IMIL: Interactive Medical Image Learning Framework

Adrit Rao[1,2] Andrea Fisher[1] Ken Chang[1] John Christopher Panagides[1] Katherine McNamara[1]
{adritrao, atfisher, changk1, jpanagid, kpogrebn}@stanford.edu

Joon-Young Lee[3] Oliver Aalami[1]
jolee@adobe.com aalami@stanford.edu
[1]Stanford University [2]Palo Alto High School [3]Adobe Research

Abstract

Data augmentations are widely used in training medical image deep learning models to increase the diversity and size of sparse datasets. However, commonly used augmentation techniques can result in loss of clinically relevant information from medical images, leading to incorrect predictions at inference time. We propose the Interactive Medical Image Learning (IMIL) framework, a novel approach for improving the training of medical image analysis algorithms that enables clinician-guided intermediate training data augmentations on misprediction outliers, focusing the algorithm on relevant visual information. To prevent the model from using irrelevant features during training, IMIL will 'blackout' clinician-designated irrelevant regions and replace the original images with the augmented samples. This ensures that for originally mispredicted samples, the algorithm subsequently attends only to relevant regions and correctly correlates them with the respective diagnosis. We validate the efficacy of IMIL using radiology residents and compare its performance to state-of-the-art data augmentations. A 4.2% improvement in accuracy over ResNet-50 was observed when using IMIL on only 4% of the training set. Our study demonstrates the utility of clinician-guided interactive training to achieve meaningful data augmentations for medical image analysis algorithms .

1. Introduction

The applications of computer vision to medical image analysis have been numerous in recent years [5, 11]. This can be attributed to major advancements in deep learning and increased availability of large, open-access medical imaging datasets [21]. These algorithms have the potential to significantly improve the efficiency and accuracy of disease diagnosis in various medical imaging modalities [23, 28, 30, 31].

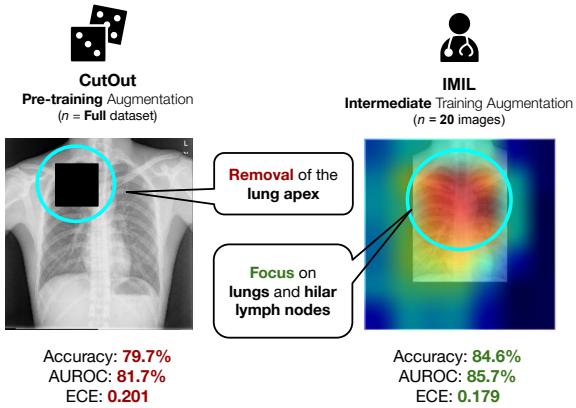

Figure 1. **IMIL versus CutOut [9].** In this study, we propose an interactive medical image learning framework (IMIL), a callback framework that solicits clinician input to generate human-guided, intermediate training data augmentations. Compared to *random* pre-training augmentations, IMIL prevents the removal of clinically relevant visual features. When IMIL is used on just **20** CXR images, it surpasses the performance and calibration of CutOut for tuberculosis classification (see Results 4 - IMIL + Res1).

As these models are translated into the clinical setting, it is important to consider the interaction between clinicians and algorithms [6]. In addition to training algorithms for accuracy, practitioners must also ensure that a model identifies and concentrates on clinically relevant features (correct *focus*) and performs consistently and dependably in various settings (*reliability*). A focused algorithm should not only be able to make correct predictions but should make these predictions based on relevant regions of the medical image and at the same time, avoid spurious correlations from irrelevant parts of the image [39, 42]. Algorithms can have difficulty focusing on clinically relevant regions, as they often lack the ability to independently sort information for clinical relevance like a clinician would [8, 32]. Furthermore,

model reliability is achieved by providing confidence levels for predictions, which enable clinicians to understand the true certainty of a prediction. Standard convolutional neural networks (CNN) are often prone to overconfidence and miscalibration, creating the need for confidence calibration [17]. Although visual explainability and calibration are important, they are not reflected in accuracy or performance-based metrics and can often go undetected.

Image augmentations are used in medical image analysis to counteract a lack of diversity or scarcity of clinical data. Typically, these are performed before training in a randomized manner through image manipulations (e.g. crop, zoom, flip) [7]. Recent, modern augmentations include CutMix, MixUp, and CutOut [9, 41, 43]. These methods aim to improve the performance, robustness, and calibration of CNNs. Although their efficacy has been established for medical image analysis [33], their underlying effects pose significant challenges and may necessitate further optimization. When training algorithms on clinical data, it is most efficient to retain all relevant information. However, data augmentations developed for non-medical use may run the unintended risk of removing clinically relevant information from images. Doing so may not always be directly reflected in accuracy but can significantly impact focus and visual explainability. For example, CutOut [9] (which performs randomized pre-training dropouts in the visual space), may remove the lung apex in a chest x-ray which is critical to identify tuberculosis (as shown in Figure 1). The algorithm instead focuses on external and irrelevant visual features. With limited medical image datasets, it is important for the algorithm to correctly associate class labels with relevant visual indicators. Even small shifts in this relation can cause downstream effects on attention and domain shifting.

To improve model focus and reliability, we explore the usage of clinician feedback during training, rather than before, to perform clinically meaningful augmentations on the most challenging cases for the algorithm. We aim to focus the algorithm on relevant visual information, increase the safety of augmentations, and minimize clinician annotation burden. The Interactive Medical Image Learning (IMIL) framework is a flexible callback that enables clinician-guided intermediate training data augmentations. During a set frequency in training, IMIL selects a predefined number of outliers from the training dataset based on the algorithm's worst mispredictions. These images are provided to a clinician along with the associated class activation map (CAM) [35], prediction with confidence, and ground-truth label. The clinician then re-directs the *attention* of the algorithm after understanding why it is making a misprediction. To provide feedback, akin to the Google reCAPTCHA method [16] where users identify relevant segments within a grid on an image, the clinician selects the region of the image that the model *should be* focusing on, using a similar grid overlay approach. Using this input, IMIL will perform a 'blackout' augmentation and remove all of the unselected grid regions. The newly augmented image only contains clinically relevant information that should be associated with the diagnostic label. The algorithm then relearns the correct visual features on the most challenging outlier cases (as shown in Figure 1). After feedback, the original training samples are replaced with the IMIL augmented images for subsequent training. Our study demonstrates that intermediate training augmentations based on clinician-guidance can significantly improve performance and calibration and can maximize the potential of smaller medical image datasets.

To validate the efficacy of IMIL, we perform a clinical user-study with three clinical radiology residents. The residents provide IMIL feedback on a tuberculosis (TB) chest x-ray (CXR) dataset, which is used to create three separate IMIL-augmented algorithms. We perform a comparative analysis of these algorithms to modern augmentations (MixUp [43], CutMix [41], and CutOut [9]) using performance and calibration-based metrics.

Our main **contributions** are summarized below:

1. We propose a novel deep learning callback framework, **IMIL**, which incorporates clinician feedback to perform intermediate training augmentations.
2. We conduct a user-study on clinical residents to understand the potential of **IMIL** deployment
3. We validate **IMIL** using deep learning algorithms trained during the user-study and compare performance against state-of-the-art modern augmentations: CutMix, MixUp, and CutOut.

2. Related Work

Medical Confidence Calibration: Various studies have shown that standard CNNs are highly prone to overconfidence which can lead to unreliable confidence estimates. Gou *et al.* [17] studied the calibration of widely used architectures, such as ResNet [19]. The study found that although increased network depth and width tends to improve accuracy, it can have negative effects on calibration. Furthermore, the use of batch normalization can also reduce calibration. Calibration level is not reflected in standard performance-based metrics which can lead to this problem going undetected. However, in the general computer vision domain, various techniques have been developed to combat miscalibration. Modern augmentations such as MixUp, CutMix, and CutOut can have significant effects on calibration. In the original studies, these augmentations showed major benefits for regularization of CNNs [9, 41, 43]. Various studies have also separately validated the calibration benefits of these algorithms showing that they can yield significant improvements. They have also been applied in

the medical domain for image classification and segmentation [10, 13]. Additionally, a study looked at benchmarking these augmentations for various medical image modalities and revealed that they can significantly improve the performance and calibration of CNNs [33].

Visual Explainability: Studies have shown that standard CNNs do not possess the ability to accurately rank the clinical relevance of visual features [32, 34]. In one study [8], the visual explainablity of a deep learning algorithm for COVID-19 detection in chest x-rays was assessed. The study found that even though performance-based metrics indicated that the algorithm was *accurate*, saliency maps revealed that predictions were being made on textual indicators rather than actual pathology. This is just one example of the potential risk associated with models that lack *focus*.

Human-in-the-loop Training: We have not identified a prior augmentation technique that incorporates clinician feedback during training. However, interactive training (human-in-the-loop) has been notably studied through active learning (AL) in the medical image domain [4]. AL often involves querying a clinical expert to label data points that will have the greatest effect on performance. This is a technique that is often leveraged when the cost of obtaining labeled data is high (time and expertise) [12, 37, 40].

3. Methods

We perform a preliminary validation of the IMIL callback framework for TB classification in CXR images [24]. We validate IMIL by performing a clinical user-study with residents to train three IMIL-augmented CNNs. We then compare the performance and confidence calibration of IMIL to state-of-the-art modern augmentations: CutMix [41], MixUp [43], and CutOut [9]. As follows is a description of the callback architecture (3.1), modern data augmentations (3.2), and study design (3.3).

3.1. IMIL Architecture

The overall architecture of the IMIL callback module is shown in Figure 2A and the usage of IMIL within a training pipeline for a CNN (as done in our study) is shown in Figure 2B. Figure 2B also shows the IMIL callback configuration used within our study. Each of the parameters are customizable within the framework and consist of `num_outliers`, `IMIL_epoch`, and `grid_size`. As follows are details regarding each step of the feedback loop.

Misprediction Outlier Selection Based on the predefined `num_outliers` parameter in the callback configuration, IMIL finds the 'most significant' mispredictions on which clinician feedback will be obtained. At the `IMIL_epoch` at which the callback is called, IMIL makes predictions across the full training set. IMIL then compiles a list of mispredictions by comparing ground-truth labels to model predictions. The mispredictions are then sorted in descending order based on confidence. The final outliers (n = `num_outliers`) are then chosen from the *highest confidence mispredictions*. This process helps direct human feedback towards the most challenging cases for the model. As such, directing human feedback to shift attention on hallucinations can potentially have more significant effects on downstream feature extraction compared to a random outlier selection strategy.

User Interpretation For each outlier, the clinician is provided with a visual dialogue during training to interpret the misprediction and then provide feedback. Before the clinician is prompted for feedback, they are provided with information to inform their decision-making: the original image, CAM heat-map [35], predicted label (with confidence), and ground-truth (as shown after misprediction selection in Figure 2A). The clinician should interpret all datapoints to understand the cause for the misprediction. For example, the heat-map can direct the clinicians attention to the region of the image where the model extracted features. The predicted label and associated confidence can help the clinician determine how these features are being mapped in the CNN. As the clinician becomes more familiar with this interpretation flow, they may start to notice biases in the model's focus (e.g. consistent attention towards the soft tissues instead of lung in CXR) and thus tailor their feedback to better correct the model.

Localization Grid Feedback Clinician feedback is obtained using a localization grid overlaid on the original image. The grid is made up of equally sized squares of modifiable dimensions (based on `grid_size`). The clinician selects relevant squares on the grid overlay that correspond to clinically relevant visual features (as shown in the sparse user input stage in Figure 2A). This step re-focuses the model and prevents it from falsely associating the ground-truth class label to irrelevant features. In the current prototype implementation of IMIL, this is done by assigning each grid region a number and mapping to a console input. However, the underlying callback function can be applied to a variety of different interactive feedback interfaces.

Callback Output Augmentation The region-based sparse input is used to generate the augmented output. IMIL removes unselected regions ('blackout') from the image (similar to *visual dropout* [9]), and retains the selected regions. The original training sample is then replaced with the augmented IMIL output at the specified `IMIL_epoch`. The replacement of the original image during training

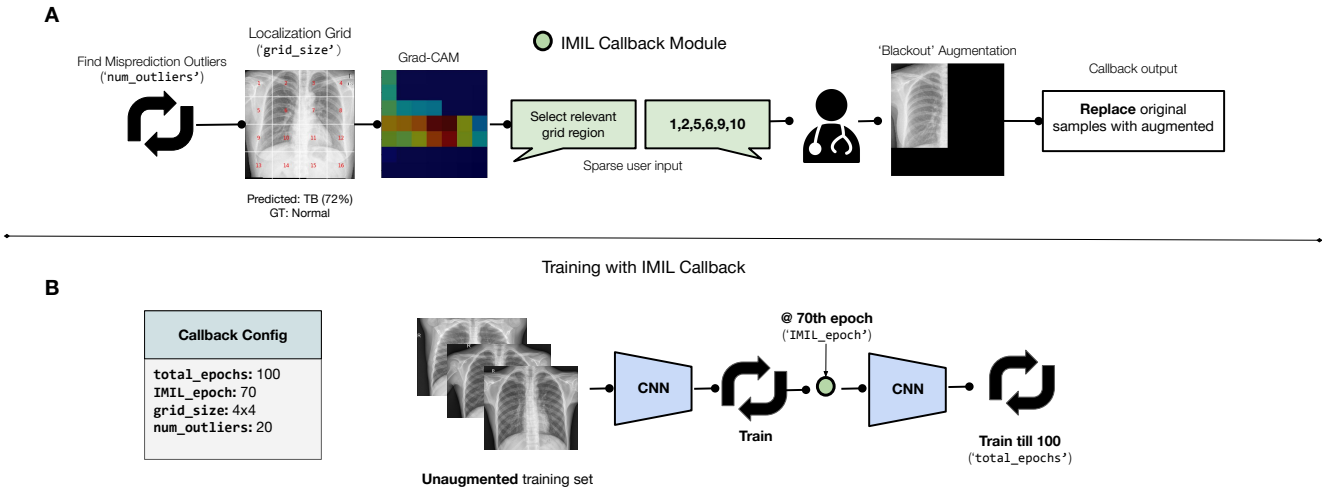

Figure 2. **Interactive Medical Image Learning (IMIL)** callback framework architecture (A) and implementation within a standard CNN training pipeline (along with callback configuration) (B). The IMIL framework allows the clinician to perform guided-augmentations ('blackout') *during* training to re-focus the algorithm on clinically relevant regions. The callback consists of various configuration parameters allowing for customized usage of IMIL during training.

allows the algorithm to re-associate the ground-truth label with relevant visual features for the remainder of training epochs. After replacment, training continues (from the 70th to 100th epoch; Figure 2B).

3.2. Modern Data Augmentations

We compare IMIL against MixUp [43], CutMix [41], and CutOut [9] which have been widely studied and applied for medical image analysis. Out of these augmentations, IMIL's 'blackout'-like augmentation is similar to CutOut. However, instead of random visual dropout it is clinician-guided and performed during training on a limited sample size. As follows are brief details surrounding the formulation for each augmentation.

MixUp Zhang *et al.* [43] proposed MixUp as a modern augmentation technique for training neural networks on a *blend* between a pair of images and labels based on convex combinations. MixUp has demonstrated benefits in terms of increasing robustness of neural networks when learning from corrupt labels and adversarial examples. The original formulation of MixUp from the original paper [43] is:

$$\begin{aligned} \tilde{x} &= \lambda x_i + (1-\lambda)x_j \\ \tilde{y} &= \lambda y_i + (1-\lambda)y_j, \end{aligned} \tag{1}$$

where $\boldsymbol{x}_i, \boldsymbol{y}_i$ are raw randomly sampled input vectors and $\boldsymbol{x}_j, \boldsymbol{y}_j$ are the corresponding one-hot label encodings. λ are values in the range [0, 1] which are randomly sampled from the Beta distribution for each augmented example.

CutMix Yun *et al.* [41] introduced CutMix, an augmentation built upon the original formulation of MixUp and the idea of combining samples. CutMix removes a patch from an image and swaps it for a region of another image generating a locally natural unseen sample. The formulation for CutMix is as follows:

$$\begin{aligned} \tilde{x} &= Mx_i + (1-M)x_j \\ \tilde{y} &= \mu y_i + (1-\mu)y_j, \end{aligned} \tag{2}$$

where M indicates the binary mask used to perform the cutout and fill-in operation from two randomly drawn images. μ are values (in [0,1]) randomly drawn from the Beta distribution.

CutOut This technique was proposed by DeVries *et al.* [9] and is a simple augmentation technique for improving the regularization of CNNs. CutOut was formulated based on the idea of extending dropout [20] to a spatial prior in the input space. CutOut performs occlusions of an input image similar to the idea proposed in [3]. Rather than partially occluding portions of an image [3], CutOut performs fixed-size zero-masking to fully obstruct a random location of an image. CutOut differentiates from dropout as it is an augmentation technique and visual features are dropped at the input stage of the CNN whereas in dropout, this occurs in intermediate layers. The goal of CutOut is to not only improve regularization of CNNs but improve robustness to occluded samples in real-world applications.

3.3. Study Design

As follows is a description of each component of the user-study for feedback-guided augmentations. The dataset used for this study is described in 3.3.1, the configuration of IMIL in 3.3.2, the CNN architectures used in 3.3.3, training details in 3.3.4, clinical user-study structure in 3.3.5, and lastly the evaluation metrics used to measure performance and calibration of the models in 3.3.6.

3.3.1 Tuberculosis Dataset

To validate the efficacy of IMIL, we focused on TB diagnosis in CXR. TB is a highly prevalent lung condition resulting in more than 1 million deaths per year worldwide; thus, significant international attention has been paid to prompt diagnosis and treatment of the disease [15]. Chest x-ray is an effective and cost-efficient modality for pulmonary TB diagnosis, making it a vital clinical tool in low-resource settings where TB is most prevalent [38]. Moreover, labeled CXR datasets are available, and AI for TB diagnosis in CXR has been validated in previous research [14]. The dataset selected for this study was from the U.S. National Library of Medicine (NIH) Shenzhen No. 3 People's Hospital in China [24]. The dataset consists of 662 frontal chest x-ray (CXR) images labeled with each respective patient's TB positive or negative diagnosis. The limited sample size helps demonstrate the effective use of data through IMIL's feedback-based augmentations. The dataset has a class distribution of 49% positive and 51% TB negative. In terms of demographics, the dataset only included sex: 69% male and 31% female. The models in our study were trained on 80% (n = 530) of the dataset and evaluated on 20% (n = 132).

3.3.2 IMIL Configuration

For our study, we configured the IMIL callback to launch at the 70th epoch (`IMIL_epoch`) of training. We also used an IMIL grid size of 4x4 (`grid_size`). Lastly, we use 20 for the `num_outliers` parameter. In summary, IMIL will function by sequentially providing the clinician with 20 misprediction outliers (which accounts for 3.77% of the training dataset) at the 70th epoch and after feedback is provided, training will continue till the 100th epoch using the dataset with the newly augmented samples.

3.3.3 Model Architecture

The CNN architecture used in this study was ResNet-50 [19] which is widely used for medical image analysis, notably through transfer learning [22, 27, 36]. The ResNet model was implemented based on the standard Tensorflow Keras [1] applications plugin [1]. The implementations of the MixUp, CutMix, and CutOut followed the original formulations [2]. For CutOut we used a 50x50 pixel mask size.

3.3.4 Training Details

Each of the seven ResNet-50 models (baseline, CutMix, CutOut, MixUp, IMIL + Resident 1/2/3) were trained for 100 epochs. Each input is trained to two output logits corresponding to the normal and TB class labels. Experiments are performed using the stochastic gradient descent (SGD) optimizer [25], a batch size of 64, and learning rate 0.001. All input images are uniformly scaled to 224x224.

3.3.5 Clinical User-Study

Our clinical user-study to validate the efficacy of IMIL was performed with three radiology residents at Stanford Medicine. These trainees will be referred to as residents 1, 2, and 3. Resident 1 is a PGY-2 diagnostic radiology resident, resident 2 is a PGY-2 interventional radiology resident, and resident 3 is a PGY-1 interventional radiology resident. Three separate ResNet-50 models were trained for each resident. The total time of interaction between each resident and the feedback mechanism lasted less than 30 minutes. Before each training session, two main components of IMIL were described to the resident. The first component is how to interpret the predicted label and CAM output jointly to understand where the model was focusing on to reach its prediction. The second component described is the how to perform the region selection and how the grid overlay functions. The resident was informed about the objective of the feedback: to shift the model's focus from irrelevant to clinically significant features that correlate to the ground-truth label. The residents were instructed to prioritize a single specific region for feedback, despite the possibility of multiple relevant areas.

3.3.6 Evaluation Metrics

Below are descriptions of the statistical metrics used to validate the CNN. The first two metrics are used to validate the performance of the algorithm. The third metric is used to evaluate confidence calibration which is not reflected in performance-based metrics.

Accuracy and AUROC To evaluate the performance of the algorithm, we use the test set accuracy and area under the receiver operating characteristic (AUROC). The accuracy measures the fraction of predictions that were made correctly across the test set after training. AUROC is a robust measure of the ability for the binary classifier to discriminate between class labels [18].

[1] https://keras.io/api/applications

[2] DataAugmentationTF Repository

Expected Calibration Error (ECE) The ECE is commonly used to quantify the level of confidence calibration for algorithms. This approach provides a scalar summary statistic of calibration by grouping a model's predictions into equally spaced bins (B). The weighted average of the difference between accuracy and confidence across the bins is outputted. The formulation of ECE from [17] is shown:

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{n_b}{N} |\,\mathrm{acc}(b) - \mathrm{conf}(b)|, \tag{3}$$

where n represents the number of samples. Gaps in calibration or miscalibration are represented by the difference between acc and conf. In our study, we use a bin size of 15.

4. Results

IMIL data augmentations yielded improvements in performance and calibration over baseline and CutOut. Performance-based metrics are discussed in 4.1 and calibration in 4.2. In 4.3, CAM visualizations from the different algorithms are presented as well as IMIL interaction samples from each clinical resident to assess usability.

4.1. Performance

The performance-based metrics for the different ResNet-50 [19] data augmentation variants are shown in Table 1 in the first two columns (accuracy and AUROC). The baseline ResNet-50 model demonstrated accuracy of 80.4% and AUROC of 83.2% on the CXR test set (Row 1). Of the modern data augmentations, MixUp presented the most significant performance improvements in accuracy and AUROC with accuracy of 83.4% (+3% over ResNet-50) and AUROC of 87.8%. CutMix showed less significant improvements in performance. CutOut reduced accuracy to 79.7% (-0.7% compared to ResNet-50) and AUROC score to 81.7% (-1.5% compared to ResNet-50). Rows 5-7 present the performance of the IMIL-augmented models trained separately for each resident's feedback loop. All IMIL models present performance improvements over baseline with more consistency observed in AUROC compared to accuracy. The first IMIL model (with Resident 1) presented the most significant improvement in accuracy compared to baseline and all other augmentations at 84.6% (+4.2% over ResNet-50). The third IMIL model (with Resident 3) showed the highest improvement in AUROC at 85.8% (+2.6% over ResNet-50). A summary of AUROC performance is also shown in the Figure 3 ROC Curve.

4.2. Calibration

The ECE scores for each algorithm are shown in the last column of Table 1. Lower ECE scores represent higher levels of confidence calibration. The baseline ResNet-50 had an ECE score of 0.2. Every augmented model decreased the ECE except for CutOut, which had almost no effect on it (+0.001 compared to baseline). The most significant decrease in ECE was observed in CutMix at 0.15 (-0.05 less than ResNet-50). MixUp also had significant calibration improvements with an ECE of 0.179 (-0.021 less than ResNet-50). Considering the limited outlier count, IMIL also presented significant and consistent improvements in ECE, although it did not outperform CutMix and CutOut in terms of calibration. The most significant improvement was observed in the second IMIL model (with Resident 2), which received an ECE of 0.175 (-0.025 below ResNet-50).

4.3. Visualizations

To gauge the effect of the various augmentations on explainability, we generate CAM visualizations on four random test set samples images (two TB and two normal) as shown in Figure 4. We examine visualizations for the IMIL Resident 1 model as it yielded the highest test set accuracy. As shown in the figure, there are significant variations in the model's CAM heat-maps based on the augmentation technique used. The heat-map for the baseline seems to be distributed across the lung and hilar lymph node region region, with greatest attention paid to the lower and middle lung zones. MixUp pays preferential attention to the upper zones. CutMix has a patchy heterogenous focus on the lungs. Visually, CutOut appears to focus the least reliably on lung fields, in some cases focusing preferentially on shoulder and thorax soft tissues. In contrast, Res1 IMIL heat-maps are well-focused on all of the lung fields and hilar lymph nodes.

We also present three notable sample interactions between the resident and IMIL to demonstrate the feedback mechanism in Figure 5. The IMIL outlier heat-map and selection grid are presented to the resident along with the predicted label, confidence, and ground-truth. Clinician-selected grid regions are shown along with the augmented output. Each resident also provides a reasoning in the form "The model was focused on .... so I switched the focus towards ...." to understand their decision-making process behind the feedback and effort to shift focus.

5. Discussion

The main objective of this study is developing a human-in-the-loop learning framework and validating the efficacy of clinical feedback during training. We choose a CutOut-like "blackout" augmentation for two reasons. First, within IMIL, asking clinicians to select relevant regions of the image to shift the model's focus is highly intuitive; refocusing the model on lymphadenopathy or a lung opacity as opposed to clavicle or a gastric bubble was easily understood by all participants. Second, it allows for a clear validation of the framework in comparison with the CutOut baseline: random blackout vs. human-guided blackout. Although

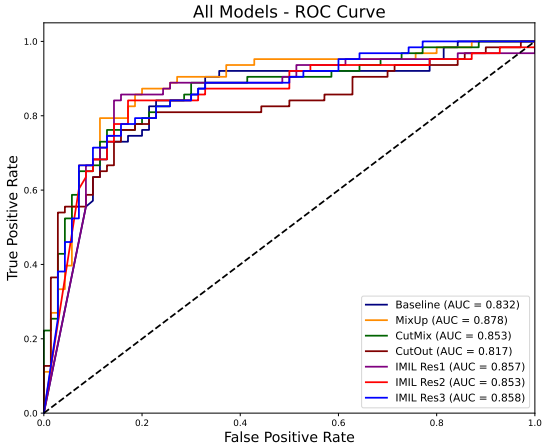

Figure 3. **ROC Curve.** The curve represents the comparison of AUC between the baseline, modern augmentations (MixUp [43], CutMix [41], and CutOut [9]), as well as the three IMIL-trained resident algorithms (IMIL Res1, Res2, and Res3).

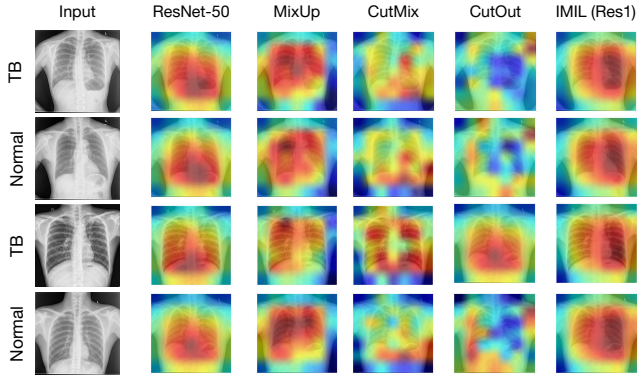

Figure 4. **CAM Visualizations.** Shown above are CAM heat-maps comparing the baseline with the modern augmentations (MixUp [43], CutMix [41], and CutOut [9]) as well as IMIL (for Resident 1 which had the highest test set accuracy).

Model	Accuracy	AUROC	Expected Calibration Error (ECE)
Baseline (ResNet-50 [9])	0.804	0.832	0.2
CutMix [41]	0.812 (+0.008)	0.853 (+0.021)	0.15 (-0.05)
CutOut [9]	0.797 (-0.007)	0.817 (-0.015)	0.201 (+0.001)
MixUp [43]	0.834 (+0.03)	0.878 (+0.046)	0.161 (-0.039)
IMIL + Resident 1	**0.846 (+0.042)**	0.857 (+0.025)	0.179 (-0.021)
IMIL + Resident 2	0.835 (+0.031)	0.853 (+0.021)	**0.175 (-0.025)**
IMIL + Resident 3	0.824 (+0.02)	**0.858 (+0.026)**	0.183 (-0.017)

Table 1. **Validation statistics of baseline, augmentations, and IMIL framework (for the three residents).** Statistics consist of test set accuracy, AUROC score, and Expected Calibration Error (ECE) ($M = 15$ *bins*) for the TB CXR dataset. Bold values represent the best performing IMIL resident model.

this type of augmentation was chosen with the usability of the framework in mind, it is worth considering how human feedback could be integrated into a CutMix or MixUp style of augmentation. CutMix and MixUp seemed to outperform CutOut in our investigation, which suggests that the benefits of IMIL clinician input adapted to these two paradigms could deliver even greater performance and calibration improvements. Future work may investigate whether IMIL could drive accuracy gains in non-CNN models as well.

In the current study, each of the three IMIL experiments were done on a single resident respectively. Each experiment shows performance and calibration improvements over baseline and CutOut. Although the improvements for AUROC are fairly consistent, we do notice more significant variations in accuracy between residents.

Avenues of future research should include variations on dataset/sample size and combinations of clinical participants. We validated IMIL on a single smaller scale dataset where limited feedback (20 outliers) had a significant effect. It is worth studying the effect of the IMIL outlier count on performance, as too few outliers can be insignificant, but too many can overly corrupt the data and shift the domain. It remains to be determined what the optimal ratio of clinician feedback-to-dataset size is when expanding this framework to significantly larger datasets. We hope to apply IMIL to different imaging modalities (CT, MRI) and different disease entities to validate its utility across clinical use cases. We also plan to conduct more user-studies on diagnosticians in various stages of practice (all participating residents in this study have significant familiarity in developing AI for radiology applications) to ensure our framework remains intuitive for a wide swathe of clinical users. For large datasets, it may be necessary to combine feedback from several clinicians into the same model. We would need to test whether a multi-radiologist trained algorithm produces valid results, or whether inter-radiologist

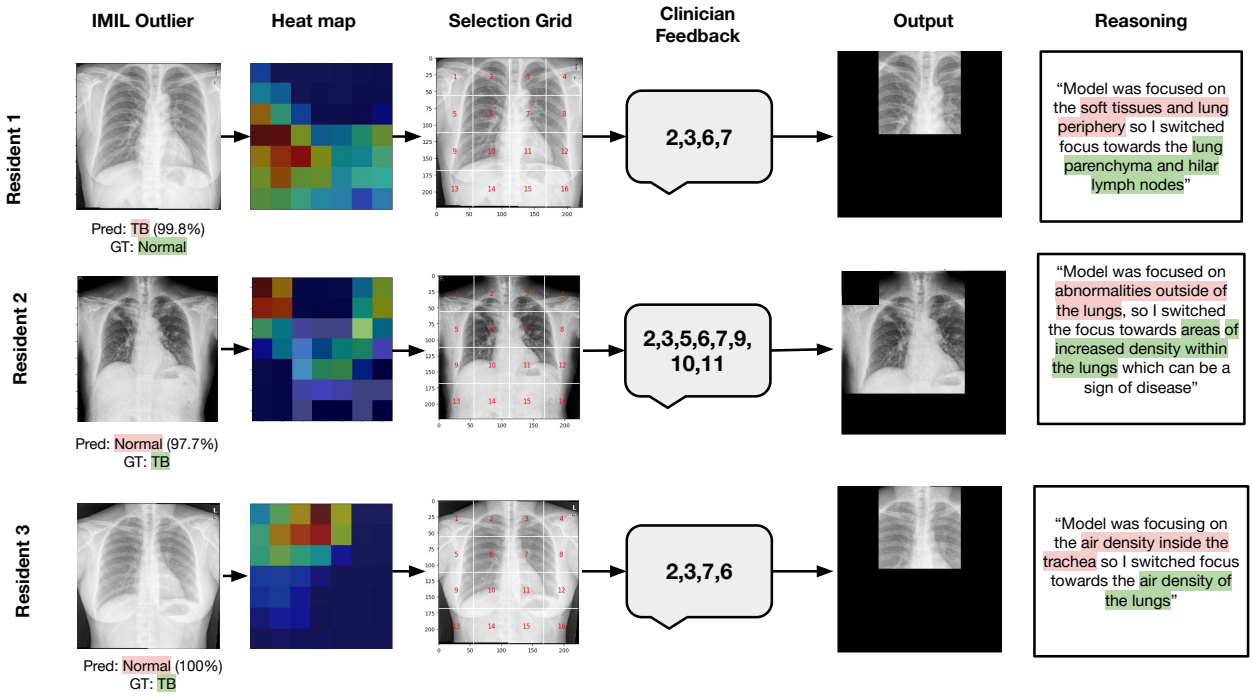

Figure 5. **IMIL Feedback Samples**. End-to-end interaction samples are shown above with the IMIL outlier and the predicted label (with confidence) and the ground-truth. Next to that is the heat-map and selection grid which the resident uses to provide feedback in the form of a numerical input. This is then used to perform the 'blackout' augmentation. Each resident also provided clinical reasoning for their feedback decisions in the samples shown.

variability would serve as a confounder.

Finally, we only launch the IMIL callback at a single instance during training (70th epoch), but the callback may be employed multiple times during training. Therefore, when training models for a greater number of epochs, we could compare the timing and number of IMIL occurrences during training (e.g. comparing a single feedback loop for 20 outliers versus having two loops for 10 outliers each). Optimal grid size is also unknown and may depend on the disease and imaging modality. In some cases, having more grid regions to choose from may improve performance and allow the clinician to provide even more detailed feedback.

From a clinical perspective, IMIL holds the potential to be easily integrated into a clinical radiologist's workflow. Because AI is not yet so diagnostically reliable (and trusted by providers, patients, and the public) that it can be used in absence of a radiologist [2], most AI tools for triaging images to improve radiologist efficiency instead of outright replacing radiologists [26]. Humans therefore still oversee final imaging reports. Discrepancies between model and radiologist interpretations could therefore be fed back into a training loop along with IMIL-style image augmentations to improve subsequent model performance. As patient populations, disease presentations, and diagnostic criteria evolve over time, ensuring AI solutions for radiologists remain flexible will be paramount [29]. Interactive learning for medical imaging can therefore be critical in training better out-of-the-box models for image analysis and easily adapting these models to best suit the specific needs of the hospitals and patient populations in which they are deployed.

6. Conclusion

Using interactive clinician-guided intermediate training feedback from three radiology residents, the IMIL callback framework and data augmentation demonstrates significant and reliable improvements in accuracy, AUROC, and calibration compared with ResNet-50 baseline and CutOut data augmentation for pulmonary tuberculosis classification on chest x-rays. The IMIL framework thus holds great potential for improving performance of computer vision models for medical imaging tasks hampered by small dataset size. Future investigation is needed to elucidate optimal outlier-to-dataset ratio, epoch timing, and image masking techniques, and to validate the model for use among different practitioners, patient populations, and imaging modalities.

References

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org. 5

[2] Manisha Bahl. Artificial intelligence in clinical practice: implementation considerations and barriers. *Journal of Breast Imaging*, 4(6):632–639, 2022. 8

[3] Yoshua Bengio, Frédéric Bastien, Arnaud Bergeron, Nicolas Boulanger-Lewandowski, Thomas Breuel, Youssouf Chherawala, Moustapha Cisse, Myriam Côté, Dumitru Erhan, Jeremy Eustache, et al. Deep learners benefit more from out-of-distribution examples. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 164–172. JMLR Workshop and Conference Proceedings, 2011. 4

[4] Samuel Budd, Emma C Robinson, and Bernhard Kainz. A survey on active learning and human-in-the-loop deep learning for medical image analysis. *Medical image analysis*, 71: 102062, 2021. 3

[5] Heang-Ping Chan, Ravi K Samala, Lubomir M Hadjiiski, and Chuan Zhou. Deep learning in medical image analysis. *Deep Learning in Medical Image Analysis*, pages 3–21, 2020. 1

[6] Haomin Chen, Catalina Gomez, Chien-Ming Huang, and Mathias Unberath. Explainable medical imaging ai needs human-centered design: guidelines and evidence from a systematic review. *NPJ digital medicine*, 5(1):156, 2022. 1

[7] Phillip Chlap, Hang Min, Nym Vandenberg, Jason Dowling, Lois Holloway, and Annette Haworth. A review of medical image data augmentation techniques for deep learning applications. *Journal of Medical Imaging and Radiation Oncology*, 65(5):545–563, 2021. 2

[8] Alex J DeGrave, Joseph D Janizek, and Su-In Lee. Ai for radiographic covid-19 detection selects shortcuts over signal. *Nature Machine Intelligence*, 3(7):610–619, 2021. 1, 3

[9] Terrance DeVries and Graham W Taylor. Improved regularization of convolutional neural networks with cutout. *arXiv preprint arXiv:1708.04552*, 2017. 1, 2, 3, 4, 7

[10] Zach Eaton-Rosen, Felix Bragman, Sebastien Ourselin, and M Jorge Cardoso. Improving data augmentation for medical image segmentation. 2018. 3

[11] Andre Esteva, Katherine Chou, Serena Yeung, Nikhil Naik, Ali Madani, Ali Mottaghi, Yun Liu, Eric Topol, Jeff Dean, and Richard Socher. Deep learning-enabled medical computer vision. *NPJ digital medicine*, 4(1):5, 2021. 1

[12] Mélanie Gaillochet, Christian Desrosiers, and Hervé Lombaert. Active learning for medical image segmentation with stochastic batches. *Medical Image Analysis*, 90:102958, 2023. 3

[13] Adrian Galdran, Gustavo Carneiro, and Miguel A González Ballester. Balanced-mixup for highly imbalanced medical image classification. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 323–333. Springer, 2021. 3

[14] C Geric, ZZ Qin, CM Denkinger, SV Kik, B Marais, A Anjos, PM David, F Ahmad Khan, and A Trajman. The rise of artificial intelligence reading of chest x-rays for enhanced tb diagnosis and elimination. *The International Journal of Tuberculosis and Lung Disease*, 27(5):367–372, 2023. 5

[15] Philippe Glaziou, Dennis Falzon, Katherine Floyd, and Mario Raviglione. Global epidemiology of tuberculosis. In *Seminars in respiratory and critical care medicine*, pages 003–016. Thieme Medical Publishers, 2013. 5

[16] Google. About recaptcha. https://www.google.com/recaptcha/about/, 2024. Accessed: 2024-03-31. 2

[17] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR, 2017. 2, 6

[18] James A Hanley and Barbara J McNeil. The meaning and use of the area under a receiver operating characteristic (roc) curve. *Radiology*, 143(1):29–36, 1982. 5

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2, 5, 6

[20] Geoffrey E Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan R Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint arXiv:1207.0580*, 2012. 4

[21] Corey Horien, Stephanie Noble, Abigail S Greene, Kangjoo Lee, Daniel S Barron, Siyuan Gao, David O'Connor, Mehraveh Salehi, Javid Dadashkarimi, Xilin Shen, et al. A hitchhiker's guide to working with large, open-source neuroimaging datasets. *Nature human behaviour*, 5(2):185–193, 2021. 1

[22] Md Belal Hossain, SM Hasan Sazzad Iqbal, Md Monirul Islam, Md Nasim Akhtar, and Iqbal H Sarker. Transfer learning with fine-tuned deep cnn resnet50 model for classifying covid-19 from chest x-ray images. *Informatics in Medicine Unlocked*, 30:100916, 2022. 5

[23] Shih-Cheng Huang, Tanay Kothari, Imon Banerjee, Chris Chute, Robyn L Ball, Norah Borus, Andrew Huang, Bhavik N Patel, Pranav Rajpurkar, Jeremy Irvin, et al. Penet—a scalable deep-learning model for automated diagnosis of pulmonary embolism using volumetric ct imaging. *NPJ digital medicine*, 3(1):1–9, 2020. 1

[24] Stefan Jaeger, Sema Candemir, Sameer Antani, Yì-Xiáng J Wáng, Pu-Xuan Lu, and George Thoma. Two public chest x-ray datasets for computer-aided screening of pulmonary diseases. *Quantitative imaging in medicine and surgery*, 4(6): 475, 2014. 3, 5

[25] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 5

[26] Claudia Mello-Thoms and Carlos AB Mello. Clinical applications of artificial intelligence in radiology. *The British Journal of Radiology*, 96(1150):20221031, 2023. 8

[27] Quang H Nguyen, Binh P Nguyen, Son D Dao, Balagopal Unnikrishnan, Rajan Dhingra, Savitha Rani Ravichandran, Sravani Satpathy, Palaparthi Nirmal Raja, and Matthew CH Chua. Deep learning models for tuberculosis detection from chest x-ray images. In *2019 26th international conference on telecommunications (ICT)*, pages 381–385. IEEE, 2019. 5

[28] Allison Park, Chris Chute, Pranav Rajpurkar, Joe Lou, Robyn L Ball, Katie Shpanskaya, Rashad Jabarkheel, Lily H Kim, Emily McKenna, Joe Tseng, et al. Deep learning–assisted diagnosis of cerebral aneurysms using the headxnet model. *JAMA network open*, 2(6):e195600–e195600, 2019. 1

[29] Oleg S Pianykh, Georg Langs, Marc Dewey, Dieter R Enzmann, Christian J Herold, Stefan O Schoenberg, and James A Brink. Continuous learning ai in radiology: implementation principles and early applications. *Radiology*, 297(1):6–14, 2020. 8

[30] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brandon Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, et al. Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. *arXiv preprint arXiv:1711.05225*, 2017. 1

[31] Pranav Rajpurkar, Allison Park, Jeremy Irvin, Chris Chute, Michael Bereket, Domenico Mastrodicasa, Curtis P Langlotz, Matthew P Lungren, Andrew Y Ng, and Bhavik N Patel. Appendixnet: Deep learning for diagnosis of appendicitis from a small dataset of ct exams using video pretraining. *Scientific reports*, 10(1):1–7, 2020. 1

[32] Adrit Rao, Jongchan Park, Sanghyun Woo, Joon-Young Lee, and Oliver Aalami. Studying the effects of self-attention for medical image analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3416–3425, 2021. 1, 3

[33] Adrit Rao, Joon-Young Lee, and Oliver Aalami. Studying the impact of augmentations on medical confidence calibration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2462–2472, 2023. 2, 3

[34] Adriel Saporta, Xiaotong Gui, Ashwin Agrawal, Anuj Pareek, Steven QH Truong, Chanh DT Nguyen, Van-Doan Ngo, Jayne Seekins, Francis G Blankenberg, Andrew Y Ng, et al. Benchmarking saliency methods for chest x-ray interpretation. *Nature Machine Intelligence*, 4(10):867–878, 2022. 3

[35] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 2, 3

[36] Sadia Showkat and Shaima Qureshi. Efficacy of transfer learning-based resnet models in chest x-ray image classification for detecting covid-19 pneumonia. *Chemometrics and Intelligent Laboratory Systems*, 224:104534, 2022. 5

[37] Asim Smailagic, Pedro Costa, Hae Young Noh, Devesh Walawalkar, Kartik Khandelwal, Adrian Galdran, Mostafa Mirshekari, Jonathon Fagert, Susu Xu, Pei Zhang, 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)*, pages 481–488. IEEE, 2018. 3

[38] MRA Van Cleeff, LE Kivihya-Ndugga, H Meme, JA Odhiambo, and PR Klatser. The role and performance of chest x-ray for the diagnosis of tuberculosis: a cost-effectiveness analysis in nairobi, kenya. *BMC infectious diseases*, 5:1–9, 2005. 5

[39] Julia K Winkler, Christine Fink, Ferdinand Toberer, Alexander Enk, Teresa Deinlein, Rainer Hofmann-Wellenhof, Luc Thomas, Aimilios Lallas, Andreas Blum, Wilhelm Stolz, et al. Association between surgical skin markings in dermoscopic images and diagnostic performance of a deep learning convolutional neural network for melanoma recognition. *JAMA dermatology*, 155(10):1135–1141, 2019. 1

[40] Lin Yang, Yizhe Zhang, Jianxu Chen, Siyuan Zhang, and Danny Z Chen. Suggestive annotation: A deep active learning framework for biomedical image segmentation. In *Medical Image Computing and Computer Assisted Intervention-MICCAI 2017: 20th International Conference, Quebec City, QC, Canada, September 11-13, 2017, Proceedings, Part III 20*, pages 399–407. Springer, 2017. 3

[41] Sangdoo Yun, Dongyoon Han, Seong Joon Oh, Sanghyuk Chun, Junsuk Choe, and Youngjoon Yoo. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6023–6032, 2019. 2, 3, 4, 7

[42] John R Zech, Marcus A Badgeley, Manway Liu, Anthony B Costa, Joseph J Titano, and Eric Karl Oermann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. *PLoS medicine*, 15(11):e1002683, 2018. 1

[43] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 2, 3, 4, 7