The Effects of Masking in Melanoma Image Classification with CNNs Towards International Standards for Image Preprocessing

Fabrizio Nunnari[1,2(✉)], Abraham Ezema[1,2], and Daniel Sonntag[1,2]

[1] German Research Center of Artificial Intelligence, Kaiserslautern, Germany
{Fabrizio.Nunnari,Abraham_Obinwanne.Ezema,Daniel.Sonntag}@dfki.de
[2] Oldenburg University, Oldenburg, Germany
http://www.dfki.de/iml

Abstract. The classification of skin lesion images is known to be biased by artifacts of the surrounding skin, but it is still not clear to what extent masking out healthy skin pixels influences classification performances, and why. To better understand this phenomenon, we apply different strategies of image masking (rectangular masks, circular masks, full masking, and image cropping) to three datasets of skin lesion images (ISIC2016, ISIC2018, and MedNode). We train CNN-based classifiers, provide performance metrics through a 10-fold cross-validation, and analyse the behaviour of Grad-CAM saliency maps through an automated visual inspection. Our experiments show that cropping is the best strategy to maintain classification performance and to significantly reduce training times as well. Our analysis through visual inspection shows that CNNs have the tendency to focus on pixels of healthy skin when no malignant features can be identified. This suggests that CNNs have the tendency of "eagerly" looking for pixel areas to justify a classification choice, potentially leading to biased discriminators. To mitigate this effect, and to standardize image preprocessing, we suggest to crop images during dataset construction or before the learning step.

Keywords: Skin cancer · Convolutional neural networks · Masking · Reducing bias · AI standardization roadmap · Preprocessing

1 Introduction

As reported in the 2019 USA cancer statistics, skin diseases have been steadily increasing over the years, whereby skin cancer represents 7% of the total cancer cases. As of 2019, there were 104,350 expected cases of skin cancer, of which 96,480 were melanomas. The importance of promptly detecting skin cancer is evident from the high percentage of survival (92%) after surgery resulting from early detection [19].

J. Ye et al. (Eds.): MobiHealth 2020, LNICST 362, pp. 257–273, 2021.
https://doi.org/10.1007/978-3-030-70569-5_16

The classification of skin lesions using computer vision algorithms has been a subject of recent research [6,10,13]. One of the breakthroughs being the publication of Esteva et al. [8], reporting a better performance than expert dermatologists using transfer learning on a deep convolutional neural network (CNN). The network was first trained on a set of about one million diverse images, and then fine-tuned with more than 100k images of skin lesions.

Given the promising progress of computer vision algorithms in aiding skin lesion classification, the ISIC[1] (International Skin Imaging Collaboration) hosts a competition for the automated analysis of skin lesions. In the years 2016 [12], 2017 [5], and 2018 [4], the challenge included three tasks: segmentation, attribute extraction, and classification. These tasks replicate the procedure usually followed by dermatologists: to identify the contour of the skin lesion, highlight the areas in the lesion that suggest malignancy, and classify the specific type of lesion.

To accomplish these tasks, the ISIC challenge provides a public dataset that has grown from 900 images as of 2016 to more than 33,000 images for the 2020 edition. This is the largest publicly available dataset of dermoscopic images, and is widely used by many researchers throughout the world.

Masking skin lesion images, i.e., using segmentation to remove the pixels pertaining to the healthy skin and retaining the pixels belonging to the lesion, is an image pre-processing technique that is supposed to help the classification of skin lesions by removing unneeded, unwanted image artifacts.

In fact, Winkler et al. [23] found that the presence of *gentian violet* ink, often used by dermatologists to mark the skin in proximity to suspicious lesions, can disrupt the correct classification and lower the specificity of commercial DSS (Diagnosis Support Systems). Moreover, recently, Bissoto et al. [2] found a strong bias in the ISIC dataset; by completely removing 70% of the central part of the images (hence removing the totality of pixels containing the skin lesions), the CNN model was still able to reach 0.74 AUC (with respect to 0.88 AUC reached with full images). This suggests a strong bias of the dataset at its borders.

To date, while there seem to be clear advantages of masking out the skin surrounding the lesion area, it is not clear to what extent masking images influences (positively or negatively) the quality of classification (e.g., by removing bias). And what are other consequences for the process of training classifiers?

In this paper, we present a further investigation on image masking by, first, assessing the presence of biases at the dataset images' borders, and, second, comparing the classification performances when applying several types of masks. Third, we analyse the bias patterns through a visual inspection of Grad-CAM saliency maps [18]. This analysis employs four types of *masks* (see Fig. 1):

1. Rectangular Mask (RM) removes 30% of the image surface around the border. This is a direct contrast to the masking utilized by Bissoto et al. [2] to show the presence of bias at the borders and its influence on model performance. With this masking type, we verify whether removing the border affects the performance of a classifier.

[1] https://www.isic-archive.com/.

2. Circular Mask (CM) draws a circle at the middle of images. Here, we evaluate if removing the corners of the images and inspecting only its central part retains model performance.
3. Full Mask (M) reveals only the lesions and a fraction of the surrounding skin. It is used to reveal whether completely removing the skin surrounding a lesion improves prediction performance.
4. Finally, a rectangular cropping (CR) of the image is applied, which removes the image borders and increases the quantity of information passed to the classifiers.

In the rest of this paper, we conduct experiments on three popular skin lesion image datasets (ISIC 2016, ISIC 2018, and MedNode), each evaluated through a 10-fold cross validation approach to reduce biases by randomization. All the maskings were implemented using a dedicated U-NET (de-)convolutional neural network, following the procedure described in [14].

With the standardization roadmap for artificial intelligence, a comprehensive analysis of the current state of and need for international standards and specifications has been published [22]. Data bias and the bias of classifiers is a key factor. As a result of our experiments, we suggest to crop images during dataset construction or before the learning step, towards a process to standardize image preprocessing in CNN contexts.

Section 2 gives an overview of related work and on the importance of masking to avoid biases. Section 3 describes the method used to train and test the data material. Section 4 describes the experiments measuring the difference in performances among different masking conditions. Section 5 reports on the analysis of our results with the help of saliency extraction (visual explanation). Section 6 discusses the results, and Sect. 7 concludes the paper.

2 Related Work

The classification of skin lesions through the use of CNNs has increased in popularity since the publication of Esteva et al. [8]. Their CNN-based model matched the performance of experienced dermatologists. To this end, all performant neural-network-based solutions for skin lesion classification are based on a transfer learning approach [21]: a baseline deep CNN is pre-trained for example on the ImageNet dataset [7], and the transfer-learning steps consists of substituting the final fully-connect layers of the network with a few randomly initialized ones, then to continue training the model on skin lesion images. In our work, we perform transfer learning using pre-trained versions of VGG16 [20].

Rather than focusing on benchmarks [10], our goal in this contribution to investigate the change of performance between using plain images and segmented or normalized ones for the classification task. To train our reference classifier, we rely on three publicly available datasets: ISIC 2016 [12], ISIC 2018 [4], and MedNode [9]. All of them were used to train several models, each on a number masking methods, as later explained in Sect. 3.

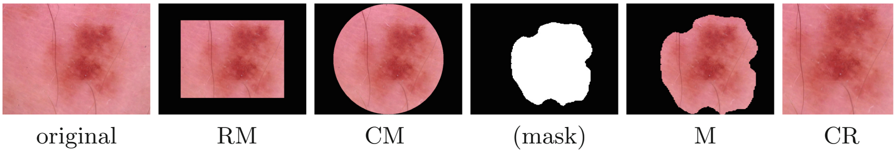

Fig. 1. Masking examples, from left to right: the original full image (ISIC_0024307), rectangular mask (RM), circular mask (CM), the segmentation mask, the segmented image (M), the image cropped on the mask bounding box (CR).

Burdick et al. [3] performed a systematic study on the importance of masking images used for training CNN models. They compared the performance of the CNN model using the full images compared to applying the masks on several levels: from fully masking out the surrounding skin to exposing some portion of the skin surrounding the lesion. Tests show best results when only a limited portion of the surrounding skin is kept for training. The hypothesis is that masking the healthy skin helps in classification while showing all the healthy skin in the image "confuses" the network, that is, it becomes more probable that the network learns image artefacts. Following the results in Burdick et al. [3], for each image, we also extend the lesion mask from segmentation to 110% of its original area, in order to expose a bit more of the surrounding skin areas during training than the original mask shows.

Binary masking of an image defines a black/white area within it, whereas white is associated with the pixels of interest, and black is associated with non-interesting or the confounding part of the image to be discarded in subsequent processing steps. This segmentation techniques have been significantly improved by the use of deep learning models. Ronneberger et al. [17] first proposed the application of the convolution-deconvolution network (U-Net) for medical image segmentation. The U-Net architecture applies stacks of convolutional layers with downsampling to extract latent image features and deconvolutional layers with upsampling within the network. This method of segmentation has been very successfully applied to medical image segmentation.

Variants of this model have shown to be very effective in the ISIC segmentation challenge in the past, with a Jaccard index score of 0.765 and 0.802 in the ISIC2017 and ISIC2018 editions, respectively, see [1,16]. In this paper, we implement a transparent segmentation model to show the effects of masking in melanoma images by using the approach described in [14] and using the data provided for Task 1 of the ISIC 2018 challenge [4].

3 Method Overview

Figure 2 illustrates the method we follow to test the effectiveness of the different masking conditions on prediction performance. The method is composed of three phases: preparation of the segmentation model, masked images construction, and training of the classification models. They are discussed in more detail in the following.

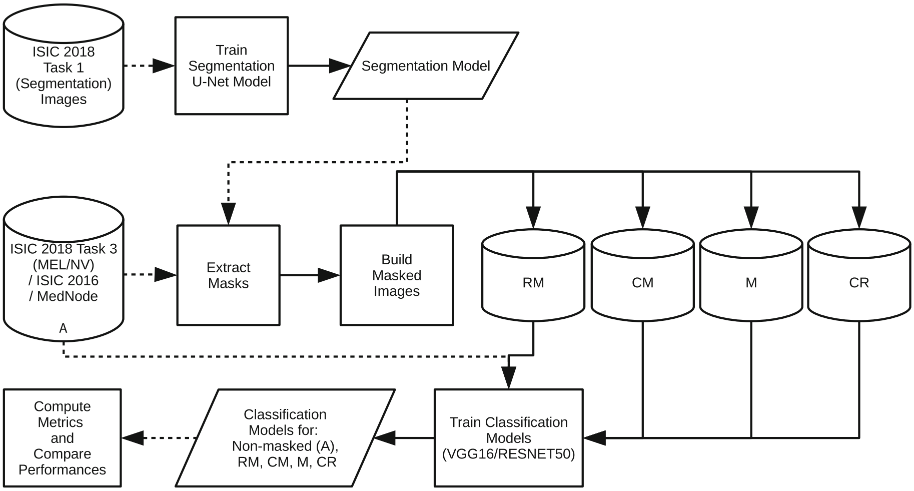

Fig. 2. Methodology overview. The top blocks depict the training of the segmentation model. The middle blocks are related to the preparation of the masked images, and the bottom blocks represents the training of the classification models.

3.1 Segmentation Model

We utilize the images from Task 1 of the ISIC 2018 to train a masking model based on the U-Net architecture [17]. This dataset comprised of 2594 RGB skin lesion images, and for each sample, the ground truth is a binary mask in the same resolution as the input image.

Figure 3 shows the U-Net architecture together with a sample input and output (binary mask). The architecture is composed of 9 convolution blocks, where each of them is a pair of 2D *same* convolution with a kernel size of 3 × 3 × 3. Downsamplig is the result of a max-pooling with size 2 × 2. Upsampling is the result of a 2 × 2 transposed 2D *same* convolution. After each upsampling step, the convolution is performed on the concatenation of the upsampling result and the output of the downsampling with corresponding resolution. The initial number of filters (32) doubles at each downsampling. For this work, we used an input/output resolution of 160 × 160 pixels.

3.2 Masked Image Datasets

The segmentation model described above is used to extract *masks* for Melanoma and Nevus images of the ISIC 2018 Task 3 [4], ISIC 2016 [12], and MedNode [9] datasets. From ISIC 2018 Task 3, we selected only nevus (NV) and melanoma (MEL) classes because these are the exactly the same classes which used as ground truth for the masks in Task 1. After an initial visual inspection, we realized that applying the mask prediction to any of the other 6 classes of the Task 3 dataset often leads to erroneous results due to the very different nature of the lesions.

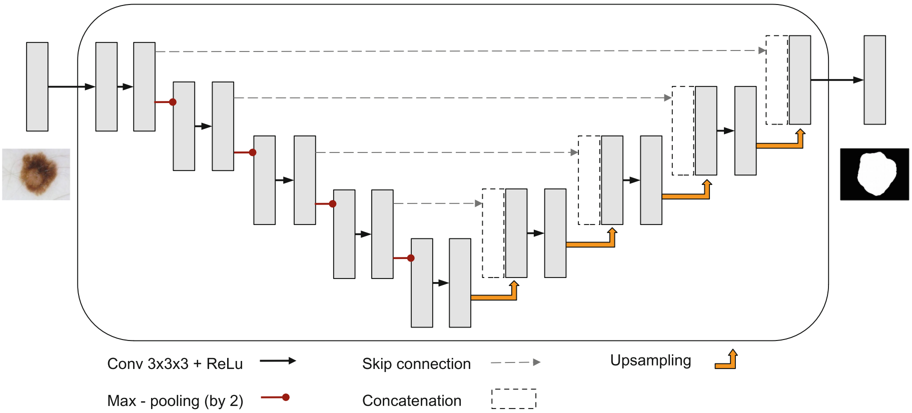

Fig. 3. The U-Net architecture used for lesion segmentation. The input image is 3-channel RGB, while the output image is 1-channel gray-scale with the same resolution.

In total we define five sets of pre-processed images: **A** (the full image, containing all of the pixels), and the four types already described in the introduction **RM**, **CM**, **M**, and **CR**. See Fig. 1.

The M and CR datasets are obtained through the composition between the original images and the extracted masks. For the M dataset, the mask is first scaled around its center by a factor of 1.1 to reveal a portion of the surrounding skin (as suggested by [3]). The composition setting converts all pixels outside the lesion mask to black. For the CR dataset, the mask is utilized to identify a rectangular cropping region containing the lesions contour.

The CR datasets contains a few less samples than the others, because an initial inspection revealed that the masks of samples with a thin lesion–foreground pixel variation result in very small (mostly inaccurate) lesion blobs. Hence, we automatically filtered these defective images from our samples based on an automated comparison between the area of the masks and the total image size. Images whose mask areas was less than $\frac{1}{8}$ of the picture were discarded.

3.3 Binary Classifiers

For each of the 3 datasets and 5 masking conditions, we trained 10 binary classification models using a 10-fold splitting strategy. Each fold was composed using 10% of the dataset for testing and another random 10% for validation. While splitting, we ensured to preserve the proportion between classes. In the rest of this paper, we report the mean and the standard deviation among the 10 folds.

The performance of the binary classifiers in discriminating *nevi* from *melanomas* are reported in terms of accuracy, specificity, sensitivity, and ROC AUC (Receiver Operating Curve - Area Under the Curve) on the test set, where the positive case is associated with the malignant melanoma.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
input_1 (InputLayer)         (None, 227, 227, 3)       0
_________________________________________________________________
block1_conv1 (Conv2D)        (None, 227, 227, 64)      1792
_________________________________________________________________
block1_conv2 (Conv2D)        (None, 227, 227, 64)      36928
_________________________________________________________________
block1_pool (MaxPooling2D)   (None, 113, 113, 64)      0
_________________________________________________________________
[... 13 more layers ...]
_________________________________________________________________
block5_conv3 (Conv2D)        (None, 14, 14, 512)       2359808
_________________________________________________________________
block5_pool (MaxPooling2D)   (None, 7, 7, 512)         0
_________________________________________________________________
flatten (Flatten)            (None, 25088)             0
_________________________________________________________________
fc1 (Dense)                  (None, 4096)              102764544
_________________________________________________________________
dropout_1 (Dropout)          (None, 4096)              0
_________________________________________________________________
fc2 (Dense)                  (None, 4096)              16781312
_________________________________________________________________
dropout_2 (Dropout)          (None, 4096)              0
_________________________________________________________________
predictions (Dense)          (None, 3047)              12483559
=================================================================
Total params: 146,744,103
Trainable params: 146,744,103
Non-trainable params: 0
_________________________________________________________________
```

Listing 1.1. An excerpt of the VGG16 architecture used for the binary classification task.

As already successfully employed in previous research (e.g., [8]), all of the binary classifiers are based on the transfer learning approach [21] with CNNs. The base CNN model is the VGG16 [20] architecture pre-trained on ImageNet [11]. We then substituted the original three final fully connected layers with a sequence of two fully connected layers, each followed by a dropout of 0.5, and a final 2-class discrimination softmax layer. Listing 1.1 shows an excerpt of the architecture.

Each model was trained for a maximum of 100 epochs and optimized for accuracy. Input images were fed to the network with an 8× augmentation factor, where each image was horizontally flipped and rotated by 0, 90, 180, and 270 degrees. To avoid the generation of black bands, images were rotated after scaling to the CNN input resolution. Class imbalance was taken into account using a compensation factor in the loss-function (parameter `class_weight` in the `fit` method of the Keras framework). For each model, we also report what epoch returned the most accurate model.

All training was performed on Linux workstations using our toolkit for Interactive Machine Learning (TIML)[2], which uses the Keras[3] (v2.2.4) framework with Tensorflow[4] (v1.13.1) as backend. Our reference Hardware is an 8-core Intel 9th-gen i7 CPU with 64 GB RAM and an NVIDIA RTX Titan 24 GB GPU.

4 Experiments

The following three sections report details on the analysis performed on the three datasets: ISIC2016, MedNode, and ISIC2018. For each dataset, the metrics of the binary classification are reported for each of the five masking conditions described before: A, RM, CM, M, and CR. The analysis focuses on determining a potential bias from the border of the images.

4.1 ISIC2018

Table 1 shows the distribution of the samples in the ISIC2018 dataset. Training a full model (6256 samples, 100 epochs) takes about 9 h on our reference hardware. Table 2 show the results of the tests.

Table 1. Distribution of the 7818 images from the ISIC2018 dataset.

conditions	samples	MEL	NV	train	val	test
A, RM, CM, M	7818	1113 (14.2%)	6705 (85.8%)	6256	781	781
CR	7645	1099 (14.3%)	6546 (85.9%)	6119	763	763

In order to measure the statistical significance of the metric among conditions, we run a set of t-tests for independent samples between the no-mask condition (A) against all the others. The results of the test are reported in Table 3. The tests compare the results across the 10 folds (N = 10). The table reports the compared conditions, followed by the different statistic metrics, their absolute and relative difference, and the significance code for the p-value (+: $p < .1$; *: $p < .05$; **: $p < .01$; ***: $p < 0.001$).

When applying a rectangular mask, the results show a significant reduction on almost all metrics. For example, accuracy drops by 2.99%. Also circular masks and full masking decrease accuracy by 2.85% and 4.37%, respectively. Only cropping shows some improvewd accuracy values. Although not significant, we report a positive tendency of 4.73% increase in sensitivity.

This results suggest that there is indeed a bias in the surrounding skin; the other explanation is that exposing a large portion of the surrounding skin helps

[2] https://github.com/DFKI-Interactive-Machine-Learning/TIML.

[3] https://keras.io/.

[4] https://www.tensorflow.org/.

Table 2. Results of the test on the ISIC2018 dataset.

set	testacc	testspec	testsens	testauc	epch
A	.909 (.014)	.933 (.017)	.763 (.062)	.948 (.010)	90.3 (6.7)
RM	.882 (.011)	.899 (.018)	.781 (.059)	.937 (.011)	41.8 (3.9)
CM	.883 (.017)	.899 (.021)	.789 (.066)	.938 (.012)	41.5 (6.1)
M	.870 (.013)	.884 (.014)	.785 (.034)	.930 (.011)	39.0 (10.7)
CR	.911 (.014)	.929 (.017)	.799 (.057)	.955 (.010)	40.7 (7.8)

Table 3. Significant differences between masking conditions in the ISIC2018 dataset.

Condition	Metric	Difference	Diff. pct	Signif.
A vs RM	ACC	−0.027	−2.99%	***
A vs RM	SPEC	−0.035	−3.73%	***
A vs RM	AUC	−0.011	−1.21%	*
A vs RM	EPOCH	−48.5	−53.71%	***
A vs CM	ACC	−0.026	−2.85%	**
A vs CM	AUC	−0.1	−1.11%	+
A vs CM	EPOCH	−48.8	−54.04%	***
A vs M	ACC	−0.04	−4.37%	***
A vs M	SPEC	−0.05	−5.33%	***
A vs M	AUC	−0.018	−1.90%	**
A vs M	EPOCH	−51.3	−56.81%	***
A vs CR	SENS	0.036	4.73%	0.2152
A vs CR	EPOCH	−49.6	−54.93%	***

in the classification to some extent. For the cropping condition, such deficiency might be compensated by higher quantity of information passed to the neural network. In fact, when the image is cropped, almost all of the 277 × 277 pixels of the image are covered by the lesion–hence increasing the quantity of detail attributed to the skin.

A common aspect across all our comparisons is the significant and consistent drop (more than 50%) of the number of epochs needed to train the model.

4.2 MedNode

Table 4 shows the distribution of the samples in the MedNode dataset. Training one fold of the full dataset (ca. 136 samples, 100 epochs) takes about 15 minutes on our reference hardware. Table 5 show the results.

In comparison to A, we observed a considerable decrease in the performance when applying a rectangular mask, e.g., accuracy −0.053 (−6.58%), and mild loss in performance for all other conditions. However, none of the differences is

Table 4. Distribution of the 170 images of the MedNode dataset.

conditions	samples	MEL	NV	train	val	test
A, RM, CM, M	170	70 (41.2%)	100 (58.8%)	136	17	17
CR	169	70 (41.4%)	99 (58.6%)	137	16	16

Table 5. Results of the test on the MedNode dataset.

set	testacc	testspec	testsens	testauc	epch
A	.806 (.123)	.870 (.100)	.714 (.181)	.869 (.131)	34.1 (12.9)
RM	.753 (.094)	.800 (.118)	.686 (.189)	.860 (.073)	36.1 (23.5)
CM	.818 (.140)	.830 (.135)	.800 (.194)	.890 (.114)	40.5 (21.3)
M	.806 (.112)	.830 (.090)	.771 (.214)	.880 (.111)	43.6 (24.7)
CR	.768 (.144)	.820 (.087)	.700 (.328)	.843 (.120)	55.9 (32.2)

significant according to our t-tests, likely because of the high variance in the measurements among the 10 folds from the limited number of samples.

4.3 ISIC 2016

Table 6 shows the distribution of the samples in the ISIC2016 dataset. Training one fold of the full dataset (722 samples, 100 epochs) takes about 1 h 30 m on our reference hardware. Table 7 show the results.

Table 6. Distribution of the 900 images of the ISIC2016 dataset. In the right columns, the mean of the number of samples among the 10 folds used for validation (standard deviation is ≤0.6).

conditions	samples	MEL	NV	train	val	test
A, RM, CM, M	900	173 (19.2%)	727 (80.8%)	722	89	89
CR	884	173 (19.6%)	711 (80.4%)	708	88	88

The only statistically significant difference stems from the specificity between the A and CM conditions (-0.026, -2.92%, $p < 0.1$). We can also observe a noticeable drop in sensitivity between A and CR conditions ($-0{,}092$, -22.18%), but it is not significant for our tests (p = 0.2352).

As for the MedNode dataset, the reduced number of samples led to a high variance during the cross-fold validation, making it thus impossible to validate the differences among conditions using our statistical method.

Table 7. Results of the on the ISIC2016 dataset.

set	testacc	testspec	testsens	testauc	epch
A	.806 (.028)	.898 (.031)	.416 (.163)	.773 (.074)	45.1 (38.4)
RM	.794 (.037)	.878 (.069)	.445 (.146)	.756 (.063)	49.2 (36.2)
CM	.788 (.026)	.872 (.030)	.432 (.151)	.790 (.059)	52.7 (26.9)
M	.784 (.035)	.866 (.065)	.445 (.181)	.774 (.066)	57.7 (28.7)
CR	.805 (.044)	.923 (.045)	.324 (.155)	.755 (.078)	46.3 (35.1)

5 Visual Inspection

In order to visually explain the characteristics that influenced model predictions, we leveraged the Grad-CAM method [18] to generate the saliency maps of *attention*. Figure 4 shows a nevus and a melanoma images from ISIC2018 and their relative attention maps on all masking conditions. All the saliency maps were extracted from the last convolutional layer of the VGG16 architecture (`block5_conv3`).

Two contrasting patterns emerge, thus giving additional details about the model's discrimination strategy. The saliency is higher on the skin lesion pixels (focused towards the center) for images correctly predicted as melanoma. In contrast, the saliency is higher on the skin pixels (towards the borders) in pictures correctly classified as nevus. The opposite happens when images are wrongly classified, with the attention for wrongly classified nevus towards the center and the attention for wrongly classified melanomas towards the border.

It is worth pointing out that the attention of the CNN moves towards the border regardless of the kind of masking strategy used. To systematically quantify this behaviour, we recorded the occurrence of this pattern in relation to the classification results, categorizing images according to whether the salient pixels are accumulated towards the (B)order or towards the (C) enter. The discrimination was made by a processing routine in terms of a pixel-level analysis. When the activation value for the pixel along the image borders (*left, top, bottom, right*) is very low (<0.1), then the image saliency map is considered as centered. For opposite cases, that is, when high activation values are present along the borders, an image-centred square patch covering $\frac{1}{16}th$ of the total image size is evaluated to confirm border images. As a result, when the patch is dominated by low activation values, a border case is recorded for the image, while a centered image is recorded for the opposite characteristic.

Table 8 shows the results on the ISIC2018 dataset. The observed behaviour (saliency is at the center for correct melanoma and wrong nevus, otherwise at the border) is prominent in the A, CM, and RM masking conditions, but less prominent for the M and CR conditions.

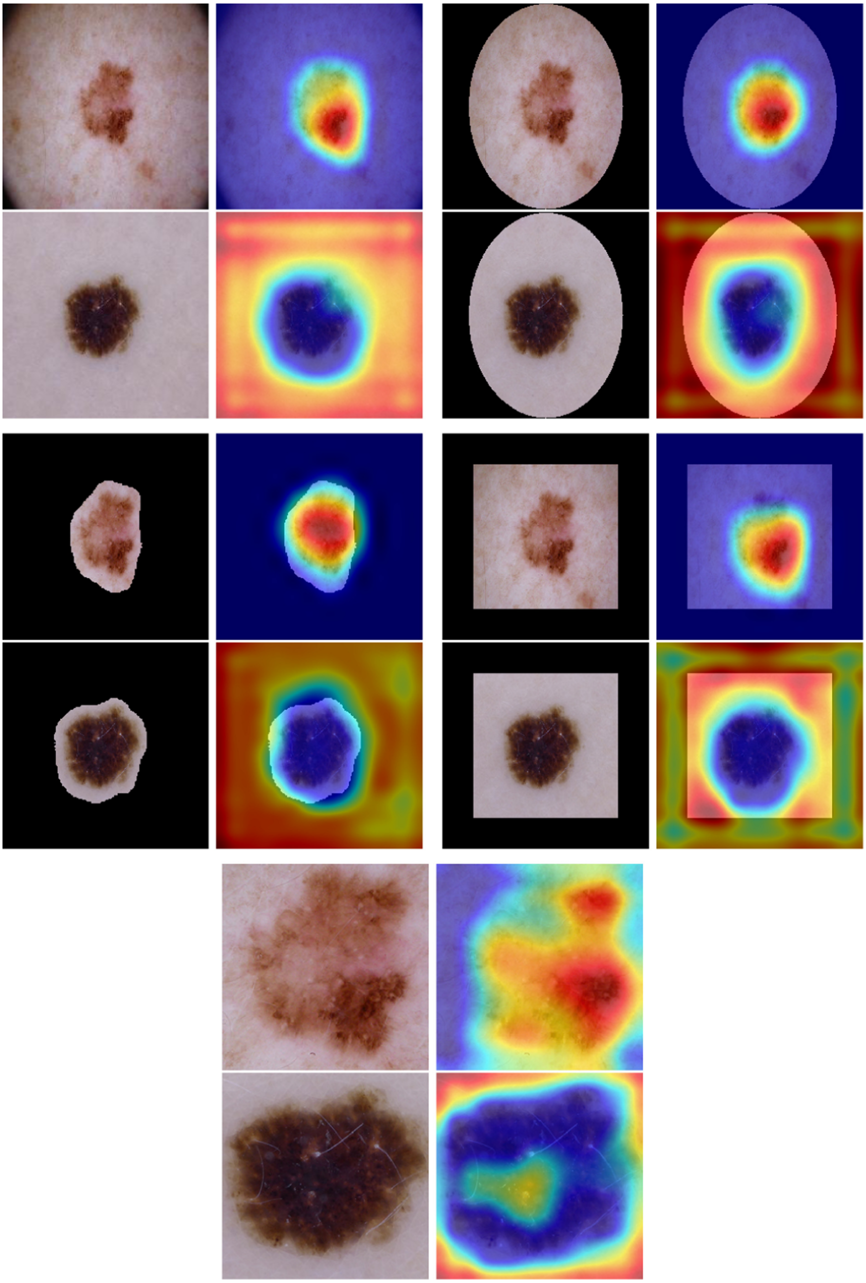

Fig. 4. ISIC 2018: colored saliency maps (aka heatmaps) extracted by Grad-CAM from the last convolution layer of the VGG16 model. The heatmaps are relative to one melanoma (top) and one nevus (bottom), both correctly classified. For each of the five masking conditions we show both the input image and its composition with the heatmap. Notice that for melanoma the heatmaps concentrate towards the center, while for nevus the model focuses on the border, regardless of the mask type.

Table 8. Results of the automatized saliency map inspection: counts of images with saliency map concentration at (B) order or (C), divided for Melanoma (MEL) and Nevus (NV), further split in (C)orrectly or (W)rongly classified.

ISIC 2018					
Mask	Concentration	MEL-C	MEL-W	NV-C	NV-W
A	B	54	258	5430	20
	C	795	6	830	425
CM	B	0	235	5974	0
	C	878	54	0	677
RM	B	0	244	6024	0
	C	869	0	2	679
M	B	163	219	2986	71
	C	711	20	2938	710
CR	B	325	185	3599	162
	C	554	36	2483	302

6 Discussion

Here we summarize our observations on the use of the different masking conditions arising from the classification results and from the visual inspections. As the low number of samples does not lead to statistically significant results for the MedNode and ISIC2016 datasets, we focus our analyses on the results obtained on the ISIC 2018 dataset.

From the classification results (Sect. 4.1) it is clear that masking the images affects the overall performance, likely because this eliminates any biases of the image borders. Moreover, we can notice a slight improvement in the sensitivity for images cropped to contain their lesions (hence, somehow zoomed), likely because of a increased quantity of details passed to the CNN model. We thus propose CR as the preferred condition which takes away potential biases in the data and forces the model to learn more from salient details (lesion area). We expect the models trained on CR to generalize better to unseen data and deviate from the learning process of models with high bias in data. This way, we can potentially improve data quality and reduce overfitting, although this means a slight drop in performance on the closed world test set.

It is worth noting that with all masking conditions, the number of epochs needed to converge to the best predicting model, decreases by over 50%. This happens not only when blackening out significant parts of the image, thus providing the CNN with flat-valued uniform color areas, but also when zooming into the image and maximizing the number of pixels belonging to lesions. This suggest that the network is indeed *learning faster* thanks to the high quantity of meaningful, focused, information.

From the visual inspection of the saliency maps (Sect. 5), it appears that when images are classified as melanoma, the network concentrates most of its "attention" in the central part of the image, as a human practitioner would do. In contrast, when images are classified as nevus, the saliency map is more spread towards the border. This last phenomenon is less regular in the M and CR conditions, where most of the healthy skin is absent, suggesting that the CNN (when classifiyng a nevus) tends to activate on the small areas of skin around the lesion. Notably, this happens regardless of the correctness of the prediction, showing that in fact the CNN learned to search for the features characterizing the positive case (melanoma) within the lesion area.

However, it seems that in absence of visual elements characterizing a melanoma, the network has the tendency to find a "reason" for the competing class (nevus) elsewhere in the image, either on blacked-out areas, which are surely non-discriminating, but also on healthy skin areas. This might in part question and refute the conclusions of Burdick et al. [3], who stated that extending the masks of a lesion allows us to take advantage of the contrast between the lesioned and healthy skin. Differently, it seems that CNNs really need an "area of alternative attention", which we could define as the portions of the image on which the CNN needs to concentrates the activation of its layers when predicting a negative case (nevus).

7 Conclusions

In this paper, we presented a comprehensive investigation on the effect of masking on the binary classification of skin lesions between nevus and melanoma towards international standards for image preprocessing to reduce bias and increase data quality.

We performed our analyses on three datasets (ISIC 2018, MedNode, and ISIC 2016) using a 10-fold cross validation procedure. Then, in order to discard shallow conclusions due to the intrinsic randomness of CNN training procedures, we considered only those differences that have been confirmed as significant through statistical tests.

Inspired by the work of Bisotto et al. [2], who discovered the possibility of classifying skin lesions still after covering 70% of the internal surface of the images, we verified that prediction power indeed diminishes when removing 30% around the border, thus confirming the existence of some kind of bias.

Further experiments, with other types of masking, confirmed the bias at the border, and also showed that the best non-biased performances can be achieved through automated cropping.

The cropping condition also leads to 50% shorter training times, suggesting that the presence of healthy skin is noisy information that slows the convergence of the training process.

Finally, an automated analysis of the saliency maps extracted from the CNN classifier via Grad-CAM led us to formulate an hypothesis of *area of alternative attention*. In fact, the analysis leads to the following informal argument: while it is true that one should better maximize the area of the image with visual features able to identify a (positive) class, at the same time some of the pixels should be left free for the network to “justify” the complementary (negative) class. Future work, with more fine-tuned masking along the border of the lesion, and on other datasets, should be conducted to confirm this hypothesis.

In fact, it is worth noticing that most of the research in image classification has been conducted on databases of images where the objects of interests occupy only a relatively small portion of an image. Consequently, visual explanation methods like GradCAM [18] and RISE [15] have been developed and tested with the goal of identifying the relatively small subset of pixels justifying a classification. Differently, in the domain of skin cancer detection very often the majority of the pixels of an image are associated to a single entity, and this case has been so far received very little attention.

In general, the outcome of this investigation supports the idea that the creation of systems for skin lesion classification should go through a cropping process, either automated or manual, for both the creation of training data and for samples classification. This would both increase prediction performances (at least on sensitivity) and would significantly reduce the computational power needed for training—towards a process to standardize image preprocessing in CNN contexts [22].

Acknowledgements. The research has been supported by the Ki-Para-Mi project (BMBF, 01IS19038B), the pAItient project (BMG, 2520DAT0P2), and the Endowed Chair of Applied Artificial Intelligence, Oldenburg University. We would like to thank all student assistants that contributed to the development of the platform (see https://iml.dfki.de/).

References

1. Berseth, M.: ISIC 2017 - skin lesion analysis towards melanoma detection. CoRR abs/1703.00523 (2017). http://arxiv.org/abs/1703.00523
2. Bissoto, A., Fornaciali, M., Valle, E., Avila, S.: (De)Constructing bias on skin lesion datasets. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops, June 2019
3. Burdick, J., Marques, O., Weinthal, J., Furht, B.: Rethinking skin lesion segmentation in a convolutional classifier. J. Digit. Imaging **31**(4), 435–440 (2017). https://doi.org/10.1007/s10278-017-0026-y
4. Codella, N., Rotemberg, V., Tschandl, P., Celebi, M.E., et al.: Skin Lesion Analysis Toward Melanoma Detection 2018, February 2019. http://arxiv.org/abs/1902.03368
5. Codella, N.C.F., Gutman, D., Celebi, M.E., Helba, B., et al.: Skin lesion analysis toward melanoma detection: a challenge at the 2017 International symposium on biomedical imaging. In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018), Washington, DC, pp. 168–172. IEEE, April 2018. https://doi.org/10.1109/ISBI.2018.8363547

6. Curiel-Lewandrowski, C., Novoa, R.A., Berry, E., Celebi, M.E., et al.: Artificial intelligence approach in melanoma. In: Melanoma, pp. 1–31. Springer, New York, New York, NY (2019). https://doi.org/10.1007/978-1-4614-7322-0_43-1
7. Deng, J., Dong, W., Socher, R., Li, L.J., et al.: ImageNet: a large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami, FL, pp. 248–255. IEEE, June 2009. https://doi.org/10.1109/CVPR.2009.5206848
8. Esteva, A., Kuprel, B., Novoa, R.A., Ko, J., Swetter, S.M., et al.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**, 115, January 2017. https://doi.org/10.1038/nature21056
9. Giotis, I., Molders, N., Land, S., Biehl, M., et al.: MED-NODE: A computer-assisted melanoma diagnosis system using non-dermoscopic images. Expert Syst. Appl. **42**(19), 6578–6585 (2015). https://doi.org/10.1016/j.eswa.2015.04.034
10. Kawahara, J., Hamarneh, G.: Visual Diagnosis of Dermatological Disorders: Human and Machine Performance, June 2019. http://arxiv.org/abs/1906.01256
11. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, vol. 25, pp. 1097–1105, Curran Associates, Inc., (2012)
12. Marchetti, M.A., Codella, N.C., Dusza, S.W., Gutman, D.A., et al.: Results of the 2016 international skin imaging collaboration international symposium on biomedical imaging challenge. J. Am. Acad. Dermatol. **78**(2), 270–277.e1 (2018). https://doi.org/10.1016/j.jaad.2017.08.016
13. Masood, A., Ali Al-Jumaily, A.: Computer aided diagnostic support system for skin cancer: a review of techniques and algorithms. Int. J. Biomed. Imaging **2013**, 1–22 (2013). https://doi.org/10.1155/2013/323268
14. Nguyen, D.M.H., Ezema, A., Nunnari, F., Sonntag, D.: A visually explainable learning system for skin lesion detection using multiscale input with attention U-Net. In: Schmid, U., Klügl, F., Wolter, D. (eds.) KI 2020. LNCS (LNAI), vol. 12325, pp. 313–319. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58285-2_28
15. Petsiuk, V., Das, A., Saenko, K.: RISE: randomized input sampling for explanation of black-box models. In: Proceedings of the British Machine Vision Conference (BMVC) (2018)
16. Qian, C., Liu, T., Jiang, H., Wang, Z., et al.: A detection and segmentation architecture for skin lesion segmentation on dermoscopy images. CoRR abs/1809.03917 (2018). http://arxiv.org/abs/1809.03917
17. Ronneberger, O., Fischer, P., Brox, T.: U-Net: convolutional networks for biomedical image segmentation. In: Navab, N., Hornegger, J., Wells, W.M., Frangi, A.F. (eds.) MICCAI 2015. LNCS, vol. 9351, pp. 234–241. Springer, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
18. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., et al.: Grad-CAM: visual explanations from deep networks via gradient-based localization. In: The IEEE International Conference on Computer Vision (ICCV), October 2017
19. Siegel, R.L., Miller, K.D., Jemal, A.: Cancer statistics. CA: Cancer J. Clin. **69**(1), 7–34, January 2019. https://doi.org/10.3322/caac.21551
20. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition, September 2014. http://arxiv.org/abs/1409.1556
21. Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., Liu, C.: A survey on deep transfer learning. In: Kůrková, V., Manolopoulos, Y., Hammer, B., Iliadis, L., Maglogiannis, I. (eds.) ICANN 2018. LNCS, vol. 11141, pp. 270–279. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01424-7_27

22. Wahlster, W., Winterhalter, C.: German Standardization Roadmap on Artificial Intelligence. Technical Report, DIN e.V. and German Commission for Electrical, Electronic and Information Technologies of DIN and VDE (2020)
23. Winkler, J.K., Fink, C., Toberer, F., Enk, A., et al.: Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. JAMA Dermatol. **155**(10), 1135 (2019). https://doi.org/10.1001/jamadermatol.2019.1735