On the overlap between Grad-CAM saliency maps and explainable visual features in skin cancer images

Fabrizio Nunnari[1 [0000−0002−1596−4043]], Md Abdul Kadir[1,2 [0000−0002−8420−2536]], and Daniel Sonntag[1,2 [0000−0002−8857−8709]]

[1] German Research Center for Artificial Intelligence (DFKI)
Saarland Informatics Campus D3 2
{fabrizio.nunnari,md_abdul.kadir,daniel.sonntag}@dfki.de
[2] Oldenburg University

Abstract. Dermatologists recognize melanomas by inspecting images in which they identify human-comprehensible visual features. In this paper, we investigate to what extent such features correspond to the saliency areas identified on CNNs trained for classification. Our experiments, conducted on two neural architectures characterized by different depth and different resolution of the last convolutional layer, quantify to what extent thresholded Grad-CAM saliency maps can be used to identify visual features of skin cancer. We found that the best threshold value, i.e., the threshold at which we can measure the highest Jaccard index, varies significantly among features; ranging from 0.3 to 0.7. In addition, we measured Jaccard indices as high as 0.143, which is almost 50% of the performance of state-of-the-art architectures specialized in feature mask prediction at pixel-level, such as U-Net. Finally, a breakdown test between malignancy and classification correctness shows that higher resolution saliency maps could help doctors in spotting wrong classifications.

Keywords: skin cancer · visual features · explainable AI · saliency maps

1 Introduction

The recognition of skin cancer from digital pictures is a task that has received much attention in the last years [3,23,22]. Many evaluations show that convolutional neural networks (CNNs) are capable of distinguishing between malignant skin cancer and benign lesions with higher accuracy than experienced practitioners [11,2].

When neural networks are employed as classification models, decisions come by default without a human-comprehensible explanation—an issue affecting the adoption of neural networks in medical applications both for legal reasons as well as for the lack of trust in such systems. On the contrary, dermatologists diagnose skin cancer in the basis of widely recognized *visual features*, i.e., areas of the skin, or regions of interest (ROIs), characterized by well-defined visual patterns associated with medical concepts [21]. Figure 1 shows some examples. Such visual features can be present in both benign as well as malignant lesions. When their visual presence is significant, clinical guidelines suggest to assume malignancy.

New algorithms in the field of eXplainable Artificial Intelligence (XAI) allow for the extraction of saliency maps from classification models; Grad-CAM [27] is one of

Fig. 1. The picture of a melanoma (ISIC_0000013, from the picture archive of the International Skin Imaging Collaboration) and its segmentation, followed by annotations for globules, pigment network, streaks, and the union of the three.

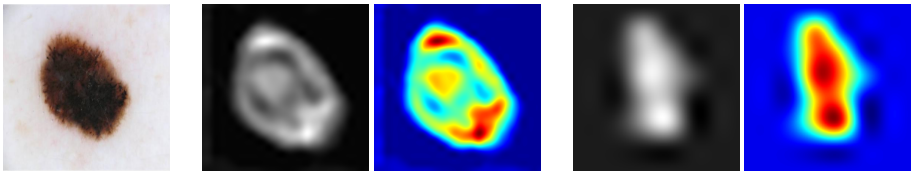

Fig. 2. The picture of a correctly classified melanoma (ISIC_0000013) (left), the saliency map and its colored heatmap computed on a VGG16 (middle) and on a RESNET50 model (right).

the most popular algorithms. Saliency maps are images that indicate the pixels areas contributing to a certain classification decision. Saliency maps are normally encoded as greyscale images or converted to heatmaps for visual inspection. Figure 2 provides an example.

Intuitively, it can be expected and observed an overlap between regions with high saliency and regions of interest (that practitioners would identify as signs of malignancy) occurs. However, this relationship has never been investigated in detail.

In this paper, we present a study measuring to what extent saliency maps can be used to identify visual features of skin lesions. In particular, we investigate the behavior of the involved deep learning architectures on actual data, in order to extract reference measurement values, reference thresholds, and to identify the limits of this approach.

The remainder of this paper first describes in section 2 related work in the field of feature extraction and XAI. In section 3, we describe the two classification models used for our experiments, able to discriminate (among others) between nevus and melanoma. Section 4 describes the skin lesion images and masking data used for the experiments. Then, section 5 describes the first experiment, aiming to find the best threshold value maximizing the overlap between saliency maps and ground truth regions of interest, which is not granted to be the usual 0.5. The second experiment (section 6), investigates the difference in overlpping when distinguishing between correctly vs. wrongly classified lesions, showing significant differences. Finally, section 7 discusses the results of the experiments and section 8 describes future work.

2 Related Work

The experiments presented in this paper are conducted on the dataset presented for the Task 2 (feature extraction) of the ISIC 2018 challenge (https://challenge2018.isic-archive.com) [5]. The dataset contains 2386 dermoscopy images, all of them annotated with

binary masks highlighting the presence of five "features" at pixel-level, i.e., patterns on skin lesions unanimously recognized as indicators of potential malignancy [21]. Namely, the are: globules, streaks, pigment network, negative network, and milia-like cysts.

The performances on the feature extraction task are measured using the Jaccard index. Given two 2-color (black-white) image masks of the same resolution, the index J returns the ratio between the count of the common white pixels (correctly classified) and the union of all the white pixels. This is also known as the "intersection over union" ratio. To give a reference on the performances of the best feature extraction models, the first three ranks of the ISIC 2018 challenge were taken by the NMN-Team, with three approaches reaching J=0.307, 0.305, and 0.304 respectively [18]. One of the goals of the analysis presented in this paper is to assess to what extent thresholded saliency maps can identify skin lesion features, to compare performances with the best ISIC 2018 systems, and to provide a reference performance baseline of a XAI-based system.

The recent experiments [8] on skin cancer detection focus on image classification only: they cannot produce explanations. Esteva et al. [11] also show that Inception v3 works very well in skin lesion detection and outperforms doctors. Even though the algorithm outperforms doctors, it cannot explain its decision accurately. Han et al. [13] fine-tuned the ResNet-152 model for cutaneous tumor detection. The classification performance of the network was comparable with that of 16 dermatologists, and they also exploit Grad-CAM to explain the classifications. There is a lot of success in explaining algorithmic output, but to increase the performance of an explainable model, we need a way to evaluate the quality of an explanation [16].

Concerning visual explanation techniques, GradCAM [27] is an analytical technique that, applied to convolutional neural classifiers, is able to highlight the areas of a picture contributing to the classification choice. The method is fast, as it needs just one forward and one backward propagation step, and then builds the saliency map from an analysis of the activation values of an intermediate chosen convolutional layer. Best XAI results are obtained by analyzing the last convolutional layer of a network [28,10]; however, the results are often low-resolution images.

A method that provides higher-resolution images is RISE [25], which is based on a stochastic approach. Input images are iteratively altered via random noise, and the final saliency map is composed by accumulating the partial estimations. However, its application requires much more computational power, as it needs to run hundreds of thousands of prediction cycles. Additionally, from a set of initial tests, it seems that RISE is not able to highlight regions of interest of skin lesion images with the same reliability as on pictures of real-world objects. The experimentation using RISE (together with other visual XAI variants like Grad-CAM++ [4] and SmoothGrad [30]) is deferred to future work.

Arun et al. [1] measured the overlapping between saliency maps and human-traced ground truth, but in the domain of chest X-rays, and used only very deep networks (InceptionV3 and DenseNet121), which provide very low resolution maps. Interestingly, they found the best XAI method being XRAI [19], which we plan to include in future work using also the evaluation methodology suggested by Sun et al. [31].

Several works focus on the use of saliency map to perform lesion segmentation (i.e., distinguish the lesioned from the healthy skin area) before passing it to a classifier. Among them, Gonzalez-Diaz proposes DermaKNet [12], which follows several pre-processing steps before the classification of skin lesion. In the first step, it creates a segmentation mask and applies it to the dermoscopic image. Secondly, it creates a structure segmentation mask to identify the structure of the dermoscopic image. After masking, the original segmented image and some nonvisual metadata are fed into a convolutional neural network for classification. Khan et al. [20] propose a channel enhancing technique to increase the contrast of lesion area. As a result, there is an improvement in the quality of the segmentation mask. Jahanifar et al. [17] also propose a modified DRFI (Discriminative Regional Feature Integration) technique for a similar task for multi-level segmentation task. By combining multiple segmentation masks, they produce a more accurate mask. During the generation of the mask, they use a threshold value of 0.5, but they did not provide a reason for which they choose this value.

In our work, we rather focus on the specific degree of overlap between the saliency maps and the visual features that dermatologists search for a diagnosis.

3 Classification Architectures and Models

In this section, we describe the two classification models used for our experiments. The two models are based on two different architectures: VGG16 [29] and RESNET50 [14]. We selected these two architectures to monitor the saliency map generation process according to two important model differences: the classification performances and the resolution of the last convolutional layer. Interestingly, while the RESNET50 architecture definitely results in the better classifier, the resolution of its last convolution layer (`res5c_branch2c`) is limited to 8x8 pixels, which is much less than the 28x28 pixels resolution of the last convolution layer (`block5_conv3`) in the VGG16 architecture. Our first hypothesis is that although more prone to classification errors, the VGG16 could still deliver better "visual explanations" because of its higher resolution.

Following a transfer learning approach, both the VGG16 and the RESNET50 models are pre-trained on the Imagenet dataset [9], and their final layers are substituted with randomly initialized fully connected layers of 2048 nodes and a final 8-level softmax. Then, both models are trained using 20k images of the ISIC2019 challenge (https://challenge2019.isic-archive.com) [33,6,7], which contains 8 classes (MEL, NV, BCC, AK, BLK, DF, VASC, SCC). VGG16 was configured with an input resolution of 450x450 pixels, while RESNET50 at 227x227 pixels.

The models are tested on 2529 held-out images, and we report the following class specific metrics.

For VGG16, accuracies are MEL: 0.845, NV: 0.827, BCC: 0.934, AK: 0.964, BLK: 0.912, DF: 0.991, VASC: 0.995, SCC: 0.977, and the class sensitivities are MEL: 0.659, NV: 0.755, BCC: 0.783, AK: 0.593, BLK: 0.626, DF: 0.826, VASC: 0.880, and SCC: 0.661. Overall accuracy is 0.722 and average balanced accuracy (mean sensitivity, the metric for the ISIC challenge) is 0.723.

For RESNET 50, accurracies are MEL: 0.873, NV: 0.857, BCC: 0.947, AK: 0.972, BLK: 0.920, DF: 0.992, VASC: 0.995, SCC: 0.977, and the class sensitivities are MEL: 0.675, NV: 0.812, BCC: 0.834, AK: 0.628, BLK: 0.672, DF: 0.739, VASC: 0.800, and SCC: 0.726. Overall accuracy is 0.767 and the average balanced accuracy is 0.736

As a reference, the 2nd placed at the ISIC challenge, which is the best approach not using external data (therefore comparable to our approach), measured an average balanced accuracy of 0.753 [34].

All of the model training and testing was performed using our Toolkit for Interactive Machine Learning (TIML) [24], which operates on top of the Keras and Tensorflow frameworks.

4 Data Preparation

To generate the saliency maps for our experiments, we run the two classification models (VGG16 and RESNET50) on the images of the ISIC Challenge 2018 Task 2 (see related work). The dataset contains 2386 RGB skin lesion images (519 melanomas, 1867 nevi), each associated to five ground truth black-white feature maps: globules, streaks, pigment network, negative network, and milia-like cysts. As an additional feature, we compute the pixels-wise *union* of all the features (see figure 1). The resolution of the ground truth feature maps is consistent with their corresponding colored picture.

Table 1. Counts of non-black masks for each feature class on the 2386 total samples.

Globules	Mil.	Neg. net.	Pig. net.	Streaks	Union
601	574	188	1502	98	1963

Some of the ground truth feature maps are completely black, as the dermatologists did not find any region of the corresponding class during the annotation. As can be seen in table 1, only 1963 pictures have at least one non-black feature map. In our experiments, we ignore the skin lesion samples with no features.

The *generation of the saliency maps* consists of running the Grad-CAM algorithm [27] on each skin lesion picture with non-black union mask. The saliency is generated for the predicted class. We repeat the procedure for both the VGG16 and the RESNET50 models, generating the $\mathcal{S}_V$ and $\mathcal{S}_R$ greyscale picture sets, where $\|\mathcal{S}_V\| = \|\mathcal{S}_R\| = 1963$. Saliency maps have a resolution of 24x24 pixels for $\mathcal{S}_V$ and 8x8 for $\mathcal{S}_R$, and their pixels are normalized in the range $[0, 1]$.

To *compare the saliency maps with ground truth maps*, we scaled up $\mathcal{S}_V$ and $\mathcal{S}_R$ to the resolution of the original images using a nearest neighbour filter. Figure 3 shows the histogram distribution of the Jaccard indices J computed between the features class (plus union) and $\mathcal{S}_V$ at threshold 0.5 ($\mathcal{S}_V^{0.5}$). We can observe that all distributions are strongly right skewed, and all Js are mostly below 0.2, with the exception of a peak in performance for the pigment network class. A similar profile could be observed for $\mathcal{S}_R^{0.5}$. The next step is to investigate whether 0.5 is the best value to use for thresholding saliency maps.

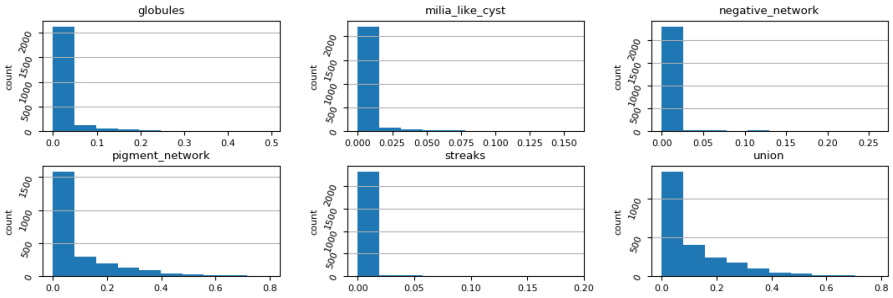

Fig. 3. Distribution of the measured Jaccard indices (horizontal axis), computed on $\mathcal{S}_V^{0.5}$.

Fig. 4. VGG16: the saliency map for sample ISIC_0000013 thresholded for 0.0=all-white, 0.1, 0.2, ..., 1.0=all-black. Corresponding Jaccard indices between ground truth and union feature are 0.064, 0.150, 0.171, 0.189, 0.221, 0.242, 0.243, 0.216, 0.137, 0.059, and 0.000.

5 First Experiment

With the first experiment we aim at identifying the threshold value that leads to a maximization of the overlap between saliency maps and ground truth. To do so, we converted each saliency map into 11 binary maps using thresholds from 0.0 to 1.0 with steps of 0.1. For example, for VGG16, we define 11 sets $\mathcal{S}_V^t, t \in 0.0, 0.1, ..., 0.9, 1.0$. Figures 4 and 5 show examples of this threshold process. Then, we proceed by computing the Jaccard indices J between the ground truth and all of the processed saliencies $\mathcal{S}_V^x$ and $\mathcal{S}_R^x$.

Tables 2 and 3 report the summary of the *threshold analysis* for VGG16 and RESNET50, respectively, on which we report the threshold leading to the highest average Jaccard index.

Table 2. For VGG16, the best performing masking threshold together with corresponding Jaccard index data.

Feature	Best Thr.	J-min	J-mean (SD)	J-max
Globules	0.600	0.000	0.067 (0.078)	0.428
Mil.	0.600	0.000	0.019 (0.032)	0.236
Neg. net.	0.400	0.000	0.044 (0.048)	0.201
Pig. net.	0.500	0.000	0.141 (0.146)	0.797
Streaks	0.700	0.000	0.062 (0.062)	0.271
Union	0.500	0.000	0.132 (0.137)	0.784

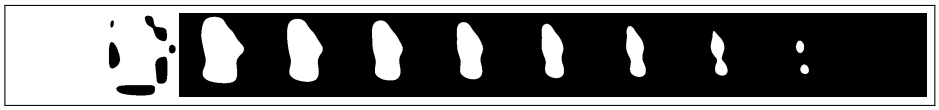

Fig. 5. RESNET50: the saliency map for sample ISIC_0000013 thresholded for 0.0=all-white, 0.1, 0.2, ..., 1.0=all-black. Corresponding Jaccard indices between ground truth and union feature are 0.064, 0.104, 0.156, 0.174, 0.174, 0.158, 0.127, 0.101, 0.072, 0.009, and 0.000.

Table 3. For RESNET50, the best performing masking threshold together with corresponding Jaccard index data.

Feature	Best Thr.	J-min	J-mean (SD)	J-max
Globules	0.500	0.000	0.079 (0.090)	0.591
Mil.	0.700	0.000	0.032 (0.043)	0.288
Neg. net.	0.600	0.000	0.100 (0.102)	0.526
Pig. net.	0.300	0.000	0.133 (0.134)	0.720
Streaks	0.600	0.000	0.041 (0.050)	0.265
Union	0.300	0.000	0.136 (0.130)	0.720

For VGG16, among the features classes, the best threshold ranges between 0.4 and 0.7. The minimum J index is 0.0 on all categories, meaning that among all samples there is always at least one map with zero-overlap with the ground truth. The highest average (J=0.141) and maximum (J=0.797) belong to the pigmented network class. The union of all features lowers the scores to average J=0.132 and max J=0.784 at threshold 0.5.

When switching to RESNET50, the best thresholds range between 0.3 and 0.7. With respect to VGG16, pigmented network and streaks present the worse performance, while the average J increases for the other three classes. Overall, the union class has slightly higher average performance (average J=0.136) at threshold 0.3.

Surprisingly, the Jaccard indices measured with the RESNET50 maps, which have a resolution limited to 8x8 pixels, are comparable to the ones extracted from the VGG16 models (24x24 pixels). The second hypothesis is that the lower resolution of the RESNET50 maps is compensated by the higher accuracy of the classification model, i.e., a better overall overlap.

6 Second Experiment

We proceed with a deeper analysis by further diving the samples into Melanoma and Nevus, and into correctly vs. wrongly classified samples. The goal is to observe the correlation between the measured J and the correctness of the classification. Here, the Jaccard indices are calculated using the *union* feature and using the best threshold identified in the first experiment, hence on $\mathcal{S}_V^{0.5}$ and $\mathcal{S}_R^{0.3}$. Tables 4 and 5 report the results for VGG16 and RESNET50, respectively.

For VGG16, we can observe that the mean J for correctly classified melanomas (0.135) is similar to the union class average (0.132). However, when melanomas are wrongly classified, the Jaccard index drops to 0.086, meaning that the saliency maps

Table 4. For VGG16, statistics for the union feature as measured by splitting the $\mathcal{S}_V^{0.5}$ dataset in MELanoma and NeVus, either correctly or wrongly classified.

Feature	count	Best Thr.	J-mean (SD)	J-max
MEL-correct	279	0.500	0.135 (0.108)	0.553
MEL-wrong	158	0.500	0.086 (0.089)	0.495
NV-correct	1165	0.500	0.134 (0.147)	0.784
NV-wrong	361	0.500	0.143 (0.145)	0.666

Table 5. For RESNET50, statistics for the union feature as measured by splitting the $\mathcal{S}_R^{0.3}$ dataset MELanoma and NeVus, either correctly or wrongly classified.

Feature	count	Best Thr.	J-mean (SD)	J-max
MEL-correct	314	0.200	0.114 (0.109)	0.564
MEL-wrong	123	0.400	0.132 (0.120)	0.554
NV-correct	1259	0.400	0.144 (0.135)	0.706
NV-wrong	267	0.300	0.127 (0.120)	0.517

diverges from the ground truth. This could effectively help doctors is spotting a wrong classification. The idea is that: if the classifier tells the doctor that the sample is a melanoma, but then the reported saliency areas diverge a lot from what would be manually marked, then doctors can be more easily induced to think that the system is misclassifying the image. For correctly classified nevi, the average J (0.134) is also similar to the full class average (0.132), and for wrongly classified nevi the average J increases to 0.143. This suggests that, for nevi, doctors can better rely of the suggested saliency areas, which helps them in identifying the true area of interest.

To verify if these differences between correct vs. wrong classification are statistically significant, we ran a set of tests on the J indices measured on all items. As the distributions are not normal, we used the Mann-Whitney U-test. Table 6, top, shows that for the VGG16 maps the difference between the two conditions is statistically significant for $\alpha = 0.05$. The same tests are inconclusive for the RESNET50 model (table 6, bottom), for which we couldn't identify a statistical significance.

7 Discussion

Our experiments show that the generation of features masks from threshold saliency maps performs, on the union of the features, at maximum J=0.136. Among the five features, only Pigment Network reaches the same level of accuracy of the union class. This value is less that the half with respect to state-of-the-art networks specialized for pixel-level classification such as U-Net [26] or pyramid pooling [18]. Nevertheless, when considering the union of all the features, threshold saliency maps could still be a valid alternative to ad-hoc pixel-level feature extraction when dedicated features data sets are not available. In fact, the creation of ground truth datasets for feature extraction requires a considerable amount of work, involving experts in tracing the contour of

Table 6. Results of a Mann-Whitney U-test on the Jaccard indices between correctly and wrongly classified classes.

Model	Vs.	U	p-value
VGG16	MEL-cor vs MEL-wr	28747.5	$\mathbf{1.2E-7}$
VGG16	NV-cor vs NV-wr	172027.0	$\mathbf{0.038}$
RESNET50	MEL-cor vs MEL-wr	49693.5	0.8620
RESNET50	NV-cor vs NV-wr	151981.5	0.393

regions of interests, or labeling super-pixels [5]. This is a huge annotation overhead when compared to labeling images with their diagnose class.

The value of the threshold to reach the best J index varies among datasets and features. Since it is not possible to analytically foresee the best threshold of a given dataset, we suggest the development of interactive exploratory visual interfaces, where dermatologists can autonomously control the saliency threshold value in an interactive fashion for exploration.

Our second hypothesis, that higher resolution saliency maps would lead to a higher Jaccard index than lower resolution ones, cannot be confirmed. However, from a decomposition between classes and correctness of classification, it appears that, for higher resolution maps (24x24 pixels on VGG16), saliency maps overlap much better with ground truth features when the classifier is correctly classifying a melanoma (J=0.135) and performance drops when the prediction is incorrect (J=0.086).

In summary, it seems that for the VGG16 model, in case of misclassification of melanoma, the saliency maps have the tendency to draw the attention of the observer to areas that they would rather ignore, thus inducing doctors to question the choice of the machine. This holds only for the VGG16 architecture, whereas this is not true in case of a low the resolution maps produced by RESNET50 (8x8 pixels), thus supporting our first hypothesis (i.e., higher resolution layers deliver better visual explanations).

8 Conclusions and Future Work

In this paper we presented an investigation on how saliency maps (an explainable AI technique) could be used to identify regions of interest in the diagnosis of skin cancer.

Our experiments show that thresholded saliency maps extracted from classifiers perform, in terms of Jaccard index, almost the half w.r.t. deep neural networks specialized for mask prediction. This applies only when using architectures with high resolution saliency. On the contrary, very deep architectures, usually characterized by very low resolution at the last convolution layer, would lead to the generation of maps with less explanatory power.

The long term goal of this research is the development of an interactive reinforcement learning approach involving human practitioners and their feedback to improve attribute detection. Due to the existence of uncertainty and incompleteness in data, the traditional approach of data-driven algorithms fails. In such a scenario, the "human-in-the-loop" approach can retrain a classification model to increase performance based

on the knowledge of domain experts [15]. Starting from a base classifier, trained from a wide set of labeled images, whenever a dermatologist recognizes a wrong classification, he or she provides the correct class and marks the image with the regions of interests (features) that he or she recognizes. The human feedback could then be used to improve the automatic classification performance by comparing the human feedback with the saliency map of the CNN. The measured discrepancy between the two maps could be used to fine-tune the architecture towards higher accuracy [32].

Further, we would like to investigate on better options for thresholding. In this paper, a global threshold, in the range of 0.0 to 1.0, was simultaneously searched and applied to all the saliency map. This allows for an "emersion" of the most relevant region of interests of a *global* scale. However, there might be regions of saliency below the global threshold which are relevant with respect to the *local* surrounding area. To spot local maxima, we could split the maps into tiles, or super-pixels, and iteratively identify multiple local threshold values based on the range of saliency values of each region.

Finally, the current implementation of Grad-CAM returns saliency maps whose range $[0, 1]$ is filled by stretching the range of activation values of the target convolution layer. Each saliency map is forced to use the full activation range, independent of other samples. In so doing, regions of interests are "forced" to emerge, even when the activation values of the inner layer are lower when compared to other images. As future work, we could consider performing saliency normalization according to global statistics (mean and variance) on the tested set.

Acknowledgements

This research is partly funded by the pAItient project (BMG) and the Endowed Chair of Applied Artificial Intelligence (Oldenburg University).

References

1. Arun, N., Gaw, N., Singh, P., Chang, K., Aggarwal, M., Chen, B., Hoebel, K., Gupta, S., Patel, J., Gidwani, M., Adebayo, J., Li, M.D., Kalpathy-Cramer, J.: Assessing the (un)trustworthiness of saliency maps for localizing abnormalities in medical imaging (2020) 3
2. Brinker, T.J., Hekler, A., Enk, A.H., et al.: Deep learning outperformed 136 of 157 dermatologists in a head-to-head dermoscopic melanoma image classification task. European Journal of Cancer **113**, 47–54 (May 2019). https://doi.org/10.1016/j.ejca.2019.04.001 1
3. Brinker, T.J., Hekler, A., Utikal, J.S., et al.: Skin Cancer Classification Using Convolutional Neural Networks: Systematic Review. Journal of Medical Internet Research **20**(10), e11936 (Oct 2018). https://doi.org/10.2196/11936 1
4. Chattopadhay, A., Sarkar, A., Howlader, P., Balasubramanian, V.N.: Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. 2018 IEEE Winter Conference on Applications of Computer Vision (WACV) (Mar 2018). https://doi.org/10.1109/wacv.2018.00097 3
5. Codella, N., Rotemberg, V., Tschandl, P., et al.: Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC) (Feb 2019), arXiv: 1902.03368 2, 9

6. Codella, N.C.F., Gutman, D., Celebi, M.E., et al.: Skin Lesion Analysis Toward Melanoma Detection: A Challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), Hosted by the International Skin Imaging Collaboration (ISIC) (Oct 2017), arXiv: 1710.05006 4
7. Combalia, M., Codella, N.C.F., Rotemberg, V., et al.: BCN20000: Dermoscopic Lesions in the Wild. arXiv:1908.02288 [cs, eess] (Aug 2019), arXiv: 1908.02288 4
8. Curiel-Lewandrowski, C., Novoa, R.A., Berry, E., Celebi, M.E., Codella, N., Giuste, F., Gutman, D., Halpern, A., Leachman, S., Liu, Y., Liu, Y., Reiter, O., Tschandl, P.: Artificial Intelligence Approach in Melanoma. In: Fisher, D.E., Bastian, B.C. (eds.) Melanoma, pp. 1–31. Springer New York, New York, NY (2019). https://doi.org/10.1007/978-1-4614-7322-0_43-1 3
9. Deng, J., Dong, W., Socher, R., 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 4
10. Donahue, J., Jia, Y., Vinyals, O., et al.: DeCAF: A Deep Convolutional Activation Feature for Generic Visual Recognition. In: Proceedings of the 31st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 32, pp. 647–655. PMLR, Bejing, China (Jun 2014) 3
11. Esteva, A., Kuprel, B., Novoa, R.A., et al.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**, 115 (Jan 2017), https://doi.org/10.1038/nature21056 1, 3
12. Gonzalez-Diaz, I.: DermaKNet: Incorporating the Knowledge of Dermatologists to Convolutional Neural Networks for Skin Lesion Diagnosis. IEEE journal of biomedical and health informatics **23**(2), 547–559 (Mar 2019). https://doi.org/10.1109/JBHI.2018.2806962 4
13. Han, S.S., Kim, M.S., Lim, W., Park, G.H., Park, I., Chang, S.E.: Classification of the Clinical Images for Benign and Malignant Cutaneous Tumors Using a Deep Learning Algorithm. Journal of Investigative Dermatology **138**(7), 1529–1538 (Jul 2018). https://doi.org/10.1016/j.jid.2018.01.028, https://linkinghub.elsevier.com/retrieve/pii/S0022202X18301118 3
14. He, K., Zhang, X., Ren, S., Sun, J.: Deep Residual Learning for Image Recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (Jun 2016) 4
15. Holzinger, A.: Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Informatics **3**(2), 119–131 (Jun 2016). https://doi.org/10.1007/s40708-016-0042-6, https://doi.org/10.1007/s40708-016-0042-6 10
16. Holzinger, A., Carrington, A., Müller, H.: Measuring the Quality of Explanations: The System Causability Scale (SCS). KI - Künstliche Intelligenz **34**(2), 193–198 (Jun 2020). https://doi.org/10.1007/s13218-020-00636-z, https://doi.org/10.1007/s13218-020-00636-z 3
17. Jahanifar, M., Tajeddin, N.Z., Asl, B.M., Gooya, A.: Supervised Saliency Map Driven Segmentation of Lesions in Dermoscopic Images. IEEE Journal of Biomedical and Health Informatics **23**(2), 509–518 (Mar 2019). https://doi.org/10.1109/JBHI.2018.2839647, conference Name: IEEE Journal of Biomedical and Health Informatics 4
18. Jahanifar, M., Tajeddin, N.Z., Koohbanani, N.A., et al.: Segmentation of skin lesions and their attributes using multi-scale convolutional neural networks and domain specific augmentations (2018) 3, 8
19. Kapishnikov, A., Bolukbasi, T., Viegas, F., Terry, M.: Xrai: Better attributions through regions. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) (October 2019) 3
20. Khan, M.A., Akram, T., Sharif, M., Saba, T., Javed, K., Lali, I.U., Tanik, U.J., Rehman, A.: Construction of saliency map and hybrid set of features for efficient segmentation and

classification of skin lesion. Microscopy Research and Technique **82**(6), 741–763 (Jun 2019). https://doi.org/10.1002/jemt.23220 4
21. Mishra, N.K., Celebi, M.E.: An Overview of Melanoma Detection in Dermoscopy Images Using Image Processing and Machine Learning (Jan 2016), arXiv: 1601.07843 1, 3
22. Nunnari, F., Bhuvaneshwara, C., Ezema, A.O., Sonntag, D.: A study on the fusion of pixels and patient metadata in cnn-based classification of skin lesion images. In: International Cross-Domain Conference on Machine Learning and Knowledge Extraction CD-MAKE. Springer (2020). https://doi.org/10.1007/978-3-030-57321-8_11 1
23. Nunnari, F., Sonntag, D.: A CNN toolbox for skin cancer classification. CoRR **abs/1908.08187** (2019) 1
24. Nunnari, F., Sonntag, D.: A software toolbox for deploying deep learning decision support systems with xai capabilities. In: Companion of the 2021 ACM SIGCHI Symposium on Engineering Interactive Computing Systems. p. 44–49. EICS '21, Association for Computing Machinery, New York, NY, USA (2021). https://doi.org/10.1145/3459926.3464753, https://doi.org/10.1145/3459926.3464753 5
25. 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) 3
26. 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.) 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 8
27. Selvaraju, R.R., Cogswell, M., Das, A., et al.: Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In: The IEEE International Conference on Computer Vision (ICCV) (Oct 2017) 1, 3, 5
28. Sharif Razavian, A., Azizpour, H., Sullivan, J., Carlsson, S.: CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops (Jun 2014) 3
29. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition (Sep 2014), arXiv: 1409.1556 4
30. Smilkov, D., Thorat, N., Kim, B., et al.: Smoothgrad: removing noise by adding noise (2017) 3
31. Sun, J., Chakraborti, T., Noble, J.A.: A comparative study of explainer modules applied to automated skin lesion classification. In: Atzmüller, M., Kliegr, T., Schmid, U. (eds.) Proceedings of the First International Workshop on Explainable and Interpretable Machine Learning (XI-ML 2020) co-located with the 43rd German Conference on Artificial Intelligence (KI 2020), Bamberg, Germany, September 21, 2020 (Virtual Workshop). CEUR Workshop Proceedings, vol. 2796. CEUR-WS.org (2020), http://ceur-ws.org/Vol-2796/xi-ml-2020_sun.pdf 3
32. Teso, S.: Toward faithful explanatory active learning with self-explainable neural nets. Interactive Adaptive Learning **2444**, 13 (2019) 10
33. Tschandl, P., Rosendahl, C., Kittler, H.: The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific Data **5**(1) (Dec 2018). https://doi.org/10.1038/sdata.2018.161 4
34. Zhou, S., Zhuang, Y., Meng, R.: Multi-Category Skin Lesion Diagnosis Using Dermoscopy Images and Deep CNN Ensembles (2019) 5