CROCODILE: Causality aids RObustness via COntrastive DIsentangled LEarning

Gianluca Carloni[1,2][0000−0002−5774−361X], Sotirios A. Tsaftaris[3][0000−0002−8795−9294], and Sara Colantonio[1][0000−0003−2022−0804]

[1] National Research Council, Pisa, Italy
gianluca.carloni@isti.cnr.it
[2] University of Pisa, Pisa, Italy
[3] The University of Edinburgh, Edinburgh, UK

Abstract. Deep learning image classifiers often struggle with domain shift, leading to significant performance degradation in real-world applications. In this paper, we introduce our CROCODILE framework, showing how tools from causality can foster a model's robustness to domain shift via feature disentanglement, contrastive learning losses, and the injection of prior knowledge. This way, the model relies less on spurious correlations, learns the mechanism bringing from images to prediction better, and outperforms baselines on out-of-distribution (OOD) data. We apply our method to multi-label lung disease classification from chest X-rays (CXRs), utilizing over 750000 images from four datasets. Our bias-mitigation method improves domain generalization, broadening the applicability and reliability of deep learning models for a safer medical image analysis. Find our code at: https://github.com/gianlucarloni/crocodile.

Keywords: Domain shift robustness · Out-of-distribution · Causality

1 Introduction

Domain shift bias is the problem of machine learning (ML) models performing not consistently across *in-distribution* (ID) and *out-of-distribution* (OOD) data. The former are independent and identically distributed (i.i.d) to the data on which the model was trained. Conversely, data are OOD when their distribution essentially differs from the source one, such as chest X-rays (CXR) coming from a different hospital than the training one [18,7,28]. Traditional ML models still tend to rely on spurious correlations seen during training for predicting the outcome and spectacularly fail when those shortcut associations are not present in OOD data, for instance, due to variations in scanner settings, image artifacts, or patient demographics [6,20,1,8]. For this reason, the field of domain generalization (DG) has searched for ways to make deep learning (DL) models learn robust features that could generalize better to unseen domains [11,13,25,29].

Conceptually, we could think of a set of features that causally determine the outcome and are invariant to shifts in non-relevant attributes, as well as a separate set of features that are spuriously correlated with the outcome but

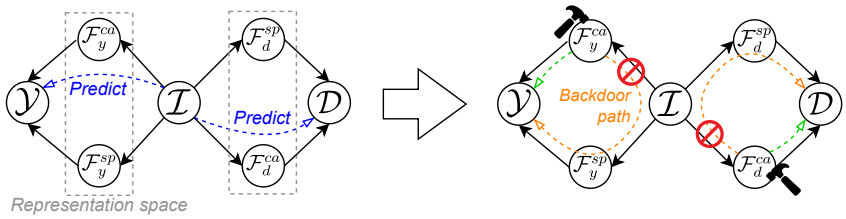

Fig. 1: A causal view on classifying medical images $\mathcal{I}$ coming from different domains $\mathcal{D}$ for the presence of diseases $\mathcal{Y}$. By applying the latent causal intervention (hammer), the backdoor path through the spurious features is cut off.

do not have a causal effect. Some works have proposed using tools from causal inference to achieve this disentanglement [26,21,12]. The common idea is that using the causal instead of the spurious features would allow a model to learn the underlying mechanism and be more robust on new data. However, these efforts try to model domain shifts implicitly, with a scope limited to the disease prediction task, disregarding the wealth of information on possible domain shifts from different source data sets.

In this work, we advance this causal/spurious feature disentanglement on a cross-domain level by leveraging information from different datasets in a contrastive learning setting. We conceive a domain-prediction branch along the disease-prediction branch to instill domain awareness into the model's representations. Moreover, we propose a new way to inject background medical knowledge, effectively designing a task prior to guiding learning and fostering DG.

2 Methodology

We define a structural causal model (SCM) [15] for medical image classification in Fig 1. Given the input images $\mathcal{I}$, such as CXRs, and the disease classification $\mathcal{Y}$, we obtain two sets of features via feature extraction. We denote $\mathcal{F}_y^{ca}$ the causal features that truly determine the outcome (e.g., the patchy airspace opacification typical in pneumonia). Similarly, we denote $\mathcal{F}_y^{sp}$ the spurious features, determined by data bias's confounding effect, which are unrelated to a disease (e.g., metal tokens on the image corners). Ideally, $\mathcal{Y}$ should be caused only by $\mathcal{F}_y^{ca}$, but is naturally confounded by $\mathcal{F}_y^{sp}$, as both types of features usually coexist in medical data. Unfortunately, conventional models tend to learn the correlation $P(\mathcal{Y}|\mathcal{F}_y^{ca})$ via the shortcut (backdoor) path $\mathcal{F}_y^{ca} \leftarrow \mathcal{I} \rightarrow \mathcal{F}_y^{sp} \rightarrow \mathcal{Y}$ instead of the desired $\mathcal{F}_y^{ca} \rightarrow \mathcal{Y}$. As we detail next, we exploit the *do-calculus* from causal theory [16] on the causal features to block the backdoor path, estimating $P(\mathcal{Y}|do(\mathcal{F}_y^{ca}))$. Following the same idea, we conceive two other sets of features extracted from $\mathcal{I}$, this time concerning the trivial task of predicting from which source domain come the data $\mathcal{D}$: $\mathcal{F}_d^{ca}$ would be the features that are relevant to distinguish different domains, and $\mathcal{F}_d^{sp}$ the confounding features.

2.1 Disease-branch and Domain-branch

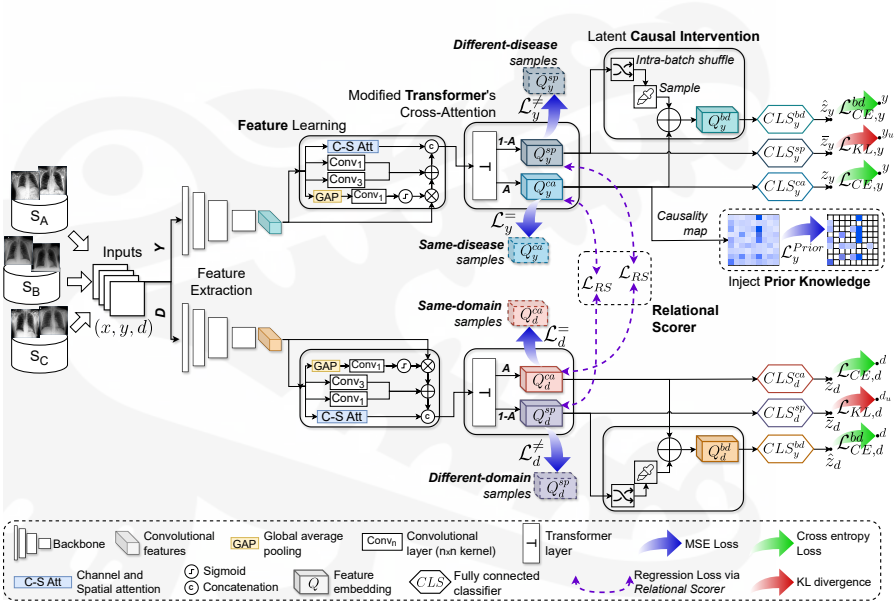

Fig. 2: CROCODILE involves two branches to learn robust, invariant features for predicting the labels from medical images (e.g., multi-label findings from CXRs) while disregarding confounding features. We disentangle *causal* features determining the label from *spurious* features associated with the label due to domain shift. We exploit images from multiple domains in a contrastive learning scheme and propose a new way to inject prior knowledge. Best seen in color.

We present our overall framework in Fig 2. A *disease prediction* branch learns to extract useful image features to predict the medical finding (e.g., pneumothorax or atelectasis in a CXR), regardless of the different domains. On another parallel branch for *domain prediction*, the image features that are useful for the trivial task of predicting the domain the images come from are learned (regardless of the different diseases). The architecture is trained end-to-end. Each branch involves a feature extraction backbone followed by a block to enhance features via channel- and spatial- attention [14]. Then, a Transformer network [24] yields the set $\mathbf{A}$ of attention scores, typically in the range 0-1, that identifies the portion of the input that is causally relevant to the task of interest (i.e., *what knowledge does the network use to make predictions*). Given an arbitrary set $\mathbf{A}$, we modify the Transformer's cross-attention mechanism to yield also the complementary set $\mathbf{1} - \mathbf{A}$ ($\mathbf{1}$ is the all-one matrix), representing the trivial/spurious aspects of the input. This way, we encode disentangled causal and spurious fea-

ture embeddings, Q^{ca} and Q^{sp}, by modulating the features by $\mathbf{A}$ and $\mathbf{1} - \mathbf{A}$, respectively. Finally, three classifiers connect the features Q to the classification logits z. In the following sections, we design specific contrastive learning losses and introduce a novel way to inject prior knowledge about the medical task.

2.2 Feature Disentanglement and Causal Intervention

For each branch, we need to make Q^{ca} and Q^{sp} capture the authentic and trivial aspects from the input samples. To achieve the correctness of the predictions, we impose two cross-entropy (CE) loss terms, $\mathcal{L}_{CE,y}$ and $\mathcal{L}_{CE,d}$, over the classification logits z_y and z_d from the causal features Q_y^{ca} and Q_d^{ca}, supervised by the disease labels y and domain labels d, respectively.

To make Q^{sp} features encode the trivial patterns that are unnecessary for classification, we push its predictions $\bar{z}_y$ and $\bar{z}_d$ evenly to all respective categories. We define the uniform classification losses $\mathcal{L}_{KL,y}$ and $\mathcal{L}_{KL,d}$ as the KL-divergence between the spurious features and the respective uniform distribution (y_u or d_u).

To alleviate the confounding effect, we implement the backdoor adjustment by performing a latent causal intervention [21,12]: we stratify the spurious features appearing from training data and pair the causal set of features with those stratified spurious features to compose the *intervened* graph. This way, we fit the concept of *borrowing from others* (i.e., "if everyone has it, it is as if no one has it"). We impose CE losses $\mathcal{L}_{CE,y}^{bd}$ and $\mathcal{L}_{CE,d}^{bd}$ between the logits $\hat{z}_y$ and $\hat{z}_d$ obtained from the corresponding intervened features Q^{bd} and the same ground-truth label for the causal features. This way, we push the predictions of such intervened images to be invariant and stable across different stratifications due to shared causal features. Practically, we approximate this operation with an intra-batch shuffling of Q^{sp} followed by random sampling (with 0.3 drop probability) and addition to Q^{ca}. By combining the supervised CE loss, the KL loss, and the backdoor CE loss for each branch, we obtain the two following equations:

$$\mathcal{L}_y = -(\lambda_1 \underbrace{y^\top \log(z_y)}_{\mathcal{L}_{CE,y}} + \lambda_2 \underbrace{KL(y_u, \bar{z}_y)}_{\mathcal{L}_{KL,y}} + \lambda_3 \underbrace{y^\top \log(\hat{z}_y))}_{\mathcal{L}_{CE,y}^{bd}} \tag{1}$$

$$\mathcal{L}_d = -(\lambda_4 \underbrace{d^\top \log(z_d)}_{\mathcal{L}_{CE,d}} + \lambda_5 \underbrace{KL(d_u, \bar{z}_d)}_{\mathcal{L}_{KL,d}} + \lambda_6 \underbrace{d^\top \log(\hat{z}_d))}_{\mathcal{L}_{CE,d}^{bd}} \tag{2}$$

2.3 Contrastive Learning

To attain cross-domain robustness, we posit there should also exist an alignment between the *causal* features that determine the *disease* and the *spurious* features for the *domain* prediction task. And the converse should also be true. For instance, we want the regions of the image that determine the presence of pneumonia to be unrelated to what contributes to discerning different domains (e.g., spurious metal tokens). Conversely, the image aspects determining which

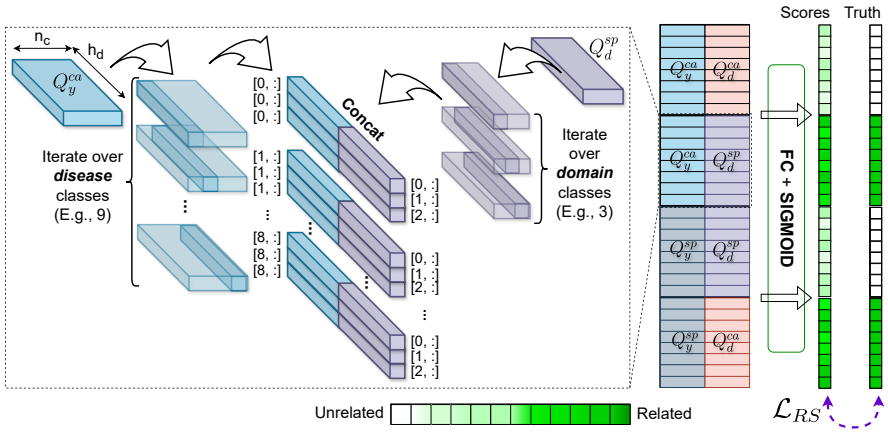

Fig. 3: Our *Relational Scorer* stratifies and concatenates every combination of *causal* and *spurious* features across both tasks. With a fully connected layer and a consecutive sigmoid$(\cdot)$, it maps each pair to a *relational score* between 0 and 1. We use an MSE loss regressing the relational scores to the ground truth. The model *learns to compare* the four sets of disentangled features. Best in color.

domain the image comes from should be unrelated to what determines disease prediction.

However, we are interested in measuring the *relational* alignment rather than the structural similarity of the representations. Matched (mismatched) pairs should "inform" ("repel") each other. Therefore, inspired by the concept of *learning to compare* [22,3], we design a new module named **Relational Scorer** (RS) to learn which image representations' pairings are semantically related and which are not (Fig 3). Our RS stratifies and combines each possible cross-branch pairing $p \in P = \{Q_y^{ca} \times Q_d^{ca} \cup Q_y^{ca} \times Q_d^{sp} \cup Q_y^{sp} \times Q_d^{sp} \cup Q_y^{sp} \times Q_d^{ca}\}$ and then maps them to a *relational score* between 0 and 1. We use an MSE loss regressing the relational scores r to the ground truths r^{GT}: matched pairs have a similarity of 1, and the mismatched pair have a similarity of 0.

Although this problem may seem to be a *classification* problem with label space $\{0, 1\}$, we are predicting relation scores, which can be considered a *regression* problem (with $r^{GT} \in \{0, 1\}$ generated by construction). We set the ground truth to 1 for the Q_y^{ca}-Q_d^{sp} and Q_y^{sp}-Q_d^{ca} pairings, and 0 otherwise. The resulting regression loss term is:

$$\mathcal{L}_{RS} = -\lambda_7 \sum_{i=1}^{|P|} (r_i - r_i^{GT})^2 \tag{3}$$

Moreover, we conceive other loss terms to enforce consistency/separation of medical image representations in a contrastive setting at a *batch* level:

- $\mathcal{L}_{\mathbf{y}}^{=}$: samples exhibiting a **common** radiological **finding** should lie close in *disease-causal* feature space Q_y^{ca}, regardless of the source domain.
- $\mathcal{L}_{\mathbf{y}}^{\neq}$: samples exhibiting **different** radiological **findings** should lie close in *disease-spurious* feature space Q_y^{sp}, regardless of the source domain.
- $\mathcal{L}_{\mathbf{d}}^{=}$: samples from the **same dataset** should lie close in *domain-causal* feature space Q_d^{ca}, regardless of the diseases.
- $\mathcal{L}_{\mathbf{d}}^{\neq}$: samples from **different datasets** should lie close in *domain-spurious* feature space Q_d^{sp}, regardless of the diseases.

We implement each of such terms via an MSE loss between the representation Q of each sample in the batch and the corresponding average representation $\tilde{Q}$ of samples with the same/different label:

$$\mathcal{L}_y^{batch} = -(\lambda_8 \underbrace{\sum_{y \in \mathcal{Y}} (Q_y^{ca} - \tilde{Q}_y^{ca})^2}_{\mathcal{L}_y^{=}} + \lambda_9 \underbrace{\sum_{y \in \mathcal{Y}} (Q_y^{sp} - \tilde{Q}_{not(y)}^{sp})^2)}_{\mathcal{L}_y^{\neq}} \tag{4}$$

$$\mathcal{L}_d^{batch} = -(\lambda_{10} \underbrace{\sum_{d \in \mathcal{D}} (Q_d^{ca} - \tilde{Q}_d^{ca})^2}_{\mathcal{L}_d^{=}} + \lambda_{11} \underbrace{\sum_{d \in \mathcal{D}} (Q_d^{sp} - \tilde{Q}_{not(d)}^{sp})^2)}_{\mathcal{L}_d^{\neq}} \tag{5}$$

where $\mathcal{Y}$ and $\mathcal{D}$ are the possible disease and domain labels seen in the batch. To compute those losses correctly, we design a custom sampler favoring consistent batches where the class prevalence is respected.

2.4 Injecting Prior Knowledge

Motivated by the high interclass similarity and hierarchical structure of CXR findings [19,27], we propose a new method to inject prior (medical) knowledge into the model to guide its learning (Fig. 4). Differently from solutions as *conditional training* [17], which rely on data, our proposal is desirable to capture semantic priors without relying on data. We define a causal graph representing the relationship between the CXR findings and propose a novel formulation of the *causality map* concept [5,4] to model the co-occurrence of CXR findings in the images. As we have seen, each Q_y^{ca} representation has shape $n_c \times h$, where n_c is the number of classes (e.g., nine CXR findings) and h is the hidden dimension of the embeddings. After normalizing Q_y^{ca} by their global maximum batch-wise, they lie in the range 0-1, and we interpret their values as probabilities of the CXR findings to be present in the image. Indeed, given two embeddings Q^i and Q^j, to compute the effect of the former on the presence of the latter, we estimate the ratio between their joint and marginal probabilities as:

$$P(Q^i|Q^j) = \frac{P(Q^i, Q^j)}{P(Q^j)} \approx \frac{(\max_h Q_h^i) \cdot (\max_h Q_h^j)}{\sum_h Q_h^j}, \forall i, j \in 1 \leq i, j \leq n_c \tag{6}$$

thus obtaining the relationships between embeddings Q^i and Q^j, since, in general, $P(Q^i|Q^j) \neq P(Q^j|Q^i)$. By computing these quantities for every pair i, j, we

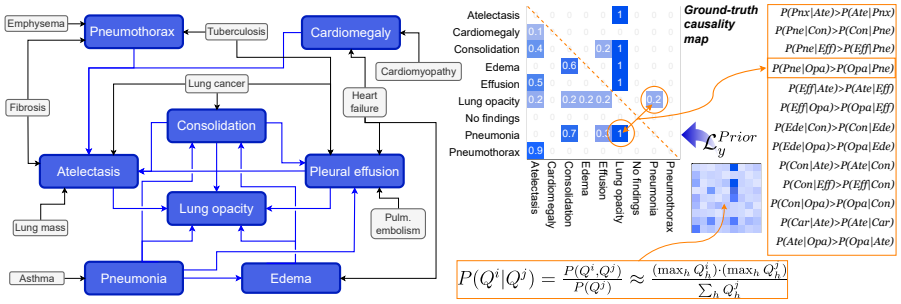

Fig. 4: Causal graphical model among the CXR findings of interest (blue) and the ground-truth *causality map* defined from that graph. Gray boxes represent additional findings or risk factors (not investigated in this study) that might be associated with the desired ones.

obtain the $n_c \times n_c$ map C_y. We interpret asymmetries across estimates opposite the main diagonal in C_y as causality signals between their activation. Accordingly, the representation of a CXR finding causes the activation of another when $P(Q^i|Q^j) > P(Q^j|Q^i)$, that is $Q^i \rightarrow Q^j$. We design our **Task-Prior loss** as an MSE loss to push the causality map C_y obtained from the learned representations to the ground-truth causality map C_y^{GT}, which we defined by estimating frequencies based on medical knowledge about the possible co-occurrence of CXR findings:

$$\mathcal{L}_y^{Prior} = -\lambda_{12}(C_y - C_y^{GT})^2 \tag{7}$$

Overall, the training objective of our CROCODILE framework is defined as the sum of the losses defined in Equations 1, 2, 3, 4, 5 and 7:

$$\mathcal{L}_{TOT} = \mathcal{L}_y + \mathcal{L}_d + \mathcal{L}_{RS} + \mathcal{L}_y^{batch} + \mathcal{L}_d^{batch} + \mathcal{L}_y^{prior}. \tag{8}$$

3 Experimental Setup

We classify eight radiological findings (plus the *No finding* class) from frontal CXR images of four popular data sets in both ID and OOD settings. After cleaning, the number of images for each set is: 112110 for ChestX-ray14 [27], 183453 for CheXpert [9], 95452 for PadChest [2], and 365737 for MIMIC-CXR [10]. For the first dataset, we create the *Lung opacity* class as OR logic across the *consolidation*, *effusion*, *edema*, *pneumonia*, and *atelectasis* classes. We resize the images to 320×320 and adjust their contrast in the range 0-255. For ID experiments, we combine images of ChestX-ray14, CheXpert, and PadChest, split them into 80-20% train and validation sets, and assess the multi-label classification performance via the area under the ROC curve (AUC) and the average precision (AP) scores for each category and their average. We test the best-performing ID model on the external, never-before-seen MIMIC-CXR dataset to evaluate

OOD generalization abilities. In all the experiments, we adopted ResNet50 backbones, Adam optimizer, a learning rate of 1e-6, a batch size of 12, and trained the model in early-stopping on a multi-node multi-gpu cluster with NVIDIA 64 GB cards. For the lambda hyperparameters, we tried out different values that would counterbalance the average values of the losses with unit weights. We thus conducted a random grid search and chose the following: set $\lambda_1, \lambda_3, \lambda_4, \lambda_6$ to 1; λ_2 to 10; λ_9, λ_{11} to 15; λ_8, λ_{10} to 25; λ_5 to 80; and λ_7, λ_{12} to 100. We compare to a regular ResNet50 architecture, a ResNet50 version of Nie *et al.* [12] corresponding to discarding domain-branch and task-prior information from our method, our method without contrastive learning (CL) ($\mathcal{L}_{RS}$, $\mathcal{L}_y^{batch}$, $\mathcal{L}_d^{batch}$), and our method without the task prior ($\mathcal{L}_y^{prior}$).

4 Results and Discussion

The results of our ID and OOD investigations (Table 1) reveal our method is behind its ablated versions and [12] on i.i.d. data (ID) while is the best-performing model on the external never-before-seen data (OOD). Notably, our method is the most effective in reducing the ID-to-OOD drop in performance. This significant result points to a necessary trade-off between in-domain accuracy and out-of-domain robustness on real-world data, supporting recent work [23]. As expected, models not contrasting the information from the two branches ([12] and ours without CL) find associations that make them perform better on the ID data, where they remain faithful. Then, however, they fail to perform as well on OOD data, where many spurious correlations due to the domain no longer exist, suggesting those associations are still based mainly on shortcut features. On the contrary, adopting our contrastive learning scheme first leads to lower performance on ID data (as if the representation power on such data were 'spoiled' compared to the above). Still, it leads to better results on OOD data. This suggests that our method learns image-to-prediction mechanics that are more transportable and generalizable, relying less on confounding factors and breaking down barriers between domains.

Moreover, injecting prior task knowledge helped the model with specific findings. For instance, we know *effusion* is likely an effect of *pneumonia* or *consolidation* and one of five aspects defining *lung opacity*. We also know that patients with heart failure typically feature both *cardiomegaly* and *effusion*, but there is no causal effect of one aspect onto the other (Fig 4). Thus, when the model was equipped with this knowledge during training, it learned to pay attention to such co-occurrences more and ultimately could detect more *effusion* cases in OOD data, possibly disregarding the confounding effect of heart failure.

Among the limitations of this work, we have utilized the same architecture type for feature extraction on the two branches, and we implicitly optimized the network on ID validation data. We will improve by trying different backbones and allowing an ID test set.

Finding	ResNet50	Nie *et al.* [12]	**Ours** *w/o* CL	**Ours** *w/o* TP	**Ours**
In-distribution (ID) data					
Atelectasis	65.74/24.98	76.81/30.04	**77.13**/30.26	77.07/30.37	77.04/**30.37**
Cardiomegaly	81.53/51.21	92.43/56.56	**92.92**/**56.60**	92.29/56.20	92.27/56.17
Consolidation	69.74/8.71	80.89/13.85	80.62/**14.10**	**81.13**/13.82	81.10/13.86
Edema	77.34/17.62	88.49/**23.01**	88.21/22.53	**88.73**/22.02	88.72/22.05
Effusion	77.69/51.26	88.68/56.31	**89.08**/56.46	88.92/**56.65**	88.93/56.65
Lung opacity	69.81/39.27	81.20/44.62	**81.20**/**44.66**	80.60/44.10	80.55/44.08
No finding	68.75/68.08	**80.14**/73.46	79.68/**73.47**	79.38/73.22	79.35/73.22
Pneumonia	67.76/20.74	78.05/**26.13**	**79.15**/25.73	77.65/24.86	77.63/24.85
Pneumothorax	78.86/32.78	89.87/**38.17**	**90.25**/37.69	88.79/37.02	89.86/37.03
Mean [↑]	73.02/34.96	84.06/**40.24**	**84.25**/40.17	83.95/39.81	83.94/39.81
Out-of-distribution (OOD) data					
Atelectasis	62.79/31.56	74.02/36.69	74.11/36.63	74.15/**36.89**	**74.18**/36.83
Cardiomegaly	61.43/31.84	71.44/36.22	71.86/36.42	**72.82**/37.16	72.80/**37.17**
Consolidation	66.41/7.20	77.01/11.97	77.38/**12.53**	77.46/12.13	**77.80**/12.07
Edema	74.04/36.12	84.52/40.48	83.95/40.46	**85.43**/41.43	85.39/41.45
Effusion	75.10/59.66	86.16/64.60	86.04/64.87	86.01/64.85	**86.49**/**64.99**
Lung opacity	56.92/28.52	67.86/33.49	67.43/33.10	68.30/33.83	**68.31**/**33.85**
No finding	67.39/63.72	78.53/68.66	78.72/68.99	**78.78**/69.02	78.74/**69.05**
Pneumonia	53.64/7.47	63.96/12.29	64.62/12.52	65.01/12.76	**65.03**/**12.80**
Pneumothorax	64.72/12.39	74.89/16.76	75.41/17.65	75.48/17.70	**76.11**/**17.72**
Mean [↑]	64.71/30.94	75.38/35.68	75.50/35.91	75.94/36.20	**76.09**/**36.21**
ID-OOD drop	11.38/11.50	10.33/11.33	10.38/10.60	9.54/9.07	**9.35**/**9.04**

Table 1: AUC and AP scores obtained on each CXR finding on ID and OOD data. CL: contrastive learning, TP: task prior. ID-OOD drop is the average percent drop in scores from ID to OOD settings.

5 Conclusion

We have presented the CROCODILE framework, a new approach to enhance a medical image classifier's generalization and OOD robustness, addressing the problem of removing confounders. Our solution learns what to focus on/suppress by borrowing from multiple sub-disciplines: latent causal intervention, graphical models, causality maps, feature disentanglement, the *learning to compare* idea and enforcing representation consistency. Our bias-mitigation proposal is general and can be applied to tackle domain shift bias in other computer-aided diagnosis applications, fostering a safer and more generalizable medical AI.

Acknowledgments. This study has received funding from the European Union's Horizon 2020 program under grant No 952159 (ProCAncer-I) and the Tuscany Project PAR-FAS PRAMA. The funders had no role in the design of the study, the collection, analysis, and interpretation of data, or writing the manuscript. We acknowledge the CINECA award under the ISCRA initiative for the availability of high-performance computing resources and support.

Disclosure of Interests. The authors have no competing interests to declare.

References

1. Bercean, B., Buburuzan, A., Birhala, A., Avramescu, C., Tenescu, A., Marcu, M.: Breaking down covariate shift on pneumothorax chest x-ray classification. In: International Workshop on Uncertainty for Safe Utilization of Machine Learning in Medical Imaging. pp. 157–166. Springer (2023)
2. Bustos, A., Pertusa, A., Salinas, J.M., De La Iglesia-Vaya, M.: Padchest: A large chest x-ray image dataset with multi-label annotated reports. Medical image analysis **66**, 101797 (2020)
3. Cao, C., Zhang, Y.: Learning to compare relation: Semantic alignment for few-shot learning. IEEE Transactions on Image Processing **31**, 1462–1474 (2022)
4. Carloni, G., Colantonio, S.: Exploiting causality signals in medical images: A pilot study with empirical results. Expert Systems with Applications p. 123433 (2024)
5. Carloni, G., Pachetti, E., Colantonio, S.: Causality-driven one-shot learning for prostate cancer grading from mri. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2616–2624 (2023)
6. Castro, D.C., Walker, I., Glocker, B.: Causality matters in medical imaging. Nature Communications **11**(1), 3673 (2020)
7. Cohen, J.P., Hashir, M., Brooks, R., Bertrand, H.: On the limits of cross-domain generalization in automated x-ray prediction. In: Medical Imaging with Deep Learning. pp. 136–155. PMLR (2020)
8. Hartley, J., Sanchez, P.P., Haider, F., Tsaftaris, S.A.: Neural networks memorise personal information from one sample. Scientific Reports **13**(1), 21366 (2023)
9. Irvin, J., Rajpurkar, P., Ko, M., Yu, Y., Ciurea-Ilcus, S., Chute, C., Marklund, H., Haghgoo, 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. Johnson, A.E., Pollard, T.J., Greenbaum, N.R., Lungren, M.P., Deng, C.y., Peng, Y., Lu, Z., Mark, R.G., Berkowitz, S.J., Horng, S.: Mimic-cxr-jpg, a large publicly available database of labeled chest radiographs. arXiv preprint arXiv:1901.07042 (2019)
11. Li, Z., Cui, Z., Wang, S., Qi, Y., Ouyang, X., Chen, Q., Yang, Y., Xue, Z., Shen, D., Cheng, J.Z.: Domain generalization for mammography detection via multi-style and multi-view contrastive learning. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VII 24. pp. 98–108. Springer (2021)
12. Nie, W., Zhang, C., Song, D., Bai, Y., Xie, K., Liu, A.A.: Chest x-ray image classification: A causal perspective. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 25–35. Springer (2023)
13. Ouyang, C., Chen, C., Li, S., Li, Z., Qin, C., Bai, W., Rueckert, D.: Causality-inspired single-source domain generalization for medical image segmentation. IEEE Transactions on Medical Imaging **42**(4), 1095–1106 (2022)
14. Pan, X., Ge, C., Lu, R., Song, S., Chen, G., Huang, Z., Huang, G.: On the integration of self-attention and convolution. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 815–825 (2022)
15. Pearl, J.: Causal inference. Causality: objectives and assessment pp. 39–58 (2010)
16. Pearl, J.: Interpretation and identification of causal mediation. Psychological methods **19**(4), 459 (2014)

17. Pham, H.H., Le, T.T., Tran, D.Q., Ngo, D.T., Nguyen, H.Q.: Interpreting chest x-rays via cnns that exploit hierarchical disease dependencies and uncertainty labels. Neurocomputing **437**, 186–194 (2021)
18. Pooch, E.H., Ballester, P., Barros, R.C.: Can we trust deep learning based diagnosis? the impact of domain shift in chest radiograph classification. In: Thoracic Image Analysis: Second International Workshop, TIA 2020, Held in Conjunction with MICCAI 2020, Lima, Peru, October 8, 2020, Proceedings 2. pp. 74–83. Springer (2020)
19. Rajaraman, S., Antani, S.: Training deep learning algorithms with weakly labeled pneumonia chest x-ray data for covid-19 detection. MedRxiv (2020)
20. Sanchez, P., Voisey, J.P., Xia, T., Watson, H.I., O'Neil, A.Q., Tsaftaris, S.A.: Causal machine learning for healthcare and precision medicine. Royal Society Open Science **9**(8), 220638 (2022)
21. Sui, Y., Wang, X., Wu, J., Lin, M., He, X., Chua, T.S.: Causal attention for interpretable and generalizable graph classification. In: Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. pp. 1696–1705 (2022)
22. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P.H., Hospedales, T.M.: Learning to compare: Relation network for few-shot learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1199–1208 (2018)
23. Teney, D., Lin, Y., Oh, S.J., Abbasnejad, E.: Id and ood performance are sometimes inversely correlated on real-world datasets. Advances in Neural Information Processing Systems **36** (2024)
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
25. Wang, J., Lan, C., Liu, C., Ouyang, Y., Qin, T., Lu, W., Chen, Y., Zeng, W., Philip, S.Y.: Generalizing to unseen domains: A survey on domain generalization. IEEE transactions on knowledge and data engineering **35**(8), 8052–8072 (2022)
26. Wang, T., Zhou, C., Sun, Q., Zhang, H.: Causal attention for unbiased visual recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3091–3100 (2021)
27. Wang, X., Peng, Y., Lu, L., Lu, Z., Bagheri, M., Summers, R.M.: Chestx-ray8: Hospital-scale chest x-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2097–2106 (2017)
28. Zhang, J., Xue, P., Gu, R., Gu, Y., Liu, M., Pan, Y., Cui, Z., Huang, J., Ma, L., Shen, D.: Learning towards synchronous network memorizability and generalizability for continual segmentation across multiple sites. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 380–390. Springer (2022)
29. Zunaed, M., Haque, M.A., Hasan, T.: Learning to generalize towards unseen domains via a content-aware style invariant model for disease detection from chest x-rays. IEEE Journal of Biomedical and Health Informatics (2024)