Journal of Machine Learning for Biomedical Imaging. 2021:020. pp 1-26 Submitted 07/2021; Published 12/2021

ENHANCE (ENriching Health data by ANnotations of Crowd and Experts): A case study for skin lesion classification

Ralf Raumanns ralf.raumanns@fontys.nl
Fontys University of Applied Science, Eindhoven, The Netherlands
Eindhoven University of Technology, Eindhoven, The Netherlands

Gerard Schouten g.schouten@fontys.nl
Fontys University of Applied Science, Eindhoven, The Netherlands
Eindhoven University of Technology, Eindhoven, The Netherlands

Max Joosten m.p.h.joosten@student.tue.nl
Eindhoven University of Technology, Eindhoven, The Netherlands

Josien P. W. Pluim j.pluim@tue.nl
Eindhoven University of Technology, Eindhoven, The Netherlands

Veronika Cheplygina vech@itu.dk
IT University of Copenhagen, Denmark

Abstract

We present **ENHANCE**, an open dataset with multiple annotations to complement the existing ISIC and PH2 skin lesion classification datasets. This dataset contains annotations of visual ABC (asymmetry, border, color) features from non-expert annotation sources: undergraduate students, crowd workers from Amazon MTurk and classic image processing algorithms. In this paper we first analyze the correlations between the annotations and the diagnostic label of the lesion, as well as study the agreement between different annotation sources. Overall we find weak correlations of non-expert annotations with the diagnostic label, and low agreement between different annotation sources. Next we study multi-task learning (MTL) with the annotations as additional labels, and show that non-expert annotations improve the diagnostic performance of (ensembles of) state-of-the-art convolutional neural networks. We hope that our dataset can be used in and inspires further research into multiple annotations and/or MTL.

All data and models are available on: `https://github.com/raumannsr/ENHANCE`.

Keywords: Open data, Crowdsourcing, Multi-task learning, Skin cancer, Ensembles, Overfitting

1. Introduction

Machine learning offers many opportunities, but medical imaging datasets, for example for skin lesion diagnosis, are limited, and overfitting can occur. To illustrate, Winkler and colleagues found that superimposed scale bars (Winkler et al., 2021) or skin markings (Winkler et al., 2019) in dermoscopic images may impair the diagnostic performance of the convolutional neural network (CNN) when unintentionally overfitting these artifacts during model training.

©2021 Raumanns, Schouten, Joosten, Pluim and Cheplygina. License: CC-BY 4.0.
https://www.melba-journal.org/papers/2021-020.html.

A promising approach to generalize better in small sample size settings is multi-task learning (MTL), where the model has to learn different tasks simultaneously. This approach showed improved performance in various medical applications, for example, for breast lesions (Shi et al., 2019; Liu et al., 2018). However, when moving from single-task to multi-task models, we need additional annotations. Applying MTL is challenging because datasets typically do not have such additional annotations. Furthermore, building a medical image dataset from scratch with expert annotations is time-consuming and costly.

We present a dataset of additional annotations for skin lesion diagnosis based on non-expert annotations on three dermatoscopic criteria: asymmetry, border and color (so-called ABC criteria). In dermatology, the use of the ABCDE (asymmetry, border, color, diameter, and evolution or elevation) rule is widespread. However, scoring the diameter (D) and evolution or elevation (E) are more complex tasks and therefore less suitable for non-expert annotation. The term *non-expert* is defined here as annotations provided by three different annotation sources: undergraduate students, crowd workers from Amazon MTurk and automated annotations through classic image processing algorithms.

We study the quality of non-expert annotations from different viewpoints. Firstly, we determine the discriminative power of ABC features for diagnosis. We show to what extent the ABC annotations correlate to the diagnosis, and we study how we can use ABC annotations to improve the performance of a CNN. Secondly, the inter agreement level for A, B and C feature between the different annotation sources. The study extends our research on the topic (Raumanns et al., 2020) by using automated annotations as well as comparing the performance on three open source CNN architectures, in particular: VGG-16 (Simonyan and Zisserman, 2015), Inception v3 (Szegedy et al., 2016) and ResNet50 (He et al., 2016) encoders. Further, we investigate whether MTL is also beneficial for automated annotations and show that the performance benefits from using multiple annotations in MTL.

Besides addressing the lack of expert annotations using non-expert ones, we make the dataset with collected ABC annotations and code open, eliminating obstacles for future research. More specifically, the investigation addresses the following research questions:

1. **What is the correlation between the ABC annotations and the diagnostic label?**

2. **How can we use the ABC annotations to improve the performance of a CNN?**

3. **What is the inter agreement level for A, B and C feature between the different annotation sources?**

4. **How can CNN performance benefit from using multiple annotations?**

Our results give valuable insights into the quality of non-expert annotations. Using the collected non-expert annotations in three different CNNs, we show that these are of added value for the performance of the models. This suggests that the use of non-expert annotations might be promising for application in similar domains.

2. Related Work

The shortage of publicly available medical datasets suitable for machine learning is a widely shared problem. Less suitable datasets are small and contain few attributes. Several initiatives have been launched to disseminate and make more medical datasets accessible to alleviate this shortcoming. For example, Tschandl et al. (2018) released the HAM10000 ("Human Against Machine with 10000 training images") dataset containing dermatoscopic images from different populations. Pacheco et al. (2020) released a skin lesion benchmark composed of clinical images collected from smartphone devices and a set of patient clinical data containing up to 21 features.

In machine learning, we usually concentrate on training a single-task model, focusing on a single output. These single-task models have been successful in medical imaging, but there is a potential to make them even more generalizable, thereby preventing overfitting by using MTL that share representations between related tasks. More specific MTL uses inductive transfer with domain-specific information to improve model generalization (Caruana, 1997).

MTL has been successfully in various applications in medical imaging. Hussein et al. (2019) presented a framework for the malignancy determination of lung nodules based on an MTL method using features provided by experts (radiologists). Dhungel et al. (2017) used automatically extracted features in a two-step training process for mass classification in mammograms using a CNN and Random Forest model. Murthy et al. (2017) included shape information in classifying glioma nuclei and showed that they could improve CNN performance compared to a baseline model. Recently we (Raumanns et al., 2020) found that MTL ensembles with the VGG-16 encoder, combined with crowdsourced features, lead to improved generalization for skin lesion classification. For more background information on MTL in deep learning, we refer the reader to an overview written by Ruder (2017).

In this work, we further investigate the use of non-expert annotations for skin lesion classification by comparing non-expert annotations and using them as additional output in three different multi-task models based on commonly employed CNN architectures. Note that the MTL approach allows using ABC annotations on the output side of the network, different from traditional approaches such as Giotis et al. (2013); Cheplygina and Pluim (2018), where they are used as classifier inputs.

3. Dataset creation

This section describes the two public image datasets that we used and details the annotation procedures used to collect additional annotations for these datasets.

3.1 Image datasets

We used two publicly available datasets of dermoscopic images, the training set of the ISIC 2017 challenge dataset (Codella et al., 2017) and the PH2 dataset (Mendonca et al., 2015). A summary of the both datasets is provided in Table 1. There are no other meta-data available than the characteristics presented.

3.1.1 ISIC dataset

The ISIC 2017 challenge was to develop lesion segmentation and lesion classification algorithms. The training dataset contains 2000 dermoscopic images with lesions of three classes: 374 melanoma, 1372 nevi, and 254 seborrheic keratosis lesions. Next to the diagnostic labels, the dataset also contains age and sex for some of the patients, binary lesion segmentation masks, and superpixel segmentation masks (the result of dividing a lesion image with the help of the SLIC algorithm (Achanta et al., 2012)).

3.1.2 PH2 dataset

The PH2 dataset was acquired at the Dermatology Service of Hospital Pedro Hispano, Matosinhos, Portugal. The objective was to create an open dataset to help comparative studies on dermoscopic images' segmentation and classification algorithms. The PH2 dataset contains 200 dermoscopic images, of which 80 common nevi, 80 atypical nevi, and 40 melanoma lesions. The dataset also contains information related to histology and diagnosis, and expert annotations of lesion characteristics including color, asymmetry, and pigment network.

Table 1: Summary of the lesion characteristics in the ISIC and PH2 datasets. A dash (-) means that the information was not provided.

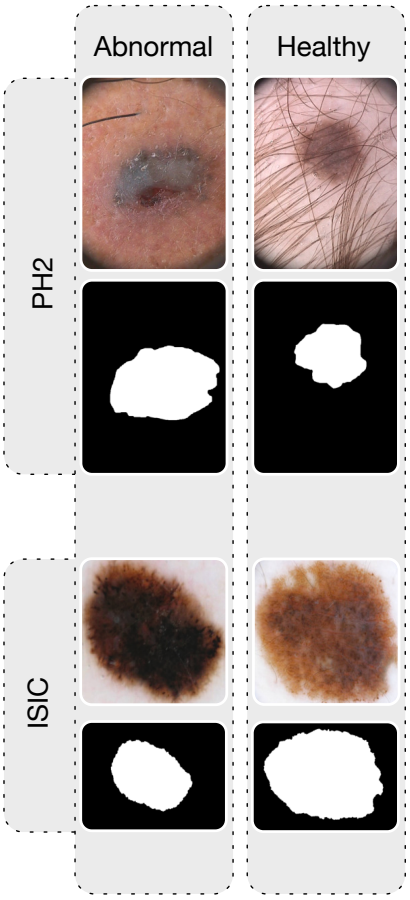
Characteristics		ISIC	PH2
Diagnosis	Benign nevi	1372	160
	Melanoma	374	40
	Seborrheic keratosis	254	0
Patient sex	Female	871	-
	Male	900	-
	Not provided	229	200
Age group	0-10	4	-
	10-20	147	-
	20-30	136	-
	30-40	225	-
	40-50	305	-
	50-60	270	-
	60-70	307	-
	70-80	213	-
	80-90	137	-
	Not provided	256	200
Fitzpatrick phototype	II or III	-	200
	Not provided	2000	0
Expert annotations	Asymmetry	0	200
	Border	0	0
	Color	0	200

3.2 Annotations

We collected annotations from three types of sources:

1. Students of the undergraduate biomedical engineering program at TU Eindhoven,
2. Crowd workers on the Amazon Mechanical Turk platform, and
3. Classical image processing algorithms.

Each source assessed the ABC attributes (Abbasi et al., 2004) as are commonly used by dermatologists: A for asymmetrical shape, B for border irregularity, and C for color of the assessed lesion.



	Asymmetry	Border	Color
Abnormal			
Automated	2	10	5
Crowd	1,2,0	4,8,0	4,4,3
Expert	2	-	2
Healthy			
Automated	2	7	6
Crowd	0,0,0	8,5,7	2,1,1
Expert	0	-	1

	Asymmetry	Border	Color
Abnormal			
Automated	2	50	5
Crowd	1,1,2	8,2,7	1,1,1
Student	1,1,2	5,3,6	4,4,4
Healthy			
Automated	2	62	6
Crowd	0,0,1	7,6,6	3,3,1
Student	2,2,2	5,3,8	3,3,3

Figure 1: The left part shows examples of skin lesions and corresponding masks (black-white image). The top two lesions are from the PH2 dataset; the bottom two are from the ISIC dataset. The right part shows examples of scores from different annotation sources are also given. The student and crowd annotations indicate that at least three different annotators assessed each image.

When training algorithms, multiple annotations per image were combined as follows: we normalized the annotation scores per annotator (resulting in standardized data with mean

0 and standard deviation 1) and then averaged these standardized scores per lesion. Early experiments by (Cheplygina and Pluim, 2018) showed that this was a justifiable strategy.

3.2.1 Student annotations

Undergraduate students without prior knowledge of rating dermoscopic images gathered annotations following a similar protocol as in (Cheplygina and Pluim, 2018), in the context of a medical image analysis course. Each group annotated visual features of 100 images. The students were not blinded to the diagnostic labels, as ISIC is a public dataset. Each group could decide which visual features they wanted to focus on, and the type of annotation scale they wanted to use. However, each visual feature had to be annotated by at least three students from the group. All groups annotated at least some of the ABC features, resulting in 1631 images with asymmetry, border, and color annotations. As we only included the PH2 dataset in the current study after this educational activity was finished, the students did not annotate images from the PH2 dataset.

3.2.2 Crowd annotations

Crowd workers on Amazon Mechanical Turk received instructions of how to rate the ABC features based on three examples. Figures 8, 9 and 10 in (Appendix D. Amazon MTurk annotation) illustrate how the annotation task is presented to crowd workers. The annotation scales were as follows:

- Asymmetry: Zero for symmetrical, one of half-symmetrical, two for asymmetrical.
- Border: The degree of irregularities in the skin lesion's edge for the border score, with a maximum score of eight. A score of eight indicates irregularities in the entire edge, a score of four in the half-edge, and a zero score indicates no irregularities.
- Color: The number of the following colors being present in the lesion: light brown, dark brown, white, blue-gray, black, and red.

We requested three different crowd workers to annotate each image. The diagnostic labels were not provided, but due to the public nature of the dataset, could be potentially be known. Only workers who had done more than 500 approved tasks previously with an approval rate higher than 90% could participate. We paid the workers $0.05 per task. The workers annotated 1250 lesions from ISIC, and all 200 lesions from PH2 with asymmetry, border, and color annotations.

3.2.3 Automated annotations

We used classical image processing algorithms from the literature (Kasmi and Mokrani (2016), Jaworek-Korjakowska (2015) and Achanta et al. (2012)) to automatically extract asymmetry, border and color features. We implemented these algorithms based on the papers, as code was not provided. We explain the different steps of the automated asymmetry, border and color algorithms in Appendix C. Automatic annotation.

Note that we used the provided segmentation masks, and we observed a wide variation of mask boundary definitions: ranging from precisely defined to more loose boundary ones,

including clipped boundaries, therefore not every algorithm could be applied to all images. Table 2 summarizes the number of collected annotations for each annotation source and type.

Table 2: Summary of collected annotations.

	ISIC	PH2
Automated		
Asymmetry (AutoA)	1970[1]	165[1]
Border (AutoB)	1996[1]	0
Color (AutoC)	2000	200
Crowd		
Asymmetry (CrowdA)	1250	200
Border (CrowdB)	1250	200
Color (CrowdC)	1250	200
Student		
Asymmetry (StudA)	1631	0
Border (StudB)	1631	0
Color (StudC)	1631	0
Expert		
Asymmetry (ExpA)	0	200
Border (ExpB)	0	0
Color (ExpC)	0	200

4. Experiments

To investigate the added value of our annotations, we address a binary classification problem: healthy (nevi) vs abnormal (melanoma and seborrheic keratosis).

4.1 Annotation analysis

To gain insight into the characteristics of the collected annotations, we use visual representations that provide detail and overview at the same time. In particular, we created raincloud plots (Allen et al., 2019) with A, B, and C annotations for each annotation source.

We compared the agreement level between the three types of annotation sources. We first averaged the image's scores separately for A, B, and C annotations and then standardized these mean scores, resulting in z-scores. Based on the z-scores, we calculated the agreement level using Pearson correlations.

To quantitatively measure how related the annotation sources and the diagnostic labels (abnormal and healthy) are, we computed the Pearson's correlation coefficient between them for all collected annotations for the ISIC and the PH2 dataset.

We interpret the Pearson's correlation coefficient ρ as follows: weak correlation for $0 \leq \rho < 0.3$, moderate correlation for $0.3 \leq \rho < 0.5$, strong correlation for $0.5 \leq \rho < 1$.

4.2 Multi-task learning

To understand how the collected annotations can contribute to training deep learning algorithms, we performed several experiments where we compared a baseline model which is trained only on the diagnosis, and a MTL model which is trained on both the diagnosis and one or more of the collected ISIC based ABC annotations (Fig. 2). We compared these for three different CNN architectures: VGG-16 (Simonyan and Zisserman, 2015), Inception v3 (Szegedy et al., 2016), and ResNet50 (He et al., 2016).

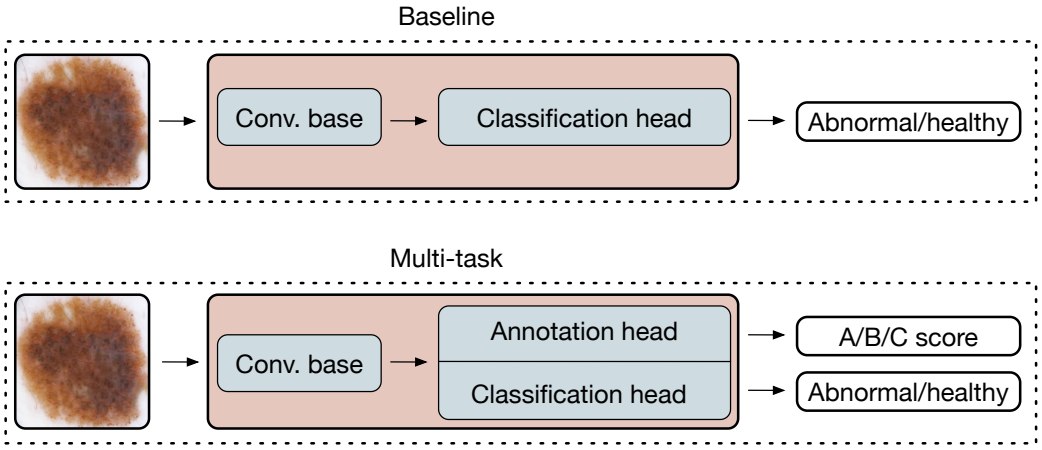


Figure 2: Architecture of baseline (top) and MTL (bottom) models. All models are on top of the VGG-16, Inception v3 or ResNet50 convolutional base.

The **baseline model** extends the convolutional base with two fully connected layers with a sigmoid activation function. During training, we use class weights to pay more attention to samples from the underrepresented class. We use cross-entropy as a loss function. All choices for the baseline model were made based on only the training and validation set. We fine-tune the baseline model until the performance on the validation dataset was within the performance of the ISIC 2017 challenge.

The **multi-task model** extends the convolutional base with three fully connected layers. The model has two outputs with different network heads: one head is the classification output, the other represents the visual feature (such as asymmetry). We use a customized mean squared error loss L_{MTL} and a last layer linear mapping for the annotation-regression task. We used a vector a_i in which we store per lesion whether or not an annotation is available. If an annotation is present, then a_i is equal to 1; otherwise, a_i is 0. This vector is used in a custom loss function that calculates the mean squared error. In case no annotation is available, it does not affect the weights, and the error is not propagated. For the binary classification task, we used a cross-entropy loss L_c and sigmoid activation function of the nodes. The contribution of the losses is equal – we did not optimize the weighting parameter, but early experiments showed that the contribution of the classification loss should not be too low. The resulting loss values are summed and minimized during network training. For our MTL models, the loss function (1) is as follows:

$$L_{MTL} = 0.5(\frac{1}{A}\sum_{i=1}^{N} a_i(y_i - \hat{y}_i)^2) + 0.5L_c \quad (1)$$

where A denotes the number of available annotations, N represents the number of lesions, and y_i and $\hat{y}_i$ denotes respectively the prediction and the expected outcome for lesion i.

To investigate the value of the additional annotations, we compare baseline networks with multi-task networks to which we add one of the following nine types of annotations individually:

- Asymmetry: AutoA, CrowdA, StudA
- Border: AutoB, CrowdB, StudB
- Color: AutoC, CrowdC, StudC

We also combine the networks above into ensembles, by averaging the network predictions with equal weights (earlier experiments (Raumanns et al., 2020) showed limited effectiveness of optimizing the weights). We tested the following three ensembles:

- StudA + StudB + StudC (named studABC)
- CrowdA + CrowdB + CrowdC (named crowdABC)
- AutoA + AutoB + AutoC (named autoABC)

Finally, we trained the ResNet50 MTL models with synthetic annotations. We replaced the original A, B and C scores with uniformly randomized ones in the range [0.0, 1.0] and combined them into ensembles.

The experimental procedure is as follows: we trained all CNN models using 70% as training, 17.5% as validation and 12.5% as test proportions, in a five-fold cross-validation approach. We trained all layers, including the convolutional base, starting with the pre-trained ImageNet weights. For training, we used 30 epochs with a batch size of 20 using the default backpropagation algorithm RMSprop (Tieleman and Hinton, 2012) as the optimizer, with a learning rate of 2.0e−5.

To determine a reasonable learning rate, we evaluated six learning rates on the baseline VGG-16 model (varying over 2 orders of magnitude, between 1.0e-3 and 1.0e-5). We did not vary the learning rate across the different model architectures to keep the number of experiments reasonable while still having a systematic comparison. A description of all hyperparameters is available in Table 6 (Appendix B. Hyperparameters).

We performed the ensemble of the predictions of multi-task models through averaging; we based the lesion's classification on the predictions of the three multi-task models, with each model's prediction having an equal weight (a third).

We compared the average area under the receiver operating characteristic curve (AUC) of the baseline model to the average AUC scores of the different multi-task models and ensembles. The average AUC score was calculated per experiment, taking the average of the AUC score of each fold.

We realized our deep learning models in Keras using the TensorFlow backend (Géron, 2019).

5. Results

5.1 Correlations between annotations and the diagnostic label

The first research question is whether annotations for ABC criteria are discriminative for diagnosis. We show the distributions of asymmetry, border and color versus the label (normal or abnormal) for each annotation source in Fig. 3 (ISIC) and Fig. 4 (PH2). Table 3 lists Pearson correlation coefficient measures between annotation type and diagnostic label for ISIC annotations.

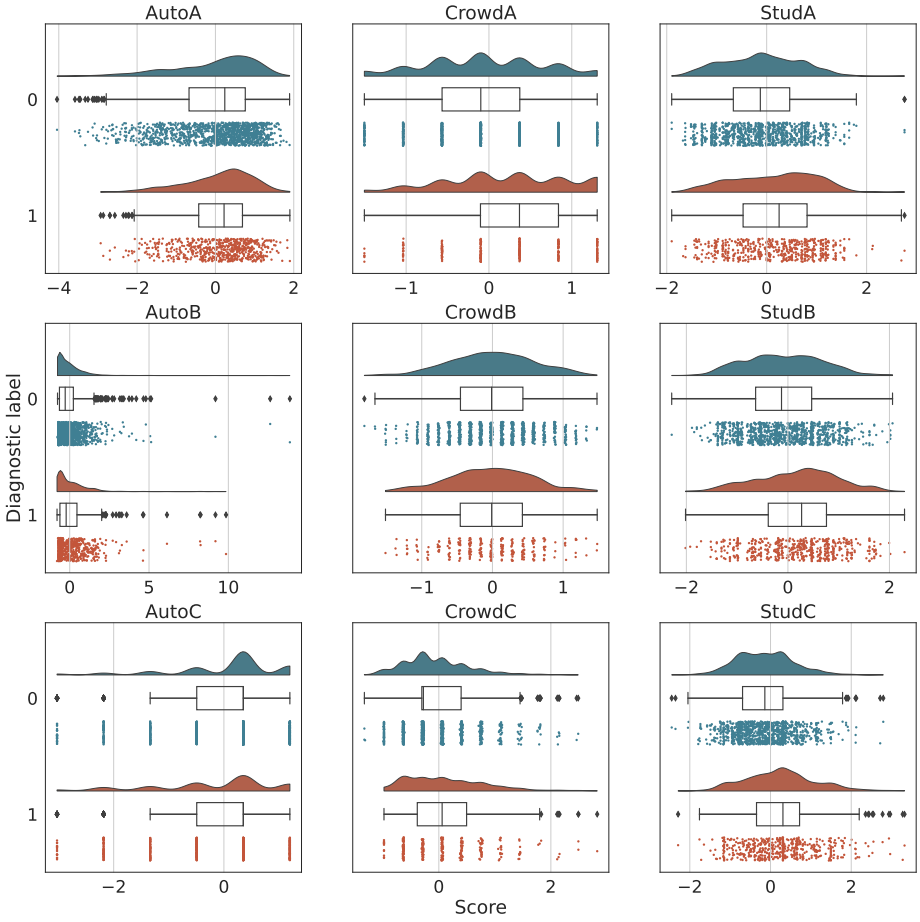


Figure 3: Raincloud plots of automated, crowd, and student ISIC annotations. Diagnostic label 1 stands for abnormal, label 0 for healthy lesions.

The distributions of the A, B and C scores for each diagnostic label overlap considerably. StudA, StudB, StudC, CrowdA, and CrowdC sample medians for abnormal lesions were greater than sample medians for normal ones. AutoA, AutoB, AutoC, and CrowdB sample medians for abnormal and normal lesion are almost identical, although the underlying nature of raw AutoA observations is different. The width of the AutoA annotations belonging to the normal distribution is greater than those of the abnormal distribution.

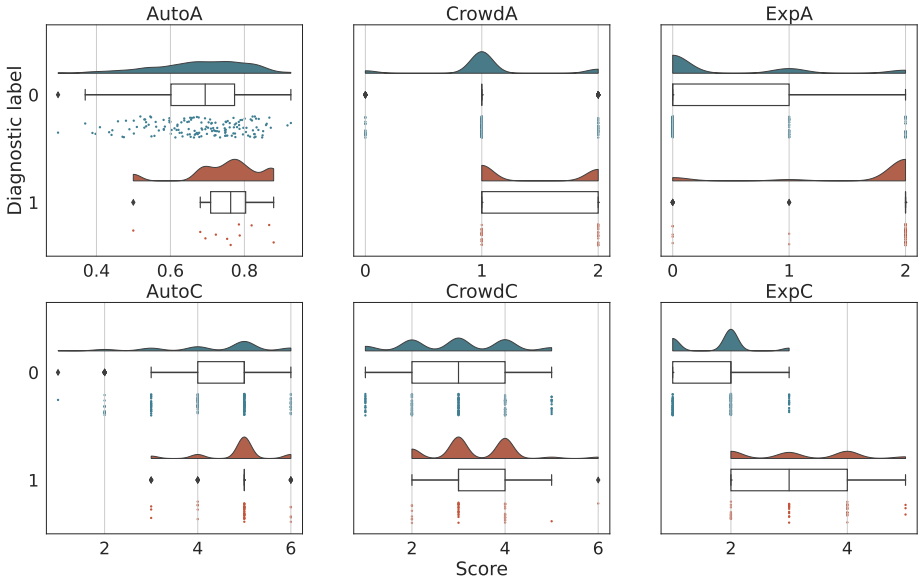


Figure 4: Raincloud plots of automated, crowd, and expert PH2 annotations. Diagnostic label 1 corresponds to abnormal lesions.

Table 3: Pearson correlation coefficient measures between annotation type and diagnostic label

Type	A	B	C
Automated	0.046	0.071	-0.122
Crowd	0.166	-0.017	0.103
Student	0.151	0.161	0.230

The data of ExpA are right-skewed for normal lesions and left-skewed for abnormal ones; their normal and abnormal distributions do not overlap. CrowdA for normal lesions is unimodal; the data is centered around the middle. CrowdA for abnormal lesions is bimodal. The CrowdA's abnormal and normal distributions partially overlap. AutoA's distributions for abnormal and normal lesions overlap considerably, and the sample median for abnormal is greater than the sample median for normal ones.

The distributions of ExpC, CrowdC and AutoC are multimodal. The ExpC's abnormal and normal distributions partially overlap. CrowdC's and AutoC's distribution for abnormal and normal lesions considerably overlap.

The Pearson's coefficient (Table 3) indicates there is no correlation between the crowd-B annotations and the diagnostic label. No correlation also applies to the AutoA and AutoB annotations. We see a weak correlation between all three types of student annotations (StudA, StudB, and StudC), the AutoC annotations and the diagnostic label.

5.2 Inter agreement level between the different annotation sources

The second research question examined the level of agreement between annotation sources for the A, B and C criteria. In order to answer this question, the results of the analysis of agreement level using Pearson correlations were examined. Figure 5 illustrates these correlations graphically.

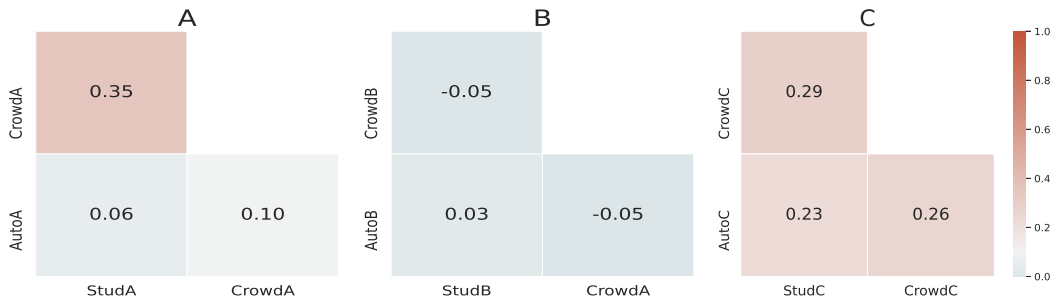


Figure 5: Pearson's correlation coefficient measures for A, B and C features

For asymmetry, we observed a moderate positive correlation between student and crowd annotations. For color, we observed a weak correlation between all annotation sources. We observed no correlation between the different sources for border annotations.

5.3 Effect on CNN performance for multiple annotations usage

Our last research question examines the performance benefit of using multiple annotations in a multi-task model. We used the annotations together with skin lesion images for training multi-task CNNs. We explored how they affect the model's performance compared to a baseline model without using annotations.

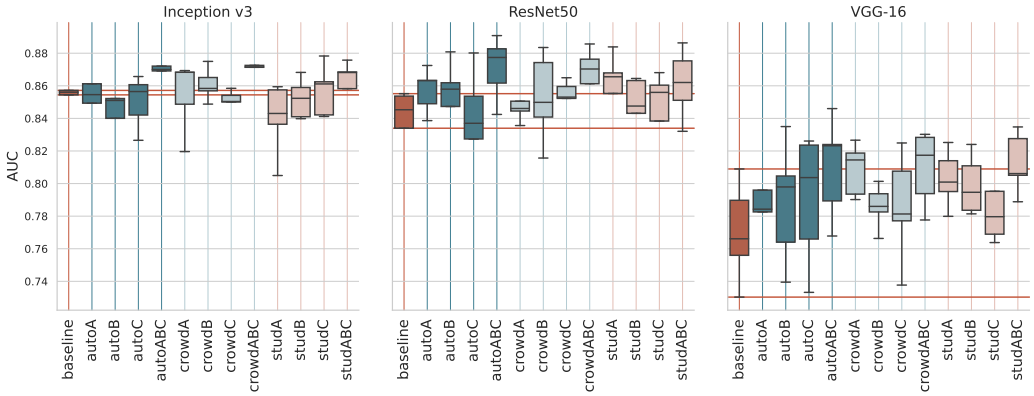


Figure 6: AUCs of baseline and multi-task models (asymmetry, border and color) per CNN architecture.

The AUCs of the baseline and multi-task models are shown in Fig. 6 (the numerical results are also available in Appendix A. AUC performances). Table 4 compares the AUCs of the ensembles based on randomized annotations and original annotations.

Table 4: Mean AUCx100 performances of ResNet50 MTL ensembles trained with original and randomized annotations

Annotations	autoABC	crowdABC	studABC
Original	87.1 ± 1.9	86.3 ± 2.4	86.1 ± 2.1
Randomized	85.0 ± 2.4	84.4 ± 2.8	84.6 ± 2.4

The AUC values for each CNN architecture are close together. The Inception v3 CNN outperformed the ResNet50 and VGG-16 architectures for all three features, but for studA, autoA and autoB multi-task models, ResNet50 performs slightly better. The three ensembles show improved performance when compared to the other models in the same architecture. The best performing model is the autoABC ResNet50 ensemble model. For VGG-16 and ResNet50 the multi-task models outperform the baseline models. For Inception v3 only studC, studABC, crowdA, crowdB, crowdABC, autoA and autoABC models outperform the baseline model. Table 4 shows that ResNet50 MTL ensembles trained with retrieved annotations perform better than ensembles with randomized annotations.

6. Discussion

6.1 Summary of results

6.1.1 At most moderate correlations between annotations and the diagnostic label

In general, the acquired non-expert annotations (especially the student annotations) show a weak correlation with the diagnostic label (Table 3). At the same time, automated annotations show only a correlation for the color feature. Results in a study by Giotis et al. (2013) support this. They show that the color feature has a discriminative power, but the border feature not. In their study, they did not assess asymmetry.

6.1.2 Weak to moderate inter agreement level between the different annotation sources

We observed low agreement between student and crowd for asymmetry and color scores and comparable agreement level between automated and non-expert color features. In contrast, we observed disagreement for the border feature between non-experts and automated scores. Based on these agreement levels, annotating the asymmetry and color seems more straightforward than annotating the border.

Measures with disagreement can still be informative. According to Cheplygina et al., "Although removing disagreement leads to the best performances, we show that disagreement alone leads to better-than-random results. Therefore, the disagreement of these crowd labels could be an advantage when training a skin lesion classifier with these crowd labels as additional outputs" Cheplygina and Pluim (2018).

6.1.3 A mainly positive effect on CNN performance for multiple annotations usage

Our preprint (Raumanns et al., 2020) showed that VGG-16 based multi-task models with individual crowdsourced features have a limited effect on the model. Still, when combined in ensembles, it leads to improved generalization. This paper has further confirmed this improved generalization for Inception v3 CNN and the ResNet50 based multi-task models. We also showed that the Inception v3 CNN outperformed the ResNet50 and VGG-16 architectures for all three annotations. The ResNet50 showed improved performance compared to the VGG-16 model.

One perhaps surprising result is that noisy annotations positively contribute to the performance. Our annotations are highly noisy, as shown by the correlation coefficients in Table 3 and the raincloud plots in Fig. 3. We explain the success of these annotations by several factors. First, annotations that are perfectly correlated with the diagnostic label would not add information to the classifier. Second, training multiple (biased) classifiers and combining these in an ensemble is a well-known result that can lead to improved generalization (Jain et al., 2000). We also show that this is the case for fully randomized annotations (Table 4), but that this is less effective than our gathered annotations. Finally, the classifier also learns from the diagnostic labels; learning from only the annotations is not feasible.

6.2 Limitations

We demonstrated the added value of using non-expert and automated annotations for model learning. Yet, we have not optimally tuned the algorithms for automatically scoring the ABC features. Also, the algorithms depend on the available masks. The quality of these masks differs and affects the automated score of asymmetry and border irregularity. Furthermore, we did not optimize the multi-task model to achieve the highest performance.

There are three main differences between the annotation by the crowd and the students. Firstly, the student annotations were done in an educational setting, and students were encouraged to decide their annotation protocols based on existing literature. Letting the crowd decide on their protocol was not feasible due to the distributed nature of crowdsourcing: the crowd workers do not interact with each other, and most annotate only a few images. Secondly, students were not blinded to the diagnostic label while annotating. Since it is a public dataset, anybody could try to find the label, but this is less likely for MTurk because the MTurk jobs are time-limited. Thirdly, only the ISIC skin lesions were part of the student assignment. The PH2 dataset was only added to this study after an informal review round after the subject was no longer taught.

Thompson et al. (2016) conclude that personal characteristics (such as sex and age) influence healthcare-seeking behavior; thus, the datasets might have an unintended bias against subjects from specific groups. Abbasi-Sureshjani et al. (2020) showed that even when the training set is relatively balanced, there is a variable performance on subgroups based on age and sex. For the ISIC dataset, the distribution between men and women is roughly the same. However, the dataset is skewed towards lower age groups. Unfortunately, the sex and age characteristics are missing for the PH2 dataset. Besides these, the Fitzpatrick phototype (Sachdeva, 2009) is another characteristic that would be relevant in this

prediction problem. The PH2 dataset contains lesions with phototype II or III (Mendonca et al., 2015), but it is unspecified per lesion, and the ISIC database does not have phototype information. The discussed unintended bias or missing demographic characteristics might lead to representation bias, ultimately leading to an insufficiently representative model.

6.3 Opportunities

In this study, we used automated annotations based on best in class image algorithms. However, the automated annotations are not as discriminative as the expert annotations for diagnosis. A future study could concentrate on tuning the used image algorithms or finding new algorithms that mimic expert annotations more closely.

Another important direction for future research is explainability. We showed that multi-task models with non-expert annotations have a limited effect on the model performance. However, the multi-task prediction of ABC scores together with a diagnosis helps to explain the diagnostic outcome. For example, when the model predicts abnormality and high ABC scores, the latter could be interpreted as warning signs/explanation for abnormality. The use of the model in this way needs further research. For instance, what kind of model architectures fit the best for this explanation task? What types of annotations are suitable for this task? What kind of model produces gradual class activation maps (Selvaraju et al., 2017) that explain the model outcome to dermatologists?

We used three different sources to collect annotations for the ISIC and PH2 dataset to create a new open dataset that incorporates these additional annotations and then used this dataset to see whether a multi-task model can take advantage of the extra annotations. A future study, for example, might consider clarifying the multi-task model performance benefit of additional annotations for mammography in the domain of breast cancer diagnosis or studying the quality of the extra annotations in another field.

The annotators' expertise levels will probably be different. Depending on the level of expertise, an annotation can have more or less influence during the model's training procedure. This could be possible via different weights or by modeling the expertise of each annotator, similar to Guan et al. (2017). However, this might be more applicable for few annotators who annotated all images, rather than many annotators who each annotate a few images, as in crowdsourcing.

This study used the ground truth masks' segmentation masks for extracting asymmetry, border irregularity, and color characteristics. We see a wide variation in the given segmentation masks, with boundaries ranging from precisely defined to more loose ones. Some masks also have clipped boundaries. Optimizing these masks need to be considered in future studies.

7. Conclusions

This paper presented ENHANCE, an open dataset with additional annotations, gathered by three different annotation sources, for ISIC 2017 and PH2 images of skin lesions, describing the lesions' asymmetry, border, and color. We tested the dataset extensively with three different network architectures (VGG-16, ResNet50 and Inception c3) and show that the often claimed enhancement of multiple tasks is marginal.

The asymmetry and color annotations from different annotation sources show their discriminative power to diagnose healthy or abnormal. At the same time, there is a moderate agreement level between the annotation sources. Nevertheless, this disagreement could be advantageous when training a skin lesion classifier with the annotations as additional output.

Acknowledgments

We would like to acknowledge all students and crowd workers who have contributed to this project with image annotation. We gratefully acknowledge financial support from the Netherlands Organization for Scientific Research (NWO), grant no. 023.014.010.

Ethical Standards

The work follows appropriate ethical standards in conducting research and writing the manuscript, following all applicable laws and regulations regarding treatment of animals or human subjects.

Conflicts of Interest

We declare we don't have conflicts of interest.

References

Naheed R Abbasi, Helen M Shaw, Darrell S Rigel, Robert J Friedman, William H McCarthy, Iman Osman, Alfred W Kopf, and David Polsky. Early diagnosis of cutaneous melanoma: revisiting the ABCD criteria. *JAMA*, 292(22):2771–2776, December 2004.

Samaneh Abbasi-Sureshjani, Ralf Raumanns, Britt E J Michels, Gerard Schouten, and Veronika Cheplygina. Risk of training diagnostic algorithms on data with demographic bias. In *Interpretable and Annotation-Efficient Learning for Medical Image Computing*, pages 183–192. Springer International Publishing, 2020.

Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(11):2274–2282, November 2012.

Micah Allen, Davide Poggiali, Kirstie Whitaker, Tom Rhys Marshall, and Rogier A Kievit. Raincloud plots: a multi-platform tool for robust data visualization. *Wellcome Open Res*, 4:63, April 2019.

Rich Caruana. Multitask learning. *Mach. Learn.*, 28(1):41–75, July 1997.

Veronika Cheplygina and Josien P W Pluim. Crowd disagreement about medical images is informative. In *Intravascular Imaging and Computer Assisted Stenting and Large-*

Scale Annotation of Biomedical Data and Expert Label Synthesis, pages 105–111. Springer International Publishing, 2018.

Noel C F Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, and Allan Halpern. 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). *arXiv preprint arXiv:1710.05006*, October 2017.

Neeraj Dhungel, Gustavo Carneiro, and Andrew P Bradley. A deep learning approach for the analysis of masses in mammograms with minimal user intervention. *Medical Image Analysis*, 37:114–128, 2017.

Aurélien Géron. *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*. "O'Reilly Media, Inc.", September 2019.

Ioannis Giotis, Margaretha Visser, Marcel Jonkman, and Nicolai Petkov. Discriminative power of visual attributes in dermatology. *Skin Res. Technol.*, 19(1):e123–31, February 2013.

Melody Y Guan, Varun Gulshan, Andrew M Dai, and Geoffrey E Hinton. Who said what: Modeling individual labelers improves classification. *arXiv preprint arXiv:1703.08774*, 2017.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Sarfaraz Hussein, Pujan Kandel, Candice W Bolan, Michael B Wallace, and Ulas Bagci. Lung and pancreatic tumor characterization in the deep learning era: Novel supervised and unsupervised learning approaches. *IEEE Transactions on Medical Imaging*, 38(8): 1777–1787, 2019.

Anil K Jain, Robert P. W. Duin, and Jianchang Mao. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(1):4–37, 2000.

Joanna Jaworek-Korjakowska. Novel method for border irregularity assessment in dermoscopic color images. *Comput. Math. Methods Med.*, 2015:496202, October 2015.

Reda Kasmi and Karim Mokrani. Classification of malignant melanoma and benign skin lesions: implementation of automatic ABCD rule. *IET Image Processing*, 10(6):448–455, 2016.

Jiali Liu, Wanyu Li, Ningbo Zhao, Kunlin Cao, Youbing Yin, Qi Song, Hanbo Chen, and Xuehao Gong. Integrate domain knowledge in training CNN for ultrasonography breast cancer diagnosis. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2018*, pages 868–875. Springer International Publishing, 2018.

T F Mendonca, M E Celebi, T Mendonca, and J S Marques. Ph2: A public database for the analysis of dermoscopic images. *Dermoscopy image analysis*, 2015.

Veda Murthy, Le Hou, Dimitris Samaras, Tahsin M Kurc, and Joel H Saltz. Center-focusing multi-task CNN with injected features for classification of glioma nuclear images. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 834–841. IEEE, 2017.

Andre G C Pacheco, Gustavo R Lima, Amanda S Salomão, Breno Krohling, Igor P Biral, Gabriel G de Angelo, Fábio C R Alves, Jr, José G M Esgario, Alana C Simora, Pedro B C Castro, Felipe B Rodrigues, Patricia H L Frasson, Renato A Krohling, Helder Knidel, Maria C S Santos, Rachel B do Espírito Santo, Telma L S G Macedo, Tania R P Canuto, and Luíz F S de Barros. PAD-UFES-20: A skin lesion dataset composed of patient data and clinical images collected from smartphones. *Data Brief*, 32:106221, October 2020.

Ralf Raumanns, Elif K Contar, Gerard Schouten, and Veronika Cheplygina. Multi-task ensembles with crowdsourced features improve skin lesion diagnosis. *arXiv preprint arXiv:2004. 14745*, 2020.

Sebastian Ruder. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*, 2017.

Silonie Sachdeva. Fitzpatrick skin typing: applications in dermatology. *Indian J. Dermatol. Venereol. Leprol.*, 75(1):93–96, January 2009.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

Jialin Shi, Department of Electrical Engineering, Tsinghua University, Beijing, China, Ji Wu, Ping Lv, and Jiajia Guo. BreastNet: Entropy-Regularized transferable multi-task learning for classification with limited breast data, 2019.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2015.

Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016.

Ashley E Thompson, Yvonne Anisimowicz, Baukje Miedema, William Hogg, Walter P Wodchis, and Kris Aubrey-Bassler. The influence of gender and other patient characteristics on health care-seeking behaviour: a QUALICOPC study. *BMC Fam. Pract.*, 17:38, March 2016.

Tijmen Tieleman and Geoffrey Hinton. Rmsprop: Divide the gradient by a running average of its recent magnitude. coursera: Neural networks for machine learning. *COURSERA Neural Networks Mach. Learn*, 2012.

Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific Data*, 5(1), 2018.

Julia K. Winkler, Christine Fink, Ferdinand Toberer, Alexander Enk, Teresa Deinlein, Rainer Hofmann-Wellenhof, Luc Thomas, Aimilios Lallas, Andreas Blum, Wilhelm Stolz, and Holger A. Haenssle. 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):1135–1141, 10 2019. ISSN 2168-6068. doi: 10.1001/jamadermatol.2019.1735. URL `https://doi.org/10.1001/jamadermatol.2019.1735`.

Julia K Winkler, Katharina Sies, Christine Fink, Ferdinand Toberer, Alexander Enk, Mohamed S Abassi, Tobias Fuchs, and Holger A Haenssle. Association between different scale bars in dermoscopic images and diagnostic performance of a market-approved deep learning convolutional neural network for melanoma recognition. *Eur. J. Cancer*, 145: 146–154, March 2021.

Appendix A. AUC performances

Table 5 shows the mean AUCx100 of 5-fold cross-validation of Inception v3, ResNet50 and VGG-16 CNN baseline and multi-task models. We trained the multi-task models on all features (A, B and C) from three different annotation sources (student, crowd, and automated). The AUC of the best performing model per CNN architecture is emphasized in bold.

Table 5: Mean AUCx100 performances of 5-fold cross-validation on all features (**A**symmetry, **B**order and **C**olor) for three different CNNs.

Type	Inception v3	ResNet50	VGG-16
baseline	85.3 ± 1.2	83.6 ± 2.7	77.0 ± 3.0
autoA	85.4 ± 2.3	85.7 ± 1.3	79.0 ± 2.6
autoB	84.8 ± 2.0	85.3 ± 2.4	78.8 ± 3.7
autoC	85.0 ± 1.6	84.5 ± 2.2	79.1 ± 4.0
autoABC	86.4 ± 1.8	**87.1** ± 1.9	81.0 ± 3.1
crowdA	85.5 ± 2.2	84.8 ± 1.0	80.9 ± 1.6
crowdB	86.1 ± 1.0	85.3 ± 2.7	78.6 ± 1.3
crowdC	85.1 ± 0.6	84.2 ± 3.5	78.6 ± 3.3
crowdABC	**86.7** ± 1.4	86.3 ± 2.4	80.9 ± 2.3
studA	84.0 ± 2.2	85.8 ± 2.6	80.3 ± 1.7
studB	85.2 ± 1.2	84.6 ± 2.1	79.9 ± 1.8
studC	85.7 ± 1.6	84.4 ± 2.7	78.9 ± 2.9
studABC	86.2 ± 1.4	86.1 ± 2.1	**81.2** ± 1.9

Appendix B. Hyperparameters

The table below describes the hyperparameter choices for baseline and MTL experiments.

Table 6: Describes each parameter, its value, and, when applicable, the name in the Python code

Hyperparameter	Value baseline	Value MTL	Name in code
Learning rate	2.0e−5	2.0e−5	
Optimizer	RMSprop	RMSprop	
Batch size	20	20	BATCH_SIZE
Activation function	Sigmoid	Sigmoid (classification) Linear (annotation)	
Number of epochs	30	30	EPOCHS
Steps per epoch	100	100	STEPS_PER_EPOCH
Steps to validate	50	50	VALIDATION_STEPS
Steps to yield from pred.generator	20	20	PREDICTION_STEPS
Class weights	balanced	balanced	
Input shape	(384, 384, 3)	(384, 384, 3)	INPUT_SHAPE
Include top	False	False	
Pre-training	Imagenet	Imagenet	

Appendix C. Automatic annotation

This appendix explains the different steps of automated asymmetry, border and color algorithms. Fig. 7 presents a schematic overview of the algorithm annotation steps.

As a preprocessing step, we first rotate and center the provided segmentation mask, such that the major axis (longest diameter of the lesion) passes horizontally through the segmentation center of mass. This is the major axis, the minor axis is then defined as the diameter perpendicular to the major axis and also passing through the center of mass.

The asymmetry algorithm is based on shape asymmetry (Kasmi and Mokrani, 2016). The asymmetry score algorithm steps are:

1. flip the rotated image over both axes separately to measure the overlap in pixels between the mask areas on either side of the axis;

2. determine the shape symmetry ratio for an axis by dividing the overlap in pixels by the total amount of pixels in the lesion's mask;

3. calculate the score by averaging the symmetry ratios of both axes.

The border score is based on (Jaworek-Korjakowska, 2015) and assesses the border irregularity as follows:

1. calculate a bounding box around the segmentation and connect four lines between the bounding box's vertices and the center of mass;

2. locate the border pixels that are on each of the four lines (resulting in a total of four border pixels);
3. divide the border into four parts with the help of the four pixels (clockwise: top, right, bottom and left part border pixels);
4. calculate the shortest distance between the border pixel and the edge's image for each of the four parts (the direction in which the distances are calculated varies per part; for the top border pixels in the upward direction, the right part border pixels in the rightward direction, the bottom border pixels in the downward direction and the left border pixels in the leftward direction);
5. create a borderline function based on the pixel location on the border and the calculated distances;
6. smooth the borderline function with a Gaussian filter;
7. count the number of turning points in the smoothed signal, resulting in the actual border score.

The color algorithm is based on (Kasmi and Mokrani, 2016) and computes the number of the suspicious colors present in the lesion: light brown, dark brown, white, blue-gray, black, and red, as follows:

1. extract the region of interest from the lesion image using the mask;
2. segment the region of interest into superpixels by using the SLIC superpixels algorithm (Achanta et al., 2012);
3. find all unique colors (having different RGB values) that are present in the SLIC superpixels;
4. measure the normalized Euclidean distance between each unique color in the set of SLIC superpixels and the six predefined values for the suspicious colors;
5. link a unique color to a suspicious color when its distance is less than a fixed threshold (0.4);
6. count a linked unique color as a suspicious color when the number of pixels belonging to that unique color exceeds five per cent of the total amount of pixels of the region of interest;

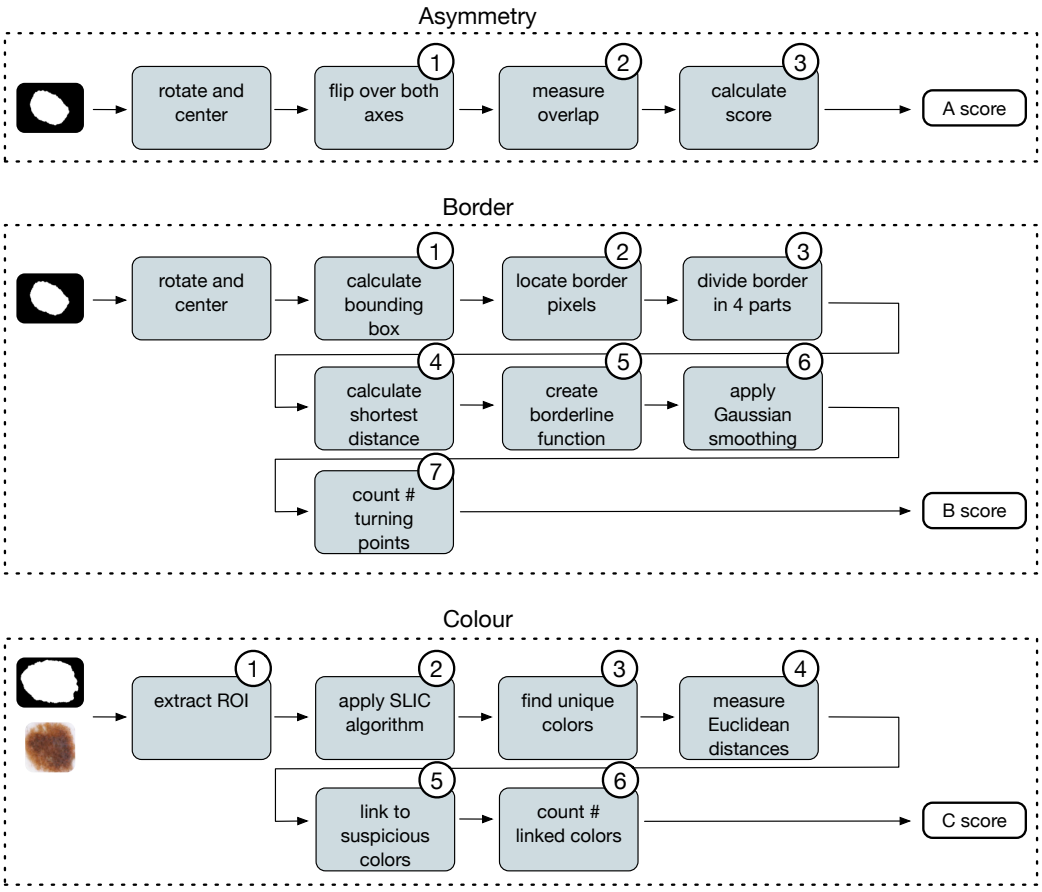


Figure 7: A schematic overview of the different steps of the automated annotation of asymmetry, border and color.

Appendix D. Amazon MTurk annotation

The following three figures represent the screens presented to the crowd workers by the online AmazonMechanical Turk (`www.mturk.com`) crowdsourcing tool. First, the workers receive instruction (Figure 8) to assess the skin lesions. Second, the tool presents three examples (Figure 9) that illustrate scoring asymmetry, border and color. Third, the crowdsourcing tool presents one by one the skin lesions (Figure 10). The crowd workers annotate each presented skin lesion by using the scoring sliders and pressing the submit button. With each skin lesion presented, the MTurk crowdsourcing tool shows the instructions and examples.

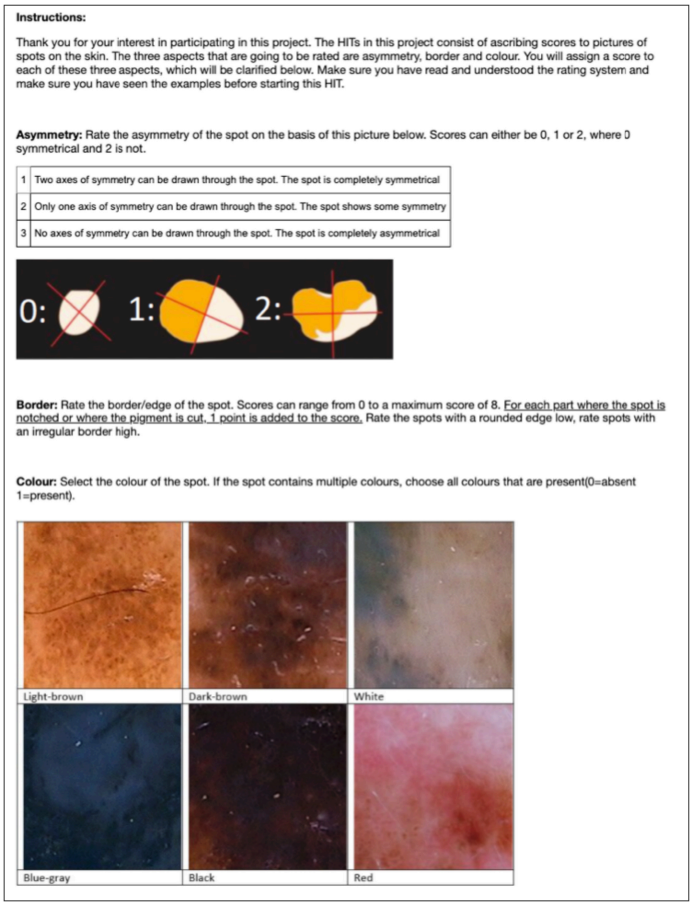


Figure 8: An instruction screen with the task description.

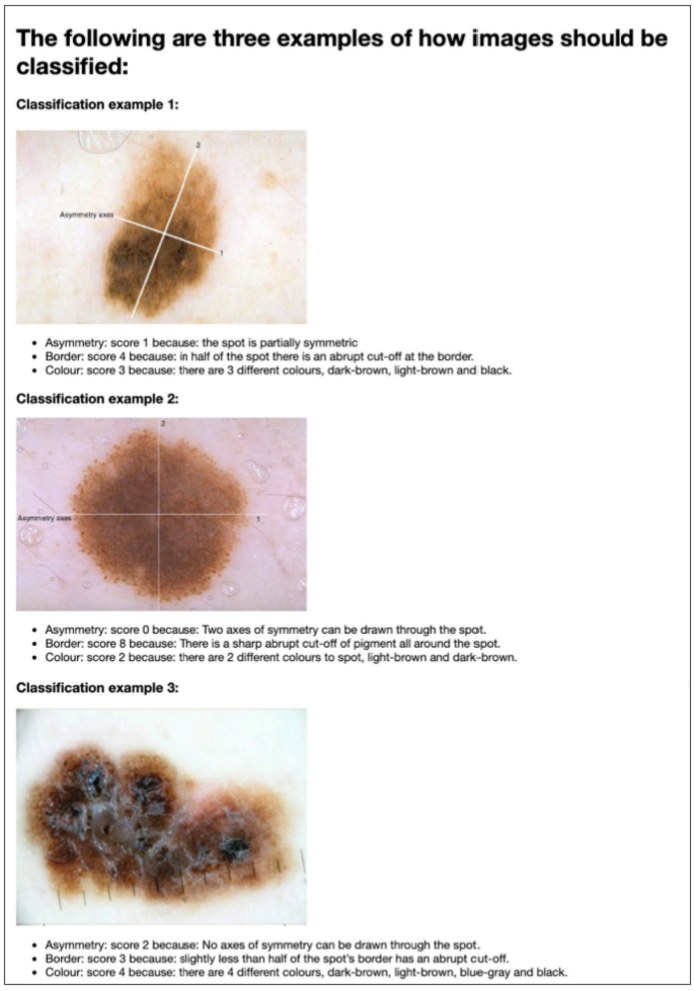


Figure 9: A screen with three examples showing how to score skin lesions.

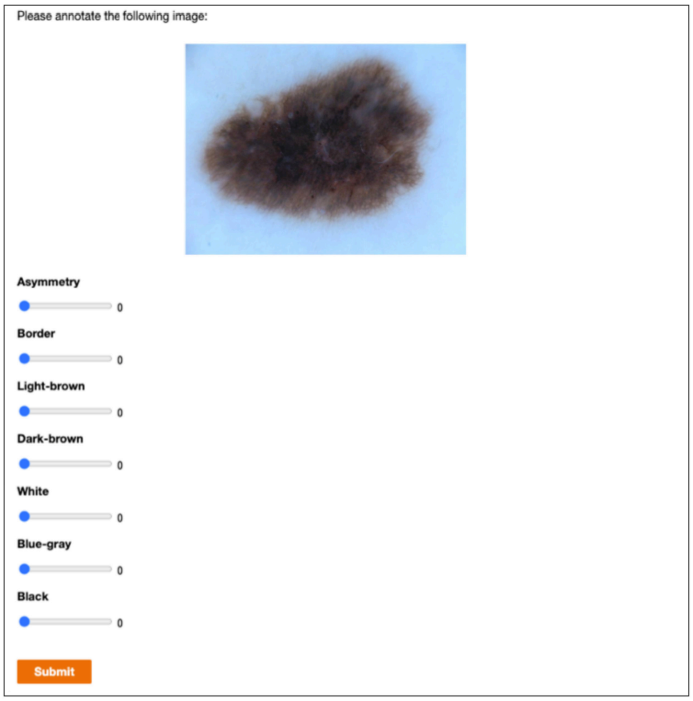


Figure 10: A screen with a skin lesion to be scored using the scoring sliders and pressing the submit button.