Beyond Accuracy on Medical Imaging

Ashka Shah*
ashkashah@cs.toronto.edu
MScAC, University of Toronto
Toronto, Ontario, Canada

Hongcheng Wei*
hongchengw@cs.toronto.edu
MScAC, University of Toronto
Toronto, Ontario, Canada

Abstract

Machine Vision Components (MVCs) using Machine Learning (ML) have shown promising results in the field of medical imaging for tasks such as diagnosis, segmentation, and classification [1]. For successful integration into the healthcare domain, it is crucial to assess the reliability and safety of the MVCs being deployed [6]. However, the industry lacks precise definitions of what it means for an ML-based MVC to be "correct." In 2022, Hu et al. introduced a promising definition of a reliability requirement for generic MVCs, using human performance as a baseline. The definition is as follows: "if the changes in an image do not affect a human's decision, neither should they affect the MVC's" [5].

In this paper, we explore how this definition of reliability can be applied to the medical image classification space. We develop a pipeline to assess the reliability of an ML-based MVC focused on mole classification, used in the context of an early detection smartphone app. Specifically, we provide:

(1) Three realistic image transformations that reflect typical consumer usage scenarios
(2) A cost-effective methodology to phrase the question in a way that allows gathering data for establishing the "prediction-preservation" [5] baseline for mole classification in the absence of trained medical professionals.
(3) Human performance data on a range of transformations that can impact their perception of mole images.
(4) An evaluation of the MVC's performance in mole classification using this adapted reliability framework.

Finally, we discuss the limitations and shortcomings of our approach in applying the above definition of reliability to the medical imaging domain.

1 Introduction

Machine Vision Components (MVCs) powered by Machine Learning (ML) have shown remarkable performance in image analysis and processing. They are becoming widely deployed and integrated into medical specialties like radiology and pathology for diagnosis, classification and segmentation tasks. Given the widespread adoption of ML-based MVCs and the safety-criticalness of the medical imaging domain, assessing the quality and safety of these becomes crucial to the trustworthiness of these MVCs. Establishing requirements for verifying MVCs is harder compared to typical software tools due to their use of Machine Learning (ML). One of the major challenges in the deployment of MVCs is assessing their technical robustness, which is defined as the need for a standardized quality control process for reliability measurement [4] . Reliability can be defined as the consistency and correctness of the MVC's prediction in real-world scenarios. Current approaches focus on using held-out datasets to evaluate the performance of the MVCs using a single aggregated statistic. Limitations of these approaches include (1) Overestimation as the held-out dataset comes from the same distribution as the training set (2) A single aggregated statistic makes it hard to understand where the model is failing (3) Held-out datasets are quite small and not comprehensive [8]. Many studies have explored small-neighborhood (pixel based) transformations to original images that are usually imperceptible to humans. Yet they don't entirely represent the real-world situations in the healthcare domain [5]. The application of pixel based adversarial attacks may not be obvious in the medical imaging domain.

In 2022, inspired by human perception, Hu et al. introduced a promising definition of reliability requirements that could serve as a baseline for assessing generic MVCs. Since the goal of MVCs is to at least outperform humans, it makes sense to compare their performance to that of humans. They consider a range of transformations—such as brightness, contrast, defocus, and others—that do not affect human performance in recognizing car images. If these transformations do not affect human performance, they should not affect MVC performance either. They then define two classes of reliability requirements: (1) correctness preservation (t_c)— which requires that the prediction remains correct even after the transformation, and (2) prediction preservation (t_p)— which requires that the prediction on the original image is the same as the prediction on the transformed image. To quantify the visual change introduced by these transformations, they define a metric Δv (Figure 1) that utilizes: (1) VIF (Visual Information Fidelity) — a metric that evaluates perceived image quality and is empirically closest to human opinions on image quality, returning values between 0 (completely degraded) and 1 (perfect quality), and (2) VSNR (Visual Signal to Noise Ratio) — a metric that assesses the visibility of changes in the image, returning ∞ when the change is invisible to humans [5]. In this paper, we explore how this definition of reliability can be applied to the medical image classification space. We specifically

Definition 1: Visual change Δ_v

Let an image x, an applicable transformation T_X with a parameter domain C and a parameter $c \in C$, s.t. $x' = T_X(x, c)$ be given. $\Delta_v(x, x')$ is a function defined as follows:

$$\begin{cases} 0 & \text{If } \mathrm{VSNR}(x, x') = \infty \\ & \text{or } \mathrm{VIF}(x, x') > 1 \\ 1 - \mathrm{VIF}(x, x') & \text{Otherwise} \end{cases}$$

Figure 1: The definition of Δv as provided by Hu et al [5] in the paper - If a Human Can See It, So Should Your System: Reliability Requirements for Machine Vision Components

*Both authors contributed equally to this research.

focus on mole classification, a task critical for the early detection of skin cancer. We built 3 Mole Classification Models (MCM) by finetuning a pre-trained ResNet50 (on ImageNet) on the 2017 International Skin Imaging Collaboration (ISIC) dataset (2000 training images, 600 test images and 150 validation images) that classifies the image as either benign or malignant. Each model was trained for 16, 27 and 39 epochs respectively on a batch size of 32 using the Stochastic Gradient Descent optimization process (α (learning rate) = 0.001). Each model had a different validation and test accuracy (refer to Table 2). We evaluate the MVCs on "prediction-preservation (t_p)" as defined by Hu et al. We assess the MVC in the context of an early cancer detection smartphone app. Thus, we consider three pixel-level realistic image transformations that reflect typical consumer usage scenarios - (1) Defocus blur (2) Brightness (3) Dimness. Our choice of transformations were constrained by the definition of Δv which utilizes VIF. Transformations like rotation, crop, zooming and flipping were not considered, since VIF does not work on these. Given the time and budget constraint of the study, we develop a cost-effective method for posing questions that can gather human performance data without requiring trained medical professionals, establishing the human boundary for prediction-preservation of mole classification tasks. This report contributes to understanding if it is feasible to evaluate an MVC specifically in the context of mole classification using the apporach outline by Hu et al. The methodology for collecting human performance data is detailed in section 2 whereas section 3 outlines the results of the human performance evaluation, along with the assessment of the prediction-preservation in the MVC. Lastly, section 4 described the strengths and limitations of this study.

2 Research Methodology

In this section we talk about the research methodology used in the project. Section 2.1 details how the survey with human participants is designed, generated, and collected. Section 2.2 describes the analysis method.

2.1 Survey with Human Participants

Recall our research problem: **How can we apply the reliability requirement defined by Hue et al. [5] in a medical imaging task?** While the research goal is the same, the problem domain and research budget are key differences between the two studies. Applying the original study's method to our study poses key challenges.

(1) The original study focuses on common object classification task (i.e. cars). Our study is centered on mole classification. An untrained human would not be expected to differentiate between malignant and benign types. This means the human participant of the survey would ideally need to be a trained professional on the task. However, due to the constrain of our study, it's not feasible to hire trained medical professional for the survey.
(2) Given the constraints of our study, verifying correctness-preservation is not possible, since it would require the human participant to give a prediction (malignant or benign). Similarly, the original study's approach of verifying prediction-preservation also wouldn't work in our case.

Ultimately, we decided to limit our scope to only verifying the **prediction-preservation** requirement with **ordinary human participants** untrained in the mole classification task.

2.1.1 Survey Design. With untrained human, the survey cannot request the participants to assign a class label (malignant or benign) to each mole image. Therefore, we have designed the survey to ask a different question.

Each survey consists of 30 questions that each contains two images: one original and one transformed. Each question asks participants to determine whether the two images shown to them contain the same set of features or not. Figure 3 shows an example of one survey question.

The rationale behind this is that the decision regarding the "same features" is expected to be positively correlated with the decision about the "same prediction" (aka prediction preservation). In other words, if a participant identifies the same set of features in both images, they are likely to assign the same class label to both images as well. This approach allows us to include untrained participants in our survey to predict the human boundary for prediction preservation (t_p) without them assigning class labels to any image.

However, there are two major concerns with this approach. First, the concept of "feature" is inherently abstract, and individual interpretations of what constitutes a feature may differ considerably, particularly when facing an unfamiliar task. In this study, we have defined a standardized set of features that we asked the participants to look for when comparing the two images, according to the ABCDEs[1] rule of self-examination a melanoma mole. In particular, we asked the participants to consider the mole's *shape, texture, size, edges, and surroundings* as features and disregard other unimportant details in the image. Deriving the definition of features from a well-established self-examination guide ensures that non-professionals can easily identify them and that these features are important in mole classification.

The second concern is with the human biases. The survey design in original study follows the forced-choice image categorization task [10], where humans are tasked to categorize an image (car or not car) within 200 ms. This is to alleviate the effect of subjective biases and ensure fairness against machine classifiers [3]. We applied the same principle when designing our study's survey, with some adjustments. First, the time limit is adjusted to 5 seconds for each question because our participants are unfamiliar with mole images. Second, we ask the participant to follow their instincts and not modify the answers after a decision has been made. Unfortunately, due to the limitation of the survey platform (Google Form), we have no way to enforce these rules. Therefore, this is left for the participants to self-enforce.

Another human bias is referred to as the "boiling frog" effect, a metaphor describing that exposure to continuous small changes can lead to biased results. In our survey, participants might gradually adapt to subtle differences in the transformed images, making it harder for them to accurately recognize or report changes as the survey progresses, potentially skewing the results. To reduce the effect of this bias, we deliberately added 5 error questions in our surveys, where sometimes the transformed image is the same as

[1]American Academy of Dermatology Association "What to look for: ABCDEs of melanoma" https://www.aad.org/public/diseases/skin-cancer/find/at-risk/abcdes [7]

the original image ($\Delta v = 0$), and sometimes the transformed image is some other completely different image ($\Delta v = 1$). These error questions are randomly scattered in different positions of the survey and serve the purpose of resetting the potential cognitive adaptation to subtle changes.

2.1.2 Survey Generation. We generated the survey using a Python script (see Table 4). The generation process has three steps.

Step 1 begins by iterating over the test dataset and applying all three transformations to each image. We calculated the corresponding Δv value[2] for each transformation applied. If $\Delta v \leq 0$, we discard the transformation and attempt it again to ensure all resulting transformations are valid. The output of step 1 is a set of transformed images with $0 < \Delta v \leq 1$.

In Step 2, we sample the image pairs to be used in the surveys. The transformed images from Step 1 are split into two groups: no_error entries and error entries. We begin by creating the error entries. Recall that there are two types of error. For error questions that display the same original image as the transformation, we changed the transform image pointer to point to the original image, setting $\Delta v = 0$ to indicate no changes. For error questions that display completely different images as the transformation, the transform image pointer is changed to point to the next image in the list and set $\Delta v = 1$ to indicate the images are completely different. The proportion of the two error types is 34% same-image errors and 66% different-image errors.

Each survey consists of 30 questions, which includes 5 error questions. The remaining 25 questions are selected from the no_error entries. We implemented the below criteria during the sampling process to ensure fairness and minimize biases.

- Criterion 1: Each survey contains an equal number of transformations for each type.
- Criterion 2: Each survey includes transformations with Δv values scattered evenly across all ranges.
- Criterion 3: Each survey avoids displaying the same images more than once, even if they have different transformations.

These criteria help to ensure that the surveys do not favor a particular transformation type or Δv range and minimize the potential effect of subjective biases. We achieved this by grouping the entries by transformation type and Δv bin index. We then sampled evenly without replacement from all groups while keeping track of the names of the images of the entries already selected, discarding any selection if the same image was previously selected. Figure 4 illustrates the result of this sampling process.

To create the surveys, we shuffle a combination of 25 sampled entries from the non_error group and 5 from the error group to construct 30 questions for one survey. This process was repeated 40 times to generate all the surveys. Before starting one survey, we removed the sampled entries from the previous surveys from the dataset to ensure there are no duplicates. The output of Step 2 was a list of 40 surveys, each containing 30 questions.

Step 3 is to convert the surveys into PDFs. We utilize an HTML template and HTML-to-PDF converter to achieve this step. The output of Step 3 is a list of 40 PDF files, each corresponding to a survey. The PDF file contains 30 pages, each page contains one entry similar to Figure 3.

Finally, we uploaded these PDFs to Google Drive and utilized Google App Script to generate Google Forms for each survey PDF. Each Google Form includes the link to its corresponding survey PDF in the header section, along with instructions and the survey rules. Following the header, there are Yes/No selectors for each of the 30 questions in the survey. At the end of the form, there is a comment section where participants can provide additional feedback. The form for survey 0 [3] is listed here as an example.

2.1.3 Survey Collection. We targeted UofT students from the Fall 2024 CSC2125 course and 2024 MScAC program as the demographic for our study. We published the survey sign-up link on November 11, 2024, to collect email addresses from participants who expressed willingness to join. Ultimately, we distributed a total of 40 surveys and received 38 responses. Each survey consisted of 30 questions, of which 25 questions were valid. In total, we collected $38 \times 30 = 1140$ data points and $38 \times 25 = 950$ valid data points. The responses on the error questions are not considered valid data points and were not included in the following analysis. To compile the survey results, we used a Google App Script that automatically populated a CSV file with the collected data.

2.2 Analysis Method

To estimate the human tolerated range of transformation for prediction preservation (t_p) and to evaluate the Mole Classification Models we adapt the method defined by Hue et al. [5]. We define s as the **percent prediction preservation (ppp)**. Given the setup of our survey, the ppp would be the percentage of "Yes" responses. Now, we need to determine a baseline ppp, s_0. So, we consider it for very small transformations $\Delta v \leq \epsilon$. We sort all the transformed images by Δv and then assign ϵ to be the lower 5th quantile of this ordering. Thus, we define s_0 as ppp where $\Delta v \leq \epsilon$. Next, we divide all the Δv into bins of size 0.01. We generate a univariate spline on this data to remove outliers and run a binomial test on s_k, the ppp for the k^{th} bin. The Δv (lower bound of the k^{th} bin) at which $s_k < s_0$, with a p-value of 0.05 is considered the t_p for that transformation. We follow this procedure for defocus blur, brightness and dimness.

We evaluate the models using the human decision boundary t_p of each transformation. A bootstrap approach is applied to the 600 test images, where batches of 50 images are sampled 100 times with replacement. Every image in the batch is transformed such that $\Delta v \leq t_p$. The s_0 and s_{tp}, the ppp when $\Delta v \leq \epsilon$ and $\Delta v \leq t_p$ respectively are calculated. The ϵ is calculated as the 50th quantile (median). For each model, we end up with 100 s_0 and $s_t p$ values. We obtain the mean ($\hat{s_0}, \hat{s_{tp}}$) and standard ($\sigma_{s_0}, \sigma_{s_{tp}}$) deviations of the population. Using the mean s_0 and s_{tp} values, we calculate the **reliability distance (Δs)**, defined as $\Delta s = \hat{s_0} - \hat{s_{tp}}$. The standard deviation of the reliability distance is defined as $\sigma_{\Delta s} = \sqrt{\sigma_{s_0}^2 + \sigma_{s_{tp}}^2}$. A $\Delta s < 0$ indicates the model has achieved the target prediction preservation for transformations with $\Delta v \leq t_p$. A positive Δs signifies that the model failed the prediction preservation test. To

[2] We used the original MATLAB implementation of VIF and VSNR calculation, found at https://github.com/sattarab/image-quality-tools [9]

[3] https://docs.google.com/forms/d/1ZcETo3airzos1lnWtuByGmxHt7vSHfPNE_3yG16OCtU/viewform

ensure that $\Delta s \leq 0$ with a confidence of 95% we calculate the right handed confidence interval (**z-value** = $\Delta s + z_{0.05} \cdot \sigma_{\Delta s}$ where $z_{0.05} = 1.644854$). If the **z-value** > 0, the model fails the test, otherwise it passes the test.

3 Results

3.1 Evaluating Mole Classification Models on Held-out Accuracy

As per the single aggregated test accuracy statistic (evaluated against ground truth) Model 87 performs the best at 83.28% (refer to Table 2). We highlight that even though the validation accuracies of each of the models vary significantly, the test accuracies remain in the same neighborhood.

3.2 Human Performance on Defocus, Dimness and Brightness

Participants only loose their prediction ability when the image is defocused with a Δv of 0.87. When the Δv is less, they can preserve their prediction 93.75% (s_0) of the time, demonstrating strong resilience to the defocus transformations (Table 3). It must be noted that there is a dip in the percent preserved prediction (ppp) around Δv of 0.6 when defocus is applied (Figure 2). Participants loose their prediction ability a lot earlier in the case of a dimness transformation. They are only able to preserve their prediction less than 75% (s_0) of the time when $\Delta v \geq 0.40$. This indicates that humans may have a reduced ability to preserve their prediction under dimming conditions. It can be seen in Figure 2 that participants are not able to notice the similarity between the original and transformed image at all after the image has been dimmed by a $\Delta v \geq 0.7$. For brightness transformations, participants preserved their predictions 70.00% of the time, for $\Delta v < 0.93$. The t_p for brightness was higher than both dimness and defocus. These results suggest that human perception is most robust to brightness and most vulnerable to dimness (Table 3).

3.3 Evaluating Mole Classification Models on Prediction Preservation (t_p)

The results in Table 1 demonstrate that despite such a high test and validation accuracy (refer to Table 2), none of the models **pass/satisfy** the prediction preservation (t_p) test. The $\Delta s < 0$ in 5 cases, but the large $\sigma_{\Delta s}$ value causes the the 95% confidence interval (z-value) to span beyond 0. Thus, if the confidence interval were decreased, the models could have passed/satisfied the prediction preservation (t_p) test. The reliability distance (Δs) was always negative for the brightness transformation for all models. The brightness seemed to affect the models' prediction preservation (t_p) much less compared to all the other transformations. This same observation (less sensitivity to brightness transforms) was seen in the case of humans (refer to Table 3). For both the defocus and the brightness transformations, Model 78 was the best at preserving the prediction despite its low validation accuracy and test accuracy (refer to Table 2). Model 87 was best at preserving the prediction after the dimness transform amongst all the models. The dimness transformations had the worst z-values (higher) overall indicating that the models really struggled to deal with dimness. The same pattern was seen with humans, they were the most sensitive to dimness ($t_p = 0.40$). Despite having test accuracies in a similar neighbourhood, Model 78 (less epochs) was a lot closer to passing the prediction-preservation than any of the other models. This really underscores the point that only one aggregated accuracy statistic may not be representative of the strengths and weaknesses of the model.

4 Discussion & Limitation

Our study aims to establish a human decision boundary on medical imaging to verify the robustness of MVCs. However, the study has flaws in its setup. Due to budget and time constraints, we conducted the survey with untrained individuals. We aimed to ensure that untrained individuals could participate effectively and that human perceptions of features remained consistent. However, it is important to acknowledge that complete elimination is inherently unattainable. People's perceptions of what constitutes a feature are diverse, and their standard of what is identified as similar and different can vary. This can be observed from the responses of the survey's error questions. These questions should be trivial to answer, in which the same image or a completely different image is shown. However, a noteworthy observation was that in 5/127 instances, participants thought two very different images were the same. In 2/63 instances, they thought that two of the exact same images were different. In their feedback, many individuals also mentioned being grossed out by the mole images. This may have contributed to them hurrying up to finish the survey and may have resulted in these lapses of judgment.

Another limitation that we share with the original study is on the visual change metric Δv. VIF and VSNR are good measurement for changes of pixel-level transformations (defocus, brightness-change, color shift). However, for geometric transforms (flipping, zooming, rotation, crop), these measurements falls apart. Future work can employ other metrics to improve on this.

Our approach offers a cost-effective method for phrasing questions to gather data, enabling the evaluation of prediction-preservation in the absence of medical professionals. However, this method is not applicable when the correctness of the prediction is required, for example when verifying correctness-preservation.

5 Conclusion

To conclude, we developed a pipeline to assess the reliability of an ML-based MVC for mole classification within the context of an early detection smartphone app. We successfully applied the concept of prediction-preservation (t_p), as defined by Hu et al. to the medical imaging space. However, we were unable to implement correctness-preservation (t_c) due to limitations in time, budget, and the absence of medical professionals. For future work, ones can aim to conduct a survey with medical professionals to establish a human correctness-preservation baseline (t_c) for these transformations. Additionally, ones can to explore natural transformations, rather than the artificial ones used in this study. This would require exploring a different metric (other than Δv) to more effectively in capturing the changes in the images in those settings.

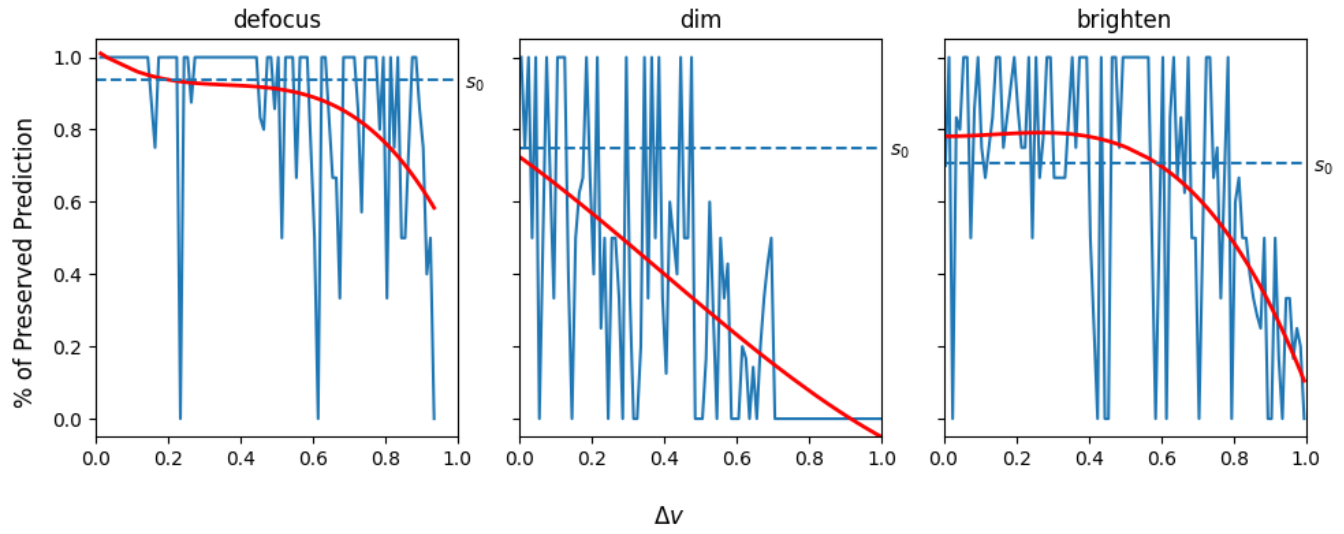


Figure 2: The result from human survey for the prediction-preservation on mole images used to derive the human decision boundary t_p. The x-axis is Δv grouped into a bin size 0.01. The y-axis is the percentage of "Yes" in the human responses of the particular bin (aka, the percent prediction preservation). The s_0 baseline is shown for each transform type, which is the percentage of "Yes" responses of the lower 5th quantile of the Δv. The red line is the fitted spline used to remove the outliers.

Model	Transform	$\hat{s_0}$	σ_{s_0}	$\hat{st_p}$	σ_{st_p}	Δs	$\sigma_{\Delta s}$	z-value	Decision
Model 78	defocus	0.868000	0.074512	0.8790	0.054873	**-0.011000**	0.092536	0.141209	Not Satisfied
Model 78	dim	0.820700	0.074234	0.7942	0.057222	0.026500	0.093729	0.180670	Not Satisfied
Model 78	brighten	0.906383	0.066252	0.9232	0.041155	**-0.016817**	0.077994	0.111472	Not Satisfied
Model 83	defocus	0.802400	0.072844	0.7774	0.055689	0.025000	0.091692	0.175820	Not Satisfied
Model 83	dim	0.830717	0.087732	0.8004	0.058377	0.030317	0.105379	0.203649	Not Satisfied
Model 83	brighten	0.878367	0.067370	0.8844	0.046868	**-0.006033**	0.082070	0.128959	Not Satisfied
Model 87	defocus	0.755200	0.097534	0.7380	0.064094	0.017200	0.116709	0.209169	Not Satisfied
Model 87	dim	0.854767	0.070527	0.8554	0.048154	**-0.000633**	0.085398	0.139834	Not Satisfied
Model 87	brighten	0.877100	0.072888	0.8860	0.051108	**-0.008900**	0.089020	0.137525	Not Satisfied

Table 1: The result of verifying the MVC's prediction-preservation requirement of a each transform, based on the human decision boundary t_p. Showing all test statistics: $(\hat{s_0}, \sigma_{s_0})$ and $(\hat{st_p}, \sigma_{st_p})$ are the estimate of the mean and standard deviation of s_0 and s_{t_p} given the results from the bootstrap. The reliability distance and standard deviation are the Δs and $\sigma_{\Delta s}$. The z-value is the right-handed confidence interval that $\Delta s \leq 0$, the row highlighted in green and red shows the transformation that the model is best and worst at, respectively. The test decision is shown in the last column. Note that a few Δs values are negative (shown in bold), indicating those tests would pass if the standard deviation were lower.

Model Name	Epochs	Validation Acc.(%)	Test Acc. (%)
Model 87	16	87.25%	83.28%
Model 83	27	83.89%	82.78%
Model 78	39	78.52%	81.44%

Table 2: The table shows accuracies obtained from evaluating 3 Mole Classification Models (same architecture, but trained for different epochs) on the held-out validation and test sets.

	Defocus	Dim	Brighten
s_0	93.75%	75.00%	70.00%
t_p (Δv)	0.87	0.40	0.93

Table 3: Shows the calculated human boundary for prediction-preservation (t_p) in a mole classification task. Here, the t_p is the Δv value after which humans can no longer maintain a % prediction preservation $\geq s_0$ (baseline % prediction preservation).

References

[1] Ana Barragán-Montero, Umair Javaid, Gilmer Valdés, Dan Nguyen, Paul Desbordes, Benoit Macq, Siri Willems, Liesbeth Vandewinckele, Mats Holmström, Fredrik Löfman, Steven Michiels, Kevin Souris, Edmond Sterpin, and John A.

Lee. 2021. Artificial intelligence and machine learning for medical imaging: A technology review. *Physica Medica* 83 (2021), 242–256. https://doi.org/10.1016/j.ejmp.2021.04.016

[2] Alexander Buslaev, Vladimir I. Iglovikov, Eugene Khvedchenya, Alex Parinov, Mikhail Druzhinin, and Alexandr A. Kalinin. 2020. Albumentations: Fast and Flexible Image Augmentations. *Information* 11, 2 (2020). https://doi.org/10.3390/info11020125

[3] Chaz Firestone. 2020. Performance vs. competence in human-machine comparisons. *Proc Natl Acad Sci U S A* 117, 43 (Oct. 2020), 26562–26571.

[4] N. Hasani, M. A. Morris, A. Rhamim, R. M. Summers, E. Jones, E. Siegel, and B. Saboury. 2022. Trustworthy Artificial Intelligence in Medical Imaging. *PET Clinics* 17, 1 (2022), 1–12. https://doi.org/10.1016/j.cpet.2021.09.007

[5] Boyue Caroline Hu, Lina Marsso, Krzysztof Czarnecki, Rick Salay, Huakun Shen, and Marsha Chechik. 2022. If a human can see it, so should your system: reliability requirements for machine vision components. In *Proceedings of the 44th International Conference on Software Engineering* (Pittsburgh, Pennsylvania) *(ICSE '22)*. Association for Computing Machinery, New York, NY, USA, 1145–1156. https://doi.org/10.1145/3510003.3510109

[6] Alain Jungo and Mauricio Reyes. 2019. Assessing Reliability and Challenges of Uncertainty Estimations for Medical Image Segmentation. In *Medical Image Computing and Computer Assisted Intervention – MICCAI 2019*, Dinggang Shen, Tianming Liu, Terry M. Peters, Lawrence H. Staib, Caroline Essert, Sean Zhou, Pew-Thian Yap, and Ali Khan (Eds.). Springer International Publishing, Cham, 48–56.

[7] American Academy of Dermatology Association. [n. d.]. What to look for: ABCDEs of melanoma — aad.org. https://www.aad.org/public/diseases/skin-cancer/find/at-risk/abcdes. [Accessed 01-12-2024].

[8] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP models with CheckList. arXiv:2005.04118 [cs.CL] https://arxiv.org/abs/2005.04118

[9] Abdullah Sattar. 2014. Image Quality Tools. https://github.com/sattarab/image-quality-tools. Accessed: 2024-12-01.

[10] Felix A. Wichmann, David H. J. Janssen, Robert Geirhos, Guillermo Aguilar, Heiko H. Schütt, Marianne Maertens, and Matthias Bethge. 2017. Methods and measurements to compare men against machines. *Electronic Imaging* 29, 14 (2017), 36–36. https://doi.org/10.2352/ISSN.2470-1173.2017.14.HVEI-113

Appendix A: Survey Example

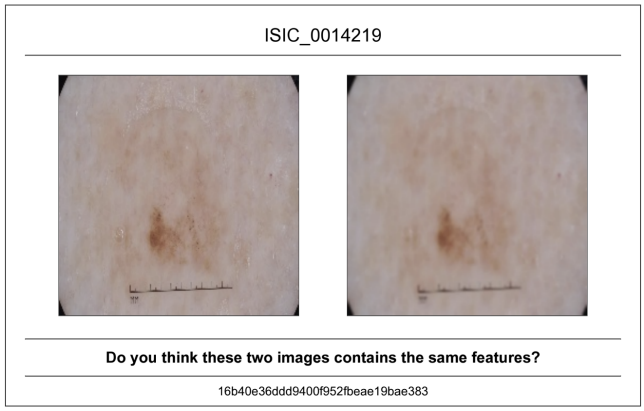
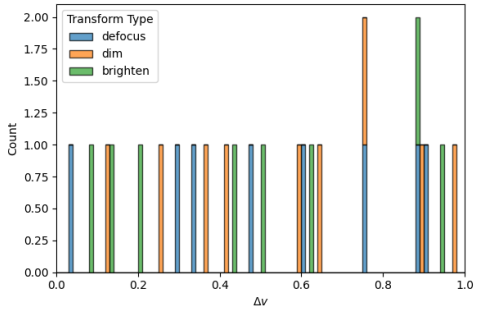
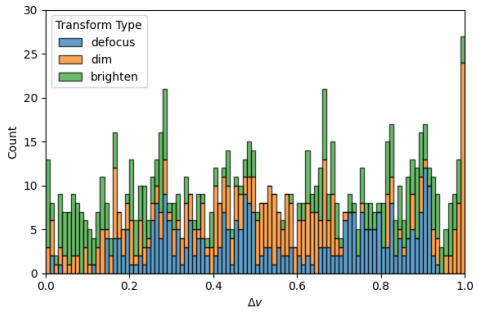


Figure 3: An example survey question shown to the participant. The left shows the original mole image and the right shows the transformed image. The survey question follows below the two images. For debugging and reference purposes, the source name of the original image and the transformation ID are displayed at top and bottom of the box, respectively.

Appendix B: Survey Sample Distribution



(a) sampling result for survey 0



(b) sampling result for all surveys

Figure 4: Stacked histograms show Δv distribution and transformation type for sampled transformed images. X-axis: Δv in 0.01-sized bins. Y-axis: count of images per bin, categorized by transformation type (defocus, dim, brighten). Graph (a) is for survey 0, graph (b) shows the overall result with all surveys combined. Note in (a), samples are evenly distributed across bins, with nearly equal amount of each transform types. In (b), every bin has entries, most including all transformation types.

Appendix C: Project Dependency

Module	Dependencies
Transform Image	albumentations[2]
Δv (VIF, VSNR)	matlabengine github.com/sattarab/image-quality-tools
PDF Generation	django, WeasyPrint
Survey	Google Form, Google Apps Script
Model Training	pytorch
Data Analysis	pandas, numpy, scipy, matplotlib

Table 4: Dependencies for each modules of the project, including python packages and other dependencies.