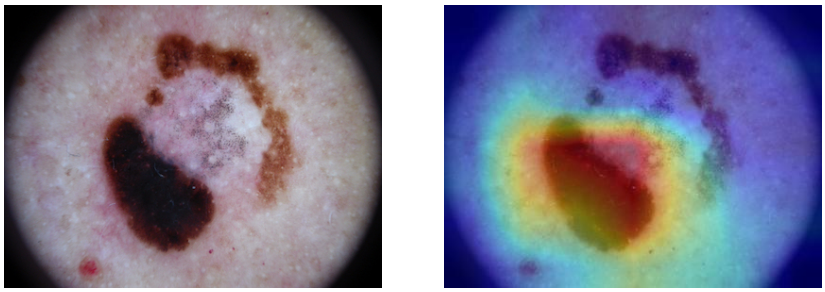


Image: ISIC_0032604

Class: MEL

Classified as: BKL (76.8 %)



6 Summary

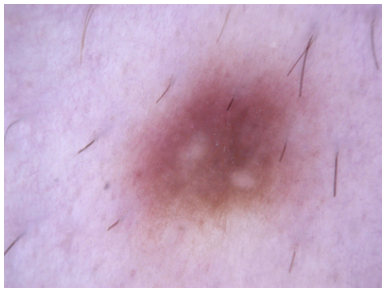
ISIC is a good example for a biased dataset. AUCMEDI's metric overview 3 contains not only the widely popular F1 score, but also precision and sensitivity, building blocks for the F1 score. That allows users to visually analyse, how the F1 score is composed. In combination with the class distribution, AUCMEDI helps us understand our model's performance.

By comparing the original sample images with their Grad-CAM images, we are able to gain insights into our model's decision making process. We can, for instance, learn which parts of a skin patch are more relevant than others. That may even open up new perspectives in medical research. Nevertheless, interpretation of Grad-CAM images are challenging, especially when trying to understand an incorrect classification.

Image: ISIC_0033808

Class: DF

Classified as: DF (98.6 %)



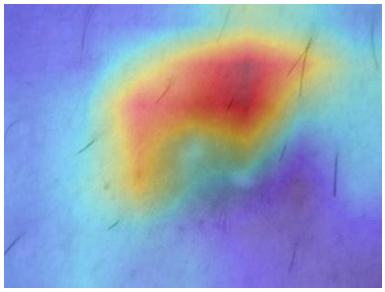
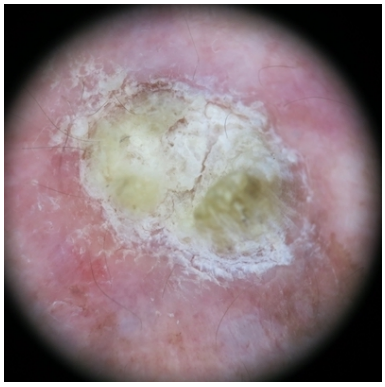


Image: ISIC_0056787

Class: SCC

Classified as: SCC (66.9 %)



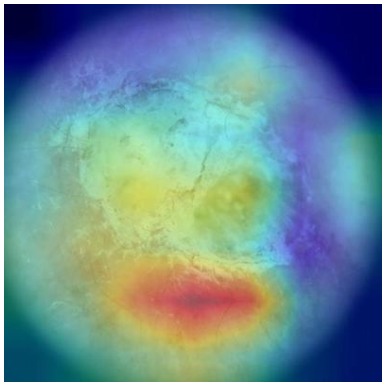


Image: ISIC_0069544

Class: AK

Classified as: AK (76.8 %)

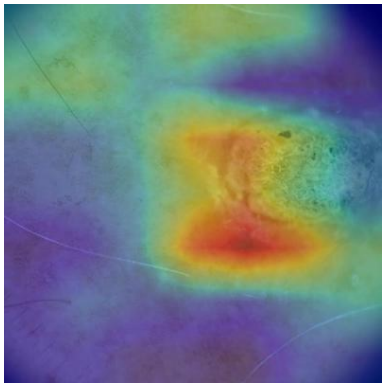
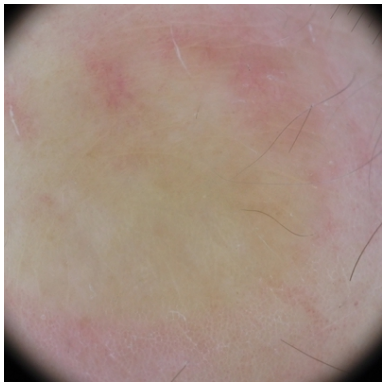


Image: ISIC_0070103

Class: VASC

Classified as: BCC (32.1 %)



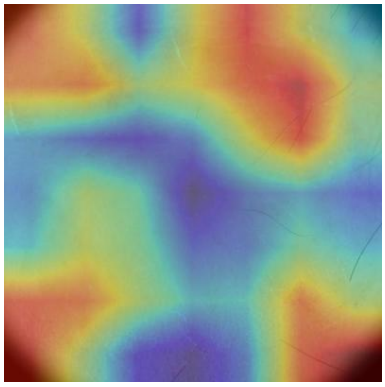


Image: ISIC_0072427

Class: BCC

Classified as: BCC (81.0 %)