Unlearning Spurious Correlations in Chest X-ray Classification

Misgina Tsighe Hagos[1,2][0000−0002−9318−9417], Kathleen M. Curran[1,3][0000−0003−0095−9337], and Brian Mac Namee[1,2][0000−0003−2518−0274]

[1] Science Foundation Ireland Centre for Research Training in Machine Learning
misgina.hagos@ucdconnect.ie
[2] School of Computer Science, University College Dublin, Ireland
brian.macnamee@ucd.ie
[3] School of Medicine, University College Dublin, Ireland
kathleen.curran@ucd.ie

Abstract. Medical image classification models are frequently trained using training datasets derived from multiple data sources. While leveraging multiple data sources is crucial for achieving model generalization, it is important to acknowledge that the diverse nature of these sources inherently introduces unintended confounders and other challenges that can impact both model accuracy and transparency. A notable confounding factor in medical image classification, particularly in musculoskeletal image classification, is skeletal maturation-induced bone growth observed during adolescence. We train a deep learning model using a Covid-19 chest X-ray dataset and we showcase how this dataset can lead to spurious correlations due to unintended confounding regions. eXplanation Based Learning (XBL) is a deep learning approach that goes beyond interpretability by utilizing model explanations to interactively unlearn spurious correlations. This is achieved by integrating interactive user feedback, specifically feature annotations. In our study, we employed two non-demanding manual feedback mechanisms to implement an XBL-based approach for effectively eliminating these spurious correlations. Our results underscore the promising potential of XBL in constructing robust models even in the presence of confounding factors.

Keywords: Interactive Machine Learning · eXplanation Based Learning · Medical Image Classification · Chest X-ray

1 Introduction

While Computer-Assisted Diagnosis (CAD) holds promise in terms of cost and time savings, the performance of models trained on datasets with undetected biases is compromised when applied to new and external datasets. This limitation hinders the widespread adoption of CAD in clinical practice [21,16]. Therefore, it is crucial to identify biases within training datasets and mitigate their impact on trained models to ensure model effectiveness.

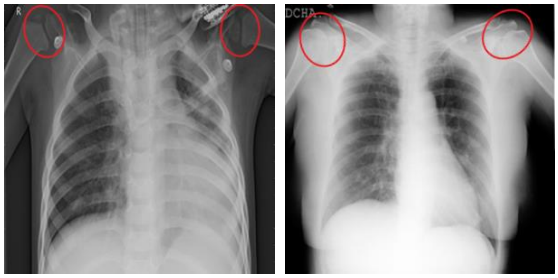

Fig. 1. In the left image, representing a child diagnosed with Viral pneumonia, the presence of Epiphyses on the humerus heads is evident, highlighted with red ellipses. Conversely, the right image portrays an adult patient with Covid-19, where the Epiphyses are replaced by Metaphyses, also highlighted with red ellipses.

For example, when building models for the differential diagnosis of pathology on chest X-rays (CXR) it is important to consider skeletal growth or ageing as a confounding factor. This factor can introduce bias into the dataset and potentially mislead trained models to prioritize age classification instead of accurately distinguishing between specific pathologies. The effect of skeletal growth on the appearance of bones necessitates careful consideration to ensure that a model focuses on the intended classification task rather than being influenced by age-related features.

An illustrative example of this scenario can be found in a recent study by Pfeuffer et al. [12]. In their research, they utilized the Covid-19 CXR dataset [4], which includes a category comprising CXR images of children. This dataset serves as a pertinent example to demonstrate the potential influence of age-related confounders, given the presence of images from pediatric patients. It comprises CXR images categorized into four groups: Normal, Covid, Lung opacity, and Viral pneumonia. However, a notable bias is introduced into the dataset due to the specific inclusion of the Viral pneumonia cases collected exclusively from children aged one to five years old [9]. This is illustrated in Figure 1 where confounding regions introduced due to anatomical differences between a child and an adult in CXR images are highlighted. Notably, the presence of Epiphyses in images from the Viral pneumonia category (which are all from children) is a confounding factor, as it is not inherently associated with the disease but can potentially mislead a model into erroneously associating it with the category. Addressing these anatomical differences is crucial to mitigate potential bias and ensure accurate analysis and classification in pediatric and adult populations.

Biases like this one pose a challenge to constructing transparent and robust models capable of avoiding spurious correlations. Spurious correlations refer to image regions that are mistakenly believed by the model to be associated with a specific category, despite lacking a genuine association.

While the exact extent of affected images remains unknown, it is important to note that the dataset also encompasses other confounding regions, such as

texts and timestamps. However, it is worth mentioning that these confounding regions are uniformly present across all categories, indicating that their impact is consistent throughout. For the purpose of this study, we specifically concentrate on understanding and mitigating the influence of musculoskeletal age in the dataset.

eXplanation Based Learning (XBL) represents a branch of Interactive Machine Learning (IML) that incorporates user feedback in the form of feature annotation during the training process to mitigate the influence of confounding regions [17]. By integrating user feedback into the training loop, XBL enables the model to progressively improve its performance and enhance its ability to differentiate between relevant and confounding features [6]. In addition to unlearning spurious correlations, XBL has the potential to enhance users' trust in a model [5]. By actively engaging users and incorporating their expertise, XBL promotes a collaborative learning environment, leading to increased trust in the model's outputs. This enhanced trust is crucial for the adoption and acceptance of models in real-world applications, particularly in domains where decisions have significant consequences, such as medical diagnosis.

XBL approaches typically add regularization to the loss function used when training a model, enabling it to disregard the impact of confounding regions. A typical XBL loss can be expressed as:

$$L = L_{CE} + L_{expl} + \lambda \sum_{i=0} \theta_i^2 \ , \tag{1}$$

where L_{CE} is categorical cross entropy loss that measures the discrepancy between the model's predictions and ground-truth labels; λ is a regularization term; θ refers to network parameters; and L_{expl} is an explanation loss. Explanation loss can be formulated as:

$$L_{expl} = \sum_{i=0}^{N} M_i \odot Exp(x_i) \ , \tag{2}$$

where N is the number of training instances, $x \in X$; M_i is a manual annotation of confounding regions in the input instance x_i; and $Exp(x_i)$ is a saliency-based model explanation for instance x_i, for example generated using Gradient weighted Class Activation Mapping (GradCAM) [17]. GradCAM is a feature attribution based model explanation that computes the attention of the learner model on different regions of an input image, indicating the regions that significantly contribute to the model's predictions [18]. This attention serves as a measure of the model's reliance on these regions when making predictions. The loss function, L_{expl}, is designed to increase as the learner's attention to the confounding regions increases. Overall, by leveraging GradCAM-based attention and the associated L_{expl} loss, XBL provides a mechanism for reducing a model's attention to confounding regions, enhancing the interpretability and transparency of a model's predictions.

As is seen in the inner ellipse of Figure 2, in XBL, the most common mode of user interaction is image feature annotation. This requires user engagement

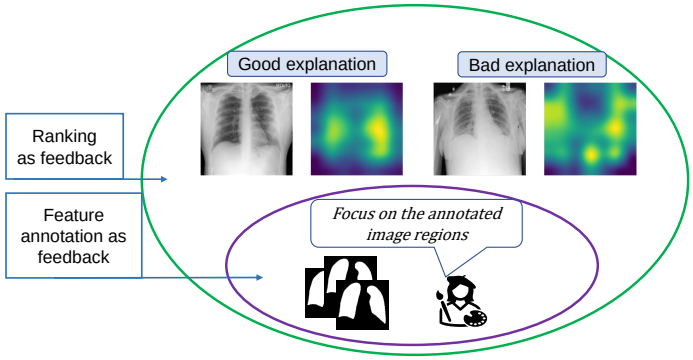

Fig. 2. The inner ellipse shows the typical mode of feedback collection where users annotate image features. The outer ellipse shows how our proposed approach requires only identification of one good and one bad explanation.

that is considerably more demanding than the simple instance labeling that most IML techniques require [22] and increases the time and cost of feedback collection. As can be seen in the outer ellipse of Figure 2, we are interested in lifting this pressure from users (feedback providers) and simplifying the interaction to ask for identification of two explanations as exemplary explanations and ranking them as good and bad explanations. This makes collecting feedback cheaper and faster. This kind of user interaction where users are asked for a ranking instead of category labels has also been found to increase inter-rater reliability and data collection efficiency [11]. We incorporate this feedback into model training through a contrastive triplet loss [3].

The main contributions of this paper are:

1. We propose the first type of eXplanation Based Learning (XBL) that can learn from only two exemplary explanations of two training images;
2. We present an approach to adopt triplet loss for XBL to incorporate the two exemplary explanations into an explanation loss;
3. Our experiments demonstrate that the proposed method achieves improved explanations and comparable classification performance when compared against a baseline model.

2 Related Work

2.1 Chest x-ray classification

A number of Covid-19 related datasets have been collated and deep learning based diagnosis solutions have been proposed due to the health emergency caused by Covid-19 and due to an urgent need for computer-aided diagnosis (CAD) of the disease [8]. In addition to training deep learning models from scratch,

transfer learning, where parameters of a pre-trained model are further trained to identify Covid-19, have been utilized [20]. Even though the array of datasets and deep learning models show promise in implementing CAD, care needs to be taken when the datasets are sourced from multiple imaging centers and/or the models are only validated on internal datasets. The Covid-19 CXR dataset, for example, has six sources at the time of writing this paper. This can result in unintended confounding regions in images in the dataset and subsequently spurious correlations in trained models [16].

2.2 eXplanation Based Learning

XBL can generally be categorized based on how feedback is used: (1) augmenting loss functions; and (2) augmenting training datasets.

Augmenting Loss Functions. As shown in Equation 1, approaches in this category add an explanation loss, L_{expl}, during model training to encourage focus on image regions that are considered relevant by user(s), or to ignore confounding regions [7]. Ross et al. [14] use an L_{expl} that penalizes a model with high input gradient model explanations on the wrong image regions based on user annotation,

$$L_{expl} = \sum_{n}^{N} \left[M_n \odot \frac{\partial}{\partial x_n} \sum_{k=1}^{K} \log \hat{y}_{nk} \right]^2, \tag{3}$$

for a function $f(X|\theta) = \hat{y} \in R^{N \times K}$ trained on N images, x_n, with K categories, where $M_n \in \{0,\ 1\}$ is user annotation of confounding image regions. Similarly, Shao et al. [19] use influence functions in place of input gradients to correct a model's behavior

Augmenting Training Dataset. In this category, a confounder-free dataset is added to an existing confounded training dataset to train models to avoid learning spurious correlations. In order to unlearn spurious correlations from a classifier that was trained on the Covid-19 dataset, Pfeuffer et al. [12] collected feature annotation on 3,000 chest x-ray images and augmented their training dataset. This approach, however, doesn't target unlearning or removing spurious correlations, but rather adds a new variety of data. This means models are being trained on a combination of the existing confounded training dataset and the their new dataset.

One thing all approaches to XBL described above have in common is the assumption that users will provide feature annotation for all training instances to refine or train a model. We believe that this level of user engagement hinders practical deployment of XBL because of the demanding nature and expense of feature annotation that is required [22]. It is, therefore, important to build an XBL method that can refine a trained model using a limited amount of user interaction and we propose eXemplary eXplanation Based Learning to achieve this.

3 eXemplary eXplanation Based Learning

User annotation of image features, or M, is an important prerequisite for typical XBL approaches (illustrated in Equation 1). We use eXemplary eXplanation Based Learning (eXBL) to reduce the time and resource complexity caused by the need for M. eXBL simplifies the expensive feature annotation requirement by replacing it with identification of just two exemplary explanations: a *Good explanation* (C_{good_i}) and a *Bad explanation* (C_{bad_j}) of two different instances, x_i and x_j. We pick the two exemplary explanations manually based on how much attention a model's explanation output gives to relevant image regions. A good explanation would be one that gives more focus to the lung and chest area rather than the irrelevant regions such as the Epiphyses, humerus head, and image backgrounds, while a bad explanation does the opposite.

We choose to use GradCAM model explanations because they have been found to be more sensitive to training label reshuffling and model parameter randomization than other saliency based explanations [1]; and they provide accurate explanations in medical image classifications [10].

We then compute product of the input instances and the Grad-CAM explanation in order to propagate input image information towards computing the loss and to avoid a bias that may be caused by only using a model's GradCAM explanation,

$$C_{good} := x_i \odot C_{good_i} \tag{4}$$

$$C_{bad} := x_j \odot C_{bad_j} \tag{5}$$

We then take inspiration from triplet loss [3] to incorporate C_{good} and C_{bad} into our explanation loss, L_{expl}. The main purpose of L_{expl} is to penalize a trainer according to similarity of model explanations of instance x to C_{good} and its difference from C_{bad}. We use Euclidean distance as a loss to compute the measure of dissimilarity, d (loss decreases as similarity to C_{good} is high and to C_{bad} is low).

$$d_{xg} := d(x \odot GradCAM(x), C_{good}) \tag{6}$$

$$d_{xb} := d(x \odot GradCAM(x), C_{bad}) \tag{7}$$

We train the model f to achieve $d_{xg} \ll d_{xb}$ for all x. We do this by adding a $margin = 1.0$ and translating it to: $d_{xg} < d_{xb} + margin$. We then compute the explanation loss as:

$$L_{expl} = \sum_{i}^{N} \max(d_{x_i g} - d_{x_i b} + margin, 0) \tag{8}$$

In addition to correctly classifying X, which is achieved through L_{CE}, this L_{expl} (Equation 8) trains f to output GradCAM values that resemble the good explanations and that differ from the bad explanations, thereby refining the model to focus on the relevant regions and to ignore confounding regions. L_{expl}

is zero, for a given sample x, unless $x \odot GradCAM(x)$ is much more similar to C_{bad} than it is to C_{good}—meaning $d_{xg} > d_{xb} + margin$.

4 Experiments

4.1 Data Collection and Preparation

To demonstrate eXBL we use the Covid-19 CXR dataset [4,13] described in Section 1. For model training we subsample 800 x-ray images per category to mitigate class imbalance, totaling 3,200 images. For validation and testing, we use 1,200 and 800 images respectively. We resize all images to 224 × 224 pixels. The dataset is also accompanied with feature annotation masks that show the lungs in each of the x-ray images collected from radiologists [13].

4.2 Model Training

We followed a transfer learning approach using a pre-trained MobileNetV2 model [15]. We chose to use MobileNetV2 because it achieved better performance at the CXR images classification task at a reduced computational cost after comparison among pre-trained models. In order for the training process to affect the GradCAM explanation outputs, we only freeze and reuse the first 50 layers of MobileNetV2 and retrain the rest of the convolutional layers with a custom classifier layer that we added (256 nodes with a ReLu activation with a 50% dropout followed by a Softmax layer with 4 nodes).

We first trained the MobileNetV2 to categorize the training set into the four classes using categorical cross entropy loss. It was trained for 60 epochs[4]. We refer to this model as the Unrefined model. We then use the Unrefined model to select the good and bad explanations displayed in Figure 2. Next, we employ our eXBL algorithm using the good and bad explanations to teach the Unrefined model to focus on relevant image regions by tuning its explanations to look like the good explanations and to differ from the bad explanations as much as possible. We use Euclidean distance to compute dissimilarity in adopting a version of the triplet loss for XBL. We refer to this model as the eXBL_{EUC} model and it was trained for 100 epochs using the same early stopping, learning rate, and optimizer as the Unrefined model.

For model evaluation, in addition to classification performance, we compute an objective explanation evaluation using Activation Precision [2] that measures how many of the pixels predicted as relevant by a model are actually relevant using existing feature annotation of the lungs in the employed dataset,

$$AP = \frac{1}{N} \sum_{n}^{N} \frac{\sum(T_\tau(GradCAM_\theta(x_n)) \odot A_{x_n})}{\sum(T_\tau(GradCAM_\theta(x_n)))} , \tag{9}$$

[4] The model was trained with an early stop monitoring the validation loss at a patience of five epochs and a decaying learning rate = 1e-04 using an Adam optimizer.

where x_n is a test instance, A_{x_n} is feature annotation of lungs in the dataset, $GradCAM_\theta(x_n)$ holds the GradCAM explanation of x_n generated from a trained model, and T_τ is a threshold function that finds the (100-τ) percentile value and sets elements of the explanation, $GradCAM_\theta(x_n)$, below this value to zero and the remaining elements to one. In our experiments, we use $\tau = 5\%$.

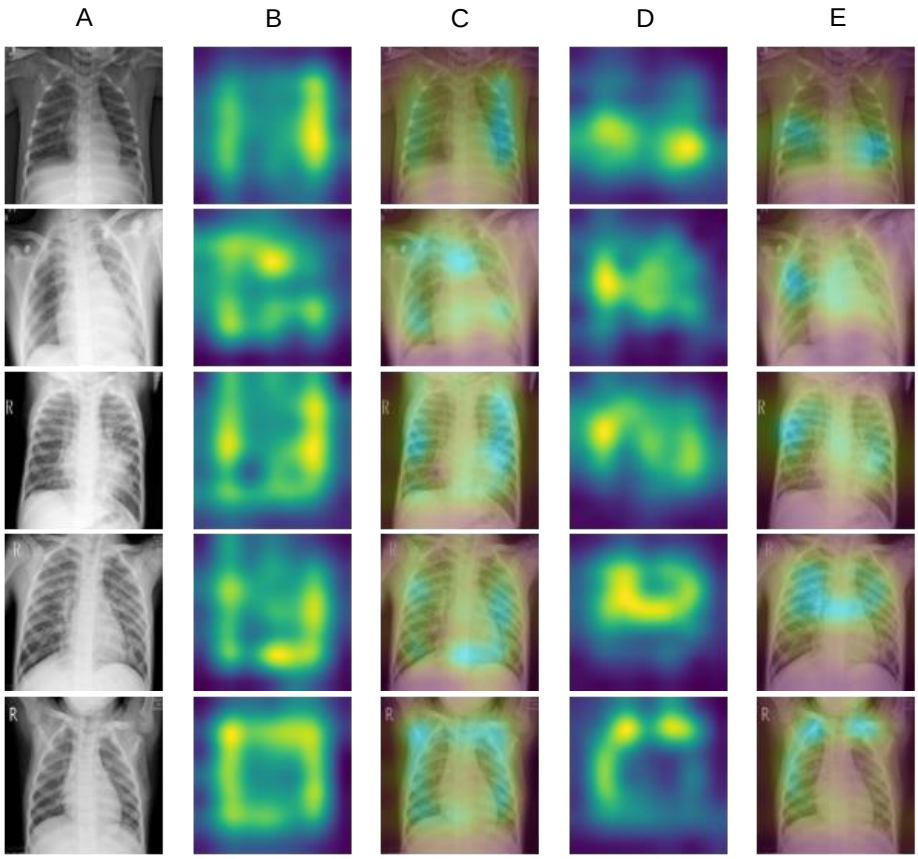

Fig. 3. Sample outputs of Viral Pneumonia category. (A) Input images; (B) GradCAM outputs for Unrefined model and (C) their overlay over input images; (D) GradCAM outputs for eXBL$_{EUC}$ and (E) their overlay over input images.

Table 1. Classification and explanation performance.

Models	Accuracy		Activation Precision	
	Validation	Test	Validation	Test
Unrefined	0.94	0.95	0.32	0.32
eXBL_{EUC}	0.89	0.90	0.34	0.35

5 Results

Table 1 shows classification and explanation performance of the Unrefined and eXBL_{EUC} models. Sample test images, GradCAM outputs, and overlaid GradCAM visualizations of x-ray images with Viral pneumonia category are displayed in Figure 3. From the sample GradCAM outputs and Table 1, we observe that the eXBL_{EUC} model was able to produce more accurate explanations that avoid focusing on irrelevant image regions such as the Epiphyses and background regions. This is demonstrated by how GradCAM explanations of the eXBL_{EUC} model tend to focus on the central image regions of the input images focusing around the chest that is relevant for the classification task, while the GradCAM explanations generated using the Unrefined model give too much attention to areas around the shoulder joint (humerus head) and appear angular shaped giving attention to areas that are not related with the disease categories.

6 Conclusion

In this work, we have presented an approach to debug a spurious correlation learned by a model and to remove it with just two exemplary explanations in eXBL_{EUC}. We present a way to adopt the triplet loss for unlearning spurious correlations. Our approach can tune a model's attention to focus on relevant image regions, thereby improving the saliency-based model explanations. We believe it could be easily adopted to other medical or non-medical datasets because it only needs two non-demanding exemplary explanations as user feedback.

Even though the eXBL_{EUC} model achieved improved explanation performances when compared to the Unrefined model, we observed that there is a classification performance loss when retraining the Unrefined model with eXBL to produce good explanations. This could mean that the initial model was exploiting the confounding regions for better classification performance. It could also mean that our selection of good and bad explanations may not have been optimal and that the two exemplary explanations may be degrading model performance.

Since our main aim in this study was to demonstrate effectiveness of eXBL_{EUC} based on just two ranked feedback, the generated explanations were evaluated using masks of lung because it is the only body part with pixel-level annotation in the employed dataset. However, in addition to the lung, the disease categories might be associated with other areas of the body such as the throat and torso. For this reason, and to ensure transparency in practical deployment of such systems in clinical practice, future work should involve expert end users for evaluation of the classification and model explanations.

Acknowledgements

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6183. For the purpose of Open Access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

1. Adebayo, J., Gilmer, J., Muelly, M., Goodfellow, I., Hardt, M., Kim, B.: Sanity checks for saliency maps. arXiv preprint arXiv:1810.03292 (2018)
2. Barnett, A.J., Schwartz, F.R., Tao, C., Chen, C., Ren, Y., Lo, J.Y., Rudin, C.: A case-based interpretable deep learning model for classification of mass lesions in digital mammography. Nature Machine Intelligence **3**(12), 1061–1070 (2021)
3. Chechik, G., Sharma, V., Shalit, U., Bengio, S.: Large scale online learning of image similarity through ranking. Journal of Machine Learning Research **11**(3) (2010)
4. Chowdhury, M.E., Rahman, T., Khandakar, A., Mazhar, R., Kadir, M.A., Mahbub, Z.B., Islam, K.R., Khan, M.S., Iqbal, A., Al Emadi, N., et al.: Can ai help in screening viral and covid-19 pneumonia? IEEE Access **8**, 132665–132676 (2020)
5. Dietvorst, B.J., Simmons, J.P., Massey, C.: Overcoming algorithm aversion: People will use imperfect algorithms if they can (even slightly) modify them. Management Science **64**(3), 1155–1170 (2018)
6. Hagos, M.T., Curran, K.M., Mac Namee, B.: Identifying spurious correlations and correcting them with an explanation-based learning. arXiv preprint arXiv:2211.08285 (2022)
7. Hagos, M.T., Curran, K.M., Mac Namee, B.: Impact of feedback type on explanatory interactive learning. In: International Symposium on Methodologies for Intelligent Systems. pp. 127–137. Springer (2022)
8. Islam, M.M., Karray, F., Alhajj, R., Zeng, J.: A review on deep learning techniques for the diagnosis of novel coronavirus (covid-19). Ieee Access **9**, 30551–30572 (2021)
9. Kermany, D.S., Goldbaum, M., Cai, W., Valentim, C.C., Liang, H., Baxter, S.L., McKeown, A., Yang, G., Wu, X., Yan, F., et al.: Identifying medical diagnoses and treatable diseases by image-based deep learning. Cell **172**(5), 1122–1131 (2018)
10. Marmolejo-Saucedo, J.A., Kose, U.: Numerical grad-cam based explainable convolutional neural network for brain tumor diagnosis. Mobile Networks and Applications pp. 1–10 (2022)
11. O'Neill, J., Delany, S.J., Mac Namee, B.: Rating by ranking: An improved scale for judgement-based labels. In: IntRS@ RecSys. pp. 24–29 (2017)
12. Pfeuffer, N., Baum, L., Stammer, W., Abdel-Karim, B.M., Schramowski, P., Bucher, A.M., Hügel, C., Rohde, G., Kersting, K., Hinz, O.: Explanatory interactive machine learning. Business & Information Systems Engineering pp. 1–25 (2023)
13. Rahman, T., Khandakar, A., Qiblawey, Y., Tahir, A., Kiranyaz, S., Kashem, S.B.A., Islam, M.T., Al Maadeed, S., Zughaier, S.M., Khan, M.S., et al.: Exploring the effect of image enhancement techniques on covid-19 detection using chest x-ray images. Computers in Biology and Medicine **132**, 104319 (2021)
14. Ross, A.S., Hughes, M.C., Doshi-Velez, F.: Right for the right reasons: Training differentiable models by constraining their explanations. arXiv preprint arXiv:1703.03717 (2017)

15. Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., Chen, L.C.: Mobilenetv2: Inverted residuals and linear bottlenecks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4510–4520 (2018)
16. Santa Cruz, B.G., Bossa, M.N., Sölter, J., Husch, A.D.: Public covid-19 x-ray datasets and their impact on model bias–a systematic review of a significant problem. Medical Image Analysis **74**, 102225 (2021)
17. Schramowski, P., Stammer, W., Teso, S., Brugger, A., Herbert, F., Shao, X., Luigs, H.G., Mahlein, A.K., Kersting, K.: Making deep neural networks right for the right scientific reasons by interacting with their explanations. Nature Machine Intelligence **2**(8), 476–486 (2020)
18. Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D.: Grad-cam: Visual explanations from deep networks via gradient-based localization. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 618–626 (2017)
19. Shao, X., Skryagin, A., Stammer, W., Schramowski, P., Kersting, K.: Right for better reasons: Training differentiable models by constraining their influence functions. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 9533–9540 (2021)
20. Yousefzadeh, M., Esfahanian, P., Movahed, S.M.S., Gorgin, S., Rahmati, D., Abedini, A., Nadji, S.A., Haseli, S., Bakhshayesh Karam, M., Kiani, A., et al.: ai-corona: Radiologist-assistant deep learning framework for covid-19 diagnosis in chest ct scans. PloS One **16**(5), e0250952 (2021)
21. Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J., Oermann, E.K.: Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Medicine **15**(11), e1002683 (2018)
22. Zlateski, A., Jaroensri, R., Sharma, P., Durand, F.: On the importance of label quality for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 1479–1487 (2018)