Medical Image Analysis 81 (2022) 102551

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

Interpretability-Guided Inductive Bias For Deep Learning Based Medical Image

Dwarikanath Mahapatra [a,*], Alexander Poellinger [b,c], Mauricio Reyes [d]

[a] Inception Institute of AI, Abu Dhabi, United Arab Emirates
[b] Department of Diagnostic, Interventional and Pediatric Radiology, Inselspital, Bern University Hospital, Bern, Switzerland
[c] University of Bern, Switzerland
[d] ARTORG Center for Biomedical Engineering Research, University of Bern, Bern, Switzerland

ARTICLE INFO

Keywords:
Interpretability
Inductive bias
Medical image classification
Medical image segmentation

ABSTRACT

Deep learning methods provide state of the art performance for supervised learning based medical image analysis. However it is essential that trained models extract clinically relevant features for downstream tasks as, otherwise, shortcut learning and generalization issues can occur. Furthermore in the medical field, trustability and transparency of current deep learning systems is a much desired property. In this paper we propose an interpretability-guided inductive bias approach enforcing that learned features yield more distinctive and spatially consistent saliency maps for different class labels of trained models, leading to improved model performance. We achieve our objectives by incorporating a class-distinctiveness loss and a spatial-consistency regularization loss term. Experimental results for medical image classification and segmentation tasks show our proposed approach outperforms conventional methods, while yielding saliency maps in higher agreement with clinical experts. Additionally, we show how information from unlabeled images can be used to further boost performance. In summary, the proposed approach is modular, applicable to existing network architectures used for medical imaging applications, and yields improved learning rates, model robustness, and model interpretability.

1. Introduction

Deep learning (DL) is highly effective in medical image analysis and has shown state-of-the-art performance on a wide variety of tasks such as disease classification, segmentation, localization, etc. Liu et al. (2019); Litjens et al. (2017); Aggarwal et al. (2021). One important factor in guaranteeing high performance of DL models is the availability of large curated datasets. For medical imaging applications, having access to large collections of imaging datasets is a true challenge due to diversity of protocols, vendors, inter-rater variability, data protection and heterogeneous data governance regulations, etc. In order to address this challenge, different approaches have been proposed, including methods such as data augmentation, active learning, semi-supervised learning, and self supervised learning. Common to all these approaches, it is essential that trained models extract clinically relevant features for the downstream tasks. This includes modelling and incorporating appropriate inductive biases (i.e., set of assumptions used by a learner to predict outputs) (Griffiths et al., 2010; Hessel et al., 2019; Goyal and Bengio, 2020) in order to avoid spurious correlations leading to shortcut learning (Geirhos et al., 2020). Shortcuts stem from spurious correlations as deep features that perform well on standard benchmarks but fail to generalize in real world scenarios (e.g. (DeGrave et al., 2021)). A robust and effective inductive bias reduces shortcut learning by injecting knowledge about desired properties of a model and its outputs. This can be done at different levels by considering model architecture, training data selection, training cost functions, and model optimization, as described below.

Recent works have used attention mechanisms aiming at an inductive bias that employs channel-wise attention mechanisms such as Squeeze and Excitation networks (SENet) (Hu et al., 2018), or spatially-based global attention (Bello et al., 2019; Woo et al., 2018a). SENet learns channel-wise relationships and proposes a novel architectural unit that adaptively recalibrates channel-wise feature responses by explicitly modelling interdependencies among channels. In (Bello et al., 2019), the authors use self-attention for discriminative visual tasks as an alternative to convolutions, which have been very successful in medical

* Corresponding author.
E-mail address: dwarikanath.mahapatra@inceptioniai.org (D. Mahapatra).

https://doi.org/10.1016/j.media.2022.102551
Received 17 December 2021; Received in revised form 5 July 2022; Accepted 13 July 2022
Available online 22 July 2022

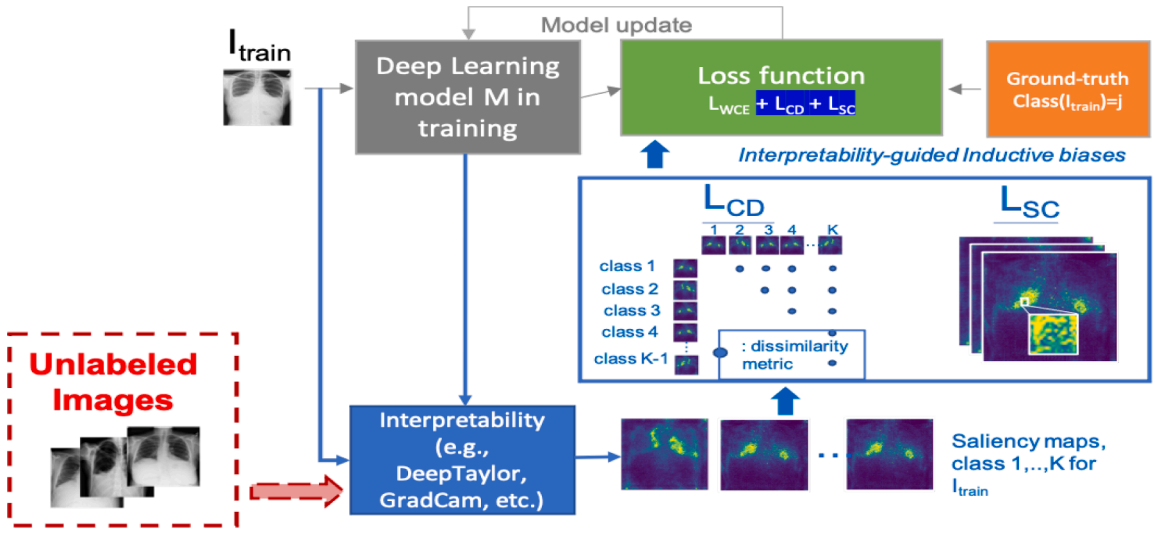

Fig. 1. Proposed pipeline for interpretability-guided inductive bias, denoted as *SIBNet* (Saliency Inductive Bias Network). Given a training image I_{train}, and model M, saliency maps describing current model's interpretation for each class label are used to yield class-distinctiveness (L_{CD}) and a spatial coherent loss terms (L_{SC}). These terms can be used in conjunction with existing loss terms (e.g. Weighted Cross-Entropy L_{WCE}), to inject desired properties of the saliency maps for improved model performance and interpretability. Gray- and blue-coloured components describe the standard and new proposed learning pipeline, respectively. Our method is, and when combined with the unsupervised information from unlabeled images (red-dotted line) it is denoted as *SIBNet* + . Note: Illustration for the case of image classification. For image segmentation tasks the same principle applies, with the difference that class labels for ground-truth data (orange block) are used in this case as a proxy classification task to yield improved learned segmentation features (see Section 3.3 for more details). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

image applications. In this work, a novel two-dimensional relative self-attention mechanism is used that reliably replaces convolutions for image classification. In (Woo et al., 2018a), a Convolutional Block Attention Module (CBAM) for feed-forward convolutional neural networks was proposed. For an intermediate feature map, CBAM sequentially infers attention maps along the channel and spatial dimensions separately. The attention maps are then multiplied to the input feature map for an adaptive feature refinement. Previous self-attention methods require modification of existing architectures to include additional attention layers (or build on fully attention models (Ramachandran et al., 2019)), consequently one cannot utilize existing pre-trained models that have been well studied and proven to be accurate for different tasks. Indeed, many studies have shown that transfer learning for medical image analysis applications using pre-trained networks on other datasets performs well compared to models trained from scratch (Tajbakhsh et al., 2016; Weatheritt et al., 2020). Other attention based approaches include self-supervised contrastive learning (Chen et al., 2020), where targeted data augmentations along with a contrastive loss is used to inform the model on variations in the data that should not be considered for feature learning. However, contrastive learning assumes implicit knowledge of downstream task invariances (Chen et al., 2020), which can be challenging to design for medical applications.

The importance of utilizing effective inductive biases is exacerbated in the field of medical image analysis due to the typically low sample size of training datasets. We propose an interpretability-guided approach that incorporates an inductive bias for medical image analysis tasks in order to simultaneously improve model performance, its robustness and interpretability. The proposed approach, coined hereafter as SIBNet, for Salient Inductive Bias Network, is complementary to previously proposed attention-based approaches; it leverages findings from the area of interpretability (Reyes et al., 2020; McCrindle et al., 2021; Kitamura and Marques, 2021; Fuhrman et al., 2022), and is motivated by the following observation: A trained radiologist learns to perform differential diagnosis on medical images based on disease-specific image patterns or characteristics. Consequently, during model training we propose to incorporate a novel inductive bias in the loss term of the model such that learned features yield more class-distinctive and spatially coherent interpretability saliency maps.

The proposed interpretability-guided inductive bias acts directly on the loss function of the model being trained, it is modular and easy to implement, and can be utilized in conjunction with other existing loss functions, as well as on existing classification model architectures without modifications.

We show further benefits of the proposed approach by using information from unlabeled datasets, which is otherwise not possible for attention-based approaches that necessitate ground-truth information to

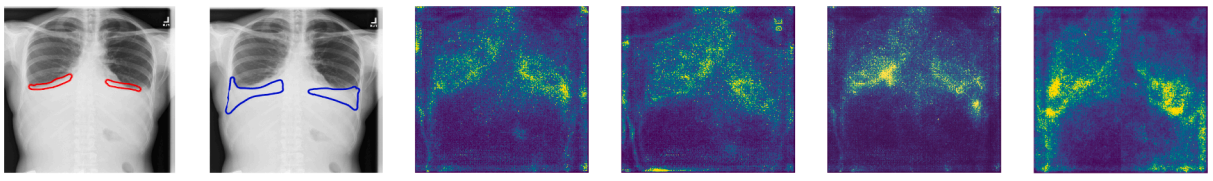

Fig. 2. Motivation and seminal observation of SIBNet. Original image with expert annotated regions corresponding to (a) Pleural Effusion; (b) Atelectasis. Saliency maps (Deep Taylor) of individual classifiers for (c) Pleural Effusion ($AUC = 0.922$) and (d) Atelectasis ($AUC = 0.848$). Saliency maps obtained with joint classifier for (e) Pleural Effusion ($AUC = 0.939$) and (f) Atelectasis ($AUC = 0.869$). Along with improved performance, it can be observed that in comparison to Figures (c) and (d), saliency maps generated from a model classifying both conditions are more distinctive, more spatially coherent, and more in line with the expected areas of radiological interest describing each condition.

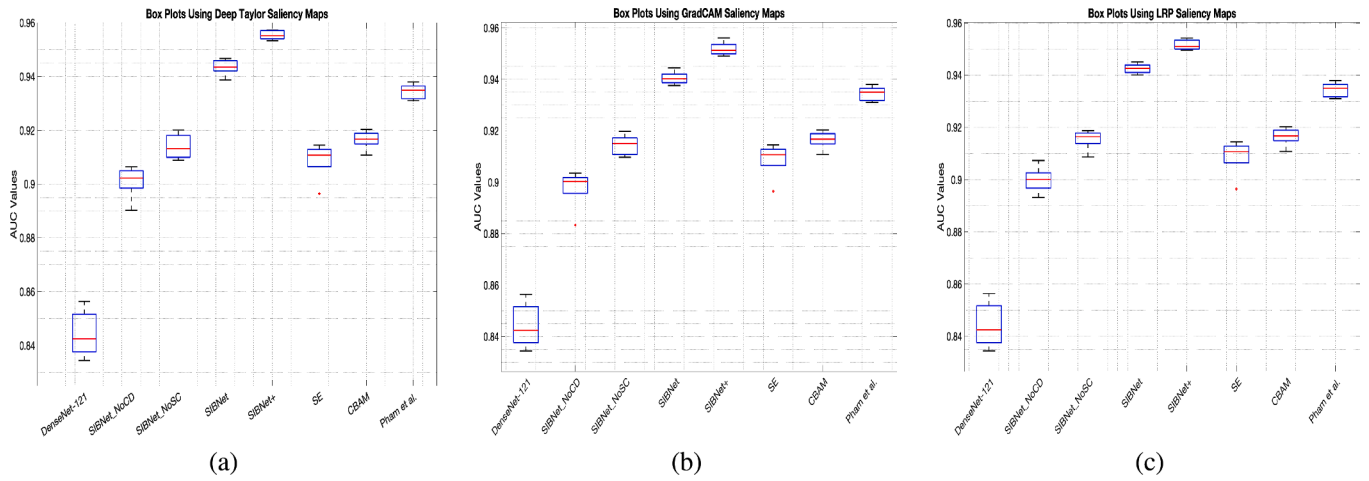

Fig. 3. (a) Mean AUC from 5-fold validation for all 5 labels using: (a) Deep Taylor Saliency Maps values of $\lambda_1 = 0.9; \lambda_2 = 1.1$; (b) GradCAM based saliency maps. The values of $\lambda_1 = 1.4; \lambda_2 = 1.0$ are used; (c) LRP saliency maps (Alber et al., 2019), values of $\lambda_1 = 1.1; \lambda_2 = 0.8$. Results are shown for the competing methods and ablation studies.

back propagate gradients. We present results on multi-class classification and medical image segmentation, comparing the proposed approach to different state-of-the-art methods, as well as several ablation studies conducted to demonstrate the added value of the proposed approach in terms of model performance, robustness and improved model training. We also show quantitatively how the proposed approach yields improved saliency maps in better agreement with expert-annotated saliency maps. As means to present and evaluate the new proposed interpretability-guided inductive bias, we demonstrate its effectiveness on publicly available datasets for chest X-ray classification and histopathology segmentation.

1.1. Summary of the proposed method

Fig. 1 shows the main pipeline of the proposed approach. Given a training set, a model is trained with a combination of standard classification loss(es) (e.g., weighted cross-entropy) and the proposed interpretability-guided inductive bias terms. As the training process evolves, learned features yield more class-distinctive and spatially coherent saliency maps, which enforces class separability, and hence expected improved performance, as well as an enhanced spatial attention to the class label describing the targeted medical condition/disease.

In the following we describe the basic motivation and observation leading to the design of SIBNet, as well as a summary of its main components.

Motivation Fig. 2(a,b) show a patient image with expert annotated regions corresponding to pleural effusion and atelectasis, two typically diagnosed lung conditions, along the AUROC values for different trained models. Fig. 2(c,d) show the corresponding saliency maps obtained on classification models trained to detect only pleural effusion and atelectasis, respectively. In contrast, Fig. 2(e,f) show saliency maps generated from a model classifying both conditions. It can be observed that in comparison to Fig. 2(c) and (d), saliency maps generated from a model classifying both conditions are more distinctive, more spatially coherent, and more in line with the expected areas of radiological interest describing each condition. Based on this initial observation we hypothesized whether improved model performance and interpretability could be attained through an inductive bias reflecting enhanced distinctiveness and spatial consistency of saliency maps.

Deep learning model We note that any deep learning model can be used with the proposed interpretability-guided inductive bias. For the selected case of lung disease classification we trained a DenseNet-121 architecture due to its known good performance on this task (Irvin et al., 2019; Rajpurkar et al., 2017). Similarly, for medical image segmentation we trained the popular U-Net (Ronneberger et al., 2015) architecture (see Section 3.3). Given a set of N labeled training images $\{(x_i, y_i) : 1 \leq i \leq N\}, x_i \in \mathbb{R}^d$, being the training images, and $y_i \in \{1, \ldots, K\}$, the corresponding class labels, a deep learning model M is commonly updated by minimizing a standard loss term, such as the weighted cross entropy loss (L_{WCE}). We note that the modularity of the approach enables utilization of other existing loss functions, as illustrated in the loss function block of Fig. 1.

Interpretability saliency maps Saliency maps are a popular interpretability approach developed initially for natural images, and later used for interpretability of deep learning models in medical image applications (Cardoso et al., 2020; Reyes et al., 2020; Fuhrman et al., 2022; Budd et al., 2021; Kitamura and Marques, 2021; McCrindle et al., 2021; Mahapatra et al., 2022; 2021a). Saliency maps have been used for many medical image analysis application such as image registration (Mahapatra and Sun, 2011; 2008), joint registration and segmentation (Mahapatra and Sun, 2012; 2010), active learning (Mahapatra and Buhmann, 2015), image quality assessment (Mahapatra et al., 2016), medical image super resolution (Mahapatra et al., 2017). In this study we selected DeepTaylor decomposition (Montavon et al., 2017) due to its popularity and previous uses in other medical image applications (Silva et al., 2020; Mahapatra et al., 2021b; Eitel et al., 2019). DeepTaylor is a method to explain neural network's predictions in terms of input variables. It operates by running a backward pass on the network in order to produce a decomposition of the neural network's output on the input variables. Each neuron of a deep network is viewed as a function that can be expanded and decomposed on its input variables. The decompositions of multiple neurons are then aggregated or propagated backwards, resulting in a saliency map (e.g. Fig. 2).

Given an input image I, and model M, a saliency map $S_{I,c} \in \mathbb{R}^d$ identifies relevant regions of interest in I to be classified as label c. The proposed inductive bias aims at enhancing the distinctiveness of saliency maps $S_{I,c=i}$ and $S_{I,c=j}(j \neq i)(\forall i.j \in \{1, \ldots, K\})$, as well as its spatial coherence to identify the area of interest used by model M. We note that the modularity of the approach enables utilization of other interpretability approaches without loss of generalization. In the results section we also show results obtained employing GradCAM (Selvaraju et al., 2017) in order to illustrate this point (see Fig. 3, and supplementary).

2. Methods

Given an input image I, and model M, a saliency map $S_{I,c} \in \mathbb{R}^d$ identifies relevant regions of interest in I to be classified as label c. The proposed inductive bias aims at enhancing the distinctiveness between

saliency maps $S_{I,c=i}$ and $S_{I,c=j}(j \neq i)(\forall i,j \in \{1,\ldots,K\})$, as well as its spatial coherence to effectively identify the area of interest used by model M to perform the task. In the next section we describe how these two properties are modelled within the loss term.

2.1. Interpretability-guided inductive bias: loss terms

2.1.1. Class distinctiveness

Given training image I and model M, we yield during model training the set of saliency maps $\{S_{I,c}\}_{c=1}^{K}$ identifying map explanations for each class c. Deep latent representations has been effectively used as an image perception similarity metric (Zhang et al., 2018), as well as in combination with interpretability methods for image retrieval (Silva et al., 2020). Consequently, we calculate the corresponding latent representations of the saliency maps $\{Z_{S_{I,c}}\}_{c=1}^{K}$ from the second to last layer of the current classification model. The latent representation vectors Z hence encode the *current understanding or perception* of model M to the calculated saliency maps.

In order to enhance distinctiveness of saliency maps for different classes, we calculate the following class distinctiveness loss term (L_{CD}), as follows:

$$L_{CD} = \frac{2}{K(K-1)} \sum_{c_1=1}^{K-1} \sum_{c_2=c_1+1}^{K} cosine\_similarity\left(Z_{S_{I,c_1}}, Z_{S_{I,c_2}}\right), \quad (1)$$

where *cosine_similarity(·)* is the cosine similarity metric used to compare latent representations (Zhang et al., 2018; Silva et al., 2020), with values ranging from 0 (i.e. maximum dissimilarity) and 1 (i.e. minimum dissimilarity). We note that other similarity metrics could be used here. The idea is to ensure that the latent representations Z for each class are as dissimilar as possible. Since the objective function aims to minimize the overall loss, the cosine similarity among different latent representations Z needs to be minimized towards zero.

Eq. (1) therefore enforces distinctiveness of the different K saliency maps for each label class generated by the model, promoting with this distinctiveness of learned features.

2.1.2. Spatial coherence

The spatial coherence loss term complements the class distinctiveness loss term, and aims at regularizing the spatial distribution of saliency maps. From observations, and as shown in the example in Fig. 2, saliency maps tend to be disperse and not spatially consistent in relation to expert-annotated saliency maps identifying regions of interest used to perform clinical diagnosis. We propose a spatial coherence loss term (L_{SC}), as follows:

$$L_{SC} = \sum_{p} \sum_{n_p \in \mathcal{N}(p)} \| x_p - x_{n_p} \|^2, \quad (2)$$

where x_p is the pixel intensity in saliency map $S_{I,c}$ and x_{n_p} is the set of pixels in the 9×9 neighborhood $\mathcal{N}(p)$ belonging to the same cluster as p. A large neighborhood size will lead to merging pixels from different regions and increase the computational complexity, while too small a neighborhood size does not provide adequate context information. We explore different neighborhood sizes ranging from 3×3 to 15×15, and found that 9×9 neighborhood size provides the best trade-off between computation time and accuracy. We identify clusters on each saliency map by connected component analysis. We note that Eq. (2) does not impose a pixel-wise local smoothing effect (as in a total-variation loss term used in denoising), but its cluster analysis penalizes the presence of spurious clusters.

We present in the result section a quantitative evaluation showing how these proposed terms yield saliency maps in better agreement with an expert radiologist - who manually annotated regions of pixel attribution- than other benchmarked models.

The total loss is then defined as:

$$L_{Total} = L_{WCE} + \lambda_1 L_{CD} + \lambda_2 L_{SC} \quad (3)$$

In the results section we present a sensitivity analysis of parameters λ_1 and λ_2, and describe their role through ablation experiments.

2.1.3. Leveraging inductive bias from unlabeled data

Due to its design the proposed SIBNet approach comes with an implicit benefit, as it can be utilized on unsupervised data. As described above, saliency maps $\{S_{I,c}\}_{c=1}^{K}$ can be calculated to yield map explanations *for each class c*, and independently of the actual true class of the interpreted sample. This is possible since interpretability approaches, such as the ones investigated here, do not rely on knowing class labels to generate saliency maps. This enables us to employ SIBNet on unsupervised data to further promote learning of features leading to distinctive and spatially coherent saliency maps.

Given a set of unlabeled images, we generate their saliency maps and their corresponding latent representations $\mathbf{Z}_{S_{I,c}}$ in the same manner as described above. Similar to Eqs. (1) and (2), we calculate the class distinctiveness and spatial coherence loss terms, which are combined into Eq. (3). In this manner we highlight that the proposed approach enables supervised and unsupervised learning. In the results section we refer to SIBNet+ to models trained with additional unlabeled data.

2.1.4. SIBNet for segmentation tasks

For deep learning based medical image segmentation the common choice for the loss function is a combination of cross entropy and Dice loss (Isensee et al., 2021). For segmentation there are no explicit methods to calculate saliency maps. However, we propose to incorporate inductive bias into segmentation problems using the following rationale, which is also used in self-supervised learning: Enhancing class distinctiveness of structures being segmented (e.g, segmenting tumoral and benign cells), leads to improved segmentation performance. We list below the steps to implement this extension to segmentation tasks.

1. A separate classification model (e.g. DenseNet-121) is trained to predict class labels of training images. From this trained model, saliency maps for each class label can be computed and corresponding latent representation vectors are extracted by forward passing saliency maps till the second to last layer of the classification model. The proposed loss terms L_{CD} and L_{SC} are then calculated, in addition to the weighted cross entropy.
2. The encoder block of the trained classification model is used as a pre-trained encoder for a UNet model. This step enables guidance of the U-Net's encoder via the proposed SIBNet inductive bias.
3. The decoder block of the UNet is then initialized with random weights, and its training is conducted. The weights of the encoder section are frozen without any updates. We denote this network as $UNet_{SIBNet}$, to refer to a UNet with pre-trained encoder using our SIBNet loss functions.

In our experiments we used a combination of weighted cross entropy and Dice loss.

2.2. Implementation details

Our SIBNet method was implemented in PyTorch. For classification, we trained DenseNet-121 models (Huang et al., 2016), although we note that other networks can also be used. For segmentation tasks we used DenseNet-121 for the classification model, as described above in Section 2.1.4, and the U-Net architecture (Ronneberger et al., 2015) for the segmentation model.

We used Adam optimizer (Kingma and Ba, 2014) with $\beta_1 = 0.93$, $\beta_2 = 0.999$, and batch normalization, with learning rate of 10^{-3}, and 10^5 update iterations, and early stopping based on the validation accuracy. In all experiments with DeepTaylor $\lambda_1 = 1.2, \lambda_2 = 0.9$.

For the generation of interpretability saliency maps, we used default

parameters of the iNNvestigate implementation of DeepTaylor (Alber et al., 2019). In the results section we also present results with two other interpretability approaches, GradCAM (Selvaraju et al., 2017), and LRP (Bach et al., 2015) with default parameters. In this study we selected DeepTaylor decomposition (Montavon et al., 2017) due to its popularity and previous uses in other medical image applications (Eitel et al., 2019; Silva et al., 2020; Mahapatra et al., 2021b). DeepTaylor is a method to explain neural network's predictions in terms of input variables. It operates by running a backward pass on the network in order to produce a decomposition of the neural network output on the input variables. Each neuron of a deep network is viewed as a function that can be expanded and decomposed on its input variables. The decompositions of multiple neurons are then aggregated or propagated backwards, resulting in a saliency map.

Training and test was performed on a NVIDIA Titan X GPU having 12 GB RAM. The input images were all of size 320 × 320 pixels. Training the baseline DenseNet-121 with L_{WCE} for 50 epochs took 13 h, and for our method it took 14.5 h (extra 11.5% time). All the reported results are on the test set and an average of 3 different runs to ensure a fair and robust assessment.

3. Results

We demonstrate the benefits of SIBNet on two common applications for medical image analysis - classification and segmentation. For classification we employed chest X-rays (a multi-class classification problem) and for segmentation we employed digital histopathology interstinal glandular images of colorectal adenocarcinoma patients (a multi-object segmentation problem).

In the following sections: 1) we first describe the different datasets and the evaluation metrics used; 2) we present classification results for the CheXpert and NIH chest Xray dataset; 3) segmentation results on histopathology images are then presented; 4) we show results of multiple experiments to determine the importance of saliency maps; and 5) also report results of ablation studies.

3.1. Dataset description and evaluation details

Classification dataset We used the CheXpert dataset (Irvin et al., 2019) consisting of 224,316 chest radiographs of 65,240 patients labeled for the presence of common chest conditions. The complete training set has 223,414 images (i.e, for all 14 class labels), while validation and test set have 200 and 500 images, respectively. We adopt a five-fold validation strategy using a 80/20 ratio for validation and testing. The validation ground-truth was obtained using majority voting from annotations of 3 board-certified radiologists. Test images were labeled by consensus of 5 board-certified radiologists. The test set evaluation protocol, as designed by the dataset creators, is based on 5 disease labels: *Atelectasis, Cardiomegaly, Consolidation, Edema*, and *Pleural Effusion*, which were selected in order to compare to (Pham et al., 2020), which is the 2nd ranked method in the CheXpert challenge (same AUC as 1st ranked approach and with an available implementation). Since the validation and test sets have images from 5 disease labels we trained our method on images from the same labels, using 9000 training images for each of the 5 labels.

Furthermore, in order to evaluate the saliency maps yielded by every benchmarked model, we asked a lung radiologist with over 15 years of experience to manually annotate salient regions describing diagnosed conditions on a subset of 25 randomly selected cases. Additionally, in order to show the ability of the proposed approach to use unlabeled data, we present results obtained when incorporating an additional set of 15,450 unlabeled images (3090 images from each class) from the NIH dataset (Wang et al., 2017). To set values for λ_1, λ_2 we performed an exhaustive grid search by varying the values of λ_1, λ_2 in the range of [0, 2] in steps of 0.05 and use the combination with the best performance on a separate dataset of 10,000 images (refer to Section 3.5.3).

Segmentation dataset For histopathology segmentation we used the public GLAS digital histopathology image dataset (Sirinukunwattana et al., 2017) that has manual segmentation maps of glands in 165 *H&E* stained images derived from 16 histological sections from different patients with stage *T*3 or *T*4 colorectal adenocarcinoma. The slides were digitized with a Zeiss MIRAX MIDI Slide Scanner having pixel resolution of 0.465 μm. The WSIs were rescaled to a pixel resolution of 0.620 μm (equivalent to 20× magnification).

A total of 52 visual fields from malignant and benign areas from the WSIs were selected by the challenge organizers to cover a wide variety of tissues. An expert pathologist graded each visual field as either "benign" or "malignant". Further details of the dataset can be found in Sirinukunwattana et al. (2017).

3.1.1. Comparison methods and ablation experiments

Classification results are shown for the following methods:

1. Our proposed method SIBNet: **S**alient **I**nductive **B**ias **Net**work, which includes all loss terms, as in Eq. (3).
2. Our first comparison is the method by Pham et al. (2020), which is the 2nd ranked method in the CheXpert challenge (same AUC as 1st ranked approach and with an available implementation), and uses directed acyclic graphs (DAGs) to learn the relationship between diseases for improved performance.
3. Our second comparison method is Squeeze and Excitation (SE) (Hu et al., 2018), which uses a channel-wise attention mechanism.
4. The third method is CBAM (Convolutional Block Attention module) (Woo et al., 2018b), which combines channel-, and spatial-wise attention mechanisms.

The above-mentioned approaches are based on attention mechanisms. In addition, we include further baselines and ablated variations of our proposed approaches:

1. *DenseNet*: uses only L_{WCE} for training;
2. SIBNet$_{NoL_{SC}}$: Excludes the spatial coherence term in Eq. (2);
3. SIBNet$_{NoL_{CD}}$: Excludes the class distinctiveness term of Eq. (1).
4. SIBNet+: Same as SIBNet but utilizing 15,450 unlabeled datasets from the NIH dataset, as mentioned above.
5. SIBNet-GradCAM & SIBNet-LRP : Same as SIBNet but with GradCAM (Selvaraju et al., 2017) or LRP (Layer-wise Relevance Propagation) (Bach et al., 2015) for saliency maps. This experiment aims at showing the modularity and effectiveness of SIBNet using alternative interpretability saliency methods.

Evaluation metrics The trained classification models were assessed via the Area Under the ROC (AUC-ROC) curve, and the Area Under the Precision Recall curve measures. We also report validation loss curves to show the effectiveness of the proposed approach for improved model training. In addition, we compared quantitatively the saliency maps produced by the different benchmarked models with those produced by an experienced lung radiologist by means of commonly used metrics, such as Dice Coefficient (DC), Hausdorff Distance 95%(HD_{95})(Reinke et al., 2021), and Structural Similarity Index Measure (SSIM) (Wang et al., 2004). To calculate these metrics, we binarized saliency maps using the ConvexHull function of SciPy with default parameters.

Due to the small dataset size of the provided validation set (200 samples), and following best practices in training models, we show distribution of classification results for all labels in the CheXpert validation set of 200 images, using five-fold validation, with a 80%/20% ratio in order to show consistency of the performance across different folds and models (i.e. boxplots in Fig. 3(a)). We refrained from comparing to previously reported ensemble results (as in Pham et al., 2020), which are typically oriented to challenges, in order to make a more clear and fair comparison of benchmarked models.

Table 1
Mean area under precision recall curves (AUC_{PR}) from 5-fold validation CheXpert dataset.

Baselines				Proposed		Ablation		
DenseNet-121	SE	CBAM	Pham	SIBNet	SIBNet+	SIBNet NoL_{CD}	SIBNet NoL_{SC}	SIBNet-GradCAM
84.1(4.1)	90.4(4.0)	91.4(4.2)	93.2(4.2)	94.1(3.4)	95.6(3.9)	90.3(4.3)	91.5(4.2)	91.2(4.2)

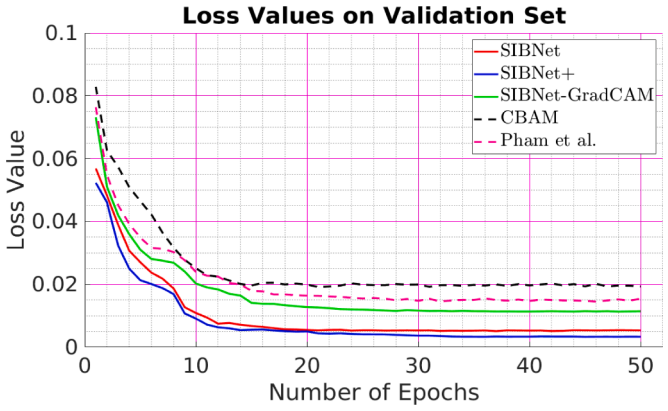

Fig. 4. Validation loss values for different epochs during training. Note: Deep Taylor decomposition was used to compute the saliency maps.

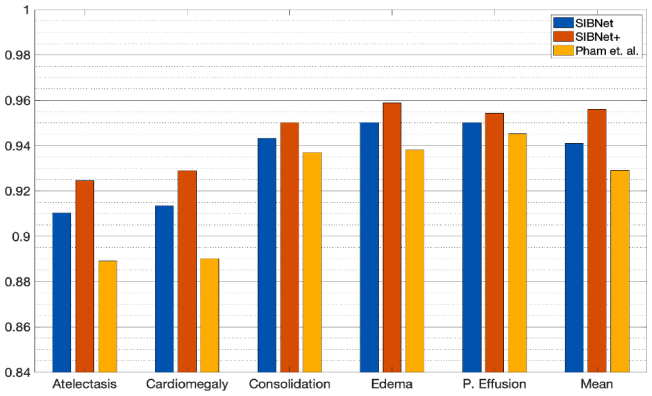

Fig. 5. Bar plots showing average AUC_{ROC} values for each pathology from the 5-fold evaluation CheXpert validation set, using the three-best methods: SIBNet, SIBNet+, Pham (Pham et al., 2020).

3.2. Classification results for chest xray

3.2.1. CheXpert dataset

Fig. 3 shows AUC-ROC (AUC_{ROC}) box plots for all 5 conditions in the CheXpert 5-fold validation set using different saliency map methods such as DeepTaylor (Fig. 3(a)), GradCAM (Fig. 3(b)) and LRP (Bach et al., 2015) (Fig. 3(c)). The proposed SIBNet shows the highest mean AUC followed (Pham et al., 2020; Woo et al., 2018b; Hu et al., 2018). Note that SIBNet+ has superior performance for all saliency methods due to the inclusion of unlabeled data from the NIH dataset. This demonstrates the ability of the approach to utilize, where available, unlabeled images to improve model performance, which is not possible with other approaches necessitating ground-truth data to back propagate gradients.

The AUC_{ROC} is a popular metric for most classification challenges. However, due to the potential bias towards majority classes, we also report the area under precision recall curves (AUC_{PR}). Table 1 summarizes the AUC_{PR} values for different methods, showing a similar trend as for AUC_{ROC} values.

Training performance Fig. 4 shows validation loss curves for different approaches. The results show that SIBNet (and SIBNet+) has a lower initial loss and better convergence than other methods utilizing different attention mechanisms. Fig. 5 shows the AUC_{ROC} values for each of the individual 5 diseases, as well as the overall mean for the 5-fold evaluation on the validation dataset. Results are shown for the three-best methods: SIBNet, SIBNet++, and Pham et al. (2020). SIBNet consistently outperforms (Pham et al., 2020) for all lung conditions, while SIBNet++ improves upon SIBNet due to the availability of additional information from unlabeled images.

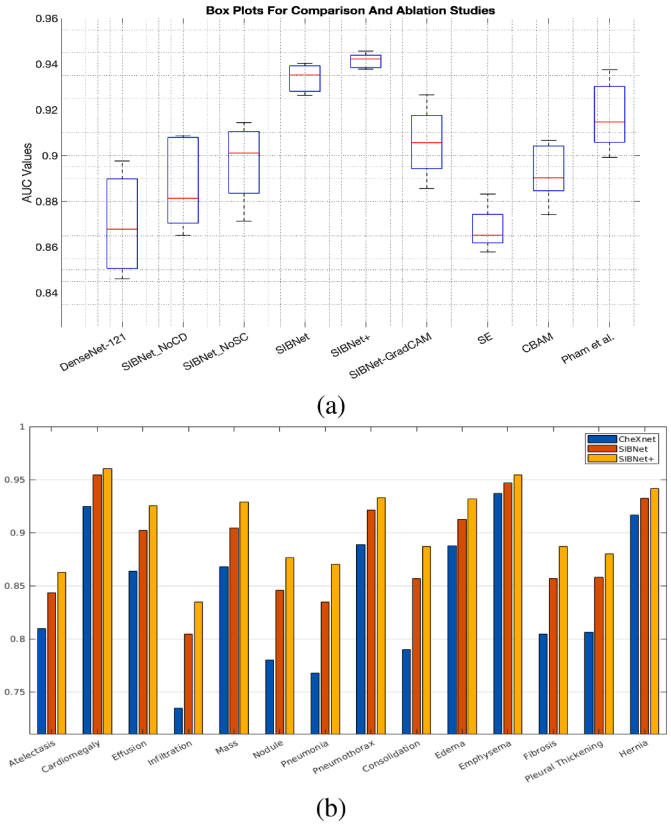

Fig. 6. Results for NIH dataset: (a) AUC_{ROC} Box Plots for average values across all pathologies; (b) AUC_{ROC} values for all 14 pathologies for 3 methods.

3.2.2. NIH dataset

We also conducted experiments on a second and different testing dataset using the NIH Chest X-ray dataset (Wang et al., 2017), which has 112,120 expert-annotated frontal-view X-rays from 30,805 unique patients, and annotations of 14 pulmonary conditions.

For this experiment we compared SIBNet to Chexnet (Rajpurkar et al., 2017) training both models with the same training dataset as described above for the CheXpert dataset, so the only difference between both models is in the loss term used by Chexnet (cross entropy) and by SIBNet (cross entropy plus class distinctiveness and spatial coherence terms). This experiment hence serves the purpose to test for model generalization, and to assess the added benefit of the proposed SIBNet loss terms on a different dataset. Results using DeepTaylor and GradCAM are summarized in Fig. 6(a). Fig. 6(b) shows AUC_{ROC} results for

Table 2
Segmentation performance on the Glas Segmentation Challenge - Mean (standard deviation), using Dice Metric (DSC); F1- F1 score; HD_{95}-95th percentile Hausdorff distance in mm.

	DenseNet 121		ResNet101		Competing methods		
	$UNet_{SIBNet}$	UNet	$UNet_{SIBNet}$	UNet	Rank 1	Rank 2	(Xie et al., 2020)
DSC	92.2(3.4)	87.6 (4.1)	91.8(3.8)	87.0 (4.5)	89.9	89.6	90.6
HD_{95}	52.8(4.2)	60.6 (4.9)	53.9(4.3)	61.1 (5.3)	55.9	62.8	55.1
F1	91.8(3.5)	87.0 (4.3)	91.3(4.0)	86.5 (4.8)	89.4	88.9	89.0

individual pathologies from three methods, including Chexnet (Rajpurkar et al., 2017) baseline results, our proposed SIBNet and SIBNet++. It is observed that SIBNet and SIBNet++ improves upon Chexnet for all 14 different classes.

3.3. Segmentation of pathological structures from histopathology images

We compared the performance of $UNet_{SIBNet}$ with a conventionally trained UNet from scratch using cross entropy and Dice loss for training. To get saliency maps for training we train a separate DenseNet-121 model using the manually provided "benign" or "malignant" labels. The trained model is able to generate the desired saliency maps for each label which is used for training the different UNet models. Table 2 summarizes the segmentation performance on GLAS Challenge for $UNet_{SIBNet}$ and *UNet* using different backbone networks. We observe that $UNet_{SIBNet}$ leads to considerable higher segmentation performance than *UNet* (DSC 92.2 vs. 87.6). Similarly, in terms of Hausdorff distance metrics, $UNet_{SIBNet}$ yielded more robust results than its counterpart (HD_{95} 52.8 mm vs. 60.6 mm). This can be attributed to $UNet_{SIBNet}$ being able to capture class-specific imaging features (i.e., benign vs. malign) that lead the model to better delineate the structures.

In Table 2 we also show results for the top-2 ranked methods, according to challenge results reported in Sirinukunwattana et al. (2017), and comparative results from Xie et al. (2020) a recent deep learning

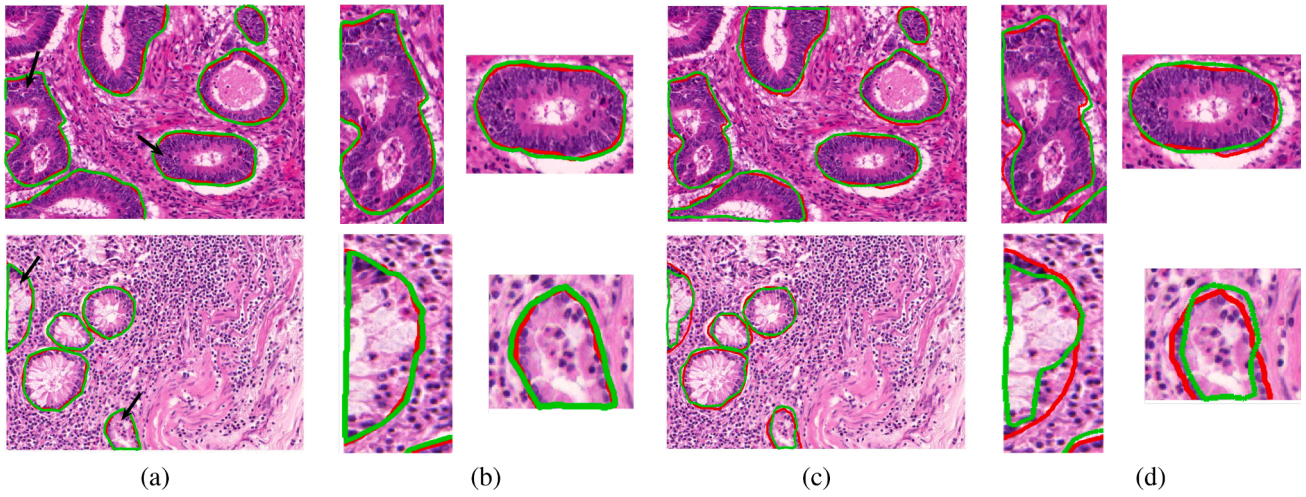

Fig. 7. Segmentation results for the GlaS Segmentation Challenge. Original image and contours of different segmentation methods are shown. The manual segmentation is shown as a red contour and each algorithm's output is shown in green for: (a) $UNet_{Pre-SIBNet}$; (b) Cropped regions highlighted by black arrows in (a); (c) UNet; (d) Cropped regions highlighted by black arrows in (c). Rows correspond to different images. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

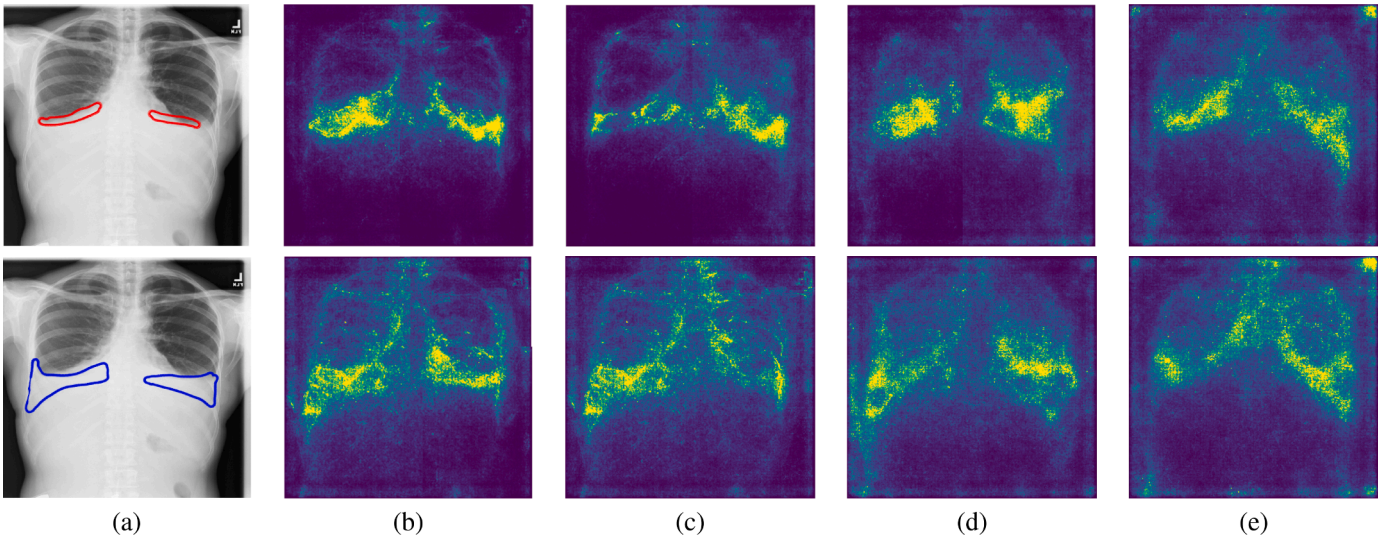

Fig. 8. Comparison with Radiologist's Saliency Maps.(a) Original image with expert-annotated outlines of diagnosed conditions. Saliency Maps for different methods: (b) SIBNet; (c) $SIBNet_{NoL_{SC}}$; (d) $SIBNet_{NoL_{CD}}$; (e) Pham et al. (2020). Top row: Pleural effusion; Bottom row: Atelectasis.

Table 3
Similarity of saliency maps with expert maps-Mean (standard deviation), using SSIM-structural similarity index (Wang et al., 2004); DM-Dice Metric; HD_{95}-95th percentile Hausdorff distance in mm. SSIM,DM $\in [0, 100]$.

	Baselines				Proposed		Ablation	
	DenseNet-121	SE (Hu et al., 2018)	CBAM (Woo et al., 2018b)	Pham (Pham et al., 2020)	SIBNet	SIBNet+	SIBNet NoL_{CD}	SIBNet NoL_{SC}
SSIM	52.1(4.1)	62.2(4.0)	64.3(4.1)	68.5(4.2)	72.1(3.7)	74.2(3.5)	61.2(4.1)	64.2(4.2)
DM	69.1(4.4)	72.3(3.9)	77.4(3.7)	80.2(3.7)	85.9(3.2)	88.2(3.4)	73.8(3.8)	76.7(3.7)
HD_{95}	14.7(3.5)	12.4(3.4)	11.8(3.1)	10.9(3.2)	10.3(2.9)	9.9(2.8)	12.5(2.9)	11.6(3.2)

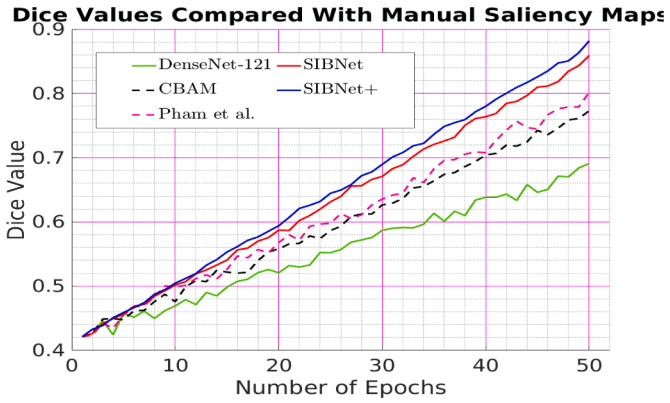

Fig. 9. Agreement of saliency maps between methods and expert annotations, assessed via Dice values over epochs.

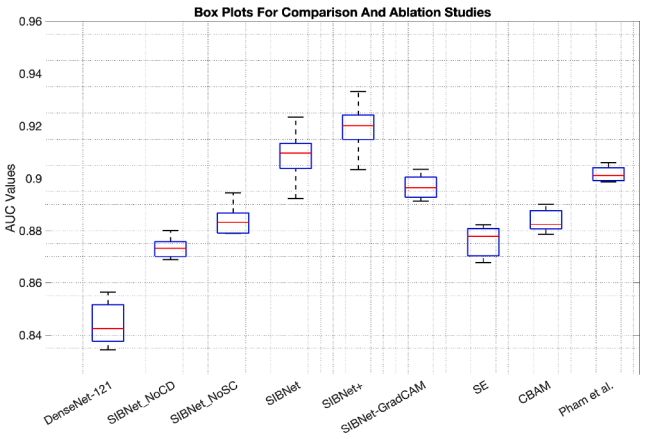

Fig. 10. Box Plots of AUC_{ROC} values when using latent representations of original images instead of saliency maps.

method that outperforms these two methods. Rank-1, Rank-2 are both deep learning based methods with specific pre-processing steps designed to improve method robustness, while (Xie et al., 2020) employs a pair-wise relation-based semi-supervised (PRS2) model for gland segmentation. Fig. 7 shows example segmentation outputs of $UNet_{SIBNet}$ and UNet. The results clearly show the improvements over the baseline UNet. We note that visual results for the other approaches are not available due to the lack of code availability.

3.4. Investigation of saliency maps' performance

3.4.1. Enhanced interpretability saliency maps

Fig. 8 shows the resulting saliency maps on the CheXpert dataset for the proposed, ablated, and best compared approach from Pham et al. (2020), along with the corresponding saliency map generated by the expert lung radiologist. Fig. 8(a) shows the expert delineated regions for pleural effusion (red outline, top row) and atelectasis (blue outline, bottom row). Salient maps are shown for SIBNet (Fig. 8(b)), $SIBNet_{NoL_{SC}}$ (Fig. 8(c)), and $SIBNet_{NoL_{CD}}$ (Fig. 8(d)), and by Pham et al. (2020) (Fig. 8 (e)). Maps corresponding to the baseline DenseNet are shown in Fig. 2. Excluding the spatial coherence term L_{SC} results in more dispersed salient regions, with image borders and corners also seemingly used by the model, alluding to a potential shortcut learning. On the other hand excluding the class distinctiveness L_{CD} results in more similar saliency maps for different disease labels. Employing all loss terms (Eq. (3)) yields more distinctive and spatially coherent salient regions for each class label.

3.4.2. Comparison with radiologist's saliency maps

We compared saliency maps from SIBNet and benchmarked approaches, with those generated by a radiologist with over 15 years of experience. Results in Fig. 8 show that saliency maps yielded by SIBNet are most similar to the radiologist's maps. Although we do not expect a perfect alignment, a good saliency map should highlight a majority of the regions of interest characterizing the lung condition. Quantitative measures for saliency overlap in Table 3 also highlight SIBNet's superior performance.

We also report performance of models during their training. Fig. 9 shows the average Dice metric of the resulting saliency maps after every epoch. All methods show dice values that are uniformly increasing, but values of SIBNet and SIBNet+ are notably more stable, while all other methods produce values that keep fluctuating.

3.5. Ablation studies

3.5.1. Ablation studies for CheXpert dataset

Fig. 3 shows results for the ablation experiments excluding different terms in the loss function. The results show that the proposed SIBNet loss terms are complementary and important contributors to the overall model performance. Excluding any one term results in a performance drop.

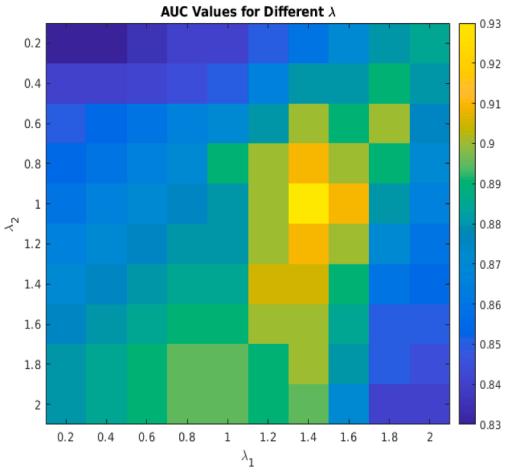

Fig. 11. Change of AUC values for different λ_1, λ_2 using Deep taylor Saliency maps.

Table 4
Effect of adding noise on the Chexpert dataset. AUC_{ROC} values for added Gaussian noise of $\mu = 0$ and different σ.

	Baselines				Proposed		Ablation		
	DenseNet-121	SE	CBAM	Pham	SIBNet	SIBNet+	SIBNet NoL_{CD}	SIBNet NoL_{SC}	SIBNet-GradCAM
$\sigma = 0$	84.4(4.6)	90.8(4.3)	91.6(4.3)	93.2(4.1)	94.3(3.9)	95.5(3.8)	90.1(4.3)	91.4(4.1)	92.1(4.0)
$\sigma = 0.01$	83.9(4.8)	90.2(4.4)	91.0(4.5)	92.6(4.3)	93.6(4.1)	95.0(3.9)	89.6(4.4)	91.0(4.3)	91.7(4.2)
$\sigma = 0.05$	81.1(5.0)	89.0(4.7)	89.7(4.6)	91.1(4.4)	92.1(4.2)	93.4(4.1)	88.1(4.5)	89.6(4.2)	90.3(4.3)
$\sigma = 0.1$	78.4(5.1)	86.2(4.8)	87.3(4.5)	88.4(4.5)	89.8(4.2)	90.7(4.2)	85.0(4.5)	86.1(4.3)	87.1(4.2)

Table 5
Effect of using features with different classifiers.

	SVM-Linear	SVM-Gaussian	Random forests
AUC	94.1(4.3)	94.2(4.2)	94.0(4.3)

3.5.2. Using latent representations of original images instead of saliency maps

As further ablation study, we report performance metrics when using the original images instead of saliency maps to calculate the proposed losses. Fig. 10 shows the box plots of AUC_{ROC} values when using the latent feature vectors of the original images instead of the saliency maps. Although SIBNet still outperforms the other approaches, there is a significant reduction in performance when using the original images (see Fig. 3). This behaviour can be attributed to the fact that saliency maps highlight information regarding the pathology, which is in turn the target of the classification model being explained. In contrast, the original X-ray image includes other sources of information, including the overall anatomy, that is of much lower relevance to the trained model. This result highlights the benefit of using saliency maps as a natural and interpretable attention mechanism.

3.5.3. Setting λ values

Fig. 11 illustrates the change of λ_1 and λ_2 values on a coarse scale for DeepTaylor saliency maps. For the full range of values of λ_1, λ_2 there are 1600 combinations. Since this is difficult to illustrate, we show in Fig. 11 variations with steps of 0.2, which gives 100 possible combinations. Best performance was found for $\lambda_1 = 1.4$ and $\lambda_2 = 1.0$, showing that both terms contribute to improved results (i.e. none of the terms reduced to zero). Similar behaviour was observed for saliency maps based on GradCAM and LRP.

3.5.4. Robustness to noise

In an attempt to quantify the models robustness to noise we added Gaussian noise of $\mu = 0$ and different $\sigma \in \{0.005, 0.01, 0.05, 0.1\}$ to input images. Table 4 shows for the CheXpert dataset, the AUC_{ROC} values for different methods at different levels of added noise. While the performance of all methods decreases with added noise, SIBNet performance is more robust to noise variations.

3.5.5. Using simpler classifiers

To determine whether the observed superior performance of SIBNet is due to more discriminative learned features, we performed an alternate classification where we used learned features to train simple classifiers such as SVMs (with different kernels) and Random Forests. Results are summarized in Table 5. We observed that the linear classifiers yielded similar results as the SIBNet method (see Table 1). This suggests that the use of inductive biases results in more discriminative features that can be used with different types of classifiers. We note that the proposed inductive bias is not restricted by the deep learning framework, and nor does it require further use of non-linear kernels in SVM classifiers.

3.5.6. Effect of reduced/increased network capacity

As another way of evaluating model generalization capability, we reduced and increased the number of parameters across different layers of benchmarked models and investigated the effect on validation and test losses. Fig. 12 summarizes our findings. These experiments were carried out on the CheXpert dataset. We denote our reference model (i.e., without any reduction/increase) as having P number of parameters. Fig. 12 shows the loss value with varying number of parameters denoted as multiples of P. With reduced number of parameters we observe an increase in validation and test error. Conversely, when increasing the number of parameters beyond P, training and validation errors remain constant up to $1.8P$, when we notice that although the validation error remains mostly constant, the test error increases, particularly for CBAM and (Pham et al., 2020). These results indicate a lower tendency of SIBNet to overfit as well as an improved generalization capability.

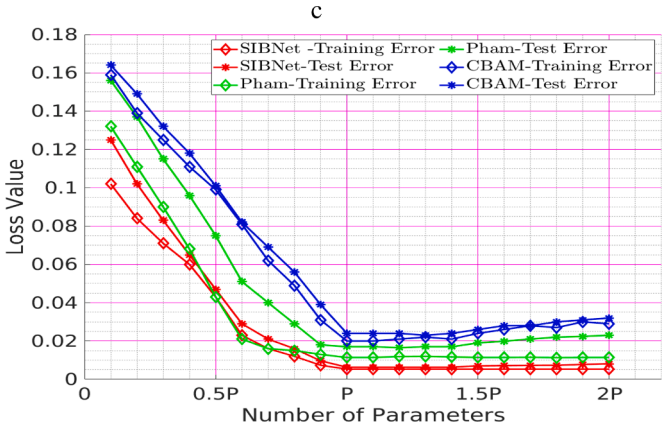

Fig. 12. Change of validation and test error with varying number of parameters.

4. Discussion

Inductive bias plays an important role in machine and deep learning as it allows us to inject domain knowledge and specific sought-out characteristics that a trained model should have. In medical image analysis applications this need is exacerbated by the relatively smaller datasets, and different confounders present in real world clinical datasets, increasing the likelihood of shortcut learning (Geirhos et al., 2020; DeGrave et al., 2021), where models use spurious correlations in the data. Convolutional Neural Networks (LeCun et al., 1989) are a successful example of inductive bias for image recognition tasks, where the convolution operator implicitly encodes the known a-priori information regarding neighboring pixel relationships in an image. Recently, visual image transformer networks (Dosovitskiy et al., 2020) have demonstrated high performance levels by making the inductive bias more general but at the cost of necessitating more training data than CNN networks. In medical imaging, hybrid approaches are emerging in an effort to combine the benefits of both approaches (Chen et al., 2021; Wu et al., 2021). In this study we introduce a complementary inductive bias building on findings from the area of interpretability of deep learning. Interpretability in medical imaging applications has been signaled as essential to ensure a safe, trustable and effective adoption of deep

learning technologies in the clinics (Cardoso et al., 2020; Reyes et al., 2020; Fuhrman et al., 2022; Budd et al., 2021; Kitamura and Marques, 2021; McCrindle et al., 2021). Beyond the originally-defined objective of yielding model interpretation, in this study we demonstrate how interpretability can be used to produce an effective inductive bias mechanism, simultaneously leading to improved model performance and model interpretability. Through experimentation with two common medical imaging problems of classification and segmentation, several ablation and robustness tests, we show the interesting properties of the proposed Saliency Inductive Bias Network (SIBNet) approach.

Modularity of SIBNet SIBNet enables utilization of different interpretability approaches. In this study we experimented with known methods such as DeepTaylor, GradCam and LRP, and demonstrate that irrespective of the saliency map method, improved model performance is attained. Although we show results for DeepTaylor, GradCAM and LRP, we emphasize that other approaches can be used interchangeably. This flexibility allows one to adopt new approaches being developed in the evolving field of interpretability (Doshi-Velez and Kim, 2017; Eitel et al., 2019). Secondly, since SIBNet does not require modification of the model architecture, it allows re-utilization of available pre-trained models on large datasets. Furthermore, the proposed SIBNet loss terms can be used in conjunction with other loss terms commonly used to train deep learning models, allowing further exploration of losses (Ma et al., 2021).

On the effectiveness of saliency map latent representations The proposed class distinctiveness term of SIBNet utilizes the latent representation of the whole saliency map, which simultaneously encodes location, shape and other information. In this sense, this term has the ability of characterizing diseases or clinical conditions that might even present spatial overlap, as is the case presented in this study for lung conditions. We attribute this to the effectiveness of latent representations studied in Zhang et al. (2018). Complementing this term, the spatial coherence term of SIBNet regularizes incorrect local variations, which from our experience occur often in medical imaging, potentially as a result of shortcut learning, leading to sparse saliency maps not being consistent with the analysis performed by radiologists (DeGrave et al., 2021).

Furthermore, the superiority of latent representations derived from saliency maps, over those directly generated from the input images, aligns with previous findings in Silva et al. (2020), where latent representations of saliency maps were used for medical image retrieval purposes, and in sample selection for active learning (Mahapatra et al., 2021b). Intuitively, we attribute this superiority to the fact that saliency maps focus on the information regarding the pathology, which is in turn the target of the model being explained. In contrast, latent representations of the input images encode other sources of information, including the overall anatomy, which is typically of much lower relevance for the task, but can potentially be misused by a model operating in shortcut learning mode (Geirhos et al., 2020; DeGrave et al., 2021).

Improved learning rate, robustness and generalization potential The proposed interpretability-guided inductive bias jointly aims at improving model performance and interpretability by explicitly guiding the learning process to generate features yielding distinctive saliency maps across classes, as well as to promote spatial coherence of saliency maps generated by the model. Our findings suggest that the proposed inductive bias is not only able to yield improved performance of trained models, but also yields a faster learning rate, as evidenced in the validation curves in Fig. 4. We note that this behaviour was found across the different saliency maps utilized in our experiments (i.e. DeepTaylor, LRP and GradCam). This characteristic is particularly important in medical imaging applications where large annotated training datasets are challenging and time consuming to create. We believe these results align with recent findings regarding the importance of effective inductive bias (Locatello et al., 2019). Furthermore, high learning rates are particularly important in clinical scenarios where imaging devices and protocols are updated, and thus model retraining, or active learning, needs to take place in a time and resource effective manner.

In terms of robustness, we analyzed SIBNet under different levels of image noise, model capacity, and its performance on a secondary unknown dataset. Compared to the benchmarked approaches, SIBNet shows a superior generalization capability and robustness levels, however our experiments are limited here and we cannot guarantee superiority across tasks (no free lunch theorem). We advocate that the proposed inductive bias is well suited for a large set of medical imaging tasks where the focus of the task relates to image areas where the medical condition of interest is radiologically seen.

Improved interpretability Enhanced interpretability of deep learning systems for medical image applications has been an important area of discussion and research in the last few years. The proposed interpretability-guided inductive bias explicitly enforces sought-out characteristics of class distinctiveness and spatial coherence, leading to saliency maps in better agreement with expert-generated saliency maps (see Table 3 and Fig. 9). The proposed class distinctiveness loss term of SIBNet across potential classes, aligns with the differential diagnosis that radiologists need to be trained on. This is essential in order to correctly diagnose cases in the presence of similar imaging patterns that could otherwise confuse a non-expert reader. Compared to other approaches proposing self-attention maps, such as CBAM (Woo et al., 2018b) and channel-wise attention mechanisms (Hu et al., 2018), SIBNet can be utilized with any network architecture without need for modifications. Similarly, in comparison to contrastive learning (Chen et al., 2020), SIBNet does not require artificial construction of contrastive samples via data augmentation, but naturally utilizes the expected class-distinctiveness of saliency maps per data point to drive the learning process.

Semi-supervised uses Beyond supervised learning we investigated the use of unlabeled data, and reported results of SIBNet++, which can use unlabeled data to further guide a model during training. In these regards, an interesting area of further research would be to employ this finding to perform quality control of already deployed models on newly unseen (i.e. unlabeled) datasets, utilizing the levels of class distinctiveness and spatial coherence, observed during training, as quality reference points for testing time.

Some limitations and potential further areas of research Some limitations are worth mentioning. Our study was limited to classification and segmentations tasks, and although we foresee interesting advantages, we cannot guarantee similar findings on tasks such localization and other regression tasks. We also limited our study to medical imaging datasets, but we see interesting applications to other medical multi-omics scenarios where existing interpretability approaches for non-imaging datasets (Lundberg and Lee, 2017) could be used in combination with saliency maps for medical images to drive the learning process of such multi-omics models.

5. Conclusions

In this paper we propose a novel interpretability-guided inductive bias approach that is motivated by the hypothesis that enhanced class-distinctiveness and spatial coherence of saliency maps, injected during model training, leads to improved model performance and improved interpretability of model's predictions. Experiments and comparisons with state of the art approaches on publicly available datasets show the added benefits of the proposed approach. We highlight its ability to operate on any existing architecture without need of modifications and its modularity, which is exemplified by using different interpretability approaches and unlabeled data. Beyond the presented results, we believe the proposed approach can be extended for multi-omics problems where different data types (e.g. imaging, text reports, clinical laboratory, etc.) can naturally be handled by available interpretability approaches able to interpret data of different nature, leading to a flexible multi-omics interpretability-guided inductive bias framework.

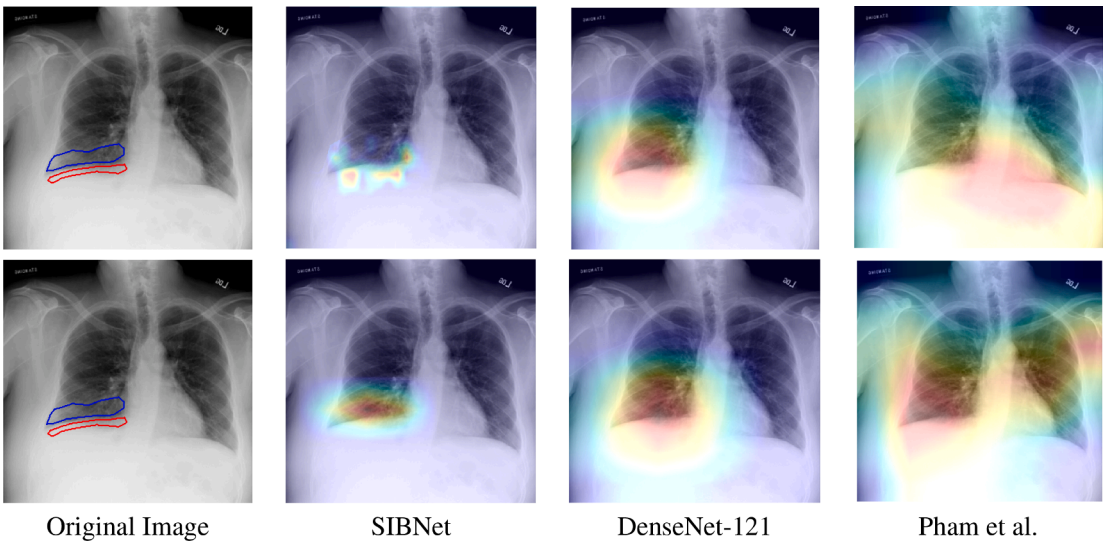

Fig. A1. Saliency map comparison with radiologist's saliency maps using **GradCAM** for a patient diagnosed with Pleural effusion (red contour) and Atelectasis (blue contour). From left to right: Original image with expert-annotated outlines of diagnosed conditions, saliency maps for SIBNet, DenseNet-121, Pham et al. (2020). Top row: Saliency maps for Pleural effusion (red contour); Bottom row: Saliency maps for Atelectasis (blue contour). (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

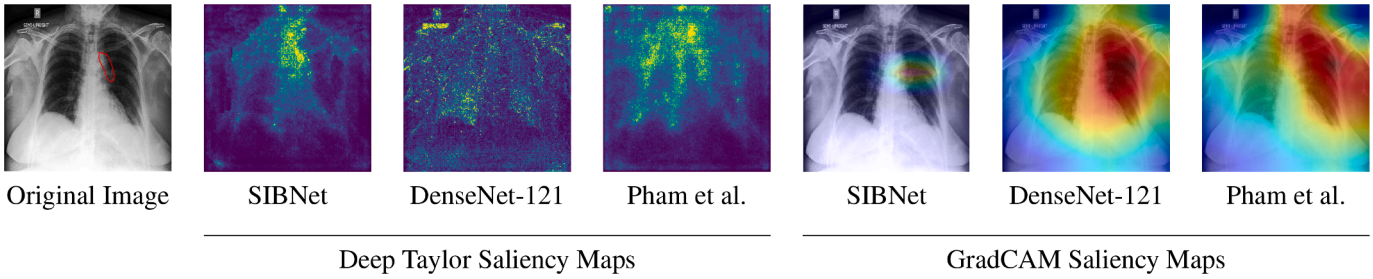

Fig. A2. Saliency map comparison with radiologist's saliency maps using **Deep Taylor** and **GradCAM** for a patient with **Edema**. From left to right: Original image with expert-annotated outline of Edema, DeepTaylor-based saliency maps for SIBNet, DenseNet-121, Pham et al. (2020), and GradCAM-based saliency maps for SIBNet, DenseNet-121, Pham et al. (2020).

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Swiss National Foundation grant number 198388, and Innosuisse grant number 31274.1.

Appendix A. Additional results For SIBNet saliency maps

Fig. A.13 shows the resulting *GradCAM-based* saliency maps for SIBNet, DenseNet-121, and best compared approach from Pham et al. (2020), along with the corresponding saliency map generated by the expert lung radiologist, for a case jointly presenting Pleural effusion and Atelectasis. Fig. A.14 shows saliency maps for a case with Edema. We show results for SIBNet, DenseNet and best compared approach from Pham et al. (2020). In addition, we show results for both Deep Taylor and GradCAM. Similar to the results presented for Deep Taylor Maps (Ref Fig. 8), we observe that inclusion of our novel loss terms makes the saliency maps more compact and greater aligned with diseased regions.

References

Aggarwal, R., Sounderajah, V., Martin, G., Ting, D.S., Karthikesalingam, A., King, D., Ashrafian, H., Darzi, A., 2021. Diagnostic accuracy of deep learning in medical imaging: a systematic review and meta-analysis. NPJ Digit. Med. 4 (1), 1–23.

Alber, M., Lapuschkin, S., Seegerer, P., Hägele, M., Schütt, K.T., Montavon, G., Samek, W., Müller, K.-R., Dähne, S., Kindermans, P.-J., 2019. Innvestigate neural networks. J. Mach. Learn. Res. 20 (93), 1–8.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., Samek, W., 2015. On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. PLoS One 10 (7), e0130140.

Bello, I., Zoph, B., Vaswani, A., Shlens, J., Le, Q.V., 2019. Attention augmented convolutional networks. Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 3286–3295.

Budd, S., Robinson, E.C., Kainz, B., 2021. A survey on active learning and human-in-the-loop deep learning for medical image analysis. Med. Image Anal. 71 (4), 102062.

Cardoso, J., Van Nguyen, H., Heller, N., Abreu, P.H., Isgum, I., Silva, W., Cruz, R., Amorim, J.P., Patel, V., Roysam, B., et al., 2020. Interpretable and annotation-efficient learning for medical image computing. Third International Workshop, iMIMIC 2020, Second International Workshop, MIL3ID 2020, and 5th International Workshop, LABELS 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 4–8, 2020, Proceedings.

Chen, J., Lu, Y., Yu, Q., Luo, X., Adeli, E., Wang, Y., Lu, L., Yuille, A. L., Zhou, Y., 2021. Transunet: transformers make strong encoders for medical image segmentation. arXiv preprint arXiv:2102.04306.

Chen, T., Kornblith, S., Norouzi, M., Hinton, G., 2020. A simple framework for contrastive learning of visual representations. International Conference on Machine Learning. PMLR, pp. 1597–1607.

DeGrave, A.J., Janizek, J.D., Lee, S.-I., 2021. Ai for radiographic COVID-19 detection selects shortcuts over signal. Nat. Mach. Intell. 3, 610–619.

Doshi-Velez, F., Kim, B., 2017. Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., Houlsby, N., 2020. An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929.

Eitel, F., Ritter, K., et al., for the Alzheimer Disease Neuroimaging Initiative (ADNI), 2019. Testing the robustness of attribution methods for convolutional neural networks in mri-based alzheimer's disease classification. Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support. Springer, pp. 3–11.

Fuhrman, J.D., Gorre, N., Hu, Q., Li, H., El Naqa, I., Giger, M.L., 2022. A review of explainable and interpretable ai with applications in COVID-19 imaging. Med. Phys. 49 (1), 1–14. https://doi.org/10.1002/mp.15359.Epub 2021 Dec 7. PMID: 34796530; PMCID: PMC8646613.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., Wichmann, F.A., 2020. Shortcut learning in deep neural networks. Nat. Mach. Intell. 2 (11), 665–673.

Goyal, A., Bengio, Y., 2020. Inductive biases for deep learning of higher-level cognition. arXiv preprint arXiv:2011.15091.

Griffiths, T.L., Chater, N., Kemp, C., Perfors, A., Tenenbaum, J.B., 2010. Probabilistic models of cognition: exploring representations and inductive biases. Trends Cogn. Sci. 14 (8), 357–364.

Hessel, M., van Hasselt, H., Modayil, J., Silver, D., 2019. On inductive biases in deep reinforcement learning. arXiv preprint arXiv:1907.02908.

Hu, J., Shen, L., Sun, G., 2018. Squeeze-and-excitation networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7132–7141.

Huang, G., Liu, Z., van der Maaten, L., Weinberger, K., 2016. Densely connected convolutional networks. arXiv:1608.06993.

Irvin, J., Rajpurkar, P., et al., 2019. Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. arXiv preprint arXiv:1901.07031.

Isensee, F., Jaeger, P.F., Kohl, S.A., Petersen, J., Maier-Hein, K.H., 2021. nnU-Net: a self-configuring method for deep learning-based biomedical image segmentation. Nat. Methods 18 (2), 203–211.

Kingma, D., Ba, J., 2014. Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Kitamura, F.C., Marques, O., 2021. Trustworthiness of artificial intelligence models in radiology and the role of explainability. J. Am. Coll. Radiol. 18 (8), 1160–1162.

LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., Jackel, L., 1989. Handwritten digit recognition with a back-propagation network. Adv. Neural Inf. Process. Syst. 2, 396–404.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., Van Der Laak, J.A., Van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in medical image analysis. Med. Image Anal. 42, 60–88.

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

Locatello, F., Bauer, S., Lucic, M., Raetsch, G., Gelly, S., Schölkopf, B., Bachem, O., 2019. Challenging common assumptions in the unsupervised learning of disentangled representations. International Conference on Machine Learning. PMLR, pp. 4114–4124.

Lundberg, S.M., Lee, S.-I., 2017. A unified approach to interpreting model predictions. Proceedings of the 31st International Conference on Neural Information Processing Systems, pp. 4768–4777.

Ma, J., Chen, J., Ng, M., Huang, R., Li, Y., Li, C., Yang, X., Martel, A.L., 2021. Loss odyssey in medical image segmentation. Med. Image Anal. 71, 102035.

Mahapatra, D., Bozorgtabar, B., Ge, Z., 2021. Medical image classification using generalized zero shot learning. IEEE CVAMD 2021, pp. 3344–3353.

Mahapatra, D., Bozorgtabar, S., Hewavitahranage, S., Garnavi, R., 2017. Image super resolution using generative adversarial networks and local saliencymaps for retinal image analysis. Proc. MICCAI, pp. 382–390.

Mahapatra, D., Buhmann, J., 2015. Visual saliency based active learning for prostate MRI segmentation. Proc. MLMI, pp. 9–16.

Mahapatra, D., Ge, Z., Reyes, M., 2022. Self-supervised generalized zero shot learning for medical image classification using novel interpretable saliency maps. IEEE Trans. Med. Imaging 1. https://doi.org/10.1109/TMI.2022.3163232.

Mahapatra, D., Poellinger, A., Shao, L., Reyes, M., 2021. Interpretability-driven sample selection using self supervised learning for disease classification and segmentation. IEEE Trans. Med. Imaging 40 (10), 2548–2562. https://doi.org/10.1109/TMI.2021.3061724.

Mahapatra, D., Roy, P., Sedai, S., Garnavi, R., 2016. Retinal image quality classification using saliency maps and CNNs. Proc. MICCAI-MLMI, pp. 172–179.

Mahapatra, D., Sun, Y., 2008. Nonrigid registration of dynamic renal MR images using a saliency based MRF model. Proc. MICCAI, pp. 771–779.

Mahapatra, D., Sun, Y., 2010. Joint registration and segmentation of dynamic cardiac perfusion images using MRFs. Proc. MICCAI, pp. 493–501.

Mahapatra, D., Sun, Y., 2011. MRF based intensity invariant elastic registration of cardiac perfusion images using saliency information. IEEE Trans. Biomed. Eng. 58 (4), 991–1000.

Mahapatra, D., Sun, Y., 2012. Integrating segmentation information for improved MRF-based elastic image registration. IEEE Trans. Image Proc. 21 (1), 170–183.

McCrindle, B., Zukotynski, K., Doyle, T.E., Noseworthy, M.D., 2021. A radiology-focused review of predictive uncertainty for AI interpretability in computer-assisted segmentation. Radiology 3 (6), e210031.

Montavon, G., Lapuschkin, S., Binder, A., Samek, W., Muller, K.-R., 2017. Explaining nonlinear classification decisions with deep taylor decomposition. Pattern Recognit. 65, 211–222.

Pham, H. H., Le, T. T., Tran, D. Q., Ngo, D. T., Nguyen, H. Q., 2020. Interpreting chest X-rays via CNNs that exploit hierarchical disease dependencies and uncertainty labels. arXiv preprint arXiv:1911.06475.

Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., Lungren, M. P., Ng, A., 2017. Chexnet: radiologist-level pneumonia detection on chest X-rays with deep learning. arXiv preprint arXiv:1711.05225.

Ramachandran, P., Parmar, N., Vaswani, A., Bello, I., Levskaya, A., Shlens, J., 2019. Stand-alone self-attention in vision models. arXiv preprint arXiv:1906.05909.

Reinke, A., Eisenmann, M., Tizabi, M. D., Sudre, C. H., Rädsch, T., Antonelli, M., Arbel, T., Bakas, S., Cardoso, M. J., Cheplygina, V., et al., 2021. Common limitations of image processing metrics: apicture story. arXiv preprint arXiv:2104.05642.

Reyes, M., Meier, R., Pereira, S., Silva, C.A., Dahlweid, F.-M., von Tengg-Kobligk, H., Summers, R.M., Wiest, R., 2020. On the interpretability of artificial intelligence in radiology: challenges and opportunities. Radiology 2 (3), e190043. https://doi.org/10.1148/ryai.2020190043.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-net: Convolutional networks for biomedical image segmentation. Proc. MICCAI, pp. 234–241.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2017. Grad-CAM: visual explanations from deep networks via gradient-based localization. Proc. ICCV, pp. 618–626.

Silva, W., Poellinger, A., Cardoso, J.S., Reyes, M., 2020. Interpretability-guided content-based medical image retrieval. International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 305–314.

Sirinukunwattana, K., et al., 2017. Gland segmentation in colon histology images: the GlaS challenge contest. Med. Image Anal. 35, 489–502.

Tajbakhsh, N., Shin, J.Y., Gurudu, S.R., Hurst, R.T., Kendall, C.B., Gotway, M.B., Liang, J., 2016. Convolutional neural networks for medical image analysis: full training or fine tuning? IEEE Trans. Med. Imaging 35 (5), 1299–1312.

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

Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P., 2004. Image quality assessment: from error visibility to structural similarity. IEEE Trans. Image Process. 13 (4), 600–612.

Weatheritt, J., Rueckert, D., Wolz, R., 2020. Transfer learning for brain segmentation: Pre-task selection and data limitations. Medical Image Understanding and Analysis. Springer International Publishing, pp. 118–130.

Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. CBAM: convolutional block attention module. Proceedings of the European Conference on Computer Vision (ECCV).

Woo, S., Park, J., Lee, J.-Y., Kweon, I.S., 2018. CBAM: convolutional block attention module. Proceedings of the European Conference on Computer Vision (ECCV).

Wu, H., Xiao, B., Codella, N., Liu, M., Dai, X., Yuan, L., Zhang, L., 2021. CVT: introducing convolutions to vision transformers. arXiv preprint arXiv:2103.15808.

Xie, Y., Zhang, J., Liao, Z., Verjans, J., Shen, C., Xia, Y., 2020. Pairwise relation learning for semi-supervised gland segmentation. Proc. MICCAI, pp. 417–427.

Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O., 2018. The unreasonable effectiveness of deep features as a perceptual metric. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 586–595.