
TASK-AGNOSTIC CONTINUAL LEARNING FOR CHEST RADIOGRAPH CLASSIFICATION

Muthu Subash Kavitha, Anas Zafar, Amgad Muneer, Jia Wu*

Department of Imaging Physics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA

*Corresponding author: email@mdanderson.org

ABSTRACT

Clinical deployment of chest radiograph classifiers requires models that can be updated as new datasets become available without retraining on previously observed data or degrading validated performance. We study, for the first time, a task-incremental continual learning setting for chest radiograph classification, in which heterogeneous chest X-ray datasets arrive sequentially and task identifiers are unavailable at inference. We propose a continual adapter-based routing learning strategy for Chest X-rays (CARL-XRay) that maintains a fixed high-capacity backbone and incrementally allocates lightweight task-specific adapters and classifier heads. A latent task selector operates on task-adapted features and leverages both current and historical context preserved through compact prototypes and feature-level experience replay. This design supports stable task identification and adaptation across sequential updates while avoiding raw-image storage. Experiments on large-scale public chest radiograph datasets demonstrate robust performance retention and reliable task-aware inference under continual dataset ingestion. CARL-XRay outperforms joint training under task-unknown deployment, achieving higher routing accuracy (75.0% vs. 62.5%), while maintaining competitive diagnostic performance with AUROC of 0.74 in the oracle setting with ground-truth task identity and 0.75 under task-unknown inference, using significantly fewer trainable parameters. Finally, the proposed framework provides a practical alternative to joint training and repeated full retraining in continual clinical deployment.

1 INTRODUCTION

Deep learning for chest radiograph classification achieves strong multi-label performance for common findings, especially with large backbones trained on curated datasets Rajpurkar et al. (2017); Irvin et al. (2019); Johnson et al. (2019). This standard practice trains on a fixed distribution and retrains as new data arrive. However, it is less suitable for clinical deployment.

A practical clinical system needs sequential updates across years while preserving earlier diagnostic knowledge. The requirement is threefold: new datasets should be added incrementally, earlier performance should remain stable without continuous access to historical images, and the update cost should remain small. Continual learning formalizes this stability via plasticity setting Parisi et al. (2019); De Lange et al. (2021). Several traditional continual learning approaches remain difficult to apply to large scale medical backbones. A complete network training updates increase interference across tasks and storage burden Kirkpatrick et al. (2017); Rebuffi et al. (2017); Lopez-Paz & Ranzato (2017). Data access constraints and update cost often determine method choice in medical imaging.

Foundation radiograph models have demonstrated improved transferability and robustness under distribution shift Wang et al. (2022); Tiu et al. (2022); Zhang et al. (2020); Azizi et al. (2021). However, adaptation to new clinical sources still typically relies on full layers fine-tuning or joint multi-dataset training Kulkarni et al. (2025); Chambon et al. (2022); Ma et al. (2025). Systematic evaluation of performance retention across repeated updates is essential for reliable clinical decision support. This work introduces a continual chest radiograph classifier that supports sequential dataset ingestion with retention control. A Swin Transformer encoder is used as a frozen backbone to provide stable hierarchical features across time Liu et al. (2021). Each dataset is assigned a lightweight

adapter and a task-specific head. Recent work in medical vision–language learning also supports frozen-backbone designs with lightweight modules to reduce training cost while preserving prior knowledge Qin et al. (2024). A latent task selector is trained to infer the most appropriate task context from adapted feature representations, guided by compact task prototypes. Selector stability is maintained through feature-level experience replay, which stores a bounded set of adapted feature vectors across tasks. This mechanism preserves prior-task evidence for routing across updates and scales naturally as new datasets are added. The resulting system supports continual extension to additional datasets over time while maintaining stable performance and reliable task routing under label-free deployment. Our contributions are as follows:

- We introduce the first task-incremental continual learning formulation for chest radiograph classification, a setting for which no standardized evaluation protocols currently exist. The formulation reflects realistic clinical deployment, where heterogeneous datasets arrive sequentially and task identifiers are unavailable at inference.
- We propose CARL-XRay, a continual learning framework in which task-specific adapters are trained in isolation to preserve clear task representations under sequential updates, while feature-level experience replay stabilizes a latent task selector by maintaining consistent decision boundaries across tasks. Together, these mechanisms enable reliable task-unknown routing without degrading classification performance.
- We show that CARL-XRay supports efficient sequential adaptation without joint access to previously seen data or repeated full retraining, substantially reducing trainable parameter growth and computational cost compared to joint training and multi-dataset models.
- We conduct a large-scale task-incremental evaluation on public chest radiograph datasets, reporting task-wise AUROC, catastrophic forgetting, routing accuracy, and trainable parameter requirements. This evaluation protocol enables principled comparison of continual radiograph learning strategies under realistic clinical deployment conditions.

2 RELATED WORK

2.1 CONTINUAL LEARNING AND PARAMETER ISOLATION

Continual learning addresses sequential task updates by mitigating catastrophic forgetting through regularization, rehearsal, and architectural expansion strategies Kirkpatrick et al. (2017); Rebuffi et al. (2017); Lopez-Paz & Ranzato (2017); Parisi et al. (2019); De Lange et al. (2021). While rehearsal-based methods can be effective, storing raw images is often impractical under clinical governance, privacy, and storage constraints Rebuffi et al. (2017); Lopez-Paz & Ranzato (2017). Parameter-isolation approaches, such as progressive networks and adapter-based transfer, preserve prior knowledge by freezing shared components and incrementally adding lightweight task-specific modules Rusu et al. (2016); Houlsby et al. (2019). Additional model drifting methods are provided in Appendix A.1.

2.2 TASK AWARENESS AND MODULAR CONTINUAL UPDATES

Following this principle, recent work has explored freezing shared vision backbones, such as Swin Transformers Liu et al. (2021), while updating only dataset-specific adapters and classifier heads to limit representational drift and maintain efficient continual updates. Compute-aware continual learning in medical imaging further supports modular strategies that reduce update cost. Task awareness at inference remains critical in multi-hospital deployment, as task identifiers are often unavailable. Learned routing methods address this challenge by selecting task-specific pathways from feature representations Rusu et al. (2016). Selector-based approaches using compact task prototypes and bounded feature-level replay have been proposed to stabilize task identification while maintaining low storage overhead and better aligning with clinical data access and privacy constraints than input-level rehearsal Rebuffi et al. (2017); Lopez-Paz & Ranzato (2017).

3 METHODOLOGY

In this work, we are interested in training a continual chest radiograph classification model that supports sequential ingestion of heterogeneous clinical datasets without joint access to previously observed training data. The setting reflects realistic deployment scenarios in which institutions curate datasets independently and label spaces partially overlap. Figure 1 presents an overview of the proposed framework. The design keeps a high-capacity backbone encoder fixed to ensure representational stability, while each new task allocates only lightweight task-specific adaptation modules. A task-awareness component infers the appropriate task context at inference in the absence of explicit task identifiers.

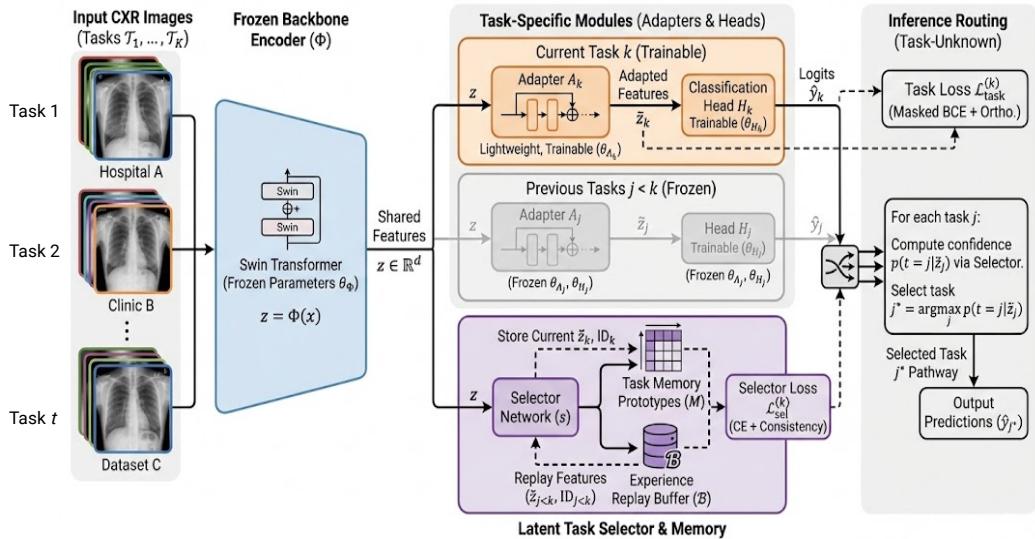


Figure 1: Proposed continual learning framework.

3.1 PROBLEM SETTING

Continual radiograph classification is formulated as an ordered sequence of supervised multi-label tasks $\{\mathcal{T}_1, \dots, \mathcal{T}_K\}$. Each task \mathcal{T}_k corresponds to a dataset from a distinct clinical source or labeling pipeline. Task \mathcal{T}_k provides $\mathcal{D}_k = \{(x_i^{(k)}, y_i^{(k)})\}_{i=1}^{N_k}$, where $x_i^{(k)}$ is a radiograph and $y_i^{(k)} \in \{0, 1, -1, \text{NaN}\}^{C_k}$ is a multi-label vector over C_k findings. Training proceeds sequentially by learning task \mathcal{T}_k using access to \mathcal{D}_k and parameters learned from tasks $\{1, \dots, k-1\}$. The goal is to maintain strong performance on all observed tasks while efficiently adapting to each new task under incremental access. Algorithm 1 summarizes the full training and inference procedure of the proposed CARL-XRay framework is presented in Appendix B.1.

3.2 MODEL ARCHITECTURE

Let $\Phi(\cdot; \theta_\Phi)$ be an image encoder mapping x to a feature vector $z \in \mathbb{R}^d$: A swin transformer Liu et al. (2021) backbone is used to capture both local and global radiographic structure. The backbone parameters θ_Φ remain frozen throughout continual training. This constraint promotes representational stability and reduces interference during updates. For each task k , a task-specific adapter $A_k(\cdot; \theta_{A_k})$ and classification head $H_k(\cdot; \theta_{H_k})$ are allocated. The adapter transforms shared features into task-adapted features:

$$\tilde{z}_k = A_k(z; \theta_{A_k}), \quad (1)$$

and the head produces logits over the task label set:

$$\hat{y}_k = H_k(\tilde{z}_k; \theta_{H_k}). \quad (2)$$

Only θ_{A_k} and θ_{H_k} are updated when learning task k . All previously learned $\{\theta_{A_j}, \theta_{H_j}\}_{j < k}$ remain frozen. This isolate-then-freeze strategy reduces catastrophic forgetting without storing raw images.

Adapter modules: Inspired by long-term memory systems, we focus on motivating the design of adapter modules as persistent memory-like components and on how they can be effectively integrated into the overall architecture Behrouz et al. (2024; 2023). To study this, we employ multiple adapter designs, including a simple single-layer MLP, a Continuum memory system, and a Hope memory system, to examine their ability to learn and retain task-specific representations. Simple adapter uses a bottleneck residual MLP, $A(z) = z + \text{MLP}(z)$. Continuum adapter increases capacity via multiple residual MLP branches, $A(z) = z + \sum_{m=1}^3 \text{MLP}_m(z)$. A Hope adapter appends an attention like residual transform before the Continuum block. These variants trade off trainable parameters, memory usage, and routing separability under task-unknown inference.

3.3 TRAINING OBJECTIVE AND LABEL HANDLING

Each task is trained with a masked multi-label binary cross-entropy objective. Let $\hat{y}_{i,c}^{(k)}$ denote the logit for class c and sample i in task k . Let $y_{i,c}^{(k)} \in \{0, 1, -1, \text{NaN}\}$ be the corresponding target. Valid labels are defined as

$$\Omega_k = \{(i, c) : y_{i,c}^{(k)} \neq \text{NaN}\}. \quad (3)$$

Entries with $y = \text{NaN}$ are excluded from the loss. Uncertain labels ($y = -1$) are not treated as negative. Instead, a soft target $\tilde{y} \sim \mathcal{U}(\alpha, \beta)$ is used, which prevents overconfident updates on uncertain clinical annotations. The masked BCE loss is

$$\mathcal{L}_{\text{BCE}}^{(k)} = \frac{1}{|\Omega_k|} \sum_{(i,c) \in \Omega_k} \text{BCEWithLogits}(\hat{y}_{i,c}^{(k)}, \tilde{y}_{i,c}^{(k)}). \quad (4)$$

In addition, an orthogonality regularizer is applied on adapted features to reduce redundancy. Given a batch $\tilde{Z}_k \in \mathbb{R}^{B \times d}$, features are ℓ_2 -normalized and a cosine similarity matrix $S = \tilde{Z}_k \tilde{Z}_k^\top$ is formed. The off-diagonal similarity is penalized as

$$\mathcal{L}_{\text{ortho}}^{(k)} = \frac{1}{B(B-1)} \sum_{i \neq j} S_{ij}. \quad (5)$$

Thus the adapter task specific loss is defined as

$$\mathcal{L}_{\text{task}}^{(k)} = \mathcal{L}_{\text{BCE}}^{(k)} + \lambda_{\text{ortho}} \mathcal{L}_{\text{ortho}}^{(k)}. \quad (6)$$

3.4 LATENT TASK SELECTOR WITH PROTOTYPE MEMORY

A latent task selector predicts the task context from intermediate features. Let $s(\cdot; \theta_S)$ be an MLP producing logits over tasks, is denoted as

$$\ell = s(\tilde{z}; \theta_S) \in \mathbb{R}^K, \quad p(t | \tilde{z}) = \text{softmax}(\ell). \quad (7)$$

A learnable memory matrix $M \in \mathbb{R}^{K \times d}$ is maintained, where M_k is a prototype embedding for task k . During training on task k , the selector is optimized using cross-entropy to predict the task is defined as,

$$\mathcal{L}_{\text{sel_CE}}^{(k)} = \text{CE}(\ell, k), \quad (8)$$

and a prototype consistency loss is

$$\mathcal{L}_{\text{mem}}^{(k)} = \|\tilde{z}_k - M_k\|_2^2. \quad (9)$$

The selector loss is

$$\mathcal{L}_{\text{sel}}^{(k)} = \mathcal{L}_{\text{sel_CE}}^{(k)} + \lambda_{\text{mem}} \mathcal{L}_{\text{mem}}^{(k)}. \quad (10)$$

Adapters and heads are task-isolated, whereas the selector is shared and updated at every task. Therefore, the selector is the main source of task-identity drift under continual updates. To stabilize task identification, experience replay is applied at the feature level. A replay buffer \mathcal{B} stores a bounded number of adapted feature vectors \tilde{z} from previous tasks with their task IDs. During training on task k , selector optimization uses a mixed batch of current-task features and replayed

features from $\{1, \dots, k-1\}$. Cross-entropy is computed on the mixed labels. In contrast, the prototype consistency loss is applied only to current-task features. Both the task prototypes and the replay buffer are constructed from task-adapted features produced by the corresponding adapters, ensuring consistency between selector inputs, prototype representations, and replayed samples. This design targets selector forgetting while respecting clinical data-governance constraints.

3.5 TASK UNKNOWN INFERENCE AND ROUTING

If task identity is known, prediction uses the corresponding (A_k, H_k) . If task identity is unknown, each test image must be routed to a task-specific pathway. The primary routing is selector-based and adapter-conditioned. For each task j , adapted features are computed as $\tilde{z}_j = A_j(\Phi(x))$. These adapted features are then passed to the selector, and the diagonal probability $p(t = j | \tilde{z}_j)$ is used as the confidence score for task j . The predicted task is selected by argmax over tasks, and the corresponding head output is used for prediction. This procedure is consistent with training. The selector operates on adapted features, not backbone features. In addition, two alternative routing signals are evaluated as non-parametric inference-time ablations. Memory-based routing uses cosine similarity between \tilde{z}_j and the prototype M_j . Entropy-based routing selects the task whose head yields the lowest mean predictive entropy. These variants quantify reliance on learned selector scores versus prototype matching and uncertainty cues.

4 EXPERIMENTAL SETTINGS AND RESULTS

We evaluate task-incremental continual chest radiograph classification under sequential dataset ingestion using two major public datasets: MIMIC-CXR (377,110 images) and CheXpert (224,316 images), each annotated with 14 clinical findings. In the continual setting, Task 1 corresponds to MIMIC-CXR classification and Task 2 corresponds to CheXpert classification. Performance is assessed using AUROC and macro F1 averaged across datasets, together with catastrophic forgetting, routing accuracy under task-unknown inference, and comparison to joint training. Additional ablations analyze adapter-based routing strategies, replay buffer size, and trainable parameter requirements. Further details on data processing and implementation are provided in Appendix B.

4.1 EVALUATING DIAGNOSTIC PERFORMANCE

Evaluation Metrics. We report per-task AUROC for classification performance and per-task routing accuracy for task identification. Per-task routing accuracy is computed separately on samples from each dataset. Overall routing accuracy is measured over the union of all test samples across tasks, weighted by dataset size, and therefore reflects deployment performance rather than the arithmetic mean of per-task accuracies. As a result, high accuracy on a larger dataset can dominate the overall score even when per-task accuracies differ substantially.

Task Definition. In our two-task evaluation: **Task 1** = MIMIC-CXR (377,110 images, first task learned) and **Task 2** = CheXpert (224,316 images, second task learned). Training proceeds sequentially: Task 1 → Task 2. At inference, task identity is unknown unless explicitly stated.

Sequential Continual Learning Performance. In the main setting with task prototypes and feature-level replay, inference-time routing is performed exclusively by the learned latent selector. As shown in Table 1, CARL-XRay maintains strong diagnostic performance with minimal forgetting under sequential updates. After learning Task 1 (MIMIC-CXR), the model achieves an AUROC of 0.752 ± 0.003 . Following the addition of Task 2 (CheXpert), the model maintains strong performance on the previously learned task, achieving an AUROC of 0.740 ± 0.004 on Task 1, while reaching 0.748 ± 0.003 on Task 2. The resulting forgetting on Task 1 is limited to 0.012, indicating effective retention under sequential updates. The combined adapters and selector introduce only 2.3 MB of additional parameters (0.08% of the backbone), corresponding to approximately $1250 \times$ fewer trainable parameters than full backbone fine-tuning.

Joint Training Upper-Bound Analysis. We compare the proposed sequential continual setting against a joint training baseline that assumes simultaneous access to both datasets. Additional joint

Table 1: Performance of CARL-XRay Under Sequential Continual Learning.

Training Phase	MIMIC AUC	CheXpert AUC	Forgetting	Memory (MB)
Task 1 (MIMIC)	0.752 ± 0.003	–	–	1.1
Task 2 (CheXpert)	0.740 ± 0.004	0.748 ± 0.003	0.012	2.3

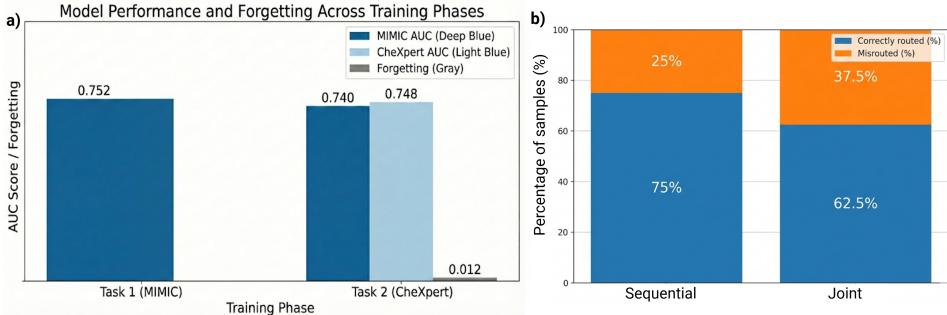


Figure 2: Diagnostic performance of CARL-XRay. (a) Forgetting on Task 1 is 0.012, indicating strong retention under sequential updates. (b) Under task unknown inference, sequential training achieves 75% routing accuracy by maintaining task specific structure through isolated adapters and feature level replay. In contrast, joint training attains 62.5% routing accuracy, suggesting weaker task separability when both datasets are optimized simultaneously.

training setup are provided in Appendix C. With task identity provided at inference, joint training achieves an AUROC of 0.74 on MIMIC-CXR (Task 1) and 0.73 on CheXpert (Task 2), comparable to sequential results (0.74 and 0.75, respectively), within a 2% margin. However, under task-unknown deployment, which is common in multi-hospital environments, routing performance under joint training degrades substantially (62.5% vs. 75.0%) (Fig. 2). While joint training matches AUROC when task identity is known, it substantially degrades task-unknown routing, which is the critical requirement for real-world deployment.

This degradation arises because joint training optimizes on both datasets simultaneously, which reduces the separation between task-specific representations and weakens task boundary cues required for reliable routing. In contrast, the sequential continual setting preserves clearer task-specific structure by learning from incrementally arriving datasets. Task-specific adapters are trained in isolation, and feature-level replay maintains stable decision boundaries in the selector. This design leads to improved task-unknown routing while maintaining comparable classification performance.

Task Order Sensitivity. We analyze the sensitivity of the proposed framework to task ordering by reversing the training sequence (Task 2→Task 1) and evaluating routing performance under task-unknown inference. Routing accuracy decreases from approximately 75.0% under the core order (MIMIC-CXR→CheXpert) to approximately 70.0% under the reversed order, corresponding to a 5% absolute drop. Despite this reduction, the framework remains robust to task ordering, indicating that feature-level replay preserves usable task identities in both directions.

The slightly stronger performance under the core order is expected, as MIMIC-CXR is larger and more diverse and therefore provides a stronger initialization for subsequent adaptation. Overall, the observed order effect is limited, suggesting that task ordering is a practical consideration rather than a fundamental limitation of the proposed approach.

Performance Under Task Known and Task Unknown Deployment. In the oracle setting with known task identity, CARL-XRay attains an AUROC of 0.74, which is comparable to reported performance of multi-dataset pretrained chest radiograph models Kulkarni et al. (2025); Chambon et al. (2022); Ma et al. (2025). Under the more realistic task-unknown inference setting, CARL-XRay maintains an AUROC of 0.75, demonstrating that strong diagnostic performance can be preserved even when task identifiers are unavailable and with substantially fewer trainable parameters. Refer Appendix C.1

Direct comparison with prior work is not feasible because standardized continual learning benchmarks for chest radiograph classification are currently unavailable and reported settings vary widely. Prior studies often use large scale multi dataset pretraining with image text supervision and heavy architectures. In contrast, CARL-XRay is trained sequentially in a task incremental setting using only image level supervision, reports AUROC as a macro average across all clinical findings, and achieves competitive performance even under task unknown deployment, a scenario not addressed in prior work.

4.2 ABLATION STUDY

Effect of Experience Replay. This experiment evaluates sequential continual learning without experience replay, serving as a lower-bound baseline for task-unknown inference (Table 2). Routing performance degrades catastrophically without replay. After learning Task 2, the selector largely forgets Task 1, routing almost all samples to the most recent task. The low overall routing accuracy (14.3%) indicates that task discrimination becomes unreliable without access to representations from previously learned tasks.

In contrast, incorporating experience replay leads to a marked improvement in routing performance. Selector-based routing reaches 75.0% accuracy, corresponding to a 60.7 % gain over the no-replay setting. This large performance gap demonstrates that replay is the primary mechanism enabling stable task identification in the proposed framework. By retaining a bounded set of adapter-adapted features and their associated task labels, replay mitigates task-identity drift in the shared selector and prevents catastrophic forgetting of task identities across sequential updates.

Table 2: Impact of experience replay on task-unknown routing performance.

Setting	Routing	Overall Routing Acc (%)
Prototypes only	Selector-based (MLP)	14.3
Prototypes + Replay	Selector-based (MLP)	75.0
Prototypes only	Entropy-based	57.3

Inference-Time Routing Strategies. Automatic task routing is required at inference, as task identifiers are often unavailable in multi-hospital deployment. We evaluated two alternative routing strategies—memory-based and entropy-based—as inference-time ablations (Table 3). Memory-based routing exhibits severe task imbalance, routing the majority of samples to a single task (88.0% MIMIC, 11.0% CheXpert). This collapse toward the MIMIC prototype suggests that cosine similarity to task prototypes alone provides insufficient discrimination between task-adapted feature distributions, resulting in near-total routing bias that fails to identify CheXpert (Task 2) samples.

In contrast, entropy-based routing achieves balanced performance across both datasets (66.0% MIMIC, 65.0% CheXpert) with substantially higher overall routing accuracy (65.0%) and classification AUC (0.740). Entropy routing leverages prediction confidence directly from the task-specific heads, providing a more robust signal for task discrimination than prototype similarity alone. These results demonstrate that confidence-based routing is more robust for task identification, leading to fewer misrouted cases and better classification performance under task-unknown deployment.

Table 3: Task-unknown inference routing strategies performance.

Routing Strategy	MIMIC Acc (%)	CheXpert Acc (%)	Overall Acc (%)	AUC
Memory-based	88.0	11.0	50.0	0.721
Entropy-based	66.0	65.0	65.0	0.740

Note: Acc denotes routing per task and overall accuracy.

Effect of Experience Replay Capacity. Table 4 reports routing performance across different replay buffer sizes. Introducing feature-level replay consistently improves routing behavior. Routing accuracy denotes fraction of correctly routed samples. Moderate buffer capacities of 1,000 and 2,500

achieve overall routing accuracies of 0.674 and 0.575, respectively. The highest overall routing accuracy is observed with a buffer size of 5,000 (0.748), reflecting a favorable balance between retaining representative features from earlier tasks and incorporating new task information. Further increasing the buffer capacity to 10,000 reduces overall accuracy to 0.702, indicating that larger replay buffers do not necessarily yield additional gains and may introduce stale or less representative features. Overall, these results show that feature-level replay is effective for stabilizing task-unknown routing by preserving representative embeddings from earlier updates, while replay capacity directly influences routing stability through the diversity and relevance of stored features.

Table 4: Routing accuracy under different replay buffer capacities (task-unknown inference).

Buffer Capacity	MIMIC Acc	CheXpert Acc	Overall Routing Acc
0	0.520	0.835	0.556
1000	0.690	0.549	0.674
2500	0.546	0.795	0.575
5000	0.778	0.523	0.748
10000	0.726	0.517	0.702

Note: Acc denotes routing per task and overall accuracy.

Effect of Adapter Design. We evaluated three adapter variants to study the impact of adapter design on classification and routing performance (Table 5). The Continuum adapter achieves the highest routing accuracy (0.710) and the strongest overall classification performance, including the best CheXpert AUROC (0.788) while maintaining competitive MIMIC-CXR AUROC (0.747), with modest memory requirements (4.61 MB). The Simple adapter is the most memory-efficient (1.51 MB) but yields lower routing accuracy (0.660), indicating insufficient capacity to separate task contexts.

In contrast, the Hope adapter substantially increases memory usage (40.65 MB) without improving routing performance (0.575). This suggests that excessive adapter complexity can increase feature overlap across tasks and destabilize task identification. Overall, these findings support the Continuum adapter as the most effective design for continual deployment, balancing representational capacity, routing reliability, and parameter efficiency.

Table 5: Performance comparison of adapter designs.

Adapter Type	MIMIC AUC	CheXpert AUC	Overall Routing Acc.	Memory (MB)
Simple	0.745	0.760	0.660	1.51
Continuum	0.747	0.788	0.710	4.61
Hope	0.732	0.770	0.575	40.65

Note: Acc denotes overall routing accuracy. Memory denotes trainable parameters (adapters + heads + selector).

5 CONCLUSIONS

We presented CARL-XRay, an continual adapter-based routing learning for task-incremental chest radiograph classification designed for realistic clinical deployment. Allocating lightweight task-specific adapters and heads in a high-capacity freezing backbone, CARL-XRay maintains strong diagnostic performance while limiting catastrophic forgetting and computational overhead. A latent task selector operating on adapter-adapted features enables reliable task-unknown inference, and feature-level experience replay is shown to be essential for preserving task identity across sequential updates in clinical settings. Ablation studies further highlight the importance of replay capacity and adapter design for stable inference-time routing. Future work will extend CARL-XRay to longer task sequences, evaluate continual learning across additional institutions and modalities, and investigate adaptive replay strategies to further support safe and scalable clinical deployment.

REFERENCES

- Shekoofeh Azizi, Basil Mustafa, Fiona Ryan, Zachary Beaver, Jan Freyberg, Jonathan Deaton, Aaron Loh, Alan Karthikesalingam, Simon Kornblith, Ting Chen, and Mohammad Norouzi. Big self-supervised models advance medical image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3478–3488, 2021.
- Ali Behrouz, Meisam Razaviyayn, Peilin Zhong, and Vahab Mirrokni. Nested learning: The illusion of deep learning architecture. *arXiv preprint*, 2023.
- Ali Behrouz, Peilin Zhong, and Vahab Mirrokni. Titans: Learning to memorize at test time. *arXiv preprint*, 2024.
- Pierre Chambon, Christian Bluethgen, Jean-Benoit Delbrouck, Rogier Van der Sluijs, Małgorzata Połacina, Juan Manuel Zambrano Chaves, Tanishq Mathew Abraham, Shivanshu Purohit, Curtis P. Langlotz, and Akshay Chaudhari. Roentgen: Vision-language foundation model for chest x-ray generation. *arXiv preprint arXiv:2211.12737*, 2022.
- Matthias De Lange, Rahaf Aljundi, Marc Masana, Sarah Parisot, Xu Jia, Ales Leonardis, Gregory Slabaugh, and Tinne Tuytelaars. A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.
- Samuel G. Finlayson, Adarsh Subbaswamy, Karan Singh, et al. The clinician and dataset shift in artificial intelligence. *New England Journal of Medicine*, 385(3):283–286, 2021.
- Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 2790–2799, 2019.
- Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Christopher Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. CheXpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 590–597, 2019.
- Alistair E. W. Johnson, Tom J. Pollard, Seth J. Berkowitz, Nathan R. Greenbaum, Matthew P. Lungren, Chih-Ying Deng, and Roger G. Mark. MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports. *Scientific Data*, 6:317, 2019.
- James Kirkpatrick, Razvan Pascanu, Neil C. Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, 2017.
- Pranav Kulkarni, Adway Kanhere, Paul H. Yi, and Vishwa S. Parekh. From isolation to collaboration: Federated class-heterogeneous learning for chest x-ray classification. In *Proceedings of the 4th Machine Learning for Health Symposium*, volume 259 of *Proceedings of Machine Learning Research*, pp. 623–635. PMLR, 2025.
- Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9992–10002, 2021.
- David Lopez-Paz and Marc’Aurelio Ranzato. Gradient episodic memory for continual learning. In *Advances in Neural Information Processing Systems*, pp. 6467–6476, 2017.
- DongAo Ma, Jiaxuan Pang, Michael B. Gotway, and Jianming Liang. A fully open ai foundation model applied to chest radiography. *Nature*, 643:488–498, 2025. doi: 10.1038/s41586-025-09079-8.
- German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, 2019.

-
- Jiuming Qin, Che Liu, Sibo Cheng, Yike Guo, and Rossella Arcucci. Freeze the backbones: A parameter-efficient contrastive approach to robust medical vision-language pre-training. *arXiv preprint arXiv:2401.01179*, 2024.
- 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.
- Sylvestre-Alvise Rebuffi, Alexander Kolesnikov, Georg Sperl, and Christoph H. Lampert. iCaRL: Incremental classifier and representation learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2001–2010, 2017.
- Andrei A. Rusu, Neil C. Rabinowitz, Guillaume Desjardins, Hubert Soyer, James Kirkpatrick, Koray Kavukcuoglu, Razvan Pascanu, and Raia Hadsell. Progressive neural networks. *arXiv preprint arXiv:1606.04671*, 2016.
- Ekin Tiu, Ellie Talius, Pujan Patel, Curtis P. Langlotz, Andrew Y. Ng, and Pranav Rajpurkar. Expert-level detection of pathologies from unannotated chest x-ray images via self-supervised learning. *Nature Biomedical Engineering*, 6(12):1399–1406, 2022.
- Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3876–3887, 2022.
- John R. Zech, Marc A. Badgeley, MiaoSong Liu, Allison B. Costa, Joseph J. Titano, and Eric K. Germann. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, 15(11):e1002683, 2018.
- Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and Curtis P. Langlotz. Contrastive learning of medical visual representations from paired images and text. *arXiv preprint arXiv:2010.00747*, 2020.

A APPENDIX

A.1 MODEL UPDATING UNDER DISTRIBUTION SHIFT

Clinical models often require re-validation after updates and can show unexpected performance drops under distribution shift Zech et al. (2018); Finlayson et al. (2021). Recent work has highlighted that identifying the type of shift is important for safe deployment ?. Although self-supervised and multimodal pretraining have improved radiograph feature learning and enabled stronger transfer and zero-shot interpretation Zhang et al. (2020); Azizi et al. (2021); Tiu et al. (2022); Wang et al. (2022), adapting models to new clinical sources still mainly relies on fine-tuning complete layers or joint training across datasets, which is often limited by practical constraints. This motivates continual update methods that support sequential dataset ingestion while preserving previously validated performance.

B DATA PROCESSING AND IMPLEMENTATION DETAILS

All chest radiographs are resized to 384×384 . During training, data augmentation includes random resized cropping, horizontal flipping, and small-angle rotation. At evaluation time, only resizing is applied. Images are normalized using ImageNet mean and standard deviation. All experiments use a frozen Swin-Large backbone pretrained on ImageNet. Only task-specific adapters, classifier heads, and the latent task selector are updated during continual training. Models are trained for 20 epochs per task with a batch size of 32. Adapter parameters are optimized with a learning rate of 1×10^{-4} , while the selector uses a learning rate of 5×10^{-4} , with weight decay 1×10^{-4} . An orthogonality regularizer with weight 0.05 and a prototype consistency loss weighted by 0.5 are applied. The adapter bottleneck dimension is set to 64, and the selector uses a hidden dimension of 256 with dropout 0.1. All experiments use a fixed random seed of 1337. Experiments are conducted on a DGX-H100 system with 8 NVIDIA H100 GPUs per run. Training is executed in a containerized Kubernetes environment with sufficient system memory to support efficient data loading and feature-level replay.

B.1 ALGORITHMIC DESCRIPTION OF CARL-XRAY

C JOINT TRAINING SETTING AS AN UPPER BOUND

This experiment defines a joint training setting that serves as an upper bound for continual learning performance. A single model is trained simultaneously on the MIMIC and CheXpert datasets using a shared backbone architecture. Separate task specific adapters and classification heads are maintained for each dataset. During training, data from both tasks are available, concurrently and the task selector is optimized with access to samples from both datasets. This setting is not imposed any sequential ordering of tasks and mechanisms to mitigate forgetting, as all tasks are learned together from the beginning. The joint training configuration represents an idealized multi-task learning scenario and is included as a reference for comparison with the sequential continual learning experiments presented in the main paper.

C.1 TASK ROUTING UNDER TASK UNKNOWN INFERENCE.

Figure 3 compares oracle evaluation with task-unknown routed inference and presents the corresponding confusion matrix. Oracle and routed evaluations achieve comparable macro AUROC on both MIMIC and CheXpert, indicating that diagnostic performance is largely preserved under task-unknown deployment. Relative to the task-known oracle setting, routed inference shows only modest performance changes.

For true MIMIC samples, the selector correctly routes 3,383 out of 5,159 cases (65.6%), while for true CheXpert samples, 432 out of 668 cases (64.7%) are routed to the correct task. Misrouting is balanced in both directions, showing that the proposed selector maintains robust task separation without bias toward a dominant dataset.

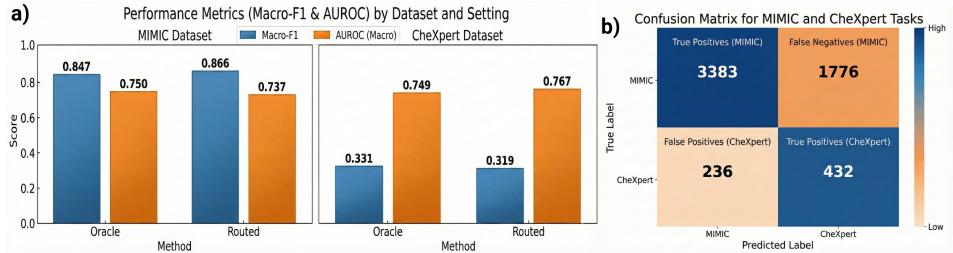


Figure 3: Task-unknown inference analysis. (a) Comparison of oracle (task-known) and routed (task-unknown) diagnostic performance on MIMIC and CheXpert. (b) Confusion matrix showing task routing accuracy between MIMIC and CheXpert under task-unknown inference.

Algorithm 1 CARL-XRay: Continual Adaptive and Task Unknown Routing for Chest Radiographs

Require: Ordered tasks $\{\mathcal{T}_1, \dots, \mathcal{T}_K\}$ with datasets \mathcal{D}_k
Require: Frozen encoder $\Phi(\cdot; \theta_\Phi)$; adapter family $A(\cdot; \theta_{A_k})$; head $H(\cdot; \theta_{H_k})$
Require: Task selector $s(\cdot; \theta_S)$; prototypes $M \in \mathbb{R}^{K \times d}$; replay buffer \mathcal{B}
Require: Hyperparameters $\lambda_{\text{ortho}}, \lambda_{\text{mem}}$, uncertain-label bounds (α, β) , buffer size B_{\max}
Ensure: Task-specific modules $\{A_k, H_k\}_{k=1}^K$, shared selector s , prototypes M , routing rule for task-unknown inference

- 1: Initialize θ_S ; initialize empty \mathcal{B} ; freeze θ_Φ
- 2: **for** $k = 1$ to K **do**
 - 3: Initialize $\theta_{A_k}, \theta_{H_k}$; initialize prototype M_k (e.g., zeros or random) modules (adapters + head) while keeping all previous modules frozen ▷ Train current task
 - 4: **for** epoch = 1 to E **do**
 - 5: **for all** mini-batches $(x, y) \sim \mathcal{D}_k$ **do**
 - 6: $z \leftarrow \Phi(x; \theta_\Phi)$ ▷ Frozen feature extraction
 - 7: $\tilde{z} \leftarrow A_k(z; \theta_{A_k})$ ▷ Task-specific adaptation
 - 8: $\hat{y} \leftarrow H_k(\tilde{z}; \theta_{H_k})$ ▷ Task-specific logits
 - 9: Construct masked targets \tilde{y} :
 - 10: **for all** labels c **do**
 - 11: **if** $y_c = \text{NaN}$ **then**
 - 12: mark (\cdot, c) as invalid (exclude from loss)
 - 13: **else if** $y_c = -1$ **then**
 - 14: $\tilde{y}_c \sim \mathcal{U}(\alpha, \beta)$ ▷ Uncertain label handling
 - 15: **else**
 - 16: $\tilde{y}_c \leftarrow y_c$
 - 17: **end if**
 - 18: **end for**
 - 19: $\mathcal{L}_{\text{BCE}} \leftarrow \text{MASKEDBCEWITHLOGITS}(\hat{y}, \tilde{y})$
 - 20: $\mathcal{L}_{\text{ortho}} \leftarrow \text{OFFDIAGCOSINEPENALTY}(\tilde{z})$
 - 21: $\mathcal{L}_{\text{task}} \leftarrow \mathcal{L}_{\text{BCE}} + \lambda_{\text{ortho}} \mathcal{L}_{\text{ortho}}$
 - 22: Update $\theta_{A_k}, \theta_{H_k}$ by descending $\nabla(\mathcal{L}_{\text{task}})$
 - 23: **end for**
 - 24: **end for** ▷ Update shared selector using current-task features + replayed features
 - 25: **for** epoch = 1 to E_S **do**
 - 26: **for all** mini-batches $(x, y) \sim \mathcal{D}_k$ **do**
 - 27: $z \leftarrow \Phi(x; \theta_\Phi)$
 - 28: $\tilde{z}_k \leftarrow A_k(z; \theta_{A_k})$ ▷ Selector consumes adapted features
 - 29: Sample a batch $\{(\tilde{z}_r, t_r)\}$ from replay buffer \mathcal{B} ▷ Feature-level replay
 - 30: Form mixed selector batch $\mathcal{Z} \leftarrow \{\tilde{z}_k\} \cup \{\tilde{z}_r\}$ with task labels $\mathcal{T} \leftarrow \{k\} \cup \{t_r\}$
 - 31: $\ell \leftarrow s(\mathcal{Z}; \theta_S)$; $p(t | \mathcal{Z}) \leftarrow \text{SOFTMAX}(\ell)$
 - 32: $\mathcal{L}_{\text{sel,CE}} \leftarrow \text{CE}(\ell, \mathcal{T})$
 - 33: $\mathcal{L}_{\text{mem}} \leftarrow \|\tilde{z}_k - M_k\|_2^2$ ▷ Prototype consistency (current task only)
 - 34: $\mathcal{L}_{\text{sel}} \leftarrow \mathcal{L}_{\text{sel,CE}} + \lambda_{\text{mem}} \mathcal{L}_{\text{mem}}$
 - 35: Update θ_S by descending $\nabla(\mathcal{L}_{\text{sel}})$
 - 36: Update prototype M_k (e.g., EMA): $M_k \leftarrow (1 - \eta)M_k + \eta \text{MEAN}(\tilde{z}_k)$
 - 37: Add $\{(\tilde{z}_k, k)\}$ to replay buffer \mathcal{B} ; if $|\mathcal{B}| > B_{\max}$, evict oldest
 - 38: **end for**
 - 39: **end for**
 - 40: Freeze $\theta_{A_k}, \theta_{H_k}$ ▷ Isolate-then-freeze to reduce interference
 - 41: **end for** ▷ Task-unknown inference (routing)
 - 42: **function** PREDICT(x)
 - 43: $z \leftarrow \Phi(x; \theta_\Phi)$
 - 44: **for** $j = 1$ to K **do**
 - 45: $\tilde{z}_j \leftarrow A_j(z; \theta_{A_j})$
 - 46: $\ell^{(j)} \leftarrow s(\tilde{z}_j; \theta_S)$; $p^{(j)} \leftarrow \text{SOFTMAX}(\ell^{(j)})$
 - 47: $score_j \leftarrow p^{(j)}(t = j)$ ▷ Diagonal confidence $p(t = j | \tilde{z}_j)$
 - 48: **end for**
 - 49: $j^* \leftarrow \arg \max_j score_j$
 - 50: **return** $H_{j^*}(\tilde{z}_{j^*}; \theta_{H_{j^*}})$
 - 51: **end function**