Agreement Between Saliency Maps and Human-Labeled Regions of Interest: Applications to Skin Disease Classification

Nalini Singh[*]
MIT
nmsingh@mit.edu

Kang Lee
Google Research
kanlig@google.com

David Coz
Google Health
dcoz@google.com

Christof Angermueller
Google Research
christofa@google.com

Susan Huang[†]
susanhuang@google.com

Aaron Loh
Google Health
aaronloh@google.com

Yuan Liu
Google Health
yuanliu@google.com

Abstract

We propose to systematically identify potentially problematic patterns in skin disease classification models via quantitative analysis of agreement between saliency maps and human-labeled regions of interest. We further compute summary statistics describing patterns in this agreement for various stratifications of input examples. Through this analysis, we discover candidate spurious associations learned by the classifier and suggest next steps to handle such associations. Our approach can be used as a debugging tool to systematically spot difficult examples and error categories. Insights from this analysis could guide targeted data collection and improve model generalizability.

1. Introduction

A growing body of work [4, 5, 10, 15, 35, 38] investigates the use of deep learning to classify skin conditions from medical images. Historically, the machine learning community has developed models that achieve high classification accuracy on a test set, ostensibly drawn from the same distribution that the model is trained on. However, in practice, the data that the model is trained on frequently differs from the data on which the model is deployed [19]. Patient populations between different sites vary in demographics and disease presentation, and these characteristics shift in time. Further, datasets are frequently sourced from a few sites with specific image acquisition procedures that may not generalize to other sites [34].

For this reason, beyond optimizing performance on a pre-defined test set, it is useful to understand *how* a model makes predictions. First, this offers intuition to practitioners as to when the model will fail. Such intuition allows further model improvement by targeting data collection towards difficult out-of-distribution samples or by optimizing model architectures or losses to mitigate such failures. Moreover, when deploying the model, one could use model explanations to “gate” the use of the machine learning system. For example, when the model makes a prediction for an inappropriate reason, the system could refer patients to clinicians instead of presenting the prediction.

Several model explanation methods have been introduced in the literature to analyze how convolutional neural networks (CNNs) use input features to make predictions. We focus on saliency maps, a class of model explanations which highlight pixels in an input image that the network deems important. Saliency maps provide no information about *how* pixels are used to arrive at a classification, so appropriate image regions being highlighted does not *prove* that the network used the right logic for that input. However, a saliency map that highlights regions that differ from human intuition indicates either: (1) the model learned an inappropriate strategy for classifying the image, (2) human intuition regarding relative importance of features for classification is misguided, or (3) the saliency map was generated with an inappropriate technique. Given a saliency map generated via an appropriate method, qualitative examination makes it easy to identify when either (1) or (2) are occurring. However, it is difficult to extend this technique to large datasets as one has to sift through and visually assess many images to catch inputs of interest.

In this work, we propose that, in addition to assessing model classification accuracy, model explanations should be systematically evaluated to efficiently identify potential spurious correlations learned by the model. To that end, we quantitatively evaluate **model-human explanation agreement**, the extent to which saliency maps agree with human-labeled regions of interest in images. We use two metrics

[*] Work done as an intern at Google Health.
[†] Work done at Google Health via Advanced Clinical.

that can be easily applied to input examples and aggregated across arbitrary stratifications of large datasets, making it easy to both spot individually troublesome input examples and analyze broad trends in network performance. This strategy can be applied to saliency maps generated via any technique, and thus is broadly applicable even as methods for generating accurate saliency maps continue to improve.

We demonstrate our approach on a model trained to make diagnoses on the dataset of dermatology images in [15]. In particular, we answer three questions:

1. **On which input examples is model-human explanation agreement lowest?** For many cases with low model-human explanation agreement, the model explanations frequently focus on lips, hair, and fingernails unrelated to the skin pathology present in the image, as confirmed by a board-certified dermatologist. This finding holds for both correctly and incorrectly classified input images.

2. **How does model-human explanation agreement vary across different skin condition classes?** Model-human explanation agreement is lowest for *different* conditions when considering incorrectly classified and correctly classified images. Model-human explanation agreement is low for incorrectly-classified seborrheic keratosis and melanoma images and for correctly-classified androgenetic alopecia and acne images.

3. **How does model-human explanation agreement vary across different demographic groups?** Model-human explanation agreement is consistent across age and skin tone, but varies with sex.

These results suggest that further development of the model in question may involve exposure to a larger, more varied set of images that include normal anatomy such as fingernails and lips or design of loss terms to account for them, in order to encourage the model to learn from image regions directly related to the disease being studied. This is particularly true for seborrheic keratosis and melanoma, two conditions where we found the current model to suffer the most in this regard.

Though we perform this analysis on a specific dermatology dataset, the techniques described in this paper are general and could be applied to any CNN in order to identify spurious correlations learned by the model.

2. Background

2.1. Dermatology Disease Classification Techniques

There has been significant prior work on detecting skin conditions from images. Several groups [4, 35] developed skin lesion classification and segmentation algorithms for dermoscopic images. Recently, interest has increased in automated skin lesion classification from photographs taken by consumer grade cameras [1, 18, 21]. These photographs are easier to acquire by non-specialists and are used for a wider range of dermatology conditions, but also typically exhibit greater variability in composition.

In [10], the authors developed a CNN to distinguish cancerous and benign lesions from clinical photographs. [38] developed a skin disease recognition system designed to mimic a dermatologist's decision-making process to diagnose roughly 200 different skin lesions. [15] uses a deep network trained on a much larger and more diverse set of clinical skin photographs to diagnose 26 common skin conditions referred to a teledermatology service. [15] improves on previous models by incorporating a variable number of images and additional clinical metadata about each case, providing a differential diagnosis instead of a single classification, and demonstrating comparable performance to dermatologists. For those reasons, this paper analyzes model-human explanation agreement for the model in [15].

2.2. Dermatology Dataset Diversity and Fairness

In [15], significant variation exists in the composition of the photographs that the model was trained and evaluated on. Besides the skin pathology, the image content often contains background objects including furniture, clothing and accessories, as well as body parts with no skin pathology. Further, dermatology images may contain skin markings and scale references, which have been shown to significantly increase the false positive rate on a melanoma detection task [17, 36]. Thus, it is important to characterize the impact of these photograph attributes on model behavior.

In addition to diversity in photo composition, the patient cohort used in [15] included images from patients of varying age, biological sex, and skin tone, which often exhibit different visual appearance and prevalence of dermatology conditions. Previous work [3] has demonstrated disparities in model performance across sex and skin tone on a facial recognition task, while [14] found no measurable correlation between model performance and skin tone values on skin disease classification tasks. The model in [15] showed slight variations on model accuracy across these demographic characteristics. To the best of our knowledge, however, there has been no previous analysis of how agreement in classification explanations by humans and models vary within such a diverse patient cohort.

Figure 1 shows sample images demonstrating the variability in image content, skin tone, and patient age.

2.3. Saliency Maps and Integrated Gradients

What constitutes a useful model explanation is a topic of significant debate [9, 11, 13]. One subset of these techniques explain how a model processes particular inputs [23, 25, 26,

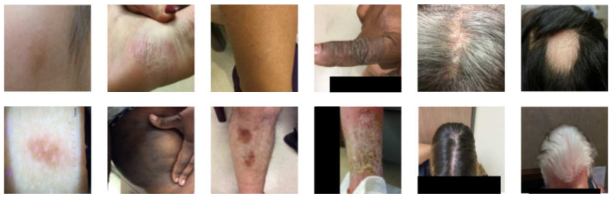

Figure 1: Sample images in the dataset. Some images are masked to omit sensitive information; the model had access to the full image, including this information.

30,31,39]. Other techniques describe what each component of a neural network does [2,16,24,40]. Yet another set alters the network to explicitly include an interpretability component or to generate explanations [12,20,37]. Our work uses saliency maps, a model explanation technique from the first of these subsets. A saliency map assigns a score to each input pixel specifying how 'important' it was for the model performing the task.

In this paper, we use Integrated Gradients [31] to generate saliency maps. In this technique, a baseline image x' is specified. For each input x, the gradient of the model output score function F is computed at each of m images on the line segment between x' and x. The average gradient is scaled by the difference in pixel intensity to produce an importance score for each pixel i in the image:

$$IG_i(x) = (x_i - x'_i)\frac{1}{m}\sum_{k=1}^{m}\frac{\partial F(x' + \frac{k}{m}(x - x'))}{\partial x_i}$$

We chose Integrated Gradients because this technique satisfies the properties of sensitivity and implementation invariance [31], unlike other explanation techniques mentioned above. Internal experiments on a single-image classification model on the dataset from [15] showed minimal differences between saliency maps generated via Integrated Gradients and GradCAM [23], suggesting that our conclusions are not highly sensitive to the saliency map generation method.

We combine Integrated Gradients with the SmoothGrad [27] technique, in which saliency maps are averaged over noisy versions of each input.

2.4. Quantitative Explanation Evaluation

Previous work on quantitative evaluation of model explanations has focused on the quality of the model explanation as a representation of the inner workings of the model. [2] quantifies how well the activations of neurons segment human-labeled concepts within an image. This allows fine-grained understanding of individual neurons within the network, but it is challenging to transfer this approach to new models because a large number of explicit concept-labels must be hand-crafted and collected for new datasets. Alternatively, [22] proposed a perturbation-based approach that removes pixels with the highest attributions and examines the effect on the network's classification score. Relatedly, [6] proposed to crop input images to bounding boxes containing all salient regions and determine whether the network is still able to classify appropriately. This class of metrics quantifies whether saliency maps accurately represent the image pixels used to make network predictions, but does not measure whether such pixels agree with human intuition behind the decision-making process.

[17] takes an important step toward this goal by quantitatively comparing saliency maps with segmentations of an explicitly specified potential source of bias (ink markings on skin images). However, this requires pre-existing knowledge of image components that might bias the model. Instead, we quantitatively compare saliency maps with human segmentations of regions that we expect *should* be used for model learning. This allows us to identify potential sources of bias without hypothesizing a priori that might exist.

3. Methods

3.1. Image Dataset

We used the dataset in [15], comprised of 19,870 adult patient dermatology cases submitted to a teledermatology company from 2010 to 2018. Each case comprised 1-6 clinical images taken by medical assistants, along with 45 metadata fields specifying the patient's demographic information, medical history, and symptoms. All data were de-identified according to HIPAA Safe Harbor prior to transfer to study authors. The protocol was reviewed by Advarra IRB (Columbia, MD), which determined that it was exempt from further review under 45 CFR 46.

The reference standard skin condition for each case was based on aggregated opinions of multiple dermatologists from a panel of U.S. and Indian board-certified dermatologists. When reviewing the images and additional medical information associated with a case, each dermatologist independently generated a differential diagnosis, comprised of a set of diagnosis codes and their respective confidence. Diagnosis codes were drawn from the Systematized Nomenclature of Medicine-Clinical Terms (SNOMED-CT) or entered as free-text if no appropriate SNOMED term was found. These diagnosis codes were manually mapped to a short list of 419 conditions of the appropriate granularity by three U.S. board-certified dermatologists. Ultimately, the 26 most prevalent conditions on this list were chosen, along with an 'Other' category assigned to the remaining conditions. Cases that contained multiple conditions or that were deemed not diagnosable were removed from the dataset.

Cases collected between 2010 and 2017 were designated as the development set to train and optimize the model (16,114 cases), whereas cases collected afterwards were used as the test set to conduct all evaluations (3,756 cases).

No patients were present in both the development and test set. More details about the construction of the dermatology dataset are available in [15].

3.2. Classification Model

We used the network from [15] to predict condition labels from the images and metadata associated with each input case. Up to 6 input images per case were fed as input to a neural network with the Inception-v4 [33] architecture. The pre-logit layers of each of these sub-networks were averaged and concatenated with a one-hot vector encoding of the case metadata; this final representation was fed to a softmax layer outputting a vector of classification scores for each of the 27 classes (26 conditions + 'Other' category).

The weights of each Inception-v4-like arm of the network architecture were initialized to weights used for classification of the ImageNet dataset [7]. Then, network weights were optimized using stochastic gradient descent on the training cases, augmented by random cropping, rotation, flipping, and color perturbation to each image. Training proceeded for 100,000 steps with a batch size of 8.

The ensemble version of the model was shown to perform comparably with dermatologists and superior to primary care physicians and nurse practitioners on a test subset, with top-1 accuracy of 66%, 63%, 44%, and 40% respectively. Additional details about the model architecture, training, and evaluation procedure are available in [15].

3.3. Pathology ROI Label Collection

We collected at least one 'region-of-interest' (ROI) label for 1907 images from 1309 cases in the test set, sampled randomly from the distribution of conditions present in that dataset. For each image, three dermatologist-trained graders specified polygon-shaped regions containing pathology as binary masks; pixels with pathology were assigned to 1, and all other pixels were assigned to 0.

Graders first indicated whether a skin condition was visible within each image. For images with clearly-visible skin conditions, graders then labeled polygonal ROIs on each image. Some images were ambiguous to segment into a single ROI; for example, rash-like conditions often present as diffuse markings dispersed across large sections or several patches across the image. Graders determined whether there were fewer than five distinct pathology ROIs in the image. If so, graders individually outlined each ROI. If more than five ROIs were present in the image, graders indicated this and did not provide an ROI for the image. All subsequent analysis was done only on images for which all ROIs were labeled. For each image, ROI labels of different graders were combined by pixelwise majority vote to produce a consensus ROI label.

We collected consensus ROI labels from three graders out of a pool of 50 graders for each image in the evaluation set. The Fleiss kappa value, which measures inter-rater agreement, was 0.65±0.27 across the dataset; its distribution is shown in Figure 2. Images with Fleiss kappa below 0.4 were discarded, resulting in a filtered dataset of 1526 images from 1083 cases, with an average Fleiss kappa value of 0.76±0.16. This dataset is henceforth referred to as the 'saliency evaluation dataset'. A diagram demonstrating how ROIs were generated and used to compute model-human explanation agreement is shown in Figure 3.

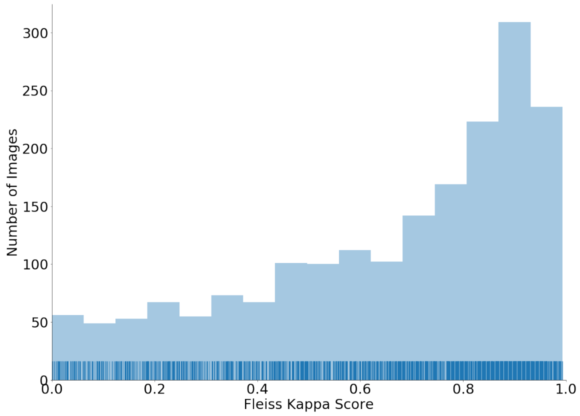

Figure 2: Distribution of Fleiss kappa scores indicating inter-rater agreement of ROI labels. Dark blue vertical lines represent individual datapoints within each histogram bin.

3.4. Saliency Map Generation

For each image in the saliency evaluation dataset, we used Integrated Gradients, described in Section 2.3, to generate saliency maps. Each saliency map was generated using SmoothGrad [27] on a 50-step path between a black baseline and input image. The final saliency map was normalized to range from -1 to 1. All saliency maps generated are visualized using the techniques described in [32].

3.5. Quantifying Explanation Agreement

We used two metrics to quantify model-human explanation agreement on each image: a thresholded Dice score and Spearman's rank correlation coefficient. We then used case-level agreement metrics to compute dataset summary statistics as in Sections 4.2 and 4.3. In particular, we computed these statistics on only the image with the highest DS and SRCC across the case, in order to avoid overweighting cases with multiple images and to understand the model's behavior on the most informative images.

The thresholded Dice score (DS) was determined by computing the Dice score [8,28] between a thresholded version of the continuous valued Integrated Gradients scores and the binary grader-based ROI labels. Choosing an appropriate threshold is non-trivial; the salient regions of different images vary in size and relative intensity. We chose the

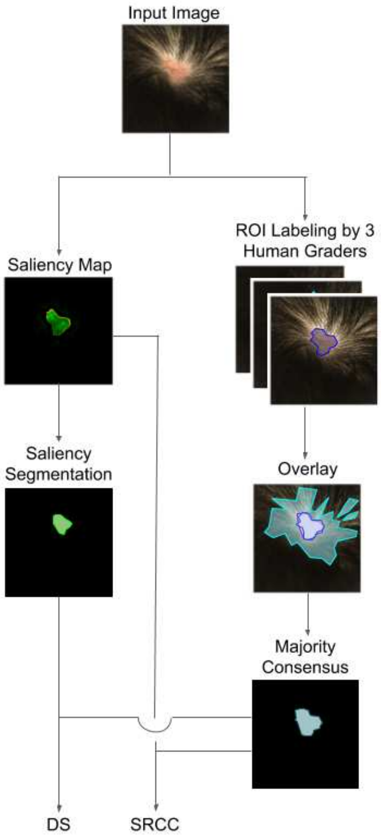

Figure 3: Process for generating explanation agreement metrics. ROIs are labeled by 3 dermatology-trained graders. Majority consensus selects the final ROI that is compared to both the raw saliency map and a binary version of this map.

threshold for each image to be the multiple of 0.01 between 0 and 1 that maximizes the computed Dice score for that particular image; since we follow this procedure for every image, relative rankings between examples remain valid.

We used Spearman's rank correlation coefficient [29] (SRCC) as a complementary metric that does not depend on the choice of a threshold and explicitly characterizes the relative rankings of attributions in the saliency map. SRCC is determined by computing the Pearson correlation coefficient between the ranks of the continuous valued scores produced by integrated gradients for each pixel and the corresponding pixel-wise binary human-graded ROI labels.

The pixelwise nature of the saliency maps yields metrics that are lower than the values in the segmentation literature, even when the generated saliency maps and the human-labeled regions of interest qualitatively agree. To provide context for what different agreement scores mean qualitatively, Figure 4 shows sample images, saliency maps, and human labeled ROIs, for different values of the two metrics.

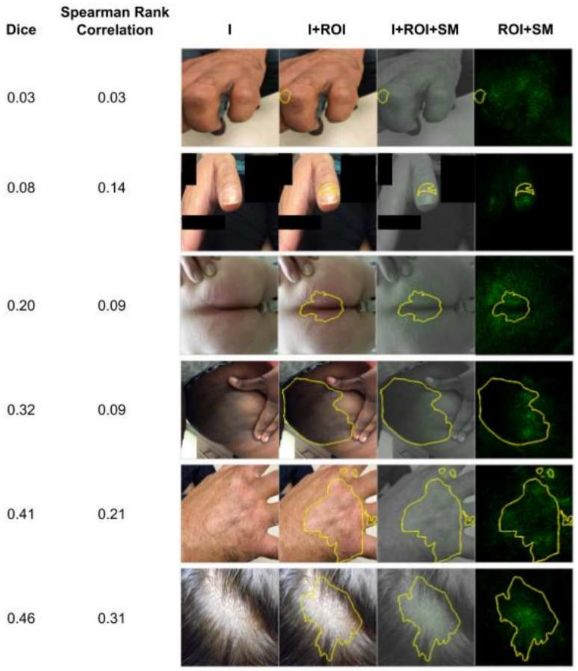

Figure 4: Images with varying thresholded Dice scores and Spearman's rank correlations. Each row shows an example with the original image (I), image overlaid with ROIs (I+ROI), image overlaid with ROIs and saliency map (I+ROI+SM), and saliency map overlaid with ROIs (ROI+SM). The Dice score metric ranges from 0 to 1, but the maximum value seen in our dataset was approximately 0.5. Some images are masked to omit sensitive information.

4. Experiments & Results

4.1. On which input examples is model-human explanation agreement lowest?

We ranked all images by DS and SRCC and qualitatively examined images with the best and worst model-human explanation agreement amongst correctly- and incorrectly-classified examples. Figure 5 shows example images for which the DS and SRCC fall within either the top 10% or the bottom 10% of the dataset; these consist of both correctly and incorrectly-classified images. To further understand low model-human agreement cases, we collected body part labels for images within the bottom 10% of the dataset. Trained human graders were instructed to label all body parts present in each image. The representation of various body parts present in the images within the bottom 10% of the dataset is shown in Figure 6.

4.2. How does model-human explanation agreement vary across different skin conditions?

We analyzed the DS and SRCC across cases stratified by disease type. We further stratified this comparison by pre-

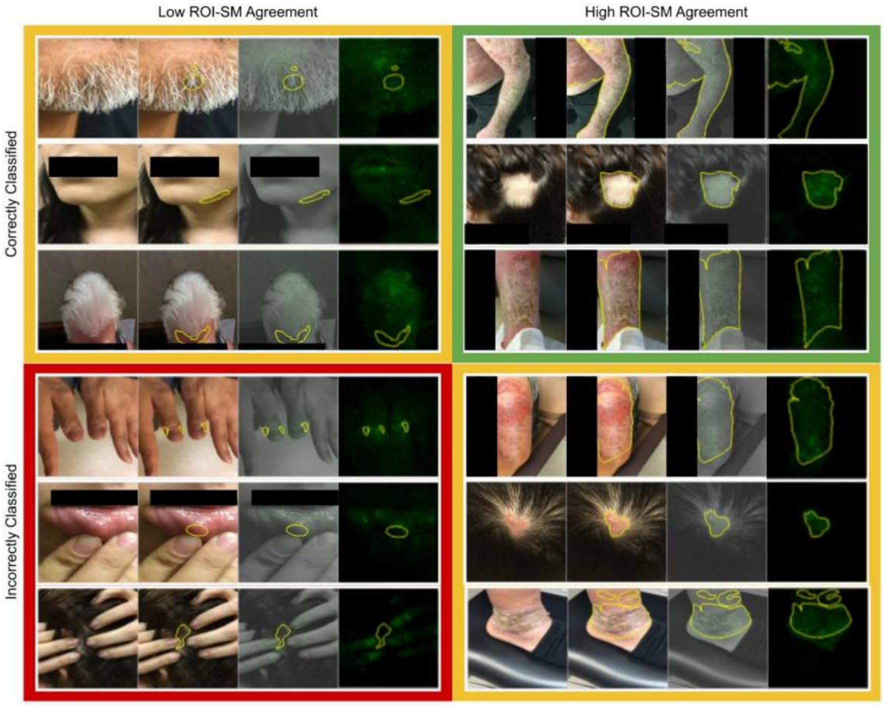

Figure 5: Examples with lowest (left) and highest (right) model-human explanation agreement between the model saliency map (SM) and human-labelled ROI, as measured by both thresholded Dice score and Spearman's rank correlation. Examples are further divided as correctly (top) or incorrectly (bottom) classified. Some images are masked to omit sensitive information.

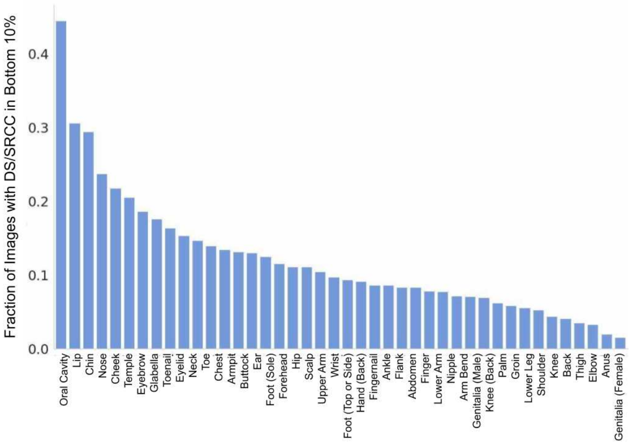

Figure 6: The fraction of examples containing a particular body part in the bottom 10% of the dataset in terms of both Dice score and Spearman's rank correlation. A large proportion of images with facial features (oral cavity, lip, and chin) had low model-human explanation agreement.

diction accuracy. Figure 7 shows these results for conditions with at least 5 cases containing ROI-labeled images.

Figure 7 also shows example images, ROIs, and saliency maps from the bottom-three conditions in terms of both DS and SRCC. Amongst images from correctly classified cases, those with androgenetic alopecia and acne demonstrated notably lower mean Dice scores compared to other conditions. Amongst images from incorrectly classified cases, melanoma and seborrheic keratosis demonstrated notably lower mean Dice scores compared to other conditions.

4.3. How does model-human explanation agreement vary across different demographic groups?

We performed statistical analysis to compare case-level model-human explanation agreement for sex, skin tone, and age, each in terms of DS and SRCC. For these six analyses, we used a Bonferroni-adjusted $\alpha = 0.05/6 = 0.0083$. The results in Figure 8 indicate a significant difference in model-human explanation agreement based on sex (Two-sample t-test | DS: *t=-3.67, p<0.001*; SRCC: *t=-3.36, p<0.001*), but not based on skin tone (1-way ANOVA | DS: *f=0.42, p=0.83*; SRCC: *f=2.66, p=0.02*) or age (Pearson correlation | DS: *ρ=0.10, p<0.001*; SRCC: *ρ=0.04, p=0.16*).

5. Discussion

The left column of Figure 5 suggests that images with certain body parts (e.g., lips, hair, and fingernails) demonstrate the starkest differences between saliency maps and human-labeled ROIs. This finding is further verified via the

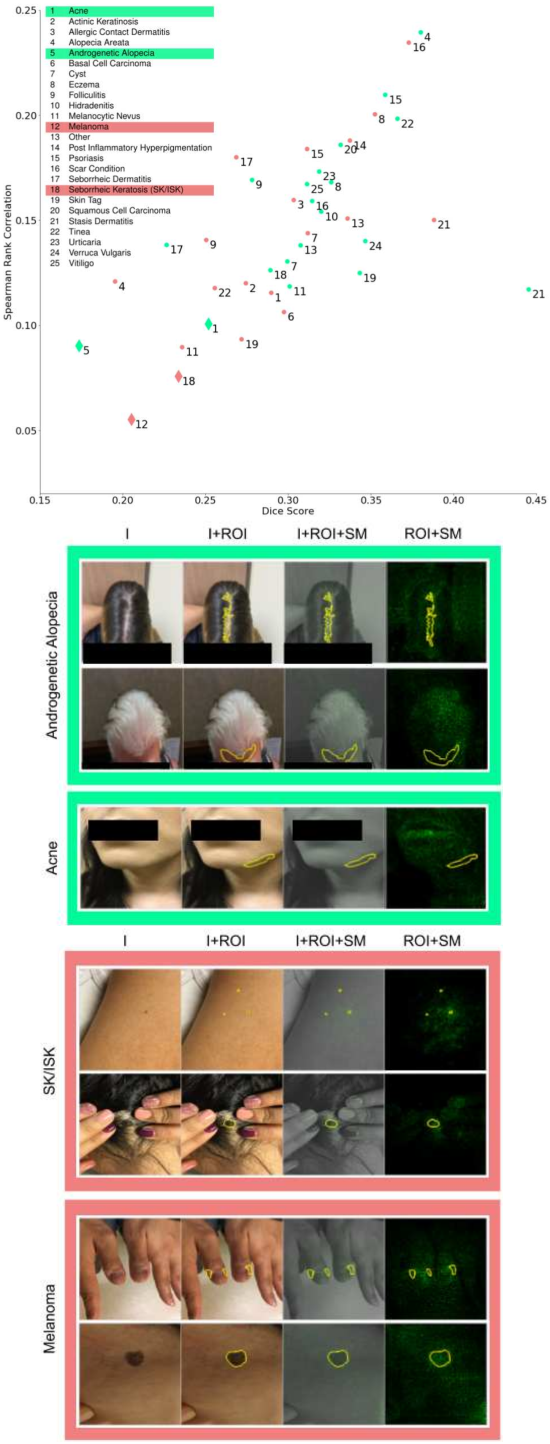

Figure 7: Model-human explanation agreement by skin condition classes. The scatterplot compares thresholded Dice scores and Spearman's rank correlation examples of correctly (green) and incorrectly (red) classified cases. Diamonds denote conditions amongst the bottom-3 lowest in terms of both DS and SRCC; sample images, ROIs, and saliency maps for these conditions are shown below. Some images are masked to omit sensitive information.

body part distribution for low agreement cases in Figure 6 (note that this figure plots *all* present body parts in an image, not just the ones highlighted by its corresponding saliency map). The model attends more strongly to this "normal anatomy" even when clear skin pathology is present in a spatially distinct location. After examining the example images, a board-certified dermatologist confirmed the conclusion that this anatomy is inappropriate as primary criteria to determine skin conditions for those cases. Though our analysis does not determine what *causes* specific model predictions, the repeated presence of these elements in images with low model-human explanation agreement suggests that the model might associate these features with certain conditions, instead of focusing on the pathology of interest itself.

In contrast, the right column of Figure 5 shows cases where the model correctly focused on pathology. This accurate saliency localization makes intuitive sense for correctly classified examples. However, for incorrectly classified examples, the model focuses on the correct spatial locations but may be misinterpreting the texture. Model explanations beyond saliency maps are needed to understand this issue.

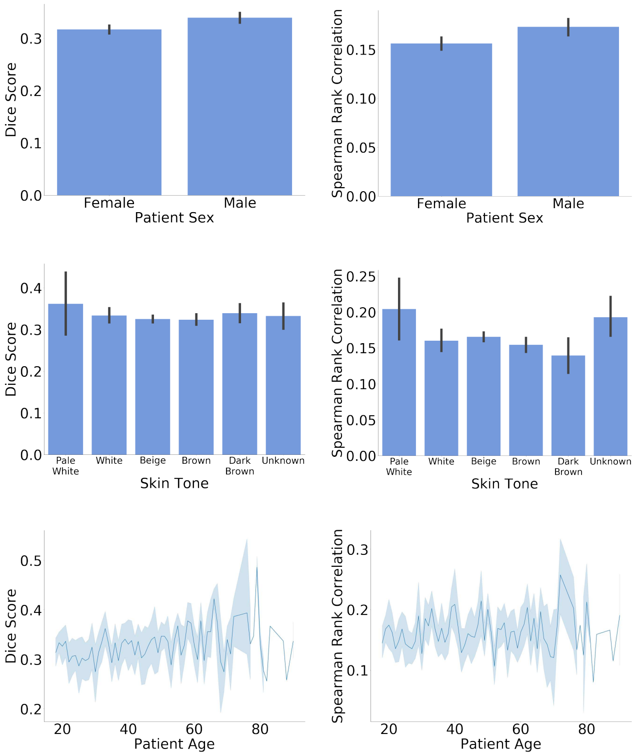

Figure 8: Model-human explanation agreement by demographic groups. From top to bottom, the plots represent the thresholded Dice score (left) and Spearman's rank correlation coefficient (right) for different sexes, skin tones, and ages respectively. In all bar plots, error bars indicate a 95% bootstrap confidence interval.

When stratified by model prediction accuracy, androgenetic alopecia and acne images have lower mean human-model explanation than other conditions, a finding consistent with the qualitative observation that the model frequently misattends to hair and lips in images. The androgenetic alopecia finding indicates that either: (1) the model correctly uses hair as context with which to compare hair loss that occurs in androgenetic alopecia, or (2) the model has learned that images that have hair in them are more likely to be androgenetic alopecia, regardless of the specific hair loss pattern. Though we cannot definitively accept either of these hypotheses, qualitative analysis of the examples shows that the model attends to regions where the hair is present and not where the hair is absent, which is inconsistent with the first hypothesis. Similarly, many of the acne images in our dataset show acne on the face. This, combined with the attention paid by the model to the patient's lips instead of the affected lesion, suggest that the model might have learned to associate facial features with acne.

Understanding why certain conditions exhibit particularly low human-model explanation agreement on incorrectly classified exmamples is more difficult. Though sample images from these classes confirm the qualitative observation that the model occasionally attends to fingernails (and other "normal anatomy" image features) instead of relevant lesions, there is no intuitive connection between these characteristics and the corresponding disease categories (melanoma and seborrheic keratosis). These are both conditions with low representation in the training dataset (0.6% and 4.4%, respectively); one hypothesis is that, without enough training examples to construct a representation of the underlying pathology, the model instead focuses on auxiliary image components. If so, training with additional melanoma and seborrheic keratosis images could improve both the model-human explanation agreement and the accuracy on these conditions. Alternatively, these results suggest that incorporating an object-detection or pre-processing algorithm to identify or remove these components of images could encourage the model to generalize better.

Finally, stratification by demographic groups confirmed that skin tone and age do not significantly affect human-model explanation agreement, even though these characteristics affect the visual appearance of skin and hair. On the other hand, stratification by sex *did* yield a slight, statistically significant difference in human-model explanation agreement. This might be due to the difference in natural distributions of conditions with varying levels of agreement.

In all of our experiments, we observed similar trends in the thresholded Dice score and Spearman's rank correlation. While these metrics differ in specific examples as shown in Figure 4, the summary statistics across different data stratifications demonstrate similar trends for both metrics.

One concern about our method is that collected ROIs represent only pathological regions, and that it is reasonable for the model to attend to other image context (e.g. to understand skin tone, compare to normal skin characteristics, etc.). We address this issue via adaptive thresholding of the Dice score. In particular, while we might expect the model to attend to external context, it seems unlikely that these "non-pathology ROIs" should be *more* important than pathology ROIs. We pick a threshold on the saliency map to produce a 'saliency segmentation' that maximizes the Dice score with the labeled pathology ROI. If the model is behaving correctly, we expect there to exist a threshold at which the saliency segmentation includes pathology ROIs but not non-pathology ROIs, yielding a high Dice score. Thus, even if the model does use non-pathology ROIs, if it selects the correct the pathology ROIs, we would expect a high Dice score. By a similar argument, we would expect a high Spearman's rank correlation, since that metric is explicitly based on pixel rankings; even if the model uses both non-pathology and pathology ROIs, we would expect the pathology ROIs to rank higher. However, the assumption that pathology ROIs should be more important than non-pathology ROIs might not hold. Future work to address this would increase the robustness of our technique.

Another potential concern is that our method is sensitive to the underlying saliency map generation technique. However, our strategy is applicable to maps generated via *any* technique. Thus, the high-level method in this paper is applicable even as saliency map generation techniques continue to improve. Further, this high-level method could be applied to maps generated via a suite of techniques and used to identify persistent trends.

6. Conclusion

We conducted a quantitative assessment of the agreement between model-based saliency maps and human-labeled regions-of-interest in a skin condition classification task from consumer-grade camera images. We also computed statistics that summarize trends in this agreement for different skin conditions and demographics. We found that several examples for which model-human explanation agreement were lowest were cases in which the model identified normal anatomy (e.g. lips, hair, and fingernails without pathology) as important for diagnosing disease, particularly for correctly-classified androgenetic alopecia and acne examples and for incorrectly-classified melanoma and seborrheic keratosis examples. Further, we found significant differences in model-human explanation agreement between different sexes, but not between groups of different age or skin tone. These findings suggest future data collection and model development strategies that could improve network performance and generalizability.

References

[1] Wiem Abbes and Dorra Sellami. High-level features for automatic skin lesions neural network based classification. In *2016 International Image Processing, Applications and Systems (IPAS)*, pages 1–7. IEEE, 2016. 2

[2] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. Network dissection: Quantifying interpretability of deep visual representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6541–6549, 2017. 3

[3] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91, 2018. 2

[4] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, pages 168–172. IEEE, 2018. 1, 2

[5] Angel Alfonso Cruz-Roa, John Edison Arevalo Ovalle, Anant Madabhushi, and Fabio Augusto González Osorio. A deep learning architecture for image representation, visual interpretability and automated basal-cell carcinoma cancer detection. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 403–410. Springer, 2013. 1

[6] Piotr Dabkowski and Yarin Gal. Real time image saliency for black box classifiers. In *Advances in Neural Information Processing Systems*, pages 6967–6976, 2017. 3

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 4

[8] Lee R Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945. 4

[9] Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017. 2

[10] Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115, 2017. 1, 2

[11] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. Explaining explanations: An overview of interpretability of machine learning. In *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*, pages 80–89. IEEE, 2018. 2

[12] Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating visual explanations. In *European Conference on Computer Vision*, pages 3–19. Springer, 2016. 3

[13] Frank C Keil. Explanation and understanding. *Annu. Rev. Psychol.*, 57:227–254, 2006. 2

[14] Newton M Kinyanjui, Timothy Odonga, Celia Cintas, Noel CF Codella, Rameswar Panda, Prasanna Sattigeri, and Kush R Varshney. Estimating skin tone and effects on classification performance in dermatology datasets. *arXiv preprint arXiv:1910.13268*, 2019. 2

[15] Yuan Liu, Ayush Jain, Clara Eng, David H Way, Kang Lee, Peggy Bui, Kimberly Kanada, Guilherme de Oliveira Marinho, Jessica Gallegos, Sara Gabriele, et al. A deep learning system for differential diagnosis of skin diseases. *arXiv preprint arXiv:1909.05382*, 2019. 1, 2, 3, 4

[16] Anh Nguyen, Alexey Dosovitskiy, Jason Yosinski, Thomas Brox, and Jeff Clune. Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In *Advances in Neural Information Processing Systems*, pages 3387–3395, 2016. 3

[17] Jacob Pfau, Albert T Young, Maria L Wei, and Michael J Keiser. Global saliency: Aggregating saliency maps to assess dataset artefact bias. *arXiv preprint arXiv:1910.07604*, 2019. 2, 3

[18] Viraj Prabhu, Anitha Kannan, Murali Ravuri, Manish Chaplain, David Sontag, and Xavier Amatriain. Few-shot learning for dermatological disease diagnosis. In *Machine Learning for Healthcare Conference*, pages 532–552, 2019. 2

[19] Joaquin Quionero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. The MIT Press, 2009. 1

[20] Andrew Slavin Ross, Michael C Hughes, and Finale Doshi-Velez. Right for the right reasons: Training differentiable models by constraining their explanations. *arXiv preprint arXiv:1703.03717*, 2017. 3

[21] Afsah Saleem, Muhammad Naeem A Bhatti, Muhammad A Ashraf, Muhammad Zia, and Hasan Mehmood. Segmentation and classification of consumer-grade and dermoscopic skin cancer images using hybrid textural analysis. *Journal of Medical Imaging*, 6(3):034501, 2019. 2

[22] Wojciech Samek, Alexander Binder, Grégoire Montavon, Sebastian Lapuschkin, and Klaus-Robert Müller. Evaluating the visualization of what a deep neural network has learned. *IEEE transactions on neural networks and learning systems*, 28(11):2660–2673, 2016. 3

[23] 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. 2, 3

[24] Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813, 2014. 3

[25] Avanti Shrikumar, Peyton Greenside, and Anshul Kundaje. Learning important features through propagating activation differences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3145–3153. JMLR. org, 2017. 2

[26] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013. 2

[27] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017. 3, 4

[28] Thorvald Sørensen. A method of establishing groups of equal amplitude in plant sociology based on similarity of species content and its application to analyses of the vegetation on danish commons. 1948. 4

[29] Charles Spearman. The proof and measurement of association between two things. 1961. 5

[30] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. Striving for simplicity: The all convolutional net. *arXiv preprint arXiv:1412.6806*, 2014. 2

[31] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3319–3328. JMLR. org, 2017. 2, 3

[32] Mukund Sundararajan, Jinhua Xu, Ankur Taly, Rory Sayres, and Amir Najmi. Exploring principled visualizations for deep network attributions. In *IUI Workshops*, 2019. 4

[33] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander A Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 4

[34] Jayaraman J Thiagarajan, Deepta Rajan, and Prasanna Sattigeri. Can deep clinical models handle real-world domain shifts? *arXiv preprint arXiv:1809.07806*, 2018. 1

[35] 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*, 5:180161, 2018. 1, 2

[36] Julia K Winkler, Christine Fink, Ferdinand Toberer, Alexander Enk, Teresa Deinlein, Rainer Hofmann-Wellenhof, Luc Thomas, Aimilios Lallas, Andreas Blum, Wilhelm Stolz, et al. 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, 2019. 2

[37] Tianjun Xiao, Yichong Xu, Kuiyuan Yang, Jiaxing Zhang, Yuxin Peng, and Zheng Zhang. The application of two-level attention models in deep convolutional neural network for fine-grained image classification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 842–850, 2015. 3

[38] Jufeng Yang, Xiaoxiao Sun, Jie Liang, and Paul L Rosin. Clinical skin lesion diagnosis using representations inspired by dermatologist criteria. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1258–1266, 2018. 1, 2

[39] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European conference on computer vision*, pages 818–833. Springer, 2014. 2

[40] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Object detectors emerge in deep scene cnns. *arXiv preprint arXiv:1412.6856*, 2014. 3