2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV) | 978-1-6654-0915-5/22/$31.00 ©2022 IEEE | DOI: 10.1109/WACV51458.2022.00161

X-MIR: EXplainable Medical Image Retrieval

Brian Hu, Bhavan Vasu, and Anthony Hoogs
Kitware, Inc.
Clifton Park, NY
{brian.hu, bhavan.vasu, anthony.hoogs}@kitware.com

Abstract

Despite significant progress in the past few years, machine learning systems are still often viewed as "black boxes," which lack the ability to explain their output decisions. In high-stakes situations such as healthcare, there is a need for explainable AI (XAI) tools that can help open up this black box. In contrast to approaches which largely tackle classification problems in the medical imaging domain, we address the less-studied problem of explainable image retrieval. We test our approach on a COVID-19 chest X-ray dataset and the ISIC 2017 skin lesion dataset, showing that saliency maps help reveal the image features used by models to determine image similarity. We evaluated three different saliency algorithms, which were either occlusion-based, attention-based, or relied on a form of activation mapping. We also develop quantitative evaluation metrics that allow us to go beyond simple qualitative comparisons of the different saliency algorithms. Our results have the potential to aid clinicians when viewing medical images and addresses an urgent need for interventional tools in response to COVID-19. The source code is publicly available at: https://gitlab.kitware.com/brianhhu/x-mir.

1. Introduction

Machine learning has made significant progress in the past few years, particularly in the area of deep learning, with increasing adoption in the medical imaging domain [37, 54, 20]. Deep learning has the potential to help human experts in the interpretation of medical images, a process which can be time-consuming, expensive, and prone to errors due to visual fatigue. Despite the success of deep learning systems, their "black box" lack of interpretability is a serious barrier for use in high-stakes situations such as healthcare, criminal justice, and autonomous driving [25, 17, 50]. For computer vision models, various saliency algorithms have been proposed as forms of explainable AI (XAI) which can highlight regions of the input image responsible for the model's output decision [68, 47, 53, 22]. A classic example is when saliency maps helped reveal that an algorithm used stray radiologist scribbles instead of lung regions to detect certain diseases in chest X-rays. Even so, extreme care must be taken in the interpretation of saliency maps, as several studies have questioned the usefulness of some techniques [31, 2].

Here, we focus on the problem of medical image retrieval (Figure 1). Image retrieval refers to finding similar images in a large image archive given only a query image. For example, this might correspond to finding relevant medical images that contain the same disease pathology. The retrieved images can facilitate case-based reasoning and discovery of underlying patterns in the data. We first test our approach on a COVID-19 chest X-ray dataset [61]. COVID-19 can cause progressive respiratory failure in patients, leading to hospitalization and potential death [14]. While reverse transcriptase-polymerase chain reaction testing is considered the gold standard for diagnosing COVID-19 [62], early detection via radiography examination may help prevent the spread of the disease and lead to better patient outcomes. In addition to COVID-19, we also apply our approach to a skin lesion dataset which contains examples of benign cases and cancerous melanoma [12]. This is also an important problem, as the most prevalent form of cancer in the United States is skin cancer, resulting in over 9,000 deaths a year [55]. As a result, computer-aided dermascopy has the potential to help with early detection of melanoma and improve patient outcomes.

Critically, an end user may not understand why a given image is considered relevant and returned by the retrieval system, motivating the need for explanations. These explanations can be used to build model understanding and trust. Research on visual explanations in the form of saliency maps has largely focused on image classification problems. Here, we adapt saliency for image retrieval through a similarity-based rather than classification-based formulation. This straightforward change enables a user to see which parts of a result image match the query image, *i.e.* what an algorithm pays attention to during the image retrieval process. Despite the fact that multiple similarity-

978-1-6654-0915-5/21/$31.00 ©2021 IEEE
DOI 10.1109/WACV51458.2022.00161

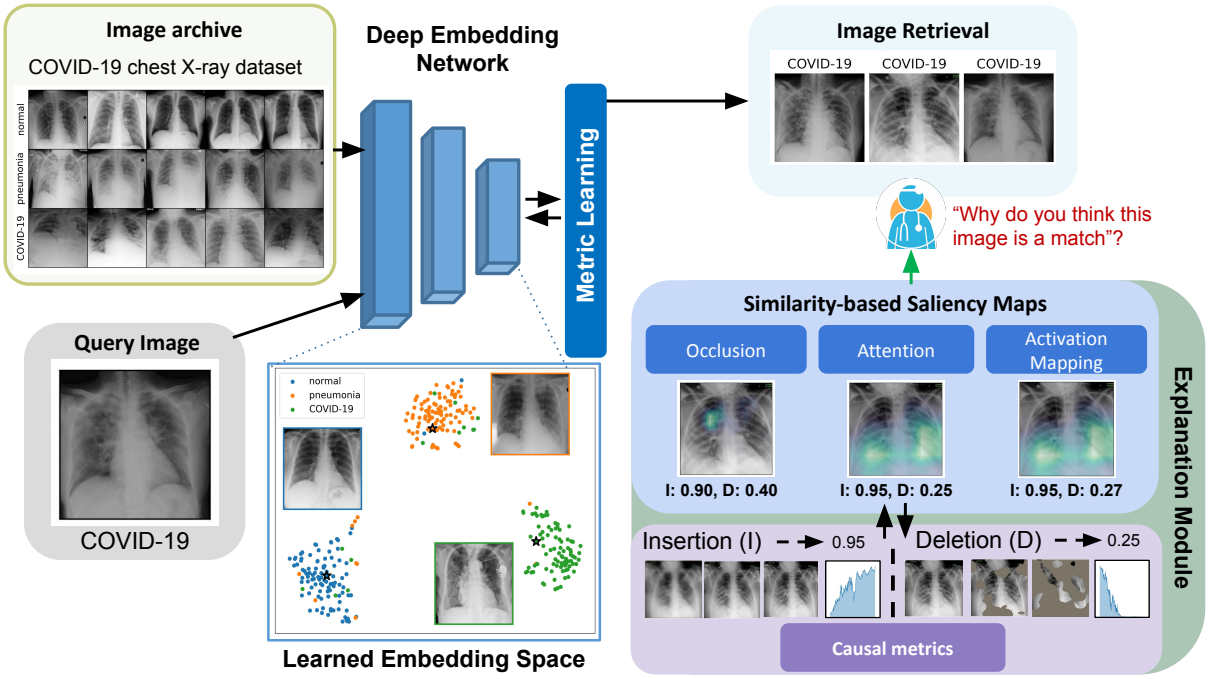

Figure 1. Using deep metric learning, we train a neural network on a large image archive (shown here as the COVID-19 chest X-ray dataset). The learned embedding forces different classes to cluster into different regions of the latent space, shown with a t-SNE visualization. For each class, we show the closest image in the dataset to the class centroid (marked with stars). We use the trained model for image retrieval (light blue box) and leverage explainable AI (XAI) to reason about the model's decisions. We use multiple types of similarity-based saliency maps to highlight regions in a result image which are most similar to the query image (blue boxes). To quantitatively evaluate the generated saliency maps, we use the causal metrics of insertion and deletion (purple box).

based saliency algorithms have been proposed [16, 70, 57, 71], there is a lack of existing benchmarks and quantitative metrics to compare these different methods.

In our paper, we make the following contributions:
(1) We develop a benchmark for image retrieval on two different publicly available medical imaging datasets embodying different problems.
(2) We apply similarity-based saliency maps to the medical imaging domain, providing visual explanations for deep metric learning models trained on medical images.
(3) We adapt a set of causal metrics to image similarity models to quantitatively evaluate different saliency algorithms on image retrieval.
(4) We show a form of self-similarity and differential saliency which can highlight regions of images responsible for different disease conditions.

2. Related Work

2.1. Medical imaging with deep learning

Deep learning has shown remarkable success on image recognition tasks, which largely involve natural image datasets such as ImageNet [51]. Deep learning is now also increasingly being used in the medical imaging domain, with applications to radiology [33, 26], dermatology [19], opthalmology [15], and pathology [8]. With increasing amounts of data, deep learning allows computers to automatically learn patterns in the data that can be useful for prediction and diagnosis. Developing deep learning models for healthcare has also become easier, taking advantage of techniques such as transfer learning [10, 46].

For the specific problem of COVID-19 detection, both chest X-rays (CXR) [61, 4] and computed tomography (CT) [69, 23] have emerged as leading candidates for early detection. While CT can produce 3D volumetric data with greater image resolution, CXR also provides many advantages, including rapid triaging, availability and accessibility, and portability [61]. Numerous open-source COVID-19 datasets with crowd-sourced images or images scraped from medical papers online are now available [13, 59]. However, care should be taken when using these datasets, as several biases in the dataset have been found which can impact study conclusions [38, 58, 48]. This is a good use case of XAI which may help identify situations when the model is

actually "right for the right reasons" [34].

For automated skin lesion classification, the application of deep convolutional neural networks has already shown great success [19, 56]. Using an Inceptionv3 architecture pretrained on ImageNet and a large annotated dataset of clinical images, [19] demonstrated dermatologist-level performance on melanoma classification. Recently, more attention has been paid to deploying skin lesion classification models on mobile devices, which can aid diagnosis in resource-limited environments [56]. Despite these successes, a recent study has shown that deep neural networks may actually rely on the presence of surgical skin markers in dermascopic images to make their predictions [65]. This is yet again another example of how XAI may help identify unintended biases when the model is displaying "Clever Hans"-like abilities [34].

Traditionally, medical image retrieval involved matching local features computed densely across images [36, 39]. More recently, deep neural networks have emerged as strong baselines for generating semantically rich features at multiple levels [45, 3]. Recent work has also studied universal image retrieval from multiple domains [21] and tools for human-machine teaming on image retrieval tasks [24, 7].

2.2. Explainable artificial intelligence (XAI)

XAI is the field of machine learning that tries to make deep learning models more interpretable [52, 1, 5, 60]. Although several different taxonomies of explanation methods have been proposed, explanations generally fall into different categories based on their scope and mechanism. Local explanations provide interpretations of individual data points (*e.g*. images), while global explanations try to summarize models at the dataset level. Explanations can either be white-box or black-box, depending on the amount of access to the model the explanation requires. Black-box methods are model agnostic and can be applied more generally, while white-box methods often require the computation of model gradients. As an alternative to post-hoc explanation methods, models can be made to be interpretable in the first place [17, 50]. Along these lines, several methods have tried to learn "prototypes," or representative examples that capture information about the underlying data distribution. Recent techniques include prototypical part networks [10] and concept bottleneck networks [32], where models are made more interpretable via specialized loss functions.

2.3. Saliency maps as visual explanations

We focus specifically on the use of visual explanations in the form of saliency maps, which attempt to provide insight into which image regions a model uses to arrive at its output prediction. Most similar to our work, saliency maps have been used to visualize the image features used by COVID-19 classification models [11] or joint classification and segmentation models [66] on chest CT images. While most XAI techniques involving saliency have been developed for classification tasks [68, 47, 53, 22], there has been an increasing push to create explanations for other image understanding tasks, including object detection [43] and image similarity [16, 57, 71, 64, 18, 9]. Dong *et al*. [16] proposed a black-box method for computing similarity-based saliency maps using occlusion. In contrast, methods based on similarity activation mapping use the last convolutional feature map before pooling to compute point-wise similarity maps [57, 71]. Zheng *et al*. [70] extended Grad-CAM to image similarity models by using a triplet-like loss, showing that this form of similarity attention can also be used to regularize models during training. Similarly, Chen *et al*. [9] proposed a more efficient extension of this method using weight-transfer. Other works use layer-wise relevance propagation for image similarity [18] or incorporate explainability for face recognition and matching tasks [64].

3. Methods

The core of our approach is comprised of deep metric learning, image retrieval, and saliency map computation (Figure 1). We describe these in the following sections, but first we introduce the two datasets used in order to provide examples and motivation for the approach.

3.1. COVID-19 chest X-ray dataset

We use the COVIDx dataset [61], a publicly available COVID-19 chest X-ray dataset. The dataset consists of images and other associated metadata, including patients' ages, gender, hospitalization status, etc. The dataset contains approximately 14,000 training images across three classes: normal, pneumonia, and COVID-19 cases. We ignore the fine-grained differences between different causes of pneumonia (*e.g*. bacterial vs. viral). The dataset is highly unbalanced, with about only 500 COVID-19 cases. Example images from the dataset are shown in Figure 2. For ease of comparison, we use the same train and test splits as the original authors.

3.2. ISIC 2017 skin lesion dataset

The International Skin Imaging Collaboration (ISIC) releases data every year for machine learning challenges centered around skin lesion analysis [12]. These challenges focus on skin lesion segmentation, feature extraction, and classification. We use data from the 2017 version of the ISIC challenge, focused specifically on lesion classification [12]. Images belong to one of three categories: benign nevi, sebborheic keratosis, and melanoma (Figure 2). The first two conditions are benign, while the last is cancerous. There are a total of 2000 training images in the dataset, which is again highly unbalanced with fewer cases of keratosis and melanoma. To create a balanced test set, we ran-

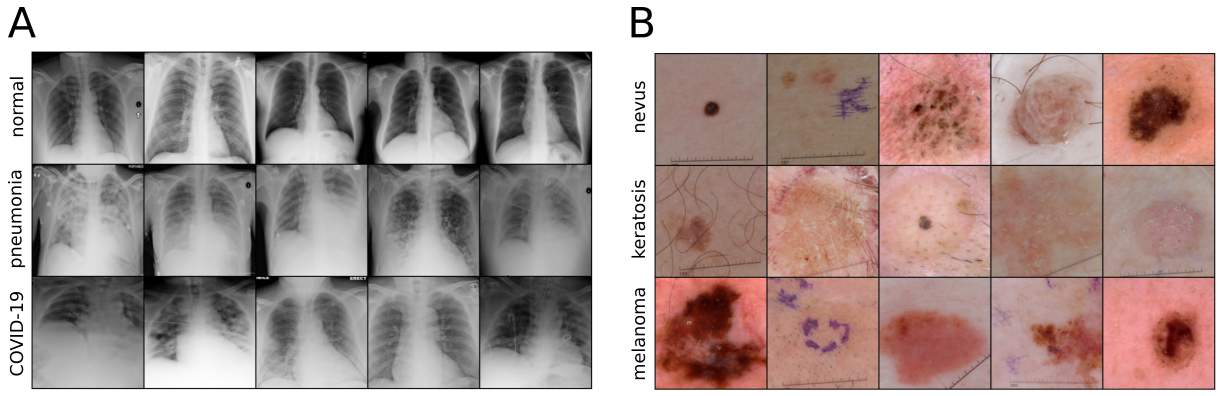

Figure 2. (A) Example images from the COVID-19 chest X-ray dataset which includes normal (top row), pneumonia (middle row), and COVID-19 (bottom row). (B) Example images from the ISIC 2017 skin lesion dataset which includes benign nevi (top row), seborrheic keratosis (middle row), and melanoma (bottom row).

domly subselected 90 examples of nevi and melanoma to match the number of available keratosis test examples.

3.3. Deep metric learning and image retrieval

We use a deep metric learning framework, which is a supervised learning approach where the model learns to embed images of the same class closer together in a low-dimensional latent space and further away from images not of the same class [40, 49]. Details about the exact training and testing procedure used are provided in the Supplement. Training results in a generalizable latent space where different classes are well separated (see Supplement). For the COVIDx dataset, we started with a pretrained DenseNet-121 model[1], which has been shown to perform well on the classification of different pathologies in chest X-rays [29]. We also tried training the model with either ImageNet pre-trained weights or randomly initialized weights, but these runs yielded poorer results so we did not pursue them any further. We used a triplet loss with a margin of 0.2, using all possible image pairs in a given batch. We sampled triplets for each batch using a 16-3 strategy (16 samples randomly from each of 3 classes). We used standard data augmentation (*i.e*. resize to 256 pixels, center or random resized crop to 224×224 pixels, and random horizontal flip). All embeddings were L_2 normalized. We also explored adding a 256-dimension linear layer for the final embedding (linear layers of other dimensions did not produce better results). We trained our models for 20 epochs, using the Adam optimizer with default beta values (0.9, 0.999) and a learning rate of $1e-4$. We did not use weight decay or any other forms of regularization.

For the ISIC 2017 dataset, we found that the DenseNet-121 architecture also performed well. Given the size of the dataset, training models with more parameters (*e.g*. ResNet-50) yielded poorer performance. We initialized the model with ImageNet pre-trained weights as models pre-trained on skin lesion classification were not publicly available. Unless otherwise stated, we used the same set of hyper-parameters as for the COVID-19 experiments. We did not overly-optimize for best model performance, as our study is more concerned with explaining models rather than obtaining state-of-the-art results on the given datasets.

Given an input image $\mathbf{x}$, the learned embedding network extracts an embedding feature vector $\mathbf{f}(\mathbf{x})$. In the image retrieval context, this is done for each query image $\mathbf{q}$ and each retrieved image $\mathbf{r}$, resulting in feature vectors $\mathbf{f_q}$ and $\mathbf{f_r}$, respectively, which are both D-dimensional vectors. To rank the retrieved images, a similarity score s between $\mathbf{f_q}$ and $\mathbf{f_r}$ is used. The similarity score s is calculated using cosine similarity as

$$s(\mathbf{f_q}, \mathbf{f_r}) = \frac{\mathbf{f_q} \cdot \mathbf{f_r}}{\|\mathbf{f_q}\| \|\mathbf{f_r}\|}. \quad (1)$$

The similarity score s indicates how similar each retrieved image is to the query image.

3.4. Similarity-based saliency maps

We explored three types of similarity-based saliency maps in the current work: occlusion-based [16], attention-based [70], and a form of activation mapping [57, 71]. These methods take as input a query image and a retrieved image, and indicate which regions of the retrieved image are most similar to the query image. The occlusion-based method is a black-box approach in which a small box is moved across the retrieved image to occlude parts of the image [16]. We measure the distance between the query embedding and the image embedding as a result of the occlusion. Occluded image regions which cause a large increase

[1]https://github.com/arnoweng/CheXNet

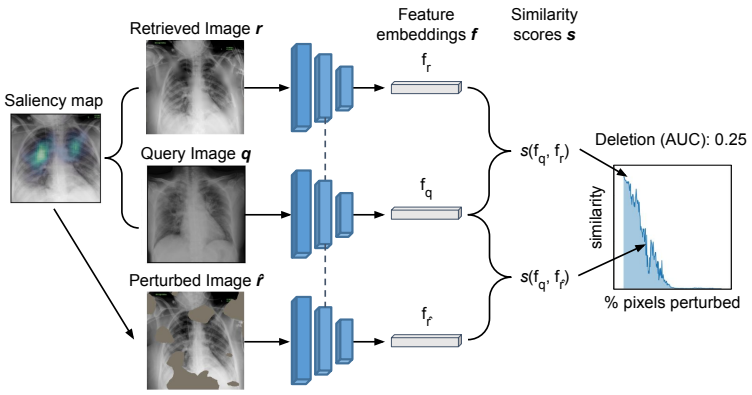

Figure 3. Metrics for quantifying similarity-based saliency maps. The retrieved image r (top) and query image q (middle) are passed through a network with shared weights (denoted by blue bars) to compute feature embeddings f_r and f_q, respectively. These feature embeddings (gray bars) are first used to compute a saliency map (left). The saliency map indicates important image regions in the retrieved image, which are then used to perturb the retrieved image ($\hat{r}$) and generate perturbed feature embeddings ($f_{\hat{r}}$). For each level of perturbation (in this example, deletion using a constant gray value), a similarity score $s(f_q, f_{\hat{r}})$ is calculated and used to compute the final AUC metric.

in distance are important for the computation of image similarity, while regions which do not change the distance are less important. This can be converted into a heat map representing saliency. We used a window size of 24×24 pixels with a stride of 5 pixels. We also explored a white-box, gradient-based method which uses the triplet loss to compute a form of similarity attention [70]. This method can take as input image triplets and explain why a set of images are either similar or dissimilar. The authors showed that the generated saliency maps can be used during training as a form of regularization to improve generalization performance. Finally, we tested a saliency method that requires access to features from the last convolutional layer of the model [57, 71]. In contrast to the occlusion-based method, where feature vectors are obtained after pooling, these methods compute similarity between unpooled features at each of the spatial locations in the last convolutional layer (7×7 spatial grid for most architectures). To compute this point-wise similarity, a single point in one feature map can be chosen, and the similarity is averaged across all other locations in the other feature map. To ensure the validity of the computed saliency methods, we also performed a model randomization test as proposed in [2], where we randomly re-initialized the weights of the model and re-computed saliency maps. We found that all methods passed this sanity check (see Supplement).

We also developed a form of *self-similarity* saliency, where we computed the similarity of an image with respect to itself. This is equivalent to applying one of the similarity-based saliency algorithms described above where the query image and the retrieved image are purposefully chosen to be the same. The resulting saliency map indicates which regions of the image are important for the computation of its associated feature embedding. This provides an unbiased method to evaluate saliency which is not dependent on the exact choice of query image or retrieved images. We show example average self-similarity saliency maps in Figure 6.

3.5. Insertion and deletion metrics

To evaluate the quality of different saliency maps, we adapted a set of causal metrics first introduced in [42]. In the classification context, the insertion and deletion metrics measure the increase or decrease in output classification probability as a result of changes to the input image. Here, we instead measure changes in image similarity as a result of changes to the input image (Figure 3). Deletion measures the drop in image similarity to the query image as more image pixels are removed from the retrieved image. We gradually mask out pixels on the retrieved image with a constant gray value from highest relevance to lowest based on the computed saliency map and measure the cosine similarity between the query image and the masked images. In contrast, insertion measures the increase in image similarity as more pixels are gradually introduced to the retrieved image. Starting from a blurred version of the original retrieved image, we gradually reveal high-resolution pixels from highest to lowest relevance and again measure the cosine similarity between the query image and unmasked images.

We compute the similarity score s between the query image $\mathbf{q}$ and perturbed versions of the retrieved image $\hat{\mathbf{r}}$. As stated above, perturbations are either in the form of insertion onto a blurred image or deletion using a constant gray

value. Extending Equation 1, the similarity score s is calculated using cosine similarity as

$$s(\mathbf{f_q}, \mathbf{f_{\hat{r}}}) = max(0, \frac{\mathbf{f_q} \cdot \mathbf{f_{\hat{r}}}}{\|\mathbf{f_q}\|\|\mathbf{f_{\hat{r}}}\|}). \quad (2)$$

To ensure non-negative outputs when using the cosine similarity metric, we rectified all similarity values to a minimum value of zero. For both insertion and deletion, we sweep over a range of perturbation values and use the area-under-the-curve (AUC) as a measure of the 'goodness' of the saliency maps, where higher AUC values are better for insertion (*i.e*. the added pixels rapidly increase image similarity) and lower AUC values are better for deletion (*i.e*. the removed pixels rapidly decrease image similarity).

3.6. Implementation details

All models were trained using Pytorch [41]. Training was done on a single Nvidia Titan Xp GPU, which was part of a GPU cluster. The saliency map methods were either implementations (if source code was unavailable) or re-implementations of author-released public versions. To ensure the robustness of our results, all reported results are the average over three random initializations.

4. Experimental Results

4.1. Image retrieval results

Using the deep metric learning framework, the learned embeddings generally separate into different clusters corresponding to the different classes in the COVID-19 dataset (Figure 1, inset). The centroids associated with each class also highlight the key differences between the different classes, with normal cases showing relatively clear (dark) lung regions, and pneumonia and COVID-19 cases showing more cloudy lung regions. We report our image retrieval results in Table 1. We use the standard image retrieval metrics of mean average precision (mAP) and mean precision (P@K) with $K = 1, 5$. We used DenseNet-121 [27] as our baseline model. *RR* refers to the use of random resizing data augmentation during training. *256-d* refers to the addition of an extra 256-dimension linear projection layer. Both of these techniques have been observed to help in deep metric learning on other problem domains [49]. However, we did not observe any clear trends in performance across the different combinations of models and augmentation methods. We also observed that image retrieval performance was higher on the COVID-19 dataset compared to the ISIC 2017 dataset, most likely due to use of a pretrained model and the larger COVIDx dataset size.

Dataset	Model	mAP ↑	P@1 ↑	P@5 ↑
COVID-19	DenseNet-121	87.6	90.0	89.9
	DenseNet-121+*RR*	86.6	89.4	88.8
	DenseNet-121+*256-d*	88.4	90.8	90.3
	DenseNet-121+*RR*+*256-d*	86.2	88.1	89.2
ISIC 2017	DenseNet-121	57.5	66.3	64.6
	DenseNet-121+*RR*	61.6	69.6	69.2
	DenseNet-121+*256-d*	56.0	64.7	62.4
	DenseNet-121+*RR*+*256-d*	59.7	67.0	66.2

Table 1. Image retrieval results. For image retrieval, we used mAP and P@1,5 metrics. *RR*: random resizing, *256-d*: 256-dimension linear projection layer.

4.2. Qualitative evaluation of saliency maps

We compare qualitative examples of similarity-based saliency maps generated by three different algorithms: occlusion [16], attention [70], and activation mapping [57, 71]. In Figure 4, we show example query images along with their top-three retrieved images on the COVID-19 dataset. In Figure 4A, we show examples where the model was correct, and in Figure 4B, we show examples where the model failed to retrieve images of the correct class. We find that for query images where the model was correct, saliency was generally more focused on the lung regions. In contrast, when the model was incorrect, saliency revealed that the model tended to focus more on non-lung regions such as the shoulder blades or the heart. We also show similar qualitative results on the ISIC 2017 dataset (Figure 5). Due to space constraints, other combinations of models and saliency algorithms are shown in the Supplement.

In general, we found that occlusion-based saliency maps are typically higher resolution, as the other saliency methods are computed based on the last convolutional layer, which is typically a $32\times$ downsampling of the original input image size. We also note many counter-examples where saliency is not well-localized even for correctly retrieved results or well-localized despite retrieving incorrect results. As a case in point, for the ISIC 2017 dataset, saliency seems to always be focused on the lesion irregardless of whether the returned image was of the correct class. This type of qualitative evaluation of saliency maps (and usually only of good examples, and not failure cases) is typically what is presented in papers. As these qualitative results are oftentimes misleading, we believe that quantitative measures of saliency are needed, which are discussed in the next section.

4.3. Quantitative evaluation of saliency maps

We report our results as areas-under-the curve (AUC) values for both the insertion and deletion metrics, where AUC should be higher for insertion and lower for deletion (Table 2). On the COVID-19 dataset, we found that occlusion-based saliency had the lowest deletion scores, while the attention-based and activation mapping saliency methods had higher insertion scores. On the ISIC 2017 dataset, we found that occlusion-based saliency performed the best on the insertion metric and the attention-based and activation mapping saliency methods performed the best on the deletion metric. As a control, using the saliency maps

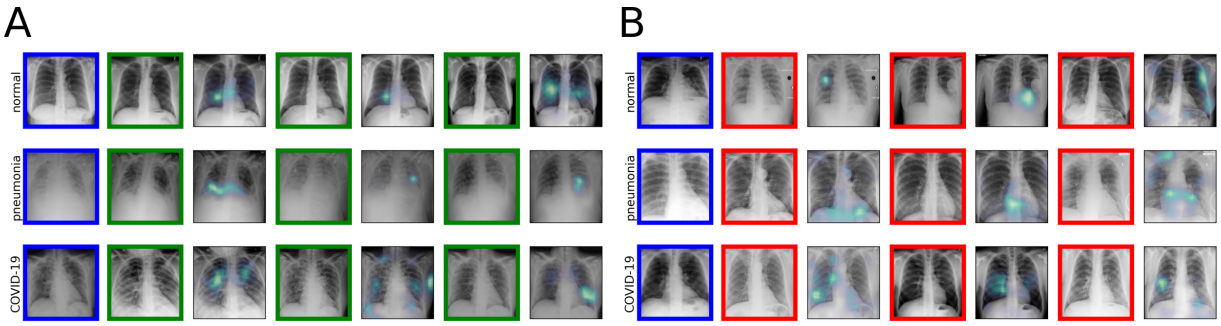

Figure 4. Example occlusion-based saliency maps [16] on the COVID-19 dataset, showing normal, pneumonia, and COVID-19 cases. (A) Queries where the top-3 retrieved images are correct. (B) Queries where the top-3 retrieved images are incorrect. Query images are marked in blue, correct results are marked in green, and incorrect results are marked in red.

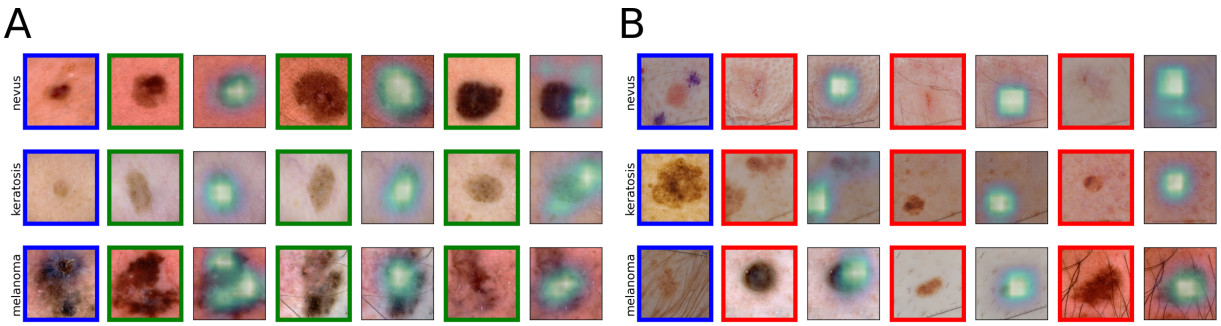

Figure 5. Example activation-based saliency maps [57, 71] on the ISIC 2017 dataset, showing nevus, keratosis, and melanoma examples. The same convention is used as in Figure 4.

Dataset	Method	Insertion ↑	Deletion ↓
COVID-19	Occlusion	83.2	64.8
	Attention	85.4	67.6
	Activation mapping	85.5	67.5
	Occlusion (random)	78.0	71.6
	Attention (random)	79.2	74.6
	Activation mapping (random)	79.1	74.5
ISIC 2017	Occlusion	75.8	58.2
	Attention	75.1	56.8
	Activation mapping	75.2	56.6
	Occlusion (random)	71.0	60.3
	Attention (random)	72.1	61.4
	Activation mapping (random)	72.1	61.1

Table 2. Quantitative evaluation of saliency maps using insertion (higher score is better) and deletion (lower score is better). *Random* refers to a model randomization test using randomly initialized weights [2].

generated by models with randomly initialized weights (denoted as *random* in Table 2) resulted in lower insertion and higher deletion scores on both datasets, indicating that the random saliency maps did not pick up on salient image features. This suggests that the proposed insertion and deletion metrics adapted to image similarity are appropriate for evaluating saliency maps as they capture differences between algorithms and are sensitive to trained and untrained models. We also note that the attention-based and activation mapping methods produced nearly identical saliency maps and associated insertion and deletion scores. This suggests a possible connection between these two methods which has not been shown before previously in the literature.

4.4. Self-similarity and differential saliency

Average images are often used to visualize the salient features associated with each class within a dataset [44]. We extend this approach by averaging the *self-similarity* saliency maps computed for each class. These saliency maps reveal image features used by the model to compute the learned feature embeddings. As a result, this approach highlights class-specific features in saliency space instead of image space, and can be viewed as a form of global explanation as it operates over the entire dataset (split by class), instead of individual data examples. In addition, we used a form of differential saliency to highlight which image features were more likely to correspond to the different conditions. The differential saliency maps were computed

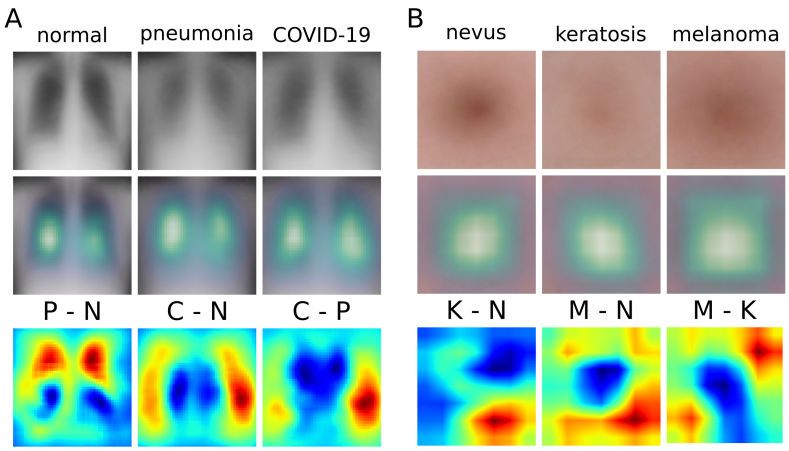

Figure 6. Self-similarity and differential saliency to distinguish between different conditions. (A) COVID-19 chest X-ray dataset. The average images (top row), average self-similarity saliency maps (middle row), and differential saliency maps are shown (bottom row). **P**: pneumonia, **C**: COVID-19, **N**: normal. (B) Same analysis for the ISIC 2017 skin lesion dataset. **K**: keratosis, **M**: melanoma, **N**: nevus.

by taking differences in the self-similarity saliency maps corresponding to the different conditions (*e.g*. COVID-19 - normal or C - N). In Figure 6, we show the average images, average self-similarity saliency maps, and differential saliency maps for each dataset. For the COVID-19 dataset, we find that both pneumonia and COVID-19 show slightly brighter lung regions in the average images. The COVID-19 cases also show zoomed in radiographs on average, which appears as differential saliency towards the image boundaries. On the ISIC 2017 dataset, the nevus examples have the highest contrast and are more centrally localized. As a result, the differential saliency maps show offsets for keratosis and melanoma relative to the nevus examples. While these results are preliminary, we believe this simple tool can also be used to set more meaningful baselines in attribution methods (*e.g*. comparing the sensitivity of saliency maps for the top-1 and top-2 predictions).

5. Discussion

We have shown that deep neural networks can be used for image retrieval on a COVID-19 chest X-ray dataset and a skin lesion dataset. In the present work, we did not focus on achieving state-of-the-art for these tasks, but instead demonstrate that it is possible to use different forms of saliency maps as XAI techniques in the medical imaging domain. In addition to qualitative comparisons, we also proposed novel quantitative measures of similarity-based saliency maps using the insertion and deletion metrics based on the cosine similarity between feature embeddings. Finally, we used our differential saliency method to reveal which image features were more likely to correspond to COVID-19 (or melanoma), which warrants further investigation and validation.

To test our explainable image retrieval approach, we would like to run user studies with radiologists or dermatologists to understand whether the provided saliency maps can improve their performance on image retrieval tasks [67]. Future work should explore the combination of saliency with other forms of explanations, *e.g*. textual explanations or medical report generation [30, 35]. Saliency maps could also be applied to other medical imaging problems, such as anomaly detection [63, 6, 28]. Our proposed methods are also general in that they can be applied to other datasets and models to study similar issues of using saliency maps for image similarity, such as in the face verification or person re-identification domains.

Acknowledgments

This material is based on research sponsored by Air Force Research Laboratory and DARPA under Cooperative Agreement number N66001-17-2-4028. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of Air Force Research Laboratory and DARPA or the U.S. Government. Distribution Statement 'A' (Approved for Public Release, Distribution Unlimited).

References

[1] Amina Adadi and Mohammed Berrada. Peeking inside the black-box: A survey on explainable artificial intelligence (xai). *IEEE Access*, 6:52138–52160, 2018.

[2] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pages 9505–9515, 2018.

[3] Mauro Annarumma and Giovanni Montana. Deep metric learning for multi-labelled radiographs. In *Proceedings of the 33rd Annual ACM Symposium on Applied Computing*, pages 34–37, 2018.

[4] ID Apostolopoulos and TA Mpesiana. Covid-19: automatic detection from x-ray images utilizing transfer learning with convolutional neural networks. *Physical and Engineering Sciences in Medicine*, pages 1–6, 2020.

[5] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai. *Information Fusion*, 58:82–115, 2020.

[6] Christoph Baur, Benedikt Wiestler, Shadi Albarqouni, and Nassir Navab. Deep autoencoding models for unsupervised anomaly segmentation in brain mr images. In *International MICCAI Brainlesion Workshop*, pages 161–169. Springer, 2018.

[7] Carrie J Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S Corrado, Martin C Stumpe, et al. Human-centered tools for coping with imperfect algorithms during medical decision-making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2019.

[8] Pornpimol Charoentong, Francesca Finotello, Mihaela Angelova, Clemens Mayer, Mirjana Efremova, Dietmar Rieder, Hubert Hackl, and Zlatko Trajanoski. Pan-cancer immunogenomic analyses reveal genotype-immunophenotype relationships and predictors of response to checkpoint blockade. *Cell reports*, 18(1):248–262, 2017.

[9] Lei Chen, Jianhui Chen, Hossein Hajimirsadeghi, and Greg Mori. Adapting grad-cam for embedding networks. In *The IEEE Winter Conference on Applications of Computer Vision*, pages 2794–2803, 2020.

[10] Po-Hsuan Cameron Chen, Yun Liu, and Lily Peng. How to develop machine learning models for healthcare. *Nature materials*, 18(5):410, 2019.

[11] Xiaocong Chen, Lina Yao, Tao Zhou, Jinming Dong, and Yu Zhang. Momentum contrastive learning for few-shot covid-19 diagnosis from chest ct images. *Pattern recognition*, 113:107826, 2021.

[12] 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.

[13] Joseph Paul Cohen, Paul Morrison, and Lan Dao. Covid-19 image data collection. *arXiv preprint arXiv:2003.11597*, 2020.

[14] CDC Covid and Response Team. Severe outcomes among patients with coronavirus disease 2019 (covid-19)—united states, february 12–march 16, 2020. *MMWR Morb Mortal Wkly Rep*, 69(12):343–346, 2020.

[15] Jeffrey De Fauw, Joseph R Ledsam, Bernardino Romera-Paredes, Stanislav Nikolov, Nenad Tomasev, Sam Blackwell, Harry Askham, Xavier Glorot, Brendan O'Donoghue, Daniel Visentin, et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nature medicine*, 24(9):1342–1350, 2018.

[16] Bo Dong, Roddy Collins, and Anthony Hoogs. Explainability for content-based image retrieval. In *CVPR Workshops*, pages 95–98, 2019.

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

[18] Oliver Eberle, Jochen Büttner, Florian Kräutli, Klaus-Robert Müller, Matteo Valleriani, and Grégoire Montavon. Building and interpreting deep similarity models. *arXiv preprint arXiv:2003.05431*, 2020.

[19] 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):115–118, 2017.

[20] Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature medicine*, 25(1):24–29, 2019.

[21] Yang Feng, Yubao Liu, and Jiebo Luo. Universal model for multi-domain medical image retrieval. *arXiv preprint arXiv:2007.08628*, 2020.

[22] Ruth C Fong and Andrea Vedaldi. Interpretable explanations of black boxes by meaningful perturbation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3429–3437, 2017.

[23] Stephanie A Harmon, Thomas H Sanford, Sheng Xu, Evrim B Turkbey, Holger Roth, Ziyue Xu, Dong Yang, Andriy Myronenko, Victoria Anderson, Amel Amalou, et al. Artificial intelligence for the detection of covid-19 pneumonia on chest ct using multinational datasets. *Nature Communications*, 11(1):1–7, 2020.

[24] Narayan Hegde, Jason D Hipp, Yun Liu, Michael Emmert-Buck, Emily Reif, Daniel Smilkov, Michael Terry, Carrie J Cai, Mahul B Amin, Craig H Mermel, et al. Similar image search for histopathology: Smily. *NPJ digital medicine*, 2(1):1–9, 2019.

[25] Andreas Holzinger, Chris Biemann, Constantinos S Pattichis, and Douglas B Kell. What do we need to build explainable ai systems for the medical domain? *arXiv preprint arXiv:1712.09923*, 2017.

[26] Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H Schwartz, and Hugo JWL Aerts. Artificial intelligence in radiology. *Nature Reviews Cancer*, 18(8):500–510, 2018.

[27] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4700–4708, 2017.

[28] Yotam Intrator, Gilad Katz, and Asaf Shabtai. Mdgan: Boosting anomaly detection using multi-discriminator generative adversarial networks. *arXiv preprint arXiv:1810.05221*, 2018.

[29] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, 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, pages 590–597, 2019.

[30] Donghyun Kim, Kuniaki Saito, Kate Saenko, Stan Sclaroff, and Bryan A Plummer. Self-supervised visual attribute learning for fashion compatibility. *arXiv preprint arXiv:2008.00348*, 2020.

[31] Pieter-Jan Kindermans, Sara Hooker, Julius Adebayo, Maximilian Alber, Kristof T Schütt, Sven Dähne, Dumitru Erhan, and Been Kim. The (un) reliability of saliency methods. *arXiv preprint arXiv:1711.00867*, 2017.

[32] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. *arXiv preprint arXiv:2007.04612*, 2020.

[33] Thijs Kooi, Geert Litjens, Bram Van Ginneken, Albert Gubern-Mérida, Clara I Sánchez, Ritse Mann, Ard den Heeten, and Nico Karssemeijer. Large scale deep learning for computer aided detection of mammographic lesions. *Medical image analysis*, 35:303–312, 2017.

[34] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. Unmasking clever hans predictors and assessing what machines really learn. *Nature communications*, 10(1):1–8, 2019.

[35] Xin Li, Rui Cao, and Dongxiao Zhu. Vispi: Automatic visual perception and interpretation of chest x-rays. *arXiv preprint arXiv:1906.05190*, 2019.

[36] Zhongyu Li, Xiaofan Zhang, Henning Müller, and Shaoting Zhang. Large-scale retrieval for medical image analytics: A comprehensive review. *Medical image analysis*, 43:66–84, 2018.

[37] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[38] Gianluca Maguolo and Loris Nanni. A critic evaluation of methods for covid-19 automatic detection from x-ray images. *arXiv preprint arXiv:2004.12823*, 2020.

[39] Henning Müller, Nicolas Michoux, David Bandon, and Antoine Geissbuhler. A review of content-based image retrieval systems in medical applications—clinical benefits and future directions. *International journal of medical informatics*, 73(1):1–23, 2004.

[40] Kevin Musgrave, Serge Belongie, and Ser-Nam Lim. A metric learning reality check. *arXiv preprint arXiv:2003.08505*, 2020.

[41] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[42] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018.

[43] Vitali Petsiuk, Rajiv Jain, Varun Manjunatha, Vlad I Morariu, Ashutosh Mehra, Vicente Ordonez, and Kate Saenko. Black-box explanation of object detectors via saliency maps. *arXiv preprint arXiv:2006.03204*, 2020.

[44] Jean Ponce, Tamara L Berg, Mark Everingham, David A Forsyth, Martial Hebert, Svetlana Lazebnik, Marcin Marszalek, Cordelia Schmid, Bryan C Russell, Antonio Torralba, et al. Dataset issues in object recognition. In *Toward category-level object recognition*, pages 29–48. Springer, 2006.

[45] Adnan Qayyum, Syed Muhammad Anwar, Muhammad Awais, and Muhammad Majid. Medical image retrieval using deep convolutional neural network. *Neurocomputing*, 266:8–20, 2017.

[46] Maithra Raghu, Chiyuan Zhang, Jon Kleinberg, and Samy Bengio. Transfusion: Understanding transfer learning for medical imaging. In *Advances in neural information processing systems*, pages 3347–3357, 2019.

[47] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. " why should i trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1135–1144, 2016.

[48] Michael Roberts, Derek Driggs, Matthew Thorpe, Julian Gilbey, Michael Yeung, Stephan Ursprung, Angelica I Aviles-Rivero, Christian Etmann, Cathal McCague, Lucian Beer, et al. Common pitfalls and recommendations for using machine learning to detect and prognosticate for covid-19 using chest radiographs and ct scans. *Nature Machine Intelligence*, 3(3):199–217, 2021.

[49] Karsten Roth, Timo Milbich, Samarth Sinha, Prateek Gupta, Bjoern Ommer, and Joseph Paul Cohen. Revisiting training strategies and generalization performance in deep metric learning. *arXiv preprint arXiv:2002.08473*, 2020.

[50] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

[51] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015.

[52] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296*, 2017.

[53] 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.

[54] Dinggang Shen, Guorong Wu, and Heung-Il Suk. Deep learning in medical image analysis. *Annual review of biomedical engineering*, 19:221–248, 2017.

[55] R Siegiel, K Miller, and A Jemal. Cancer statistics, 2017. *CA Cancer J Clin*, 67(1):7–30, 2017.

[56] Luis R Soenksen, Timothy Kassis, Susan T Conover, Berta Marti-Fuster, Judith S Birkenfeld, Jason Tucker-Schwartz, Asif Naseem, Robert R Stavert, Caroline C Kim, Maryanne M Senna, et al. Using deep learning for dermatologist-level detection of suspicious pigmented skin lesions from wide-field images. *Science Translational Medicine*, 13(581), 2021.

[57] Abby Stylianou, Richard Souvenir, and Robert Pless. Visualizing deep similarity networks. In *2019 IEEE winter conference on applications of computer vision (WACV)*, pages 2029–2037. IEEE, 2019.

[58] Enzo Tartaglione, Carlo Alberto Barbano, Claudio Berzovini, Marco Calandri, and Marco Grangetto. Unveiling covid-19 from chest x-ray with deep learning: a hurdles race with small data. *arXiv preprint arXiv:2004.05405*, 2020.

[59] Maria de la Iglesia Vayá, Jose Manuel Saborit, Joaquim Angel Montell, Antonio Pertusa, Aurelia Bustos, Miguel Cazorla, Joaquin Galant, Xavier Barber, Domingo Orozco-Beltrán, Francisco Garcia, et al. Bimcv covid-19+: a large annotated dataset of rx and ct images from covid-19 patients. *arXiv preprint arXiv:2006.01174*, 2020.

[60] Giulia Vilone and Luca Longo. Explainable artificial intelligence: a systematic review. *arXiv preprint arXiv:2006.00093*, 2020.

[61] Linda Wang, Zhong Qiu Lin, and Alexander Wong. Covid-net: A tailored deep convolutional neural network design for detection of covid-19 cases from chest x-ray images. *Scientific Reports*, 10(1):1–12, 2020.

[62] Wenling Wang, Yanli Xu, Ruqin Gao, Roujian Lu, Kai Han, Guizhen Wu, and Wenjie Tan. Detection of sars-cov-2 in different types of clinical specimens. *Jama*, 323(18):1843–1844, 2020.

[63] Qi Wei, Yinhao Ren, Rui Hou, Bibo Shi, Joseph Y Lo, and Lawrence Carin. Anomaly detection for medical images based on a one-class classification. In *Medical Imaging 2018: Computer-Aided Diagnosis*, volume 10575, page 105751M. International Society for Optics and Photonics, 2018.

[64] Jonathan R Williford, Brandon B May, and Jeffrey Byrne. Explainable face recognition. *arXiv preprint arXiv:2008.00916*, 2020.

[65] 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.

[66] Yu-Huan Wu, Shang-Hua Gao, Jie Mei, Jun Xu, Deng-Ping Fan, Rong-Guo Zhang, and Ming-Ming Cheng. Jcs: An explainable covid-19 diagnosis system by joint classification and segmentation. *IEEE Transactions on Image Processing*, 30:3113–3126, 2021.

[67] Yao Xie, Melody Chen, David Kao, Ge Gao, and Xiang'Anthony' Chen. Chexplain: Enabling physicians to explore and understand data-driven, ai-enabled medical imaging analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–13, 2020.

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

[69] Kang Zhang, Xiaohong Liu, Jun Shen, Zhihuan Li, Ye Sang, Xingwang Wu, Yunfei Zha, Wenhua Liang, Chengdi Wang, Ke Wang, et al. Clinically applicable ai system for accurate diagnosis, quantitative measurements, and prognosis of covid-19 pneumonia using computed tomography. *Cell*, 2020.

[70] Meng Zheng, Srikrishna Karanam, Terrence Chen, Richard J Radke, and Ziyan Wu. Learning similarity attention. *arXiv preprint arXiv:1911.07381*, 2019.

[71] Sijie Zhu, Taojiannan Yang, and Chen Chen. Visual explanation for deep metric learning. *arXiv preprint arXiv:1909.12977*, 2019.