Interpretability-Guided Human Feedback During Neural Network Training*

Pedro Serrano e Silva[1,2,3] [0009−0007−9032−4875],
Ricardo Cruz[2,3] [0000−0002−5189−6228], ASM Shihavuddin[4] [0000−0002−4137−9374],
and Tiago Gonçalves[2,3] [0000−0003−4744−9174]

[1] NILG.AI, Portugal
pedro.serrano@nilg.ai
[2] Faculty of Engineering, University of Porto, Portugal
{rpcruz,tiagofs}@fe.up.pt
[3] INESC TEC, Portugal
[4] Green University, Bangladesh
shihav@eee.green.edu.bd

Abstract. When models make wrong predictions, a typical solution is to acquire more data related to the error: an expensive process known as active learning. Our supervised classification approach combines active learning with interpretability so the user can correct such mistakes during the model's training. At the end of each epoch, our training pipeline shows examples of mistaken cases to the user, using interpretability to allow the user to visualise which regions of the images are receiving the model's attention. The user can then guide the training through a regularisation term in the loss function. This approach differs from previous works where the user's role was to annotate unlabelled data since, in this proposal, the user directly influences the training procedure through the loss function. Overall, in low-data regimens, the proposed method returned lower loss values in the predictions made for all three datasets used: 0.61, 0.47, 0.36, when compared with fully automated training methods using the same amount of data: 0.63, 0.52, 0.41, respectively. We also observed higher accuracy values in two datasets: 81.14% and 92.58% over the 78.41% and 92.52% seen in fully automated methods.

Keywords: artificial neural networks · active learning · explainable artificial intelligence · human feedback · interpretability

1 Introduction

The performance of deep neural networks has challenged human performance in many cases [14]; yet, they can fail spectacularly at surprisingly simple cases [25]. When a neural network makes a wrong prediction, a popular solution is to retrain it with more related cases. However, this approach may present high optimisation

* Supported by the Portuguese Foundation for Science and Technology - FCT within PhD grant and 2020.06434.BD.

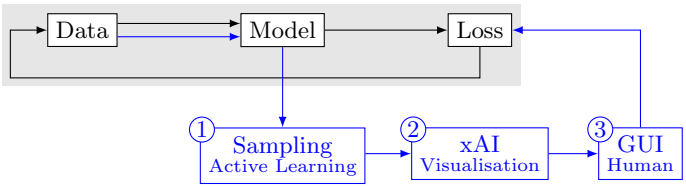


Fig. 1: Overview of the proposed pipeline: **black** lines represent the traditional model and **blue** lines represent the proposal.

costs, while obtaining new data may not be trivial nor guarantee better results. Therefore, to overcome such problems, an alternative solution may be to let the model efficiently choose the data samples that may positively benefit its training process. In machine learning (ML) literature, this strategy is known as *active learning* [18]. The intuition behind this approach assumes that the model will react better to a given problem if we provide more examples of that instance during training. Recently, the advantages of active learning have been studied and discussed in the literature of deep learning (DL), where the main goal is to retain the powerful learning capabilities of DL while reducing the cost of sample annotation [16]. Moreover, instead of relying solely on a curated set of data samples provided by an active learning approach, one may also integrate human experts in the training loop and benefit from their domain knowledge, in line with other Human-in-the-Loop (HITL) work [28, 6]. These frameworks focus on having the human improve the model [4] (e.g., new human annotations [10], reinforcement learning policies that mimic human behaviour [15]).

In this work, we propose more direct human participation by having the human guide the training process itself. The field of explainable artificial intelligence (xAI) (or ML interpretability) generally splits into three main research lines: *pre-*, *in-*, and *post-model* strategies [7]. Pre-model strategies aim to understand the data before making any ML model (e.g., exploratory data analysis). In-model strategies focus on inherently interpretable algorithms through rules or constraints [20]. Post-model strategies aim to produce explanations after the model's training (e.g., saliency maps on the image showing the locations that contributed most to the model's output [19, 24, 24]). Interestingly, integrating human users into the optimisation processes of interpretability is already part of the literature on xAI [9], thus motivating our work and showing that this is a timely opportunity to contribute towards this research line.

Our proposal, illustrated in Fig. 1, uses post-model explanations to show the user which parts of the image seem to contribute more to the prediction. The user interface shows rectangles around the most salient regions, and the user clicks on the image regions that should not be relevant to the model's prediction. We integrate this user input into the loss function as a regularisation term. Our proposal uses sampling strategies inspired by active learning methods to decide which images to show the user. This framework aims to promote transparency

by allowing users to visualise how their feedback impacts the model's decision iteratively.

Besides this Introduction, the remainder of this paper is organised as follows: Section 2 discusses other approaches of HITL that relate to our proposal; Section 3 extensively describes our proposal; Section 4 presents the experimental work; Section 5 shows the obtained results; Section 6 provides a detailed discussion of the outcomes of our proposal; and Section 7 concludes the paper and suggests future work directions.

2 Related Work

Liu et al. [11] proposed a reinforcement learning method based on HITL, which avoids pre-labelling steps and keeps the model upgrading with progressively collected data. Their strategy uses a *deep reinforcement active learning* method to guide an agent in selecting training samples by an oracle. In this case, the uncertainty value of each human who picked a training sample works as a reinforcement reward. Uehara et al. [27] developed a novel process for object detection in satellite images based that uses active learning combined with Gradient-weighted Class Activation Mapping (Grad-CAM) [17] to select the best images from an unlabelled pool of samples to train a feature extractor and classifier efficiently. Using Grad-CAM, this application queries users about the features in a given image, who must select the regions they consider to be related to a given object. Le et al. [10] proposed a simple and efficient *interactive self-annotation* framework to cut down time and human labour costs for video object bounding box annotation. The interactive recurrent annotation relies on a human annotator that gives feedback to the bounding-box detector. Later, Adhikari and Huttunen [1] improved on this framework by modifying the role of humans, which transitioned from performing full annotations to correcting errors.

On autonomous driving, Rajendran et al. [15] use HITL learning for vehicle self-exploration with HITL learning, in two phases. In the first phase (non-exploratory), the AI system learns from available historical data and the imitation of an oracle. In the second phase (exploratory), the AI system interacts with the environment, updating its policy through a long short-term memory network (LSTM), restricted by the anomaly detector trained in the first phase with the aid of a human annotator. Smailagic et al. [23] innovated the task of medical image classification by introducing MedAL, a novel way of selecting relevant samples from an unlabelled image pool. This method maximises the average distance to all training set examples in a learned feature space. The main goal is to reduce the amount of data required for state-of-the-art results by generating an optimal initial training set (i.e., with the most information about the overall data distribution) without needing a trained model. Regarding interpretability-guided frameworks, we highlight the works developed by Silva et al. [22, 21], where they proposed the use of interpretability methods to localise relevant regions of images, promoting more focused feature representations and, consequently, improve medical image retrieval; Mahapatra et al. [13], where they

proposed an interpretability-guided method that enforces that learned features yield more distinctive and spatially consistent saliency maps for different class labels of trained models; and, lastly, Mahapatra et al. [12], where they presented a sample selection approach based on graph analysis to identify informative samples in a multi-label setting. While this literature served as inspiration, in this work, we propose a more direct approach at HITL by having the user guide the model directly through its loss function.

3 Proposal

In our supervised classification proposal, we approach the neural network and the human as two entities that speak different languages. This way, xAI methods establish the bridge between these two entities (i.e., a translator). From active learning and HITL, we can learn the most effective way of mediating this communication and make it more efficient by finding the least amount of information required to convey any ideas in the conversation. The pipeline is summarised in Algorithm 1 and includes the following steps that are detailed in the following subsections:

1. Active learning sampling to find the most egregious model mistakes (line 3).
2. Use xAI to detect which regions of the images are being considered by the model (line 4).
3. The user feedback is integrated into the loss function (lines 5–6).

Algorithm 1 Pseudocode of our training method.

1: **function** TRAIN(model, images, labels)
2: preds $\leftarrow$ model(images)
3: images $\leftarrow$ EntropySample(images)
4: saliencies $\leftarrow$ xAI(images, preds)
5: $W_{i,j} \leftarrow$ UserInterface(saliencies)
6: loss $\leftarrow$ CE(preds, labels) $+ \sum_{i,j} \lambda W_{i,j} \frac{\partial \text{preds}}{\partial \text{images}_{i,j}}$
7: **end function**

3.1 Sampling

Data sampling and querying play a crucial role in the proposed workflow. In active learning, one may use several traditional *uncertainty sampling* methods to identify the most relevant data points in a dataset. The logic behind these sampling methods is that the data points which are hard to classify must contain relevant information in an active learning workflow. Although several uncertainty sampling exists (e.g., *least confident strategy*, *margin sampling*, and *entropy sampling*), Settles [18] concluded that entropy sampling is more suited

if the objective is to decrease logarithmic loss. Since one of our goals is to minimise model overfitting and considering that the solution should generalise for tasks with any number of classes, we selected entropy sampling as the best option (x^*_H), as described in Eq. 1:

$$x^*_H = \arg\max_x -\sum_i P_\theta(y_i \mid x) \log P_\theta(y_i \mid x), \tag{1}$$

where y_i is the label for observation i, x is the image, and P_θ is the probability of y_i given x, predicted by the model. Cases with high entropy are relevant since those are the cases where the model has higher uncertainty of its prediction. We then select the top-X examples for user feedback.

3.2 Interpretability

In this work, we use DeepLIFT [19] to generate the saliency maps the human user will assess. This xAI method decomposes the output prediction of a neural network on a specific input by performing backpropagation of the contributions [3] of all neurons in the network to every feature of the input signal (i.e., the image). It defines *importance* as a function of differences from a reference state, where the reference is chosen based on the problem at hand. In this work, we used the variation DeepLIFT-Rescale, included in the Captum library for Python programming language [8]. DeepLIFT has two optional, although important, parameters to be defined: the target and the reference. The target is related to which class (i.e., label) DeepLIFT will compute the pixel attributions. As the HITL process should happen in a clean, single-image display, we must execute DeepLIFT for a single target class. Our proposal aims to fix the model's prediction. Hence, it is logical to target the label predicted by the model since we are basing human interaction on this prediction. The reference refers to the reference image previously explained. The default reference is a white image of the same size as the input image. However, Shrikumar et al. [19] recommend a different strategy for more complex datasets such as the ones used in our experiments: using a blurred version of the input image as the reference to obtain more well-defined explanations.

3.3 User Feedback

After selecting the most relevant samples, it is necessary to show information to the annotator (i.e., the user) in a human-understandable manner. Simply showing the DeepLIFT saliency map to the annotator may be too cryptic, and overlaying them on top of the original can damage the perception of the image. To overcome these visualisation issues, rectangles over the regions with higher attribution scores are identified by dividing the saliency map in a grid (see Fig. 2). Besides, to allow for a more informed decision, our interface also shows the ground-truth label of the image, the model's prediction, and the image index.

The human feedback requires the following steps:

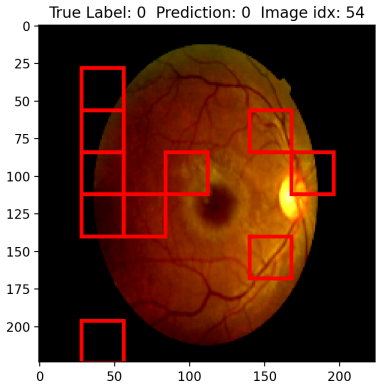


Fig. 2: Example of a query when training the model on the APTOS2019 dataset. The interface displays the sampled image, its ground-truth label, the model's prediction, and the image index. The users can click on the squared regions they perceive as less relevant for the classification task. After finishing their selection, the users can submit it and close the image.

1. The annotator must select which of the presented regions the model should disregard by clicking the square.
2. When clicked, the rectangle is highlighted, and a second time can deselect it.
3. After selecting all the desired squares, the user can close the image. We store this information in the form of a weight tensor $W_{i,j} \in \{0, 1\}$ related to the locations of the selected pixels i, j. This tensor is 1 for the pixels chosen by the user and 0 otherwise.

To backpropagate the errors identified by the user, we penalise the selected pixels (that is, where $W_{i,j}$ is 1). A regularisation term is added to the loss function that penalises the model if the pixels are being used to produce the model output. That is, the derivative of the output relative to the input ($\frac{\partial \hat{y}}{\partial x_{i,j}}$) should be zero for those cases. The final loss is described in Eq. 2:

$$L(\hat{y}, y) = \text{CE}(\hat{y}, y) + \lambda \sum_{i,j} W_{i,j} \left(\frac{\partial \hat{y}}{\partial x_{i,j}} \right)^2, \tag{2}$$

where λ is a hyper-parameter weighting the new loss term.

4 Experiments

4.1 Data

The three datasets used by the experiments are described below with a summary in Table 1. **ISIC2017** is a skin cancer dataset from the ISIC2017 Challenge [5][5],

[5] https://challenge.isic-archive.com/landing/2017/

Table 1: Datasets used in the experiments.

Dataset	N	Classes used
ISIC2017	2,000	1 vs 2
APTOS2019	3,600	multiclass (0-4)
NCI	74,820	0 vs 1

that had participants attempt to build automated solutions for the diagnosis of melanoma from dermoscopic images. We have used Task 3: Disease Classification. **APTOS2019** is a diabetic retinopathy dataset[6] originates from a competition held on Kaggle, where participants were tasked with automating the identifying of different stages of diabetic retinopathy, to detect indications of blindness. **NCI** is a cervical cancer dataset from the American National Cancer Institute (NCI). Besides the medical image data (i.e., cervigrams), the dataset includes the patient's age, HPV test, and histology results. The labels for each cervigram are related to the neoplasia progression level and divided into the following categories: normal, CIN1, CIN2, CIN3, and cancer. This dataset is not publicly available.

4.2 Model Architecture

After preliminary experiments, we decided to use EfficientNet [26] as the model's architecture. We initialised this model with the weights from the training on the ImageNet dataset. In this proposal, the weight tensor $W_{i,j}$ only stores information about the non-augmented version of the images. Hence, to include data augmentation methods during training, we must compute the backward pass twice: the first pass computes the ordinary loss for the augmented images, and the second pass computes our regularisation term using the original image.

4.3 Training Phase

Data Augmentation: We used the following data transformations: vertical and horizontal flipping, 10% cropping, free 360^{o} rotation, and 10% colour brightness. We resized all images to 224×224.

Baseline: We start by training a baseline model for 10 epochs without human feedback (i.e., we optimised this hyper-parameter through preliminary studies). From the progression of the entropy values, we noticed that most data points shared the same values, meaning high entropy sampling would work as random sampling. By running more epochs, we are letting the model *figure out* which data points are harder to classify, and at this point, we can correctly apply high entropy sampling. The learning rate for the baseline was 1.0×10^{-4}, the optimiser was adaptive moment estimation (Adam), and the batch size was 4.

[6] https://www.kaggle.com/c/aptos2019-blindness-detection

HITL Models: The HITL models were initialised with the baseline's weights and trained for 20 epochs (i.e., 10 with human input and 10 without). The number of queries was 20, which amounts to roughly 7% of the training data, and the λ value was 10^6. To compare the HITL training with automated training, we trained another model without human feedback for 20 epochs. Both models used a learning rate of 10^{-4}, Adam as the optimiser, and a batch size of 4.

The code related to the experimental work is publicly available in a GitHub repository[7].

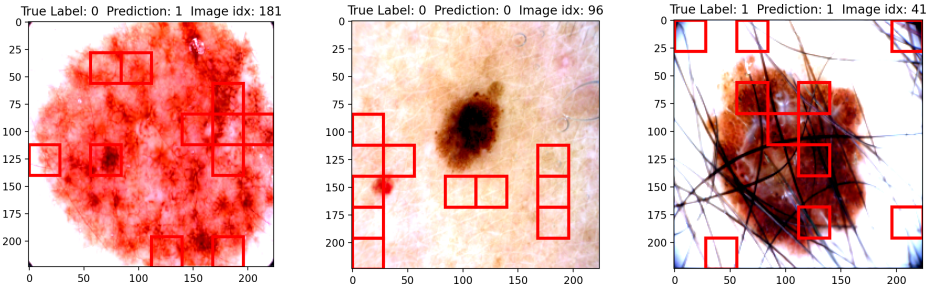
5 Results

Table 2: Results of the proposed framework.

Dataset	Method	Accuracy	Min Loss
ISIC2017	Baseline 10%	**85.91**	0.52
	HITL 10%	83.87	**0.47**
	Baseline 100%	88.59	0.34
APTOS2019	Baseline 10%	78.41	0.63
	HITL 10%	**81.14**	**0.61**
	Baseline 100%	85.11	0.47
NCI	Baseline 0.5%	92.52	0.41
	HITL 0.5%	**92.58**	**0.36**
	Baseline 50%	**93.18**	0.23
	HITL 50%	93.16	**0.22**

The performance of the proposed pipeline relates to the prediction accuracy and the minimum loss value achieved by the model. The test scenarios use only a percentage of the dataset so that the impact of the method is made more visible. Models trained with the proposed human feedback approach (i.e., HITL) are contrasted with models trained without human feedback (i.e., baseline). Table 2 shows the raw performance of the methods for the most relevant experiments performed in this work. Regarding the ISIC2017 and APTOS2019 datasets, we achieved the best performance when training on 10% of the data. We also present the baseline results when using 100% of the data. Concerning the NCI dataset, we achieved the best performance when training on 0.5% of the data. To confirm the low impact of the HITL method in scenarios with high data availability, we performed other experiments as well. These tests were performed on 50% of the data, using models pre-trained on 0.5%, with the HITL method, and compared with no feedback training on 50% of the dataset. We achieved the lowest min-loss using the proposed HITL training method in all tests. Regarding accuracy values, the HITL-trained model outperformed the baseline, in the same task, on two different occasions. Moreover, models trained with HITL and low data

[7] https://github.com/PedroSerran0/ig-human-feedback-nn

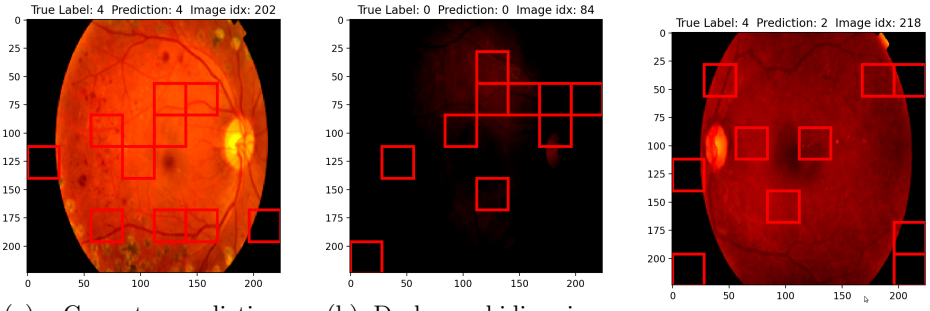


(a) Example of an image hard to interpret.

(b) DeepLIFT attributions poorly summarized.

(c) Presence of hair in the image.

Fig. 3: Example of the problems encountered when annotating query images while training the HITL approach on the ISIC2017 dataset.



(a) Correct prediction with unclear attributions.

(b) Darkness hiding important features.

(c) Possibly wrong label.

Fig. 4: Examples of the problems encountered when annotating query images while training the HITL approach on the APTOS2019 dataset.

availability could not achieve similar accuracy values to models trained with 100% or 50% of the data. Testing on 100% and 50% of the datasets showed that using feedback or not achieves similar results, albeit when using feedback, there is more stochasticity in the training progression. These results suggest a delay in the model overfitting, resulting, by hypothesis, in lower min-loss values. Interestingly, we detected a similar pattern in most low-data training scenarios.

6 Discussion

Fig. 3 presents examples of some of the difficulties found when working with the ISIC2017 dataset. From these, we realise that there are cases where the model focuses on irrelevant features (e.g., higher attribution values outside the area of the lesion). Therefore, we argue that these are the situations where the training of the models may benefit from the HITL framework. One of the drawbacks of

working with this dataset is the difficulty in understanding the type of features the model should ignore or pay attention to (e.g., the presence of hair or the existence of multiple lesions). Naturally, this process would benefit from the aid of a domain expert. Oppositely to other use cases, the APTOS2019 dataset seems intuitive to manipulate, regarding the features that the model should ignore (i.e., the features that are always present, independently of the stage of the disease, such as background, fovea, the optic disc, and the blood vessels) or pay attention (i.e., features that should relate to the stages of the disease) when classifying an image. Hence, the HITL framework is essential in guiding the model's training to focus on the relevant objects.

Nevertheless, as shown in Fig. 4, we also identified some difficulties when working with the APTOS2019 dataset. We encountered several queries where the image was too dark (making it very hard to analyse while returning a high entropy prediction), ambiguity on the features (i.e., the features where the model should focus were not easy) or on the labels (i.e., model's prediction made more sense than the ground-truth annotation), and where the attributions did not motivate the model's prediction.

Regarding the NCI dataset, we followed the annotation rationale described by Albuquerque et al. [2], which uses a Gaussian kernel for regularisation, forcing the model to focus on a more central part of the images, as most tools are near the peripheries. Hence, assuming this prior knowledge, the HITL framework serves as a flexible refined method of this state-of-the-art approach. Besides, as with APTOS2019, we noted that this dataset was intuitive. However, we also registered some abnormal behaviour during training (e.g., the model focusing on irrelevant features such as the instruments or the textual description of the cervigram). Interestingly, after a few training epochs, when the model started to focus on relevant features (e.g., the endocervix), we could still register some cases of abnormal behaviour. We do not provide a figure with the results obtained, as this dataset is not publicly available.

7 Conclusion

In this work, we proposed a HITL framework that helps comprehend the images that models find harder to classify and where they seem to fail the most. It also succeeded in further exploring the requirements for effective interaction between neural networks and people, and its limitations, with several cases highlighted during testing. All of this information can play a big part in finding better ways to improve the training process and fine-tune models while increasing the transparency of neural networks. Experiments performed in three datasets showed some loss reduction – 0.61, 0.47, and 0.36 for the proposed pipeline versus 0.63, 0.52, and 0.41 for the baseline, respectively, for each dataset.

Further work should be devoted to fine-tuning hyper-parameters such as the number of training epochs in which we ask humans for feedback, the space between those same epochs, the number of queries per epoch, the type of sampling, the threshold for entropy sampling, the number of squares to display, the shape

of those squares, the optimal learning rate or the optimal λ for the regularisation term in the proposed loss function. Concerning user interface improvements, we aim to explore other geometric shapes besides the current squared grid. Moreover, we intend to perform additional experiments on clustering methods, connecting high-value pixels in the prediction explanations or showing the personalised feature shape to the user. Furthermore, it would be interesting to extrapolate this methodology into other tasks (e.g., segmentation or object detection since there are multiple outputs).

References

1. Adhikari, B., Huttunen, H.: Iterative bounding box annotation for object detection. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 4040–4046. IEEE (2021)
2. Albuquerque, T., Cardoso, J.S.: Embedded regularization for classification of colposcopic images. In: 2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI). pp. 1920–1923. IEEE (2021)
3. Ancona, M., Ceolini, E., Öztireli, C., Gross, M.: Towards better understanding of gradient-based attribution methods for deep neural networks. arXiv preprint arXiv:1711.06104 (2017)
4. Budd, S., Robinson, E.C., Kainz, B.: A survey on active learning and human-in-the-loop deep learning for medical image analysis. Medical Image Analysis **71**, 102062 (2021)
5. Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., et al.: Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (ISBI), hosted by the international skin imaging collaboration (ISIC). In: 2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018). pp. 168–172. IEEE (2018)
6. Fischer, M., Kobs, K., Hotho, A.: NICER: aesthetic image enhancement with humans in the loop. arXiv preprint arXiv:2012.01778 (2020)
7. Kim, B., Doshi-Velez, F.: Interpretable machine learning: the fuss, the concrete and the questions. ICML Tutorial on interpretable machine learning (2017)
8. Kokhlikyan, N., Miglani, V., Martin, M., Wang, E., Alsallakh, B., Reynolds, J., Melnikov, A., Kliushkina, N., Araya, C., Yan, S., et al.: Captum: A unified and generic model interpretability library for pytorch. arXiv preprint arXiv:2009.07896 (2020)
9. Lage, I., Ross, A., Gershman, S.J., Kim, B., Doshi-Velez, F.: Human-in-the-loop interpretability prior. Advances in neural information processing systems **31** (2018)
10. Le, T.N., Sugimoto, A., Ono, S., Kawasaki, H.: Toward interactive self-annotation for video object bounding box: Recurrent self-learning and hierarchical annotation based framework. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3231–3240 (2020)
11. Liu, Z., Wang, J., Gong, S., Lu, H., Tao, D.: Deep reinforcement active learning for human-in-the-loop person re-identification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 6122–6131 (2019)
12. Mahapatra, D., Poellinger, A., Reyes, M.: Graph node based interpretability guided sample selection for active learning. IEEE transactions on medical imaging (2022)

13. Mahapatra, D., Poellinger, A., Reyes, M.: Interpretability-guided inductive bias for deep learning based medical image. Medical image analysis **81**, 102551 (2022)
14. McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G.S., Darzi, A., et al.: International evaluation of an AI system for breast cancer screening. Nature **577**(7788), 89–94 (2020)
15. Rajendran, P.T., Espinoza, H., Delaborde, A., Mraidha, C.: Human-in-the-loop learning for safe exploration through anomaly prediction and intervention. Proceedings of SafeAI, AAAI (2022)
16. Ren, P., Xiao, Y., Chang, X., Huang, P.Y., Li, Z., Gupta, B.B., Chen, X., Wang, X.: A survey of deep active learning. ACM Comput. Surv. **54**(9) (oct 2021)
17. 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 (2017)
18. Settles, B.: Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison (2009)
19. Shrikumar, A., Greenside, P., Kundaje, A.: Learning important features through propagating activation differences. In: International conference on machine learning. pp. 3145–3153. PMLR (2017)
20. 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, pp. 133–140. Springer (2018)
21. Silva, W., Gonçalves, T., Härmä, K., Schröder, E., Obmann, V.C., Barroso, M.C., Poellinger, A., Reyes, M., Cardoso, J.S.: Computer-aided diagnosis through medical image retrieval in radiology. Scientific reports **12**(1), 20732 (2022)
22. 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. pp. 305–314. Springer (2020)
23. Smailagic, A., Costa, P., Noh, H.Y., Walawalkar, D., Khandelwal, K., et al.: MedAL: Accurate and robust deep active learning for medical image analysis. In: 2018 17th IEEE international conference on machine learning and applications (ICMLA). IEEE (2018)
24. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks. In: International Conference on Machine Learning. pp. 3319–3328. PMLR (2017)
25. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint (2013)
26. Tan, M., Le, Q.: EfficientNet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. PMLR (2019)
27. Uehara, K., Nosato, H., Murakawa, M., Sakanashi, H.: Object detection in satellite images based on active learning utilizing visual explanation. In: 2019 11th International Symposium on Image and Signal Processing and Analysis (ISPA). pp. 27–31. IEEE (2019)
28. Zhang, L., Wang, X., Fan, Q., Ji, Y., Liu, C.: Generating manga from illustrations via mimicking manga creation workflow. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5642–5651 (2021)