IT UNIVERSITY OF COPENHAGEN

Annotation Task Design in Medical Imaging

Thor Wessel Lindberg (17858, mawl@itu.dk)

26794 characters (incl. spaces)

***Abstract*–Crowdsourcing in medical image analysis widens the pool of available annotators, but also raises the question of how to best design annotation tasks. This project samples an existing annotation task design, and varies it across the scale of interaction attributes. Using an image dataset sourced from the ISIC 2017 Challenge [1], this project explores the impact of design choices on the quality of data. Quality is assessed through comparison to the ISIC ground truth. The findings suggest that similarity tasks produce the highest quality of results.**

Keywords–medical image analysis, skin lesion, melanoma classification, crowdsourcing, annotation

I INTRODUCTION

Machine Learning (ML) has established itself in the field of medical imaging, as an alternative to medical expert classification. However, medical imaging models rely heavily on training data, and scarcity of data has been a hindrance in expanding the application of ML classification.

Researchers have explored a series of algorithmic approaches to addressing medical imaging data scarcity. These include Multiple Instance Learning, Multi-Task Learning, and Transfer Learning.

Multiple Instance Learning (MIL) attempts to address the lack of labels for localized regions in medical imaging scans, which are a necessary requirement for performing supervised learning [2]. Related tasks in MTL are referred to as auxiliary tasks, which are leveraged to generalize and help understand the main task. Ruder notes that while humans typically utilize quantized training objectives, that is a discrete or binary set of labels, auxiliary tasks in MTL do not have to be quantized [2]. MTL aims to bias a machine learning model towards representations chosen by auxiliary tasks. This is called representation sharing, and commonly occurs through either hard or soft parameter sharing. MTL acts as a regularizer reducing the risk of overfitting, and thus it is beneficial to find commonalities between tasks in an environment [2].

Multi-Task Learning (MTL) is an example of leveraging multiple sources of learning. In this approach, representations between tasks related to the classification task would be shared, enabling a higher degree of generalization [2].

Transfer Learning aims to reach a singular consensus based on multiple sources of learning, or transferring learning from one training dataset to another dataset.

Crowdsourcing is a non-algorithmic approach to addressing data scarcity, focusing on providing a higher degree of annotation, without reliance on medical experts. Research has demonstrated the potential of crowdsourced annotations for providing accurate training data [3]. Crowdsourcing platforms such as MTurk [4] connect researchers with crowdsourced workers, and handle distribution of tasks amongst workers as well as payment. Researchers in medical image analysis design the annotation tasks, but have no insights into which workers provide them with annotations. These platforms provide rapid annotation of research datasets, through a large pool of workers.

As crowdsourced workers can not be expected to have the educational backgrounds or academic criteria necessary to become a medical expert in medical imaging, it is important that tasks are designed to not only annotate, but also to educate. A choice has to be made by researchers as to which knowledge should be included and excluded. Workers need enough information to delimit classes, which should be based on the same knowledge medical experts leverage in their judgments. In the following section, the medical knowledge used in diagnosing melanomas is presented.

Melanomas are classified with a binary set of labels, that is benign or malignant. A hallmark of malignant melanoma is irregularity relative to neighboring melanomas or melanomas in the same environment, as well as changes over time. Melanoma classification is based on a set of warning signs, the so-called ABCDE's (Asymmetry, Border, Color, Diameter/Dark, and Evolving) [5].

These warning signs are employed in melanoma diagnostics for similarity analysis, wherein suspected malignancy is determined based on the melanoma's appearance relative to other melanomas. Another approach is to 'half' the suspected malignant melanoma in half, to compare and contrast the two halves with respects to the ABCDE's of malignancy.

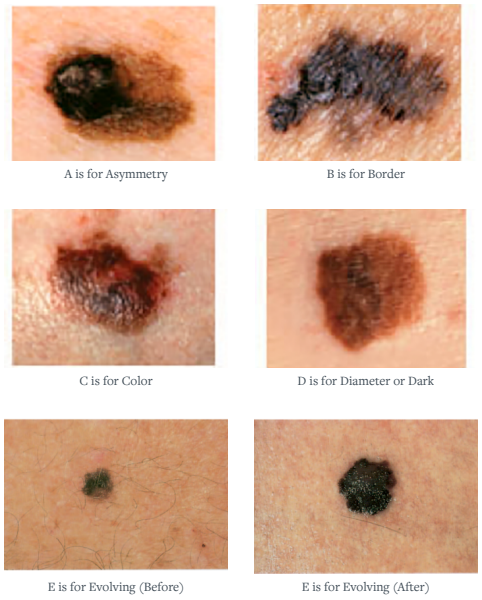

Figure 1: Warning signs of malignancy [5].

A strategy for classifying melanoma is looking for irregularities within a set of melanomas. This so-called Ugly Duckling strategy relies on the concept that moles typically appear similar in appearance, and thus a malignant melanoma will appear irregular relative to its neighboring melanomas [5].

This knowledge of crowdsourcing methodologies and melanoma classification forms the basis of this project, as well as my approach to designing annotation tasks. This project aims to develop and evaluate varied designs of annotation tasks in melanoma classification, for the purpose of determining *how design decisions are perceived by annotators and whether design impacts annotation quality.*

II RELATED WORK

A survey on crowdsourcing [6] found that several medical image analysis research papers lack description of annotation task design choices. It concludes that withholding information on the motivating factors and considerations behind designing the characteristics of annotation tasks, causes studies to be less accurately reproducible. This in turn makes crowdsourcing less approachable to resaerchers who do not have experience with conducting crowdsourced annotation. The characteristics of crowdsourcing tasks identified in the survey include *platform choice*, *number of annotators*, and *how the task is explained to annotators*. This survey acts as a literature review of studies that apply crowdsourcing in medical image analysis, and it highlights the need for increased guidance in conducting crowdsourcing, through better documentation of experiments.

Researchers in medical image analysis are typically not trained in design or development, and thus leverage existing frameworks. A popular option is Amazon's Mechanical Turk (MTurk) [4], which connects researchers with workers, while handling infrastructure and scalability. It also provides prebuilt annotation tasks, which means researchers only have to feed their training data to the platform. For these reasons, it is an obvious choice for many researchers, but it also lessens the burden of considering the aforementioned characters of crowdsourcing. Researchers do not have to describe motivating factors, as MTurk workers are financially incentivized to complete the given annotation task.

There are numerous examples of research applying crowdsourcing in medical image analysis. In the following section, examples of research approaches in crowdsourcing are presented.

Herrera et al. [7] took another approach to crowdsourcing, arguably one with less of a 'crowd'. They produced algorithmic classifications of their dataset, and then leveraged the Crowdflower platform to connect with a small group of eight medical imaging experts. These crowdsourced workers were tasked with verifying the algorithmic classifications. Each image in their dataset contained at least two crowdsourced annotations, and a third for images with opposing annotations.

These researchers found that the expert workers lacked some knowledge to complete their tasks, but that crowdsourcing was especially useful for annotating low-resolution imagery. They also conclude that while they chose a small pool of workers in the medical imaging domain, they could have taken the opposing approach by applying stricter quality control.

Maier-Hein et al. [8] leveraged MTurk in asessing the quality of crowdsourced annotations for medical endoscopic images. They supplied MTurk with their own user interface written in HTML5/JavaScript, but they do not indicate why this choice was made or whether any considerations were had about the design of this interface. They find a median annotation error of 2 px, which is twice that of medical experts, but through cluster analysis of multiple annotations per image, they reduce errors so that they are comparable to expert classifications.

In their discussion, they identify annotation tasks as consisting of multiple tasks that in combination result in a classification. Their point is that while some tasks may require expert knowledge or experience, the task they crowdsourced appears to have produced comparable classifications while leveraging non-experts.

Braindr [9] is an app for quality control of images from the Healthy Brain Network [10]. Rather than conducting crowdsourcing through a platform, its annotation tasks are custom and do not offer financial incentives. This meant the creators had to find other motivating factors, and they chose to gameify the process through a point-system and leaderboard.

Figure 2: Example of a discrete annotation task in crowdsourced medical imaging. [9]

In their evaluation, they chose to filter out image annotations that had not been repeated at least 5 times. They conclude based on the produced annotations that while some images fit well into the discrete classification of fail or pass, there are also images that do not. This raises questions about whether their task design is extensive enough, in that it evidently may not cover all scenarios during annotation. If an image is perceived as neither passing nor failing, then annotators are only given the choice to provide incorrect annotations, due to the lack of a "neither" or "unsure" option.

The quality difference between a set of interactions can be assessed by establishing their interaction profile. In Human-Computer Interaction (HCI) research exists an Interaction Vocabulary consisting of attributes that describe interfacing between humans and technology, as well as the relationship between the expectations of designers and the reality users experience [11].

Minimum	**Maximum**
slow	fast
stepwise	fluent
instant	delayed
uniform	diverging
constant	inconstant
mediated	direct
spatial separation	spatial proximity
approximate	precise
gentle	powerful
incidental	targeted
apparent	covered

Table 1: Interaction attributes

The authors conclude that interactions should be judged by whether the designer's intent matches the user's perception, meaning that no interaction attribute should be viewed as greater or more desirable than others. Instead, attributes should be viewed as appropriate relative to whether they produce a perception that matches the intent.

Based on their findings, it is hypothesised that the annotation task best perceived as it was intended must have the best fit of interaction attributes. In the context of annotation, the attributes *stepwise, instant, uniform, direct, spatial proximity, and precise* (table 1) are expected to reflect an ideal experience.

III EXPERIMENT SETUP

Following are details on the experiment setup used to test the hypothesis. The Braindr project [9] was sampled as a baseline for the design of my three annotation tasks. The baseline was varied across the scale of interaction attributes [11], by identifying its attributes and branching out from them into other or opposing attributes. The ENHANCE project [12] was used as a reference for selecting a subset of training data from the ISIC 2017 Challenge dataset [1], and to evaluate the results.

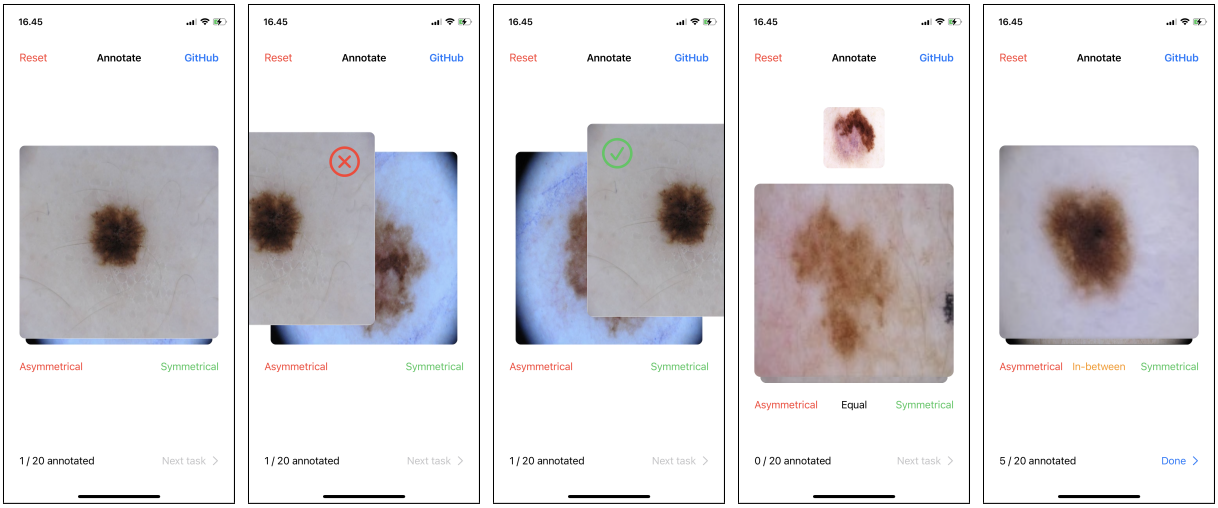

Figure 3: Screenshots of the annotation tasks.

a Image Dataset

The ISIC 2017 Challenge [1] encouraged competitors to develop image analysis tools for melanoma classification. It consists of around 2000 melanoma images, of which a training data subset was selected to match the subset in the ENHANCE project [12]. This is the only dataset used in this project, as the value of each task variant is determined by how closely the annotations produced match those produced in the ENHANCE project [12].

This subset includes 1250 images from the ISIC 2017 dataset. The ENHANCE project [13] produced 3 annotations per feature (asymmetry, border, color) for each of the 1250 images, resulting in 11250 annotations. Using the MTurk platform [4], annotation tasks were assigned to workers at random. MTurk takes a distributed approach, by assigning small tasks to a large pool of workers. This approach allows tasks to be performed concurrently, reducing the time researchers to have to wait for results, and preventing potential worker fatigue.

As this project aims to motivate annotators without a financial incentive, it was unrealistic to expect each annotator to classify the entire subset for each feature and each task variance. As such, a distributed approach similar to MTurk is necessary, and can be achieved by developing a system that assigns tasks to annotators based on the [in]completeness of the annotations dataset. As systems design is resource intensive and not the focus point of this project, the subset was reduced to the first 60 images, which were further split into three groups of 20 images, one group for each annotation task.

b Technical Implementation

The annotation tasks were implemented as a Swift application (see figure 3). It leverages the universal and Apple platform-neutral SwiftUI framework, which allows the tasks to adapt to their platform and context of use. This application was phrased as an educational component with a subsequent three-stage annotation task component. This is the same approach as was taken Braindr [9], but with the annotation divided into three tasks.

The tasks are designed to be distributed to a small pool of testers, reflected in the fact that a small (n=60, divided in 3) subset of images was selected. If a larger pool of crowdsourced workers were to be leveraged, a distributed systems approach would be necessary. Through this approach, annotation progress would be tracked server-side, and annotators would be assigned random subsets of images based on [in]completeness of the produced dataset.

c Study Design

Annotators were selected based on their educational background and knowledge of the subject, as I wanted to avoid testing with people who already had medical experience or an understanding of melanoma classification. They were sampled through non-saturation and non-probability convenience sampling, as defined by Sharp et al. [14]). Non-saturation sampling meant a limited amount of test persons were necessary, as I only needed 20 annotations per image. Non-probability sampling meant test persons were not selected based on probability, but rather selected directly by me. Convenience sampling meant test persons did not volunteer, but rather they were selected because of my accessibility to them.

d Annotation Tasks

Braindr [9] leverages a simple and rapid bi-directional task, wherein users swipe/click left or right on each image in a procedually generated stack representing their dataset. A left swipe/click annotates the given image as "Fail", while a right swipe/click annotates the given images as "Pass". This interaction was sampled for this project, because it conforms well to different contexts of use, and can be easily varied across the scale of interaction attributes.

Variances of the annotation task were determined by first categorizing the Braindr [9] interaction relative to the attribute scale, and then finding ideating based on opposing or non-present attributes. This process lead to the development of three distinct variants: fail or pass, similarity, and Likert scale.

1) Fail or pass. In this task, annotators classify an image as asymmetrical or symmetrical. It is intended to reproduce the experience of Braindr [9], acting as a baseline design. Based on the interaction vocabulary, this design is fast, stepwise, instant, direct, spatially proximate, precise, and targeted.

2) Similarity. In this task, annotators classify an image based on its similarity to a randomised reference image. Annotations can be asymmetrical, symmetrical, or equal to the reference image. Based on the interaction vocabulary, this design is slower and less uniform than the baseline, as a third classification is added.

3) Likert scale. In this task, annotators classify an image as asymmetrical or symmetrical, with the option to classify as an intermediary between the two This approach is similar to a Likert scale. Based on the interaction vocabulary, this design is slower, less uniform and more spatially separated than the baseline, as a third classification is added and there is no unique direction or position for the annotator to target.

A Likert scale is a psychometric technique that measures human attitude, as defined by Joshi et al. [15]. This scale can be either symmetric or asymmetric, meaning the classifications are either even or uneven in numbers. An asymmetric scale can contain an intermediary value, that is a value reflecting neither side of a discrete or continuous scale (see figure 4).

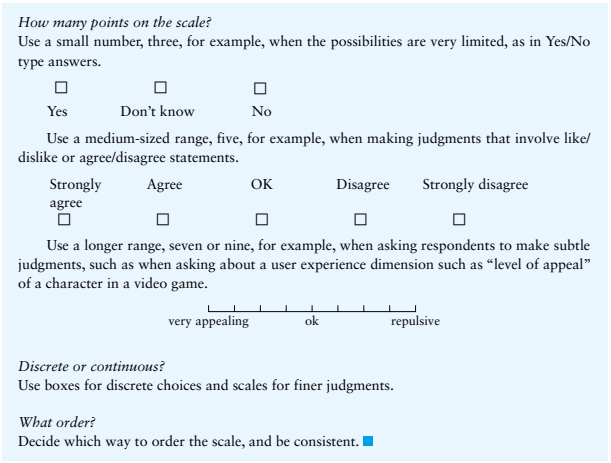

Figure 4: An asymmetric Likert scale [15]

The purpose of appropriating the asymmetric Likert scale for this project, was to give annotators agency and autonomy in deciding whether an annotation fit into neither of the discrete classifications. This is accomplished by providing an intermediary classification, which was notably absent from the Braindr project [9].

e Evaluation

Twenty annotators were recruited for participation in the annotation tasks, with each annotator producing a dataset for the 60 images from the ENHANCE project's [12] 1250 image subset of the ISIC 2017 challenge dataset [1]. In addition to producing annotations, these participants were asked to reflect on the tasks, and were encouraged to vocalise their thoughts or confusions during annotation.

The task variances are evaluated by comparing the resulting annotations to the results produced in the ENHANCE project [12]. The purpose of using a similar crowdsourcing project as reference, is to define and evaluate the data quality produced through the designed tasks.

This comparison consists of two datasets for each feature: *asymmetry*, *border*, and *color*. The first reference dataset is from the ENHANCE project [12], and the second dataset is produced through this experiment. Data quality is defined as indifference in annotations between the two datasets.

IV RESULTS

This experiment yielded a total of 1200 annotations of (0, 1, 2) for asymmetry, 20 annotations (one person participant) for each of the 60 images in the dataset. These results were adjusted for comparison with the ISIC expert classifications [1] annotations, by correcting the intermediary value to one of the binary (0, 1) values. This yields two variants of the annotations, one where the upper value (2) is corrected to the malignancy value (1), and one where the intermediary value (1) is corrected to the benign value (0) and the upper value (2) is corrected to the malignancy value (1).

subset	mturk	annotators	difference	task
1-20	0.28	0.70	0.42	1
21-40	0.48	0.65	0.17	2
41-60	0.48	0.63	0.15	3

Table 2: Accuracy scores of asymmetry annotations produced by the ENHANCE project (MTurk) [12] and the annotations produced through my experiment (figure 3. Corrected upper value (2) to malignancy value (1).

subset	mturk	annotators	difference	task
1-20	0.72	0.85	0.13	1
21-40	0.58	0.60	0.02	2
41-60	0.67	0.74	0.07	3

Table 3: Accuracy scores of asymmetry annotations produced by the ENHANCE project (MTurk) [12] and the annotations produced through my experiment (figure 3. Corrected intermediary value (1) to benign value (0) and upper value (2) to malignancy value (1).

As seen in table 2 and 3, the participating annotators produced consistently higher accuracy scores than the MTurk workers [12], when using expert classifications from the ISIC 2017 dataset [1] as reference.

It is important to stress that while this project aimed to only annotate asymmetry, the MTurk workers produced annotations for asymmetry, border, and color, which were added as weighted features to a baseline [12]. This means the accuracy scores are not representative of the MTurk workers' complete annotation work, as multiple features are available. It also means the accuracies for each subset of the ISIC 2017 dataset [1] should not be compared directly, as the accuracy scores only represent asymmetry annotations.

Data quality was defined in this experiment as indifference to the ENHANCE [12] annotations. These results thus indicate that the produced annotations are of higher quality when:

1. Intermediary classifications (1) are perceived as benign classifications (0). This is reflected in the smaller difference between MTurk annotations and this project's annotations in table 3, relative to table 2. This suggests that the crowdsourced annotators participating in this experiment annotated more accurately when their intermediary classifications are considered benign rather than malignant. This speaks to how annotators perceive the tasks, likely interpreting the intermediary classification as "not malignant, not quite benign" or "unsure".

2. Annotators are given an intermediary option between benign or malignant. This is reflected in the smaller difference between MTurk annotations and this project's annotations for task 2 and 3 in both table 2 and 3, relative to task 1. Task 2 in table 3 is especially close (0.02 difference), further suggesting that the best fit intermediary represents a benign or malignant classification, rather than an in-between classification as in task 3.

In addressing the hypothesis, these findings indicate that task 2 produces the highest quality of annotation data, out of the three designed tasks. This is reflected in the fact that task 2 and 3 share similar interaction profiles, while producing a higher quality of annotation data than task 1. It also indicates that annotation tasks perform better when their interaction profiles include the attributes: *slow, diverging and spatial proximity.* While task 3 performs slightly better according to table 2, the larger difference between task 2 and 3 according to table 3 suggests that the presence of *spatial separation* in task 3 lessens its accuracy.

These results corroborate the participants' reactions to the three designed annotation tasks. Task 1 lacks any intermediary classification, forcing annotators to classify images as benign or malignant irrespective of whether these classifications represent the annotator's perception. Task 2 addresses this issue by providing a reference image for similarity comparison and an intermediary classification that produces an annotation equal to the reference. Task 3 provides an in-between classification, which greatly improves the accuracy relative to the baseline, but may be perceived as confusing due to the uncertain nature of this classification. These in-between annotations might be better filtered out, and only provided to let the annotator continue annotating when uncertain.

V DISCUSSION

This project asks the cross-disciplinary question of "how should crowdsourced annotation tasks be designed?" rather than the status quo question of "how can annotation be crowdsourced?" In answering this question, I have attempted to synthesise knowledge, experiences and approaches from both medical image analysis and design research. As it turns out, this question encompasses more than I had initially anticipated, because facilitating annotation tasks requires infrastructure beyond the task.

In order to conduct annotation, a platform for collecting data has to be planned, developed, and distributed. Collecting annotation data is a complex and asynchronous process of identifying available data, then segmenting the data and distributing tasks to workers. Once a platform exists, researchers have to plan their approach, select or develop annotation tasks, then finally collect data through crowdsourced workers. In summary, annotation task design is only a small component of a large process, and thus it is understandable or perhaps even optimal for researchers to not document or report on their design decisions, if any are made.

I chose to design my baseline annotation task by sampling the Braindr application [9], because it already constitutes an original and custom task design, existing outside a framework/platform such as MTurk. This approach provided me full control of the task design, but also required a large resource investment into the underlying infrastructure.

I chose to develop the minimal amount of infrastructure necessary, opting not to leverage a distributed systems or networked approach, wherein tasks could be automatically assigned to participants. This would have let me test the tasks with a larger pool of participants, but requires too much of a time investment in a project focused on task design.

I will also stress the importance of educating crowdsourced workers when facilitating annotation. I can not deduce anything about this from the experiment I conducted nor the results I produced, as I chose to focus on the design of the tasks themselves. However, it became evident to me through my process that a common perception of terminology and annotation influences how well participants perform their tasks.

This could be explored through further work, by measuring the performance difference between crowdsourced workers presented to different narratives on annotation or different amounts of information. I speculate that while not enough information can cause misclassification by workers, it is also possible that too much information, in the small amount of time during facilitation of annotation tasks, leads to confusion and thus worse performance.

As noted by Herrera et al. [7], the amount of crowdsourced workers annotating influences the results of an experiment. I can not deduce from my results whether this is reflected in my experiment, but I speculate that an annotator with an incorrect perception of what constitutes malignancy is more likely to repeat misclassification of melanomas.

This could be determined by analysing trends in the annotation data, presenting to which degree each participant repeated the same type of misclassification relative to the ground truth by medical experts, in example: classifying malignant melanoma (1) as benign (0). It may be useful to build these statistics into the infrastructure, to identify misclassification during annotation tasks or immediately afterwards, allowing researchers to gather feedback on why annotators chose the wrong classification.

Finally, this project is a synthesis of medical image analysis and design research, aiming to assess the influence of design on data quality. It has become evident to me that a team of two or more people, with at least one representative from each discipline, may have been more successful at achieving the desired outcome of this project. Cross-disciplinary representation allows for the negotiation of terminology and methodology between two people with expertise in their respective discipline.

Cross-disciplinary negotiations would lessen the occurrence of roadblocks in the process, and ensure that the project and its experiment was phrased within the parameters of the medical image analysis discipline.

VI CONCLUSION

The annotation task design experiment showed that the highest data quality is achieved by a task phrased as a *slow, diverging and spatially proximate* similarity comparison with an intermediary classification equal to its reference.

Participant feedback in response to testing the task designs, showed that tasks should be phrased in a way that reflects the annotator's perception, empowering annotators to provide the most accurate and representative classifications.

References

[1] Isic 2017 challenge. https://challenge.isic-archive.com/data/#2017.

[2] Sebastian Ruder. An overview of multi-task learning in deep neural networks. *CoRR*, abs/1706.05098, 2017. http://arxiv.org/abs/1706.05098.

[3] Veronika Cheplygina. Crowddetective: Wisdom of the crowds for detecting abnormalities in medical scans. *Journal of Trial and Error*, 1(1), 12 2020. https://archive.jtrialerror.com/pub/crowddetective.

[4] Amazon mechanical turk. https://www.mturk.com/.

[5] Allan C. Halpern, Ashfaq A. Marghoob, and Ofer Reiter. Melanoma warning signs and images, Oct 2021. https://www.skincancer.org/skin-cancer-information/melanoma/melanoma-warning-signs-and-images/#panel1-2.

[6] Silas Ørting, Andrew Doyle, Arno van Hilten, Matthias Hirth, Oana Inel, Christopher R. Madan, Panagiotis Mavridis, Helen Spiers, and Veronika Cheplygina. A survey of crowdsourcing in medical image analysis, 2019.

[7] Alba García Seco de Herrera, Antonio Foncubierta, Dimitrios Markonis, Roger Schaer, and Henning Müller. Crowdsourcing for medical image classification. *Swiss Medical Informatics*, 10 2014.

[8] Lena Maier-Hein, Sven Mersmann, Daniel Kondermann, Christian Stock, Hannes Gotz Kenngott, Alexandro Sanchez, Martin Wagner, Anas Preukschas, Anna-Laura Wekerle, Stefanie Helfert, Sebastian Bodenstedt, and Stefanie Speidel. Crowdsourcing for reference correspondence generation in endoscopic images. In Polina Golland, Nobuhiko Hata, Christian Barillot, Joachim Hornegger, and Robert Howe, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014*, pages 349–356, Cham, 2014. Springer International Publishing.

[9] Braindr.us. https://braindr.us/#/.

[10] Lindsay M. Alexander, Jasmine Escalera, Lei Ai, Charissa Andreotti, Karina Febre, Alexander Mangone, Natan Vega Potler, Nicolas Langer, Alexis Alexander, Meagan Kovacs, Shannon Litke, Bridget O'Hagan, Batya Bronstein, Anastasia Bui, Marijayne Bushey, Victoria Castagna, Nicolas Camacho, Elisha Chan, Danielle Citera, Jon Clucas, Samantha Cohen, Megan Eaves, Brian Fradera, Natalie Grant-Villegas, Gabriella Green, Camille Gregory, Emily Hart, Shana Harris, Catherine Lord, Danielle Kahn, Katya Kabotyanski, Kayla Kleinman, Bonhwang Koo, Eliza Kramer, Amy Margolis, Kathleen R. Merikangas, Judith Milham, Giuseppe Minniti, Rebecca Neuhaus, Alexandra Nussbaum, Yael Osman, Lucas C. Parra, Ken R. Pugh, Amy Racanello, Anita Restrepo, Tian Saltzman, Batya Septimus, Russell Tobe, Rachel Waltz, Anna Williams, Anna Yeo, Francisco X. Castellanos, Arno Klein, Tomas Paus, Bennett L. Leventhal, Cameron R. Craddock, Harold S. Koplewicz, and Michael P. Milham. The healthy brain network biobank: An open resource for transdiagnostic research in pediatric mental health and learning disorders. *bioRxiv*, 2017. https://www.biorxiv.org/content/early/2017/06/13/149369.

[11] Eva Lenz, Sarah Diefenbach, and Marc Hassenzahl. Exploring relationships between interaction attributes and experience. 09 2013.

[12] Ralf Raumanns. Enhance, 2021. https://github.com/raumannsr/ENHANCE.

[13] Marc Hassenzahl. *Experience Design: Technology for All the Right Reasons*, volume 3. 01 2010.

[14] H. Sharp, J. Preece, and Y. Rogers. *Interaction Design: Beyond Human-Computer Interaction*. Wiley, 2019.

[15] Ankur Joshi, Saket Kale, Satish Chandel, and Dinesh Pal. Likert scale: Explored and explained. *British Journal of Applied Science Technology*, 7:396–403, 01 2015.

Appendix 1: Annotations collected in the experiment

id	e	e	e	i	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a
ISIC_0000000	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000001	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000002	2	0	2	1	1	0	0	1	1	1	0	1	1	1	0	0	1	1	0	0	1	1	0	1
ISIC_0000003	1	1	0	0	0	0	0	0	1	1	0	0	1	1	1	1	0	0	1	1	1	1	0	0
ISIC_0000004	0	1	1	1	1	1	0	1	1	1	0	1	1	1	1	0	1	1	1	1	0	1	1	1
ISIC_0000005	1	2	2	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0
ISIC_0000006	0	2	2	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0
ISIC_0000007	1	1	0	0	0	1	1	1	0	1	0	1	0	0	1	1	0	1	1	0	0	0	0	0
ISIC_0000008	1	1	1	0	0	0	0	1	0	1	0	1	0	0	1	0	0	0	1	0	0	0	0	0
ISIC_0000009	2	1	2	0	0	0	1	0	1	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0
ISIC_0000010	1	1	1	0	0	0	0	1	1	1	0	1	1	1	1	0	1	1	0	1	1	0	0	1
ISIC_0000011	0	0	1	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0
ISIC_0000012	1	1	2	1	0	1	1	0	1	0	0	0	0	0	1	0	1	1	1	0	0	0	0	0
ISIC_0000013	1	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
ISIC_0000014	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	1	0	0	0	0
ISIC_0000015	1	1	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
ISIC_0000016	2	1	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0
ISIC_0000017	1	2	2	0	0	0	0	0	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1
ISIC_0000018	2	1	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000019	2	1	1	0	0	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	0	1	1	1
ISIC_0000020	2	2	2	0	0	1	1	1	1	1	1	0	0	1	1	1	0	1	0	0	1	1	1	1
ISIC_0000021	1	0	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1
ISIC_0000022	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
ISIC_0000023	2	2	2	0	0	0	1	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0
ISIC_0000024	1	2	2	0	1	0	1	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1
ISIC_0000025	2	2	2	1	1	1	0	1	1	0	1	1	1	1	0	1	1	1	0	1	1	1	1	1
ISIC_0000026	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0
ISIC_0000027	1	1	1	0	0	0	1	0	1	1	1	1	1	1	1	1	1	0	1	0	0	0	0	0
ISIC_0000028	0	1	0	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	0	1	0	0	0
ISIC_0000029	2	2	0	1	1	1	1	0	1	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1
ISIC_0000030	2	2	2	1	1	1	1	0	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1
ISIC_0000031	1	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0
ISIC_0000032	1	1	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0
ISIC_0000033	2	1	0	1	0	1	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000034	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	0	1	1	1	1
ISIC_0000035	1	2	2	0	1	1	0	1	0	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1
ISIC_0000036	0	1	2	0	1	1	0	1	1	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1
ISIC_0000037	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0
ISIC_0000038	2	2	2	1	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	0	1	1	1	1
ISIC_0000039	2	2	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0
ISIC_0000040	1	0	1	0	0	1	1	1	1	1	1	0	0	0	0	1	0	0	1	0	1	1	0	0
ISIC_0000041	1	1	1	1	0	0	1	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0
ISIC_0000042	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0
ISIC_0000043	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0
ISIC_0000044	1	2	1	1	2	2	2	1	2	1	1	0	2	2	2	0	2	1	1	2	2	2	2	2
ISIC_0000045	0	2	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0
ISIC_0000046	0	1	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0
ISIC_0000047	1	2	0	1	1	2	1	2	1	2	1	2	2	0	2	2	1	1	0	2	2	2	2	2
ISIC_0000048	2	2	1	0	2	1	2	2	1	2	2	0	2	2	2	1	0	2	1	1	1	1	0	1
ISIC_0000049	2	2	1	0	0	0	0	1	1	1	2	1	2	2	1	2	1	1	0	0	0	2	2	2
ISIC_0000050	0	0	2	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	2	2	0	0	0	0
ISIC_0000051	1	1	1	0	2	1	1	0	0	0	0	0	2	0	2	1	1	1	1	2	0	0	2	1
ISIC_0000052	0	1	1	1	0	0	1	0	1	1	1	1	1	1	0	0	2	2	0	1	1	0	0	1
ISIC_0000053	2	1	1	0	2	0	1	1	0	1	1	1	2	0	1	0	0	0	0	0	0	0	2	0
ISIC_0000054	1	2	2	1	2	2	2	0	0	2	2	1	2	2	2	1	1	1	1	1	2	0	0	2
ISIC_0000055	1	1	0	0	2	0	0	0	1	0	2	1	0	0	0	0	2	0	1	0	1	1	0	1
ISIC_0000056	1	1	0	0	1	0	0	0	2	1	1	0	0	2	1	2	2	0	1	0	0	0	2	1
ISIC_0000057	1	0	2	0	0	1	1	0	0	2	0	0	0	0	0	0	0	0	0	2	0	0	0	0
ISIC_0000058	0	0	2	0	1	2	0	0	2	1	1	0	0	0	0	0	0	1	1	1	1	0	0	0
ISIC_0000059	1	1	1	0	0	0	1	0	0	0	0	1	0	2	0	2	1	0	0	0	0	1	0	0

Table 4: Abbreviations used are *e = ENHANCE, i = ISIC, a = Annotator.*

Appendix 2: Annotations corrected from 2 to 1

id	e	e	e	i	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a
ISIC_0000000	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000001	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000002	1	0	1	1	1	0	0	1	1	1	0	1	1	1	0	0	1	1	0	0	1	1	0	1
ISIC_0000003	1	1	0	0	0	0	0	0	1	1	0	0	1	1	1	1	0	0	1	1	1	1	0	0
ISIC_0000004	0	1	1	1	1	1	0	1	1	1	0	1	1	1	1	0	1	1	1	1	0	1	1	1
ISIC_0000005	1	1	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	1	0	1	0	0	0	0
ISIC_0000006	0	1	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0
ISIC_0000007	1	1	0	0	0	1	1	1	0	1	0	1	0	0	1	1	0	1	1	0	0	0	0	0
ISIC_0000008	1	1	1	0	0	0	0	1	0	1	0	1	0	0	1	0	0	0	1	0	0	0	0	0
ISIC_0000009	1	1	1	0	0	0	1	0	1	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0
ISIC_0000010	1	1	1	0	0	0	0	1	1	1	0	1	1	1	1	0	1	1	0	1	1	0	0	1
ISIC_0000011	0	0	1	0	0	0	0	0	1	0	0	0	0	1	1	0	0	0	0	1	1	0	0	0
ISIC_0000012	1	1	1	1	0	1	1	0	1	0	0	0	0	0	1	0	1	1	1	0	0	0	0	0
ISIC_0000013	1	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0
ISIC_0000014	1	1	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	1	0	0	0	0
ISIC_0000015	1	1	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
ISIC_0000016	1	1	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0
ISIC_0000017	1	1	1	0	0	0	0	0	1	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1
ISIC_0000018	1	1	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000019	1	1	1	0	0	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	0	1	1	1
ISIC_0000020	1	1	1	0	0	1	1	1	1	1	1	0	0	1	1	1	0	1	0	0	1	1	1	1
ISIC_0000021	1	0	1	1	1	1	0	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1
ISIC_0000022	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0
ISIC_0000023	1	1	1	0	0	0	1	0	0	0	0	0	1	0	1	0	1	0	0	0	0	0	0	0
ISIC_0000024	1	1	1	0	1	0	1	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1
ISIC_0000025	1	1	1	1	1	1	0	1	1	0	1	1	1	1	0	1	1	1	0	1	1	1	1	1
ISIC_0000026	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	1	0	0	0	0
ISIC_0000027	1	1	1	0	0	0	1	0	1	1	1	1	1	1	1	1	1	0	1	0	0	0	0	0
ISIC_0000028	0	1	0	1	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	0	1	0	0	0
ISIC_0000029	1	1	0	1	1	1	1	0	1	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1
ISIC_0000030	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1
ISIC_0000031	1	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0
ISIC_0000032	1	1	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	1	0	0	0	0	0
ISIC_0000033	1	1	0	1	0	1	0	1	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000034	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	0	1	1	1	1
ISIC_0000035	1	1	1	0	1	1	0	1	0	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1
ISIC_0000036	0	1	1	0	1	1	0	1	1	1	0	1	1	1	1	0	1	1	1	1	1	1	1	1
ISIC_0000037	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0
ISIC_0000038	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	0	1	1	1	1
ISIC_0000039	1	1	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0
ISIC_0000040	1	0	1	0	0	1	1	1	1	1	1	0	0	0	0	1	0	0	1	0	1	1	0	0
ISIC_0000041	1	1	1	1	0	0	1	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0
ISIC_0000042	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0
ISIC_0000043	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0
ISIC_0000044	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	1	1	1
ISIC_0000045	0	1	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0
ISIC_0000046	0	1	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0
ISIC_0000047	1	1	0	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1
ISIC_0000048	1	1	1	0	1	1	1	1	1	1	1	0	1	1	1	1	0	1	1	1	1	1	0	1
ISIC_0000049	1	1	1	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1
ISIC_0000050	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	1	1	0	0	0	0
ISIC_0000051	1	1	1	0	1	1	1	0	0	0	0	0	1	0	1	1	1	1	1	1	0	0	1	1
ISIC_0000052	0	1	1	1	0	0	1	0	1	1	1	1	1	1	0	0	1	1	0	1	1	0	0	1
ISIC_0000053	1	1	1	0	1	0	1	1	0	1	1	1	1	0	1	0	0	0	0	0	0	0	1	0
ISIC_0000054	1	1	1	1	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	1	0	0	1
ISIC_0000055	1	1	0	0	1	0	0	0	1	0	1	1	0	0	0	0	1	0	1	0	1	1	0	1
ISIC_0000056	1	1	0	0	1	0	0	0	1	1	1	0	0	1	1	1	1	0	1	0	0	0	1	1
ISIC_0000057	1	0	1	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
ISIC_0000058	0	0	1	0	1	1	0	0	1	1	1	0	0	0	0	0	0	1	1	1	1	0	0	0
ISIC_0000059	1	1	1	0	0	0	1	0	0	0	0	1	0	1	0	1	1	0	0	0	0	1	0	0

Table 5: Abbreviations used are *e = ENHANCE, i = ISIC, a = Annotator.*

Appendix 3: Annotations corrected from 1 to 0 and 2 to 1

id	e	e	e	i	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a	a
ISIC_0000000	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000001	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000002	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000003	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000004	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000005	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000006	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000007	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000008	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000009	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000010	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000011	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000012	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000013	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000014	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000015	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000016	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000017	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000018	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000019	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000020	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000021	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000022	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000023	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000024	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000025	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000026	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000027	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000028	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000029	1	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000030	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000031	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000032	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000033	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000034	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000035	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000036	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000037	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000038	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000039	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000040	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000041	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000042	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000043	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000044	0	1	0	1	1	1	1	0	1	0	0	0	1	1	1	0	1	0	0	1	1	1	1	1
ISIC_0000045	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000046	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000047	0	1	0	1	0	1	0	1	0	1	0	1	1	0	1	1	0	0	0	1	1	1	1	1
ISIC_0000048	1	1	0	0	1	0	1	1	0	1	1	0	1	1	1	0	0	1	0	0	0	0	0	0
ISIC_0000049	1	1	0	0	0	0	0	0	0	0	1	0	1	1	0	1	0	0	0	0	0	1	1	1
ISIC_0000050	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0
ISIC_0000051	0	0	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	1	0	0	1	0
ISIC_0000052	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0
ISIC_0000053	1	0	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0
ISIC_0000054	0	1	1	1	1	1	1	0	0	1	1	0	1	1	1	0	0	0	0	0	1	0	0	1
ISIC_0000055	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0
ISIC_0000056	0	0	0	0	0	0	0	0	1	0	0	0	0	1	0	1	1	0	0	0	0	0	1	0
ISIC_0000057	0	0	1	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0
ISIC_0000058	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
ISIC_0000059	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0

Table 6: Abbreviations used are *e = ENHANCE, i = ISIC, a = Annotator.*

Appendix 4: Accuracy scores of annotations from table 5 and 6

id	enhance (2 to 1)	annotator (2 to 1)	enhance (1 to 0 and 2 to 1)	annotator (1 to 0 and 2 to 1)
ISIC_0000000	1.00	1.00	1.00	1.00
ISIC_0000001	1.00	1.00	1.00	1.00
ISIC_0000002	0.67	0.00	0.67	0.00
ISIC_0000003	1.00	1.00	1.00	1.00
ISIC_0000004	0.00	0.00	0.00	0.00
ISIC_0000005	0.33	1.00	0.33	1.00
ISIC_0000006	0.33	1.00	0.33	1.00
ISIC_0000007	1.00	1.00	1.00	1.00
ISIC_0000008	1.00	1.00	1.00	1.00
ISIC_0000009	0.33	1.00	0.33	1.00
ISIC_0000010	1.00	1.00	1.00	1.00
ISIC_0000011	1.00	1.00	1.00	1.00
ISIC_0000012	0.33	0.00	0.33	0.00
ISIC_0000013	1.00	1.00	1.00	1.00
ISIC_0000014	1.00	1.00	1.00	1.00
ISIC_0000015	1.00	1.00	1.00	1.00
ISIC_0000016	0.67	1.00	0.67	1.00
ISIC_0000017	0.33	1.00	0.33	1.00
ISIC_0000018	0.67	1.00	0.67	1.00
ISIC_0000019	0.67	1.00	0.67	1.00
ISIC_0000020	0.00	1.00	0.00	1.00
ISIC_0000021	0.00	0.00	0.00	0.00
ISIC_0000022	1.00	1.00	1.00	1.00
ISIC_0000023	0.00	1.00	0.00	1.00
ISIC_0000024	0.33	1.00	0.33	1.00
ISIC_0000025	1.00	0.00	1.00	0.00
ISIC_0000026	1.00	1.00	1.00	1.00
ISIC_0000027	1.00	1.00	1.00	1.00
ISIC_0000028	0.00	0.00	0.00	0.00
ISIC_0000029	0.67	0.00	0.67	0.00
ISIC_0000030	1.00	0.00	1.00	0.00
ISIC_0000031	1.00	1.00	1.00	1.00
ISIC_0000032	1.00	1.00	1.00	1.00
ISIC_0000033	0.33	0.00	0.33	0.00
ISIC_0000034	0.00	0.00	0.00	0.00
ISIC_0000035	0.33	1.00	0.33	1.00
ISIC_0000036	0.67	1.00	0.67	1.00
ISIC_0000037	1.00	1.00	1.00	1.00
ISIC_0000038	1.00	0.00	1.00	0.00
ISIC_0000039	0.33	1.00	0.33	1.00
ISIC_0000040	1.00	1.00	1.00	1.00
ISIC_0000041	0.00	0.00	0.00	0.00
ISIC_0000042	1.00	1.00	1.00	1.00
ISIC_0000043	1.00	1.00	1.00	1.00
ISIC_0000044	0.33	0.65	0.33	0.65
ISIC_0000045	0.67	1.00	0.67	1.00
ISIC_0000046	1.00	1.00	1.00	1.00
ISIC_0000047	0.33	0.60	0.33	0.60
ISIC_0000048	0.33	0.55	0.33	0.55
ISIC_0000049	0.33	0.65	0.33	0.65
ISIC_0000050	0.67	0.90	0.67	0.90
ISIC_0000051	1.00	0.75	1.00	0.75
ISIC_0000052	0.00	0.10	0.00	0.10
ISIC_0000053	0.67	0.85	0.67	0.85
ISIC_0000054	0.67	0.50	0.67	0.50
ISIC_0000055	1.00	0.85	1.00	0.85
ISIC_0000056	1.00	0.75	1.00	0.75
ISIC_0000057	0.67	0.90	0.67	0.90
ISIC_0000058	0.67	0.90	0.67	0.90
ISIC_0000059	1.00	0.90	1.00	0.90

Table 7: Accuracy scores relative to the ground truth in the ISIC 2017 dataset.