Intelligent Systems with Applications 16 (2022) 200126

Contents lists available at ScienceDirect

Intelligent Systems with Applications

journal homepage: www.journals.elsevier.com/intelligent-systems-with-applications

Short communication

Deep learning-based dot and globule segmentation with pixel and blob-based metrics for evaluation☆

Anand K. Nambisan [a], Norsang Lama [a], Thanh Phan [a], Samantha Swinfard [a], Binita Lama [a], Colin Smith [b], Ahmad Rajeh [c], Gehana Patel [d], Jason Hagerty [e], William V. Stoecker [e], Ronald J. Stanley [f,*]

[a] Missouri University of Science & Technology, Rolla, MO 65209, USA
[b] A.T. Still University Medical School, Kirksville, MO 63501, USA
[c] University of Missouri School of Medicine, Columbia, MO 65212 USA
[d] University of Missouri - Columbia, Columbia, MO 65211 USA
[e] S&A Technologies, Rolla, MO 65401, USA
[f] 127 Emerson Electric Company Hall, Missouri University of Science & Technology, Rolla, MO 65209, USA

ARTICLE INFO

Keywords:
Machine learning
Deep learning
Data processing
Melanoma
Globules
Feature segmentation

ABSTRACT

Deep learning (DL) applied to whole dermoscopic images has shown unprecedented accuracy in differentiating images of melanoma from benign lesions. We hypothesize that accuracy in whole-image deep learning suffers because whole lesion analysis lacks an evaluation of dermoscopic structures. DL also suffers a "black box" characterization because it offers only probabilities to the physician and no visible structures. We propose the detection of structures called dots and globules as a means to improve precision in melanoma detection. We compare two encoder-decoder architectures to detect dots and globules: UNET vs. UNET++. For each of these architectures, we compare three pipelines: with test-time augmentation (TTA), without TTA, and without TTA but with checkpoint ensembles. We use an SE-RESNEXT encoder and a symmetric decoder. The pixel-based F1-scores for globule and dot detection based on UNET++ and UNET techniques with checkpoint ensembles were found to be 0.632 and 0.628, respectively. The blob-based UNET++ and UNET F1-scores (50 percent intersection) were 0.696 and 0.685, respectively. This agreement score is over twice the statistical correlation score measured among specialists. We propose UNET++ globule and dot detection as a technique that offers two potential advantages: increased diagnostic accuracy and visible structure detection to better explain DL results and mitigate deep learning's black-box problem. We present a public globule and dot database to aid progress in automatic detection of these structures.

1. Introduction

Invasive melanoma is a form of skin cancer with 99,780 new cases estimated in the USA in 2022 (Siegel et al., 2022). The chances of survival are high if melanoma is diagnosed early, as noted by Noone et al. (2018). Despite this, projections by Rahib et al. (2021) show that the number of melanoma cases will more than double by 2040, becoming the second most prevalent form of cancer by then. This puts emphasis on the importance of engaging in research toward raising awareness and early diagnosis of skin lesions.

In this section, an overview of digitized skin lesion analysis using machine learning techniques is presented.

In the domain of computer-aided diagnosis (CAD) for digital dermoscopy, researchers have focused on a set of tasks to analyze skin lesions for diagnosis. These tasks can be broadly categorized as image enhancement or pre-processing, lesion segmentation, feature, and finally lesion diagnosis or classification (Madooei & Drew, 2016; Adegun & Viriri, 2021). The biggest and most used public datasets for machine learning-based skin lesion research are the ISIC datasets, which were first released at the International Symposium on Biomedical Imaging (ISBI) 2016, by the International Skin Imaging Collaboration (ISIC) (Gutman et al., 2016). This was accompanied by a challenge

☆ Preprint submitted to Intelligent Systems with Applications: August 24, 2022

* Corresponding author.
E-mail address: stanleyj@mst.edu (R.J. Stanley).

https://doi.org/10.1016/j.iswa.2022.200126
Received 18 June 2022; Received in revised form 25 August 2022; Accepted 11 September 2022
Available online 17 September 2022

involving a set of tasks to push CAD-based skin research further. Since then, there have been four more iterations of the challenge, each with a set of skin lesion-related tasks and a dataset (Tschandl et al., 2018; Codella et al., 2019, 2018; Combalia et al., 2019). Many other datasets have also been used in CAD research. Some like PH2 (Mendonça et al., 2013) are small and have only 200 dermoscopic images with three diagnosis classes: common nevi, atypical nevi, and melanoma. Other datasets include the interactive atlas of dermoscopy (Argenziano et al., 2002), which has more than 1000 clinical cases with clinical images, dermoscopy images, histopathology results, and level of difficulty, it was created as a means to train medical personnel. Kassem et al. (2021) provide a succinct review of current machine learning and deep learning approaches using various datasets to highlight some of the prevalent challenges in CAD-based skin lesion research. Adegun & Viriri (2021) in their survey on deep learning techniques in CAD focused on the ISIC 2018 and ISIC 2019 datasets. They conclude that model ensembling, image pre-processing and lesion segmentation improve results for skin lesion classification.

Segmentation of dermatological features has been repeatedly highlighted as one of the more difficult lesion tasks. Codella et al. (2018) have mentioned this in both the post-challenge reviews of the ISIC 2017 () and ISIC 2018 challenges. Barata et al. (2019) provided a comprehensive survey of feature extraction in skin lesion image analysis, including clinically inspired features like negative networks, dots, globules, etc. They found that the time consumed for annotation is the prime reason for the lack of exploration of such features for image analysis. This has also hindered the incorporation of clinical features toward deep learning-based feature segmentation. Recent work done by Cassidy et al. (2022) used a curated combination of multiple ISIC datasets along with in-depth data cleaning strategies and provided benchmarks on multiple test sets. This was done to highlight the biases that occur due to noise and other artifacts in the assessment of the lesion diagnosis classification task. Work has also been done to alleviate the difficulties in creating annotations for tasks in the healthcare domain. Calisto et al. (2017) have proposed touch-based interfaces for medical annotation aimed at radiologists to help during patient follow-ups. Calisto et al. (2021, 2022) show that the integration of AI techniques improves workflow efficiency and reduces work-related cognitive fatigue. Such interfaces can be introduced into the clinic to make the annotation and data collection process across medical disciplines. This would then make it easier to standardize and collect data for later downstream machine learning tasks and statistical analysis.

1.1. Dot and globule segmentation

In this section, we focus on the clinical features of interest: dots and globules. We provide information and context to support its importance in skin lesion diagnosis.

Early diagnosis of melanoma, particularly at the *in situ* stage, yields the best prognosis (Mocellin & Nitti, 2011). However, many cases of melanoma, especially early *in situ* melanomas, are missed by domain experts (Esteva et al., 2017; Ferris et al., 2015; Marchetti et al., 2018). Machine vision techniques incorporating deep learning (DL) have shown higher diagnostic accuracy than dermatologists can achieve (Haenssle et al., 2018; Tschandl et al., 2019). However accuracy of DL methods has not been proven when applied to small-diameter melanomas. “Black dots” and “brown globules” were among the earliest dermoscopic melanoma structures detected and are still considered critical for diagnosis (Pehamberger et al., 1993). But these structures are found in both benign and malignant lesions, thus there is a need to characterize dots and globules precisely in order to use them for melanoma discrimination. These structures may be most useful for discriminating tiny melanomas from benign mimics. Pereira et al. (2022) found irregular dots and globules in 76.5% of small melanomas; these features were among the most discriminatory features for these melanomas. If dots and globules can be precisely delineated, their features can be used to predict melanoma. Xu et al. (2009) found that large globules and varying globule sizes and shapes had the highest odds ratio for melanoma. Maglogiannis and Delibasis (2015) reported automatic dot and globule detection using a multi-resolution inverse non-linear diffusion approach and found that features from the detected structures increased diagnostic accuracy by 6%, primarily by increasing specificity (true-negative rate). The study showed the potential of globules, but conclusions from this study and other studies such as (Barata et al., 2019) are limited because they analyze a limited number of lesions from nonpublic databases and lack specific metrics for assessing the detected structures. The 2017 and 2018 ISIC challenges (Codella et al., 2018, 2019) provide a globule database using Superpixel-based ground truth annotations that include extraneous areas besides globules. These masks include dermoscopic features but do not delineate them precisely. Inexact masks do not allow determination of feature-specific information like the number of instances of a feature within the lesion, variances in shape, structure, and color between instances of the feature across the lesion. Once extracted, these features can be used for other downstream tasks (explainability or classification). An example of ISIC globule annotation is shown, (Fig. 1).

This research develops precise dot and globule masks and presents a DL technique for detecting these masks automatically. We also present a blob-based metric to best ascertain model detection accuracy. The remaining sections include 2 Methodology, 3 Training, 4 Dataset Availability, 5 Hardware and Software Configuration, 6 Results, 7 Discussion, 8 Conclusion, and Future Work.

2. Methodology

In the following subsections, we break down the different steps involved in the curation of the dataset to prepare for model training, the DL models used, and the evaluation metrics to assess model performance.

2.1. Data collection and processing

To create the dataset we selected 539 images, with 501 from the ISIC 2019 dataset (Codella et al., 2018; Combalia et al., 2019; Tschandl et al., 2018) and 38 from a multi-clinic study supported by NIH grants SBIR R43 CA153927-01 and CA101639-02A2. We opted to use the ISIC 2019 dataset as this dataset also has metadata information, and we believe this can lead to the development of multimodal approaches in future work. A researcher, under the supervision of a dermatologist (W.V.S.) marked all regions in the images that contained either a dot or a globule, as defined by a consensus conference (Argenziano et al., 2003). We used a broad definition of dots and globules so that the model can extract the features and their mimics across diagnoses. Dots and globules are dark-brown, black, or gray structures with fairly sharp borders, often roughly round but sometimes irregular.

We split the dataset into 65% training set and 35% test set. This gave us 381 images in the training/validation set, of which 358 were unique images, and the rest were either the same images with more than one annotation mask (which occurs for duplicate images in the image sets) or images with the same lesion but acquired under slightly different conditions. This means that the hold-out test set had 158 images, and all the images were unique. The training/validation dataset is then split into five sets of train and validation sets for 5-fold cross-validation. During the splitting into folds, we ensure that there are no duplicates (the same image or the same lesion) across a pair of train-validation sets. All duplicates are moved into the training set; leaving only unique images in the validation dataset. To show the model’s generalization capabilities, we also tested them on a set of 160 small melanomas. We did this to determine whether the models detect similar structures across the classes. Since dots and globules are important in separating melanoma from benign nevi we want our models to be able to detect similar structures in melanomas.

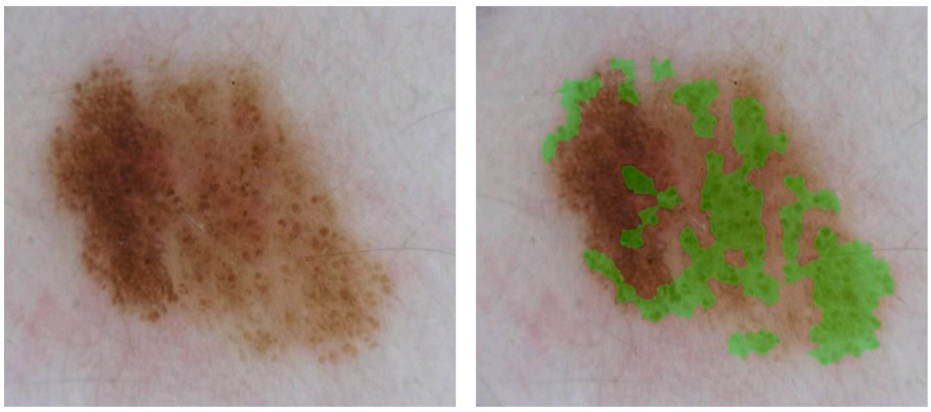

Fig. 1. Nevus (left) with Dot and globules marked (right) in the 2018 ISIC dataset.

2.2. Models

For globule segmentation, we use two variants of the UNET (Ronneberger et al., 2015) architecture, both with the same encoder network but with different decoder and skip connection structures. The two architectures are the usual UNET architecture, as in Ronneberger et al. (2015) and the UNET++ architecture (Zhou et al., 2018). The UNET++ architecture is a modification of the UNET architecture with each encoder level connected to the corresponding decoder level via nested dense convolutions. The architecture also has multiple segmentation branches, each originating from a different level of the encoder network. Two variants of the UNET++ architecture exist — based on how the outputs from the segmentation branches are processed — a fast mode where the final decoder segmentation output is selected and an accurate mode where all the segmentation masks are averaged to generate a final mask. We use the fast mode in our work. The encoder is an SE-ResNext50-32 × 4d network (Hu et al., 2018), which incorporates the squeeze-and-excitation module to learn relationships across the convolutional feature channels. The model is based on the ResNext50 (Xie et al., 2017) model, which expands upon and modifies the blocks in the Resnet model (He et al., 2016) along a new "cardinality" dimension, which in this particular case can be simplified using the notation of grouped convolutions. The decoder is a symmetric decoder based on the encoder, the output mask requirement, and the number of layers required using the implementation in (Yakubovskiy, 2022).

2.3. Evaluation metrics

In related tasks, such as the feature segmentation tasks in the ISIC 2018 and 2017 challenges, the metric used to evaluate the predicted feature masks was pixel-based IOU (Jaccard) (Codella et al., 2019, 2018). The Jaccard metric worked well for superpixel-based ground truth masks but would not be indicative of performance for our precise dataset. An example of a case where pixel-based IOU is a poor metric for dots and globule segmentation can be seen in Fig. 2.

For the PASCAL VOC challenge (Everingham et al., 2015), the intersection over union (IOU) of the predicted bounding boxes with the true boxes is the criterion used for deciding whether the predicted boxes are true-positive. A confidence score-based approach is used to construct an interpolated precision-recall curve to calculate the average precision.

We developed a blob-based approach for metrics to accommodate the subjective nature of globule boundaries. Globule boundaries marked by dermoscopy experts differ significantly between dermatologists and even between globules marked by the same expert at different times (please see discussion). Therefore, detection of an object should be considered successful if there is significant overlap (mask intersection) between the detected (predicted) mask and the ground truth mask.

We define a blob as any disjoint contour in a mask. For the globule segmentation task, a single predicted blob may overlap multiple ground truth globules. This can occur due to the subjective variation in the degree to which an anno-tator chose to combine or separate globules that appear close together.

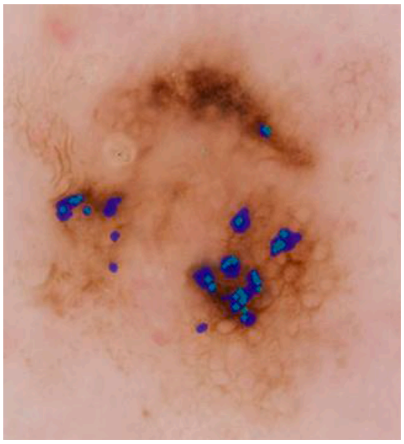

Fig. 2. Image where the models localize well, but blobs are too generous. The pixel-based IOU, precision, recall and F1-scores were 0.0947, 0.994, 0.0947, and 0.173, respectively. The blob-based precision, recall and F1-scores with an intersection threshold of 25% were 0.25, 1.0, and 0.4, respectively, indicative of good localization.

The blob-based Precision and Recall for the whole test set can be calculated using the number of true-positive (Eq. (1)), false-positive (Eq. (2)), and false-negative (Eq. (3)) blobs in an image and accumulating (Eq. (4)) them for the entire dataset.

$$\mathbf{TP} = \sum_{i=1}^{I} \sum_{j=1}^{Ni} f\left(\left|b_{j,i} \cap G_i\right| > T_1 \times \left|b_{j,i}\right|\right), \tag{1}$$

$$\mathbf{FP} = \sum_{i=1}^{I} \sum_{j=1}^{Ni} f\left(\left|b_{j,i} \cap G_i\right| \leq T_1 \times \left|b_{j,i}\right|\right), \tag{2}$$

$$\mathbf{FN} = \sum_{i=1}^{I} \sum_{j=1}^{Mi} f\left(\left|g_{k,i} \cap P_i\right| \leq T_2 \times \left|g_{k,i}\right|\right), \tag{3}$$

and f is the indicator function,

$$f\ (\text{statement}) = \begin{cases} 1, & \textit{if statement true} \\ 0, & \textit{otherwise} \end{cases} \tag{4}$$

Where $|b|$ denotes the number of pixels in the blob b, I is the total number of images in the test set, N_i is the total number of blobs in the predicted mask (P_i) for the ith image, and M_i is the total number of blobs

in the ground truth mask (G_i) for the ith image. $b_{j,i}$ is the jth blob in P_i and $g_{k,i}$ is the kth blob in G_i. Here T_1 and T_2 specify the intersection percentage thresholds (intersection thresholds) for considering a blob as true-positive (or false-positive) and false-negative, respectively, and fall in the open interval (0, 1) on the real line. The thresholds T_1 and T_2 can also be interpreted as thresholds on a predicted blob's precision and ground truth blob's recall, respectively. In our calculations, we set $T_1 = T_2 = T$ for simplicity. We can use *TP, FP,* and *FN* to calculate precision, recall, and the F1-score. A similar approach that relies on centroid distances between blobs instead of overlap was used by Xu et al. (2020). This was done to account for the fact that the ground truth was in the form of blob centroid co-ordinates rather than blob masks.

We calculated these blob-based metrics with three different percentages of intersection thresholds. Those being 25%, 50%, and 75% ($T = 0.25$, 0.5 and 0.75). Our analysis showed that F1-scores are relatively constant up to an intersection of 25% and then fall monotonically as the required overlap percentage increases. The results are shown in Tables 1 and 2.

We also calculated pixel-based metrics on the test dataset to assess segmentation quality. The pixel-based scores are calculated by accumulating the true-positive, false-positive, true-negative, and false-negative pixels across the entire dataset based on Eqs. (1-4), and using the usual definitions of precision, recall, and F1-score to calculate the metrics as done in the blob case above.

3. Training

In this section, we describe our training procedure, augmentations used, and the different testing pipelines.

To train both architectures, we trained the full model with a constant learning rate of 1e-4 using the ADAM (Kingma & Ba, 2022) optimizer and Dice loss, which has shown great promise for medical image segmentation (Zhao et al., 2020). All models are trained for 200 epochs with a 10-epoch early stopping patience on the validation loss. All the models are configured for an input with dimensions 448 × 448 and with 3 channels. During training, we first crop a 448 × 448 patch from the image (zero-padded if required), with random cropping done with a probability of 0.30 and cropping over a mask region with a probability of 0.70. Further color and spatial augmentations are performed before finally feeding it to the model. We also oversampled our dataset so that the number of training samples was 1.5 times the number of images in the training dataset. The output from the model is a 448 × 448 probability mask, which is processed further to get the final predicted mask. When performing validation during training, we employed a similar scheme as in training for crop generation, but no further augmentations were applied. We also saved the best model checkpoint within every 25-epoch window. Since all models get trained for 200 epochs, we have 8 models from training on a single fold, giving us 32 saved models in total across five folds. These saved models are then ensembled with the best model during test-time, similar to the work done in Chen et al. (2022).

Testing a single set of trained weights on the hold-out test set involved extracting overlapping patches of dimensions 448 × 448 from the image, with each patch having an overlap of 50 pixels across both the height and width dimensions. These crops are then passed to the model, which outputs probability masks of dimensions 448 × 448 for each crop. We also implemented another testing pipeline where Test Time Augmentation (TTA) is performed on each cropped patch. The augmentations used are all possible combinations of no-flip and horizontal-flip along with rotations of 0°, 90°, 180°, and 270° giving us 8 augmentations and hence 8 augmented crops. No other spatial or color-based augmentations were performed during testing. These sets of augmented crops are then passed to the model giving us 8 probability masks per crop from the image. The probability masks are de-augmented and averaged together to get the resulting probability mask. These probability masks are then patched back together with the appropriate weighting over overlapping regions to get the full probability mask having the same dimensions as the input image. We then have five probability masks, one from each model trained on one of the five folds. These are averaged to get the final probability mask for the image. A threshold of 0.5 is applied to the probability mask giving the final predicted mask. Finally, we created a testing pipeline that ensembles the checkpointed models by averaging the probability masks from each saved model and the best overall model. In this third pipeline, no TTA was done, and inputs to each model in the ensemble were processed using the same above-mentioned crop-based approach. This lets us fuse the generous predictions of the initial windows with the more precise predictions of the final ones. As in the previous pipelines, a threshold of

Table 1
Pixel-based and blob-based scores for UNET and UNET++ with different test-time approaches. The models were trained on an all-nevi dataset. The testing pipelines were: P1 (without Test Time Augmentation (TTA) or Checkpoint-ensemble approach (CE)), P2 (only TTA), and P3 (only CE). Results are for testing on nevi images. IT stands for intersection threshold. The highest value for each architecture is in bold, and the highest values across both architectures are underlined.

Architecture	UNET			UNET++		
Testing Pipeline	P1	P2	P3	P1	P2	P3
Pixel-based						
IOU	0.443	0.449	**0.458**	0.453	0.459	**0.462**
Precision	**0.676**	0.667	0.659	**0.641**	0.634	0.640
Recall	0.563	0.579	**0.601**	0.607	0.623	**0.624**
F1-score	0.614	0.620	**0.628**	0.623	0.629	**0.632**
Blob-based						
IT= 25%						
Precision	0.537	0.555	**0.577**	0.569	0.583	**0.595**
Recall	**0.856**	0.851	0.844	**0.849**	0.841	0.838
F1-score	0.660	0.672	**0.685**	0.682	0.689	**0.696**
IT = 50%						
Precision	0.502	0.525	**0.544**	0.538	0.555	**0.565**
Recall	**0.793**	0.783	0.779	**0.791**	0.781	0.776
F1-score	0.615	0.629	**0.641**	0.641	0.649	**0.654**
IT = 75%						
Precision	0.349	0.381	**0.389**	0.390	**0.421**	0.414
Recall	0.582	**0.588**	0.576	**0.597**	**0.597**	0.585
F1-score	0.436	0.462	**0.465**	0.472	**0.494**	0.485

Table 2
Pixel-based and blob-based scores for UNET and UNET++ with different test-time approaches. The models were trained on an all-nevi dataset. The testing pipelines were: P1 (without Test Time Augmentation (TTA) or Checkpoint-ensemble approach (CE)), P2 (only TTA), and P3 (only CE). Results are for testing on melanoma images. IT stands for intersection threshold. The highest value for each architecture is in bold, and the highest values across both architectures are underlined.

Architecture	UNET			UNET++		
Testing Pipeline	P1	P2	P3	P1	P2	P3
Pixel-based						
IOU	0.393	**0.401**	0.397	0.385	**0.392**	**0.392**
Precision	**0.551**	0.544	0.539	0.504	0.500	**0.507**
Recall	0.578	**0.605**	0.602	0.619	**0.645**	0.634
F1-score	0.564	**0.574**	0.568	0.556	0.563	**0.564**
Blob-based						
IT= 25%						
Precision	0.660	0.690	**0.701**	0.694	0.713	**0.724**
Recall	**0.590**	0.583	0.566	**0.563**	0.550	0.554
F1-score	0.623	**0.632**	0.626	0.544	0.621	**0.628**
IT = 50%						
Precision	0.558	0.588	**0.600**	0.604	**0.635**	0.633
Recall	**0.510**	0.505	0.492	**0.496**	0.490	0.487
F1-score	0.533	**0.544**	0.540	0.544	**0.553**	0.551
IT = 75%						
Precision	0.366	**0.411**	0.408	0.425	**0.468**	0.443
Recall	0.353	**0.367**	0.347	0.363	**0.370**	0.356
F1-score	0.360	**0.388**	0.375	0.392	**0.413**	0.395

0.5 is applied to get the final predicted mask.

4. Dataset availability

The globule masks and images for the non-public and public datasets used are publicly available and can be accessed https://scholarsmine.mst.edu/research_data/10/.

5. Hardware and software configuration

Our models were trained on an Intel(R) Xeon(R) Silver 4110 CPU (2.10 GHz) with 64 GBs of ram, along with an NVIDIA Quadro P4000 GPU with an 8 GB ram. The models were constructed using the Pytorch (Paszke et al., 2019) library as implemented in Yakubovskiy (2022).

6. Results

A discussion of the results obtained after training and testing with the different pipelines are given below. We compare and tabulate the results from the different architectures.

For blob-based evaluation, true-positive, false-positive, and false-negative blobs are extracted using three different intersection thresholds described in Section 2.3. The intersection percentages were 25%, 50%, and 75%, respectively. Once these have been calculated, we find the blob-based precision, recall, and F1-scores. Table 1 shows the blob-based and pixel-based scores for nevi for both architectures with the different testing pipelines mentioned in Section 3. Fig. 3 shows two cases with high pixel-based IOU scores (0.73 and 0.70) on our best pipeline: UNET++ with checkpoint ensembles, showing correct globule detection. Fig. 4 shows the worst globule detection results. Fig. 5 shows an image with no globules present, and none found.

Since the model was not trained on melanoma images, we would not be able to analyze the differences in the results if we treated the testing datasets (nevi and melanoma) as a single entity, hence we present them separately. The results for testing on small melanomas can be seen in Table. 2. Comparing these scores with those in Table. 1 we see that the blob-based recalls are higher for nevi compared to melanoma and blob-based precisions for melanoma are higher than for nevi. This can be attributed to the fact that the models have never been trained on melanoma images, but when they do predict globules, they are usually correct (higher precision) but fail to get all the blobs marked in the ground truth (lower recall). An example can be seen in Fig. 2.

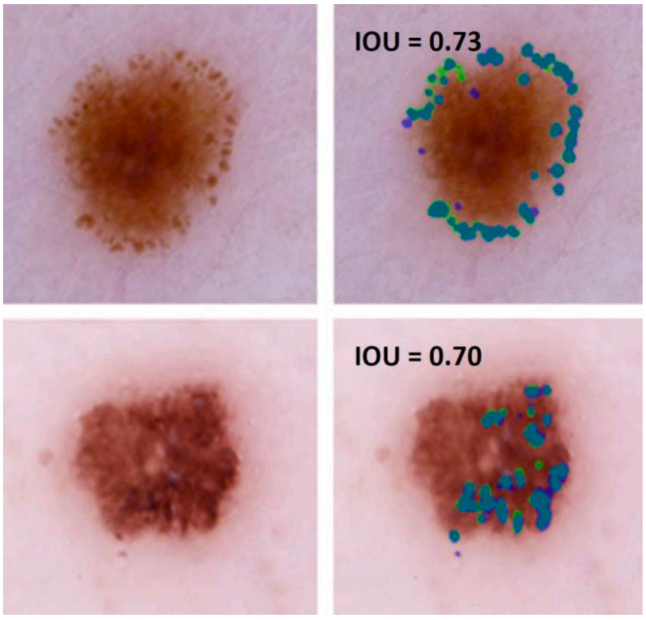

Fig. 3. Images showing two cases with the highest pixel-based IOU scores, shown. The images on the right show the lesion overlayed with both the predicted mask and the ground truth mask. The blue regions are false-positive pixels, the green are false-negatives, and the blue-green regions are true-positives for Figs. 3–5.

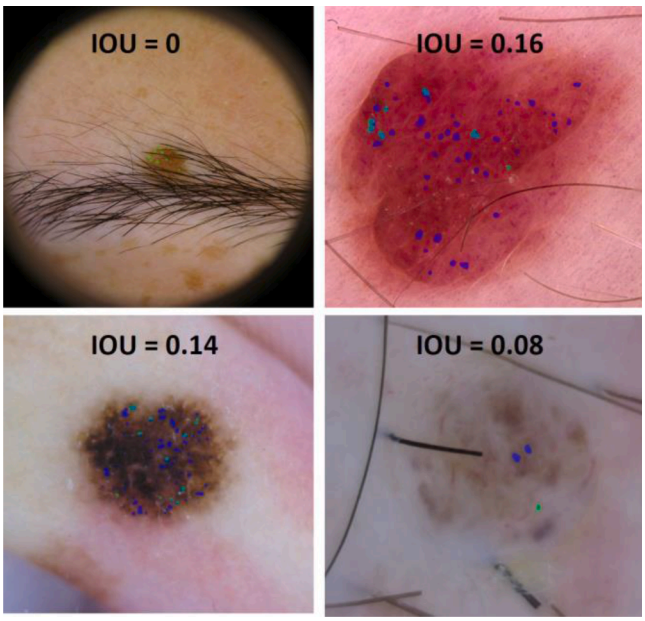

Fig. 4. Image overlays for the 4 cases with the lowest pixel-based IOU scores. The upper left image shows faint false-negative blobs (DL failed to detect). The other 3 images show false-positives for images where annotators did not mark faint blobs.

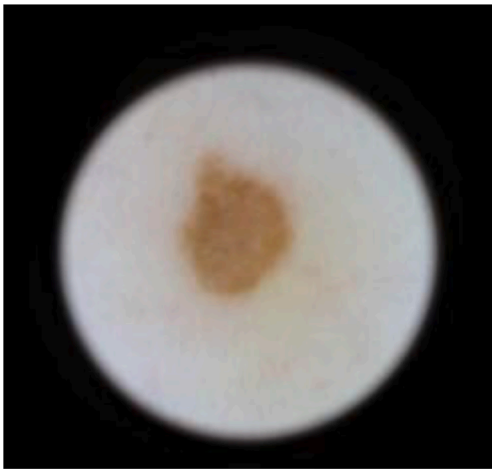

Fig. 5. Image with no dots or globules found, which equals ground truth.

7. Discussion

This study demonstrates that precisely annotated ground truth maps enable high accuracy for deep learning despite the subjective nature of dot and globule assessment. This segmentation accuracy can be achieved with a relatively small number (539) of ground truth masks. We demonstrate this accuracy with two metrics: pixel-based and blob-based. All dermoscopic structures suffer from disagreement among experts. We propose the blob-based metric as a model to better assess whether a structure has been correctly identified and approximately located. Dots and globules have a kappa statistic for interobserver agreement of 0.33 (95% confidence interval 0.32-0.34) (Argenziano et al., 2003). Our 50%

intersection blob-based dot-globule F1-score of 0.654 is over twice the dot-globule interobserver agreement score, indicating that our deep learning model for dot and globule detection shows moderate agreement with the ground truth. By ignoring pixel counting discrepancies resulting from minor shape differences, we better model small structures whose detection is critical within a hierarchical lesion classification pipeline. Our blob-based metric is but one possible biologically inspired metric (Sabbaghi Mahmouei et al., 2019) and shows how such tasks can inspire new metrics for performance evaluation of subjective modeling tasks. The construction of task-specific metrics – like the blob-based metrics presented here – is crucial to proper model assessment. Dots and globules comprise a continuum with no precise limit distinguishing the structures. Therefore, they are processed as a single class. Pigmented lesion classification suffers from high intra-class variability and low inter-class variability. Therefore, this research focuses on a single class, benign nevi, for detecting these structures. This research can be incorporated into a proposed deep learning pipeline that leverages clinical features used by dermatologists. This approach can overcome the black box nature of deep learning and can be used to present a more convincing case for deep learning models for diagnosis, as well as training, in a clinical setting. The structure-based model exploits algorithms used by clinicians to improve the accuracy of melanoma screening. These pipelines can be further analyzed using analytic methods like Grad-Cam (Selvaraju et al., 2017) and saliency maps (Simonyan et al., 2014). These clinical features can also be used within deep learning explainability paradigms such as TCAV (Lucieri et al., 2022) or used by machine learning practitioners for an explainable classification pipeline.

8. Conclusion and future work

In this section, we discuss some aspects of future work that we intend to pursue and provide some ideas on the continuation of the current work.

Prior to the advent of deep learning models for medical computer vision, researchers built highly specialized feature detection/extraction algorithms to find clinically relevant features which would then be leveraged for lesion diagnosis. These tasks can then be incorporated into a pipeline for fully automated lesion diagnosis. Existing feature-based methods that use it for skin lesion diagnosis rely on global features, which include color, shape, texture and other (usually handcrafted) information either from the entire dermoscopic image or just the lesion region (Sheha et al., 2012; Saba et al., 2019; Nasir et al., 2018). Some like Benyahia et al. (2022) have extracted the ac-tivations of intermediate layers in deep learning models (deep learning features) for classification. Hagerty et al. (2019) improved diagnostic accuracy by fusing deep learning and handcrafted features. Yap et al. (2018) used metadata information like age, gender, and lesion location and showed improvement when used with deep learning models. Similar multimodal schemes were used for breast cancer screening by Calisto et al. (2020, 2021).

The whole-image deep learning diagnostic models lack explainability. Explainability is especially desired in a medical application involving a critical decision—whether to biopsy a lesion. Detection and display of an irregular globule, a visible structure associated with melanoma, provides explainability without resorting to an interpretable model (Burkart & Huber, 2021; Molnar, 2022). Multiple researchers have noted the difficulty in using interpretable models in the medical domain. The current study, in providing automatic globule detection, provides a step toward improving explainability by structure detection. Our precise annotated masks can also be used for training deep learning models.

To address these challenges, we present an annotated database and a deep learning method that detects dots and globules with high accuracy. Another objective of the proposed work is to off-load the tedious job of annotating clinical features to deep learning models. We also intend to continue research into biologically inspired metrics like the blob-based metric proposed here, which can improve understanding of model performance on subjective databases. We will explore fuzzy-logic-based metrics and other metrics such as Taha & Hanbury, 2015), which can be used to handle multiple or subjective ground truth databases. Ground truth development will proceed for melanomas, which along with our current database can potentially improve melanoma detection accuracy, especially for small melanomas (Eqs. (1)–((4)).

CRediT authorship statement

Anand K. Nambisan: 0000-0003-4565-4609. **Norsang Lama:** 0000-0002-3580-5736. **Jason Hagerty:** NA. **Colin Smith:** NA. **Ahmad Rajeh:** NA. **Thanh Phan:** NA. **Samantha Swinfard:** NA. **Binita Lama:** NA. **Gehana Patel:** NA. **William V. Stoecker:** 0000-0003-4863-3483. **Ronald J. Stanley:** 0000-0003-0477-3388

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported in part by the National Institutes of Health (NIH) under Grant SBIR R43 CA153927-01 and CA101639-02A2 of the National Institutes of Health (NIH). Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the NIH.

References

A. Adegun, S. Viriri, Deep learning techniques for skin lesion analysis and melanoma cancer detection: A survey of state-of-the-art, Vol. 54 (2), Springer Netherlands, 2021. doi:10.1007/s10462-020-09865-y.

Argenziano, G., Soyer, H. P., Chimenti, S., Talamini, R., Corona, R., Sera, F., Binder, M., Cerroni, L., Rosa, G. D., Ferrara, G., Hofmann-Wellenhof, R., Landthaler, M., Menzies, S. W., Pehamberger, H., Piccolo, D., Rabinovitz, H. S., Schiffner, R., Staibano, S., Stolz, W., Bartenjev, I., Blum, A., Braun, R., Cabo, H., Carli, P., Giorgi, V. D., Fleming, M. G., Grichnik, J. M., Grin, C. M., Halpern, A. C., Johr, R., Katz, B., Kenet, R. O., Kittler, H., Kreusch, J., Malvehy, J., Mazzocchetti, G., Oliviero, M., Ozdemir, F., Peris, K., Perotti, R., Perusquia, A., Pizzichetta, M. A., Puig, S., Rao, B., Rubegni, P., Saida, T., Scalvenzi, M., Seidenari, S., Stanganelli, I., Tanaka, M., Westerhoff, K., Wolf, I. H., Braun-Falco, O., Kerl, H., Nishikawa, T., Wolff, K., & Kopf, A. W. (2003). Dermoscopy of pigmented skin lesions: Results of a consensus meeting via the Internet. *Journal of the American Academy of Dermatology, 48*(5), 679–693.

Argenziano, G. S. H. P., Soyer, H. P., De Giorgi, V., Piccolo, D., Carli, P., & Delfino, M. (2002). *Dermoscopy: a tutorial* (p. 16). New Media: EDRA, Medical Publishing &.

Barata, C., Celebi, M. E., & Marques, J. S. (2019). A survey of feature extraction in dermoscopy image analysis of skin cancer. *IEEE Journal of Biomedical and Health Informatics, 23*(3), 1096–1109. https://doi.org/10.1109/JBHI. 2018.2845939

Benyahia, S., Meftah, B., & Lézoray, O. (2022). Multi-features extraction based on deep learning for skin lesion classification. *Tissue and Cell, 74*, Article 101701. https://doi.org/10.1016/j.tice.2021.101701. URL https://www.sciencedirect.com/science/article/pii/S0040816621002172.

Burkart, N., & Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research, 70*, 245–317.

Calisto, F. M., Ferreira, A., Nascimento, J. C., & Gonçalves, D. (2017). Towards touch-based medical image diagnosis annotation. In *Proceedings of the 2017 ACM international conference on interactive surfaces and spaces, ISS '17* (pp. 390–395). New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3132272.3134111.

Calisto, F. M., Nunes, N., & Nascimento, J. C. (2020). Breastscreening: On the use of multi-modality in medical imaging diagnosis. In *Proceedings of the international conference on advanced visual interfaces, AVI '20*. New York, NY, USA: Association for Computing Machinery. https://doi.org/10.1145/3399715.3399744.

Calisto, F. M., Santiago, C., Nunes, N., & Nascimento, J. C. (2021). Introduction of human-centric Ai assistant to aid radiologists for multimodal breast image classification. *International Journal of Human-Computer Studies, 150*, Article 102607. https://doi.org/10.1016/j.ijhcs.2021.102607. URL https://www.sciencedirect.com/science/article/pii/S1071581921000252.

Calisto, F. M., Santiago, C., Nunes, N., & Nascimento, J. C. (2022). Breastscreening-ai: Evaluating medical intelligent agents for human-ai interactions. *Artificial Intelligence*

in Medicine, 127, Article 102285. https://doi.org/10.1016/j.artmed.2022.102285. URL https://www.sciencedirect.com/science/article/pii/S0933365722000501.

Cassidy, B., Kendrick, C., Brodzicki, A., Jaworek-Korjakowska, J., & Yap, M. H. (2022). Analysis of the isic image datasets: Usage, benchmarks and recommendations. *Medical Image Analysis, 75*, Article 102305. https://doi.org/10.1016/j.media.2021.102305. URL https://www.sciencedirect.com/science/article/pii/S1361841521003509.

H. Chen, S. Lundberg, S.I. Lee, Checkpoint ensembles: Ensemble methods from a single training process, arXiv preprint arXiv:1710.03282.(2022).

N. Codella, V. Rotemberg, P. Tschandl, M.E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, H. Kittler, A. Halpern, Skin lesion analysis toward melanoma detection 2018: A challenge hosted by the international skin imaging collaboration (ISIC) (2019). arXiv:1902.03368.

Codella, N. C., Gutman, D., Celebi, M. E., Helba, B., Marchetti, M. A., Dusza, S. W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., & Halpern, A. (2018). 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). In *, 2018-April. Proceedings of the international symposium on biomedical imaging* (pp. 168–172). IEEE. https://doi.org/10.1109/ISBI.2018.8363547. arXiv:1710.05006.

M. Combalia, N.C.F. Codella, V. Rotemberg, B. Helba, V. Vilaplana, O. Reiter, C. Carrera, A. Barreiro, A.C. Halpern, S. Puig, J. Malvehy, BCN20000: Dermoscopic lesions in the wild (2019). arXiv:1908.02288.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature, 542*(7639), 115–118. https://doi.org/10.1038/nature21056

Everingham, M., Eslami, S. M. A., Van-Gool, L., Williams, C. K. I., Winn, J., & Zisserman, A. (2015). The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision, 111*(1), 98–136.

Ferris, L. K., Harkes, J. A., Gilbert, B., Winger, D. G., Golubets, K., Akilov, O., & Satyanarayanan, M. (2015). Computer-aided classification of melanocytic lesions using dermoscopic images. *Journal of the American Academy of Dermatology, 73*(5), 769–776.

D. Gutman, N.C.F. Codella, E. Celebi, B. Helba, M. Marchetti, N. Mishra, A. Halpern, Skin lesion analysis toward melanoma detection: A challenge at the international symposium on biomedical imaging (ISBI) 2016, hosted by the international skin imaging collaboration (ISIC, 2016) arXiv preprint arXiv:1605.01397.

Haenssle, H. A., Fink, C., Schneiderbauer, R., Toberer, F., Buhl, T., Blum, A., Kalloo, A., Hassen, A. B. H., Thomas, L., & Enk, A. (2018). Others, man against machine: Diagnostic performance of a deep learning convolutional neural network for dermoscopic melanoma recognition in comparison to 58 dermatologists. *Annals of Oncology, 29*(8), 1836–1842.

Hagerty, J. R., Stanley, R. J., Almubarak, H. A., Lama, N., Kasmi, R., Guo, P., Drugge, H. S., Rhett, J., Rabinovitz, H., Oliviero, M., & Stoecker, W. V. (2019). Deep learning and handcrafted method fusion: Higher diagnostic accuracy for melanoma dermoscopy images. *IEEE Journal of Biomedical and Health Informatics, 23*(4), 1385–1391. https://doi.org/10.1109/JBHI.2019.2891049

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Hu, J., Shen, L., & Sun, G. (2018). Squeeze-and-excitation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 7132–7141).

Kassem, M. A., Hosny, K. M., Damaševičius, R., & Eltoukhy, M. M. (2021). Machine learning and deep learning methods for skin lesion classification and diagnosis: A systematic review. *Diagnostics, 11*(8), 1390.

D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980. (2022).

Lucieri, A., Bajwa, M. N., Alexander Braun, S., Malik, M. I., Den-gel, A., & Ahmed, S. (2022). On interpretability of deep learning based skin lesion classifiers using concept activation vectors. In *Proceedings of the international joint conference on neural networksar*. https://doi.org/10.1109/IJCNN48605.2020.9206946. Xiv:2005.02000.

Madooei, A., & Drew, M. S. (2016). Incorporating colour information for computer-aided diagnosis of melanoma from dermoscopy images: A retrospective survey and critical analysis. *International Journal of Biomedical Imaging*.

Maglogiannis, I., & Delibasis, K. K. (2015). Enhancing classification accuracy utilizing globules and dots features in digital dermoscopy. *Computer Methods and Programs in Biomedicine, 118*(2), 124–133.

Marchetti, M. A., Codella, N. C., Dusza, S. W., Gutman, D. A., Helba, B., Kalloo, A., Mishra, N., Carrera, C., Celebi, M. E., DeFazio, J. L., Jaimes, N., Marghoob, A. A., Quigley, E., Scope, A., Yelamos, O., & Halpern, A. C. (2018). Results of the 2016 international skin imaging collaboration international symposium on biomedical imaging challenge: Comparison of the accuracy of computer algorithms to dermatologists for the diagnosis of melanoma from dermoscopic images. *Journal of the American Academy of Dermatology, 78*(2), 270–277.e1. https://doi.org/10.1016/j.jaad.2017.08.016

Mendonca, T., Ferreira, P. M., Marques, J. S., Marcal, A. R., & Rozeira, J. (2013). Ph 2-a dermoscopic image database for research and benchmarking. In *Proceedings of the 35th annual international conference of the IEEE engineering in medicine and biology society (EMBC)* (pp. 5437–5440). IEEE.

Mocellin, S., & Nitti, D. (2011). Cutaneous melanoma *in situ*: Translational evidence from a large population-based study. *The Oncologist, 16*(6), 896–903.

Molnar, C. (2022). *Interpretable machine learning. A guide to making black box models explainable*. Munich, Germany: Leanpub.

Nasir, M., Attique Khan, M., Sharif, M., Lali, I. U., Saba, T., & Iqbal, T. (2018). An improved strategy for skin lesion detection and classification using uniform segmentation and feature selection based approach. *Microscopy Research and Technique, 81*(6), 528–543. https://doi.org/10.1002/jemt.23009. arXiv https://analyticalsciencejournals.onlinelibrary.wiley.com/doi/pdf/10.1002/jemt.23009.

A.M. Noone, N. Howlader, M. Krapcho, D. Miller, A. Brest, M. Yu, J. Ruhl, Z. Tatalovich, A. Mariotto, D.R. Lewis, H.S. Chen, E.J. Feuer, K. A.C. (eds), Seer cancer statistics review, 1975-2015, Tech. rep., National Cancer Institute, Bethesda, MD (April 2018).

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, S. Chintala, PyTorch: An imperative style, high-performance deep learning library, in: H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, R. Garnett (Eds.), Advances in neural information processing systems, Vol. 32, Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

Pehamberger, H., Binder, M., Steiner, A., & Wolff, K. (1993). *In vivo* epiluminescence microscopy: Improvement of early diagnosis of melanoma. *Journal of Investigative Dermatology, 100*(3), S356–S362.

Pereira, A. R., Corral-Forteza, M., Collgros, H., El Sharouni, M. A., Ferguson, P. M., Scolyer, R. A., & Guitera, P. (2022). Dermoscopic features and screening strategies for the detection of small-diameter melanomas. *Clinical and Experimental Dermatology, 47*(5), 932–941.

Rahib, L., Wehner, M. R., Matrisian, L. M., & Nead, K. T. (2021). Estimated projection of US cancer incidence and death to 2040. *JAMA Network Open, 4*(4), Article e214708. https://doi.org/10.1001/jamanetworkopen.2021.4708. e214708arXiv https://jamanetwork.com/journals/jamanetworkopen/articlepdf/2778204/rahib_2021_oi_210166_1617121223.53101.pdf.

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. In *Proceedings of the international conference on medical image computing and computer-assisted intervention* (pp. 234–241). Springer.

Saba, T., Khan, M. A., Rehman, A., & Marie-Sainte, S. L. (2019). Region extraction and classification of skin cancer: A heterogeneous framework of deep cnn features fusion and reduction. *Journal of Medical Systems, 43*(9), 1–19.

Sabbaghi Mahmouei, S., Aldeen, M., Stoecker, W. V., & Garnavi, R. (2019). Biologically inspired QuadTree color detection in dermoscopy images of melanoma. *IEEE Journal of Biomedical and Health Informatics, 23*(2), 570–577. https://doi.org/10.1109/JBHI.2018.2841428

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision* (pp. 618–626). https://doi.org/10.1109/ICCV.2017.74, 2017-Octob.

Sheha, M. A., Mabrouk, M. S., Sharawy, A., et al. (2012). Automatic detection of melanoma skin cancer using texture analysis. *International Journal of Computer Applications, 42*(20), 22–26.

Siegel, R. L., Miller, K. D., Fuchs, H. E., & Jemal, A. (2022). Cancer statistics, 2022. *CA: A Cancer Journal for Clinicians, 72*(1), 7–33. https://doi.org/10.3322/caac.21708. arXiv https://acsjournals.onlinelibrary.wiley.com/doi/pdf/10.3322/caac.21708.

Simonyan, K., Vedaldi, A., & Zisserman, A. (2014). Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Proceedings of the 2nd international conference on learning representations, ICLR 2014 - workshop track proceedings* (pp. 1–8). arXiv:1312.6034.

Taha, A. A., & Hanbury, A. (2015). Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool. *BMC Medical Imaging, 15*(1)), 29. https://doi.org/10.1186/s12880-015-0068-x

Tschandl, P., Codella, N., Akay, B. N., Argenziano, G., Braun, R. P., Cabo, H., Gutman, D., Halpern, A., Helba, B., Hofmann-Wellenhof, R., Lallas, A., Lapins, J., Longo, C., Malvehy, J., Marchetti, M. A., Marghoob, A., Menzies, S., Oakley, A., Paoli, J., Puig, S., Rinner, C., Rosendahl, C., Scope, A., Sinz, C., Soyer, P. H. P., Thomas, P. L., Zalaudek, I., & Kittler, H. (2019). Comparison of the accuracy of human readers versus machine-learning algorithms for pigmented skin lesion classification: An open, web-based, international, diagnostic study. *The Lancet Oncology, 20*(7), 938–947.

Tschandl, P., Rosendahl, C., & Kittler, H. (2018). The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data, 5*, Article 180161.

Xie, S., Girshick, R., Dollár, P., Tu, Z., & He, K. (2017). Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1492–1500).

Xu, J., Gupta, K., Stoecker, W. V., Krishnamurthy, Y., Rabinovitz, H. S., Bangert, A., Calcara, D., Oliviero, M., Malters, J. M., Drugge, R., Stanley, R. J., Moss, R. H., & Celebi, M. E. (2009). Analysis of globule types in malignant melanoma. *Archives of Dermatology, 145*(11), 1245–1251.

Xu, Y., Wu, T., Gao, F., Charlton, J. R., & Bennett, K. M. (2020). Improved small blob detection in 3D images using jointly constrained deep learning and Hessian analysis. *Scientific Reports, 10*(1), 1–12. https://doi.org/10.1038/s41598-019-57223-y

Yakubovskiy, P. (2022). *Segmentation models pytorch*. GitHub Repos.

Yap, J., Yolland, W., & Tschandl, P. (2018). Multimodal skin lesion classification using deep learning. *Experimental Dermatology, 27*(11), 1261–1267. https://doi.org/10.1111/exd.13777. arXiv https://onlinelibrary.wiley.com/doi/pdf/10.1111/exd.13777.

Zhao, R., Qian, B., Zhang, X., Li, Y., Wei, R., Liu, Y., & Pan, Y. (2020). Rethinking dice loss for medical image segmentation. In *Proceedings of the IEEE international conference on data mining (ICDM)* (pp. 851–860). https://doi.org/10.1109/ICDM50108.2020.00094

Zhou, Z., Rahman Siddiquee, M. M., Tajbakhsh, N., & Liang, J. (2018). Unet++: A nested u-net architecture for medical image segmentation. *Deep learning in medical image analysis and multimodal learning for clinical decision support* (pp. 3–11). Springer.