Crop it, but not too much: the effects of masking on the classification of melanoma images

Fabrizio Nunnari[1 [0000−0002−1596−4043]], Abraham Ezema[1 [0000−0002−9671−0925]], and Daniel Sonntag[1,2 [0000−0002−8857−8709]]

[1] German Research Center for Artificial Intelligence (DFKI)
Stuhlsatzenhausweg 3, Campus D3.2, 66123 Saarbrücken, Germany
{fabrizio.nunnari,abraham_obinwanne.ezema,daniel.sonntag}@dfki.de
[2] Oldenburg University, Oldenburg, Germany

Abstract. To improve the accuracy of convolutional neural networks in discriminating between nevi and melanomas, we test nine different combinations of masking and cropping on three datasets of skin lesion images (ISIC2016, ISIC2018, and MedNode). Our experiments, confirmed by 10-fold cross-validation, show that cropping increases classification performances, but specificity decreases when cropping is applied together with masking out healthy skin regions. An analysis of Grad-CAM saliency maps shows that in fact our CNN models have the tendency to focus on healthy skin at the border when a nevus is classified.

Keywords: skin cancer · convolutional neural networks · image segmentation · masking · preprocessing · reducing bias

1 Introduction

As reported in the 2019 USA cancer statistics, skin diseases have been steadily increasing over the years, whereby skin cancer (with more than 100k cases) represents 7% of the total cancer cases, of which more than 90% are classified as melanoma. The importance of detecting skin cancer is evident from the high percentage of survival (92%) after surgery resulting from early detection [21].

The classification of skin lesions using computer vision algorithms has been a subject of recent research (e.g., [14,6,11]). One of the breakthroughs is the work of Esteva et al. [8], who report a better performance than expert dermatologists (on carefully selected cases) using a deep convolutional neural network (CNN).

Given the promising progress of computer vision algorithms in aiding skin lesion classification, the ISIC [10] hosts a competition for the automated analysis of skin lesions. In the years from 2016 to 2018 (see [13,5,4]), the challenge included three tasks: segmentation, attribute extraction, and classification. These tasks replicate the procedure usually followed by dermatologists: identify the contour of the skin lesion, highlight the areas in the lesion that suggest malignancy, and classify the specific type of lesion.

Masking skin lesion images, i.e., using segmentation to remove the pixels of the healthy skin while retaining the pixels belonging to the lesion, is an image

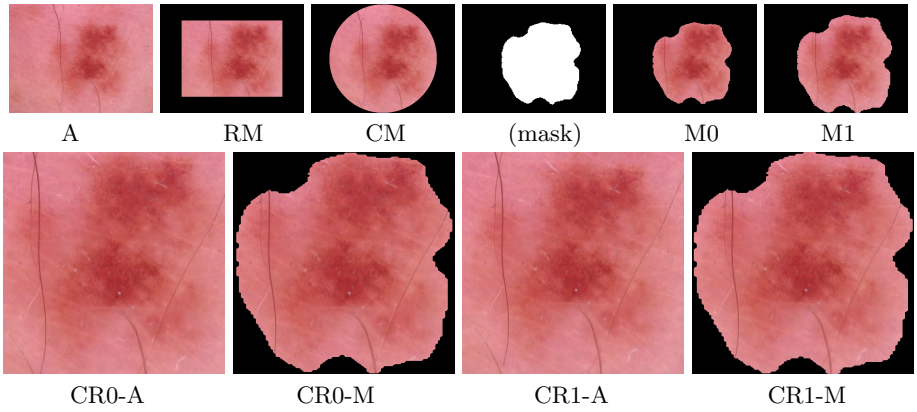

Fig. 1. Example for each of the masking policies.

pre-processing technique that is supposed to help the classification of skin lesions not only because it helps the systems to focus on the lesion itself, but also because it removes image artifacts. In fact, Winkler et al. [25] 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 Diagnosis Support Systems (DSS). Moreover, recently Bissoto et al. [2] found a strong bias in the ISIC 2018 dataset; by removing 70% of the central part of the images (hence likely removing the totality of pixels containing the skin lesions), the CNN model was still able to reach 0.74 AUC (the Area Under the Curve of the *receiver operating characteristic*), with respect to 0.88 AUC reached with full images. This suggests a strong bias of the dataset and image 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). Likewise, what are the other consequences for the process of training classifiers when for example learning the wrong, and medically irrelevant, concept?

In this paper, we present a detailed investigation and discussion 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 CNN attention patterns through a visual inspection by Grad-CAM [20] saliency maps to take (at least) visual explanations into account that account for medically relevant spatial regions (though not the semantic medical concepts behind). This analysis employs four basic types of *masks* (see figure 1, top):

1. Rectangular Mask (RM) removes 30% of the image surface around the border. This is the opposite of the masking utilized by Bissoto et al. [2] to prove a bias at the image borders.

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 (M0) reveals only the lesion pixels. It is used to reveal whether completely removing the skin surrounding a lesion improves prediction performance.
4. Extended Mask (M1) applies a mask extended by a factor 1.1 around its center, thus showing lesion pixels together with a fraction of the surrounding skin. It is used to check if providing information about the surrounding healthy skin improves prediction performances.

In addition, we investigate the change in performances when cropping the images with a rectangle circumscribing masks of type M0 and M1 (see figure 1, bottom):

- In condition CR0-A, the images are cropped at mask M0 and show the pixels of the surrounding skin.
- In condition CR0-M, the images are cropped at mask M0, but outside pixels are blacked out.
- In condition CR1-A, the images are cropped at mask M1 and show the pixels of the surrounding skin.
- Finally, in condition CR1-M, the images are cropped at mask M1, but outside pixels are blacked out.

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 overcome biases due to randomization.

To our knowledge, this is the first work measuring with such detail the role of masking and reporting that an excessive masking can in fact deteriorate performances, rather than improve them.

2 Related Work

Following the popular approach presented by Esteva et al. [8], all performant neural-network-based solutions for skin lesion classification are based on a transfer learning approach [23], where a baseline deep CNN is pre-trained for example on the ImageNet dataset [7], and the transfer-learning step 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 the VGG16 architecture [22].

Kawahara et al. [11] pointed out that much work focuses on improving benchmarks. Differently, our goal in this contribution is to investigate the change in performance when using plain images with respect to segmented ones for a classification task. To train our classifiers, we rely on three publicly available datasets: ISIC 2016 [13], ISIC 2018 [4], and MedNode [9]; all of which are used to train several models on a number of masking conditions (see Section 3).

Burdick et al. [3] performed a systematic study on the importance of masking skin lesion images. They measured the performance of a CNN using the full

images compared to applying 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 too much of the healthy skin in the image "confuses" the network, that is, it becomes more probable that the network learns image artifacts.

In general, an *image mask* is a binary black/white image wherein white is associated with pixels of interest, and black is associated with the non-interesting or the confounding part of the image to be discarded in subsequent processing steps. In skin lesion pre-processing, the identification of the contour of the contiguous area of lesioned skin is also known as *segmentation*. Ronnenberg et al. [19] first proposed the application of the convolution-deconvolution network (U-Net) for medical image segmentation. Variants of this model have shown to be very effective in past ISIC segmentation challenges, with a Jaccard index score of 0.765 and 0.802 in the ISIC2017 and ISIC2018 editions, respectively (see [1,18]). In this paper, we implement a segmentation model to show the effects of masking in melanoma images by using a variation of the Ronnenberg's method described by Nguyen et al. [15].

3 Method Overview

The experimental method is composed of three phases: preparation of the segmentation model, masked images construction, training of classification models.

As **segmentation model** we utilize the images from Task 1 of the ISIC 2018 to train a masking model based on the U-Net architecture [19] using the method of Nguyen et al. [15]. The train dataset comprised 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 2 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 3x3x3. Downsampling is the result of a max-pooling with size 2x2. Upsampling is the result of a 2x2 transposed 2D *same* convolution. After each upsampling step, the deconvolution 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 160x160 pixels.

The segmentation model described above is used to **extract the masks** for Melanoma and Nevus images of the ISIC 2018 Task 3, ISIC 2016, and MedNode datasets. From ISIC 2018 Task 3, we selected only nevus (NV) and melanoma (MEL) classes (the same used to train the segmentation model) because, after

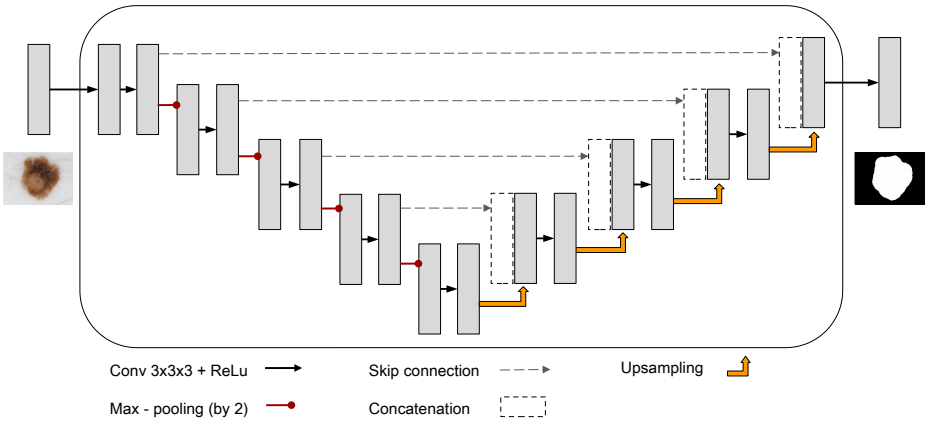

Fig. 2. The U-Net architecture used for lesion segmentation. The input image is 3-channel RGB, while the output image is 1-channel gray-scale with the same resolution.

an initial visual inspection, we realized that applying the mask prediction to any of the other 6 classes often led to erroneous results.

In total, starting from the full images (condition A, containing all of the pixels), we generated the other eight datasets described in the introduction: RM, CM, M0, M1, CR0-A, CR0-M, CR1-A, CR1-M (See figure 1). For the M1 and CR1-* datasets, 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 Burdick et al. [3]). For the CR*-* datasets, 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 visual 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 away images whose mask was less than $\frac{1}{8}$ of the picture area.

Finally, for each of the nine 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, performance metrics are reported as the mean (and the standard deviation) among the 10 folds. The performance of the binary classifiers in discriminating *nevi* (negative case) from *melanomas* (positive case) are reported in terms of accuracy, specificity, sensitivity, and ROC AUC (Receiver Operating Curve - Area Under the Curve) on the test set.

As already successfully employed in previous research (e.g., [8]), all of the binary classifiers are based on the transfer learning approach with CNNs [23]. The base CNN model is the VGG16 architecture [22], which has been pre-trained on ImageNet [12]. We then substituted the original three final fully connected layers with a sequence of two 2048-node fully connected layers, each followed by

Table 1. Results for the ISIC2018 dataset. Left: classification performances in all masking conditions, indicating with **bold** the values significantly above the baseline (condition A) and with *italic* the values significantly below the baseline. Right: significantly different masking conditions and their mutual absolute and relative variations.

set	testacc	testspec	testsens	testauc
A	.902 (.014)	.926 (.020)	.752 (.048)	.944 (.010)
RM	.898 (.011)	.920 (.012)	.768 (.053)	.940 (.012)
CM	.905 (.012)	.933 (.015)	.738 (.063)	.944 (.012)
M0	*.885 (.010)*	*.905 (.016)*	.762 (.060)	.936 (.012)
M1	*.885 (.010)*	*.905 (.012)*	.764 (.042)	*.935 (.010)*
CR0-A	**.926 (.013)**	**.947 (.016)**	**.801 (.054)**	**.961 (.009)**
CR0-M	*.833 (.018)*	*.845 (.027)*	.757 (.042)	*.915 (.010)*
CR1-A	**.922 (.007)**	**.948 (.012)**	.770 (.057)	**.958 (.009)**
CR1-M	*.884 (.008)*	*.911 (.013)*	.729 (.046)	*.921 (.014)*

Condition	Metr.	Diff.	%	p
A vs. M0	ACC	-.0172	-1.91%	**
A vs. M0	SPEC	-.0217	-2.34%	*
A vs. M1	ACC	-.0168	-1.86%	**
A vs. M1	SPEC	-.0214	-2.31%	*
A vs. M1	AUC	-.0092	-0.97%	+
A vs. CR0-A	ACC	.0239	2.65%	**
A vs. CR0-A	SPEC	.0201	2.17%	*
A vs. CR0-A	SENS	.0485	6.45%	+
A vs. CR0-A	AUC	.0171	1.81%	**
A vs. CR0-M	ACC	-.0689	-7.64%	***
A vs. CR0-M	SPEC	-.0812	-8.76%	***
A vs. CR0-M	AUC	-.0299	-3.17%	***
A vs. CR1-A	ACC	.0201	2.23%	**
A vs. CR1-A	SPEC	.0211	2.28%	*
A vs. CR1-A	AUC	.0141	1.49%	**
A vs. CR1-M	ACC	-.0173	-1.92%	**
A vs. CR1-M	SPEC	-.0160	-1.73%	+
A vs. CR1-M	AUC	-.0236	-2.50%	***

a dropout of 0.5, and a final 2-class discrimination softmax layer. Each model was trained for a maximum of 100 epochs and optimized for accuracy. Input images were fed to the network with an 8x 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 (227x227) using a nearest neighbor filter. Other training parameters are: SGD optimizer, learning rate=1e-5, decay=1e-4, momentum=0.9, nesterov=True. 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). All training was performed on Linux workstations using our toolkit for Interactive Machine Learning (TIML) [16]. Our reference Hardware is an 8-core Intel 9th-gen i7 CPU with 64GB RAM and an NVIDIA RTX Titan 24GB GPU.

4 Experiments

We report the details of the classification performances on the three datasets (ISIC2018, MedNode, and ISIC2016) for each of the nine masking conditions and the results of the statistical analyses comparing among masking conditions. The analysis focuses on determining a potential bias from the border of the images and the change in performances when masks are applied to the lesion border or are extended to reveal part of the healthy skin.

The **ISIC2018** dataset consisted of 7818 samples (7645 correctly cropped), of which 85.8% were nevi. Training a full model (6256 samples, 100 epochs) takes about 9 hours on our reference hardware. Table 1, left, show the results of the tests as mean (and standard deviation) over a 10-fold cross-validation.

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

Both rectangular (RM) and circular masks (CM) did not lead to any significant change. Differently, both the full (M0) and the extended masks (M1) lower the performances in accuracy and specificity. For M1 condition, also AUC is slightly decreasing.

This result seems to be in contrast with the observation of Bissoto et al. [2], who claims a positive bias at the border. In fact, by removing the 30% of the external image border (RM condition), we would have expected a drop in performance. We can't so far find an explanation for the loss of performances in the masking conditions M0 and M1, which could be caused by the imprecision of the segmentation algorithm or related to the following observations on the cropping conditions.

For the cropping conditions we observe a clear pattern. When the cropping is applied leaving visible healthy skin (CR0-A and CR1-A) the performances increase, up to a +2.65% accuracy and +6.45% sensitivity. This can be explained by the fact that when images are cropped almost all of the 227x227 pixels sent to the CNN are covered by the lesion–hence increasing the quantity of details attributed to the lesioned skin.

However, when the cropping is combined with a masking of the healthy pixels, the performances drop in terms of specificity. This is especially notable in the CR0-M condition, where no healthy skin is supposed to be visible.

This leads us to a question. As we cannot state that there is a bias at the images border, and that by removing all of the healthy skin pixels from the image performances drop: can we state that indeed *healthy skin contains fundamental useful information for a better classification?* This will be discussed in section 6 together with the results on the two other datasets, MedNode and ISIC2016, presented in the following.

The **MedNode** dataset consisted of 170 samples (169 correctly cropped), of which 58.8% were nevi. Training one fold of the full dataset (about 136 samples, 100 epochs) takes about 15 minutes on our reference hardware. Appendix A reports the results of the t-tests for independent samples between the no-mask condition (A) against all the others. The only notable performance difference is in the CR0-M condition, with a 10.85% drop in specificity, which is in line with the observation on ISIC2018, though with a limited significance ($p < 0.1$).

The **ISIC2016** dataset consisted of 900 samples (884 correctly cropped), of which 80.8% were nevi. Training one fold of the full dataset (722 samples, 100 epochs) takes about 1h 30m on our reference hardware. Appendix B reports the results of the t-tests for independent samples between the no-mask condition (A) against all the others. As for the ISIC2018 dataset, the masking conditions M0 and M1 lead to a drop in specificity. For this dataset, the drop is present also for the CM condition (essentially removing the angles of the images). Again in line with ISI2018, there is a significant drop in performances in the crop & mask conditions (CR0-M and CR1-M), both in accuracy and specificity.

In **summary**, for all of the three datasets, the application of a rectangular mask (RM) did not affect classification performances. As such, even though Bissoto et al. [2] claimed that the 30% of the border around the lesion is enough to reach an important classification result, with our experiments we could not state that removing the borders compromises the accuracy. In other words, although there is a bias in the images border, it is also true that when the lesioned skin is fully visible, this biased information are not affecting the performances as the network is fully concentrating on the lesioned area.

Another behavior common to the three datasets is that cropping and masking the images deteriorates performances. These are the conditions where all of the healthy skin pixels are blacked out (CR0-M, CR1-M). Hence, while it is true that removing some healthy skin pixels from the borders doesn't affect performances, removing all of the healthy skin negatively affect the classification. For the ISIC2018 dataset, we could also find that cropping but leaving healthy skin visible increases performance (CR0-A, CR1-A). This last finding couldn't be verified statistically on the MedNode and ISIC2016 datasets, possibly because of their lower amount of samples.

There seems hence to be contrast between the intuitive urge to help the neural network improving its performance by removing pixels of healthy skin, and the progressive degradation of performances as healthy skin pixels are less and less visible. To get a better understanding of this phenomenon, we ran a set of visual inspections and statistical analyses, which are presented in the next section.

5 Visual Inspection

In order to visually explain the characteristics that influenced model predictions, we leveraged the Grad-CAM method [20] to generate the saliency maps of *attention*. Figures 3 and 4 show the heatmaps of a correctly classified melanoma and a nevus in 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. For images correctly predicted as melanoma (Figure 3), the saliency is higher on the skin lesion pixels, focused towards the center of the image. In contrast, for pictures correctly classified as nevus (Fig-

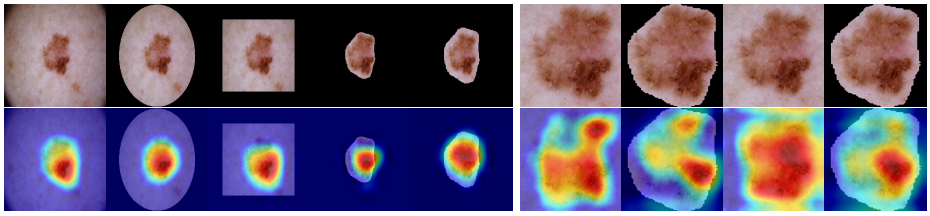

Fig. 3. Heatmaps of an ISIC2018 melanoma (ISIC_0032797) in all masking conditions. Notice how heatmaps concentrate towards the center of the image.

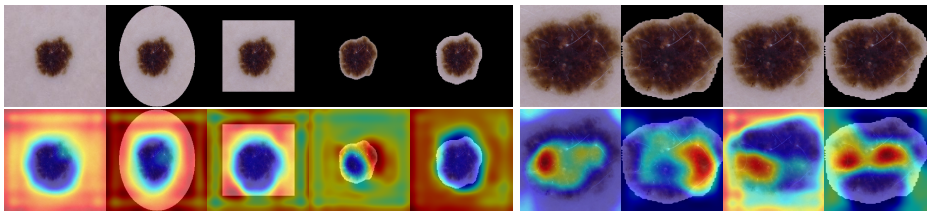

Fig. 4. Heatmaps of an ISIC2018 nevus (ISIC_0027548) in all masking conditions. Notice how heatmaps (except CR0-A) concentrate towards the border of the image.

ure 4), the saliency is higher on the skin pixels, towards the borders. The opposite happens when images are wrongly classified (not shown in the pictures), with the attention for wrongly classified nevus towards the center and the attention for wrongly classified melanomas towards the border.

From a closer look at conditions A, CM, RM, and M0 in figure 4, it seems that, as the healthy skin is progressively removed from the image, the network has the tendency to avoid black areas and tries to "justify" nevi by increasing its highest attention (red areas) towards the visible healthy skin pixels. This doesn't hold for condition M1, where the network is finally "discharging" its attention on the black borders. The same behaviour can be observed in the cropping conditions, especially for CR1-A.

Our intuition is that, in order to justify a nevus, the network needs an *area of alternative attention* to "motivate" its choice, otherwise it will be forced to look at the lesioned skin and might be induced in a wrong decision. This would explain the deterioration in terms of specificity observed in the ISIC2018 dataset when switching from CR*-A conditions to CR*-M conditions.

To confirm this intuition, we perform an analysis by measuring the degree of overlap between (thresholded) saliency maps and segmentation masks on the cropping conditions (CR0-A, CR0-M, CR1-A, CR1-M). For each of the four conditions, we apply the following procedure to each image:

1. Define $S_{0.5}$ as the saliency map thresholded at 0.5, where white pixels correspond to high saliency;
2. define L as the segmentation mask (as used for CR0-M or CR1-M), where white pixels mark the area with lesion pixels;

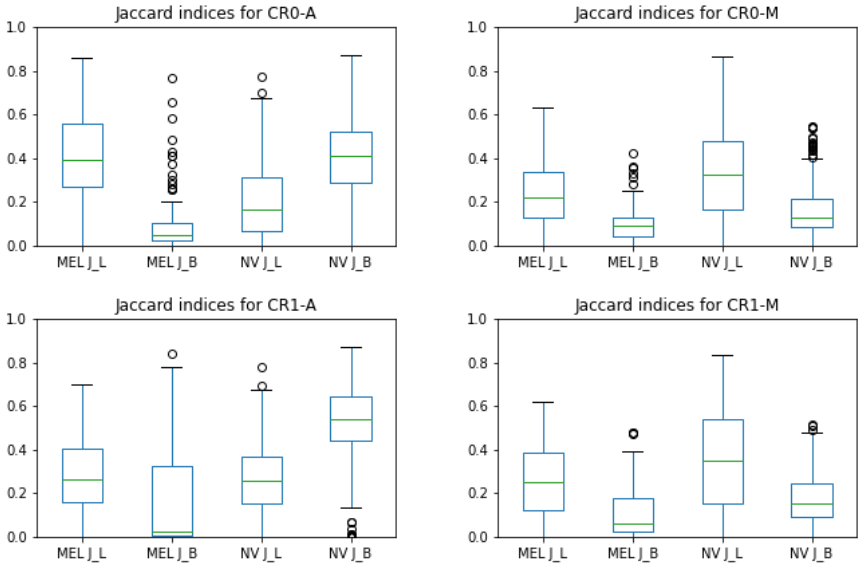

Fig. 5. Box plots of Jaccard indices between thresholded saliency and mask areas in all cropping (and masking) conditions.

3. define B as the complement of L, hence the "masked-away" area at the border;
4. compute J_L as the Jaccard index between $S_{0.5}$ and L; and
5. compute J_B as the Jaccard index between $S_{0.5}$ and B.

This leads to the creation of four datasets, each of them associating images to two Jaccard indices: J_L, representing the degree of overlap between the saliency and the lesioned skin, and J_B, representing the degree of overlap between the saliency and the healthy skin (or the black area). Figure 5 shows the box-plots of the resulting Jaccard indices J_L and J_B, for each dataset, divided by lesion type (MEL/NV).

The top-left boxplot (CR0-A, cropping but *no* black masking) shows that for melanoma (MEL), on average, J_L is higher than J_B, meaning that the saliency is more concentrated at the center. In contrast, for nevi (NV), saliency concentrates more at the border. However, for condition CR0-M (top-right, cropping and masking), for nevi (NV) the opposite happens, with the saliency more concentrated to the center. The two bottom plots report the same behavior for the condition pair CR1-A and CR1-M.

More formally, we formulated the following hypotheses:

- **H0** In all four conditions, for MEL, mean J_L is higher than mean J_B;
- **H1** In CR*-A conditions (healthy skin is visible), for NV, mean J_L is lower than mean J_B;

Table 2. Left: results of Mann-Whitney tests comparing J_L and J_B distributions. Right: results of Mann-Whitney tests comparing J_L and J_B distributions on the NV class between CR0-A and CR0-M (CR1-A vs. CR1-M, respectively).

Cond.	Class	$\overline{J_L}$	$\overline{J_B}$	U	Sig.
CR0-A	MEL	0.4010	0.1006	U=10662	***
CR0-A	NV	0.1997	0.4003	U=86403	***
CR0-M	MEL	0.2409	0.1020	U=9642	***
CR0-M	NV	0.3437	0.1562	U=336747	***
CR1-A	MEL	0.2844	0.1757	U=8692	***
CR1-A	NV	0.2662	0.5265	U=51450	***
CR1-M	MEL	0.2600	0.1182	U=9368	***
CR1-M	NV	0.3578	0.1631	U=321957	***

Crop	J	$\overline{A}$	$\overline{M}$	U	Sig.
CR0-	J_L	0.1997	0.3437	U=124265	***
CR0-	J_B	0.4003	0.1562	U=375561	***
CR1-	J_L	0.2662	0.3578	U=169012	***
CR1-	J_B	0.5265	0.1631	U=406799	***

- **H2** In CR*-M conditions (healthy skin is masked to black), for NV, mean J_L is higher than mean J_B;
- **H3** For NV, mean J_L in CR*-A is lower than mean J_L in CR*-M;
- **H4** For NV, mean J_B in CR*-A is higher than mean J_B in CR*-M.

A set of Mann-Whitney U tests (table 2) confirmed a statistically significant difference for all tests with $p < .001$, thus baking all of our hypotheses. In other words, we can conclude that *when some skin is visible, the nevi classification is accumulating the saliency on the healthy skin pixels.* In contrast, when no healthy skin is visible, the CNN attention is forced towards the center of the image, thus compromising classification specificity.

6 Discussion

The metrics measurement performed on the image datasets (ISIC2016, MedNode, and ISIC2018) shows that, with respect to using full plain images, masking decreases the performances in terms of specificity. The results on ISIC2018 show that cropping can increase the performance of the network, but cropping and masking decreases specificity.

From the visual inspection of the saliency maps (section 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 to be classified as nevus, the saliency map is more spread towards the border. This last phenomenon is less pronounced in the CR0-M and CR1-M conditions, where most of the healthy skin is absent. 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. Blacked-out areas, which are surely non-discriminating, are avoided and healthy skin areas are preferred. This leads us to re-interpret the conclusions of Burdick et al. [3], who found that (with respect to full masks) indeed extended masks increase performances, but explained it in terms of "taking advantage of the contrast between the lesioned and healthy skin". Differently, it seems that CNNs really need an "area of alternative attention", which could informally

defined as the portions of the image on which the CNN needs to focus 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 classification of skin lesions between nevus and melanoma. We performed our statistical analyses on three datasets (ISIC 2018, MedNode, and ISIC 2016) using a 10-fold cross validation procedure to discard shallow conclusions due to the intrinsic randomness of CNN training procedures.

Our experiments show that the best strategy to improve performance is to crop images around the rectangular area containing the lesion segmentation mask. Likely, performances increase thanks to the higher image detail after zooming in the lesioned area. However, specificity decreases when cropping is performed together with masking to black the healthy surrounding skin.

To better explain this behavior, we conducted an automated analysis on saliency maps and formulated the hypothesis that CNNs are more effective when an *area of alternative attention* is available. In summary, 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.

As a result of our experiments, towards a process to standardize skin lesions image preprocessing in CNN contexts (like in the standardization roadmap for artificial intelligence [24]), we suggest to apply an automated process of segmentation and cropping, but avoiding masking to black surrounding healthy skin.

Future work might investigate if the hypothesis of "area of alternative attention" generalizes to other contexts by testing classification performances after cropping images on popular non-medical databases (e.g., ImageNet [7]). In fact, it is worth noticing that most of the research in image classification has been conducted on databases where the objects of interest occupy only a relatively small portion of an image. Consequently, visual explanation methods like GradCAM [20] and RISE [17] 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 (or all) of the pixels of an image are associated to a single entity, and this condition has received so far very little attention.

Another subject of investigation would be on understanding what CNNs see on healthy skin that is invisible to human eyes and can point to new medically-relevant features.

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 (see `https://uol.de/aai/`. 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 (Jun 2019)
3. Burdick, J., Marques, O., Weinthal, J., Furht, B.: Rethinking Skin Lesion Segmentation in a Convolutional Classifier. Journal of Digital Imaging **31**(4), 435–440 (Aug 2018). 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 (Feb 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). pp. 168–172. IEEE, Washington, DC (Apr 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. pp. 248–255. IEEE, Miami, FL (Jun 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 (Jan 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 Systems with Applications **42**(19), 6578–6585 (Nov 2015). https://doi.org/10.1016/j.eswa.2015.04.034
10. ISIC: International Skin Imaging Collaboration, https://www.isic-archive.com/
11. Kawahara, J., Hamarneh, G.: Visual Diagnosis of Dermatological Disorders: Human and Machine Performance (Jun 2019), http://arxiv.org/abs/1906.01256
12. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet Classification with Deep Convolutional Neural Networks. In: Advances in Neural Information Processing Systems 25, pp. 1097–1105. Curran Associates, Inc. (2012)
13. 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. Journal of the American Academy of Dermatology **78**(2), 270–277.e1 (Feb 2018). https://doi.org/10.1016/j.jaad.2017.08.016
14. Masood, A., Ali Al-Jumaily, A.: Computer Aided Diagnostic Support System for Skin Cancer: A Review of Techniques and Algorithms. International Journal of Biomedical Imaging **2013**, 1–22 (2013). https://doi.org/10.1155/2013/323268
15. 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: 43rd German Conference on Artificial Intelligence (KI) (2020)
16. Nunnari, F., Sonntag, D.: A software toolbox for deploying deep learning decision support systems with xai capabilities. In: Proceedings of the 13th ACM SIGCHI Symposium on Engineering Interactive Computing Systems. ACM (2021). https://doi.org/10.1145/3459926.3464753

17. 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)
18. 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
19. Ronneberger, O., Fischer, P., Brox, T.: U-Net: Convolutional Networks for Biomedical Image Segmentation. In: Medical Image Computing and Computer-Assisted Intervention – MICCAI 2015, vol. 9351, pp. 234–241. Springer International Publishing, Cham (2015). https://doi.org/10.1007/978-3-319-24574-4_28
20. 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) (Oct 2017)
21. Siegel, R.L., Miller, K.D., Jemal, A.: Cancer statistics, 2019. CA: A Cancer Journal for Clinicians **69**(1), 7–34 (Jan 2019). https://doi.org/10.3322/caac.21551
22. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition (Sep 2014), http://arxiv.org/abs/1409.1556
23. Tan, C., Sun, F., Kong, T., Zhang, W., et al.: A Survey on Deep Transfer Learning. In: Artificial Neural Networks and Machine Learning – ICANN 2018. pp. 270–279. Springer International Publishing, Cham (2018)
24. Wahlster, W., Winterhalter, C.: German Standardization Roadmap on Artificial Intelligence. Tech. rep., DIN e.V. and German Commission for Electrical, Electronic & Information Technologies of DIN and VDE (2020)
25. 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 Dermatology **155**(10), 1135 (Oct 2019). https://doi.org/10.1001/jamadermatol.2019.1735

A MedNode results

Table 3 and table 4 show the results of the experiments on the MedNode dataset.

Table 3. Classification performance on the MedNode dataset. Italic text indicates values significantly below the baseline (condition A).

set	testacc	testspec	testsens	testauc
A	.806 (.123)	.870 (.100)	.714 (.181)	.869 (.131)
RM	.753 (.094)	.800 (.118)	.686 (.189)	.860 (.073)
CM	.818 (.140)	.830 (.135)	.800 (.194)	.890 (.114)
M0	.806 (.083)	.850 (.092)	.743 (.167)	.890 (.107)
M1	.806 (.112)	.830 (.090)	.771 (.214)	.880 (.111)
CR0-A	.768 (.144)	.820 (.087)	.700 (.328)	.843 (.120)
CR0-M	.739 (.085)	*.776 (.107)*	.686 (.154)	.823 (.092)
CR1-A	.823 (.111)	.870 (.090)	.757 (.203)	.882 (.093)
CR1-M	.764 (.069)	.809 (.104)	.700 (.149)	.816 (.090)

Table 4. Significant differences between masking conditions in the MedNode dataset.

Condition	**Metr.**	**Diff.**	**Diff. %**	**Sig.**
A vs. CR0-M	SPEC	-0.0944	-10.85%	+

B ISIC2016 results

Table 5 and table 6 show the results of the experiments on the ISIC2016 dataset.

Table 5. Classification performance on the ISIC2016 dataset. Italic text indicates values significantly below the baseline (condition A).

set	testacc	testspec	testsens	testauc
A	.806 (.028)	.898 (.031)	.416 (.163)	.773 (.074)
RM	.794 (.037)	.878 (.069)	.445 (.146)	.756 (.063)
CM	.788 (.026)	*.872 (.030)*	.432 (.151)	.790 (.059)
M0	.792 (.039)	*.860 (.056)*	.509 (.119)	.778 (.065)
M1	.781 (.033)	*.852 (.045)*	.486 (.152)	.773 (.065)
CR0-A	.803 (.056)	.885 (.057)	.467 (.140)	.776 (.088)
CR0-M	*.760 (.037)*	*.827 (.030)*	.485 (.091)	.754 (.060)
CR1-A	.796 (.028)	.890 (.028)	.410 (.115)	.772 (.068)
CR1-M	*.757 (.036)*	*.845 (.051)*	.391 (.148)	.732 (.080)

Table 6. Significant differences between masking conditions in the ISIC2016 dataset.

Condition	**Metr.**	**Diff.**	**Diff. %**	**Sig.**
A vs. CM	SPEC	-0.0262	-2.92%	+
A vs. M0	SPEC	-0.0384	-4.28%	+
A vs. M1	SPEC	-0.0465	-5.18%	*
A vs. CR0-M	ACC	-0.0453	-5.62%	**
A vs. CR0-M	SPEC	-0.0712	-7.93%	***
A vs. CR1-M	ACC	-0.0488	-6.06%	**
A vs. CR1-M	SPEC	-0.0529	-5.89%	*