

# Distilling BlackBox to Interpretable models for Efficient Transfer Learning

Shantanu Ghosh<sup>1</sup>(✉), Ke Yu<sup>2</sup>, and Kayhan Batmanghelich<sup>1</sup>

<sup>1</sup> Department of Electrical and Computer Engineering, Boston University, Boston, MA, USA

shawn24@bu.edu

<sup>2</sup> Intelligent Systems Program, University of Pittsburgh, Pittsburgh, PA, USA

**Abstract.** Building generalizable AI models is one of the primary challenges in the healthcare domain. While radiologists rely on generalizable descriptive rules of abnormality, Neural Network (NN) models suffer even with a slight shift in input distribution (*e.g.*, scanner type). Fine-tuning a model to transfer knowledge from one domain to another requires a significant amount of labeled data in the target domain. In this paper, we develop an interpretable model that can be efficiently fine-tuned to an unseen target domain with minimal computational cost. We assume the interpretable component of NN to be approximately domain-invariant. However, interpretable models typically underperform compared to their Blackbox (BB) variants. We start with a BB in the source domain and distill it into a *mixture* of shallow interpretable models using human-understandable concepts. As each interpretable model covers a subset of data, a mixture of interpretable models achieves comparable performance as BB. Further, we use the pseudo-labeling technique from semi-supervised learning (SSL) to learn the concept classifier in the target domain, followed by fine-tuning the interpretable models in the target domain. We evaluate our model using a real-life large-scale chest-X-ray (CXR) classification dataset. The code is available at: <https://github.com/batmanlab/MICCAI-2023-Route-interpret-repeat-CXRs>.

**Keywords:** Explainable-AI · Interpretable models · Transfer learning

## 1 Introduction

Model generalizability is one of the main challenges of AI, especially in high stake applications such as healthcare. While NN models achieve state-of-the-art (SOTA) performance in disease classification [9, 17, 24], they are brittle to small shifts in the data distribution [7] caused by a change in acquisition protocol or scanner type [22]. Fine-tuning all or some layers of a NN model on the target domain can alleviate this problem [2], but it requires a substantial amount of labeled data and be computationally expensive [12, 21]. In contrast, radiologists follow fairly generalizable and comprehensible rules. Specifically, they search for patterns of changes in anatomy to read abnormality from an image and apply logical rules for specific diagnoses. This approach is transparent and closer to an

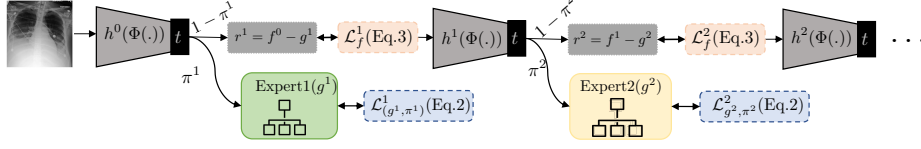
interpretable-by-design approach in AI. We develop a method to extract a mixture of interpretable models based on clinical concepts, similar to radiologists’ rules, from a pre-trained NN. Such a model is more data- and computation-efficient than the original NN for fine-tuning to a new distribution.

Standard interpretable by design method [18] finds an interpretable function (*e.g.*, linear regression or rule-based) between human-interpretable concepts and final output [14]. A concept classifier [19, 26] detects the presence or absence of concepts in an image. In medical images, previous research uses TCAV scores [13] to quantify the role of a concept on the final prediction [3, 6, 23], but the concept-based interpretable models have been mostly unexplored. Recently Posthoc Concept Bottleneck models (PCBMs) [25] identify concepts from the embeddings of BB. However, the common design choice amongst those methods relies on a single interpretable classifier to explain the entire dataset, cannot capture the diverse sample-specific explanations, and performs poorly than their BB variants.

**Our contributions.** This paper proposes a novel data-efficient interpretable method that can be transferred to an unseen domain. Our interpretable model is built upon human-interpretable concepts and can provide sample-specific explanations for diverse disease subtypes and pathological patterns. Beginning with a BB in the source domain, we progressively extract a mixture of interpretable models from BB. Our method includes a set of selectors routing the explainable samples through the interpretable models. The interpretable models provide First-order-logic (FOL) explanations for the samples they cover. The remaining unexplained samples are routed through the residuals until they are covered by a successive interpretable model. We repeat the process until we cover a desired fraction of data. Due to class imbalance in large CXR datasets, early interpretable models tend to cover all samples with disease present while ignoring disease subgroups and pathological heterogeneity. We address this problem by estimating the class-stratified coverage from the total data coverage. We then finetune the interpretable models in the target domain. The target domain lacks concept-level annotation since they are expensive. Hence, we learn a concept detector in the target domain with a pseudo labeling approach [15] and finetune the interpretable models. Our work is the first to apply concept-based methods to CXRs and transfer them between domains.

## 2 Methodology

**Notation.** Assume  $f^0 : \mathcal{X} \rightarrow \mathcal{Y}$  is a BB, trained on a dataset  $\mathcal{X} \times \mathcal{Y} \times \mathcal{C}$ , with  $\mathcal{X}$ ,  $\mathcal{Y}$ , and  $\mathcal{C}$  being the images, classes, and concepts, respectively;  $f^0 = h^0 \circ \Phi$ , where  $\Phi$  and  $h^0$  is the feature extractor and the classifier respectively. Also,  $m$  is the number of class labels. This paper focuses on binary classification (having or not having a disease), so  $m = 2$  and  $\mathcal{Y} \in \{0, 1\}$ . Yet, it can be extended to multiclass problems easily. Given a learnable projection [4, 5],  $t : \Phi \rightarrow \mathcal{C}$ , our method learns three functions: (1) a set of selectors ( $\pi : \mathcal{C} \rightarrow \{0, 1\}$ ) routing samples to an interpretable model or residual, (2) a set of interpretable models ( $g : \mathcal{C} \rightarrow \mathcal{Y}$ ), and (3) the residuals. The interpretable models are called “experts” since they



**Fig. 1.** Schematic view of our method. Note that  $f^k(.) = h^k(\Phi(.))$ . At iteration  $k$ , the selector *routes* each sample either towards the expert  $g^k$  with probability  $\pi^k(.)$  or the residual  $r^k = f^{k-1} - g^k$  with probability  $1 - \pi^k(.)$ .  $g^k$  generates FOL-based explanations for the samples it covers. Note  $\Phi$  is fixed across iterations.

specialize in a distinct subset of data defined by that iteration’s coverage  $\tau$  as shown in SelectiveNet [16]. Fig. 1 illustrates our method.

## 2.1 Distilling BB to the mixture of interpretable models

**Handling class imbalance.** For an iteration  $k$ , we first split the given coverage  $\tau^k$  to stratified coverages per class as  $\{\tau_m^k = w_m \cdot \tau^k; w_m = N_m/N, \forall m\}$ , where  $w_m$  denotes the fraction of samples belonging to the  $m^{th}$  class;  $N_m$  and  $N$  are the samples of  $m^{th}$  class and total samples, respectively.

**Learning the selectors.** At iteration  $k$ , the selector  $\pi^k$  routes  $i^{th}$  sample to the expert ( $g^k$ ) or residual ( $r^k$ ) with probability  $\pi^k(\mathbf{c}_i)$  and  $1 - \pi^k(\mathbf{c}_i)$  respectively. For coverages  $\{\tau_m^k, \forall m\}$ , we learn  $g^k$  and  $\pi^k$  jointly by solving the loss:

$$\theta_{s^k}^*, \theta_{g^k}^* = \arg \min_{\theta_{s^k}, \theta_{g^k}} \mathcal{R}^k(\pi^k(.; \theta_{s^k}), g^k(.; \theta_{g^k})) \quad \text{s.t.} \quad \zeta_m(\pi^k(.; \theta_{s^k})) \geq \tau_m^k \quad \forall m, \quad (1)$$

where  $\theta_{s^k}^*, \theta_{g^k}^*$  are the optimal parameters for  $\pi^k$  and  $g^k$ , respectively.  $\mathcal{R}^k$  is

$$\text{the overall selective risk, defined as, } \mathcal{R}^k(\pi^k, g^k) = \sum_m \frac{\frac{1}{N_m} \sum_{i=1}^{N_m} \mathcal{L}_{(g^k, \pi^k)}^k(\mathbf{x}_i, \mathbf{c}_i)}{\zeta_m(\pi^k)}$$

, where  $\zeta_m(\pi^k) = \frac{1}{N_m} \sum_{i=1}^{N_m} \pi^k(\mathbf{c}_i)$  is the empirical mean of samples of  $m^{th}$  class selected by the selector for the associated expert  $g^k$ . We define  $\mathcal{L}_{(g^k, \pi^k)}^k$  in the next section. The selectors are neural networks with sigmoid activation. At inference time,  $\pi^k$  routes a sample to  $g^k$  if and only if  $\pi^k(.) \geq 0.5$ .

**Learning the experts.** For iteration  $k$ , the loss  $\mathcal{L}_{(g^k, \pi^k)}^k$  distills the expert  $g^k$  from  $f^{k-1}$ , BB of the previous iteration by solving the following loss:

$$\mathcal{L}_{(g^k, \pi^k)}^k(\mathbf{x}_i, \mathbf{c}_i) = \underbrace{\ell(f^{k-1}(\mathbf{x}_i), g^k(\mathbf{c}_i)) \pi^k(\mathbf{c}_i)}_{\text{trainable component for current iteration } k} \underbrace{\prod_{j=1}^{k-1} (1 - \pi^j(\mathbf{c}_i))}_{\text{fixed component trained in the previous iterations}}, \quad (2)$$

where  $\pi^k(\mathbf{c}_i) \prod_{j=1}^{k-1} (1 - \pi^j(\mathbf{c}_i))$  is the cumulative probability of the sample covered by the residuals for all the previous iterations from  $1, \dots, k-1$  (i.e.,  $\prod_{j=1}^{k-1} (1 - \pi^j(\mathbf{c}_i))$ ) and the expert  $g^k$  at iteration  $k$  (i.e.,  $\pi^k(\mathbf{c}_i)$ ).

**Learning the Residuals.** After learning  $g^k$ , we calculate the residual as,  $r^k(x_i, c_i) = f^{k-1}(x_i) - g^k(c_i)$  (difference of logits). We fix  $\Phi$  and optimize the following loss to update  $h^k$  to specialize on those samples not covered by  $g^k$ , effectively creating a new BB  $f^k$  for the next iteration ( $k + 1$ ):

$$\mathcal{L}_f^k(\mathbf{x}_j, \mathbf{c}_j) = \underbrace{\ell(r^k(\mathbf{x}_j, \mathbf{c}_j), f^k(\mathbf{x}_j))}_{\text{trainable component for iteration } k} \underbrace{\prod_{i=1}^k (1 - \pi^i(\mathbf{c}_j))}_{\text{non-trainable component for iteration } k} \quad (3)$$

We refer to all the experts as the Mixture of Interpretable Experts (MoIE-CXR). We denote the models, including the final residual, as MoIE-CXR+R. Each expert in MoIE-CXR constructs sample-specific FOLs using the optimization strategy and algorithm discussed in [4].

## 2.2 Finetuning to an unseen domain

We assume the MoIE-CXR-identified concepts to be generalizable to an unseen domain. So, we learn the projection  $t_t$  for the target domain and compute the pseudo concepts using SSL [15]. Next, we transfer the selectors, experts, and final residual ( $\{\pi_s^k, g_s^k\}_{k=1}^K$  and  $f_s^K$ ) from the source to a target domain with limited labeled data and computational cost. Algorithm 1 details the procedure.

---

### Algorithm 1 Finetuning to an unseen domain.

---

- 1: **Input:** Learned selectors, experts, and final residual from source domain:  $\{\pi_s^k, g_s^k\}_{k=1}^K$  and  $f_s^K$  respectively, with  $K$  as the number of experts to transfer. BB of the source domain:  $f_s^0 = h_s^0(\Phi_s)$ . Source data:  $\mathcal{D}_s = \{\mathcal{X}_s, \mathcal{C}_s, \mathcal{Y}_s\}$ . Target data:  $\mathcal{D}_t = \{\mathcal{X}_t, \mathcal{Y}_t\}$ . Target coverages  $\{\tau_k\}_{k=1}^K$ .
  - 2: **Output:** Experts  $\{\pi_t^k, g_t^k\}_{k=1}^K$  and final residual  $f_t^K$  of the target domain.
  - 3: Randomly select  $n_t \ll N_t$  samples out of  $N_t = |\mathcal{D}_t|$ .
  - 4: Compute the pseudo concepts for the correctly classified samples in the target domain using  $f_s^0$ , as,  $\mathbf{c}_t^i = t_s(\Phi_s(\mathbf{x}_s^i))$  s.t.,  $y_t^i = f_s^0(\mathbf{x}_t^i)$ ,  $i = 1 \dots n_t$ .
  - 5: Learn the projection function  $t_t$  for target domain semi-supervisedly [15] using the pseudo labeled samples  $\{\mathbf{x}_t^i, \mathbf{c}_t^i\}_{i=1}^{n_t}$  and unlabeled samples  $\{\mathbf{x}_t^i\}_{i=1}^{N_t - n_t}$ .
  - 6: Complete the triplet for the target domain  $\{\mathcal{X}_t, \mathcal{C}_t, \mathcal{Y}_t\}$ , where  $\mathbf{c}_t^i = t_t(\Phi_s(\mathbf{x}_t^i))$ ,  $i = 1 \dots N_t$ .
  - 7: Finetune  $\{\pi_s^k, g_s^k\}_{k=1}^K$  and  $f_s^K$  to obtain  $\{\pi_t^k, g_t^k\}_{k=1}^K$  and  $f_t^K$  using equations 1, 2 and 3 respectively for 5 epochs.  $\{\pi_t^k, g_t^k\}_{k=1}^K$  and  $\{\{\pi_t^k, g_t^k\}_{k=1}^K, f_t^K\}$  represents MoIE-CXR and MoIE-CXR + R for the target domain.
- 

## 3 Experiments

We perform experiments to show that MoIE-CXR 1) captures a diverse set of concepts, 2) does not compromise BB’s performance, 3) covers “harder” in-

stances with the residuals in later iterations resulting in their drop in performance, 4) is finetuned well to an unseen domain with minimal computation.

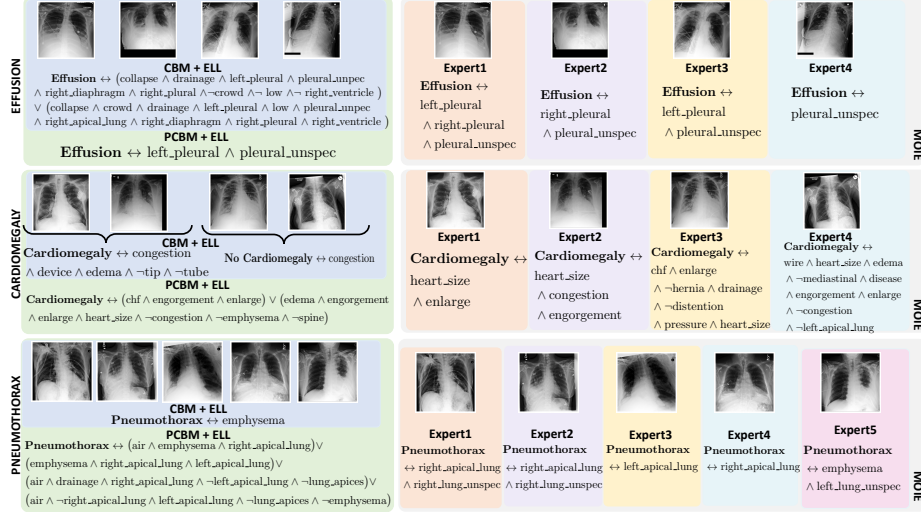


Fig. 2. Qualitative comparison of MoIE-CXR discovered concepts with the baselines.

**Experimental Details.** We evaluate our method using 220,763 frontal images from the MIMIC-CXR dataset [11]. We use Densenet121 [8] as BB ( $f^0$ ) to classify cardiomegaly, effusion, edema, pneumonia, and pneumothorax, considering each to be a separate binary classification problem. We obtain 107 anatomical and observation concepts from the RadGraph’s inference dataset [10], automatically generated by DYGIE++ [20]. We train BB following [24]. To retrieve the concepts, we utilize until the 4<sup>th</sup> Densenet block as feature extractor  $\Phi$  and flatten the features to learn  $t$ . We use an 80%-10%-10% train-validation-test split with no patient shared across splits. We use 4, 4, 5, 5, and 5 experts for cardiomegaly, pneumonia, effusion, pneumothorax, and edema. We employ ELL [1] as  $g$ . Further, we only include concepts as input to  $g$  if their validation auroc exceeds 0.7. Refer to Tab. 1 in the supplementary material for the hyperparameters. We stop until all the experts cover at least 90% of the data cumulatively.

**Baseline.** We compare our method with 1) end-to-end CEM [26], 2) sequential CBM [14], and 3) PCBM [25] baselines, comprising of two parts: a) concept predictor  $\Phi : \mathcal{X} \rightarrow \mathcal{C}$ , predicting concepts from images, with all the convolution blocks; and b) label predictor,  $g : \mathcal{C} \rightarrow \mathcal{Y}$ , predicting labels from the concepts. We create CBM + ELL and PCBM + ELL by replacing the standard classifier with the identical  $g$  of MoIE-CXR to generate FOLs [1] for the baseline.

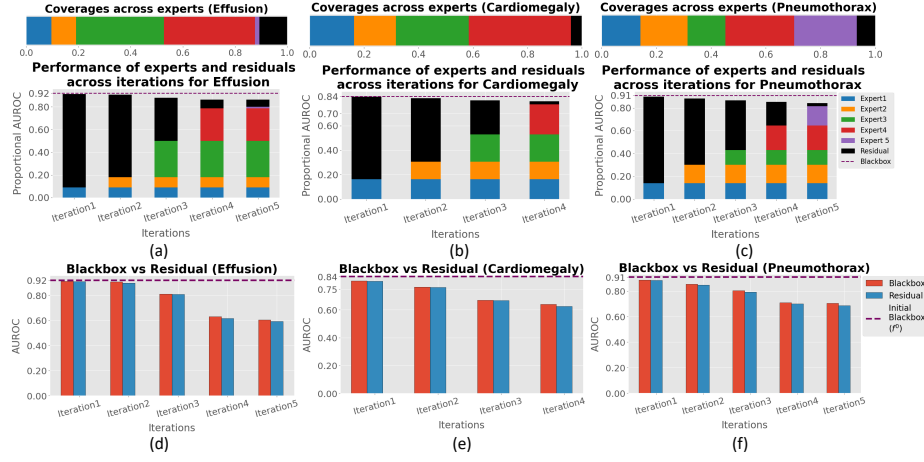
**MoIE-CXR captures diverse explanations.** Fig. 2 illustrates the FOL explanations. Recall that the experts ( $g$ ) in MoIE-CXR and the baselines are

ELLS [1], attributing attention weights to each concept. A concept with high attention weight indicates its high predictive significance. With a single  $g$ , the baselines rank the concepts in accordance with the identical order of attention weights for all the samples in a class, yielding a generic FOL for that class. In Fig. 2, the baseline PCBM + ELL uses *left\_pleural* and *pleural\_unspec* to identify effusion for all four samples. MoIE-CXR deploys multiple experts, learning to specialize in distinct subsets of a class. So different interpretable models in MoIE assign different attention weights to capture instance-specific concepts unique to each subset. In Fig. 2 expert2 relies on *right\_pleural* and *pleural\_unspec*, but expert4 relies only on *pleural\_unspec* to classify effusion. The results show that the learned experts can provide more precise explanations at the subject level using the concepts, increasing confidence and trust in clinical use.

**Table 1.** MoIE-CXR does not compromise the performance of BB. We provide the mean and standard errors of AUROC over five random seeds. For MoIE-CXR, we also report the percentage of test set samples covered by all experts as “Coverage”. We boldfaced our results and BB.

Model	Effusion	Cardiomegaly	Edema	Pneumonia	Pneumothorax
Blackbox (BB)	<b>0.92</b>	<b>0.84</b>	<b>0.89</b>	<b>0.79</b>	<b>0.91</b>
<b>INTERPRETABLE BY DESIGN</b>					
CEM [26]	0.83 $\pm$ 1e-4	0.75 $\pm$ 1e-4	0.77 $\pm$ 2e-4	0.62 $\pm$ 4e-4	0.76 $\pm$ 3e-4
CBM (Sequential) [14]	0.78 $\pm$ 1e-4	0.72 $\pm$ 1e-4	0.77 $\pm$ 5e-4	0.60 $\pm$ 1e-3	0.75 $\pm$ 6e-4
CBM + ELL [1, 14]	0.81 $\pm$ 1e-4	0.72 $\pm$ 1e-4	0.79 $\pm$ 5e-4	0.62 $\pm$ 8e-4	0.75 $\pm$ 6e-4
<b>POSTHOC</b>					
PCBM [25]	0.88 $\pm$ 1e-4	0.81 $\pm$ 1e-4	0.82 $\pm$ 1e-4	0.72 $\pm$ 1e-4	0.85 $\pm$ 7e-4
PCBM-h [25]	0.90 $\pm$ 1e-4	0.83 $\pm$ 1e-4	0.85 $\pm$ 1e-4	0.77 $\pm$ 1e-4	0.89 $\pm$ 7e-4
PCBM + ELL [1, 25]	0.90 $\pm$ 1e-4	0.82 $\pm$ 1e-4	0.85 $\pm$ 1e-4	0.75 $\pm$ 1e-4	0.85 $\pm$ 6e-4
PCBM-h + ELL [1, 25]	0.91 $\pm$ 1e-4	0.83 $\pm$ 1e-4	0.87 $\pm$ 1e-4	0.77 $\pm$ 1e-4	0.90 $\pm$ 1e-4
<b>OURS</b>					
MoIE-CXR (Coverage)	<b>0.93</b> <sup>(0.90)</sup> $\pm$ 1e-4	<b>0.85</b> <sup>(0.96)</sup> $\pm$ 1e-4	<b>0.91</b> <sup>(0.92)</sup> $\pm$ 1e-4	<b>0.80</b> <sup>(0.97)</sup> $\pm$ 1e-4	<b>0.91</b> <sup>(0.93)</sup> $\pm$ 2e-4
MoIE-CXR+R	<b>0.91</b> $\pm$ 1e-4	<b>0.82</b> $\pm$ 1e-4	<b>0.88</b> $\pm$ 1e-4	<b>0.78</b> $\pm$ 1e-4	<b>0.90</b> $\pm$ 2e-4

**MoIE-CXR does not compromise BB’s performance. Analysing MoIE-CXR:** Tab. 1 shows that MoIE-CXR outperforms other models, including BB. Recall that MoIE-CXR refers to the mixture of all interpretable experts, excluding any residuals. As MoIE-CXR specializes in various subsets of data, it effectively discovers sample-specific classifying concepts and achieves superior performance. In general, MoIE-CXR exceeds the interpretable-by-design baselines (CEM, CBM, and CBM + ELL) by a fair margin (on average, at least  $\sim 10\%$   $\uparrow$ ), especially for pneumonia and pneumothorax where the number of samples with the disease is significantly less ( $\sim 750/24000$  in the testset). **Analysing MoIE-CXR+R:** To compare the performance on the entire dataset, we additionally report MoIE-CXR+R, the mixture of interpretable experts with the final residual in Tab.1. MoIE-CXR+R outperforms the interpretable-by-design models

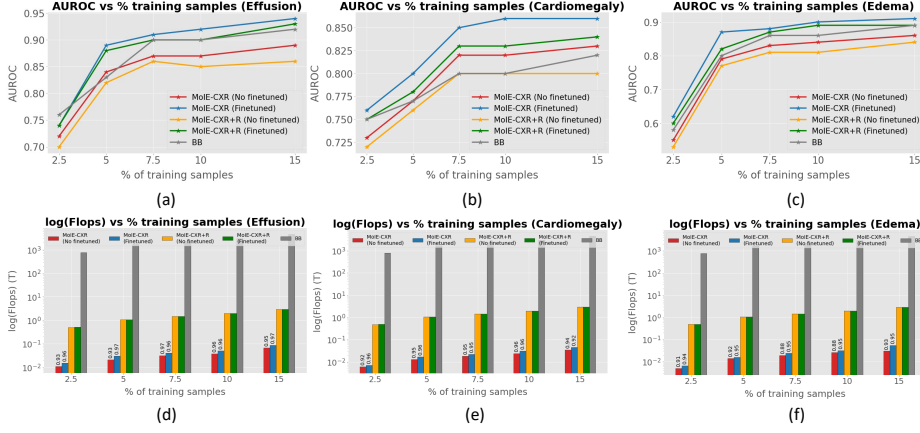


**Fig. 3.** Performance of experts and residuals across iterations. **(a-c):** Coverage and proportional AUROC of the experts and residuals. **(d-f):** Routing the samples covered by MoIE-CXR to the initial  $f^0$ , we compare the performance of the residuals with  $f^0$ .

and yields comparable performance as BB. The residualized PCBM baseline, *i.e.*, PCBM-h, performs similarly to MoIE-CXR+R. PCBM-h rectifies the interpretable PCBM’s mistakes by learning the residual with the complete dataset to resemble BB’s performance. However, the experts and the final residual approximate the interpretable and uninterpretable fractions of BB, respectively. In each iteration, the residual focuses on the samples not covered by the respective expert to create BB for the next iteration and likewise. As a result, the final residual in MoIE-CXR+R covers the “hardest” examples, reducing its overall performance relative to MoIE-CXR.

**Identification of harder samples by successive residuals.** Fig. 3 (a-c) reports the proportional AUROC of the experts and the residuals per iteration. The proportional AUROC is the AUROC of that model times the empirical coverage,  $\zeta^k$ , the mean of the samples routed to the model by the respective selector ( $\pi^k$ ). According to Fig. 3a in iteration 1, the residual (black bar) contributes more to the proportional AUROC than the expert1 (blue bar) for effusion with both achieving a cumulative proportional AUROC  $\sim 0.92$ . All the final experts collectively extract the entire interpretable component from BB  $f^0$  in the final iteration, resulting in their more significant contribution to the cumulative performance. In subsequent iterations, the proportional AUROC decreases as the experts are distilled from the BB of the previous iteration. The BB is derived from the residual that performs progressively worse with each iteration. The residual of the final iteration covers the “hardest” samples. Tracing these samples back to the original BB  $f^0$ ,  $f^0$  underperforms on these samples (Fig. 3 (d-f)) as the residual.





**Fig. 4.** Transferring the first 3 experts of MoIE-CXR trained on MIMIC-CXR to Stanford-CXR. With varying % of training samples of Stanford CXR, (a-c): reports AUROC of the test sets, (d-g) reports computation costs in terms of log (Flops) (T). We report the coverages in Stanford-CXR on top of the “finetuned” and “No finetuned” variants of MoIE-CXR (red and blue bars) in (d-g).

**Applying MoIE-CXR to the unseen domain.** In this experiment, we utilize Algo. 1 to transfer MoIE-CXR trained on MIMIC-CXR dataset to Stanford Chexpert [9] dataset for the diseases – effusion, cardiomegaly and edema. Using 2.5%, 5%, 7.5%, 10%, and 15 % of training data from the Stanford Chexpert dataset, we employ two variants of MoIE-CXR where we (1) train only the selectors ( $\pi$ ) without finetuning the experts ( $g$ ) (“No finetuned” variant of MoIE-CXR in Fig. 4), and (2) finetune  $\pi$  and  $g$  jointly for only 5 epochs (“Finetuned” variant of MoIE-CXR and MoIE-CXR + R in Fig. 4). Finetuning  $\pi$  is essential to route the samples of the target domain to the appropriate expert. As later experts cover the “harder” samples of MIMIC-CXR, we only transfer the experts of the first three iterations (refer to Fig. 3). To ensure a fair comparison, we finetune (both the feature extractor  $\Phi$  and classifier  $h^0$ ) BB:  $f^0 = h^0 \circ \Phi$  of MIMIC-CXR with the same training data of Stanford Chexpert for 5 epochs. Throughout this experiment, we fix  $\Phi$  while finetuning the final residual in MoIE+R as stated in Eq. 3. Fig. 4 displays the performances of different models and the computation costs in terms of Flops. The Flops are calculated as, Flop of (forward propagation + backward propagation)  $\times$  (total no. of batches)  $\times$  (no of training epochs). The finetuned MoIE-CXR outperforms the finetuned BB (on average  $\sim 5\% \uparrow$  for effusion and cardiomegaly). As experts are simple models [1] and accept only low dimensional concept vectors compared to BB, the computational cost to train MoIE-CXR is significantly lower than that of BB (Fig. 4 (d-f)). Specifically, BB requires  $\sim 776$ T flops to be finetuned on 2.5% of the training



data of Stanford CheXpert, whereas MoIE-CXR requires  $\sim 0.0065T$  flops. As MoIE-CXR discovers the sample-specific domain-invariant concepts, it achieves such high performance with low computational cost than BB.

## 4 Conclusion

This paper proposes a novel iterative interpretable method that identifies instance-specific concepts without losing the performance of the BB and is effectively fine-tuned in an unseen target domain with no concept annotation, limited labeled data, and minimal computation cost. Also, as in the prior work, MoIE-captured concepts may not showcase a causal effect that can be explored in the future.

## 5 Acknowledgement

This work was partially supported by NIH Award Number 1R01HL141813-01 and the Pennsylvania Department of Health. We are grateful for the computational resources provided by Pittsburgh Super Computing grant number TG-ASC170024.

## References

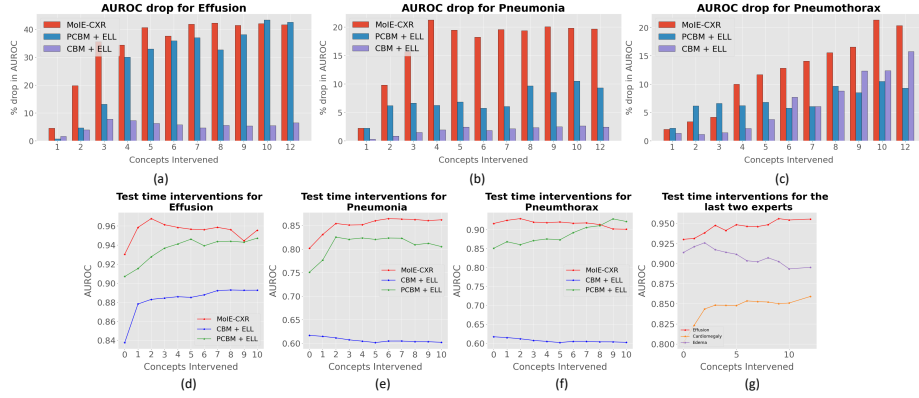
1. Barbiero, P., Ciravegna, G., Giannini, F., Lió, P., Gori, M., Melacci, S.: Entropy-based logic explanations of neural networks. In: *Proceedings of the AAAI Conference on Artificial Intelligence*. vol. 36, pp. 6046–6054 (2022)
2. Chu, B., Madhavan, V., Beijbom, O., Hoffman, J., Darrell, T.: Best practices for fine-tuning visual classifiers to new domains. In: *Computer Vision–ECCV 2016 Workshops: Amsterdam, The Netherlands, October 8–10 and 15–16, 2016, Proceedings, Part III* 14. pp. 435–442. Springer (2016)
3. Clough, J.R., Oksuz, I., Puyol-Antón, E., Ruijsink, B., King, A.P., Schnabel, J.A.: Global and local interpretability for cardiac mri classification. In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV* 22. pp. 656–664. Springer (2019)
4. Ghosh, S., Yu, K., Arabshahi, F., Batmanghelich, K.: Dividing and conquering a BlackBox to a mixture of interpretable models: Route, interpret, repeat. In: Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., Scarlett, J. (eds.) *Proceedings of the 40th International Conference on Machine Learning*. *Proceedings of Machine Learning Research*, vol. 202, pp. 11360–11397. PMLR (23–29 Jul 2023), <https://proceedings.mlr.press/v202/ghosh23c.html>
5. Ghosh, S., Yu, K., Arabshahi, F., Batmanghelich, K.: Tackling shortcut learning in deep neural networks: An iterative approach with interpretable models (2023)
6. Graziani, M., Andrearczyk, V., Marchand-Maillet, S., Müller, H.: Concept attribution: Explaining cnn decisions to physicians. *Computers in biology and medicine* **123**, 103865 (2020)
7. Guan, H., Liu, M.: Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering* **69**(3), 1173–1185 (2021)

8. Huang, G., Liu, Z., Van Der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 4700–4708 (2017)
9. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghighi, B., Ball, R., Shpanskaya, K., et al.: Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In: Proceedings of the AAAI conference on artificial intelligence. vol. 33, pp. 590–597 (2019)
10. Jain, S., Agrawal, A., Saporta, A., Truong, S.Q., Duong, D.N., Bui, T., Chambon, P., Zhang, Y., Lungren, M.P., Ng, A.Y., et al.: Radgraph: Extracting clinical entities and relations from radiology reports. arXiv preprint arXiv:2106.14463 (2021)
11. Johnson, A., Lungren, M., Peng, Y., Lu, Z., Mark, R., Berkowitz, S., Horng, S.: Mimic-cxr-jpg-chest radiographs with structured labels
12. Kandel, I., Castelli, M.: How deeply to fine-tune a convolutional neural network: a case study using a histopathology dataset. *Applied Sciences* **10**(10), 3359 (2020)
13. Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., Sayres, R.: Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). (2017). arXiv preprint arXiv:1711.11279 (2017)
14. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: International Conference on Machine Learning. pp. 5338–5348. PMLR (2020)
15. Lee, D.H., et al.: Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In: Workshop on challenges in representation learning, ICML. vol. 3, p. 896 (2013)
16. Rabanser, S., Thudi, A., Hamidieh, K., Dziedzic, A., Papernot, N.: Selective classification via neural network training dynamics. arXiv preprint arXiv:2205.13532 (2022)
17. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al.: Chexnet: Radiologist-level pneumonia detection on chest x-rays with deep learning. arXiv preprint arXiv:1711.05225 (2017)
18. Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., Zhong, C.: Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistic Surveys* **16**, 1–85 (2022)
19. Sarkar, A., Vijaykeerthy, D., Sarkar, A., Balasubramanian, V.N.: Inducing semantic grouping of latent concepts for explanations: An ante-hoc approach. arXiv preprint arXiv:2108.11761 (2021)
20. Wadden, D., Wennberg, U., Luan, Y., Hajishirzi, H.: Entity, relation, and event extraction with contextualized span representations. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 5784–5789. Association for Computational Linguistics, Hong Kong, China (Nov 2019). <https://doi.org/10.18653/v1/D19-1585>, <https://aclanthology.org/D19-1585>
21. Wang, Y.X., Ramanan, D., Hebert, M.: Growing a brain: Fine-tuning by increasing model capacity. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2471–2480 (2017)
22. Yan, W., Huang, L., Xia, L., Gu, S., Yan, F., Wang, Y., Tao, Q.: Mri manufacturer shift and adaptation: increasing the generalizability of deep learning segmentation for mr images acquired with different scanners. *Radiology: Artificial Intelligence* **2**(4), e190195 (2020)

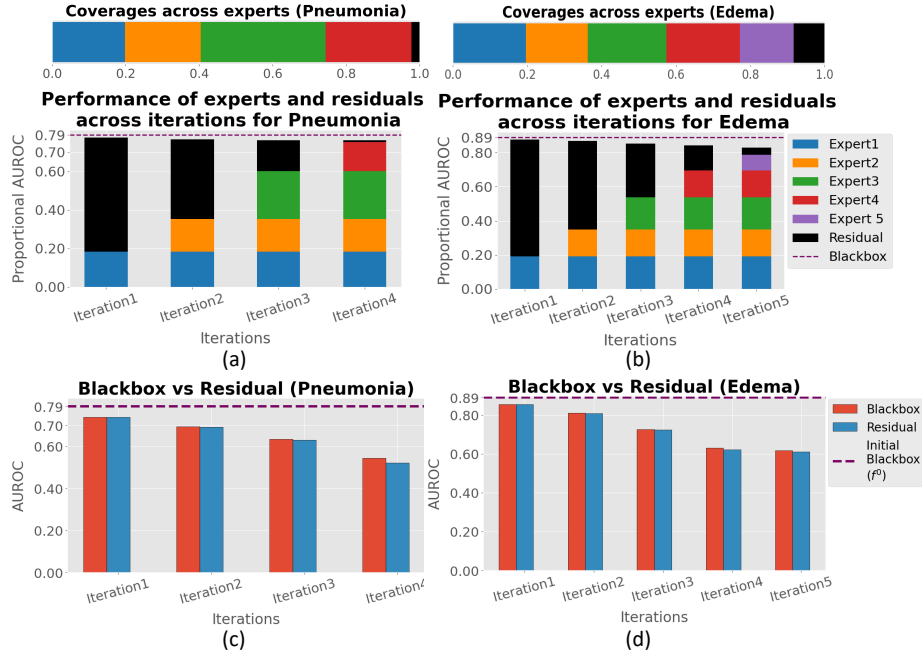
23. Yeche, H., Harrison, J., Berthier, T.: Ubs: A dimension-agnostic metric for concept vector interpretability applied to radiomics. In: Interpretability of Machine Intelligence in Medical Image Computing and Multimodal Learning for Clinical Decision Support: Second International Workshop, iMIMIC 2019, and 9th International Workshop, ML-CDS 2019, Held in Conjunction with MICCAI 2019, Shenzhen, China, October 17, 2019, Proceedings 9. pp. 12–20. Springer (2019)
24. Yu, K., Ghosh, S., Liu, Z., Deible, C., Batmanghelich, K.: Anatomy-guided weakly-supervised abnormality localization in chest x-rays. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part V. pp. 658–668. Springer (2022)
25. Yuksekgonul, M., Wang, M., Zou, J.: Post-hoc concept bottleneck models. arXiv preprint arXiv:2205.15480 (2022)
26. Zarlenga, M.E., Barbiero, P., Ciravegna, G., Marra, G., Giannini, F., Diligenti, M., Shams, Z., Precioso, F., Melacci, S., Weller, A., et al.: Concept embedding models. arXiv preprint arXiv:2209.09056 (2022)



Hyperparameter	Effusion	Cardiomegaly	Pneumothorax	Pneumonia	Edema
Batch size	1028	1028	1028	1028	1028
Learning rate	0.01	0.01	0.01	0.01	0.01
$\lambda_{lens}$	0.0001	0.0001	0.0001	0.0001	0.0001
$\alpha_{KD}$	0.99	0.99	0.99	0.99	0.99
$T_{KD}$	20	20	20	20	20
hidden neurons	30, 30	20, 20	20, 20	20, 20	20, 20
$\lambda_s$	96	1024	256	256	128
E-Lens ( $T_{lens}$ )	7.6	7.6	10	10	7.6
# Expers ( $T_{lens}$ )	5	4	5	4	5



**Fig. 2. (a-c):** Performance drop after zeroing out the concepts iteratively. The drop indicates the concepts to be more significant for prediction. **(d-g):** Test time interventions of concepts considering the ground truth concepts as an oracle on all samples (d-f), on the “hard” samples (g), covered by only the last two experts of MoIE-CXR.



**Fig. 3. (a-b):** The performances of experts and residuals across iterations for pneumonia and edema. **(c-d):** Performance comparison of the residuals and  $f^0$  for the samples covered by the successive residuals.