




OPEN

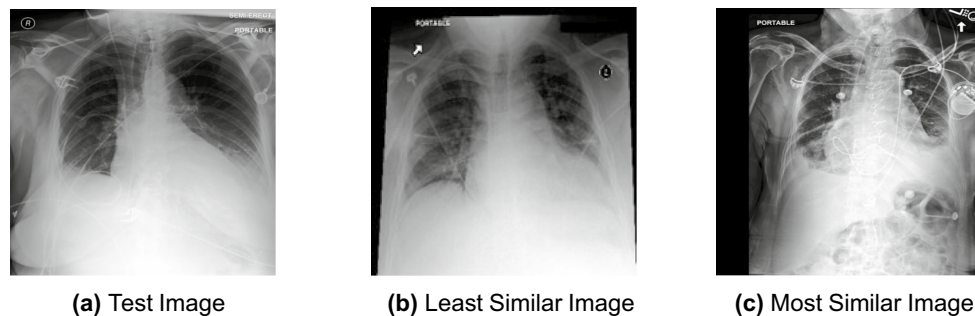
## Computer-aided diagnosis through medical image retrieval in radiology

Wilson Silva<sup>1,2</sup>, Tiago Gonçalves<sup>1,2</sup>, Kirsi Härmä<sup>3</sup>, Erich Schröder<sup>3</sup>, Verena Carola Obmann<sup>3</sup>, Maria Cecilia Barroso<sup>3</sup>, Alexander Poellinger<sup>3</sup>, Mauricio Reyes<sup>4,5</sup> & Jaime S. Cardoso<sup>1,2,5</sup>

Currently, radiologists face an excessive workload, which leads to high levels of fatigue, and consequently, to undesired diagnosis mistakes. Decision support systems can be used to prioritize and help radiologists making quicker decisions. In this sense, medical content-based image retrieval systems can be of extreme utility by providing well-curated similar examples. Nonetheless, most medical content-based image retrieval systems work by finding the most similar image, which is not equivalent to finding the most similar image in terms of disease and its severity. Here, we propose an interpretability-driven and an attention-driven medical image retrieval system. We conducted experiments in a large and publicly available dataset of chest radiographs with structured labels derived from free-text radiology reports (MIMIC-CXR-JPG). We evaluated the methods on two common conditions: pleural effusion and (potential) pneumonia. As ground-truth to perform the evaluation, query/test and catalogue images were classified and ordered by an experienced board-certified radiologist. For a profound and complete evaluation, additional radiologists also provided their rankings, which allowed us to infer inter-rater variability, and yield qualitative performance levels. Based on our ground-truth ranking, we also quantitatively evaluated the proposed approaches by computing the normalized Discounted Cumulative Gain (nDCG). We found that the Interpretability-guided approach outperforms the other state-of-the-art approaches and shows the best agreement with the most experienced radiologist. Furthermore, its performance lies within the observed inter-rater variability.

The increasing use of advanced cross-sectional imaging and the evolution of the information technology infrastructure to meet the demands of higher imaging volumes (i.e., improved computational power, storage capacity, and workflow efficiency in the picture archiving and communication system (PACS) environment), contributed to a substantial increase of the amount of images generated per examination<sup>1</sup>. Consequently, this has increased the workload of radiologists, which must now interpret more examination images in less time, thus creating the possibility for increased detection errors as a result of increased fatigue and stress, lowering the quality of the healthcare delivered by the radiologists to the patients<sup>2,3</sup>. Moreover, as the ratio of diagnostic demand to the number of radiologists increases, the diminished effective available time per diagnostic becomes a critical issue<sup>4</sup>. According to the current paradigm, in case of doubt for a suspected condition, radiologists often turn to public or internal image databases where similar disease-matching images of the diseases the radiologist has narrowed down can be searched and compared against (e.g., *Radiopaedia*). After reviewing all possible differential diagnoses (those originally considered and those that came up during the search), the radiologist weighs these diagnoses and usually gives 2–4 of them as possible diagnoses. In this process, the radiologist ranks the images and creates an ordered set of images in his/her head. This task is time-consuming and often ineffective since it requires several iterations until a proper matching image supporting the final diagnosis is found. Moreover, these databases are limited in the variability of cases presented to the users, which is exacerbated in conditions of low prevalence. Hence, it is extremely relevant to develop disease-targeted content-based image retrieval (CBIR) systems that automatically present disease-matching similar images to the one being analysed. A CBIR system usually focuses on two different tasks: feature representation, which consists of finding a low-dimensional representation of the image that is suitable for characterising it well enough; and, feature indexing and search,

<sup>1</sup>INESC TEC, Porto, Portugal. <sup>2</sup>Faculty of Engineering, University of Porto, Porto, Portugal. <sup>3</sup>Department of Diagnostic, Interventional and Pediatric Radiology, Inselspital, Bern University Hospital, University of Bern, Bern, Switzerland. <sup>4</sup>ARTORG Center for Biomedical Engineering Research, University of Bern, Bern, Switzerland. <sup>5</sup>These authors jointly supervised this work: Mauricio Reyes and Jaime S. Cardoso. ✉email: wilson.j.silva@inesctec.pt



**Figure 1.** Pleural Effusion test image and the least and most similar images of the catalogue according to our board-certified radiologist (in terms of disease and disease severity). The overall most similar image would be (b). However, such matching is not of radiological interest.

which focus on the efficiency of the retrieval process<sup>5</sup>. Our work focuses on the first step, i.e., on finding the most appropriate feature representation for the task at hand.

Finding the most appropriate feature representation is an arduous task since the clinical analysis is typically constricted to a small region of the image, discarding most of the available information. As such, finding the overall most similar image (i.e., including all pixels in the image) is not the objective, instead, we are interested in finding the most similar image in terms of disease and disease severity. As illustrated in Fig. 1, those can be quite far apart, as Fig. 1b is, overall, more similar to Fig. 1a than Fig. 1c, while in terms of disease and disease severity, it is the opposite, with Fig. 1b being the least similar image and Fig. 1c the most similar (from a catalogue of 10 images).

Given that the disease features are located in a small region of the entire image, the medical CBIR system should also be paying attention to that specific region, ignoring the remaining information. However, most CBIR systems perform their analysis taking the entire image into account, particularly the more traditional methods. Deep learning approaches have a better focus on the disease-related characteristics as they learn the appropriate feature representations to solve the classification task of interest. Thus, they represent an improvement in terms of focus when compared to the more traditional approaches. Nonetheless, we hypothesise that this can be further improved by increasing even more the focus of the network in the regions that matter to the decision, and explore two different techniques: one driven by interpretability, and another based on attention mechanisms.

This paper builds upon our work proposed in Silva et al.<sup>4</sup>. In this study we extend our previous work by (1) adding experiments with a second new dataset; (2) a second medical condition (pneumonia), and (3) a comparison to a recently proposed network, employing implicit attention mechanisms. Furthermore, we improved our comparisons to expert radiologists by adding two more board-certified radiologists to each study in order to assess the evaluated methods with respect to the inter-rater variability of these tasks.

The remainder of this paper is organised as follows: section “[Background](#)” introduces the concepts and state-of-the-art of the topics related to this study, namely, Medical Image Retrieval, Explainable Artificial Intelligence, and Attention Mechanisms; section “[Materials and methods](#)” describes the dataset used, the baselines, our methods, and the evaluation framework; section “[Results and discussion](#)” presents the quantitative and qualitative results obtained for the two conditions (pleural effusion, and pneumonia), and also a discussion of those results; section “[Conclusions](#)” sums up the conclusions drawn from this work and suggests new directions for future work in this research area.

## Background

**Medical image retrieval.** The importance of having a good medical image retrieval system to help clinicians make a diagnosis was clearly pointed out in the previous section. Here, we will focus on presenting the most relevant CBIR works available in the literature. The main difficulties in the development of CBIR systems are related to the development of algorithms that generate useful semantic representations of medical images in order to effectively retrieve the most similar examples<sup>6</sup>, and on the integration of these algorithms in end-user applications<sup>7,8</sup>. We will focus on the first difficulty. In that regard, several works were presented in the literature to find the most suitable representation to perform the retrieval: Tizhoosh<sup>9</sup> explored the use of bar code annotations as an auxiliary method for feature-based image retrieval; Srinivas et al.<sup>10</sup> implemented a clustering method that uses dictionary learning to group large medical databases and relies on different similarity measures (e.g., Euclidean) to perform image retrieval; Hofmanninger and Langs<sup>11</sup> proposed the re-mapping of visual features extracted from medical imaging data based on weak labels to obtain descriptions of local image content capturing clinically relevant information; Seetharaman and Sathiamoorthy<sup>12</sup> presented a unified learning framework for heterogeneous medical image retrieval based on a full range auto-regressive model with a Bayesian approach to extract meaningful image features; Ma et al.<sup>13</sup> created a method that consists of a weighted graph whose nodes represent the images and edges measure their pairwise similarities; Nowaková et al.<sup>14</sup> presented a novel method for fuzzy medical image retrieval using vector quantisation with fuzzy signatures in conjunction with fuzzy S-trees; Qayyum et al.<sup>15</sup>, Ayyachamy et al.<sup>16</sup> and Owais et al.<sup>17</sup> trained CNNs on multimodal and multi-class data sets, and used the learned features and the classification results to retrieve medical images; Cai et al.<sup>18</sup> used a Siamese Network in the learning process, with the CNN of each branch being used to extract features, followed by

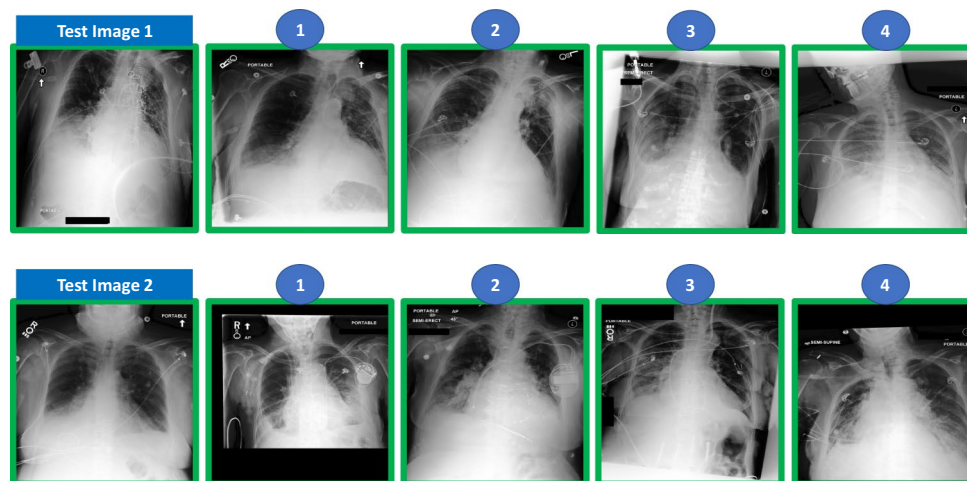
the application of a binary hash-mapping to reduce the dimensions of the feature vectors; Minarno et al.<sup>19</sup> used a CNN-based auto-encoder method in the feature extraction process to improve the results of the retrieval process; Mbilinyi et al.<sup>20</sup> used a deep metric learning approach and the triplet loss to learn a model that receives an image and a text description highlighting specific diagnoses the retrieved images should have. In summary, feature representation is performed in one of the following ways: statistical measures, hand-crafted features, learned features, or a combination of the previously mentioned strategies. However, to the best of our knowledge, none of the previously proposed approaches explicitly focuses the training process on the disease-related characteristics without requiring additional labels. In this work, we aim to utilize AI interpretability methods to guide the retrieval process, with focus on the disease and without necessitating any additional related label information.

**Explainable artificial intelligence.** In the last years, deep learning algorithms have been highly successful in medical image applications, in some cases even challenging human performance<sup>21</sup>. Nonetheless, both clinical and technical communities acknowledge that there are still several open challenges that need to be addressed. Particularly of interest for this study are the works that try to overcome the transparency and trust issues. As pointed out, most of these complex and successful models currently used to solve medical imaging problems work as black boxes (i.e., their internal logic is hidden to the user), without being able to explain their predictions in a human-understandable way<sup>22</sup>. Despite being a research field under development, there are already many approaches to obtain interpretability, or in a broader sense, to produce explanations for the decisions that models make. Interpretability research can be easily understood by looking at the three-stage categorization (pre-model, in-model and post-model) proposed by Kim and Doshi-Velez<sup>23</sup>. Pre-model methods focus on understanding the data distribution before building the model through exploratory data analysis<sup>24–27</sup>. In-model methods seek to integrate interpretability inside the model, either by relying on models based on rules<sup>28,29</sup>, based on cases<sup>30–32</sup>, through the use of regularization techniques (e.g., sparsity, monotonicity) during training<sup>33,34</sup>, by guiding the neural network into learning relevant concepts<sup>35,36</sup>, or by seeking to integrate causal knowledge into the network<sup>37,38</sup>. Finally, post-model methods are related to a posterior analysis of the model predictions, either producing saliency maps through gradient information<sup>39–41</sup>, deconvolution<sup>42,43</sup>, optimization<sup>44</sup>, decomposition<sup>45,46</sup>, or through a connection with high-level semantic concepts<sup>35,47,48</sup>. In this work, we will focus on post-model interpretability strategies, as we are interested in finding the most relevant regions for the medical decisions (explicit attention) without limiting in any way the learning process nor requiring any additional label. This can be done by identifying the areas of the image that mostly contribute to the final decision. To find these relevant regions, we used *Deep Taylor*<sup>46</sup>, which is a relevance propagation approach (similar to Layer-wise Relevance Propagation (LRP)<sup>45</sup>), that uses deep Taylor decomposition to efficiently assess the importance of single pixels in image classification problems. The choice of this interpretability method in specific was mainly driven by its recognized quality, but also because it was the method that produced the saliency maps more in-line with what our board-certified radiologist considered as relevant medical information.

**Attention mechanisms.** A different alternative to the use of post-hoc interpretability methods to focus the network into the disease-related characteristics is the use of implicit attention mechanisms. This application of attention mechanisms in deep learning algorithms was inspired by the field of psychology, according to which humans tend to selectively concentrate on a part of the information<sup>49</sup>. For instance, the human visual system tends to selectively focus on specific parts of an image while ignoring others<sup>50</sup>. The use of attention was initially proposed in Bahdanau et al.<sup>51</sup>, for the task of neural machine translation. In this work, the authors use an encoder-decoder architecture presenting two challenges: (1) the decoder needs to compress all the input information into a single fixed-length vector and pass it to the decoder; (2) ensuring model alignment between input and output sequences was not possible. Hence, it was necessary to develop an attention mechanism that could support the decoder in focusing on the relevant parts of the inputs<sup>52</sup>. Naturally, during the training phase, an extra task is added: the learning of the attention weights. Nevertheless, this approach showed improved results against the state-of-the-art and paved the way for the creation of novel attention-based methodologies. Attention models can be classified into different categories according to their input sequences, output sequences, candidate states (hidden states of the encoder) and query states (hidden states of the decoder)<sup>52</sup>. Of relevance for this work are self-attention and multi-level attention, with self-attention being when the query and candidate states belong to the same input sequence, and multi-level attention when we apply the attention mechanism on multiple levels of abstraction of the input sequence. Additional details on the attention mechanisms used in this work will be presented later when discussing their application in content-based image retrieval.

## Materials and methods

**Data.** For the experiments, we used the MIMIC Chest X-ray JPG, which is a large and publicly available database. It consists of chest radiographs already converted to JPG format, and their respective labels, which were derived from free-text radiology reports<sup>53</sup>. In this database, there are 377,110 JPG format chest radiographs with associated structured labels. Institutional approval was granted for the use of the patient datasets in research studies for diagnostic and therapeutic purposes. Approval was granted on the grounds of existing datasets. All methods were carried out in accordance with relevant guidelines and regulations. From all the available labels, we focused our attention on the following conditions: Pleural Effusion, and (Potential) Pneumonia. Even though the provided label is pneumonia, our main board-certified radiologist considers a Chest X-ray does not allow a conclusive diagnosis of pneumonia, thus, we perform our analysis referring to the cases as having potential pneumonia. To train and evaluate our models, the original splits selected by the data providers<sup>53</sup> were used, i.e., we considered the training fold for training, the validation fold to select the final model, and the test set to evaluate the performance. From these splits, we only considered frontal view images, either acquired in an AP



**Figure 2.** Example of test cases and ranking annotation for pleural effusion condition (the top 4 is shown) performed by the radiologist. The numbers on top of the image represent the ranking position of the image in the catalogue (when compared to the query/test image). The green box means pleural effusion case (according to the dataset label).

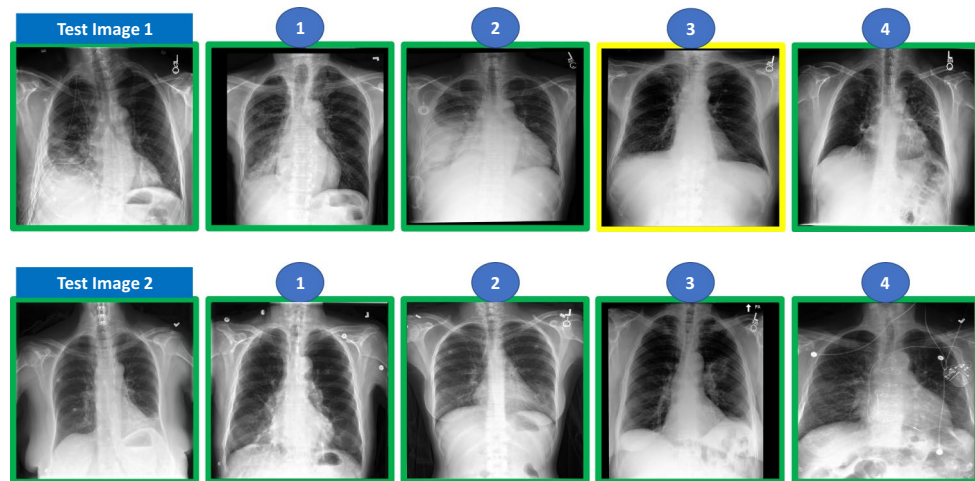
(Anterior-posterior) or PA (Posterior-anterior) setting. For completeness, one of the analysis was done with AP view images (pleural effusion) and the other with PA view images (potential pneumonia). For each condition there are four possible annotations: 1, the label was positively mentioned in the associated study; 0, the label was negatively mentioned in the associated study; -1, the label was either mentioned with uncertainty or with ambiguous language in the report; and missing, no mention of the label was made in the report. Regarding our experiments, we only considered images associated with the labels 1 and 0, thus, not introducing uncertainty in the training of the models. Given that selection, we ended up with 61203 training images, 534 validation images, and 1072 test images for pleural effusion and 18226 training images, 133 validation images, and 258 test images for pneumonia.

To evaluate the performance of the different types of methods in the retrieval task, we split the test data into query and catalogue images. For pleural effusion, we considered ten query images, each associated with ten catalogue images. For pneumonia, the test set was considerably smaller, and we only considered five query images, each associated with also ten catalogue images. For both conditions, query and catalogue images were randomly chosen. All images considered for the evaluation were labelled by our main board-certified radiologist. Besides labelling all images, our main radiologist also ranked the associated catalogue images in relation to each of the test images, considering the similarity in terms of disease severity (Fig. 2 shows an example of the ranking annotations provided for pleural effusion and Fig. 3 shows an example of the ranking annotations provided for pneumonia). In addition to our main board-certified radiologist, four other experienced radiologists also analysed and ranked the cases in order to assess the inter-rater variability in such a task.

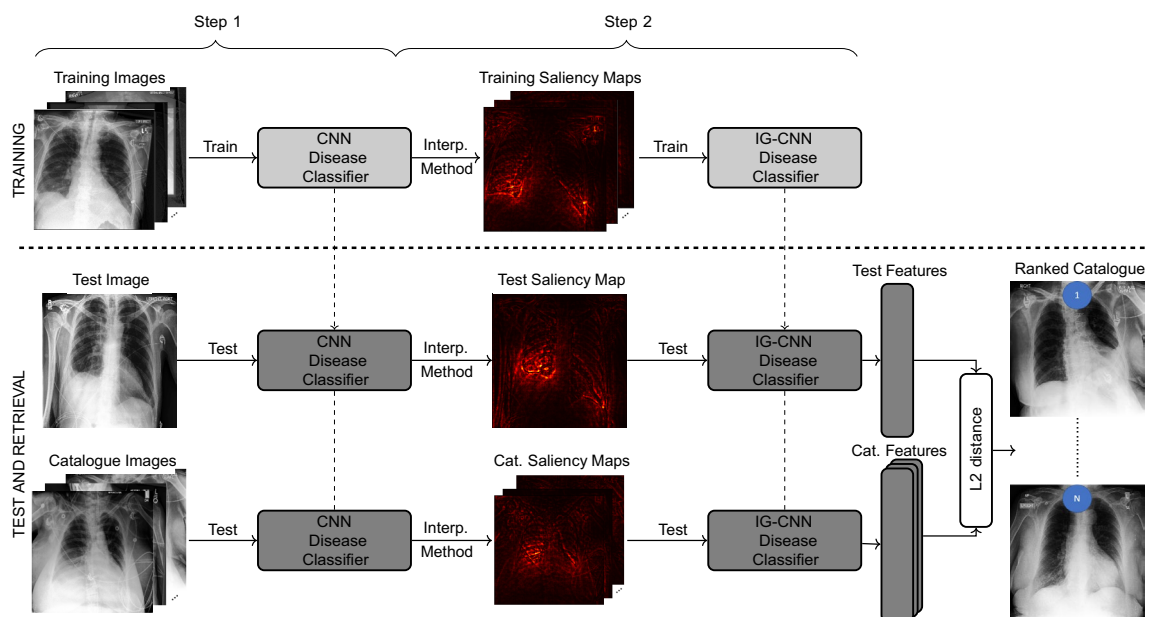
**Methods.** *Structural similarity index (SSIM).* The first method to be considered for evaluation in the retrieval task is the classic statistically-based structural similarity index (SSIM)<sup>54</sup>. As in Silva et al.<sup>4</sup>, the SSIM was computed directly between test and catalogue images, using its default values. Since higher SSIM values represent higher similarity, the top retrieved image is the one with the highest similarity index.

*Convolutional neural network (CNN).* The second method to be considered is already a deep learning based method, where the relevant features are automatically identified<sup>5,11,55,56</sup>. As in Silva et al.<sup>4</sup>, we use the DenseNet-121<sup>57</sup> as our CNN architecture. However, in this work, we do not initialize its weights with the ImageNet pre-training, but instead with the pleural effusion CheXpert CNN model from Silva et al.<sup>4</sup> for the pleural effusion condition, and with the pleural effusion model from this work for the pneumonia condition, as pre-training using data more similar to the final domain is more effective than using ImageNet pre-training<sup>58,59</sup>. Similarity between images is computed based on the Euclidean distance in the feature space of the previous to the last layer of the model. Since shorter distances represent higher similarity, the top retrieved image is the one with the shortest distance to the test image. The distance between two images is formalized in Eq. (1), where  $I_t$  represents the test image  $t$ ,  $I_c$  represents the catalogue image  $c$ ,  $\theta_{\text{CNN}}$  represents the CNN model parameters, and  $F$  represents the function that translates the original image into a latent representation constituted by the features in the previous to last layer of the network (i.e., in a vector of dimension 1024).

$$d_{\text{CNN}}(I_t, I_c) = \|F(\theta_{\text{CNN}}, I_t) - F(\theta_{\text{CNN}}, I_c)\|_2 \quad (1)$$



**Figure 3.** Example of test cases and ranking annotation for potential pneumonia condition (the top 4 are shown) performed by the radiologist. The numbers on top of the image represent the ranking position of the image in the catalogue (when compared to the query/test image). The green box means pneumonia case (according to the dataset label), yellow box means radiologist considers case as potential pneumonia and dataset label is no pneumonia.



**Figure 4.** Overview of the proposed Interpretability-guided approach. Blocks in light gray (■) mean neural networks are being trained (i.e., weights are being updated), whereas blocks in dark gray (■) represent trained neural networks (i.e., weights are fixed). In the saliency maps, brighter colors mean higher relevance. Blue circles indicate ranking positions. CNN represents the deep model used as baseline. IG-CNN represents the CNN model architecture being trained with saliency maps. The L2 distance is computed between the test image's latent features and each catalogue image's latent features.

*Interpretability-guided network (IG).* The third method being considered is the method proposed in Silva et al.<sup>4</sup> It uses the exact same architecture as the CNN model, but has as input the saliency maps, instead of the original images (Fig. 4) in order to focus the network into the disease-related characteristics. Those saliency maps are computed using the Deep Taylor interpretability method<sup>45</sup>, and are based on the previously presented CNN model. This time, the deep learning network was initialized with the CheXpert IG model from Silva et al.<sup>4</sup> for the pleural effusion condition, and with the IG pleural effusion model from this work for the pneumonia condition. As with the CNN approach, the similarity is computed based on the previous to last layer of the model. The distance between two images is formalized in Eq. (2), where  $I_t$  represents the test image  $t$ ,  $I_c$  the catalogue

image  $c$ ,  $\theta_{\text{CNN}}$  the CNN model parameters,  $\theta_{\text{IG}}$  the IG parameters,  $S$  the function that generates the saliency maps, and  $F$  the function that translates the original image into a latent representation constituted by the features in the previous to last layer of the network.

$$d_{\text{IG}}(I_t, I_c) = \|F(\theta_{\text{IG}}, S(\theta_{\text{CNN}}, I_t)) - F(\theta_{\text{IG}}, S(\theta_{\text{CNN}}, I_c))\|_2 \quad (2)$$

**Attention network (ATT).** Here, we add a fourth method to the comparison, one that is driven by attention mechanisms. Recently, a CNN with a multi-level dual-attention mechanism (MLDAM) has been proposed for macular optical coherence tomography classification<sup>60</sup>. The main novelty of this work in the context of medical image classification is the joint application of a *self-attention* and a *multi-level attention* mechanisms that allow the network to learn relevant features in coarser as well as finer sub-spaces. In their article<sup>60</sup>, the authors state that this technique enables the network to utilise the information of coarser features preventing loss of any useful information, thus enabling the network to yield more focused features and better convergence. Regarding the impact of the application of attention mechanisms in the interpretability of deep learning algorithms, Chen and Ross<sup>61</sup> proposed the joint use of a position attention module (PAM) and a channel attention module (CAM) to refine the pixel values at spatial and channel levels. These refined features are then fused through an element-wise sum. The authors performed an analysis of the saliency maps produced by the gradient-weighted class activation mapping (Grad-CAM)<sup>41</sup> and concluded that the use of attention modules had enabled the network to shift the focus on to the annular iris region.

In this work, we aimed to assure diversity in the levels and scales of the features extracted from the DenseNet-121<sup>57</sup>. Following the notation in Mishra et al.<sup>60</sup>, let  $I_A$ ,  $I_B$  and  $I_C$  be the multi-level features extracted from the backbone. We extracted features from different dense-blocks resulting in a  $I_A$  with shape [512, 28, 28], a  $I_B$  with shape [1024, 14, 14] and a  $I_C$  with shape [1024, 7, 7].

In line with the previous deep methods, the similarity is computed by measuring the Euclidean distance in the previous to last layer. The distance between two images is formalised in Eq. (3), where  $I_t$  represents the test image  $t$ ,  $I_c$  the catalogue image  $c$ ,  $\theta_{\text{ATT}}$  the Attention model parameters, and  $A$  represents the function that translates the original image into a latent representation constituted by the features in the previous to last layer of the network.

$$d_{\text{ATT}}(I_t, I_c) = \|A(\theta_{\text{ATT}}, I_t) - A(\theta_{\text{ATT}}, I_c)\|_2 \quad (3)$$

**Deep learning networks training.** All deep learning methods (i.e., CNN, IG, and ATT) were trained to solve binary classification tasks (e.g., pleural effusion vs. non-pleural effusion). Thus, we use the binary cross-entropy as our loss function (Eq. 4, where  $y$  is the binary indicator,  $\ln$  the natural logarithm,  $p$  the predicted probability, and  $\theta$  the model parameters).

$$\mathcal{L}(\theta) = -(y \ln(p(\theta)) + (1 - y) \ln(1 - p(\theta))) \quad (4)$$

For the pleural effusion condition, the deep learning models were trained for 10 epochs, with a batch size of 32, and using the *Adadelta* optimiser<sup>62</sup>. Since the data for the pleural effusion condition is highly imbalanced, the misclassifications were weighted with the inverse of the frequency of the respective class to promote a similar focus of the network in both classes<sup>63</sup>.

Regarding the pneumonia condition, the deep learning models were trained for 15 epochs, with a batch size of 32, and using the *Adam* optimiser<sup>64</sup> with a learning rate  $l_r = 1 \times 10^{-4}$ . The *Adam* optimiser was chosen over the *Adadelta* due to converging issues during the training of the CNN model, and was kept for the training of the other deep models (IG and ATT) for consistency.

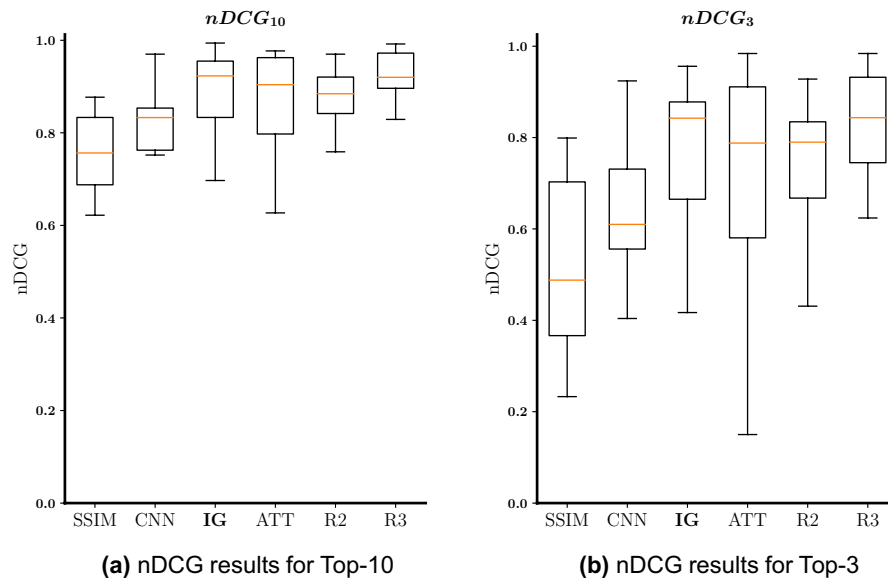
For both conditions, small rotations and translations were used as data augmentation. Hyperparameter values were empirically optimised for the CNN models and replicated for all the others. Final models were selected based on the F1 score (Eq. 5) in the validation set.

$$\text{F1} = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \quad (5)$$

We note that the training process was agnostic to the ranking task at hand. No information of ranking was provided at any point, neither in the loss function nor in the selection of the best performing model in validation.

The methods were implemented using Keras<sup>65</sup> with TensorFlow backend in a workstation equipped with an NVIDIA Tesla V100 (32 GB) GPU. For the generation of the saliency maps, we used the *iNNvestigate* toolbox<sup>66</sup> implementation of the Deep Taylor method, as in Silva et al.<sup>4</sup>.

**Evaluation and comparison.** The quality of the retrieval is evaluated by computing the normalised Discounted Cumulative Gain (nDCG)—Eq. (6), which is the normalised version of the Discounted Cumulative Gain (DCG)—Eq. (7), being it a common metric in learning to rank tasks<sup>67</sup>. The normalisation is done over the maximum possible value of the DCG metric (in our work, the maximum possible value is obtained when the ranking of the method is exactly the same as our ground-truth). The subscript  $p$  represents the number of retrieved images we are considering for the evaluation (e.g., when we perform the evaluation over the entire set of retrieved images,  $p = 10$ ). In Eq. (7),  $rel_i$  represents the relevance value assigned to the ranking position  $i$ , with the least similar image having relevance of 1 and the most similar image having relevance of 5.5 (i.e.,



**Figure 5.** Box-and-whisker plots regarding the nDCG results for the pleural effusion Top-10 (a) and Top-3 (b) retrieved images. SSIM is the statistically-based baseline, CNN is the CNN-based baseline, IG is the proposed interpretability-guided approach, ATT is the attention method, R2 is the ranking provided by the second board-certified radiologist, and R3 is the ranking provided by the third board-certified radiologist.

the relevance of two contiguous positions differs by 0.5). Thus, the first positions of the catalogue ranking have more importance than the last ones, with the importance being gradually reduced as we go from the first to last ranked image.

$$\text{nDCG}_p = \frac{\text{DCG}_p}{\text{IDCG}_p} \quad (6)$$

$$\text{DCG}_p = \sum_{i=1}^p \frac{2^{\text{rel}_i} - 1}{\log_2(i + 1)} \quad (7)$$

In order to contextualize the retrieval results of our machine learning methods, we also asked our partner board-certified radiologists to provide their similarity rankings for the pleural effusion and pneumonia conditions. Thus, we are able to check the inter-rater variability in ranking tasks, also helping us to have a more complete evaluation of our methods' quality.

## Results and discussion

**Pleural effusion.** Our first experiments were conducted for the pleural effusion condition. All images used here were frontal X-ray images acquired in an AP view fashion. Thus, experiments were performed with 61203 training images, 534 validation images, and 1072 test images. The training images were used to find the optimal set of parameters, the validation images to select the final classification model, and the test images for the assessment. To evaluate the ranking quality, ten different query images and ten catalogues of ten images each were randomly created, splitting the test data into query and catalogue images by using ten different random seeds (keeping the proportion of the classes). Afterwards, our main board-certified radiologist provided us with a ranking of those ten images in relation to the respective query image, serving as our ground-truth ranking. Moreover, we also asked two other board-certified radiologists to provide their rankings in order to compare inter-rater variability with our models' performance.

In Fig. 5a, we present the nDCG results obtained with the statistical and machine learning models (i.e., SSIM, CNN, IG, and ATT) and also the results obtained by considering the rankings provided by two other radiologists (R2, and R3) for the Top-10 retrieved images. By observing the box-and-whisker plot, we conclude that the proposed interpretability-guided approach (IG) and the attention-based method (ATT) are the ones that lead to the best nDCG results for the Top-10 retrieved images, with the interpretability-guided approach outperforming the attention-driven method. Those results are in line with those from the other radiologists, demonstrating the high-quality of both methods. Furthermore, the CNN approach leads to better results than the SSIM method, as it was expected. The same can be observed in Fig. 5b, where the nDCG results for the Top-3 retrieved images are presented (in clinical practice having the three most similar images is typically enough to help the radiologist make the diagnosis). In this scenario, nDCG values are worse than in the previous experiment due to only considering the Top-3 retrieved images, highly penalizing a "failure" in one or more of these images. This also

contributed to an increase in the variability of the results obtained, particularly in the case of the ATT method. Nonetheless, IG and ATT approaches remained the best methods and are still in line with the performance of the two radiologists.

In Fig. 6, we show the Top-4 retrieved results obtained by each of the methods, and provided by the radiologists in comparison with the ground-truth defined by our main radiologist for one split, corresponding to a specific test case and catalogue. In this split, all machine learning methods (i.e., CNN, IG, and ATT) attained extremely high nDCG results. Both CNN and IG retrieved the same Top-4 images, with the only difference being the ranking of these four images, with IG's ranking being closer to the one provided by our main radiologist than CNN's ranking. Even though the ATT's Top-4 retrieved images differ from the ones selected by our main radiologist, one of those images was also selected by one of the other radiologists (i.e., R2). SSIM was the worst method, selecting the least similar image (a non-pleural effusion case) for the Top-4 retrieved images.

In terms of classification performance in the entire test set (the 1072 test images), the interpretability-driven model was the one leading to the best results (F1-score = 0.862), followed by the attention-based model (F1-score = 0.809), and the standard CNN model (F1-score=0.790).

**(Potential) Pneumonia.** The following experiments were conducted for the pneumonia condition. All images used here were frontal X-ray images acquired in a PA view fashion. Thus, for the experiments we considered 18226 training images, 133 validation images, and 258 test images. For the ranking evaluation, five different query images and five catalogues of ten images each were created, splitting the data into query and catalogue images by using five different seeds (keeping the proportion of the classes). Afterwards, the catalogue's ten images were ranked in terms of their potential as pneumonia cases to the respective query image. Even though our dataset annotations for training and validation were pneumonia annotations, our main radiologist considers a Chest X-ray as only indicative of potential pneumonia, and not of a definitive diagnosis. Thus, catalogue images were ranked having in mind their potential as pneumonia cases. In Fig. 7a, we present the nDCG results obtained with the statistical and machine learning models and also the results obtained by considering the rankings provided by two other radiologists (R4, and R5). By observing the box-and-whisker plot, we infer that the proposed interpretability-guided approach (IG) is the method with the best retrieval ranking performance. On the contrary, for this condition, the attention-based method (ATT) had a poor ranking performance, obtaining nDCG results that were worse than the ones obtained with our deep learning baseline method (CNN), and only surpassing the performance of the statistical baseline (SSIM). The relative performance of the four methods was the same when we measured the nDCG performance for the Top-3 retrieved images (as shown in Fig. 7b). IG's results also fall within the inter-rater variability of the radiologists, which demonstrates the quality of the method.

When we compare the quantitative results obtained for the pneumonia condition with the ones obtained for the pleural effusion, we observe that they were considerably worse in general. That may be due to pneumonia being a more difficult to diagnose condition, and also to different interpretations of what a pneumonia Chest X-ray is (in several catalogue images, there was a disagreement between MIMIC-CXR label, and the diagnosis provided by our main board-certified radiologist). In Fig. 8, we present an example query case and the respective Top-4 retrieved images obtained by each of the methods, and provided by the radiologists in comparison with the ground-truth defined by our main radiologist. In this split, all deep learning models had a reasonably good ranking performance, with the interpretability-guided approach (IG), and the attention-based method (ATT) retrieving in the first position the most similar image in terms of pneumonia to the test image. As can be observed here, some images in this catalogue had different diagnoses given by MIMIC-CXR and by our main radiologist, namely the fourth and sixth ranking positions (images with orange boxes). Moreover, the direction of the disagreement was the same, with our main radiologist considering the cases as of potential pneumonia, and the MIMIC-CXR label being non-pneumonia. Nonetheless, even with this label disagreement, the performance obtained with our interpretability-guided approach (IG) was reasonably good, exceeding nDCG results of 0.88 for the Top-10 retrieved images in all but one split.

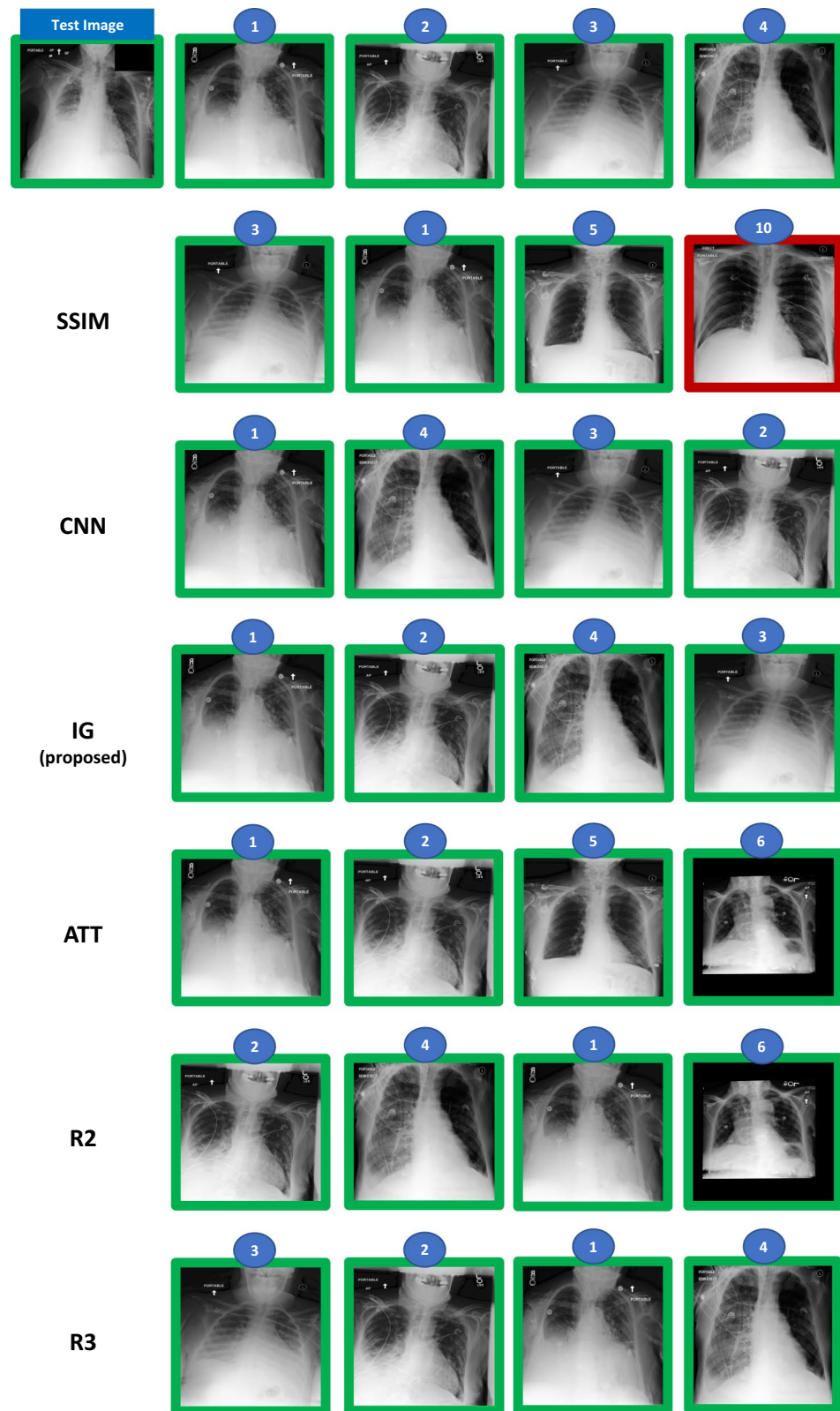
In terms of classification performance in the entire test set (the 258 test images), this time, the attention model was the one leading to the best results (F1-score = 0.655), followed by the interpretability-driven model (F1-score = 0.645), and the standard CNN model (F1-score = 0.620).

**Ablation study.** We also studied the relevance of training with the saliency maps, and not only using them to compute the features in the semantic space. In Fig. 9, we present the Top-10 nDCG results for both pleural effusion and potential pneumonia, considering the CNN baseline model and these two versions, i.e., using only the saliency maps as inputs to the CNN model—CNN(IG)—and our proposed method where we use the saliency maps both in the training and retrieval processes—IG. By observing Fig. 9, we conclude that the use of saliency maps in the training process helps to attain better results, which can be explained by an increase in the focus of the network on the relevant disease regions, and by learning this new saliency map distribution.

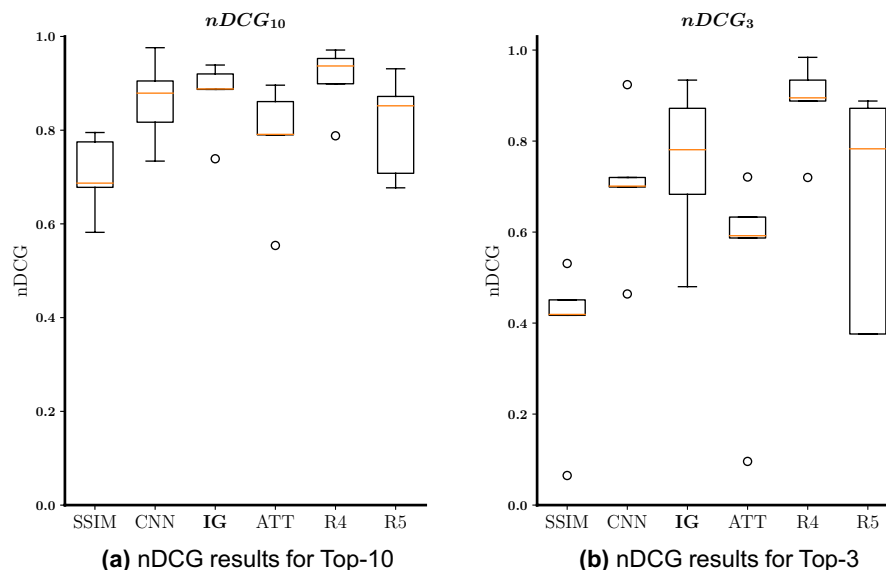
## Conclusions and future work

We have investigated the use of different content-based image retrieval methods in a Chest X-ray retrieval task, intending to study their potential to support a medical diagnosis. For radiologists, more important than having a decision support system providing prediction labels is to have a system that is able to present them with similar clinical cases, as it is usually the way they proceed when encountering a difficult diagnosis scenario. Moreover, radiologists feel more comfortable working with images than with textual descriptions, motivating the use of case-based reasoning or explainability.





**Figure 6.** Example of test case and the Top-4 retrieved images given by each of the radiologists (R1 = ground-truth, R2, and R3) and each of the machine learning methods. In this split, both the CNN, IG, and ATT obtained nDCGs (Top-10) > 0.9. The green box means pleural effusion case and the red box means no pleural effusion (according to the dataset label).

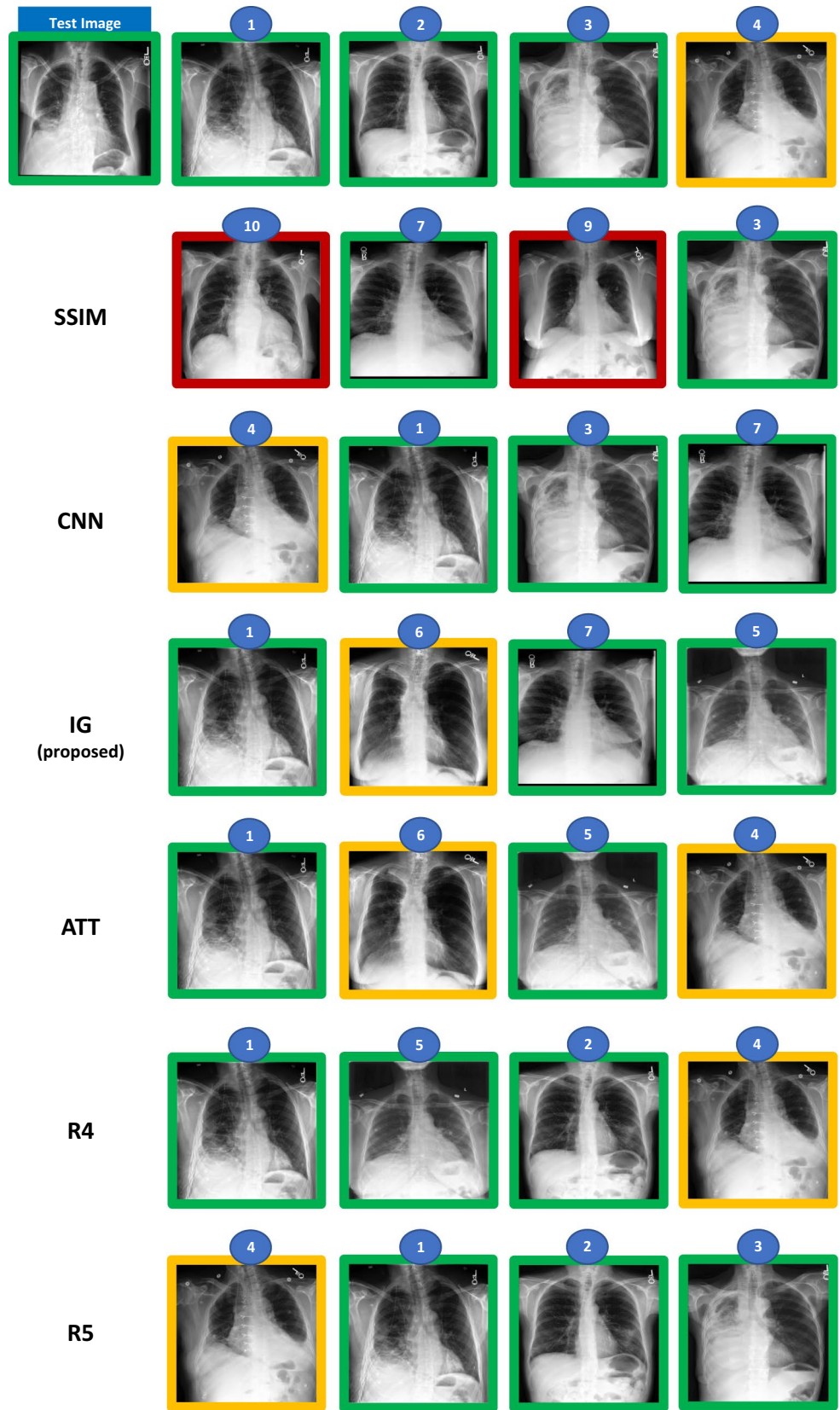


**Figure 7.** Box-and-whisker plots regarding the nDCG results for (potential) pneumonia Top-10 (a) and Top-3 (b) retrieved images. SSIM is the statistically-based baseline, CNN is the CNN-based baseline, IG is the proposed interpretability-guided approach, ATT is the attention method, R4 is the ranking provided by the fourth board-certified radiologist, and R5 is the ranking provided by the fifth board-certified radiologist.

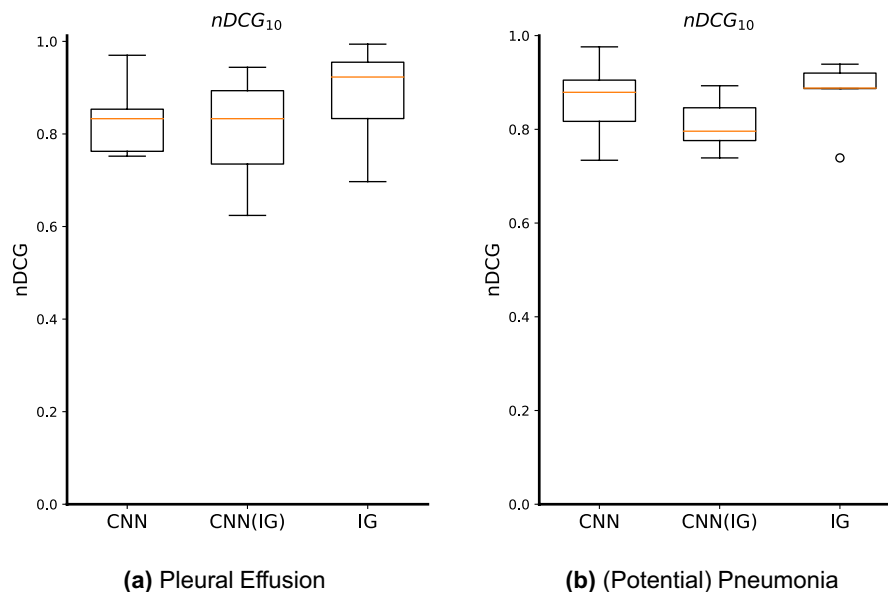
For the medical image retrieval to be successful, the comparison between images has to take into account the particular nature of medical images, i.e., that the information of interest is commonly found in a specific region of the image, while the remaining information is irrelevant. Indeed, the structural similarity index method showed the poorest performance for both conditions, demonstrating that a general image comparison does not represent disease similarity. Driven by that notion, we proposed an interpretability-guided approach and investigated the use of attention mechanisms for the retrieval task. The proposed interpretability-guided medical image retrieval approach outperformed all the other studied methods for both considered conditions. Our approach has an explicit attention mechanism that is also more intelligible than the implicit attention mechanism of the attention-driven method, leading to a more interpretable solution. Moreover, it obtained a performance in line with other human experts (board-certified radiologists) for both conditions. In turn, the attention-based medical image retrieval method had an excellent performance for the pleural effusion condition (in line with both our proposed method and the other radiologists) but failed for the pneumonia condition.

It is important to emphasize that all methods were only trained to solve binary classification tasks, not using any ranking information. This means that the annotation effort required is significantly lower than if ranking information was also needed. Even though we did not use ranking information in the training process, our proposed approach correctly captured the ranking information, obtaining excellent nDCG results for both conditions. Test and catalogue images apart from being ranked were also labelled by our main board-certified radiologist. Particularly for the pneumonia condition, we observed some disagreements in the diagnosis, which may be indicative of the usage of different definitions, and may hinder the method's performance. Nonetheless, even for the pneumonia condition, our proposed method obtained an excellent ranking performance.

This work aims to be the first step towards a deeper focus in medical image retrieval as a decision support system, helping radiologists make better and quicker decisions. However, further studies and investigations are required in order to translate these algorithms into the clinics. Considering the evaluation aspect, it is crucial to have more extensive studies, more datasets, other clinical problems, and more radiologists involved in the annotation and evaluation process. Regarding the technical side, there are several open problems or investigation opportunities, namely, the use of multimodal data, the introduction of causal knowledge<sup>68</sup>, privacy-preserving image retrieval<sup>69</sup>, and also exploring federated learning settings<sup>70</sup>. By the use of multimodal data, we mean the integration of the clinical reports in the learning process, also with the possibility of accompanying the top retrieved images with a generated clinical report to provide complementary information, which can be particularly interesting when the end-user is a general practitioner instead of a radiologist. In this work, we observed that by using post-hoc interpretability saliency maps, we were able to focus model attention into more clinically relevant regions. However, those methods are only able to capture correlations. Thus, they may also focus on confounding information<sup>71</sup>. In order to prevent this from happening, the integration of a causal structure is essential, or via effective interpretability-guided inductive biases as reported recently in Mahapatra et al.<sup>72,73</sup>. For clinical applications where personal characteristics are exposed, primarily if these systems are used for educational purposes, where the images are shown to unauthorized personnel, it is extremely important to anonymize the retrieved cases before showing them. Even though Montenegro et al.<sup>69,74</sup> already explored the use of privacy-preserving methods to anonymize medical images, further research is required in order to improve realism and preservation of clinical information. Finally, it is also relevant to explore federated learning settings



**Figure 8.** Example of test case and the Top-4 retrieved images given by each of the radiologists (R1 = ground truth, R4, and R5) and each of the machine learning methods. In this split, both the CNN, IG, and ATT obtained nDCGs (Top-10) > 0.8. The green box means potential pneumonia case, red box means no potential pneumonia, and orange box means disagreement between R1 and label, with R1 considering the case as potential pneumonia.



**Figure 9.** Box-and-whisker plots regarding the nDCG results for Top-10 retrieved images. CNN is the CNN baseline model, CNN(IG) is the CNN model having as inputs the Deep Taylor saliency maps, and IG is the proposed interpretability-based approach (i.e., it was trained (fine-tuned) with saliency maps, and has as inputs also saliency maps).

for this particular purpose, as the training process would benefit considerably from using different datasets acquired with different scanners and representing different population characteristics.

### Data availability

The data that support the findings of this study are available from PhysioNet (<https://physionet.org/content/mimic-cxr-jpg/2.0.0/>) but restrictions apply to the availability of these data, which were used under license for the current study, and so are not publicly available. Data and ranking annotations performed by the radiologists are however available from the authors upon reasonable request (through the following email contact: wilson.j.silva@inesctec.pt) and with permission of PhysioNet.

Received: 16 May 2022; Accepted: 23 November 2022

Published online: 01 December 2022

### References

- McDonald, R. J. *et al.* The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. *Acad. Radiol.* **22**, 1191–1198. <https://doi.org/10.1016/j.acra.2015.05.007> (2015).
- Lee, C. S., Nagy, P. G., Weaver, S. J. & Newman-Toker, D. E. Cognitive and system factors contributing to diagnostic errors in radiology. *Am. J. Roentgenol.* **201**, 611–617. <https://doi.org/10.2214/AJR.12.10375> (2013).
- Vosshenrich, J. *et al.* Quantifying radiology resident fatigue: Analysis of preliminary reports. *Radiology* **298**, 632–639 (2021).
- Silva, W., Poellinger, A., Cardoso, J. S. & Reyes, M. Interpretability-guided content-based medical image retrieval. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 305–314 (Springer, 2020).
- Li, Z., Zhang, X., Müller, H. & Zhang, S. Large-scale retrieval for medical image analytics: A comprehensive review. *Med. Image Anal.* **43**, 66–84 (2018).
- Das, P. & Neelima, A. An overview of approaches for content-based medical image retrieval. *Int. J. Multimedia Inf. Retrieval* **6**, 271–280 (2017).
- Zhuang, Y. *et al.* Efficient and robust large medical image retrieval in mobile cloud computing environment. *Inf. Sci.* **263**, 60–86 (2014).
- Grace, R. K., Manimegalai, R. & Kumar, S. S. Medical image retrieval system in grid using hadoop framework. In *2014 International Conference on Computational Science and Computational Intelligence*, vol. 1 144–148 (IEEE, 2014).
- Tizhoosh, H. R. Barcode annotations for medical image retrieval: A preliminary investigation. In *2015 IEEE International Conference on Image Processing (ICIP)* 818–822 (IEEE, 2015).
- Srinivas, M., Naidu, R. R., Sastry, C. S. & Mohan, C. K. Content based medical image retrieval using dictionary learning. *Neurocomputing* **168**, 880–895 (2015).
- Hofmanninger, J. & Langs, G. Mapping visual features to semantic profiles for retrieval in medical imaging. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 457–465 (2015).
- Seetharaman, K. & Sathiamoorthy, S. A unified learning framework for content based medical image retrieval using a statistical model. *J. King Saud Univ.-Comput. Inf. Sci.* **28**, 110–124 (2016).
- Ma, L. *et al.* A new method of content based medical image retrieval and its applications to ct imaging sign retrieval. *J. Biomed. Inf.* **66**, 148–158 (2017).
- Nowaková, J., Prilepok, M. & Snašel, V. Medical image retrieval using vector quantization and fuzzy s-tree. *J. Med. Syst.* **41**, 1–16 (2017).
- Qayyum, A., Anwar, S. M., Awais, M. & Majid, M. Medical image retrieval using deep convolutional neural network. *Neurocomputing* **266**, 8–20 (2017).

16. Ayyachamy, S., Alex, V., Khened, M. & Krishnamurthi, G. Medical image retrieval using resnet-18. In *Medical Imaging 2019: Imaging Informatics for Healthcare, Research, and Applications*, vol. 10954 1095410 (International Society for Optics and Photonics, 2019).
17. Owais, M., Arsalan, M., Choi, J. & Park, K. R. Effective diagnosis and treatment through content-based medical image retrieval (cbmir) by using artificial intelligence. *J. Clin. Med.* **8**, 462 (2019).
18. Cai, Y., Li, Y., Qiu, C., Ma, J. & Gao, X. Medical image retrieval based on convolutional neural network and supervised hashing. *IEEE Access* **7**, 51877–51885 (2019).
19. Minarno, A. E., Ghufiron, K. M., Sabrila, T. S., Husniah, L. & Sumadi, F. D. S. Cnn based autoencoder application in breast cancer image retrieval. In *2021 International Seminar on Intelligent Technology and Its Applications (ISITIA)* 29–34 (IEEE, 2021).
20. Mbilyinyi, A. & Schuldt, H. Retrieving chest X-rays for differential diagnosis: A deep metric learning approach. In *2021 IEEE EMBS International Conference on Biomedical and Health Informatics (BHI)* 1–4 (IEEE, 2021).
21. McKinney, S. M. *et al.* International evaluation of an AI system for breast cancer screening. *Nature* **577**, 89–94. <https://doi.org/10.1038/s41586-019-1799-6> (2020).
22. Rudin, C. *Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*. [arXiv:1811.10154](https://arxiv.org/abs/1811.10154) [cs, stat] (2019).
23. Kim, B. & Doshi-Velez, F. In *Interpretable Machine Learning: The Fuss, The Concrete and The Questions*. *ICML Tutorial on Interpretable Machine Learning* (2017).
24. Tukey, J. W. *Exploratory Data Analysis* Vol. 2 (Reading, Mass., 1977).
25. Varshney, K. R., Rasmussen, J. C., Mojsilovic, A., Singh, M. & DiMicco, J. M. Interactive visual salesforce analytics. In *ICIS* (2012).
26. Kim, B. & Shah, J. A. & Doshi-Velez, F. In *A Generative Approach to Interpretable Feature Selection and Extraction, Mind the Gap* (2015).
27. Kim, B., Khanna, R. & Koyejo, O. O. Examples are not enough, learn to criticize! criticism for interpretability. *Adv. Neural Inf. Process. Syst.* **29**, 251 (2016).
28. Rivest, R. L. Learning decision lists. *Mach. Learn.* **2**, 229–246. <https://doi.org/10.1007/BF00058680> (1987).
29. Rudin, C. & Ustun, B. Optimized scoring systems: Toward trust in machine learning for healthcare and criminal justice. *Interfaces* **48**, 449–466 (2018).
30. Kim, B., Rudin, C. & Shah, J. The bayesian case model: A generative approach for case-based reasoning and prototype classification. In *Proceedings of the 27th International Conference on Neural Information Processing Systems—Volume 2, NIPS'14 1952–1960* (MIT Press, 2014).
31. Chen, C. *et al.* *This Looks Like That: Deep Learning for Interpretable Image Recognition*. [arXiv:1806.10574](https://arxiv.org/abs/1806.10574) (2018).
32. Barnett, A. J. *et al.* *iaia-bl: A Case-Based Interpretable Deep Learning Model for Classification of Mass Lesions in Digital Mammography*. [arXiv:2103.12308](https://arxiv.org/abs/2103.12308) (2021).
33. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436–444. <https://doi.org/10.1038/nature14539> (2015).
34. Gupta, M. *et al.* Monotonic calibrated interpolated look-up tables. *J. Mach. Learn. Res.* **2016**, 563 (2016).
35. Kim, B. *et al.* Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). In *International Conference on Machine Learning* 2668–2677 (PMLR, 2018).
36. Chen, Z., Bei, Y. & Rudin, C. Concept whitening for interpretable image recognition. *Nature Mach. Intell.* **2**, 772–782 (2020).
37. Wang, T. & Rudin, C. *Causal Rule Sets for Identifying Subgroups with Enhanced Treatment Effect*. [arXiv:1710.05426](https://arxiv.org/abs/1710.05426) (2017).
38. Gan, K., Li, A., Lipton, Z. & Tayur, S. Causal inference with selectively deconfounded data. In *International Conference on Artificial Intelligence and Statistics* 2791–2799 (PMLR, 2021).
39. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep inside convolutional networks: Visualising image classification models and saliency maps. In *Workshop at International Conference on Learning Representations* (Citeseer, 2014).
40. Shrikumar, A., Greenside, P. & Kundaje, A. Learning important features through propagating activation differences. In *International Conference on Machine Learning* 3145–3153 (PMLR, 2017).
41. Selvaraju, R. R. *et al.* Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE International Conference on Computer Vision* 618–626 (2017).
42. Zeiler, M. D. & Fergus, R. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision* 818–833 (Springer, 2014).
43. Springenberg, J. T., Dosovitskiy, A., Brox, T. & Riedmiller, M. In *Striving for Simplicity: The All Convolutional Net*. [arXiv:1412.6806](https://arxiv.org/abs/1412.6806) (2014).
44. Ribeiro, M. T., Singh, S. & Guestrin, C. Why Should I trust you? Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1135–1144 (2016).
45. Bach, S. *et al.* On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS ONE* **10**, e0130140 (2015).
46. Montavon, G., Lapuschkin, S., Binder, A., Samek, W. & Müller, K.-R. Explaining nonlinear classification decisions with deep taylor decomposition. *Pattern Recogn.* **65**, 211–222 (2017).
47. Silva, W., Fernandes, K., Cardoso, M. J. & Cardoso, J. S. Towards complementary explanations using deep neural networks. In *Understanding and Interpreting Machine Learning in Medical Image Computing Applications* 133–140 (Springer, 2018).
48. Silva, W., Fernandes, K. & Cardoso, J. S. How to produce complementary explanations using an ensemble model. In *2019 International Joint Conference on Neural Networks (IJCNN)* 1–8 (IEEE, 2019).
49. Wang, F. & Tax, D. M. In *Survey on the Attention Based rnn Model and Its Applications in Computer Vision*. [arXiv:1601.06823](https://arxiv.org/abs/1601.06823) (2016).
50. Xu, K. *et al.* Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning* 2048–2057 (PMLR, 2015).
51. Bahdanau, D., Cho, K. & Bengio, Y. *Neural Machine Translation by Jointly Learning to Align and Translate*. [arXiv:1409.0473](https://arxiv.org/abs/1409.0473) (2014).
52. Chaudhari, S., Mithal, V., Polatkan, G. & Ramanath, R. An attentive survey of attention models. *ACM Trans. Intell. Syst. Technol (TIST)* **12**, 1–32 (2021).
53. Johnson, A. E. *et al.* Mimic-cxr, a de-identified publicly available database of chest radiographs with free-text reports. *Sci. Data* **6**, 1–8 (2019).
54. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **13**, 600–612 (2004).
55. Shin, H.-C. *et al.* Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE Trans. Med. Imaging* **35**, 1285–1298 (2016).
56. Wolterink, J. M., Leiner, T., Viergever, M. A. & Išgum, I. Automatic coronary calcium scoring in cardiac ct angiography using convolutional neural networks. In *International Conference on Medical Image Computing and Computer-Assisted Intervention* 589–596 (Springer, 2015).
57. Huang, G., Liu, Z., Van Der Maaten, L. & Weinberger, K. Q. Densely connected convolutional networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 4700–4708 (2017).
58. Weatheritt, J., Rueckert, D. & Wolz, R. Transfer learning for brain segmentation: Pre-task selection and data limitations. In *Annual Conference on Medical Image Understanding and Analysis* 118–130 (Springer, 2020).
59. Mustafa, B. *et al.* *Supervised Transfer Learning at Scale for Medical Imaging*. [arXiv:2101.05913](https://arxiv.org/abs/2101.05913) (2021).

60. Mishra, S. S., Mandal, B. & Puhan, N. B. Multi-level dual-attention based cnn for macular optical coherence tomography classification. *IEEE Signal Process. Lett.* **26**, 1793–1797 (2019).
61. Chen, C. & Ross, A. An explainable attention-guided iris presentation attack detector. In *WACV (Workshops)* 97–106 (2021).
62. Zeiler, M. D. *Adadelta: An Adaptive Learning Rate Method*. [arXiv:1212.5701](https://arxiv.org/abs/1212.5701) (2012).
63. Sun, Y., Kamel, M. S., Wong, A. K. & Wang, Y. Cost-sensitive boosting for classification of imbalanced data. *Pattern Recogn.* **40**, 3358–3378 (2007).
64. Kingma, D. P. & Ba, J. *Adam: A Method for Stochastic Optimization*. [arXiv:1412.6980](https://arxiv.org/abs/1412.6980) (2014).
65. Chollet, F. *et al.* Keras. <https://keras.io> (2015).
66. Alber, M. *et al.* Investigate neural networks!. *J. Mach. Learn. Res.* **20**, 1–8 (2019).
67. Fernandes, K. & Cardoso, J. S. Hypothesis transfer learning based on structural model similarity. *Neural Comput. Appl.* **31**, 3417–3430 (2019).
68. Castro, D. C., Walker, I. & Glocker, B. Causality matters in medical imaging. *Nat. Commun.* **11**, 1–10 (2020).
69. Montenegro, H., Silva, W. & Cardoso, J. S. Privacy-preserving generative adversarial network for case-based explainability in medical image analysis. *IEEE Access* **9**, 148037–148047. <https://doi.org/10.1109/ACCESS.2021.3124844> (2021).
70. Ziller, A. *et al.* Medical imaging deep learning with differential privacy. *Sci. Rep.* **11**, 1–8 (2021).
71. Geirhos, R. *et al.* Shortcut learning in deep neural networks. *Nature Mach. Intell.* **2**, 665–673 (2020).
72. Mahapatra, D., Poellinger, A. & Reyes, M. Interpretability-guided inductive bias for deep learning based medical image. *Med. Image Anal.* **81**, 102551 (2022).
73. Mahapatra, D., Poellinger, A. & Reyes, M. Graph node based interpretability guided sample selection for active learning. *IEEE Trans. Med. Imaging* **2022**, 5263 (2022).
74. Montenegro, H., Silva, W. & Cardoso, J. S. Towards privacy-preserving explanations in medical image analysis. In *1st Workshop on Interpretable Machine Learning in Healthcare at ICML2021* (2021).

## Acknowledgements

This work was supported by National Funds through the Portuguese Funding Agency, FCT–Foundation for Science and Technology Portugal, under Project LA/P/0063/2020, and also by the Portuguese Foundation for Science and Technology - FCT within PhD grants SFRH/BD/139468/2018 and 2020.06434.BD. The authors thank the Swiss National Science Foundation grant number 198388, as well as the Lindenhof foundation for their grant support.

## Author contributions

W.S. and A.P. conceived the experiments; W.S. and T.G. conducted the experiments and the analysis of the results. A.P., K.H., E.S., V.C.O., and M.C.B. annotated the Chest X-ray cases; W.S., T.G., M.R. and J.S.C. wrote the manuscript. A.P. clinically supervised the project; M.R. and J.S.C. technically supervised the project; All authors reviewed the manuscript and agreed to its published version.

## Competing interests

The authors declare no competing interests.

## Additional information

**Correspondence** and requests for materials should be addressed to W.S.

**Reprints and permissions information** is available at [www.nature.com/reprints](http://www.nature.com/reprints).

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022, corrected publication 2023