Do You See What I See? A Comparison of Radiologist Eye Gaze to Computer Vision Saliency Maps for Chest X-ray Classification

Jesse Kim [*] [1] Helen Zhou [*] [1] Zachary Lipton [1]

Abstract

We qualitatively and quantitatively compare saliency maps generated from state-of-the-art deep learning chest X-ray classification models to radiologist eye gaze data. We find that across several saliency methods, correct predictions have saliency maps more similar to the corresponding eye gaze data than the same for incorrect predictions. To incorporate eye gaze into the model training process, we create DenseNet-Aug, a simple augmentation of DenseNet which performs comparably to the state-of-the-art. Finally, we extract salient annotated regions for each label class, thereby characterizing model attribution at the dataset level. While sample-level saliency maps visibly vary, these dataset-level regional comparisons indicate that across most class labels, radiologist eye gaze, DenseNet, and DenseNet-Aug often identify similar salient regions.

1. Introduction

Deep learning for automated diagnosis has shown promise in several medical domains including radiology, opthamology, dermatology, and pathology (Pasa et al., 2019; Esteva et al., 2017; Ting et al., 2019; Campanella et al., 2019). Machine learning models can provide relatively cheap and timely predictions of the diagnosis, and, as assessed on i.i.d. holdout sets, have reached performance comparable to clinicians on several tasks of interest (Rajpurkar et al., 2018; Liu et al., 2019; Majkowska et al., 2020). Despite its significant potential benefits, this technology's adoption has met with hesitancy, especially in high-stakes settings. This reticence largely stems from concerns about limited demonstrations of external validity, a well-documented brittleness of models in more realistic non-i.i.d. deployment settings, and potential biases across different demographic groups. These concerns are compounded by a feeling among practitioners and stakeholders that they lack transparency into how the "black-box" model arrives at its prediction.

Without interpretability mechanisms built-in to these "black-boxes," post-hoc explanation methods such as saliency maps have commonly been used to characterize these models beyond their final performance metrics. While saliency maps are an active area of research and many have been shown to fail sanity checks (Adebayo et al., 2018), these methods are hoped to provide insights into (i) how the model reaches its predictions; (ii) potential reliance of models on artifacts arising from dataset creation procedures; and (iii) identifying inputs or regions of potential clinical interest.

In this work, we systematically compare saliency maps to radiologist eye gaze data, and extend a state-of-the-art chest X-ray classification model by incorporating eye gaze data into model training. In addition to saliency maps generated per sample, we identify annotated regions of interest that are salient in predictions for each label class across the entire dataset. Finally, we quantify the similarity between saliency maps and eye gaze data using a structural similarity score, finding that saliency maps for correct predictions tend to be closer to radiologist eye gaze than incorrect predictions.

1.1. Related Work

The recent releases of large public chest X-ray datasets such as MIMIC-CXR (Johnson et al., 2019a), CheXpert (Irvin et al., 2019), ChestX-ray14 (Wang et al., 2017) and Padchest (Bustos et al., 2020) have catalyzed the use of deep learning for chest X-ray classification (Rajpurkar et al., 2017; 2018; Baltruschat et al., 2019; Majkowska et al., 2020; Qin et al., 2019; Seyyed-Kalantari et al., 2020). In some cases, neural networks have been reported to attain performance on par with radiologists (Rajpurkar et al., 2018; Majkowska et al., 2020). To interpret these models, studies have often turned to saliency maps (Wang et al., 2017; Rajpurkar et al., 2017; Baltruschat et al., 2019). To our knowledge, however, there have been no studies systematically comparing radiologist eye gaze data to saliency maps produced from deep learning models for chest X-ray classification. As far as we are aware, the most similar to our work is a contemporaneous

[*] Equal contribution [1] Machine Learning Department, Carnegie Mellon University. Correspondence to: Jesse Kim <jessekim@andrew.cmu.edu>.

Workshop on Interpretable ML in Healthcare at International Conference on Machine Learning (ICML). Copyright 2021 by the author(s).

study which examines the similarity of saliency maps to segmentations provided by radiologists (Saporta et al., 2021), however this work focuses on a single saliency method, performs its evaluation using different metrics, and does not incorporate radiologist eye gaze data.

2. Methods

2.1. Data and Pre-processing

Our study uses two datasets based on the MIMIC-CXR Database v2.0.0 (Johnson et al., 2019a)(Goldberger et al., 2000): (1) MIMIC-CXR-JPG (Johnson et al., 2019b), which compresses images into JPG format and extracts labels from the free-text reports, and (2) Eye Gaze Data for Chest X-rays (Eye-Gaze-CXR) (Karargyris et al., 2021), which contains radiologist eye tracking data for 1,083 frontal chest X-rays from MIMIC-CXR. Compared to MIMIC-CXR, Eye-Gaze-CXR contains a greater proportion of the Lung Opacity, No Finding, and Pneumonia labels, and has much lower support (0-2 examples) for 6 out of the 14 labels in MIMIC-CXR (Table 1). Additionally, Eye-Gaze-CXR includes annotations for four different regions of interest: the aortic knob, right lung, left lung, and mediastanum (Figure 2).

We use the same pre-processing steps as Seyyed-Kalantari et al. (2020), to resize, normalize, and augment the data. Ground-truth radiologist eye gaze heat maps are generated using the same pre-processing as Karargyris et al. (2021), which applies Gaussian smoothing to gaze points and increases the intensity depending on time spent on each point. Figure 1 contains examples of images after pre-processing.

2.2. Saliency Map Methods

We explore six saliency map generation techniques: a simple saliency method which returns the gradient of the output respect the input (SL) (Simonyan et al., 2013), GradCAM (GC) (Selvaraju et al., 2017), DeepLift (DL) (Shrikumar et al., 2017), Layer Conductance (LC) (Dhamdhere et al., 2018), Gradient SHAP (GS) (Lundberg & Lee, 2017), and SmoothGrad (NT) (Smilkov et al., 2017). These methods were chosen for their prominence, usage in prior medical machine learning works, and variety in explanation mechanism (see supplemental material).

2.3. Models: DenseNet and DenseNet-Aug

First, we reproduce state-of-the-art chest X-ray classification on MIMIC-CXR by replicating the DenseNet results from Seyyed-Kalantari et al. (2020)[1] (Figure 1, highlighted in blue). After reproducing the state-of-the-art DenseNet model, we incorporate eye gaze data into the training process by augmenting the architecture with an arm for predicting the radiologist eye gaze heatmaps.[2]

Table 1. Support for each label in MIMIC-CXR-JPG and Eye-Gaze-CXR, given as number of samples (proportion of dataset). Excludes samples with more than one label. CM = Cardiomediastinum, Eff. = Effusion, Dev. = Devices.

Label	MIMIC-CXR-JPG (n = 377,110)	Eye-Gaze-CXR (n = 692)
Atelectasis	65,047 (0.103)	13 (0.019)
Cardiomegaly	64,346 (0.102)	53 (0.077)
Consolidation	14,675 (0.023)	9 (0.013)
Edema	36,564 (0.058)	33 (0.048)
Enlarged CM	10,042 (0.016)	0 (0.0)
Fracture	7,605 (0.012)	1 (0.001)
Lung Lesion	10,801 (0.017)	2 (0.003)
Lung Opacity	76,423 (0.121)	98 (0.142)
No Finding	143,352 (0.226)	379 (0.548)
Pleural Eff.	76,957 (0.121)	23 (0.033)
Pleural Other	3,460 (0.005)	1 (0.001)
Pneumonia	26,222 (0.041)	80 (0.116)
Pneumothorax	14,257 (0.022)	0 (0.0)
Support Dev.	84,073 (0.133)	0 (0.0)

The loss functions L for the DenseNet (DN) model and our augmented model (DenseNet-Aug, DN-Aug) are given below:

$$L_{\text{DN}} = BCE(y, \hat{y})$$

$$L_{\text{DN-Aug}} = \begin{cases} BCE(y, \hat{y}) + \lambda \cdot MSE(\hat{E}, E), & \text{if } E \text{ exists} \\ BCE(y, \hat{y}), & \text{otherwise} \end{cases}$$

Where E and $\hat{E}$ are the true and predicted 2D eye gaze images, y and $\hat{y}$ are the true and predicted 14-label probability vectors, BCE is binary cross entropy, $\lambda = 1000$, and MSE is mean squared error. The model pipeline for both the original DenseNet model and Densenet-Aug are illustrated in Figure 1. See the Supplement for model training details.

2.4. Comparing Model Saliency Maps to Radiologist Eye Gaze Ground Truth

For simplicity of interpretation, we filter out samples with more than one classification. For each class label l (e.g.

[1] This work reported results on MIMIC-CXR v1.0.0 which did not have a pre-specified test set. Since then, v2.0.0 has been released including a new pre-specified test set. Thus, after reproducing results on v1.0.0, we ultimately report results on the standardized v2.0.0 test set for reproducibility. Note that the v2.0.0 test set label distribution is different from that of training (e.g. "No Finding" in 22.8% of train vs. 9.4% of test).

[2] Note that the MIMIC-CXR v2.0.0 training set includes samples from Eye-Gaze-CXR. While DenseNet does not train on the eye gaze data, DenseNet-Aug does. This is a limitation since we analyze saliency maps from *all* samples in Eye-Gaze-CXR due to limited eye gaze data. While we do not explicitly enforce that the saliency maps be similar to eye gaze data, we do provide information about eye gaze to the DenseNet-Aug model in training.

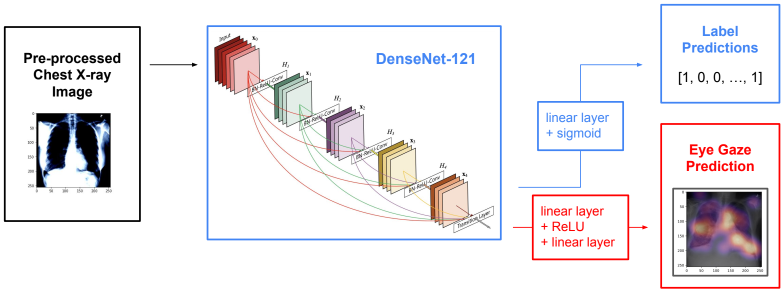


Figure 1. Model architectures for DenseNet (blue components only), and DenseNet-Aug (red component added in) which augments the DenseNet model using limited radiologist eye gaze supervision.

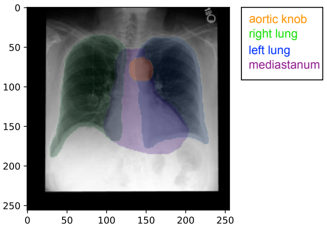


Figure 2. Annotated regions in Eye-Gaze-CXR data. Aortic knob (orange), right lung (green), left lung (blue), medistanum (purple).

pneumonia), we are interested in confident "correct" predictions, defined as samples with true label l and predicted probability among the top k predicted probabilities for l, where k is half of the training label frequency.

Table 2. AUROC on MIMIC-CXR v2.0.0 test set, with 95% confidence intervals derived from bootstrapping with 1,000 replicates. DenseNet refers to the state-of-the-art model described by Seyyed-Kalantari et al. (2020), and DenseNet-Aug refers to our augmented model with eye gaze supervision. CM = Cardiomediastinum, Eff. = Effusion, Dev. = Devices.

Label	DenseNet	DenseNet-Aug
Atelectasis	**0.759** (0.742 – 0.773)	0.751 (0.735 – 0.765)
Cardiomegaly	**0.788** (0.774 – 0.801)	0.778 (0.766 – 0.792)
Consolidation	0.745 (0.719 – 0.770)	**0.749** (0.722 – 0.776)
Edema	**0.835** (0.821 – 0.848)	0.833 (0.821 – 0.846)
Enlarged CM	0.719 (0.682 – 0.753)	**0.722** (0.685 – 0.758)
Fracture	0.676 (0.633 – 0.717)	**0.680** (0.637 – 0.724)
Lung Lesion	**0.737** (0.702 – 0.768)	0.727 (0.692 – 0.764)
Lung Opacity	0.694 (0.678 – 0.709)	**0.697** (0.681 – 0.712)
No Finding	0.793 (0.776 – 0.808)	**0.803** (0.788 – 0.817)
Pleural Eff.	**0.888** (0.879 – 0.898)	0.884 (0.874 – 0.893)
Pleural Other	0.843 (0.81 – 0.874)	**0.851** (0.824 – 0.877)
Pneumonia	0.711 (0.687 – 0.734)	**0.713** (0.689 – 0.735)
Pneumothorax	**0.832** (0.798 – 0.865)	0.816 (0.781 – 0.850)
Support Dev.	0.885 (0.875 – 0.895)	0.885 (0.875 – 0.894)

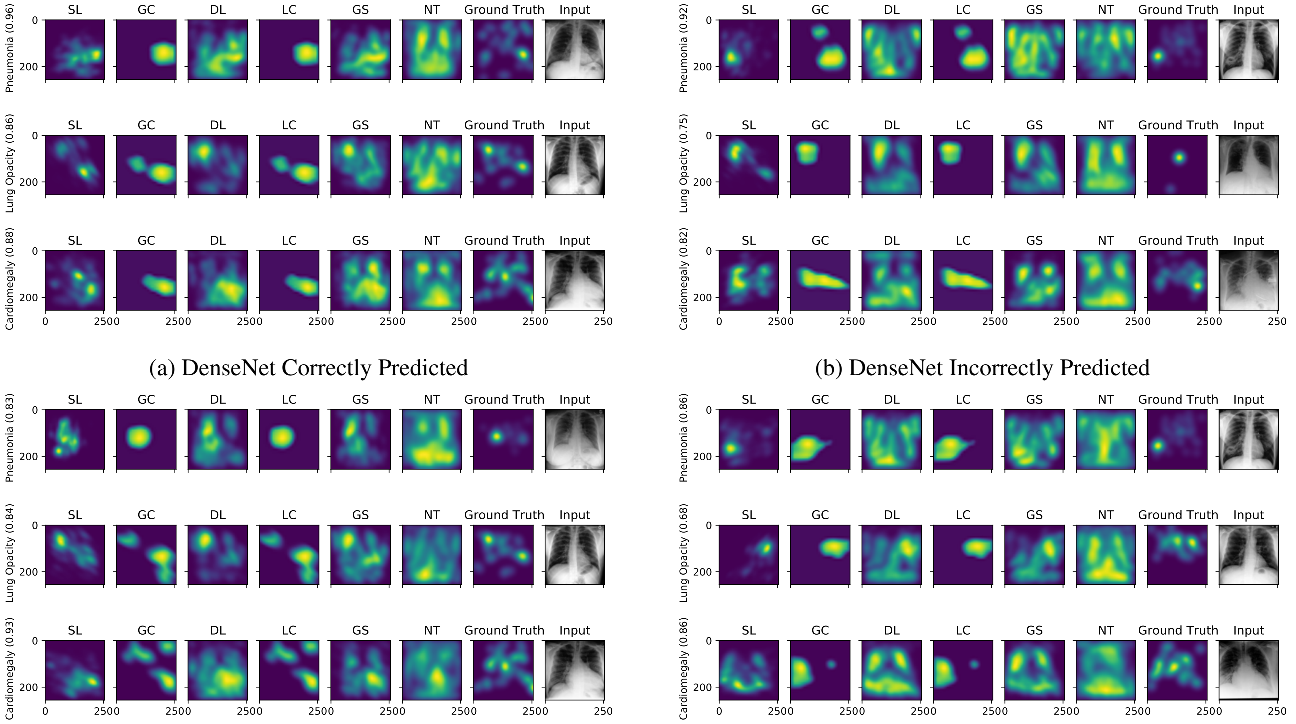
Qualitative Comparison Across six saliency methods and the three label classes with highest support in Eye-Gaze-CXR, we extract the saliency maps and corresponding ground truth for the top 1 most confident "correct" predictions as well as the most confident incorrect predictions.

Quantitative Region-level Comparison To identify the most salient regions (Figure 2) according to radiologist eye gaze and our models, for each sample we compute a z-score for the proportion of total saliency within each region. The region with the highest z-score relative to other images is considered the most salient area "predicted" by the model. Grouping samples by their class label, we can then plot distributions over most salient regions in each class.

Quantitative Image-level Comparison Across all class labels with support ≥ 10 samples and six saliency methods, we compute the average structural similarity index measure (SSIM) score between saliency maps and radiologist eye gaze data among correct and incorrect predictions. The SSIM score (see Supplement for full description) is used as the primary metric to quantitatively compare radiologist eye gaze maps to model saliency maps. In contrast to the mean squared error (MSE) metric, the SSIM score, which is computed by aggregating over sliding windows, is more robust to small geometrical changes, and takes luminance, contrast, and structures into account (Wang & Bovik, 2009).

3. Results

Overall, DenseNet and DenseNet-Aug achieve comparable test AUROCs (Table 2). Figure 3 displays model saliency maps with radiologist eye gaze heatmaps for the three diseases with the most support in Eye-Gaze-CXR. For pneumonia and lung opacity (which have highest support), the most salient points seem surprisingly consistent with radiologist eye gaze for all tested saliency methods except NT. In terms of annotated regions of interest, the most salient regions in each label class seem fairly consistent across radiologist eye gaze, DenseNet, and DenseNet-Aug (Figure 4). On average, the SSIM scores for radiologist eye gaze compared to saliency maps are higher for correct predictions than incorrect predictions (Tables 3 and 4).



(a) DenseNet Correctly Predicted

(b) DenseNet Incorrectly Predicted

(c) DenseNet-Aug Correctly Predicted

(d) DenseNet-Aug Incorrectly Predicted

Figure 3. Saliency maps of the most confident predictions per the three classes (excluding No Finding) with highest support (Pneumonia, Lung Opacity, and Cardiomegaly), compared to ground truth eye gaze data. Top: Densenet; Bottom: Densenet-Aug; Left: correct predictions, with true class on the y-axis; Right: incorrect predictions, with the (incorrect) predicted label on the y-axis.

Table 3. Average SSIM scores comparing DenseNet saliency maps from "correct" samples (✓, high predicted probability for true label) to radiologist eye gaze, and saliency maps from incorrect samples (✗, high probability for a class different from the true label) to radiologist eye gaze. True labels are across the top, and saliency methods are in the first column.

	Atelectasis		Cardiomegaly		Edema		Lung Opacity		No Finding		Pleural Effusion		Pneumonia	
	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗
SL	0.427	0.386	0.398	0.373	0.467	0.490	0.450	0.442	0.387	0.388	0.378	0.359	0.465	0.435
GC	0.363	0.285	0.351	0.353	0.383	0.397	0.362	0.332	0.314	0.317	0.417	0.394	0.402	0.381
DL	0.388	0.370	0.407	0.401	0.417	0.457	0.402	0.396	0.362	0.361	0.372	0.373	0.399	0.403
LC	0.363	0.285	0.351	0.353	0.383	0.397	0.362	0.332	0.314	0.317	0.417	0.394	0.402	0.381
GS	0.391	0.410	0.441	0.420	0.458	0.486	0.450	0.451	0.407	0.406	0.413	0.436	0.423	0.419
NT	0.397	0.435	0.443	0.428	0.433	0.459	0.406	0.406	0.399	0.398	0.456	0.464	0.358	0.371

Table 4. Average SSIM scores comparing DenseNet-Aug saliency maps from "correct" samples (✓, high predicted probability for true label) to radiologist eye gaze, and saliency maps from incorrect samples (✗, high probability for a class different from the true label) to radiologist eye gaze. True labels are across the top, and saliency methods are in the first column.

	Atelectasis		Cardiomegaly		Edema		Lung Opacity		No Finding		Pleural Effusion		Pneumonia	
	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗	✓	✗
SL	0.489	0.361	0.373	0.341	0.423	0.462	0.465	0.422	0.370	0.371	0.395	0.361	0.485	0.415
GC	0.352	0.322	0.281	0.266	0.362	0.381	0.366	0.336	0.319	0.320	0.341	0.348	0.428	0.381
DL	0.433	0.372	0.400	0.375	0.417	0.439	0.415	0.419	0.356	0.355	0.420	0.401	0.421	0.412
LC	0.352	0.322	0.281	0.266	0.362	0.381	0.366	0.336	0.319	0.320	0.341	0.348	0.428	0.381
GS	0.427	0.372	0.441	0.414	0.423	0.450	0.432	0.443	0.404	0.405	0.446	0.410	0.434	0.432
NT	0.406	0.441	0.453	0.450	0.421	0.435	0.391	0.412	0.388	0.389	0.435	0.447	0.335	0.383

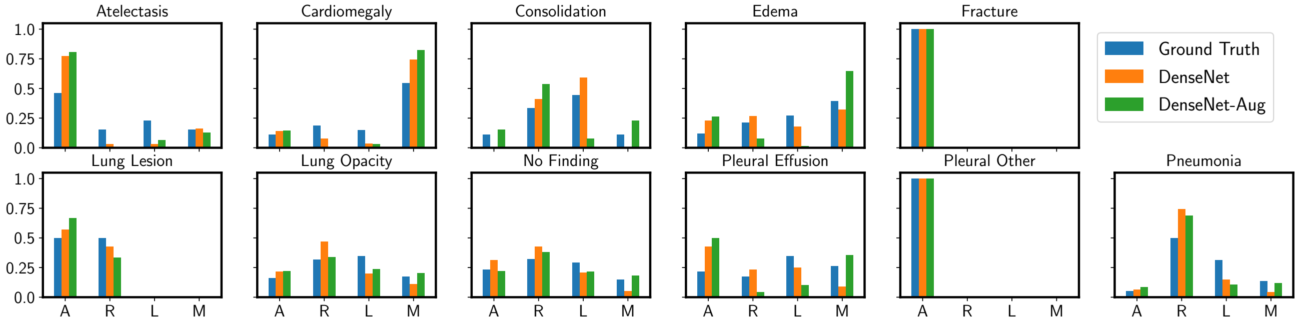


Figure 4. For each class, the proportion of samples with high attribution to each annotated region (A = aortic knob, R = right lung, L = left lung, M = mediastanum), according to the ground truth radiologist eye gaze, Densenet, and DenseNet-Aug. Proportions for Densenet and DenseNet-Aug are computed among "correct" predictions for each class. Annotated regions are shown in Figure 2.

4. Discussion

Due to the limited size of Eye-Gaze-CXR, it was unsurprising that that DenseNet and DenseNet-Aug had similar test AUROCs. To better understand the potential benefits of eye gaze data for improving predictions, future work includes collecting more eye gaze data as well as developing models which enforce a stronger prior on examples without eye gaze data, perhaps utilizing attention mechanisms to do so.

For both DenseNet and DenseNet-Aug, variation between saliency maps generated by different methods is visibly present, but most saliency maps for correct predictions tend to focus on the same areas as the ground truth radiologist eye gaze (Figure 3). GC and LC appear similar, as they are (last) layer attribution methods which require upsampling to produce coarser saliency maps. NT tends to have more spread out saliency maps, likely due to its generative process which includes adding sampled Gaussian noise to the input image and averaging across samplings. Interestingly, as shown in Figure 3, when DenseNet and DenseNet-Aug are confidently incorrect, some of their saliency maps bear a resemblance to the ground truth eye gaze data (especially DenseNet-Aug, possibly due to the auxiliary eye gaze data used to supervise DenseNet-Aug). One potential reason for this similarity could be that certain regions are distinctive and important for several possible labels. Another reason might be that the saliency method is invariant to how the model makes its predictions (Adebayo et al., 2018), a troubling property which could be sanity-checked for.

In the interpretation in terms of annotated regions (Figure 4), each sample is only counted towards its most salient region rather than taking into account the distribution over regions. While this is done for ease of interpretation, it could skew the resulting counts towards a mode and not adequately represent e.g. the second most salient region. Nevertheless, regions of highest saliency tended to be consistent with radiologist eye gaze data except for lung opacity, lung lesion, and pleural effusion. The utilization of annotated regions reduced the reliance on sample-level explanations, allowing us to characterize the nature of the model across samples.

In Tables 3 and 4, excluding Edema, NT, and No Finding (where scores are similar), both DenseNet and DenseNet-Aug's correctly predicted samples have saliency maps with higher similarity to eye gaze data than incorrectly predicted samples. This suggests potential spatial attribution information shared between saliency maps and radiologist eye gaze data, and serves as motivation for future work incorporating eye gaze data to improve predictions.

Overall, while it is inconclusive from our preliminary experiments on DenseNet and DenseNet-Aug whether incorporation of radiologist eye gaze data can improve predictive performance, we quantitatively and qualitatively compare several model saliency maps to human eye gaze data. We find that despite significant variation between the maps generated by different saliency methods, the most salient regions according to these saliency methods are often the same as those most salient in radiologist eye gaze data. Additionally, the saliency maps of correctly predicted samples tend to have higher structural similarity to radiologist eye gaze those of incorrectly predicted samples.

More broadly, however, there are several considerations for whether it is desirable for saliency maps to mimic human gaze. While human gaze can reflect key information for decision-making, eye gaze data can suffer from a central bias (Le Meur & Baccino, 2013). Furthermore, saliency maps themselves might not fully reflect "concepts" important to the model's prediction. It also is possible that models may pick up on non-intuitive characteristics that are indeed helpful for prediction. At the same time, humans might be pre-disposed to pay attention to causal features and concepts, which could be more robust to distribution shift. Ultimately, our work highlights similarity to human gaze as another lens through which models can be viewed.

References

Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., and Kim, B. Sanity checks for saliency maps. *arXiv preprint arXiv:1810.03292*, 2018.

Baltruschat, I. M., Nickisch, H., Grass, M., Knopp, T., and Saalbach, A. Comparison of deep learning approaches for multi-label chest x-ray classification. *Scientific reports*, 9 (1):1–10, 2019.

Bustos, A., Pertusa, A., Salinas, J.-M., and de la Iglesia-Vayá, M. Padchest: A large chest x-ray image dataset with multi-label annotated reports. *Medical image analysis*, 66:101797, 2020.

Campanella, G., Hanna, M. G., Geneslaw, L., Miraflor, A., Silva, V. W. K., Busam, K. J., Brogi, E., Reuter, V. E., Klimstra, D. S., and Fuchs, T. J. Clinical-grade computational pathology using weakly supervised deep learning on whole slide images. *Nature medicine*, 25(8): 1301–1309, 2019.

Dhamdhere, K., Sundararajan, M., and Yan, Q. How important is a neuron? *arXiv preprint arXiv:1805.12233*, 2018.

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., and Thrun, S. Dermatologist-level classification of skin cancer with deep neural networks. *nature*, 542(7639):115–118, 2017.

Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P., Mark, R., and Stanley, H. Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. pp. e215–e220, 2000. Circulation [Online] 101 (23).

Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, B., Ball, R., Shpanskaya, K., et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 590–597, 2019.

Johnson, A. E., Pollard, T. J., Berkowitz, S. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Mark, R. G., and Horng, S. Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):1–8, 2019a.

Johnson, A. E., Pollard, T. J., Greenbaum, N. R., Lungren, M. P., Deng, C.-y., Peng, Y., Lu, Z., Mark, R. G., Berkowitz, S. J., and Horng, S. Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. *arXiv preprint arXiv:1901.07042*, 2019b.

Karargyris, A., Kashyap, S., Lourentzou, I., Wu, J. T., Sharma, A., Tong, M., Abedin, S., Beymer, D., Mukherjee, V., Krupinski, E. A., et al. Creation and validation of a chest x-ray dataset with eye-tracking and report dictation for ai development. *Scientific data*, 8(1):1–18, 2021.

Le Meur, O. and Baccino, T. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior research methods*, 45(1):251–266, 2013.

Liu, X., Faes, L., Kale, A. U., Wagner, S. K., Fu, D. J., Bruynseels, A., Mahendiran, T., Moraes, G., Shamdas, M., Kern, C., et al. A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: a systematic review and meta-analysis. *The lancet digital health*, 1(6):e271–e297, 2019.

Lundberg, S. and Lee, S.-I. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*, 2017.

Majkowska, A., Mittal, S., Steiner, D. F., Reicher, J. J., McKinney, S. M., Duggan, G. E., Eswaran, K., Cameron Chen, P.-H., Liu, Y., Kalidindi, S. R., et al. Chest radiograph interpretation with deep learning models: assessment with radiologist-adjudicated reference standards and population-adjusted evaluation. *Radiology*, 294(2): 421–431, 2020.

Pasa, F., Golkov, V., Pfeiffer, F., Cremers, D., and Pfeiffer, D. Efficient deep network architectures for fast chest x-ray tuberculosis screening and visualization. *Scientific reports*, 9(1):1–9, 2019.

Qin, Z. Z., Sander, M. S., Rai, B., Titahong, C. N., Sudrungrot, S., Laah, S. N., Adhikari, L. M., Carter, E. J., Puri, L., Codlin, A. J., et al. Using artificial intelligence to read chest radiographs for tuberculosis detection: A multi-site evaluation of the diagnostic accuracy of three deep learning systems. *Scientific reports*, 9(1):1–10, 2019.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., and Ng, A. Y. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning, 2017.

Rajpurkar, P., Irvin, J., Ball, R. L., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C. P., et al. Deep learning for chest radiograph diagnosis: A retrospective comparison of the chexnext algorithm to practicing radiologists. *PLoS medicine*, 15(11):e1002686, 2018.

Saporta, A., Gui, X., Agrawal, A., Pareek, A., Truong, S. Q., Nguyen, C. D., Ngo, V.-D., Seekins, J., Blankenberg,

F. G., Ng, A. Y., Lungren, M. P., and Rajpurkar, P. Deep learning saliency maps do not accurately highlight diagnostically relevant regions for medical image interpretation. *medRxiv*, 2021. doi: 10.1101/2021.02.28.21252634.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, 2017.

Seyyed-Kalantari, L., Liu, G., McDermott, M., Chen, I. Y., and Ghassemi, M. Chexclusion: Fairness gaps in deep chest x-ray classifiers. In *BIOCOMPUTING 2021: Proceedings of the Pacific Symposium*, pp. 232–243. World Scientific, 2020.

Shrikumar, A., Greenside, P., and Kundaje, A. Learning important features through propagating activation differences. In *International Conference on Machine Learning*, pp. 3145–3153. PMLR, 2017.

Simonyan, K., Vedaldi, A., and Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

Smilkov, D., Thorat, N., Kim, B., Viégas, F., and Wattenberg, M. Smoothgrad: removing noise by adding noise. *arXiv preprint arXiv:1706.03825*, 2017.

Ting, D. S. W., Pasquale, L. R., Peng, L., Campbell, J. P., Lee, A. Y., Raman, R., Tan, G. S. W., Schmetterer, L., Keane, P. A., and Wong, T. Y. Artificial intelligence and deep learning in ophthalmology. *British Journal of Ophthalmology*, 103(2):167–175, 2019.

Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., and Summers, R. Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In *IEEE CVPR*, 2017.

Wang, Z. and Bovik, A. C. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE signal processing magazine*, 26(1):98–117, 2009.