HHS Public Access

Author manuscript

Phys Med Biol. Author manuscript; available in PMC 2021 June 27.

Published in final edited form as:

Phys Med Biol. ; 66(4): 04TR01. doi:10.1088/1361-6560/abcd17.

Interpretation and Visualization Techniques for Deep Learning Models in Medical Imaging

Daniel T. Huff[1], **Amy J. Weisman**[1], **Robert Jeraj**[1,2]

[1] Department of Medical Physics, University of Wisconsin-Madison, Madison WI

[2] Faculty of Mathematics and Physics, University of Ljubljana, Ljubljana, Slovenia

Abstract

Deep learning approaches to medical image analysis tasks have recently become popular; however, they suffer from a lack of human interpretability critical for both increasing understanding of the methods' operation and enabling clinical translation. This review summarizes currently available methods for performing image model interpretation and critically evaluates published uses of these methods for medical imaging applications. We divide model interpretation in two categories: (1) understanding model structure and function and (2) understanding model output. Understanding model structure and function summarizes ways to inspect the learned features of the model and how those features act on an image. We discuss techniques for reducing the dimensionality of high-dimensional data and cover autoencoders, both of which can also be leveraged for model interpretation. Understanding model output covers attribution-based methods, such as saliency maps and class activation maps, which produce heatmaps describing the importance of different parts of an image to the model prediction. We describe the mathematics behind these methods, give examples of their use in medical imaging, and compare them against one another. We summarize several published toolkits for model interpretation specific to medical imaging applications, cover limitations of current model interpretation methods, provide recommendations for deep learning practitioners looking to incorporate model interpretation into their task, and offer general discussion on the importance of model interpretation in medical imaging contexts.

Keywords

Deep learning; medical imaging; interpretation; visualization; saliency; review

*Correspondence should be addressed to: Wisconsin Institutes for Medical Research, 1111 Highland Ave, Room 1005, Madison, WI 53705. Tel: 608-263-8619; Fax: 608-262-2413, Corresponding author: Robert Jeraj, PhD; Department of Medical Physics; rjeraj@wisc.edu.

1. Introduction

Recently, artificial intelligence (AI) and, more specifically, deep learning (DL), approaches have achieved state of the art results for many medical imaging tasks including image segmentation (Hu *et al.*, 2017; Kamnitsas *et al.*, 2017; Roth *et al.*, 2015b), disease detection and diagnosis (Gao and Noble, 2017; Roth *et al.*, 2016; Kim *et al.*, 2018; Huynh *et al.*, 2016; Roth *et al.*, 2014), and image classification (Yang *et al.*, 2018; Chen and Shi, 2018; Van Molle *et al.*, 2018; Shen and Gao, 2018; Yi *et al.*, 2017). The workhorse of DL applications for medical imaging is the convolutional neural network (CNN). CNNs are a type of deep learning model that take images as input and consist of a series of convolutional layers and non-linear activations, the behavior of which are tuned by weights and biases learned throughout the model training process (Krizhevsky *et al.*, 2012).

While the popularity of using CNNs to perform medical image analysis tasks has increased rapidly in the past few years, a criticism often raised against them is their “black box” nature – meaning the internal structure of a CNN is not conducive to providing a simple explanation as to why a given input produces a corresponding output. To address this, researchers have developed model interpretation techniques and tools that aim to explain or visualize the decision-making process of CNNs. Interpretable models and interpretation methods in medical imaging have been the topic of several recent reviews and editorials (Jia *et al.*, 2019; Reyes *et al.*, 2020; Gastounioti and Kontos, 2020).

Model interpretation is of particular importance for medical imaging applications due to the complexity and high stakes nature of medical decisions. An incorrect diagnosis or failure to detect disease can be highly detrimental to patient care, so contributions of a deep learning model to medical decision-making processes need to be explainable in order to gain clinician trust (Hengstler *et al.*, 2016; Nundy *et al.*, 2019). In fact, both clinicians and patients alike have advocated for increased transparency for medical imaging applications of deep learning. In a recent review, (Hosny *et al.*, 2018) notes that the, “lack of transparency [of deep learning models] makes it difficult to predict failures, isolate the logic for a specific conclusion or troubleshoot inabilities to generalize to different imaging hardware, scanning protocols and patient populations.” In a patient perspective, (Andrews, 2017) stresses the need for model interpretation with an analogy, likening a radiologist’s use of AI to provide a medical diagnosis to a car mechanic’s use of a computer diagnostic tool to provide an automotive diagnosis. The use of the technology by both the radiologist and mechanic is helpful, but only so long as the technology can provide an explanation in terms the patient can understand.

As the use of AI tools in medicine become more widespread, additional legal requirements for model interpretability could become relevant. Article 22 of the General Data Protection Regulation (GDPR) adopted by European Union member states in 2018 contains requirements for automated decision-making, and some have argued that this could have implications for explanation of AI models in healthcare (Selbst and Powles, 2017).

The overall goal of this review is to describe existing approaches for model interpretation, provide examples of their application to medical imaging, provide recommendations for

deep learning practitioners looking to incorporate model interpretation into their task, and offer some general discussion on the importance of model interpretation in medical imaging contexts. Broadly, we divide the approaches to model interpretation into two categories: (1) understanding model structure and function (Section 2) and (2) understanding model predictions (Section 3), as described in Figure 1. Approaches in (1) primarily concern the hidden layers of the model: looking at hidden layer filters and features and visualizing or utilizing latent representations of data within the model. In contrast, approaches in (2) primarily concern the output of the model; these techniques produce heatmaps which describe which parts of an image are important to the model output.

Model interpretation closely follows the model development process as illustrated in Figure 2. Inspecting the model filters and feature activations can provide insight during the model training process (covered in Section 2.1), while techniques for reducing the dimension of high-dimensional data can come into play during both data collection and model deployment (covered in Section 2.2). Post-hoc attribution-based techniques for model interpretation, such as saliency maps or grad-CAM, provide interpretation of model output, and are most relevant during model training and at model deployment (covered in Section 3).

It is worth noting that there are several model interpretation toolkits published specifically with medical imaging in mind (covered in Section 4), which can facilitate application of interpretation methodologies. We conclude our review with specific recommendations for model interpretation best practices (Section 5), and discussion on the importance and limitations of model interpretation in medical contexts (Section 6).

2. Understanding model structure and function

When opening the “black box” of CNNs, the most direct approach to model interpretation is to look at the hidden layers of the network. This can be done in several ways, including direct inspection of the learned filters and feature maps, plotting high dimensional latent representations in two dimensions, or through employing models which learn useful latent representations.

2.1. Model filter and feature map visualization

A core component of modern CNNs is the convolutional layer (Goodfellow *et al.*, 2016). A convolutional layer takes the output of the previously layer as input, convolves it with a set of filters, sometimes also called features or kernels, and then applies a non-linear activation function. The output of the activation function is called a feature map or activation map and serves as input for the next layer (Figure 3A). CNNs learn features of varying complexity, often edges and corners in the first layers followed by more complex patterns in subsequent layers (Olah *et al.*, 2017). Understanding the features that the model learns and how those features act on images as they pass through the model can assist not only in ensuring the model is learning practical information, but also in connecting this information to patterns recognizable by humans.

2.1.1. Visual inspection of convolutional filters—After training a CNN, the learned filters can be visualized by loading the trained model and accessing the saved model

weights. The weights learned by the first convolutional layer are arguably the most useful for interpretation, as they act on the images directly. While subsequent layers may also provide useful information, the filters themselves are difficult to interpret as they are acting on feature maps from previous layers.

Within medical imaging, filter visualization has been used to compare filters across models and tasks. For example, (Roth *et al.*, 2016) compared filters learned for computer aided detection (CADe) of sclerotic metastases, lymph nodes, and colonic polyps on 3D computed tomography (CT) images. The authors observe that filters learned for lymph node detection represented a blobby texture and gradients in different orientations, whereas filters learned for colonic polyp detection were visually more diverse.

The ability to visualize the filters themselves depends on many aspects of the model architecture, most notably the filter size. This can be seen in comparing filters learned in (Yu *et al.*, 2018) with those learned in (Shin *et al.*, 2016), as shown in Figure 3B. The filters learned in (Yu *et al.*, 2018) were of size 3×3, for which only minor conclusions can be made due to the small amount of information shown in only 9 filter elements. Conversely, commonly known CNNs like AlexNet, which utilize 11×11 filters in the first layer, have been applied to medical imaging tasks such as lymph node detection and prostate segmentation (Shin *et al.*, 2016; Roth *et al.*, 2014). This allows the visualization of much more complex patterns and shapes (Figure 3C).

Filter visualization can be useful to compare filters learned after training a model with random initialization (i.e., from scratch) to those learned using transfer learning (initial weights taken from a network trained for another task). For example, filter visualization in (Shin *et al.*, 2016) indicated that AlexNet learned more blurry filters when random initialization was used, whereas transfer learning allowed more fine-tuning of higher-contrast and edge-preserving patterns (Figure 3B).

It should be noted that although larger filter sizes can arguably result in more interpretable filters, they require more memory. Commonly used 2D network architectures (e.g., AlexNet, GoogLeNet) generally utilize a combination of multiple filter sizes, whereas large filters are often not feasible in fully 3D models as the memory constraints of large kernels come at the cost of having fewer layers or fewer filters per layer (Kamnitsas *et al.*, 2017). However, memory limitations are likely to be less significant in the future as GPU memory capacity expands with time.

2.1.2. Visual inspection of feature maps—A more intuitive way of visualizing the features learned by a CNN is looking at the feature maps, sometimes referred to as activation maps. Feature maps are the output of the CNN at each layer. That is, they are the result of convolving the input of the layer with the filters of that layer and then applying an activation function. Thus, non-zero values in a feature map indicate that a feature was activated. Networks that have many feature maps that are all zero may indicate a problem with the training process.

Feature map visualization is a commonly used and straightforward model interpretation technique. It has been used in a wide variety of applications in medical imaging tasks including brain lesion segmentation on MRI images (Kamnitsas *et al.*, 2017), fetal facial plan recognition on ultrasound images (Yu *et al.*, 2018), classification of skin lesions on dermatology photographs (Van Molle *et al.*, 2018), and diagnosing Alzheimer's disease with PET/MRI (Zhang *et al.*, 2019).

Visualization of feature maps allows users to connect features that a human may learn to identify with features that the CNNs learn. For example, (Kamnitsas *et al.*, 2017) observed the network learning features to identify ventricles, cerebrospinal fluid (CSF), white, and gray matter on MRI images, indicating that differentiating between tissue types is useful for lesion segmentation. A similar finding was observed in (Van Molle *et al.*, 2018), where a CNN trained to classify skin lesions learned features corresponding to darker colors, skin types, lesion borders, and hair.

2.2. Dimensionality reduction

As discussed above, the hidden layer filters and feature maps of a CNN can be visualized. However, CNNs often have upwards of thousands of features per layer. To visualize such high dimensional data, techniques for reducing the number of dimensions while maintaining meaningful relationships between data points can be used. These methods take vectorized highly dimensional CNN features as input and produce a 2D summary that is easier to interpret. Principal component analysis (PCA) is perhaps the most well-known and widely used dimensionality reduction algorithm. PCA transforms the input data into orthogonal principal components (PCs) which are linear combinations of the original data. PCs are ordered by descending variance, such that the first few PCs often contain most of the useful information of the data and the remaining can be discarded without substantial loss of information. However, PCA relies on linear transformations, which are often not sufficient to preserve relationships of very high-dimensional data (Maaten and Hinton, 2008).

T-distributed Stochastic Neighbor Embedding (tSNE), introduced in (Maaten and Hinton, 2008), is a nonlinear dimensionality reduction technique consisting of two main stages. First, a probability distribution is constructed over the high dimensional data points, for which conditional probabilities of two objects are proportional to the similarity of those objects. Second, a similar probability distribution is created in a low-dimensional map (typically two dimensions). The Kullback-Leibler (KL) divergence is then minimized between the two distributions to ensure a good mapping to the low-dimensional space, ensuring that datapoints close together in the high dimensional space are similarly close together in the low dimensional space. tSNE is typically performed on the components of the last fully connected layer before the final classification layer. Depending on the dimension of this layer, PCA may be applied prior to tSNE to reduce the computational demands of performing tSNE.

tSNE is commonly used in visualizing deep learning models as it preserves pairwise Euclidean distances between data points. It can thus be used for several applications, for example visualizing patterns and clusters across classes and detecting outliers. This technique has been applied to the classification of abdominal ultrasound images (Cheng and

Malhi, 2017) as well as classification and anomaly detection in histopathology images (Faust *et al.*, 2018). An example of tSNE images generated from (Faust *et al.*, 2018) is shown in Figure 4A. In (Yu *et al.*, 2018), tSNE was performed not only on the components of the fully connected layers but also on vectorized 2D ultrasound images.

Another approach to visualizing high-dimensional data was proposed in (Plis *et al.*, 2014) to assess whether their CNN models were learning useful information. The authors argue that due to the complicated nature of tSNE, it is difficult to know whether a two-dimensional mapping of CNN features is of poor quality due to the tSNE process or due to the deep learning process. Instead, the authors propose a constraint-based embedding technique that uses a divide-and-conquer algorithm that recursively breaks a problem into smaller sub-problems until each sub-problem can be solved directly and explicitly outputs the constraints that are being satisfied. The constraint used in (Plis *et al.*, 2014) was that k nearest neighbors of the resulting 2D projection were the same as in the original space, where k is a tunable parameter. The authors apply their technique to visualize how well a deep belief network (DBN) can separate brain MRI images of schizophrenic and healthy patients at each layer of the network, showing increased separations at deeper layers of the network. Similarly, they also separate brain MRI images of patients with and without Huntington disease, as shown in Figure 4B. The authors also note that neither tSNE nor the constraint-based embedding was able to separate patients when applied directly to the raw data.

2.3. Autoencoders for learning latent representations

Autoencoders are a class of deep learning model common to unsupervised feature learning (Vincent *et al.*, 2008), with applications in anomaly detection (Kiran *et al.*, 2018), image compression (Cheng *et al.*, 2018; Theis *et al.*, 2017), and representation learning (Tschannen *et al.*, 2018). Autoencoders for imaging applications are similar to CNNs in that they take images as input but differ in that their output is not a label, but rather the output is equal to the input. Autoencoders consist of two stages: an encoder which converts an input image into a latent representation, and a decoder which reconstructs the image from the latent representation. Typically, the encoder and decoder are trained jointly, minimizing the reconstruction loss between input and output. However, multiple types of autoencoders with different structures and loss functions have been developed, including variational autoencoders (VAE) (Doersch, 2016) and adversarial autoencoders (AAE) (Makhzani *et al.*, 2015), among others.

Autoencoder use in medical imaging has focused predominantly on abnormality detection. In such application, an autoencoder is first trained with many examples containing no abnormality (i.e. scans of healthy patients with no pathology). This way the encoder learns a latent representation of normal images. Then, after abnormal test examples are introduced, their abnormalities are not captured in the latent representation, and the decoder will struggle to accurately reconstruct the parts of the image containing the abnormality. As a result, abnormal images can be detected by assessing the difference between the input image and model reconstruction. Simultaneously, the autoencoder can also provide a localization of abnormality by highlighting parts of the image with high reconstruction loss. In this way, autoencoders may be considered an interpretable form of deep learning model for image

analysis tasks, because they can provide an assessment of where an image differs from what is expected based on a distribution of normal images.

This approach to abnormality detection has been applied to multiple medical imaging tasks. In (Uzunova *et al.*, 2019), a VAE was trained to reconstruct OCT retinal images of healthy patients. The autoencoder was then used to classify three retinal pathologies in a separate dataset. The authors also assessed the ability of the VAE to localize the pathology in the OCT image, and compare the VAE to other visualization methods. They concluded that their proposed VAE-perturbation method was well-suited for explaining the output of their classifier. The authors also applied the same methodology to brain MRI, with similar results. An AAE is employed in (Chen *et al.*, 2020) to learn the distribution of healthy subject brain MRI images, and is then applied to test images of brain MRI containing lesions. The difference images between input and AAE reconstruction successfully localize lesions in the test images. A convolutional autoencoder was also used to detect nuclei in histopathology images by (Hou *et al.*, 2019). These authors developed an interesting approach to nuclei detection by combining learned latent representations with thresholding to separate images into a foreground containing nuclei and background containing cytoplasm.

3. Understanding model predictions

In contrast to model interpretation methods that involve visualizing intermediate network features or learned representations of data, other approaches to model interpretation try to attribute the model output to different parts of the input image. In general, they produce heatmaps that describe the importance of different parts of an image to the model decision on a pixel-by-pixel basis. Most attribution-based interpretation models function as *post-hoc* explanations – they are only meaningful when applied to a fully trained model, thus, they should be implemented after training has been completed. The available approaches generally fall into three groups: perturbation-based approaches (Section 3.1), backpropagation- or gradient-based approaches (Section 3.2), and decomposition-based approaches (Section 3.3). In addition to post-hoc attribution, so-called attention maps can be produced by trainable attention modules, which can be added to typical CNN architectures (Jetley *et al.*, 2018). This approach to model interpretation is covered in Section 3.4.

3.1. Perturbation-based methods

Perturbation-based approaches to model interpretation involve altering different parts of an image and seeing how those perturbations change the output of the model. The commonality of the approaches in this section is the underlying idea that when important parts of an image are perturbed, the output of the model is strongly affected, and when unimportant parts of an image are perturbed, the output of the model is unaffected. These approaches can be thought of as a type of sensitivity analysis to test the effect of small changes in model input on model output.

3.1.1. Occlusion—Occlusion as a means of performing model interpretation was first introduced in (Zeiler and Fergus, 2014). This technique consists of systematically occluding parts of an image and monitoring how strongly the perturbation influences model output. Image parts that, when occluded, strongly affect the output of the model are assigned high

importance, while image parts that have little effect on model output when occluded are of low importance. Kermany et al. employed occlusion as a way to perform model interpretation for the diagnosis of retinal pathologies in optical coherence tomography images (Kermany *et al.*, 2018). Several occlusion-produced heatmaps from this application are shown in Figure 5A.

One drawback of occlusion is that it amounts to performing inference on many slightly perturbed versions of the image, which requires computation. This drawback is especially pertinent if a high resolution heatmap is desired, as inference must be performed on as many perturbed images as there are occluded patches in the desired heatmap.

3.1.2. Local interpretable model-agnostic explanations (LIME)—LIME is an approach to model interpretation introduced by Ribeiro et al. (Ribeiro *et al.*, 2016). While LIME can be used to explain the prediction of any classifier, for this review we only consider LIME in the context of image models. LIME for images works by first identifying groups of contiguous pixels with similar intensities called superpixels. The image is then perturbed by turning subsets of superpixels "off" by replacing the value of all pixels in the superpixel with the mean intensity value of that superpixel. Like occlusion, changes in the model output due to the perturbation are used to identify how important each superpixel is to model output, and a heatmap highlighting the important superpixels is produced.

Seah et al. used LIME to visualize the salient portions of chest radiographs for identifying congestive heart failure (Seah *et al.*, 2018)). For their application, they find that LIME produces heatmaps that are less intelligible than their proposed Generative Visual Rationale method, as demonstrated in Figure 5B. An advantage of LIME over occlusion is that LIME uses superpixels that are more likely to correspond to semantically different parts of an image, while occlusion perturbs image patches in a systematic, uniform way, ignoring possible semantic similarity between adjacent pixels. LIME also uses less extreme perturbations than occlusion, as the intensities in the perturbed image region are replaced by the mean intensity instead of zeroes, however, there is nothing to prevent modification of either method to remove this difference.

3.1.3. Integrated gradients—Integrated gradients was first introduced in (Sundararajan *et al.*, 2017). The integrated gradients method considers the input image and a baseline image of all zeroes. Starting from the baseline image, a set of intermediate images are produced along the path from the baseline to the input image. At each step along the path, the gradient of the model output with respect to each pixel in the intermediate image is computed. Then, these gradients are summed over the path from baseline to input image. This produces the heatmap of pixelwise importance desired. Formally, the integrated gradients heatmap $IG(x)$ produced for a given input image x and baseline image x' is given by:

$$IG(x) = (x - x') \times \int_0^1 \frac{\partial F(x' + \alpha \times (x - x'))}{\partial x} d\alpha$$

where the function F: $R^n \rightarrow [0,1]$ represents the CNN model.

In its introductory paper, the authors demonstrate the use of integrated gradients for providing model explanations for detecting diabetic retinopathy on retinal fundus images. This use is expanded upon in (Sayres *et al.*, 2019), where ophthalmologists were tasked with grading the severity of diabetic retinopathy both with and without the explanatory heatmap produced by integrated gradients. Figure 5C shows an example of images provided to clinicians with and without interpretation. The authors report that readers provided with model-predicted grades and heatmaps graded patients with diabetic retinopathy more accurately than readers without any model assistance.

3.2. Backpropagation- or gradient-based methods

Backpropagation is the method by which weights in a neural network are updated during the model training process. Model interpretation methods in this section do not actually update model weights as occurs during training; rather, they rely on backpropagation to compute gradients, and these gradients are combined in different ways to visualize salient parts of an image.

3.2.1. Saliency maps—Saliency maps, introduced in 2013 by (Simonyan *et al.*, 2013), use gradients to visualize the classification of an image evaluated by a deep convolutional network. In the introductory paper, the authors offer two uses for saliency maps: class maximization visualization and image-specific class saliency maps.

Class maximization uses gradient ascent to produce an image that maximizes the activation of that class, and therefore can be interpreted as being most representative of that class. Formally, class maximization finds an image I of class c for which a class score S_c is maximized:

$$argmax S_c(I) - \lambda \|I\|_2^2$$

where λ is a regularization parameter.

In (Yi *et al.*, 2017), class maximization is used to produce visualizations of maximally malignant and maximally benign breast masses to help interpret the performance of a network trained for mammogram classification. Figure 6A contains examples of the class maximization visualizations for benign and malignant breast masses. The authors note that the maximally malignant visualization appears to contain a highly spiculated mass, a visual feature used by radiologists to identify malignant breast masses.

Image-specific class saliency maps are image- and class-specific heatmaps that represent the importance of individual pixels to the assignment of the image to a class, providing an assessment of which parts of an image are most important to the model. Saliency mapping is sometimes also referred to as “sensitivity analysis”, but it should be noted that it is a separate technique from the perturbation-based methods outlined previously in Section 3.1. Here, the heatmap $Sal_c(x)$ for a class c is computed directly as the derivative of the model output score $F_c(x)$ with respect to each pixel in the input image x through backpropagation:

$$Sal_c(x) = \frac{\partial F_c(x)}{dx}$$

Because of its simplicity, saliency mapping is one of the most widely implemented methods for model interpretation in medical imaging to date. As shown in Figure 6B, Dubost et al. employed image-specific class saliency maps as part of a weakly supervised approach to segmentation of structures on brain MRI (Dubost *et al.*, 2017; Dubost *et al.*, 2019). Other areas of application for saliency maps include heart disease classification on chest x-rays (Chen and Shi, 2018), detection of abnormalities in the spine on MRI (Jamaludin *et al.*, 2016), detection of artifacts on magnetoencephalography (Garg *et al.*, 2017), classification of breast masses in mammography (Lévy and Jain, 2016), and classification of pediatric elbow fractures on x-ray (Rayan *et al.*, 2019).

Gonzalez-Gonzalo et al. presented an expansion of class saliency maps in 2018 with the introduction of iterative saliency maps (González-Gonzalo *et al.*, 2018). The objective of iterative saliency mapping is to identify less discriminative image regions that may have been ignored in the initial saliency map. Briefly, the method works by iteratively computing a saliency map, inpainting the most salient image regions identified, and computing the saliency map again. This process repeats until the perturbed image is no longer classified as containing an abnormality, or a maximum number of iterations is reached. The final iterative saliency map is computed as a weighted sum of the saliency maps computed at each step. The authors apply their technique to the task of identifying retinal fundus image segments relevant to grading diabetic retinopathy and demonstrate higher sensitivity with iterative saliency maps compared to saliency maps without iterative refinement.

Despite their popularity, saliency mapping has the notable drawback that it provides no indication as to whether a pixel provides evidence for or against a class, only that the classification is sensitive to that pixel. Several authors have also noted that in binary classification settings, saliency maps lose their class specificity, because if a feature is important for distinguishing between two classes, it may be highlighted by a saliency map for both classes (Garg *et al.*, 2017).

3.2.2. Guided backpropagation—Guided backpropagation, introduced in (Springenberg *et al.*, 2014), is an extension to saliency maps introduced by (Simonyan *et al.*, 2013) and the ‘deconvnet’ concept introduced in (Zeiler and Fergus, 2014). The difference between these approaches lie in how backpropagation through Rectified Linear Unit (ReLU) activation layers of the network is handled. ReLU is an activation function commonly used in CNNs (Glorot *et al.*, 2011). During the forward pass, neurons with negative output are clamped to zero by ReLU by definition (ReLU(x) = max(0,x)). (Zeiler and Fergus, 2014) extended this idea to computing gradients in the backward pass by clamping to zero negative gradients. Guided backpropagation combines these two ideas, zeroing out signal through neurons that have either negative output during the forward pass or negative gradient during the backward pass. This produces a heatmap that highlights only pixels that provide positive evidence for a classification. Further discussion of the relationship between saliency maps, deconvnet, and guided backpropagation can be found in (Mahendran and Vedaldi, 2016).

Guided backpropagation has been used to visualize salient image pixels for the task of fetal heartbeat localization in ultrasound images by Gao et al. (Gao and Noble, 2017). They find that the heatmaps produced by guided backpropagation are robust to variations in heart appearance, scale, position, and contrast. Bohle et al. evaluated guided backpropagation as a method for visualizing Alzheimer's disease (AD) diagnosis on brain MRI, but found the visualizations produced by guided backpropagation to be less discriminative than those produced by other methods (Böhle *et al.*, 2019a).

3.2.3. Class activation mapping—Class activation mapping (CAM) was first introduced in 2016 by (Zhou *et al.*, 2016). Class activation mapping works by computing a weighted sum of feature maps following the final convolutional layer where the weights are provided by the fully connected layer following global average pooling, a type of pooling described in (Lin *et al.*, 2013). The class activation map $CAM_c(x)$ for a class c and image x is defined as:

$$CAM_c(x) = \sum_k w_k^c f_k(x)$$

where w_k^c are the weights for class c in the final network layer, and $f_k(x)$ is the corresponding feature map prior to global average pooling. Thus, $CAM_c(x)$ is a class-specific heatmap that indicates discriminative image segments.

Class activation mapping has seen use in both classification and localization applications in medical imaging. (Feng *et al.*, 2017) used class activation maps as part of a weakly-supervised approach to lung nodule segmentation on thoracic CT scans. First, a CNN was trained to perform binary classification of CT images as containing a nodule or not. Then, the authors show that class activation maps generated from the trained classification model successfully highlights nodule candidates. Similar weakly-supervised approaches using CAMs for chest x-ray abnormality and breast mass localization is described in (Hwang and Kim, 2016), and for ACL tear localization on knee MRI in (Liu *et al.*, 2019). Kim et al. computed class activation maps for the classification of benign vs malignant breast masses on mammograms, but found them difficult to interpret for their task, as shown in Figure 6C (Kim *et al.*, 2018). Other applications of class activation maps to medical imaging tasks include localization of diabetic retinopathy lesions in retinal fundus images (Gondal *et al.*, 2017), and weakly supervised diagnosis of tuberculosis on chest x-rays (Hwang and Kim, 2016).

A drawback of class activation mapping is that it places some restrictions on network architecture. It requires a global pooling layer, followed by a fully connected layer as the last layers before the output layer. While Zhou et al. used global average pooling when they introduced class activation mapping, related work by Oquab et al. produces similar localization score maps using global max pooling (Oquab *et al.*, 2015).

To address this limitation of class activation maps, (Selvaraju *et al.*, 2017) introduced gradient-weighted class activation maps (grad-CAM). In grad-CAM, the weights are the

gradients of the class score with respect to each feature map, instead of requiring that the weights be taken from a fully connected layer. That is:

$$gradCAM_c(x) = ReLU\left(\sum_k \alpha_k^c f_k(x)\right)$$

where the weights α_k^c are the gradients of the score for class c y_c with respect to the kth feature maps $f_k(x)$ of the preceding convolutional layer:

$$\alpha_k^c = \frac{\partial y_c}{\partial f_k(x)}$$

Selvaraju *et al.* define grad-CAM to include a ReLU activation, because they are interested only in features with a positive association with the class *c*. They also offer a further refinement with guided grad-CAM, which is the pixelwise product of grad-CAM and Guided Backpropagation (Springenberg *et al.*, 2014). More recently, (Zhao *et al.*, 2018) have added a further variant of class activation mapping with respond-weighted class activation mapping (Respond-CAM).

Garg et al. employed grad-CAM visualizations to identify discriminative regions of magnetoencephalography images in the task of detecting eye-blink artifacts (Garg *et al.*, 2017). The authors found that the regions of the eye highlighted by grad-CAM are the same regions that human experts rely on. Furthermore, (Shen and Gao, 2018) used grad-CAM to visualize areas of chest x-ray indicative of fourteen suspected diseases. In this multi-class setting, the class-specific nature of grad-CAM proved to be valuable.

3.3. Decomposition-based methods

Decomposition-based methods for model interpretation seek to decompose the prediction of the model to a heatmap that describes how much each pixel contributes to the prediction. Whereas perturbation- and gradient-based methods for interpretation highlight parts of the image that, if altered, affect the prediction of the model, decomposition-based methods identify parts of the image that directly provide evidence for the model decision.

3.3.1. Layer-wise relevance propagation—Layer-wise relevance propagation (LRP) was introduced by Bach et al. in 2015 (Bach *et al.*, 2015). Unlike saliency mapping, guided backpropagation, and grad-CAM, LRP does not rely on gradients to generate a heatmap. Instead, LRP works by computing relevance scores that distribute the output of the final layer amongst nodes in the previous layer. This process continues recursively until the input layer of the network is reached, producing a relevancy score heatmap that can be overlaid over the input image. Formally, the relevance score contribution to a neuron i in the lth layer from a neuron k in the (l+1)th layer is:

$$R_{i \leftarrow k}^{l, l+1} = R_k^{l+1} \frac{a_i w_{ik}}{\sum_h a_h w_{hk}}$$

The total relevance score for a neuron i in the lth layer is the sum of contributions over all neurons k in layer l+1 to which neuron i is connected:

$$R_i^l = \sum_k R_{i \leftarrow k}^{l, l+1}$$

Further properties of LRP and details of its theoretical basis are given in (Montavon *et al.*, 2017), and comparison of LRP to other interpretation methods can be found in (Samek *et al.*, 2016; Kohlbrenner *et al.*, 2019).

Layer-wise relevance propagation has seen some use in medical imaging applications. Eitel et al. applied LRP for providing interpretation to CNN-based diagnosis of multiple sclerosis (MS) on brain T2-weighted MRI (Eitel *et al.*, 2019). Interestingly, they show that LRP heatmaps focus on both hyperintense lesions in regions in the brain clinically associated with MS diagnosis as well as the thalamus, which is known to be affected by MS at an early disease stage (Figure 7A). The same researchers also used LRP to interpret CNN-based Alzheimer's disease (AD) diagnosis on MRI in (Böhle *et al.*, 2019a). The authors find that LRP heatmaps from their trained model highlight the hippocampal volume, which has been used to diagnose AD and predict disease progression (Figure 7B). They also compared LRP to guided backpropagation and concluded that LRP may be more valuable than guided backpropagation for their task because the difference in heatmap scores between Alzheimer's disease and healthy controls was more evident for LRP. Finally, Thomas et al. use LRP as part of their DeepLight framework for associating brain regions with different cognitive states on functional MRI (Thomas *et al.*, 2018).

3.4. Trainable attention models

In addition to techniques for interpretation that produce attribution heatmaps from a fully trained model, as covered in Sections 3.1–3.3, some research has also been done toward developing trainable mechanisms for attribution. In particular, the concept of soft attention as introduced for CNNs by (Jetley *et al.*, 2018) has seen application to medical image analysis tasks. This type of attention module can be added to any layer of a CNN to produce a fine-grain attention map which highlights salient parts of an image. Attention modules take as input the activation map output from the previous network layer (local features, l_s) and a global feature vector (g) obtained from the final network layer. The attention module computes a compatibility score c_s between the local features and the global features. In (Jetley *et al.*, 2018), two functions for computing compatibility are proposed: the dot product between local and global features ($\langle l_s, g \rangle$), and the dot product between the sum of local and global features and a learned vector u ($\langle u, l_s + g \rangle$). Intuitively, the compatibility score is high for local features which are similar to the global features from deeper in the network. Finally, the attention map a_s is produced by applying softmax normalization to the compatibility scores c_s:

$$a_s = \frac{\exp(c_s)}{\sum_{i=1}^{n} \exp(c_{s,i})}$$

The output of the attention module g_a is the local features weighted by the attention map:

$$g_a = \sum a_s l_s$$

Thus, the attention module increases signal from features with high compatibility with the global feature description of the image and suppresses signal from features with low compatibility.

Attention for 3D medical images was implemented by (Schlemper *et al.*, 2019), who introduced an Attention U-Net for organ segmentation in abdominal CT scans, and Attention Gate (AG)-Sononet for fetal ultrasound image plane classification. In both tasks, the authors demonstrated that the attention maps produced by their attention gates correctly highlighted the structures of interest. In (Li *et al.*, 2019), soft attention mechanisms were added to a traditional U-Net for segmenting breast masses on digital mammograms. For this task, the addition of attention provided a modest improvement in segmentation performance over the authors' base model. In melanoma lesion classification, the use of a regularized attention mechanism for the classification of skin lesion photographs demonstrated that attention maps generated from deeper network layers may focus more strongly on valid image regions than those from shallower layers (Yan *et al.*, 2019). Attention modules were added to a VGG-16 architecture for grading osteoarthritis severity on x-ray images of the knee (Górriz *et al.*, 2019). In contrast with the previous study, these authors obtained higher classification accuracy and more reasonable attention maps from earlier layers in the network. Other applications of attention mechanisms to medical imaging include classification of breast cancer histopathology images (Yang *et al.*, 2019), and segmentation of cardiac substructures on MRI (Sun *et al.*, 2020).

Trainable attention is particularly interesting as an interpretation strategy because not only does it provide an attribution heatmap, but it can also improve model performance (Schlemper *et al.*, 2019; Li *et al.*, 2019). However, the optimal architecture and implementation hyperparameters have yet to be determined and are likely to vary by application.

4. Toolkits for model interpretation specific to medical imaging

In Sections 2 and 3, we described methods for model interpretation that are application-agnostic. However, we highlighted examples of their use for medical imaging-specific applications. In this section, we summarize several publications that introduce tools for model interpretation designed specifically for medical imaging tasks. Some of the toolkits in this section incorporate approaches to model interpretation described in Sections 2 and 3. For example, Mimer, described in Section 4.3, makes use of grad-CAM, previously described in Section 3.2.3.

4.1. CLEAR-DR

Kumar et al. introduced CLass-Enhanced Attentive Response Discovery Radiomics (CLEAR-DR) as a framework for model interpretation in 2019 (Kumar *et al.*, 2019). CLEAR-DR is built on CLEAR, published previously by the same authors (Kumar *et al.*,

2017). CLEAR produces two types of explanatory heatmap: a dominant class attentive map, which assigns each image pixel to the class most influential at that location, and a dominant response map, which shows the dominant attentive level based on the identified dominant class. The authors apply the method to grading diabetic retinopathy (DR) in a set of more than 50,000 retinal fundus images. They produce heatmaps corresponding to evidence for each of five classes: mild, moderate, severe, and proliferative DR, as well as DR-negative. For their classification task, the authors find that in correctly classified images the CLEAR-DR maps correspond to relevant portions of the eye anatomy, and in cases that are misclassified, CLEAR-DR fails to focus on the relevant abnormality.

4.2. DeepMiner

DeepMiner, introduced in (Wu *et al.*, 2018b), is a framework for discovering interpretable representations for explaining medical imaging predictions. This framework builds on their previous work in which they demonstrate that visual patterns learned by their network correspond to relevant medical phenomena (Wu *et al.*, 2018a). DeepMiner uses a technique called network dissection (Bau *et al.*, 2017) to identify influential network features, and an expert then manually annotates these features with an application-specific semantic concept. The authors apply their framework to the task of mammogram classification. In this application, a radiologist specializing in mammography manually assigns a concept from the BI-RADS lexicon to influential features, such as “benign vascular calcification” or “spiculation”. DeepMiner then generates a report consisting of a test image overlaid with influential feature activation maps and the corresponding medical phenomena as an explanatory aid. The proposed framework is application-agnostic and could be extended to other image classification tasks where semantic concepts of image sections are of interest. However, a drawback of DeepMiner is its reliance on expert annotation of network feature activation maps. In their application, the authors find that 75% of influential features correspond to an identifiable medical phenomena, but this percentage may be task-dependent.

4.3. Mimer

Hicks et al. describes the development of an automated multimedia reporting system called Mimer in their 2018 paper (Hicks *et al.*, 2018b). The goal of Mimer is to produce an understandable and reproducible report containing text and images from a medical procedure appropriate for non-technical users. Users select an image, a network layer, and a target class, and Mimer produces a report describing the likelihood of the selected image containing the target class (model output), and a guided grad-CAM visualization of evidence for the target class (model interpretation). The authors provide example reports for performing polyp detection following a colonoscopy. They also provide clear, step-by-step instructions for installing the necessary dependencies for running Mimer either from a provided git repository, or a pre-configured Docker image.

4.4. DX-Caps

In (LaLonde *et al.*, 2019), the authors introduce a capsule network-based model for producing explainable diagnoses. Capsule networks differ from traditional CNNs in that the scalar feature maps produced at each layer in a traditional CNN are replaced with vectorized

representations (Sabour *et al.*, 2017). The authors use the capsule architecture to assign specific semantic concepts to each component of the network output. They apply DX-Caps to lung nodule classification and use a six-dimensional output to capture six attributes important to lung nodule malignancy prediction (subtlety, sphericity, margin, lobulation, spiculation, and texture). While DX-Caps does not produce a heatmap of interpretation, their approach of using a vectorized network output to provide a clinician with an explanation for the model output in terms of familiar semantic concepts is interesting and widely applicable to diagnosis tasks. However, a drawback of their approach that should be noted is its reliance on expert annotation of individual semantic concepts in training images, which may be time-consuming to produce and may differ from clinician to clinician.

4.5. MDNet

MDNet, introduced in (Zhang *et al.*, 2017), uses joint image and language models to provide diagnosis interpretation for images paired with text reports. For their image model, the authors use ResNet (He *et al.*, 2016), which they use to generate an image feature vector that is passed to the long short term memory (LSTM) language model. The LSTM also takes as input the text report. MDNet uses the attention mechanism from (Xu *et al.*, 2015) to produce heatmaps which show the image support for each word in the accompanying text. The authors apply MDNet to a bladder cancer pathology dataset of 32 whole-slide hematoxylin and eosin (H&E) stained samples from patients at risk for papillary urothelial neoplasm with paired diagnostic report text. The authors report that collaborating pathologists found the MDNet-produced attention maps to be “fairly encouraging” for highlighting informative regions of the images. While the authors only investigate applying MDNet to pathological images, radiological images would be another promising area of application, as large-scale repositories of paired images and radiologist-produced report text exist at any large academic hospital.

5. Recommendations

In this review, we have summarized current approaches to interpreting deep learning models used in medical image analysis. Here, we provide a succinct summary of recommended steps for conducting model interpretation in medical imaging. When thinking of deep learning data and model workflow, it helps to think of three distinctive steps (Figure 2): data collection, model training and model deployment.

Model interpretation can be performed during both model training and deployment. However, different approaches for interpretation are appropriate at different steps. For example, dimensionality reduction may be useful prior to model training to better understand underlying structure in a high dimensional dataset, whereas attribution-based methods for interpretation are only meaningful once a trained model is obtained, and so should be used after model training or once a model is deployed.

Before performing model interpretation, it is important to ensure that the model has been trained properly. Attempting to perform any of the interpretation techniques with a model trained with incorrect data, or a model overfit to training data, can be misleading. To help avoid common training pitfalls, we include Appendix 1 which summarizes some of the most

common issues likely to be encountered in training CNNs for medical image analysis tasks. However, as training of CNNs is a vast and complex topic, the summary is not meant to be exhaustive, but rather typical in analysis of medical images. For a full discussion on training CNNs, there are several other resources available, for example (Goodfellow *et al.*, 2016).

Filter and feature map visualization, as described in Section 2.1, is a simple way to perform model interpretation, but the value of doing so can be unclear. Some authors have been able to connect individual feature maps with human identifiable features (Van Molle *et al.*, 2018), but others find that layers contain many similar feature maps with little intuitive meaning (Zhang *et al.*, 2019). It is also important to keep in mind that feature maps are likely to become more complicated and less intuitive at deeper layers of the model (Olah *et al.*, 2017). Techniques like network dissection (Bau *et al.*, 2017), which identify important features, should be considered when pursuing feature map visualization for model interpretation.

Attribution-based methods for model interpretation is perhaps the largest research direction in model interpretation, and the existence of multiple attribution-based methods for model interpretation can make the process of choosing a method overwhelming. To better understand the relationship between attribution-based methods, we provide their “family tree” (Figure 8). Some qualities to consider when making this choice are method maturity and popularity, weaknesses described in literature, and publicly available implementations. More mature methods, such as saliency mapping (Simonyan *et al.*, 2013), have the advantages of simple intuition, and abundant examples of use in literature (Table 1), but drawbacks of the method have also been identified (Adebayo *et al.*, 2018; Rudin, 2018). Newer methods, such as grad-CAM (Selvaraju *et al.*, 2017) or layer wise relevance propagation (Bach *et al.*, 2015) have fewer examples of use in literature, but there are also fewer publications identifying their weaknesses. However, it is unclear if this is due to their being inherently more useful than previous methods, or just a product of their more recent development. Given the rapid pace of development in deep learning, there will likely be new methods for model interpretation developed in the future, which may address shortcoming of current methods.

Another point to consider when choosing a method for model interpretation is ease of implementation. Some public implementations are available. For example, the Keras Visualization Toolkit (Keras-vis) has implementations of class maximization, saliency maps, and grad-CAM. (https://raghakot.github.io/keras-vis/). Keras-explain (https://pypi.org/project/keras-explain/) is another project with implementations of multiple interpretation methods including grad-CAM, guided back-propagation, and integrated gradients. When choosing a publicly available implementation to pursue, it can be useful to check whether the developer is still actively supporting the project. Github (https://github.com/), a popular website for sharing code, shows when a project was last updated, and the Issues tab of a Github repository can be a useful indicator of whether the developer is likely to respond to questions.

6. Discussion

This review is the first to summarize both the technical and practical implementation details of model interpretability approaches for deep learning practitioners focusing on medical imaging applications. We have grouped interpretation approaches by their technical similarities, and by their relevance to different stages of the model development process. We have also provided practical advice for choosing between interpretation techniques and for implementing them. Deep learning-based approaches to medical image analysis is a rapidly expanding and exciting area of research, however the medical nature of the problems addressed in this field warrants extra emphasis on model interpretation.

When performing model interpretation, it is important to keep in mind that one of the end goals is to improve clinician trust in the model. To assess whether model interpretation has an impact on clinician behavior, comparison of clinician-alone vs clinician with model interpretation can be a valuable tool. For example, (Sayres *et al.*, 2019) compared reader performance in assigning diabetic retinopathy grade to retinal fundus images in three assistance settings: unassisted, reader provided with algorithm-predicted grade only, and reader provided with algorithm-predicted grade and integrated gradient heatmap. The authors compare the grading accuracy, reader confidence, and read time across these three assistance conditions and find that while algorithm-predicted grades improved grading accuracy and reader confidence, they also resulted in increased read time. There was also not a significant difference in grading accuracy when the integrated gradients heatmap was provided with the predicted grade as compared to the predicted grade alone. Regardless, this type of comparison is a strong example for assessing the impact of model interpretation on clinician-algorithm interaction.

It is important to recognize that the importance of model interpretation is task dependent, and different levels of model interpretation are necessary for different tasks. Table 1 organizes the reviewed literature by interpretation technique and image analysis task. The implementation of model interpretation techniques is common in detection and classification tasks, but not in segmentation tasks, despite segmentation being a large application of deep learning in medical imaging (Litjens *et al.*, 2017). This disparity may be attributed to the difference in perceived importance of interpretation by task. For detection and classification, it is natural to want to compare the parts of an image *the model* uses to make a prediction to the parts of an image *a physician* would use. In segmentation, the equivalent rationale for the importance of interpretation is less clear. For segmentation, perhaps interpretation heatmaps would highlight anatomical landmarks nearby a structure of interest, or would simply highlight the target structure itself, but this has not been fully investigated. Despite differences, all classes of application stand to gain the same benefits from model interpretation, including the investigation of model limitations, assessment of generalizability, and increased user trust.

The uses for model interpretation methods can extend beyond providing model interpretation alone. For example, (Dubost *et al.*, 2019; Dubost *et al.*, 2017) used saliency maps as part of a weakly-supervised method for detecting enlarged perivascular spaces on brain MRI. This use of saliency mapping is especially interesting because it was able to circumvent the need

for dense, pixel-by-pixel labels that would be required by a traditional, fully supervised detection approach. Dimensionality reduction techniques are commonly used to visualize network representations of data and gain intuition into how the network separates classes, but they can also be used to identify outlier data. In (Faust *et al.*, 2018), the authors apply a k nearest neighbors approach to tSNE-produced representations of patches drawn from whole slide histology images of CNS tissue samples. They find that tSNE representations of new glioma patches fall close to glioma patches in their training data. Another possible application of model interpretation techniques is to inform the design of mathematical models. For example, understanding feature maps, as described in Section 2.1, can potentially provide insight into the key dependencies describing a particular system, and could be useful inputs for modeling of the investigated system behavior. As these examples indicate, there is great potential for creative and valuable use of interpretation techniques as part of other medical image analysis approaches.

Current challenges and future directions

While current approaches to model interpretation can provide valuable insight into how a deep learning model is performing, there are important limitations that should be discussed. First, some researchers have argued that explanations provided by commonly employed attribution-based interpretation methods such as saliency mapping or class activation mapping are not reliable and can be misleading (Rudin, 2018; Adebayo *et al.*, 2018). These concerns have been voiced within medical imaging as well. For example, (Seah *et al.*, 2018) report that in detecting abnormalities in chest x-rays, several attribution-based interpretation methods (occlusion, integrated gradients, LIME) produced nonspecific heatmaps for the expected abnormality. In (Böhle *et al.*, 2019a), the authors compared guided backpropagation and layer-wise relevance propagation for producing heatmaps of explanation for separating Alzheimer's Disease from healthy controls on brain MRI. They observed that guided backpropagation failed to produce heatmaps that were visually dissimilar for Alzheimer's Disease versus healthy controls.

To avoid these weaknesses of post-hoc explanation methods, (Rudin, 2018) suggested that new types of models designed to be inherently interpretable should be used instead. One possible approach for this was described in (Hase *et al.*, 2019), where test images were classified by comparing them to a predefined hierarchical taxonomy of images that act as primitives of each classification category. As an example, the authors describe the classification of an image of a capuchin monkey: first, the image is determined to contain an animal based on similarity between the test image and an animal prototype image, then it is determined to contain a primate, and finally a capuchin. This interesting approach to interpretation could be potentially valuable in medical imaging applications where abnormalities or pathologies have a hierarchical relationship. However, this approach would likely be restricted to classification problems.

Adversarial attacks represent another limitation of current model interpretation techniques. Adversarial attacks refer to image perturbations designed to strongly affect the prediction of a deep learning model without affecting the appearance of the image to a human observer (Szegedy *et al.*, 2013; Kurakin *et al.*, 2016). Multiple researchers have demonstrated that

medical images are susceptible to adversarial attack (Finlayson *et al.*, 2018; Mirsky *et al.*, 2019). In addition to adversarial attacks designed to maximally perturb model predictions, adversarial attacks against model interpretation heatmaps have also been investigated. These perturbations are designed to maximally change the interpretation heatmap, while leaving the model prediction unaffected. (Ghorbani *et al.*, 2017) showed that small perturbations to natural images could be designed to maximally change the heatmaps produced by several commonly used interpretation methods. The adversarial perturbations could also be designed to cause the interpretation method to selectively highlight a part of the targeted image that is semantically different from the predicted label. This kind of attack would be especially damaging to physician model trust in medical settings. Given the susceptibility of medical images to adversarial attack against their predictions, it is reasonable to expect medical image interpretation to be similarly vulnerable. However, despite these limitations, many other publications summarized in this review have demonstrated useful benefit from performing model interpretation with various methods.

Finally, several growing areas of medical imaging research are ripe for application of model interpretation techniques. First, image to image synthesis tasks are an interesting area of application. For example, MRI to CT synthesis for PET/MR attenuation correction and MRI-only radiotherapy planning (Wolterink *et al.*, 2017), or low dose to high dose image synthesis for radiotracer dose reduction (Wang *et al.*, 2018; Yi and Babyn, 2018). No established interpretation methods have been consistently applied to this class of application. Additionally, several cancer imaging studies investigating the relationship between non-invasive imaging modalities and pathology or genetic information have recently been published. For example, pre-treatment PET textural features were correlated with vascular endothelial growth factor (VEGF) expression in head and neck cancer patients (Chen *et al.*, 2017). Similarly, several groups have correlated PET findings to programmed death ligand-1 (PD-L1) expression, a possible marker for patient response to cancer immunotherapies (Chen *et al.*, 2019; Takada *et al.*, 2017; Jreige *et al.*, 2019). Applying model interpretation techniques to this class of problem could potentially increase understanding of the connections between image features and the underlying biology. Early work in this direction has been done in (Wang *et al.*, 2019), where a deep learning approach to predicting epidermal growth factor receptor (EGFR) status from chest CT in lung adenocarcinoma patients made use of grad-CAM to provide visual evidence of the model decision in the image. The authors of this work provide example visualizations of both EGFR+ and EGFR- cases to demonstrate the differences in grad-CAM maps by EGFR status.

7. Summary

We have reviewed approaches to interpreting CNN-produced predictions and their use in medical imaging applications. Model interpretation can be performed by looking inside the model at the features it learns, or by looking at the output of the model and understanding which parts of an image were important to producing that output. Medical images have unique characteristics which should be considered when performing model interpretation, and several tools designed to accommodate these unique characteristics have been developed. It is now well established that deep learning models can achieve state of the art performance for a wide variety of medical image analysis tasks, but in order to better

understand the models and gain clinician trust, developing methods for providing clear and interpretable rationale for model decisions is critical. For this reason, it is imperative that developments in model interpretation progress in step with developments in model performance. It is equally important that investigators applying deep learning to medical imaging tasks rigorously implement and consistently report on model interpretation steps undertaken for their task.

Acknowledgements

Research reported in this publication was supported in part by the University of Wisconsin Carbone Cancer Center under Award Number P30CA014520, and by the National Cancer Institute of the National Institutes of Health under Award Number T32CA009206. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. The authors would like to thank Dr. Stephen Yip, Brayden Schott, and Žan Klaneček for editorial assistance.

Appendix 1:: Ensuring proper model training

While this appendix does not strictly pertain to model interpretation, we choose to include it because attempting to perform any of the interpretation techniques with an improperly trained model can be misleading. The steps described in this section can identify common issues and save time that would later be spent debugging. This appendix is not meant to serve as an exhaustive tutorial on training CNNs; it is only designed to cover some common tips with added emphasis on aspects especially relevant for medical imaging. For a full discussion on these topics, there are several other resources available, for example (Goodfellow *et al.*, 2016).

Ensuring data correctness

Before beginning to train a CNN, the dataset should be carefully checked for potential inconsistencies or artifacts. For example, any desired unit conversion should be applied consistently to the whole dataset, and in multi-modality images, spatial agreement between image channels should be checked. Image data used for training and testing CNNs might go through transformation steps that can include resampling, cropping, or normalizing image intensity. These might be part of image preprocessing, post-processing, or as part of a data augmentation strategy to increase dataset size. It is important to ensure that these steps do not introduce any inconsistencies that degrade model performance and confound model interpretation. Thus, data should be visually checked both before and after these steps are taken. For example, in training a segmentation model, it is important to ensure that any transformations applied to the input images as augmentation are also applied to the corresponding ground truth segmentation masks so that their spatial correspondence is maintained. This process is often time consuming but minimizes performance issues that would cause frustration and misinterpretation of results down the road.

Another common aspect of deep learning applications to medical imaging is the use of expert ground truth labels. For a classification task, this might take the form of images being labeled as “benign” or “malignant” by a physician, or for a segmentation task, this might be the manual contouring of an anatomical structure. To ensure ground truth consistency, it is important to communicate clearly with the physician or radiologist reviewing the images

what form the ground truth should take. Meeting with a clinical collaborator and labelling several cases together can avoid misunderstandings and minimize wasted effort. Following manual ground truth assignment, the labels should be checked for completeness and consistency.

Dataset size is often limited in medical imaging contexts, which can present a challenge for training deep learning models. Particularly, overfitting on a small dataset is a concern. Overfitting occurs when a deep learning model has the capacity to effectively memorize the training set, which leads to excellent training set performance, but poor validation or test set performance. Data augmentation strategies, such as image cropping, rotation, flipping, and intensity shifting/scaling can be employed to help increase the effective size of the training set (Shorten and Khoshgoftaar, 2019).

Model training

During the CNN training process, it is important to monitor the performance of both a training and validation set. The training set is the data used to update the weights of the model, while the validation set is used to monitor model performance and avoid overfitting. Most popular deep learning frameworks provide a way to produce learning curves, which plot training and validation set performance as a function of training iterations (Perlich, 2010). Inspecting these learning curves can reveal a large amount of information about the model training process (Google). For example, a steep increase in training set loss could indicate exploding gradients. This can be remedied by stopping and restarting network training with a decreased learning rate or implementing gradient clipping. Conversely, if the loss is not decreasing or decreasing slowly, the learning rate may be increased for faster model convergence. If the validation loss starts increasing but training loss continues to decrease, this is an indication that the model is starting to overfit the training set, and training should be stopped.

Checking model performance often during training can save a large amount of training time, allowing for faster model iteration and experimentation. Many other training parameters can also affect the training process such as choice of optimizer, choice of weight initialization scheme (Glorot and Bengio, 2010; He *et al.,* 2015), batch size, and regularization, among others (Goodfellow *et al.,* 2016). Setting these parameters appropriately is important and application-dependent, but outside the scope of this review.

Assessing model output

The appropriate method for assessing model output is dependent on the task being performed, and it is important to avoid using misleading performance metrics to evaluate model performance. For example, in a binary classification task, accuracy can be a misleading metric if the prevalence of one class is much higher than that of the other, as is often the case in medical applications. More appropriate metrics for imbalanced classes such as sensitivity, specificity, and positive and negative predictive value can be used. In evaluating segmentation performance, overlap based metrics such as Dice coefficient can be

misleadingly high for large structures, or misleadingly low for small structures. Other metrics based on surface-to-surface distances can be used in this case.

Additionally, it is important to distinguish between the metric that the network is working to minimize and the final performance metric appropriate for your application. For some applications, these two metrics may be the same, but they need not be. For example, in classification tasks, cross-entropy is a commonly used cost function used to optimize the model but calculating the cross-entropy on the test set may not be as intuitive as other performance metrics. For example, it may be more meaningful to generate a confusion matrix to see where misclassifications have occurred. Identifying instances in which the CNN performs poorly can provide direction as to how to improve the model. For example, if a classification model is consistently misclassifying a specific class, more examples for that class may be needed or class-balancing methods may need to be implemented.

Acronyms

CADe	computer aided detection
CADx	computer aided diagnosis
CAM	class activation map
CNN	convolutional neural network
CNS	central nervous system
CT	computed tomography
DBN	deep belief network
LIME	local interpretable model-agnostic explanations
LRP	layer-wise relevance propagation
LSTM	long short-term memory
MEG	magnetoencephalography
MRI	magnetic resonance imaging
PET	positron emission tomography

References

Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt Mand Kim B Advances in Neural Information Processing Systems,2018), vol. Series) pp 9505–15

Andrews D 2017 AI in Imaging: A Patient's Perspective. (American College of Radiology Data Science Institute AI Assistant Blog: American College of Radiology)

Bach S, Binder A, Montavon G, Klauschen F, Müller KR and Samek W 2015 On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation PLoS One 10 e0130140 [PubMed: 26161953]

Bau D, Zhou B, Khosla A, Oliva A and Torralba A Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2017), vol. Series) pp 6541–9

Böhle M, Eitel F, Weygandt M and Ritter K 2019a Layer-Wise Relevance Propagation for Explaining Deep Neural Network Decisions in MRI-Based Alzheimer's Disease Classification Frontiers in Aging Neuroscience 11

Böhle M, Eitel F, Weygandt M and Ritter K 2019b Visualizing evidence for Alzheimer's disease in deep neural networks trained on structural MRI data arXiv preprint arXiv:1903.07317

Chen R, Zhou X, Liu J and Huang G 2019 Relationship between the expression of PD-1/PD-L1 and 18 F-FDG uptake in bladder cancer European journal of nuclear medicine and molecular imaging 46 848–54 [PubMed: 30627815]

Chen S-W, Shen W-C, Lin Y-C, Chen R-Y, Hsieh T-C, Yen K-Y and Kao C-H 2017 Correlation of pretreatment 18 F-FDG PET tumor textural features with gene expression in pharyngeal cancer and implications for radiotherapy-based treatment outcomes European journal of nuclear medicine and molecular imaging 44 567–80 [PubMed: 27999896]

Chen X and Shi B 2018 Deep Mask For X-ray Based Heart Disease Classification arXiv preprint arXiv:1808.08277

Chen X, You S, Tezcan K C and Konukoglu E 2020 Unsupervised lesion detection via image restoration with a normative prior Medical Image Analysis 101713 [PubMed: 32492582]

Cheng P M and Malhi H S 2017 Transfer learning with convolutional neural networks for classification of abdominal ultrasound images Journal of digital imaging 30 234–43 [PubMed: 27896451]

Cheng Z, Sun H, Takeuchi M and Katto J 2018 Picture Coding Symposium (PCS),2018), vol. Series): IEEE) pp 253–7

Deng J, Dong W, Socher R, Li L-J, Li K and Fei-Fei L 2009 IEEE conference on computer vision and pattern recognition,2009), vol. Series): Ieee) pp 248–55

Doersch C. 2016Tutorial on variational autoencoders. arXiv preprint arXiv:1606.05908.

Dubost F, Adams H, Bortsova G, Ikram M A, Niessen W, Vernooij M and de Bruijne M 2019 3D regression neural network for the quantification of enlarged perivascular spaces in brain MRI Medical image analysis 51 89–100 [PubMed: 30390514]

Dubost F, Bortsova G, Adams H, Ikram A, Niessen WJ, Vernooij M and De Bruijne M International Conference on Medical Image Computing and Computer-Assisted Intervention,2017), vol. Series): Springer) pp 214–21

Eitel F, Soehler E, Bellmann-Strobl J, Brandt AU, Ruprecht K, Giess RM, Kuchling J, Asseyer S, Weygandt M, Haynes J-D, Scheel M, Paul F and Ritter K 2019 Uncovering convolutional neural network decisions for diagnosing multiple sclerosis on conventional MRI using layer-wise relevance propagation NeuroImage: Clinical 102003 [PubMed: 31634822]

Faust K, Xie Q, Han D, Goyle K, Volynskaya Z, Djuric U and Diamandis P 2018 Visualizing histopathologic deep learning classification and anomaly detection using nonlinear feature space dimensionality reduction BMC bioinformatics 19 173 [PubMed: 29769044]

Feng X, Yang J, Laine AF and Angelini ED International Conference on Medical Image Computing and Computer-Assisted Intervention,2017), vol. Series): Springer) pp 568–76

Finlayson SG, Chung HW, Kohane IS and Beam AL 2018 Adversarial attacks against medical deep learning systems arXiv preprint arXiv:1804.05296

Gao Y and Noble JA International Conference on Medical Image Computing and Computer-Assisted Intervention,2017), vol. Series): Springer) pp 305–13

Garg P, Davenport E, Murugesan G, Wagner B, Whitlow C, Maldjian J and Montillo A International Conference on Medical Image Computing and Computer-Assisted Intervention,2017), vol. Series): Springer) pp 374–81

Gastounioti A and Kontos D 2020 Is It Time to Get Rid of Black Boxes and Cultivate Trust in AI? Radiology: Artificial Intelligence 2 e200088 [PubMed: 32510520]

Ghorbani A, Abid A and Zou J 2017 Interpretation of neural networks is fragile arXiv preprint arXiv:1710.10547

Glorot X and Bengio Y Proceedings of the thirteenth international conference on artificial intelligence and statistics,2010), vol. Series) pp 249–56

Glorot X, Bordes A and Bengio Y Proceedings of the fourteenth international conference on artificial intelligence and statistics,2011), vol. Series) pp 315–23

Gondal WM, Köhler J M, Grzeszick R, Fink GA and Hirsch M 2017 IEEE International Conference on Image Processing (ICIP),2017), vol. Series): IEEE) pp 2069–73

González-Gonzalo C, Liefers B and Ginneken B v 2018), vol. Series)

Goodfellow I, Bengio Y and Courville A 2016 Deep learning (Cambridge, Massachusetts: The MIT Press)

Google Interpreting Loss Curves. In: Testing and Debugging in Machine Learning,

Górriz M, Antony J, McGuinness K, Giró-i-Nieto X and O'Connor NE 2019 Assessing Knee OA Severity with CNN attention-based end-to-end architectures arXiv preprint arXiv:1908.08856

Hase P, Chen C, Li O and Rudin C 2019 Interpretable Image Recognition with Hierarchical Prototypes arXiv preprint arXiv:1906.10651

He K, Zhang X, Ren S and Sun J Proceedings of the IEEE international conference on computer vision,2015), vol. Series) pp 1026–34

He K, Zhang X, Ren S and Sun J Proceedings of the IEEE conference on computer vision and pattern recognition,2016), vol. Series) pp 770–8

Hengstler M, Enkel E and Duelli S 2016 Applied artificial intelligence and trust—The case of autonomous vehicles and medical assistance devices Technological Forecasting and Social Change 105 105–20

Hicks S, Riegler M, Pogorelov K, Anonsen KV, de Lange T, Johansen D, Jeppsson M, Randel KR, Eskeland SL and Halvorsen P 2018 IEEE 31st International Symposium on Computer-Based Medical Systems (CBMS),2018a), vol. Series): IEEE) pp 363–8

Hicks SA, Eskeland S, Lux M, Lange T d, Randel KR, Jeppsson M, Pogorelov K, #229, Halvorsen l and Riegler M 2018b Mimir: an automatic reporting and reasoning system for deep learning based analysis in the medical domain. In: Proceedings of the 9th ACM Multimedia Systems Conference, (Amsterdam, Netherlands: ACM) pp 369–74

Hosny A, Parmar C, Quackenbush J, Schwartz LH and Aerts HJ 2018 Artificial intelligence in radiology Nature Reviews Cancer 18 500 [PubMed: 29777175]

Hou L, Nguyen V, Kanevsky AB, Samaras D, Kurc TM, Zhao T, Gupta RR, Gao Y, Chen W and Foran D 2019 Sparse autoencoder for unsupervised nucleus detection and representation in histopathology images Pattern recognition 86 188–200 [PubMed: 30631215]

Hu P, Wu F, Peng J, Bao Y, Chen F and Kong D 2017 Automatic abdominal multi-organ segmentation using deep convolutional neural network and time-implicit level sets International journal of computer assisted radiology and surgery 12 399–411 [PubMed: 27885540]

Huynh BQ, Li H and Giger ML 2016 Digital mammographic tumor classification using transfer learning from deep convolutional neural networks J Med Imaging (Bellingham) 3 034501 [PubMed: 27610399]

Hwang S and Kim H-E International Conference on Medical Image Computing and Computer-Assisted Intervention,2016), vol. Series): Springer) pp 239–46

Jamaludin A, Kadir T and Zisserman A International Conference on Medical Image Computing and Computer-Assisted Intervention,2016), vol. Series): Springer) pp 166–75

Jetley S, Lord NA, Lee N and Torr PH 2018 Learn to pay attention arXiv preprint arXiv:1804.02391

Jia X, Ren L and Cai J 2019 Clinical implementation of AI technologies will require interpretable AI models Medical physics

Jreige M, Letovanec I, Chaba K, Renaud S, Rusakiewicz S, Cristina V, Peters S, Krueger T, de Leval L and Kandalaft LE 2019 18 F-FDG PET metabolic-to-morphological volume ratio predicts PD-L1 tumour expression and response to PD-1 blockade in non-small-cell lung cancer European journal of nuclear medicine and molecular imaging 46 1859–68 [PubMed: 31214790]

Kamnitsas K, Ledig C, Newcombe V F, Simpson J P, Kane A D, Menon D K, Rueckert D and Glocker B 2017 Efficient multi-scale 3D CNN with fully connected CRF for accurate brain lesion segmentation Med Image Anal 36 61–78 [PubMed: 27865153]

Kermany DS, Goldbaum M, Cai W, Valentim CC, Liang H, Baxter SL, McKeown A, Yang G, Wu X and Yan F 2018 Identifying medical diagnoses and treatable diseases by image-based deep learning Cell 172 1122–31. e9 [PubMed: 29474911]

Kim ST, Lee J-H, Lee H and Ro YM 2018 Visually interpretable deep network for diagnosis of breast masses on mammograms Physics in Medicine & Biology 63 235025 [PubMed: 30511660]

Kiran BR, Thomas DM and Parakkal R 2018 An overview of deep learning based methods for unsupervised and semi-supervised anomaly detection in videos Journal of Imaging 4 36

Kohlbrenner M, Bauer A, Nakajima S, Binder A, Samek W and Lapuschkin S 2019 Towards best practice in explaining neural network decisions with LRP arXiv preprint arXiv:1910.09840

Krizhevsky A, Sutskever I and Hinton GE Advances in neural information processing systems,2012), vol. Series) pp 1097–105

Kumar D, Taylor GW and Wong A 2019 Discovery Radiomics With CLEAR-DR: Interpretable Computer Aided Diagnosis of Diabetic Retinopathy IEEE Access 7 25891–6

Kumar D, Wong A and Taylor GW Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops,2017), vol. Series) pp 36–44

Kurakin A, Goodfellow I and Bengio S 2016 Adversarial examples in the physical world arXiv preprint arXiv:1607.02533

LaLonde R, Torigian D and Bagci U 2019 Encoding High-Level Visual Attributes in Capsules for Explainable Medical Diagnoses arXiv preprint arXiv:1909.05926

Li S, Dong M, Du G and Mu X 2019 Attention dense-u-net for automatic breast mass segmentation in digital mammogram IEEE Access 7 59037–47

Lin M, Chen Q and Yan S 2013 Network in network arXiv preprint arXiv:1312.4400

Litjens G, Kooi T, Bejnordi BE, Setio AAA, Ciompi F, Ghafoorian M, Van Der Laak JA, Van Ginneken B and Sánchez CI 2017 A survey on deep learning in medical image analysis Medical image analysis 42 60–88 [PubMed: 28778026]

Liu F, Guan B, Zhou Z, Samsonov A, Rosas H, Lian K, Sharma R, Kanarek A, Kim J and Guermazi A 2019 Fully Automated Diagnosis of Anterior Cruciate Ligament Tears on Knee MR Images by Using Deep Learning Radiology: Artificial Intelligence 1 180091 [PubMed: 32076658]

Lévy D and Jain A 2016 Breast mass classification from mammograms using deep convolutional neural networks arXiv preprint arXiv:1612.00542

Maaten L v d and Hinton G 2008 Visualizing data using t-SNE Journal of machine learning research 9 2579–605

Mahendran A and Vedaldi A European Conference on Computer Vision,2016), vol. Series): Springer) pp 120–35

Makhzani A, Shlens J, Jaitly N, Goodfellow I and Frey B 2015 Adversarial autoencoders arXiv preprint arXiv:1511.05644

Mirsky Y, Mahler T, Shelef I and Elovici Y 2019 CT-GAN: Malicious Tampering of 3D Medical Imagery using Deep Learning arXiv preprint arXiv:1901.03597

Montavon G, Lapuschkin S, Binder A, Samek W and Müller K-R 2017 Explaining nonlinear classification decisions with deep taylor decomposition Pattern Recognition 65 211–22

Nundy S, Montgomery T and Wachter R M 2019 Promoting trust between patients and physicians in the era of artificial intelligence Jama 322 497–8 [PubMed: 31305873]

Olah C, Mordvintsev A and Schubert L 2017 Feature visualization Distill 2 e7

Oquab M, Bottou L, Laptev I and Sivic J Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2015), vol. Series) pp 685–94

Perlich C 2010 Learning curves in machine learning Encyclopedia of machine learning 577–80

Plis SM, Hjelm DR, Salakhutdinov R, Allen EA, Bockholt HJ, Long JD, Johnson HJ, Paulsen JS, Turner JA and Calhoun VD 2014 Deep learning for neuroimaging: a validation study Frontiers in neuroscience 8 229 [PubMed: 25191215]

Rayan JC, Reddy N, Kan JH, Zhang W and Annapragada A 2019 Binomial classification of pediatric elbow fractures using a deep learning multiview approach emulating radiologist decision making Radiology: Artificial Intelligence 1 e180015 [PubMed: 33937781]

Reyes M, Meier R, Pereira S, Silva CA, Dahlweid F-M, Tengg-Kobligk H v, Summers RM and Wiest R 2020 On the Interpretability of Artificial Intelligence in Radiology: Challenges and Opportunities Radiology: Artificial Intelligence 2 e190043 [PubMed: 32510054]

Ribeiro MT, Singh S and Guestrin C Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining,2016), vol. Series): ACM) pp 1135–44

Roth HR, Farag A, Lu L, Turkbey EB and Summers RM Medical Imaging 2015: Image Processing,2015a), vol. Series 9413): International Society for Optics and Photonics) p 94131G

Roth HR, Lu L, Farag A, Shin H-C, Liu J, Turkbey EB and Summers RM International conference on medical image computing and computer-assisted intervention,2015b), vol. Series): Springer) pp 556–64

Roth HR, Lu L, Liu J, Yao J, Seff A, Cherry K, Kim L and Summers RM 2016 Improving Computer-Aided Detection Using Convolutional Neural Networks and Random View Aggregation IEEE Trans Med Imaging 35 1170–81 [PubMed: 26441412]

Roth HR, Lu L, Seff A, Cherry KM, Hoffman J, Wang S, Liu J, Turkbey E and Summers RM 2014 A new 2.5D representation for lymph node detection using random sets of deep convolutional neural network observations Med Image Comput Comput Assist Interv 17 520–7 [PubMed: 25333158]

Ruderman DL 1994 The statistics of natural images Network: computation in neural systems 5 517–48

Rudin C. 2018Please stop explaining black box models for high stakes decisions. arXiv preprint arXiv:1811.10154.

Sabour S, Frosst N and Hinton GE Advances in neural information processing systems,2017), vol. Series) pp 3856–66

Samek W, Binder A, Montavon G, Lapuschkin S and Müller K-R 2016 Evaluating the visualization of what a deep neural network has learned IEEE transactions on neural networks and learning systems 28 2660–73

Sayres R, Taly A, Rahimy E, Blumer K, Coz D, Hammel N, Krause J, Narayanaswamy A, Rastegar Z, Wu D, Xu S, Barb S, Joseph A, Shumski M, Smith J, Sood AB, Corrado GS, Peng L and Webster DR 2019 Using a Deep Learning Algorithm and Integrated Gradients Explanation to Assist Grading for Diabetic Retinopathy Ophthalmology 126 552–64 [PubMed: 30553900]

Schlemper J, Oktay O, Schaap M, Heinrich M, Kainz B, Glocker B and Rueckert D 2019 Attention gated networks: Learning to leverage salient regions in medical images Medical image analysis 53 197–207 [PubMed: 30802813]

Seah JC, Tang JS, Kitchen A, Gaillard F and Dixon AF 2018 Chest radiographs in congestive heart failure: visualizing neural network learning Radiology 290 514–22 [PubMed: 30398431]

Selbst AD and Powles J 2017 Meaningful information and the right to explanation International Data Privacy Law 7 233–42

Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D and Batra D 2017 IEEE International Conference on Computer Vision (ICCV),2017), vol. Series): IEEE) pp 618–26

Shen Y and Gao M International Workshop on Machine Learning in Medical Imaging,2018), vol. Series): Springer) pp 389–97

Shin HC, Roth HR, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura D and Summers R M 2016 Deep Convolutional Neural Networks for Computer-Aided Detection: CNN Architectures, Dataset Characteristics and Transfer Learning IEEE Trans Med Imaging 35 1285–98 [PubMed: 26886976]

Shorten C and Khoshgoftaar T M 2019 A survey on image data augmentation for deep learning Journal of Big Data 6 60

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

Springenberg JT, Dosovitskiy A, Brox T and Riedmiller M 2014 Striving for simplicity: The all convolutional net arXiv preprint arXiv:1412.6806

Sun J, Darbeha F, Zaidi M and Wang B 2020 SAUNet: Shape Attentive U-Net for Interpretable Medical Image Segmentation arXiv preprint arXiv:2001.07645

Sundararajan M, Taly A and Yan Q 2017 Axiomatic attribution for deep networks arXiv preprint arXiv:1703.01365

Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I and Fergus R 2013 Intriguing properties of neural networks arXiv preprint arXiv:1312.6199

Takada K, Toyokawa G, Tagawa T, Kohashi K, Akamine T, Takamori S, Hirai F, Shoji F, Okamoto T and Oda Y 2017 Association between PD-L1 expression and metabolic activity on 18F-FDG

PET/CT in patients with small-sized lung cancer Anticancer research 37 7073–82 [PubMed: 29187498]

Theis L, Shi W, Cunningham A and Huszár F 2017 Lossy image compression with compressive autoencoders arXiv preprint arXiv:1703.00395

Thomas AW, Heekeren HR, Muller K-R and Samek W 2018), vol. Series)

Tschannen M, Bachem O and Lucic M 2018 Recent advances in autoencoder-based representation learning arXiv preprint arXiv:1812.05069

Uzunova H, Ehrhardt J, Kepp T and Handels H Medical Imaging 2019: Image Processing,2019), vol. Series 10949): International Society for Optics and Photonics) p 1094911

Van Molle P, De Strooper M, Verbelen T, Vankeirsbilck B, Simoens P and Dhoedt B 2018 Understanding and Interpreting Machine Learning in Medical Image Computing Applications: Springer) pp 115–23

Vincent P, Larochelle H, Bengio Y and Manzagol P-A Proceedings of the 25th international conference on Machine learning,2008), vol. Series) pp 1096–103

Wang S, Shi J, Ye Z, Dong D, Yu D, Zhou M, Liu Y, Gevaert O, Wang K and Zhu Y 2019 Predicting EGFR mutation status in lung adenocarcinoma on computed tomography image using deep learning European Respiratory Journal 53

Wang Y, Yu B, Wang L, Zu C, Lalush DS, Lin W, Wu X, Zhou J, Shen D and Zhou L 2018 3D conditional generative adversarial networks for high-quality PET image estimation at low dose Neuroimage 174 550–62 [PubMed: 29571715]

Wolterink JM, Dinkla AM, Savenije MH, Seevinck PR, van den Berg CA and Išgum I International Workshop on Simulation and Synthesis in Medical Imaging,2017), vol. Series): Springer) pp 14–23

Wu J, Peck D, Hsieh S, Dialani V, Lehman CD, Zhou B, Syrgkanis V, Mackey L and Patterson G Medical Imaging 2018: Computer-Aided Diagnosis,2018a), vol. Series 10575): International Society for Optics and Photonics) p 105752T

Wu J, Zhou B, Peck D, Hsieh S, Dialani V, Mackey L W and Patterson G 2018b DeepMiner: Discovering Interpretable Representations for Mammogram Classification and Explanation CoRR abs/1805.12323

Xu K, Ba J, Kiros R, Cho K, Courville A, Salakhudinov R, Zemel R and Bengio Y International conference on machine learning,2015), vol. Series) pp 2048–57

Yan Y, Kawahara J and Hamarneh G International Conference on Information Processing in Medical Imaging,2019), vol. Series): Springer) pp 793–804

Yang H, Kim J-Y, Kim H and Adhikari SP 2019 Guided soft attention network for classification of breast cancer histopathology images IEEE transactions on medical imaging 39 1306–15 [PubMed: 31634125]

Yang X, Yang JD, Hwang HP, Yu HC, Ahn S, Kim B W and You H 2018 Segmentation of liver and vessels from CT images and classification of liver segments for preoperative liver surgical planning in living donor liver transplantation Comput Methods Programs Biomed 158 41–52 [PubMed: 29544789]

Yi D, Sawyer R L, Cohn D III, Dunnmon J, Lam C, Xiao X and Rubin D 2017 Optimizing and visualizing deep learning for benign/malignant classification in breast tumors arXiv preprint arXiv:1705.06362

Yi X and Babyn P 2018 Sharpness-aware low-dose CT denoising using conditional generative adversarial network Journal of digital imaging 31 655–69 [PubMed: 29464432]

Yu Z, Tan E L, Ni D, Qin J, Chen S, Li S, Lei B and Wang T 2018 A Deep Convolutional Neural Network-Based Framework for Automatic Fetal Facial Standard Plane Recognition IEEE J Biomed Health Inform 22 874–85 [PubMed: 28534800]

Zeiler MD and Fergus R European conference on computer vision,2014), vol. Series): Springer) pp 818–33

Zhang F, Li Z, Zhang B, Du H, Wang B and Zhang X 2019 Multi-modal deep learning model for auxiliary diagnosis of Alzheimer's disease Neurocomputing 361 185–95

Zhang Z, Xie Y, Xing F, McGough M and Yang L CVPR,2017), vol. Series)

Zhao G, Zhou B, Wang K, Jiang R and Xu M MICCAI,2018), vol. Series)

Zhou B, Khosla A, Lapedriza A, Oliva A and Torralba A Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition,2016), vol. Series) pp 2921–9

Deep Learning Concepts

Activation:

a nonlinear function typically applied at each layer of a deep neural network

Backpropagation:

the process of recursively computing gradients of the model loss with respect to each weight backwards through a network to update model weights during training.

KL Divergence:

Kullback-Liebler divergence. A measure of difference between two probability distributions. $D_{KL}(P, Q) = \text{sum}(P(x)*\log(P(x)/Q(x)))$.

Natural images:

images of objects in the natural world, as opposed to man-made images or computer generated images (Ruderman, 1994). For example, the ImageNet database is comprised of natural images (Deng *et al.*, 2009).

ReLU:

Rectified Linear Unit – a commonly used activation function. $\text{ReLU}(x) = \max(0, x)$. See (Glorot *et al.*, 2011).

Superpixel:

a group of contiguous image pixels with similar intensities

Transfer learning:

a network training strategy in which initial weights are taken from a network previously trained for another task.

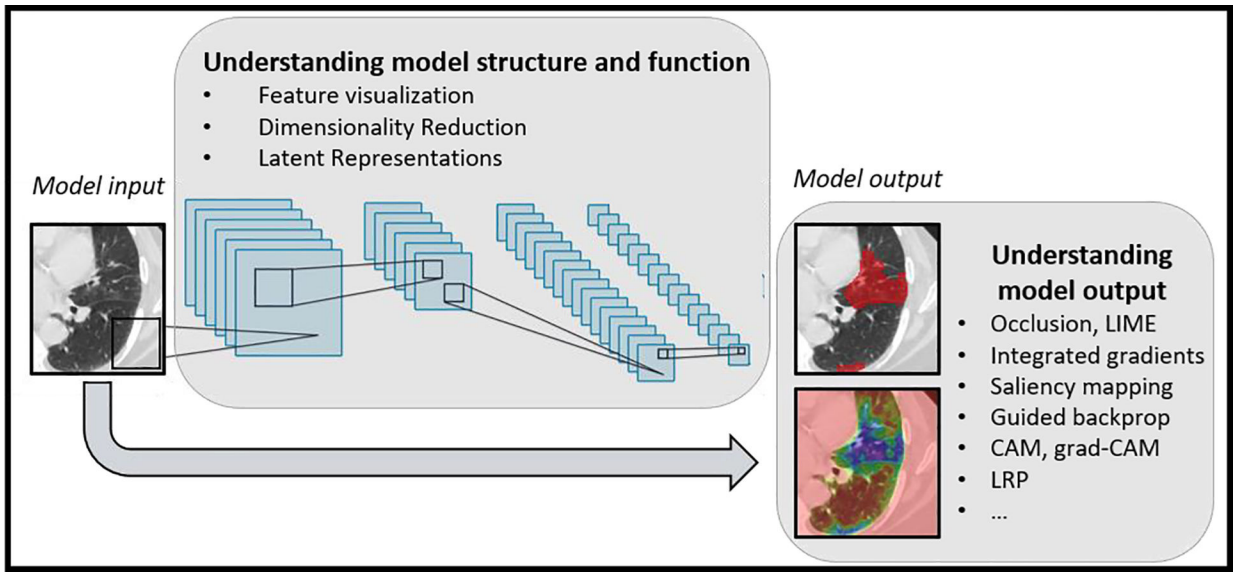

Figure 1:

Model interpretation techniques can focus either (1) on increasing understanding of internal model structure and function, or (2) on increasing understanding of model output. These two approaches to model interpretation are covered in review Section 2 and Section 3, respectively.

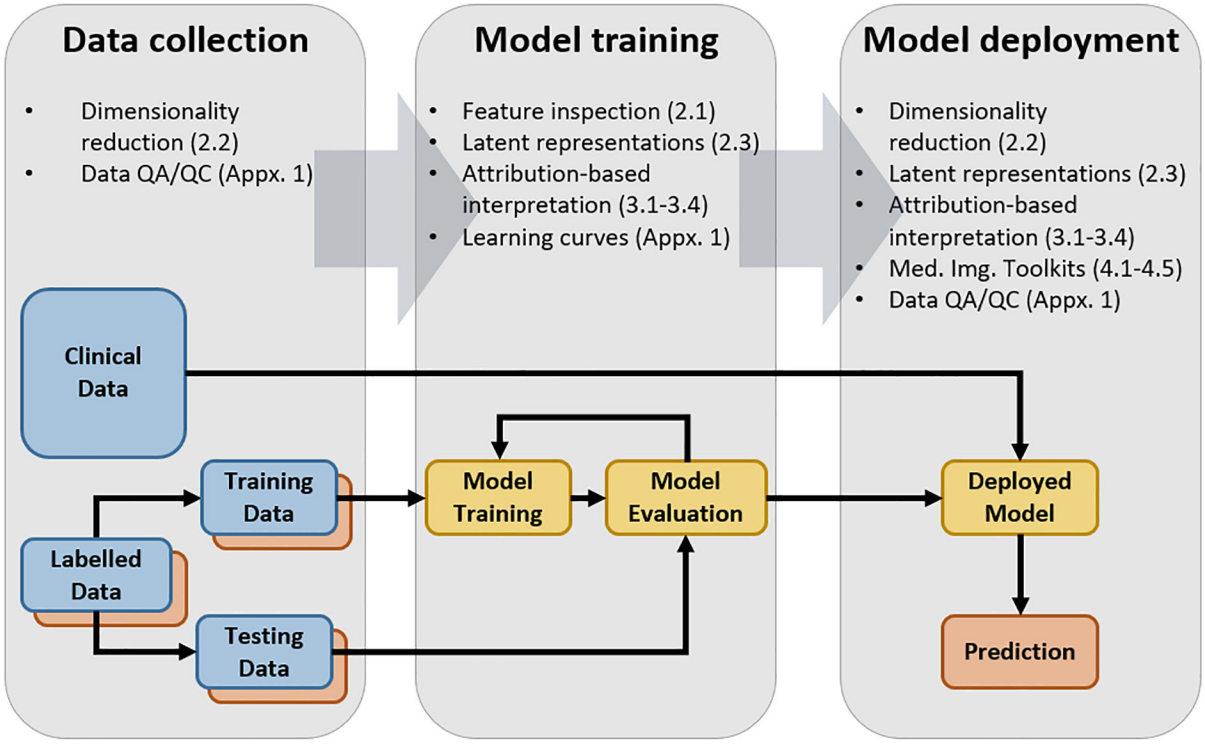

Figure 2:

Overview of the model development process (bottom), overlaid with aspects of model interpretation best suited to each phase (top), and the sections of the review in which they are covered. Different approaches to model interpretation are applicable to all steps of the model development process, from initial data collection, model training, and through to model deployment.

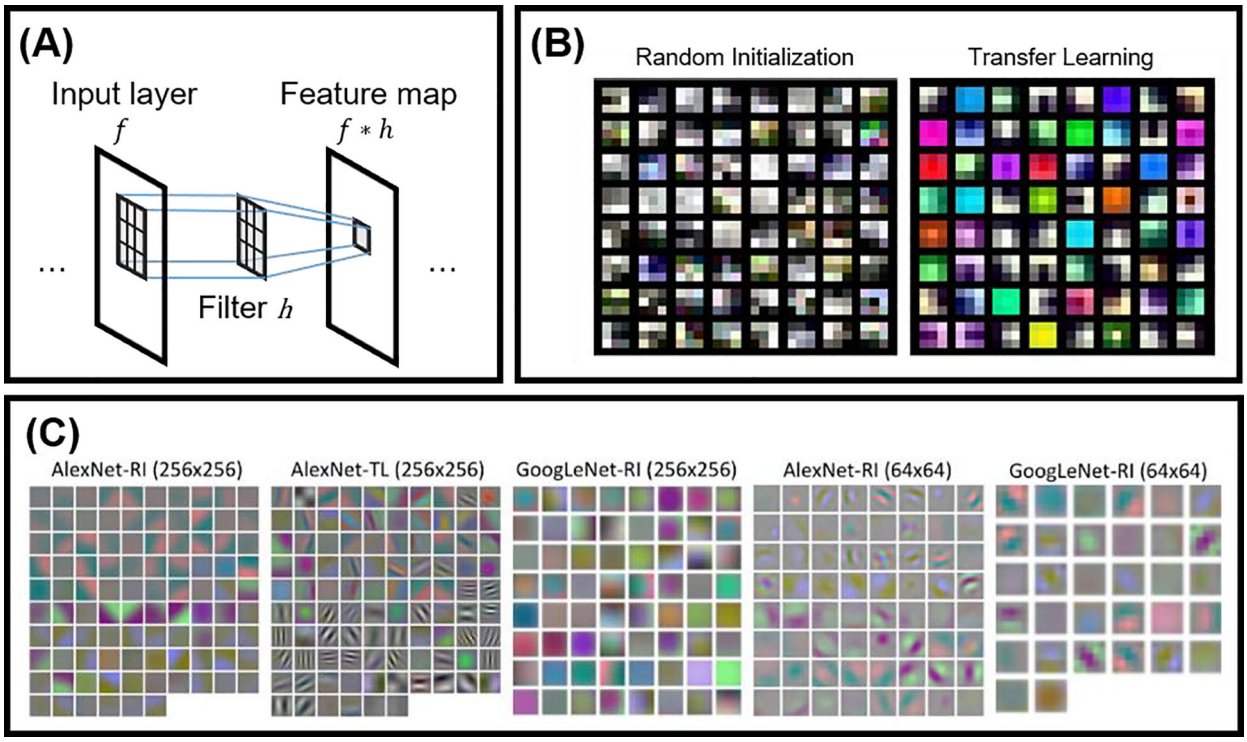

Figure 3:

The parts of a convolutional layer are shown to make clear the distinction between a convolutional filter and a feature map (A). The input layer is convolved with a set of filters and an activation function is applied to generate a feature map. Comparison of learned model filters from (Yu et al., 2018) (B) and (Shin et al., 2016) (C). Both examples show filters learned via random initialization (RI) of filters and after transfer learning (TL). Filters in (B) are of size 3×3, and filters in (C) range from size 5×5 to 11×11.

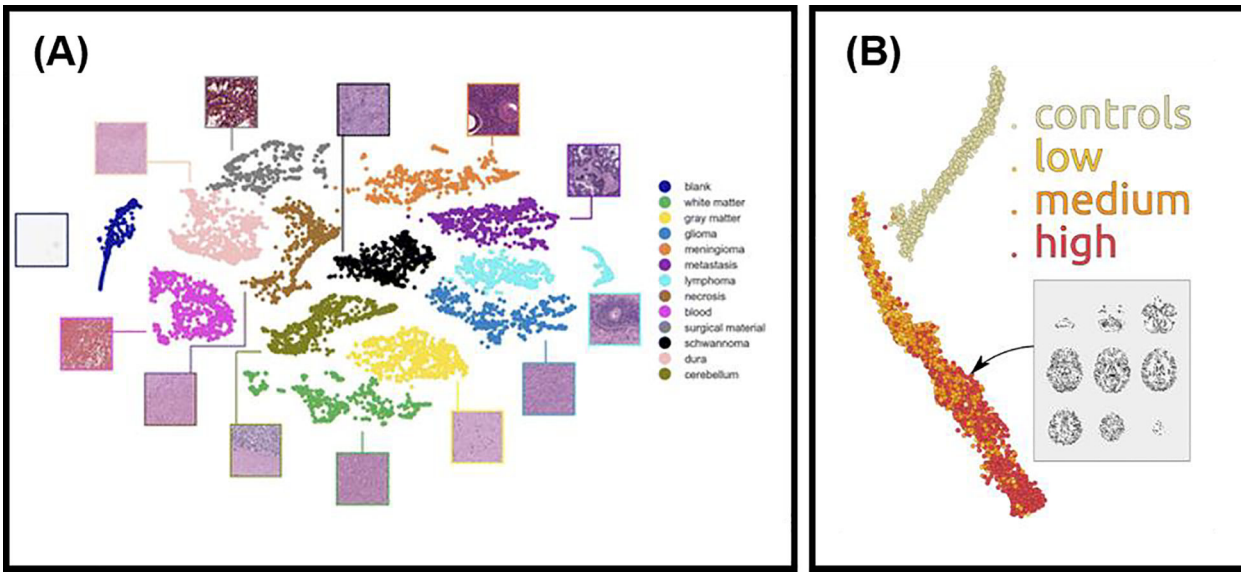

Figure 4:

Examples of visualizing high-dimensional deep learning features in a 2D projection. (A) tSNE used to classify regions of histopathology images in (Faust et al., 2018). (B) Constraint-based embedding used to visualize brain MRI images from healthy controls and patients with a varying severity of Huntington disease (Plis et al., 2014).

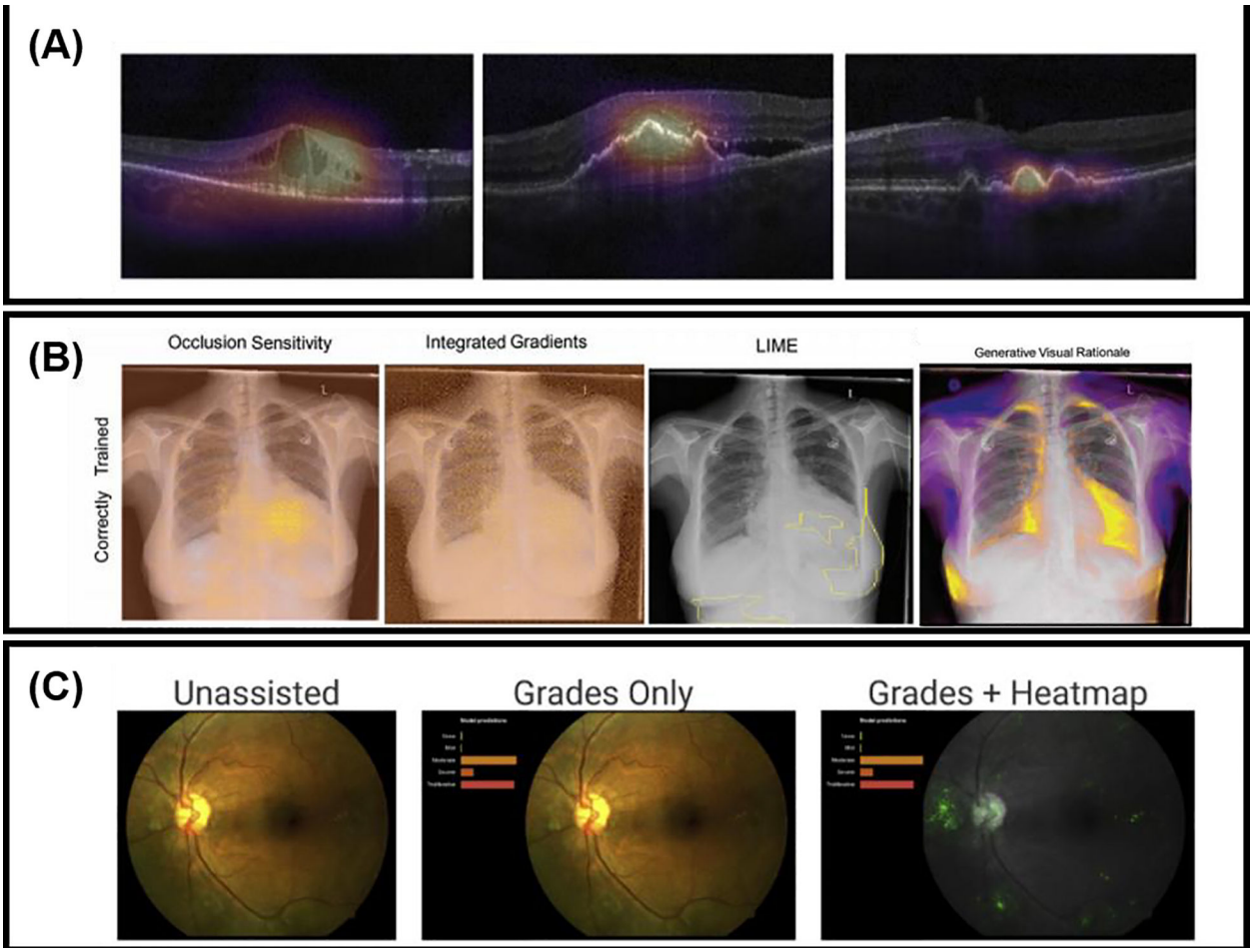

Figure 5:
Example uses of perturbation-based attribution methods for model interpretability. In (A), (Kermany et al., 2018) employed occlusion in order to visualize CNN-based diagnosis of retinal pathologies in optical coherence tomography images. (B) compares several approaches to interpretation for identifying congestive heart failure on chest x-ray (Seah et al., 2018). In (C), (Sayres et al., 2019) uses integrated gradients to visualize evidence of diabetic retinopathy on retinal fundus images.

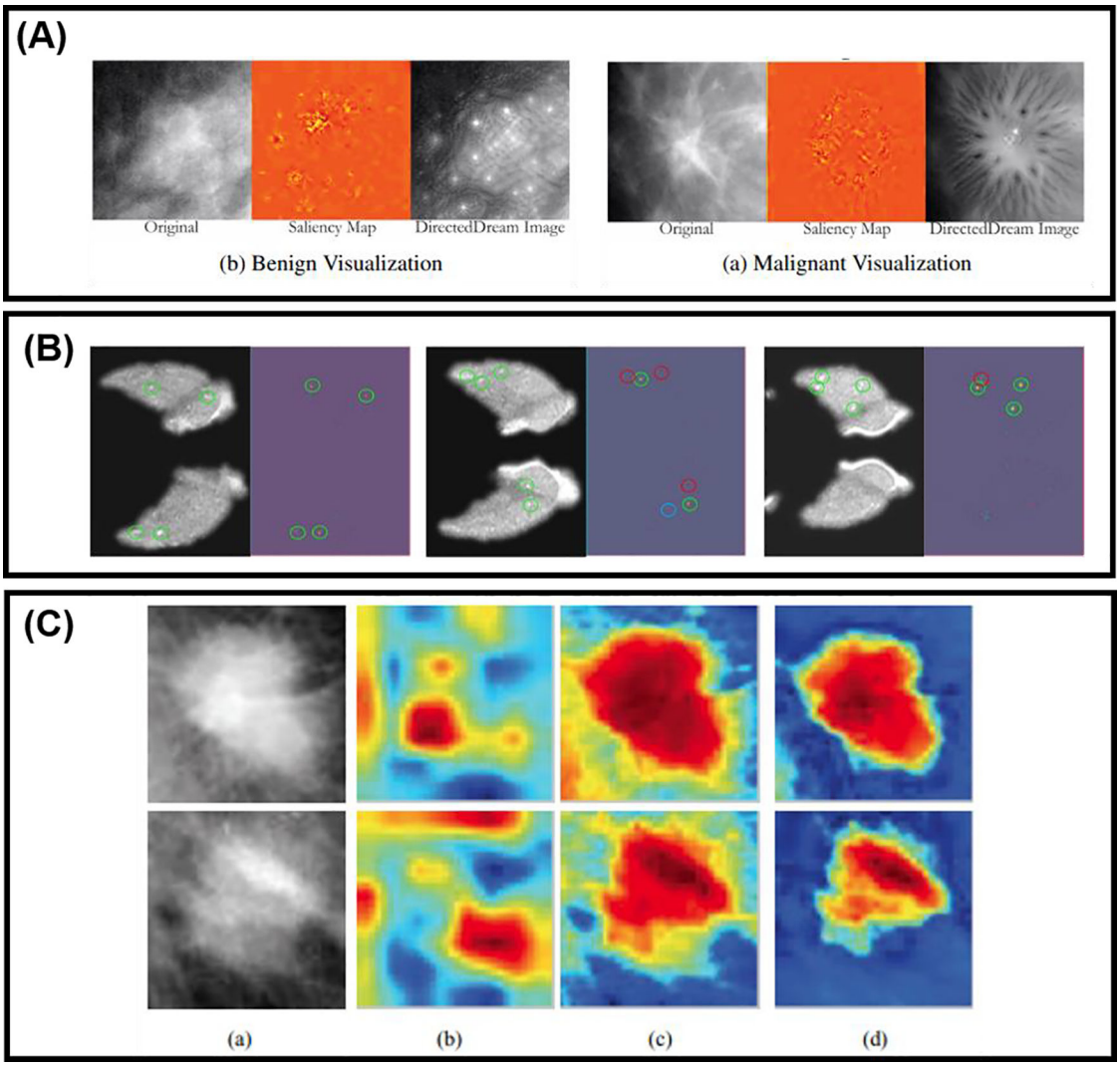

Figure 6:

Backpropagation-based approaches to model interpretation include: (A) class maximization visualization of malignant and benign breast masses on mammogram (Yi et al., 2017), (B) weakly supervised detection of extra perivascular spaces on brain MR via saliency mapping (Dubost et al., 2019), and (C) class activation map visualization for classifying breast masses on mammograms (Kim et al., 2018).

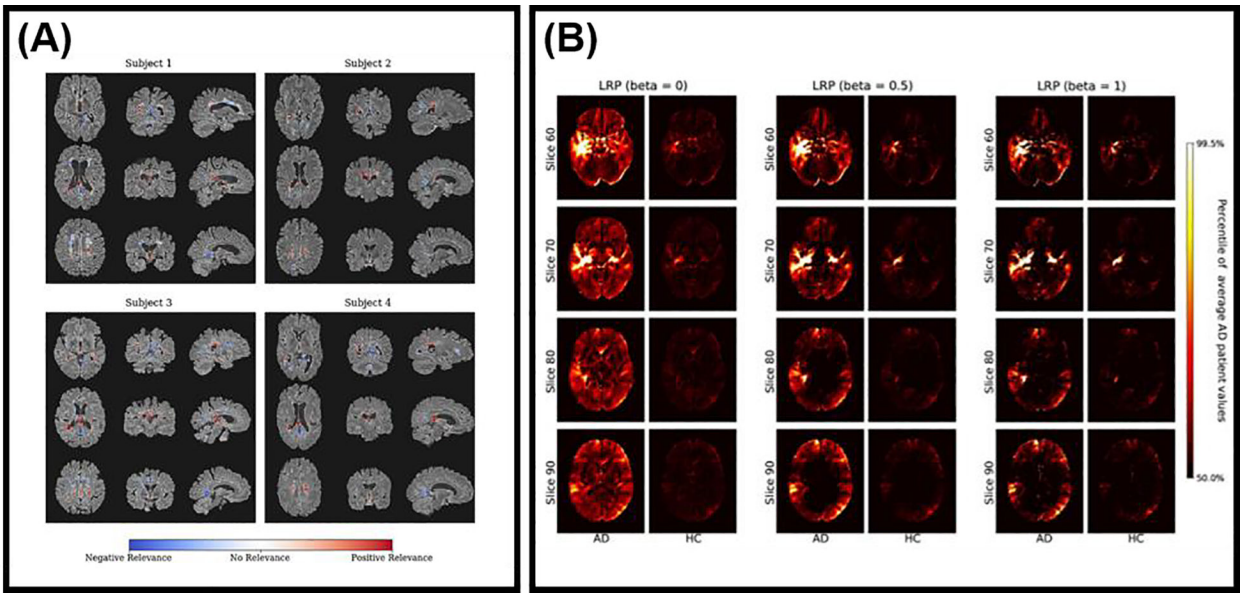

Figure 7:

Layerwise relevance propagation for model interpretation in (A) diagnosing multiple sclerosis on brain MRI (Eitel et al., 2019), and (B) visualizing evidence for Alzheimer's disease on brain MRI (Böhle et al., 2019b).

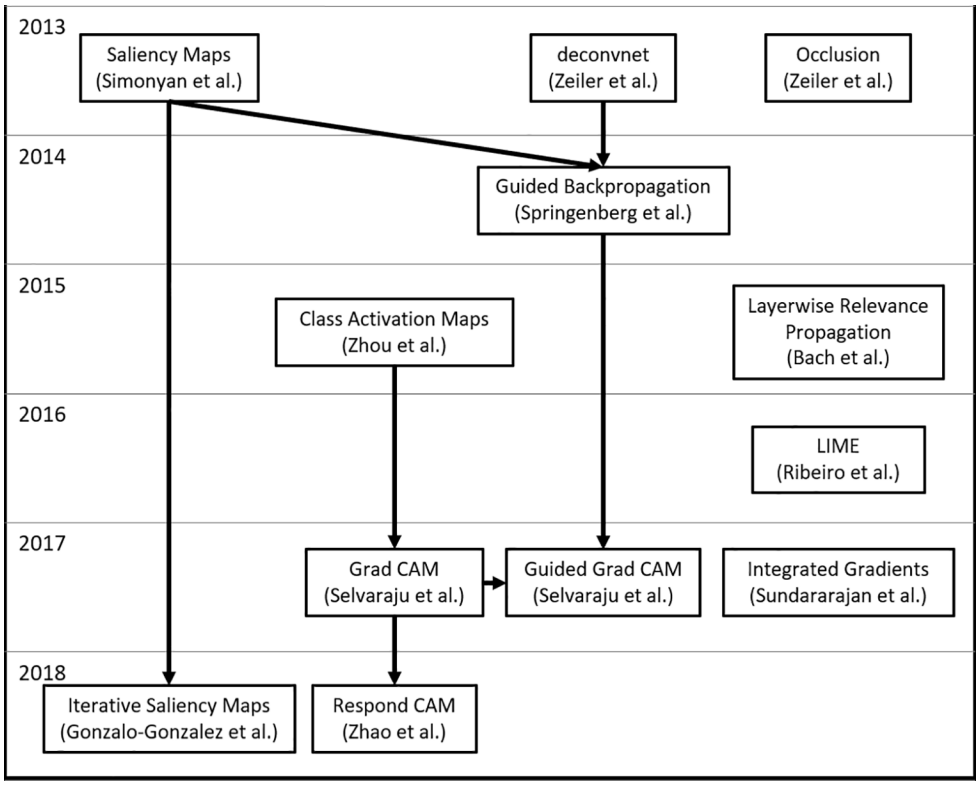

Figure 8:

Family tree of attribution-based methods for model interpretation. Dates correspond to when the method was first published. Arrows indicate methods that are developmends or refinements of previous methods.

Table 1:

Summary of uses of model interpretation methods for medical imaging tasks. Filled cells indicate which interpretation method is used by each publication. Publications that compare multiple methods have more than one filled cell in their row. Feat. Vis. = feature visualization, Dim. Red. = dimensionality reduction, AE = autoencoder, Occ. = occlusion, LIME = Local interpretable model-agnostic explanations, IG = integrated gradients, Sal. = saliency mapping, GB = guided backpropagation, CAM = class activation mapping, LRP = layer wise relevance propagation, Att. = trainable attention.

Author	Task	Modality	Feat. Vis	Dim. Red.	AE	Occ.	LIME	IG	Sal.	GB	CAM	LRP	Att.
(Yu et al., 2018)	classification	US	○	○									
(Roth et al., 2014)	detection	CT	○										
(Roth et al., 2015a)	segmentation	CT	○										
(Shin et al., 2016)	detection	CT	○										
(Van Molle et al., 2018)	classification	photo	○										
(Zhang et al., 2019)	classification	PET, MRI	○										
(Cheng and Malhi, 2017)	classification	US		○									
(Faust et al., 2018)	detection	histopath		○									
(Plis et al., 2014)	classification	MRI		○									
(Uzunova et al., 2019)	classification	OCT, MRI			○								
(Chen et al., 2020)	segmentation	MRI			○								
(Hou et al., 2019)	detection	histopath			○								
(Seah et al., 2018)	classification	x-ray				○	○	○					
(Kermany et al., 2018)	classification	OCT				○							
(Sayres et al., 2019)	classification	DR						○					
(Sundararajan et al., 2017)	classification	DR						○					
(Garg et al., 2017)	detection	MEG							○		○		
(Chen and Shi, 2018)	classification	x-ray							○				
(Dubost et al., 2019)	detection	MRI							○				
(González-Gonzalo et al., 2018)	detection	DR							○				
(Jamaludin et al., 2016)	classification	MRI							○				
(Lévy and Jain, 2016)	classification	mammo							○				
(Rayan et al., 2019)	detection	x-ray							○				
(Yi et al., 2017)	classification	mammo							○				
(Hicks et al., 2018a)	classification	colonoscopy								○	○		

Author	Task	Modality	Feat. Vis	Dim. Red.	AE	Occ.	LIME	IG	Sal.	GB	CAM	LRP	Att.
(Böhle et al., 2019a)	classification	MRI								○		○	
(Gao and Noble, 2017)	detection	US								○			
(Feng et al., 2017)	detection	CT									○		
(Gondal et al., 2017)	detection	DR									○		
(Hwang and Kim, 2016)	detection	x-ray, mammo									○		
(Kim et al., 2018)	classification	mammo									○		
(Liu et al., 2019)	detection	MRI									○		
(Shen and Gao, 2018)	detection	x-ray									○		
(Eitel et al., 2019)	classification	MRI										○	
(Thomas et al., 2018)	classification	MRI										○	
(Schlemper et al., 2019)	classification, segmentation	US, CT											○
(Li et al., 2019)	segmentation	mammo											○
(Yan et al., 2019)	classification	photo											○
(Górriz et al., 2019)	classification	x-ray											○
(Yang et al., 2019)	classification	histopath											○
(Sun et al., 2020)	segmentation	MRI											○

Phys Med Biol. Author manuscript; available in PMC 2021 June 27.