GENERALIZED CROSS-DOMAIN MULTI-LABEL FEW-SHOT LEARNING FOR CHEST X-RAYS

Aroof Aimen [⋆†¶] *Arsh Verma* [†] *Makarand Tapaswi*[†‡] *Narayanan C Krishnan*[#]

[⋆]University of Wisconsin - Madison, [†]Wadhwani AI, [‡]IIIT Hyderabad,
[#]Indian Institute of Technology - Palakkad, [¶]Indian Institute of Technology - Ropar

ABSTRACT

Chest X-ray (CXR) abnormality classification faces several challenges: (i) limited training data; (ii) training and evaluation sets that are derived from different domains; and (iii) classes that appear during training may have partial overlap with classes of interest during evaluation. We propose an integrated framework called *Generalized Cross-Domain Multi-Label Few-Shot Learning* (GenCDML-FSL), which supports class overlap during training and evaluation, cross-domain transfer, and few-shot learning for multi-label CXR image classification. Additionally, we introduce *Generalized Episodic Training* (GenET), a strategy that trains models to handle the challenges observed in GenCDML-FSL scenario. Our approach outperforms transfer learning, hybrid transfer learning, and multi-label meta-learning on multiple datasets.

***Index Terms*—** Meta-learning, Cross-domain, Few-shot.

1. INTRODUCTION

Deep neural networks (DNNs), show promise in automating chest X-ray interpretation. However, they require vast amounts of labeled data for effective training, and the time-consuming process of labeling X-rays highlights the scarcity of such data in healthcare. Few-shot learning (FSL) has emerged as a sub-field of machine learning [1] aimed at training DNNs with minimal data while maintaining generalization to new images. Meta-learning (MetaL) is effective for FSL but faces challenges in chest X-ray classification, such as overlapping labels and distributional differences between train and test datasets. While cross-domain few-shot learning (CDFSL) methods [2] address distributional issues, they assume unique labels per image and disjoint label sets between train and test data, making them less suitable for chest X-rays with multiple and overlapping abnormalities.

We propose Generalized Cross-Domain Multi-Label Few-Shot Learning (GenCDML-FSL), a problem setup where: (i) Generalized indicates partial overlap between train and test labels, inducing model bias towards the overlapping classes [3]; (ii) Cross-domain refers to domain differences between training and evaluation data [2]; (iii) Multi-label accounts for X-ray images often showing multiple abnormalities; and (iv) Few-shot learning deals with the challenge of fine-tuning on limited data. Additionally, we propose Generalized Episodic Training (GenET), a training pipeline to handle these challenges. GenET uses support, fine-tune, and query sets for training, fine-tuning, and evaluation, with possible class overlap between the sets to simulate real evaluation conditions. This helps the model adapt to both new and overlapping classes. Cross-domain differences are introduced through varied augmentations applied to each set, ensuring the model learns to adjust to domains for effective generalized learning. Further, multi-label classification is enabled by relaxing number of shots per class and episodic training improves generalizability even with limited data.

2. GenCDML-FSL FORMULATION

Preliminaries. In an FSL setup, the training, validation, and test datasets are denoted as $Train$, Val, and $Test$, with classes C_{Train}, C_{Val}, and C_{Test} from domains D_{Train}, D_{Val}, and D_{Test}. $Train$ is for model training, Val for hyperparameter tuning, and $Test$ for evaluating the model. In MetaL context, tasks (episodes) denoted as T involve randomly sampling classes from C_{Train}, C_{Val}, or C_{Test} based on a task distribution $P(\mathcal{T})$. Each task (episode) T_i is an N-way K-shot learning challenge, with N being the number of classes and K the instances per class. Each task includes a support set S_i with samples from classes $C^{S_i} \in C_{Train}$ and a query set Q_i with different samples from the same classes, i.e., $C^{Q_i} = C^{S_i}$.

2.1. GenCDML-FSL Framework

In a typical meta-learning paradigm, the train (C_{Train}), and test classes (C_{Test}) are mutually exclusive. However, in chest X-rays datasets, labels from the training set such as the NIH dataset [4] (*e.g*. *Cardiomegaly*, *Atelectasis*) may be present in test datasets like CheXpert [5]. We refer to the setup that allows overlap between classes as *Generalized*-FSL (G-FSL). Furthermore, domain disparity, *e.g*. arising from the country of data collection, between the train (*e.g*. NIH) and test (*e.g*. CheXpert) set is characterized as Cross-domain FSL (CD-FSL). Different from multi-class FSL, our formulation

allows images to be associated with multiple labels, introducing the Multi-Label FSL (ML-FSL) paradigm. All together, GenCDML-FSL integrates G-FSL, CD-FSL, and ML-FSL.

Definition 1 *We define GenCDML-FSL as a setup where: (i) train and validation have the same classes $C_{Train} = C_{Val}$; (ii) train and test may have some overlapping labels $C_{Train} \cap C_{Test} \neq \varnothing$ and (iii) domains are assumed different: $D_{Train} \neq D_{Val} \neq D_{Test}$. The paired samples in the train set are $Train = \{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^{|Train|}$. Each chest X-ray image $\mathbf{x}_i$ may be associated with one or more labels: $\mathbf{y}_i = \{y_i^c\}_{c=1}^{|C_{Train}|}$, where $y_i^c \in \{0, 1\}$ indicates absence or presence of the label c in $\mathbf{x}_i$. $|C_{Train}|$ denotes the number of classes and $|Train|$ is the number of samples in the train set. Similar multi-label definition can be adopted for the validation and test sets.*

Multi-label episodic training. To support multi-label training, we relax the constraint of exactly K shots per class while keeping N fixed. For each task, N classes (C^S) are randomly selected from the available pool (*e.g.* C_{Train}). Each selected class has *at least* K samples, with the total ranging from K to $N \times K$, based on label occurrences in other samples. This framework supports a multi-label setup, allowing an image to be selected for different classes across tasks.

3. GENERALIZED EPISODIC TRAINING (GenET)

To adapt episodic training to a generalized label space with partially overlapping classes and to address train-test domain disparities, we introduce a novel pipeline: Generalized Episodic Training (GenET). The GenET procedure organizes each training episode (or task), denoted as T_i, into three distinct sets: (i) support set $S_i = \{(\mathbf{x}_k, \mathbf{y}_k^s)_{k=1}^K\}_{s=1}^N$ to train the model; (ii) a new *fine-tune set* $F_i = \{(\mathbf{x}_p, \mathbf{y}_p^f)_{p=1}^P\}_{f=1}^N)$ for fine-tuning the model; and (iii) query set $Q_i = \{(\mathbf{x}_r, \mathbf{y}_r^q)_{r=1}^R\}_{q=1}^N)$ for evaluating the model's performance. We use $\mathbf{y}_k^s$ to denote the multi-label vector $\mathbf{y}_k$ where category s is present, *i.e.* $y_k^s = 1$, and others may or may not be 1. Similar nomenclature applies to $\mathbf{y}_p^f$ and $\mathbf{y}_r^q$. N represents classes in a task, and K, P, and R represent the minimum number of samples belonging to each class in the support, fine-tune, and query sets, respectively. During GenET, for a given task, classes within the support and query set may be partly overlapping (*i.e.* $C^{S_i} \stackrel{?}{=} C^{F_i}$ –may or may not be true). However, the fine-tune and query sets have identical labels (*i.e.* $C^{F_i} = C^{Q_i}$), but with distinct images.

Learning procedure: Handling overlapping and non-overlapping labels. The model parameters θ^t at iteration t are adapted U times on the support set S_i through standard gradient descent on the support loss L^s, with a learning rate α. We denote the adapted model as

$$\phi_i^U \leftarrow \theta^t - \alpha \nabla_{\theta^t} L^s(S_i; \theta^t)\,. \tag{1}$$

The adapted model ϕ_i^U is fine-tuned V times on F_i using a learning rate β and fine-tune loss L^f:

$$\psi_i^V \leftarrow \phi_i^U - \beta \nabla_{\phi_i^U} L^f(F_i; \phi_i^U)\,. \tag{2}$$

The fine-tuned model ψ_i^V is subsequently evaluated on the query set Q_i to obtain the query loss L^q. The query loss, together with a learning rate γ for all episodes in a batch of size B, is utilized to update the meta-model θ:

$$\theta^{t+1} \leftarrow \theta^t - \gamma \nabla_{\theta^t} \sum_{i=1}^{B} L^q(Q_i; \psi_i^V)\,. \tag{3}$$

To support the multi-label classification paradigm, we use binary cross-entropy loss over all samples and classes:

$$L = -\sum_k \sum_{j=1}^{C} y_k^j \log(\hat{y}_k^j) + (1 - y_k^j) \log(1 - \hat{y}_k^j)\,, \tag{4}$$

where, L stands for L^s, L^f, L^q while C corresponds to C^S, C^F, C^Q, the classes in the support, fine-tune and query set for the specific task. y_k^j denotes the true label for class j of sample $\mathbf{x}_k$ and $\hat{y}_k^j$ is the label probability predicted by the model. Adapting model on the support set S_i and fine-tuning it on the fine-tune set F_i trains the model for varying sets of classes. Evaluating and fine-tuning the meta-model on the query set Q_i makes the meta-model learn how to improve predictions on both overlapping and non-overlapping classes.

Handling domain shift. To address domain shift between train and test data, we simulate shifts during training with augmentations like Horizontal Flip, Vertical Flip, Random Resized Crop, *etc*. (ref Sec. 4.1). For task T_i, the augmentations for S_i, F_i, and Q_i may differ. Minimizing loss on Q_i (per Eq. 3) helps the model generalize across various augmentations, enhancing robustness in both in-domain and cross-domain scenarios. Similar augmentations are applied to baselines for a fair comparison.

4. EXPERIMENTS

We conduct experiments on the GenCDML-FSL paradigm using four popular chest X-ray datasets: NIH [4], PadChest [6], CheXpert [5], and MIMIC [7]. We selected NIH and MIMIC as source datasets because they represent extremes in sample size, with NIH having the lowest and MIMIC the highest. We perform training on one dataset followed by fine-tuning or adaptation using few samples (N=240) of the test dataset in all experiments. We train on NIH (or MIMIC) and evaluate on PadChest, CheXpert, and MIMIC (or NIH). We augment labels with a *Normal* category to indicate the absence of all abnormalities. GenET is compared against multiple baselines: (i) standard transfer learning (TL), (ii) heterogenous transfer learning (HTL) [8], (iii) multi-label MAML (MMAML) [9], and (iv) state-of-the-art multi-label meta-learning algorithm (ML-MetaL) [10].

Methods	Source: NIH			Source: MIMIC		
	CX	PC	MIMIC	CX	PC	NIH
TL	0.2956	0.1871	0.2867	0.3658	0.1895	0.2040
HTL	0.5616	0.5312	0.5331	**0.6443**	0.5108	0.5615
ML-metaL	0.4278	0.3438	0.4324	0.4353	0.3648	0.3898
MMAML	0.5332	0.5104	0.5175	0.6072	0.4871	0.4849
GenET	**0.5773**	**0.5340**	**0.5354**	0.5687	**0.5366**	**0.6985**

Table 1. Comparing all baselines against GenET using **mean Average Precision (mAP)**. CX: CheXpert, PC: PadChest. GenET achieves highest score in 5 of 6 cases.

4.1. Implementation Details

Learning rates (LR). We set the learning rate (LR) to 10^{-4} for TL based on optimal validation performance. For HTL, we use an LR of 10^{-4} for training and adaptation on the meta-test fine-tune set, following [11]. For MMAML and GenET, the support LR is 0.01 and the query LR is 0.001. All LRs were chosen via grid search in the range $[10^{-6}, 10^{-2}]$ and remain fixed across all methods and datasets. The support and fine-tune LRs in meta-train and meta-test sets are same in GenET.

Episodic training details. We use ResNet50 for all experiments, resizing images to 128×128. For GenET, MMAML, and ML-MetaL, the support adaptation steps are set to $U=5$, with $V=2$ fine-tune steps for GenET. Adaptation steps remain consistent during meta-training and meta-testing across all episodic experiments, with a 0.3 overlap between support and query/fine-tune classes.

Batch size, Episode size, epochs. We use a batch size of 24 for non-episodic training and 1 for episodic training to reduce computational burden [12]. The episodic batch corresponds to multiple tasks and cannot be directly compared to the non-episodic batch, which consists of individual samples. We set the number of epochs for non-episodic training to 40 owing to convergence. To ensure fairness in data exposure, the number of episodic epochs is calculated as $40 \times \frac{\text{Total samples}}{B\times\text{Episode Size}}$, where $B=1$ is the episodic batch size. The episode size is $N \times (K+P+R)$, with $N=4$ (classes per task), $K=1$, $P=2$, and $R=10$ (shots for support, fine-tune, and query sets, respectively). During evaluation, $P=1$.

Augmentations. We use Horizontal Flip (p: 0.5), Vertical Flip (p: 0.2), Random Resized Crop (p: 0.5, 128×128), Crop and Pad (p: 0.8, percent: [-0.3, 0.3]), and Rotation (p: 0.5) from the Albumentations library. Variations in strength and probability of selection (p) introduce stochasticity for each sample. To ensure fairness, similar augmentations are applied across all methods (TL, HTL, ML-metaL, MMAML, GenET).

Meta-test fine-tune split. To ensure fairness, we use 240 annotated samples from the test set for fine-tuning, meta-testing, or transfer learning for all methods. The fine-tune split is designed to maximize label representation in the test set. From 100 random splits, we select the split that minimizes label distribution distance for the fine-tune and test set.

4.2. GenET *vs*. Baselines

mAP evaluation. Table 1 shows GenET outperforms all baselines on mAP with NIH dataset as the source. We observe small, but consistent improvements of 2-4% on all three datasets. Interestingly, transfer from MIMIC is more imbalanced, and GenET performs best in 2 of 3 cases. From MIMIC to CX, the high score for HTL may be due to both datasets having same label space.

Evaluation at oracle threshold is presented in Table 2, that reports standard classification metrics and F1 scores. Here, GenET outperforms 5/6 baselines on F1 score when using oracle threshold, and is a close second on MIMIC to CheXpert.

Evaluation at 0.5 threshold. Analyzing the F1 score, GenET outperforms all baselines with the MIMIC dataset as source, and closely follows MMAML when source dataset is NIH.

Impact of overlapping classes. Next, we analyze GenET's performance across overlapping and non-overlapping classes separately. Fig. 1 shows that GenET outperforms baselines in 6 of 10 instances each with overlapping labels and non-overlapping labels. In particular, MIMIC to NIH shows large improvements with GenET both on overlapping and non-overlapping labels. Also, non-overlapping classes have performance comparable to overlapping classes for GenET. This suggests that GenET enables the model to learn representations that are invariant to classes, validating our hypothesis that incorporating testing conditions into the training process through the fine-tune set enhances generalization.

Comparison with SotA ML-metaL. Results presented in Table 1 and Table 2 show that GenET outperforms ML-metaL on all datasets and metrics. The diminished performance of ML-metaL stems from its inability to manage domain discrepancies and a shared label space.

Ablation Studies. The fine-tune set is important to address the GenCDML-FSL problem as demonstrated by the comparison between GenET *vs*. MMAML. Table 1 shows GenET outperforms MMAML on the mAP metric. In Table 2, F1 scores based on oracle threshold indicate that GenET outperforms MMAML in 5 of 6 cases, while at 0.5 threshold, the two approaches are closer.

5. CONCLUSION

We introduced a new few-shot learning problem appropriate for predicting chest X-ray abnormalities. *Generalized Cross-domain Multi-label Few-shot learning* (GenCDML-FSL) encompasses overlapping and non-overlapping classes, domain disparities, and multi-label instances. To address these challenges, we proposed *Generalized Episodic Training* (GenET) that simulates challenges of GenCDML-FSL during the training process. Through empirical validation, we demonstrated that adopting GenET enhances the model's ability to learn class-invariant representations, outperforming transfer learn-

Th.	Methods	NIH → CheXpert			NIH → PadChest			NIH → MIMIC			MIMIC → CheXpert			MIMIC → PadChest			MIMIC → NIH		
		F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R	F1	P	R
Oracle	TL	0.3336	0.2599	0.5093	0.2126	0.1825	0.3120	0.3225	0.2559	0.4637	0.3846	0.3343	0.4991	0.2046	0.1588	0.3823	0.2131	0.2414	0.5881
	HTL	0.5422	0.4838	0.7627	0.4746	0.4326	0.6651	0.5097	0.4389	0.7680	0.5727	0.5332	0.6828	0.4347	0.4298	0.6636	0.4928	0.4713	0.7092
	ML-metaL	0.4965	0.3872	0.8384	0.1788	0.1481	0.2883	0.3375	0.2983	0.5026	0.5297	0.3812	0.9476	0.3454	0.3198	0.5193	0.2767	0.2918	0.3734
	MMAML	0.5481	0.4452	0.8571	0.4897	0.3938	0.7593	0.5255	0.4137	0.8319	**0.5850**	0.4853	0.7805	0.4931	0.3660	0.8603	0.4932	0.3884	0.7807
	GenET	**0.5803**	0.4590	0.8898	**0.4997**	0.4144	0.7356	**0.5476**	0.4509	0.8269	0.5733	0.4449	0.8922	**0.4953**	0.4074	0.7079	**0.6363**	0.6054	0.7092
0.5	TL	0.1818	0.3815	0.1426	0.0926	0.3794	0.0887	0.1544	0.3790	0.1369	0.2930	0.5073	0.2575	0.1185	0.3107	0.1061	0.1611	0.2944	0.2930
	HTL	0.3290	0.4417	0.3247	0.2345	0.4780	0.1882	0.3064	0.4984	0.2694	0.3196	0.5221	0.2749	0.3085	0.4931	0.2697	0.3526	0.5363	0.3016
	ML-metaL	0.2082	0.2298	0.2667	0.0706	0.0752	0.1005	0.2126	0.2245	0.2738	0.1944	0.2128	0.2682	0.0466	0.0897	0.0712	0.0877	0.1555	0.1066
	MMAML	**0.4067**	0.4600	0.4258	**0.3683**	0.4355	0.3549	**0.3904**	0.4851	0.3785	0.4151	0.5358	0.3934	0.2617	0.3412	0.2511	0.3162	0.4274	0.2931
	GenET	0.3666	0.4766	0.3529	0.3327	0.4867	0.2885	0.3867	0.5152	0.3594	**0.4203**	0.5473	0.4020	**0.3303**	0.4789	0.2956	**0.5394**	0.6517	0.5033

Table 2. Comparing all baselines against GenET using threshold-based metrics such as F1, P: Precision, and R: Recall. We report results for both thresholds: the best threshold denoted as Oracle (top) and 0.5 (bottom). Overall, GenET shows good performance across all 6 setups, while achieving highest F1 score for 5 of 6 experiments with the oracle threshold.

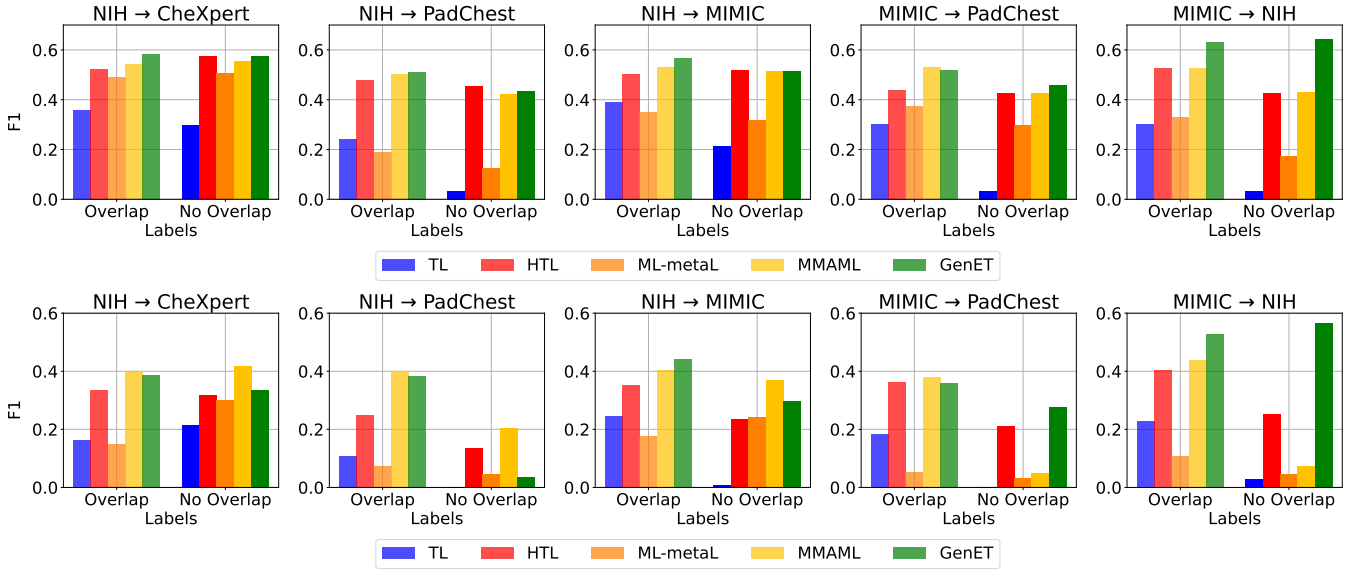

Fig. 1. Mean F1 scores for overlapping (*Overlap*) and non-overlapping (*No Overlap*) classes. **Top: oracle threshold** and **Bottom: 0.5 threshold**. We ignore MIMIC → CheXpert as both datasets contain the same set of labels. GenET outperforms baselines in majority of the cases and shows competitive performance with other meta-learning approaches.

ing and other meta-learning baselines in majority of the cases across both overlapping and non-overlapping classes.

6. REFERENCES

[1] Y Song, T Wang, T Cai, T K Mondal, and J P Sahoo, "A comprehensive survey of few-shot learning: Evolution, applications, challenges, and opportunities," *ACM Computing Surveys*, vol. 55, no. 13s, pp. 1–40, 2023.

[2] H Shao, X Zhou, J Lin, and B Liu, "Few-shot cross-domain fault diagnosis of bearing driven by task-supervised anil," *IEEE Internet of Things Journal*, 2024.

[3] W Chao, S Changpinyo, B Gong, and F Sha, "An empirical study and analysis of generalized zero-shot learning for object recognition in the wild," in *ECCV Workshops*, 2016.

[4] RM Summers, "NIH Chest X-Ray Dataset of 14 Common Thorax Disease Categories," 2019.

[5] J Irvin, P Rajpurkar, M Ko, et al., "Chexpert: A Large Chest Radiograph Dataset with Uncertainty Labels and Expert Comparison," in *AAAI*, 2019.

[6] A Bustos, A Pertusa, J Salinas, and M de la Iglesia-Vayá, "Padchest: A Large Chest X-Ray Image dataset with Multi-label annotated reports," *Medical image analysis*, vol. 66, 2020.

[7] A Johnson, TJ Pollard, S J Berkowitz, et al., "MIMIC-CXR, A De-identified publicly available database of Chest Radiographs with free-text reports," *Scientific Data*, vol. 6, no. 1, 2019.

[8] G S Dhillon, P Chaudhari, A Ravichandran, and S Soatto, "A baseline for few-shot image classification," in *ICLR*, 2020.

[9] C Finn, P Abbeel, and S Levine, "Model-agnostic meta-learning for fast adaptation of deep networks," in *ICML*, 2017.

[10] C Simon, P Koniusz, and M Harandi, "Meta-learning for multi-label few-shot classification," in *WACV*, 2022.

[11] W Chen, Y Liu, Zsolt Kira, Y F Wang, and J Huang, "A closer look at few-shot classification," in *ICLR*, 2019.

[12] J Snell, K Swersky, and R Zemel, "Prototypical networks for few-shot learning," in *NeurIPS*, 2017.