Title

Global Saliency: Aggregating Saliency Maps to Assess Dataset Artefact Bias

Permalink

https://escholarship.org/uc/item/12t368pv

Authors

Pfau, Jacob

Young, Albert T

Wei, Maria L

et al.

Publication Date

2019-10-16

DOI

10.48550/arxiv.1910.07604

Peer reviewed

Global Saliency: Aggregating Saliency Maps to Assess Dataset Artefact Bias

Jacob Pfau[1,5], **Albert T. Young**[1,5], **Maria L. Wei**[1,2,5], and **Michael J. Keiser**[3,4,5]

[1]Department of Dermatology, UCSF and Dermatology Service, SFVAMC
[2]Helen Diller Family Comprehensive Cancer Center, UCSF
[3]Institute for Neurodegenerative Diseases, and Bakar Computational Health Sciences Institute, UCSF
[4]Department of Pharmaceutical Chemistry, Department of Bioengineering and Therapeutic Sciences, UCSF
[5]{pfau, keiser}@keiserlab.org, {albert.young, maria.wei}@ucsf.edu

Abstract

In high-stakes applications of machine learning models, interpretability methods provide guarantees that models are right for the right reasons. In medical imaging, saliency maps have become the standard tool for determining whether a neural model has learned relevant robust features, rather than artefactual noise. However, saliency maps are limited to local model explanation because they interpret predictions on an image-by-image basis. We propose aggregating saliency globally, using semantic segmentation masks, to provide quantitative measures of model bias across a dataset. To evaluate global saliency methods, we propose two metrics for quantifying the validity of saliency explanations. We apply the global saliency method to skin lesion diagnosis to determine the effect of artefacts, such as ink, on model bias.

1   Introduction

Across medical imaging tasks, convolutional neural networks (CNN) have demonstrated human-level diagnostic accuracy [7], [10]. However, these models are often not robust, learning non-generalizable patterns and confounding visual artefacts with diseased tissue [2], [16]. Such models may perform well on a validation set yet fail catastrophically when applied in new contexts. The standard method to explain model predictions is to visually, individually inspect saliency maps for each image to determine whether the most salient pixels are the most relevant for diagnosis. This practice prevents some unexpected failures but remains subjective based on the reviewer's judgement. Hence, we expand on methods developed for single images and quantify model bias globally across a validation dataset, thereby facilitating inter-model comparison of bias.

Many saliency methods exist for highlighting the areas of an image most relevant to a model's prediction. Consequently, a natural validity check for saliency methods involves verifying that resulting saliency maps are compatible with known discriminatory features. Recently [1] questioned whether saliency maps explain the predictions of a model or merely highlight the foreground of an image. In response, we propose two metrics for quantifying how accurately saliency maps reflect model behavior.

[14] identified ink skin markings as a confounder for automated diagnosis and evaluated their effect by comparing predictions on inked images to predictions following ink-removal cropping. We apply and evaluate our proposed global saliency method on a skin lesion dataset with ink artefacts.

Our key contributions are as follows:

- We propose a simple procedure, termed *global saliency*, to aggregate saliency maps within and across images. Global saliency quantifies the effect of an artefact on a CNN's decisions across a dataset given segmentation masks for the artefact.
- We introduce two metrics for evaluating the quality of saliency methods: model failure prediction and dataset bias detection.
- We apply global saliency to distinguish between modes of artefact-induced model bias.

2 Background

2.1 Skin Lesion Dataset

Table 1: Training datasets and ink-co-occurrence probabilities by class

Train set	MEL	NV	BCC	SCC	AK	SK	Other
Baseline	58%	48%	74%	68%	76%	58%	66%
Unbiased	66%	66%	66%	66%	66%	66%	66%
Ink-Only	100%	100%	100%	100%	100%	100%	100%
Ablated	100%	100%	100%	100%	100%	100%	100%

The dataset consists of 12,563 clinical skin lesion images labelled by histopathology-verified diagnoses (Appendix A). The labels are: melanoma (MEL), basal cell carcinoma (BCC), squamous cell carcinoma (SCC), actinic keratosis (AK), seborrheic keratosis (SK), nevus (NV) and 'Other', under which remaining less common diagnoses are grouped together. Ink markings, used to designate the location of skin biopsy, occur as a common visual artefact across the dataset (Table 1, Row 1). We used a BiSeNet [15] semantic segmentation network to generate masks labelling the inked pixels within an image; below we denote these pixels by $A(X)$ (Appendix A). All models used in the experiments below are DenseNet-121 (Appendix A).

To address confounding via ink markings, we created three training datasets with different co-occurrence probabilities between ink and label, as shown in Table 1 (also, Appendix A). We compared saliency on ink across the three corresponding models to determine effects of artefact-induced bias on model predictions.

2.2 Saliency Methods

Given a CNN, f, and an input image, X, a *saliency map* is a function, g_f, assigning an importance $g_f(X)_{i,j} \in \mathbb{R}$ to each pixel, (i, j). A saliency map should evaluate the counterfactual importance of a subset of pixels, answering the question, "Would the model's classification change if these pixels were to be replaced by a different set of pixels?" We restrict our attention to two particular saliency maps, grad-CAM and competitive gradient⊙input [9], [12]. *Grad-CAM* visualizes saliency by summing weighted filter maps within a fixed layer where the filter weight is the spatial average of its gradient. *Competitive gradient⊙input* interprets as most salient pixels where the given class has greater absolute value than all other classes' gradients. Both grad-CAM and competitive gradient⊙input compute saliency with respect to a particular class, usually taken to be the model's predicted class on that image.

2.2.1 Aggregating Saliency Across Images and the Completeness Property

As saliency maps are constructed to compare pixels within an image, there is no guarantee that saliency values may be aggregated across images. However, certain saliency maps satisfy the *completeness property*: the pixel-wise sum of saliencies equals the model's confidence for that image, $\sum g_f(X)_{i,j} = \max(f(X))$. [9] suggests that competitive gradient⊙input empirically satisfies the completeness property.

Given a pixel-wise segmentation of an artefact (in our case, ink), we compute saliency on the subset of pixels in which the artefact is present, $A(X) \subset X$. For saliency maps satisfying the completeness

(a) Global saliency using Competitive Grad⊙Input   (b) Global saliency using grad-CAM
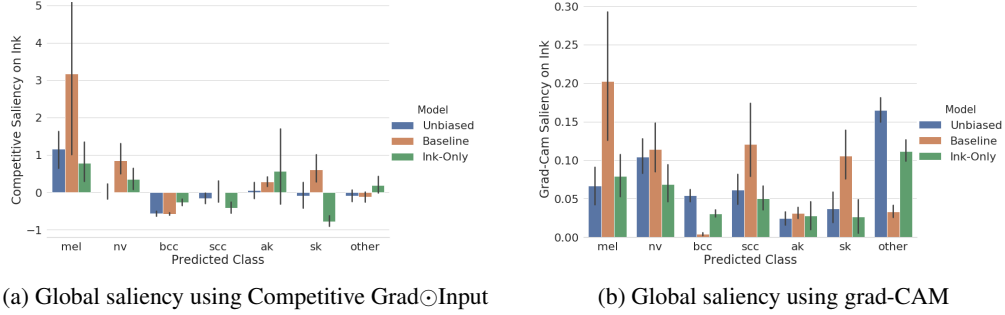
Figure 1: Global mean saliency on validation set, Z-score normalized by dataset. Mean and standard deviation computed over all pixels in all images, by dataset. Error bars are 95% CI.

property, saliency on the artefact is normalized relative to the salience on the rest of the image (Eq. 1). Note that this normalization property does not hold for perturbative saliency methods [13].

$$\max(f(X)) - \sum_{i,j \in A(X)} g_f(X)_{i,j} = \sum_{i,j \in X \setminus A(X)} g_f(X)_{i,j} \tag{1}$$

3 Global Quantification of Saliency

The simplest way to aggregate saliency is to take the mean saliency, m, over the subset of pixels corresponding to the artefact, as defined by a semantic segmentation mask.

$$m_{f,g,A}(X) = \frac{1}{|A(X)|} \sum_{i,j \in A(X)} g_f(X)_{i,j} \tag{2}$$

The values of $m_{f,g,A}(X)$ may be aggregated across a validation set, X_{val}, to quantify the global effect of an artefact on the model which we term *global saliency* (Figure 1). For saliency maps which satisfy completeness, mean aggregation may be used directly. But, for methods which do not satisfy completeness, it is necessary to normalize either within the image or across the dataset. Alternatively, we define an aggregation function invariant to re-scaling of the saliency map – e.g. rank the saliency of pixels within the image, X, and evaluate what percentage of the nth percentile of most salient pixels, $P_n(g_f(X))$, occur within the artefact, $n_{f,g,A}(X) = \frac{|P_n(g_f(X)) \cap A(X)|}{|P_n(g_f(X))|}$.

3.1 Evaluating Global Saliency

[1] showed certain saliency maps that satisfy completeness do not explain differences between trained and untrained models. For medical imaging, it is crucial that methods used to verify robustness of a model consistently detect model failure. We propose two tests for empirically evaluating the quality of saliency maps, using global saliency: (1) tracking dataset bias and (2) model failure prediction.

Tracking Dataset Bias: *Identify the (non-)existence of undesirable bias by inferring whether the underlying training dataset has spurious correlations between an artefact and labels.*

Given a training set with a visual artefact occurring across classes (e.g. Table 1), we propose a validity check for global saliency: Is artefactual global saliency correlated with increasing dataset bias? For our skin lesion dataset, this corresponds to comparing saliency on ink across the baseline, unbiased, and ink-only datasets (Table 1, Appendix A). Figure 1 shows the baseline dataset has higher inter-class ink-saliency variance relative to the unbiased and ink-only datasets – with the exception of the 'Other' class (Figure 1a). We confirm the robustness of this conclusion by comparing Figure 1 to global saliency computed using peak saliency aggregation, $n_{f,g,A}$ instead of mean saliency, $m_{f,g,A}$ (Appendix C). The peak saliency results are visually similar for grad-CAM, but not for competitive saliency. For the baseline model, MEL and possibly NV/SK, consistently show higher saliency on ink than the other classes. Possible drivers of inter-class variance in saliency on ink are discussed in Section 4.

3
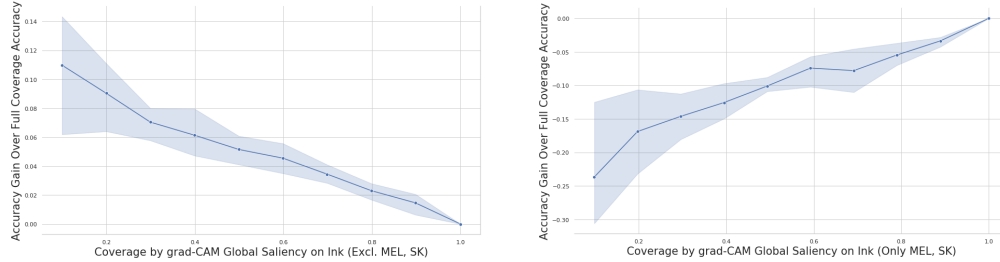
(a) Val excl. MEL+SK. Kendall's τ = -0.889, P<0.0001 (b) Val MEL+SK only. Kendall's τ = 0.806, P<0.0001

Figure 2: 3-Fold Accuracy vs ink-saliency-thresholded subsets of the validation set. Leftmost tick corresponds to the 10th percentile of images with least saliency on ink. Error margins are 95% CI. [1]

Model Failure Prediction: *Predict model error due to artefact, given excessive saliency on that artefact.*

To quantify the relationship between saliency and accuracy, we evaluated the model across validation set subsets, recording model accuracy as a function of the maximum permitted saliency on ink (Figure 2). An ideal map maximizes the area under the response rate accuracy curve (Appendix B).

Figure 1 and Appendix C suggest the model interprets ink as evidence for MEL and SK when trained on biased data. We test this hypothesis by calculating the correlation between model accuracy and ink saliency separately for the MEL+SK subset of the validation set and the rest (Figure 2). For non-MEL+SK skin lesions, excessive saliency on ink correlates with reduced accuracy whereas for MEL+SK, the baseline model spuriously interprets ink as disease-related. This relationship holds across saliency methods and aggregation schemes (Appendix B).

4   Applying Global Saliency to Understand and Correct Dataset Bias

Understanding Dataset Bias: Using global saliency as a measure of model bias, we compare training subsets to understand the ink bias effect on the model. Possible explanations for how ink biases the model include: (1) co-occurrence rates between artefact and class label, (2) variation in appearance between artefact instances, and (3) visual similarity between artefact and class object.

It appears that co-occurrence rates drive bias in our dataset (Figure 1). However, we have not disproven the competing hypothesis that global saliency is merely more sensitive to bias caused by co-occurrence than items 2 and 3. To quantify the effect of intra-artefact variance, we train a model solely on ink, with the lesion and skin blurred out (Appendix A). The lesion-ablated model performs no better than predicting the most prevalent class, with an accuracy of 28%, refuting the hypothesis that variation in ink appearance contributes to model bias. By exclusion, we propose visual similarity between skin pigment and ink as an explanation for the difference in saliency on MEL and SK compared to the other classes, but this needs to be confirmed by future experiments.

Correcting Dataset Bias: Since the co-occurrence unbiased model shows reduced saliency on ink, we propose training using a sampling procedure such that $P(c \mid Ink) = P(c \mid No\text{-}Ink)$ for every class, c. This procedure requires no post-processing of the dataset, and Figure 1 shows that a dataset which respects this property (unbiased) reduces inter-class ink-saliency variance from 0.007 to 0.003; however, this variance reduction does not reach significance with Levene's test $W = 1.04$, $P = 0.33$. It is possible that bias remains present, but is no longer detected by saliency maps. To test our sampling procedure, we evaluate the effect of lesion ablation, and these results agree with Figure 1 (Appendix A). Other solutions to model bias could involve cropping or ablating ink systematically at training time, and these augmentations could be incorporated into an unbiased sampling scheme.

Conclusion Global saliency allows for the quantitative evaluation of artefact-induced bias across models. The global saliency framework may also be used to evaluate the faithfulness of gradient-based saliency maps to model behavior.

---

[1]The results shown in this figure used models trained on an expanded version of the baseline dataset for which the corresponding global saliency figures are shown in Appendix C.

References

[1] J. Adebayo, J. Gilmer, M. Muelly, I. Goodfellow, M. Hardt, and B. Kim, "Sanity checks for saliency maps," in *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, Eds., Curran Associates, Inc., 2018, pp. 9505–9515.

[2] D. Ardila, A. P. Kiraly, S. Bharadwaj, B. Choi, J. J. Reicher, L. Peng, D. Tse, M. Etemadi, W. Ye, G. Corrado, D. P. Naidich, and S. Shetty, "End-to-end lung cancer screening with three-dimensional deep learning on low-dose chest computed tomography," en, *Nat. Med.*, vol. 25, no. 6, pp. 954–961, Jun. 2019.

[3] L. Ballerini, R. B. Fisher, B. Aldridge, and J. Rees, "A color and texture based hierarchical K-NN approach to the classification of non-melanoma skin lesions," in *Color Medical Image Analysis*, vol. 6, Jan. 2013, pp. 63–86.

[4] G. Bradski, "The OpenCV library," *Dr Dobb's J. Software Tools*, vol. 25, pp. 120–125, 2000.

[5] J. Deng, W. Dong, R. Socher, L. J. Li, K. Li, *et al.*, "Imagenet: A large-scale hierarchical image database," *2009 IEEE conference*, 2009.

[6] *DermNet NZ – all about the skin | DermNet NZ*, https://www.dermnetnz.org/, Accessed: 2019-9-14.

[7] V. Dick, C. Sinz, M. Mittlböck, H. Kittler, and P. Tschandl, "Accuracy of Computer-Aided diagnosis of melanoma: A meta-analysis," en, *JAMA Dermatol.*, Jun. 2019.

[8] I. Giotis, N. Molders, S. Land, M. Biehl, and N. Petkov, "MED-NODE: A Computer-Assisted melanoma diagnosis system using Non-Dermoscopic images," *Expert Syst. Appl.*, vol. 42, no. 19, May 2015.

[9] A. Gupta and S. Arora, "A simple saliency method that passes the sanity checks," May 2019. arXiv: 1905.12152 [cs.LG].

[10] G. Hinton, "Deep Learning-A technology with the potential to transform health care," en, *JAMA*, vol. 320, no. 11, pp. 1101–1102, Sep. 2018.

[11] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017.

[12] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-cam: Visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 618–626.

[13] J. Wagner, J. M. Kohler, T. Gindele, L. Hetzel, J. T. Wiedemer, and S. Behnke, "Interpretable and Fine-Grained visual explanations for convolutional neural networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9097–9107.

[14] J. K. Winkler, C. Fink, F. Toberer, A. Enk, T. Deinlein, R. Hofmann-Wellenhof, L. Thomas, A. Lallas, A. Blum, W. Stolz, and H. A. Haenssle, "Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition," en, *JAMA Dermatol.*, Aug. 2019.

[15] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "Bisenet: Bilateral segmentation network for real-time semantic segmentation," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 325–341.

[16] J. R. Zech, M. A. Badgeley, M. Liu, A. B. Costa, J. J. Titano, and E. K. Oermann, "Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study," en, *PLoS Med.*, vol. 15, no. 11, e1002683, Nov. 2018.

A Dataset and Model Details

A.1 Dataset Construction Details

The skin lesion dataset consists of clinical images from [3], [6], [8] and our institution. These datasets were aggregated into a 12,563 image **baseline** meta-dataset that was shuffled and then split 90/10 into train and validation subsets.

Typical examples of ink markings are shown below.

Figure 3: Two typical samples of inked skin lesions from the dataset. From left to right: input image, BiSeNet ink segmentation mask, baseline model grad-CAM saliency overlay.

The **ink-only** dataset was constructed by using only the 8,319 images which BiSeNet predicted to contain more than 100 pixels of ink.

The **unbiased** dataset was constructed by first including the ink-only images. Then for each class, c, we include a randomly chosen subset $c_{uninked}$ of uninked images of that class where $|c_{uninked}| = 0.5 * |c_{inked}|$. The 0.5 ratio was chosen to maximize the dataset size while maintaining the unbiased property.

The **ablated** dataset included the same subset of images used in the ink-only dataset. These images are then ablated by replacing pixels labelled as uninked (by BiSeNet) with the mean value pixel for that image.

As discussed in Section 4, we evaluated both the unbiased and baseline models on the ablated validation dataset. The baseline model predictions remained invariant under lesion ablation for 52% of input images, whereas the unbiased model changed predictions to 'Other' for 79% of input images. The 'Other' class may be taken as a rejection-to-predict class, and as such the behavior of the unbiased model is preferable to that of the baseline model.

A.2 DenseNet Model Details

All experiments and results used a DenseNet-121 architecture pre-trained on the ImageNet dataset [5], [11]. The DenseNet was trained for 50 epochs using SGD with an initial learning rate of 0.02 with a 0.2 decay factor applied following 15 epochs of plateaued performance. Training used standard augmentation techniques (random flipping, rotation, color jitter, and affine transformation) as well as up-sampling of low-prevalence classes to enforce class balance.

For each model we evaluated the performance on both the inked and uninked subsets of the full validation set. In order to compare between artefact-induced bias levels between models, it is desirable to hold constant model performance. Performance on each training dataset was comparable.

Table 2: Validation set macro-mean AUROC by training set [2]

Train set	Full Val	Inked Val	Uninked Val
Baseline	0.757	0.744	0.773
Unbiased	0.746	0.746	0.744
Ink-Only	0.753	0.755	0.763

---

[2]Due to time constraints, some figures in the appendices do not have confidence intervals.

A.3 BiSeNet Model Details

The BiSeNet was trained using a set of 300 openCV [4] rule-based masks selecting ink by hue and intensity. For each of these 300 images, we manually reviewed the quality of three distinct rule-based procedures and selected the mask which appeared most accurate. These 1/0 valued semantic segmentation masks were then augmented by Gaussian blurring the edges of ink regions to allow for uncertainty.

The BiSeNet was trained for 100 epochs optimized by SGD using an initial learning rate of 0.1. Learning rate was decayed by a factor of 0.2 following 12 epochs of plateaued performance. Training used standard augmentation techniques (random flipping, rotation, color jitter, and affine transformation).

Manual review of BiSeNet masks suggests that the BiSeNet performs at human level for 95% of images, but due to time constraints, no human-annotated validation set exists for evaluating the pixel-level performance of the BiSeNet masks.

B Model Failure Prediction and Response Rate Accuracy Curves

An ideal saliency map would demonstrate when the ink within a given image is confounding the model. To quantify to what extent the saliency map succeeds in identifying the degree of confounding, we calculate the area under the response rate accuracy curve (AURRAC). The AURRAC value is equivalent to weighing the model's 1/0 loss on a given image by its percentile rank saliency on ink. A greater AURRAC indicates that the saliency method provides more accurate predictions of model failure due to ink confounding. Given a trained model, for each saliency map and global aggregation scheme we show the corresponding AURRAC. Whereas Figure 2 showed response rate accuracy curves separated into MEL+SK subset and the rest, the following curves show the entire validation set. Due to time constraints, the following curves were computed on the original baseline dataset using one train/test split.
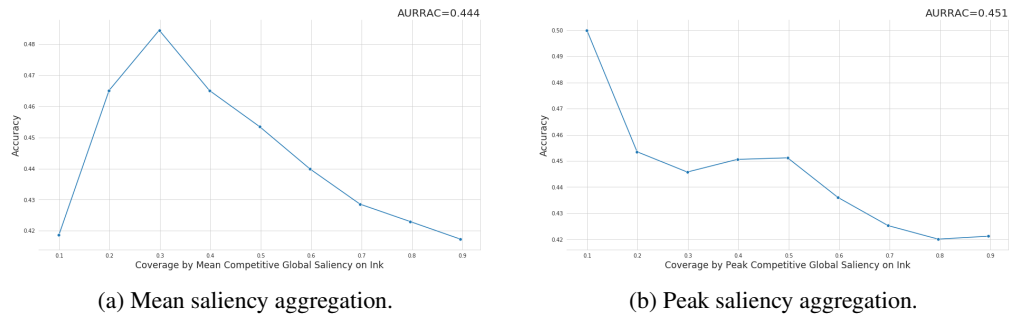

(a) Mean saliency aggregation.

(b) Peak saliency aggregation.

Figure 4: Accuracy vs ink-saliency-thresholded subsets of the validation set. Baseline model with Competitive Grad⊙Input.

(a) Mean saliency aggregation.
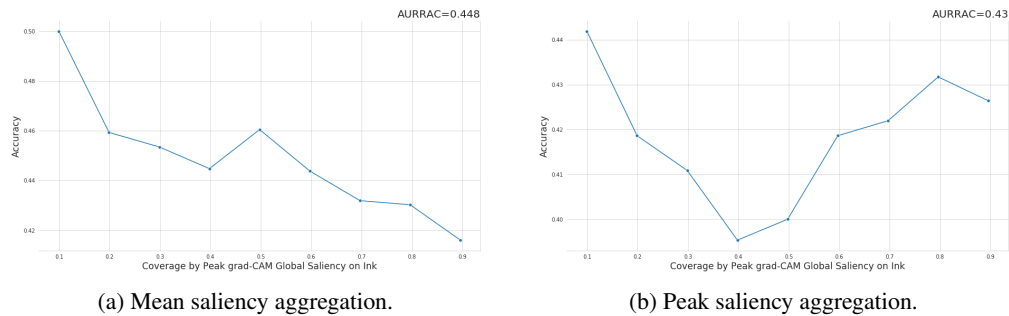
(b) Peak saliency aggregation.

Figure 5: Accuracy vs ink-saliency-thresholded subsets of the validation set. Baseline model with grad-CAM.

These preliminary results show peak competitive saliency achieves maximal AURRAC, thereby predicting baseline model failure most consistently. In contrast, the curves shown in Figure 4a and 5b do not trend downwards indicating that these global saliency methods do not predict model failure.

C   Dataset Bias and Global Saliency by Class

Inter-model comparisons of global saliency using mean aggregation assume that saliency maps take on values on the same scale. For saliency maps satisfying the completeness property, any pair of models which have the same distribution of confidence values across a dataset will also have the same distribution of global saliency values.

We show, empirically, that both grad-CAM and competitive gradient⊙input do not have comparable global saliency distributions across models. In particular, using competitive gradient⊙input, we apply the Wilcoxon signed-rank test to reject the null hypothesis that image-aggregated saliency distributions for the unbiased and ink-only models with $W = 59,967, P < 0.0001$. However, the Wilcoxon test cannot distinguish between the confidence distributions of the unbiased and ink-only models, $W = 183,745, P = 0.67$. These differences in competitive saliency distributions suggest that we must normalize saliency values for each model individually – even when confidence levels are comparable.
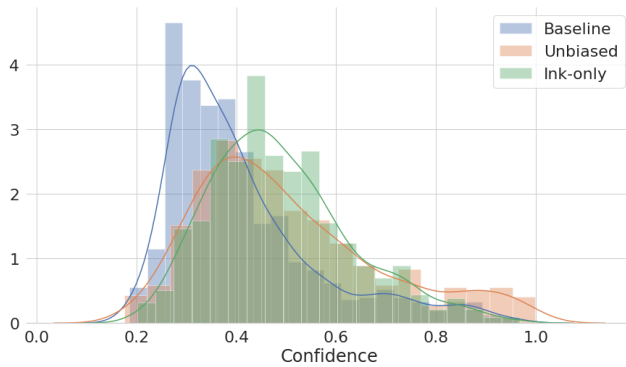

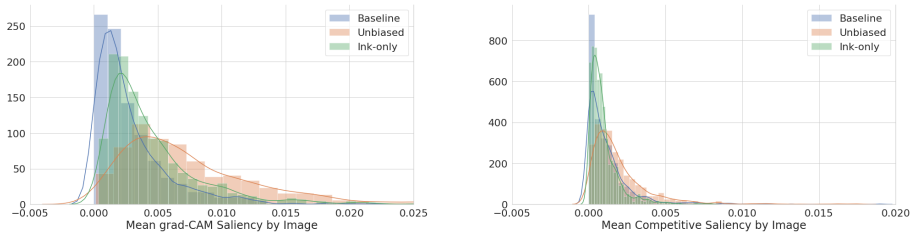Figure 6: Validation set confidence distributions by model

Figure 7: Validation set saliency distributions by model

We hypothesize that global saliency is largely invariant to the choice of aggregation function. Below we plot global saliency on the skin lesion dataset using the peak saliency function defined in Section 3,

$$n_{f,g,C}(X) = \frac{|P_{98}(g_f(X)) \cap A(X)|}{|P_{98}(g_f(X))|} \tag{3}$$

Note that Figure 8a does not track dataset bias, i.e. the baseline model appears to have comparable saliency on ink relative to the unbiased and ink-only models. Further research is required to determine why competitive saliency is not compatible with the $n_{f,g,C}$ function. Grad-CAM saliency appears compatible with the $n_{f,g,C}$ function, successfully tracking dataset bias. Taken together, the results of Figure 1 and Figure 8b suggest that melanoma and nevus consistently show higher saliency on ink in the baseline model (relative to the other classes).

(a) Global saliency using Competitive Grad⊙Input
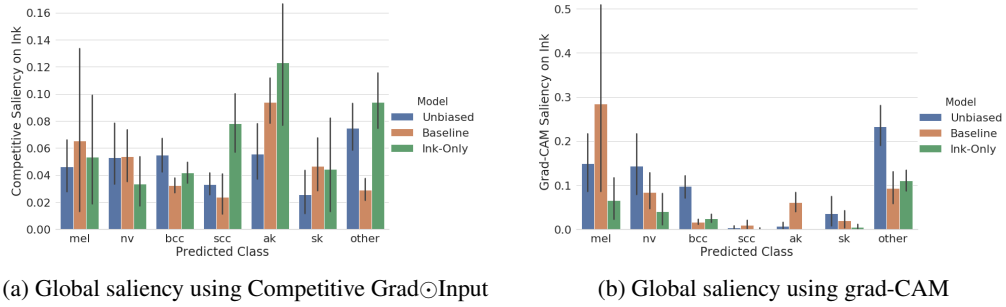
(b) Global saliency using grad-CAM

Figure 8: Global peak saliency on baseline validation set.

Figure 2 used an expanded baseline dataset (including 8,000 more recent, additional images). The global saliency by class remains similar, but the saliency on SK is higher relative to the other classes when compared to the model shown in Figure 1.
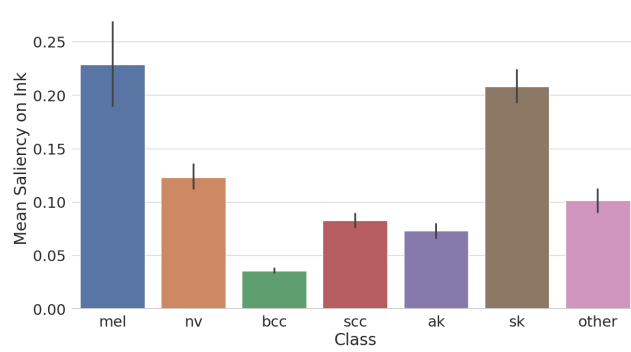

Figure 9: Expanded baseline dataset. Grad-CAM saliency aggregated across 3-folds of cross-validation.