Distribution-guided Generative Replay with Semantic Prompts for Class-Incremental Chest X-ray Diagnosis

Jayant Mahawar
mahawar.2@iitj.ac.in
Devi Prasad Maharathy
maharathy.1@iitj.ac.in
Angshuman Paul
apaul@iitj.ac.in

Department of Computer Science and Engineering,
Indian Institute of Technology Jodhpur, India

Abstract

Existing class-incremental learning (CIL) methods perform poorly for the diagnosis of medical images. Some CIL approaches require storing exemplars, which raises privacy and storage concerns. Others rely on unconditioned generative replay, compromising discriminative power. To overcome these issues, we propose a new CIL framework for chest x-ray (CXR) diagnosis that combines prompt tuning with distribution-guided generative replay at the feature level. The prompts are initialized with semantically rich embeddings and refined through training. A variational autoencoder captures the feature distribution in latent space, enabling past knowledge retention without storing raw data. To balance stability and plasticity, parts of the network are frozen after the initial phase, while others adapt to new classes. In each session, the model learns new classes using real data and replays the synthetic features of old ones to reduce forgetting. A classification module picks the closest class prompt using cosine similarity. We evaluate our method on public CXR datasets. It outperforms prior CIL methods in accuracy and retention. We achieve up to 9% improvement in average accuracy compared to SOTA methods.

1 Introduction

Chest x-ray (CXR) diagnosis faces growing challenges as respiratory diseases evolve due to changing pathogens, demographics, and environmental conditions. Recent discoveries like LayV [3] and HKU5-CoV-2 [26] highlight the necessity for adaptable diagnostic systems. Disease prevalence also varies geographically. Traditional deep learning models require full retraining to incorporate new classes, which is often infeasible in medical settings due to privacy regulations and limited expert annotations. Retraining also risks overfitting to new data and catastrophic forgetting. Class-incremental learning (CIL) frameworks address this by learning new classes while preserving past performance. Figure 1 demonstrates our proposed CIL method on the CheXpert (CXP) [7] dataset over four sessions, showing the

Ground Truth	Task 1 Top-4 Prediction Score	Task 2 Top-4 Prediction Score	Task 3 Top-4 Prediction Score	Task 4 Top-4 Prediction Score
Pleural Effusion	**Pleural Effusion:0.59**	**Pleural Effusion:0.56**	**Pleural Effusion:0.52**	**Pleural Effusion:0.64**
	Edema: 0.15	Cardiomegaly: 0.17	Consolidation: 0.18	Pneumonia:0.23
	Lung Opacity: 0.15	Edema: 0.15	Atelectasis:0.17	Consolidation: 0.07
	No Finding: 0.01	Enlarged_Cardio: 0.12	Cardiomegaly:0.13	Edema: 0.06

Figure 1: Example of prior knowledge retention in our methodology. An image from the CheXpert (CXP) dataset with the ground truth label Pleural Effusion is shown. The model, trained in a CIL setting, predicts based on the highest confidence score. The top-1 predicted labels after Task 1 to Task 4 are highlighted in bold, demonstrating consistent correct predictions across all tasks.

model's ability to retain knowledge of the "Pleural Effusion" class despite learning new ones, as it consistently predicts the correct label with the highest score.

Recent CIL efforts are broadly categorized as non-prompt-based and prompt-based methods. Non-prompt-based approaches, including replay-based [2, 23] and regularization-based [15] methods, either require storing raw images or exhibit linearly increasing memory demands with more tasks [34]. Conversely, prompt-based CIL methods utilize lightweight learnable parameters (prompts) to guide task-specific adaptation. However, most existing prompt-based methods [25, 32] heavily depend on pretrained models for learning representative embeddings, making them less suitable for medical image diagnosis.

To address these limitations, we propose a prompt-driven generative replay framework specifically for class-incremental chest x-ray (CXR) diagnosis. Our prompt-based network adapts to new data by learning small, task-specific prompts, conserving memory, and reducing overfitting. We combine prompt tuning with distribution-guided generative replay, using a Variational Autoencoder (VAE) to model the latent feature distribution and synthesize representative features for past classes. By freezing the encoder after the initial session for stability and adapting the synthesized features via learnable prompts and an MLP for plasticity, our method effectively balances knowledge retention and adaptation.

Following recent findings on the utility of semantic priors [18], we design prompts that are semantically rich. We initialize the prompt using the BioBERT model [12], trained on biomedical texts, to align learning with radiological knowledge. During training, prompt refinement captures semantic cues for class-discriminative CXR features. Once learned, prompts are stored to preserve knowledge across tasks. Our design consists of three modules:

1. **Distribution & Prompt Learner (DPL) Network:** A VAE that models the distribution of input images, and helps in learning class-specific, semantically rich prompts.
2. **Classification Network:** A lightweight module that maps input features to the query space and retrieves the most relevant class-specific prompt using cosine similarity.
3. **Generative Replay Network:** During each incremental session, synthetic features of old classes are generated using stored latent parameters and class prompts, eliminating the need to retain original training data.

In the initial session, we train the VAE to capture the data distribution and learn prompts using real data from a subset of classes called the base class. In each subsequent session, new class prompts are learned using real images, while synthetic feature replay for old classes

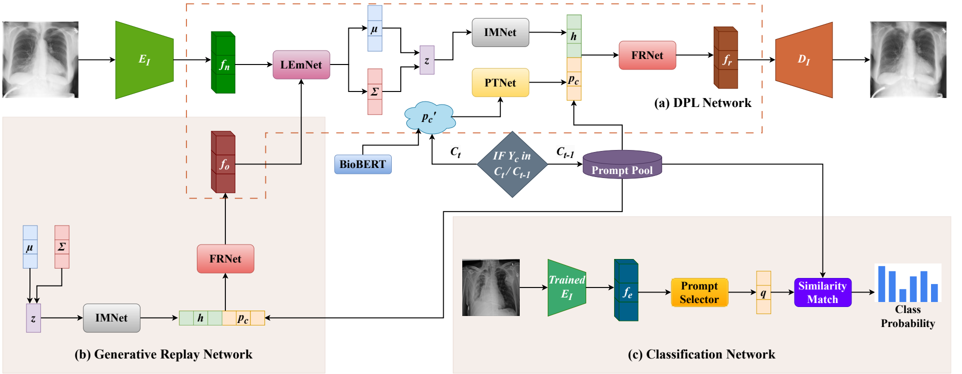

Figure 2: The block diagram of our method with an image encoder E_I, and decoder D_I. (a) The Distribution & Prompt Learner (DPL) network is built on VAE. It learns class-specific prompts (p_c) and captures the data distribution using LEmNet (latent distribution μ,Σ), IMNet (intermediate space projection), PTNet (semantic embedding to prompt), and FRNet (feature reconstruction via intermediate features and prompts). After initial training, the image encoder (E_I) and LEmNet are frozen to retain past knowledge, while other components adapt to new classes. (b) To preserve past knowledge, synthetic features are generated using stored prompts (from Prompt Pool $\mathcal{P}$) and the frozen latent distribution. (c) Classification involves computing the cosine similarity between a query vector q and learned prompts p_c.

prevents forgetting. Unlike image replay, the generative feature replay learns a simplified feature distribution [28]. The encoder is frozen after the initial session to avoid distributional drift across tasks, while prompts and intermediate representations are updated to accommodate new classes. The prompt selector is progressively fine-tuned to match queries with prompts from an expanding pool. Our contributions are as follows:

- We propose a unified framework for class-incremental CXR diagnosis that combines trained semantic prompt tuning and distribution-guided generative replay.
- We develop a novel feature-level replay mechanism that synthesizes discriminative feature representations from stored latent distributions conditioned on class-specific prompts, thus avoiding storage of raw images.
- We introduce a classification network that dynamically retrieves class-relevant prompts, enabling accurate classification across incremental sessions.
- Our method outperforms both prompt-based and non-prompt-based baselines in retaining old knowledge and learning new disease classes on benchmark CXR datasets.

The remainder of this paper is structured as follows: Section 2 discusses related work in CXR diagnosis. Section 3 details the proposed approach. Section 4 presents the experimental setup and results. Finally, Section 5 concludes the paper.

2 Related Works

Class-incremental learning allows models to learn new classes sequentially while retaining prior knowledge. Existing strategies include replay-based methods [8, 24] that store or generate past data, regularization-based approaches [5, 33] that constrain weight updates, dynamic expansion methods [4, 17] that adapt model capacity, and prompt-based methods like [31] and [30] (a subset of dynamic expansion) that learn visual prompts.

While effective generally, these methods have limited application in chest x-ray (CXR) diagnosis, where incremental learning mainly addresses domain adaptation with a fixed class set but shifting data distribution [13, 20]. To our knowledge, only Akundi et al. [1] proposed a CIL method for CXR, but it evaluated only five single-class tasks, limiting scalability. Furthermore, most replay methods require raw data access, raising privacy concerns and optimization challenges in medical contexts. Dynamic expansion increases memory demands, and prompt-based approaches often need large pre-trained models for effective initialization [27], which may not be available. These limitations hinder the direct use of current CIL techniques in privacy-sensitive domains like healthcare.

Our approach addresses these shortcomings. It is one of the first approaches to investigate CIL in a general multi-class setting with a large number of classes for CXR diagnosis. By integrating generative replay with prompt tuning, we aim to alleviate catastrophic forgetting while protecting patient privacy.

3 Methodology

In class-incremental learning (CIL), a key challenge is balancing the retention of past knowledge with the effective learning of new information. Preserving prior knowledge requires stable representations of old data as new tasks emerge. Concurrently, the model must adapt to new tasks by developing flexible representations for new classes as their data becomes available. Utilizing semantic information about classes can aid in this learning process. To this end, we propose a modular framework combining prompt tuning and generative replay, employing a Distribution & Prompt Learner (DPL) module based on a Variational Autoencoder [11] (VAE) to capture the data distribution in the latent space. This framework enables the generation of representative features for old classes without the need to store raw images.

Our model is trained through multiple sessions. In Session 0, the model is trained on a subset of classes (base classes). These base classes are used to learn the data distribution and semantically rich prompt embeddings corresponding to the base classes. In subsequent sessions, we introduce new classes called incremental classes to learn prompts corresponding to new classes. The process of learning the distribution of data is frozen after Session 0. This facilitates the retention of past knowledge. Simultaneously, allowing the IMNet to learn continuously, along with the semantically rich embeddings (prompts) from the PTNet, allows the model to adapt to new classes. This design choice is a necessary trade-off to prevent catastrophic forgetting. These prompts are initialized through semantic embedding and refined through training. Once the prompts are learned, we store them so that they do not lose the learned class information. During inference, a classification module finds the most similar prompt for a given input using cosine similarity.

3.1 Model Architecture

Our method consists of an image encoder that extracts high-dimensional feature representations from input images, a DPL network built on a VAE architecture to learn semantically rich class-specific prompts and capture the underlying data distribution, and an image decoder to reconstruct the input image. A generative replay network is added to generate data of past classes, while a prompt pool maintains a memory of learned prompts. Finally, a classification network is introduced for inference through prompt selection. Figure 2 provides an overview of the full pipeline. The complete pipeline consists of the following components.

Image Encoder: We encode input x-ray images X using a ResNet-18-based image encoder to generate a high-dimensional feature representation (FR) $f_n \in \mathbb{R}^{h \times w \times c}$. This FR serves as input to both the DPL module and the classification module. The encoder is trained only on the base classes. We assume that the base class consists of a sufficiently large diversity of data. The weights of the image encoder are frozen afterward. This preserves stable visual representations across tasks, facilitating retention of past knowledge.
Distribution & Prompt Learner (DPL) Network: CIL involves multi-phase training. The initial phase focuses on learning the data distribution and semantic prompts. Subsequent sessions learn prompts for new classes, initialized using BioBERT embeddings [12] and refined during training via the DPL network. Initialization of prompts using the BioBERT model, trained on biomedical corpora, helps in making the prompts semantically rich. The refinement of prompts during training further helps capture semantic information. DPL network comprises four MLPs: the Latent Embedding Network (LEmNet) uses two parallel MLPs to estimate the mean (μ) and diagonal covariance (Σ) of the base class features; both MLPs reduce the input feature to a lower-dimensional latent space. The Intermediate Network (IMNet) is a two-layer MLP that transforms a sampled latent vector z (drawn using μ and Σ) into an intermediate representation h of the same dimensionality as the prompt. The Prompt Tuning Network (PTNet) takes initial BioBERT-derived prompts $p_c^{'}$ and refines them into class-specific prompts p_c of the same dimension through a residual MLP. Finally, the Feature Reconstruction Network (FRNet) concatenates h and p_c and processes the combined vector through a three-layer MLP to reconstruct the original feature f_n as f_r.
Image Decoder: The image decoder D_I reconstructs the chest x-ray images X' from the reconstructed feature representation f_r. This module helps in creating better feature representation by enforcing a good reconstruction during training with base classes.
Generative Replay Network: To preserve learned knowledge, we generate some exemplars to train the model in conjunction with the new data in future training sessions. The generative replay network, a VAE decoder without PTNet, is used because prompt learning is not needed during replay. It samples data from learned distribution parameters (obtained from LEmNet). The sampled data is then combined with stored prompts (from the prompt pool) to generate class-discriminative synthetic features for old classes.
Classification Network: For inference, we have to map an image to its correct prompt from the pool. The classification network encodes the image using the image encoder E_I into a feature representation f_e. The feature representation is then encoded into a query vector q to select the best matching prompt using cosine similarity. The query prompt is compared against all prompts in the prompt pool to identify the closest match p_c. The class label Y_c associated with p_c is then assigned as the predicted label for the query sample.
Prompt Pool: We introduce a memory buffer to store class-specific prompts p_c and their labels Y_c. The prompts are learned in the learning phase and are then used by generative

replay to generate exemplars and during inference of query data.

3.2 Training

We train our model for a CIL setting through multiple sessions. In each session, we introduce new classes. To retain knowledge of previous classes, we synthetically generate data from the previous class using distribution-guided generative replay. In the incremental learning phase, the model is trained with the data of new classes and synthetic data from the old classes. We divide the dataset into T sessions. Let $\mathcal{D}_t = \{(X_a, Y_a)\}_{a=1}^{A_t}$ denote the dataset at session t, where $Y_a \in \mathcal{C}_t$. Session 0 is trained on base class set $\mathcal{C}_0$. In each subsequent session $t \geq 1$, new classes $\mathcal{C}_t$ are introduced. The model is updated to classify the entire set of seen classes $\mathcal{C}_{\leq t} = \bigcup_{k=0}^{t} \mathcal{C}_k$ without revisiting data from previous sessions.

3.2.1 Session 0: Learning the Base Classes

We begin with an empty prompt pool. Assuming a diverse base class dataset, the initial session involves the model learning the data distribution and prompts for base classes $\mathcal{C}_0$, as well as the image encoder weights. Base class training starts with prompt initialization: for each class $c \in \mathcal{C}_0$, a prompt vector $p_c^{'} \in \mathbb{R}^{\beta}$ is generated from the BioBERT embedding of the class text label. The ImageNet-pretrained image encoder E_I (ResNet-18) is fine-tuned on base class chest x-ray images to produce feature representations (FR) f_n for input images X. These f_n are then encoded into latent parameters (μ, Σ) using LEmNet, followed by latent vector sampling via the reparameterization trick [22].

The generated sample is first passed through the IMNet to get the intermediate representation h. If the input image belongs to the class c, we select the prompt $p_c^{'}$ from the prompts initialized at the beginning. The prompt $p_c^{'}$ is passed through the PTNet to learn the class-specific prompt p_c. The intermediate representation h is concatenated with the class-specific prompt p_c, expected to add class-specific features. The concatenated vector is then passed through the FRNet to reconstruct the FR f_r. To ensure that we learn meaningful data distribution parameters, the FR f_n and f_r should be the same. Adhering to the autoencoding principle, the reconstructed FR f_r is further decoded to produce the image reconstruction $X^{'}$. The image encoder and the DPL network are then trained to minimize the loss:

$$\mathcal{L}_{\text{base}} = \|X - X^{'}\|_1^2 + \lambda_1 \|f_n - f_r\|_2^2 + \lambda_2 D_{\text{KL}}(\mathcal{N}(\mu, \Sigma) \parallel \mathcal{N}(0, I)), \tag{1}$$

where, $D_{\text{KL}}(\alpha, \gamma)$ is the KL divergence [6] between two probability distributions. This training provides a class-specific prompt p_c corresponding to each class in session 0.

After initial training, the learned parameters $\{\mu, \Sigma\}$ are stored. The learned class-specific prompts and their labels Y_c are also saved in the prompt pool $\mathcal{P}$ ($\mathcal{P} = \mathcal{P} \bigcup \{Y_c, p_c\}$). The image decoder is discarded, and the weights of the image encoder and LEmNet are frozen to prevent catastrophic forgetting and stabilize feature representations during incremental learning. Freezing the VAE encoder ensures a consistent feature space for base classes, crucial for effective generative replay and prompt tuning, which rely on stable feature distributions. As demonstrated in the ablation study, changes to the encoder would render replayed features incompatible with the current output distribution. Consequently, the adaptation to new tasks/classes is primarily handled by the learned class-specific prompts.

We then train the classification network, which maps the encoded image features to a query prompt q that aligns with the correct class prompt p_c in the prompt pool. Using cosine similarity, the network encourages q to be close to p_c and distant from the prompts of other classes. The training objective is $\mathcal{L}_{\text{prompt}} = 1 - \frac{q \cdot p_c}{\|q\| \cdot \|p_c\|}$. The process is also outlined in the Algorithm given in the supplementary material.

3.2.2 Sessions 1 to T-1: Learning New Classes

In each session, new classes are learned while preserving performance on old ones via generative replay. Initially, prompts are generated using BioBERT. For new classes, their image features $\mathcal{F}_n$ are extracted with a frozen encoder. For old classes, synthetic features $\mathcal{F}_o$ are generated by sampling from learned $\{\mu, \Sigma\}$ and passing these samples along with stored prompts through IMNet and FRNet ($\mathcal{V}$ synthetic features per prompt). The DPL network is then fine-tuned on both $\mathcal{F}_n$ and $\mathcal{F}_o$. This process learns new class prompts and retains knowledge of old classes.

The combined feature $f \in \{\mathcal{F}_n \bigcup \mathcal{F}_o\}$ is mapped to latent distribution parameters (μ, Σ) by *LEmNet*, and a latent vector z is sampled using the reparameterization trick [22]. *IMNet* then processes z into an intermediate representation h. For a new class $c \in \mathcal{C}_t$, its initial prompt $p_c^{'}$ is refined by *PTNet* into p_c, which is then concatenated with h. For an old class $c \in \mathcal{C}_{t-1}$, its stored prompt p_c is directly concatenated with h, bypassing *PTNet* to prevent relearning. To mitigate catastrophic forgetting, distribution parameters and old class prompts are frozen. The combined vector $\{h, p_c\}$ is fed into *FRNet* to reconstruct the original feature f_r. The DPL network is trained by minimizing the reconstruction loss: $\mathcal{L}_{\text{new}} = \|f - f_r\|_2^2, \forall f \in \{f_n, f_o\}$. This training process yields learned, class-specific prompts p_c for new classes present in the session t. These prompts are then stored in the prompt pool.

After fine-tuning the DPL network, the learned prompt p_c and its class label Y_c for each new class are added to the prompt pool $\mathcal{P}$ ($\mathcal{P} = \mathcal{P} \cup \{Y_c, p_c\}$). Subsequently, the classification network is updated to include these new classes. For a new image of class $c \in \mathcal{C}_t$, a query prompt q is generated using the image encoder and the classification network. The classification network is then fine-tuned using the cosine similarity loss to correctly link these query prompts with their corresponding class-specific prompts in the expanded prompt pool.

The training continues in this manner over all T sessions. The complete training procedure for classes added in incremental sessions is also outlined in the Algorithm given in the supplementary material. This completes our methodology for class-incremental learning using prompt tuning and generative replay for chest x-ray diagnosis.

3.3 Inference

At test time, a query image x is passed through the image encoder to get the feature representation. The feature representation is processed through the classification network to get the query prompt. The model compares the query prompt q with each class prompt $p_c \in \mathcal{P}$ from the prompt pool using cosine similarity. The prediction score for class c is then computed using a softmax function over the similarity scores. The class with the highest score is assigned as the predicted label.

Table 1: Comparative performance in terms of average accuracy (A_{avg}, mean $\pm$ sd for five runs) and Performance Drop (PD, mean $\pm$ sd for five runs) for different approaches on the NIH, CXR, and MCXR datasets .

Method	NIH		CXP		MCXR	
	$A_{\text{avg}} \uparrow$	$PD \downarrow$	$A_{\text{avg}} \uparrow$	$PD \downarrow$	$A_{\text{avg}} \uparrow$	$PD \downarrow$
Replay	70.48 ± 2.5	14.2 ± 1.2	66.70 ± 1.9	11.7 ± 1.1	65.88 ± 1.7	15.2 ± 0.9
LwF	72.03 ± 2.9	10.8 ± 1.5	64.93 ± 1.7	**11.4 ± 1.1**	67.40 ± 1.9	15.8 ± 0.8
DER	66.15 ± 2.8	13.4 ± 1.2	62.35 ± 2.1	15.6 ± 1.9	60.25 ± 2.5	19.3 ± 2.1
L2P	65.30 ± 4.1	28.6 ± 2.3	63.05 ± 3.8	21.7 ± 1.7	62.48 ± 4.1	20.2 ± 2.7
CodaP	66.05 ± 4.6	25.7 ± 2.6	64.35 ± 3.2	19.3 ± 1.6	61.10 ± 4.0	23.3 ± 2.5
ILCXR	71.13 ± 2.2	11.2 ± 1.5	65.73 ± 1.8	11.9 ± 1.0	67.87 ± 1.8	15.6 ± 1.3
CAPrompt	65.75 ± 4.3	27.3 ± 2.4	65.15 ± 3.6	20.1 ± 1.2	62.91 ± 4.2	22.1 ± 2.6
Proposed	**74.70 ± 2.3**	**10.8 ± 0.7**	**69.48 ± 1.2**	12.1 ± 1.0	**68.50 ± 2.1**	**13.3 ± 1.3**

4 Experiments and Results

4.1 Implementation Details

Datasets: We conduct experiments on three publicly available datasets: NIHChestXray14 [29] (NIH), CheXpert [7] (CXP), and MIMIC-CXR [9, 10, 21] (MCXR). NIH contains 91,234 images with single labels across 15 categories. CXP includes 50,223 single-labeled images in 14 categories. MCXR has 187,020 images labeled with the same 14 categories as CXP. For each dataset, we use a 70:10:20 split for training, validation, and testing, employing iterative stratification to maintain label balance. We only use frontal view images.

Evaluation Protocol: Following prior CIL work [16, 19], we define Top-1 accuracy after task t across all tasks 0 to t as A_t. The average accuracy $A_{\text{avg}} = \frac{\sum_{t=0}^{T-1} A_t}{T}$ reflects overall incremental learning performance. Forgetting is quantified by the performance drop $PD = A_0 - A_{T-1}$, the accuracy difference between the initial (task 0) and final (task $T-1$) tasks.

Training Details: For experiments, we use PyTorch on NVIDIA RTX A6000 GPUs. Input images are resized to 224×224 and the model is trained for 100 epochs with the Adam optimizer (initial learning rate 0.001, prompt token length $\beta = 256$, chosen by validation performance). The base session dataset $|\mathcal{D}_0|$ is designed to have at least as many samples as any incremental session $|\mathcal{D}_t|$ ($t \geq 1$) to ensure data diversity. For NIH, 6 base classes are used; for CXP and MCXR, 5 classes are used. In all datasets, 3 new classes are added incrementally across $T = 4$ sessions. Different runs involve varying the base classes.

4.2 Performance Comparisons

We compare the performance of the proposed method with several non-prompt-based CIL methods, including Replay [23], LwF [15], DER [2], and ILCXR [1]. We also compare with SOTA prompt-based CIL approaches, including L2P [32], CodaP [25], and CAPrompt [14]. For each experiment, we conduct five runs and calculate the mean and standard deviation of each metric over the five runs. The results are presented in Table 1.

Notice that our method typically outperforms others. Prompt-based CIL approaches exhibit greater performance reductions and lower average accuracy. Non-prompt-based methods have lower average accuracy but smaller decrements. Figure 4's task-wise accuracy visualization on chest x-ray datasets suggests prompt-based methods decline due to inadequate adaptation to novel classes, while non-prompt-based methods degrade due to forgetting

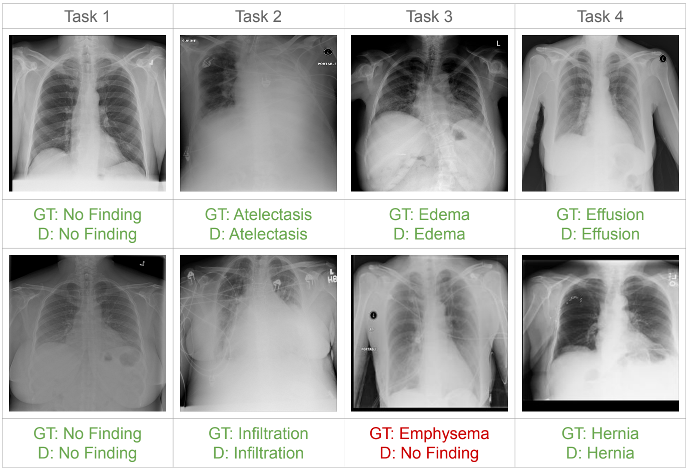

Figure 3: Classification results of the proposed model on sample images from the NIH dataset after each task, using the top-1 prediction score (GT: Ground truth, D: Prediction, Green: Correct, Red: Incorrect).

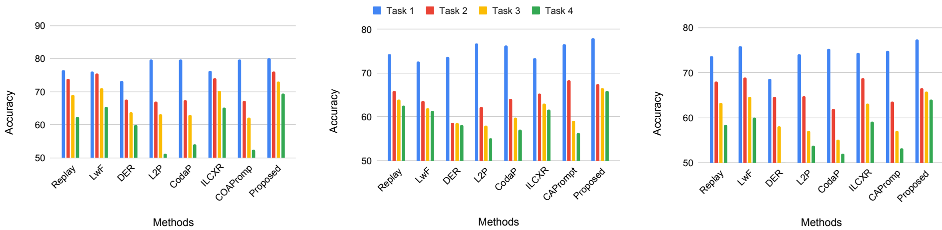

Figure 4: Comparison of task-wise accuracy across sessions for different methods on chest x-ray images from the NIH (left), CXP (middle), and MCXR (right) datasets.

old classes. Our method addresses these problems by effectively learning new classes and preserving prior knowledge, resulting in superior average accuracy and smaller performance drops.

Figure 3 shows sample classification results of our model on the NIH dataset in a class-incremental learning (CIL) setting based on the highest (top-1) score. The top row demonstrates consistent correct classification across all tasks, indicating successful knowledge retention. The bottom row shows a misclassification in Task 3, highlighting the challenge of catastrophic forgetting.

4.3 Ablation Studies

Next, we evaluate the importance of different components of the proposed method. First, we look at the role of generative replay. For that, the DPL network in session 1 to $T-1$ is trained using only the data of new classes (abbreviated as WGR), in contrast to the proposed method, where FR is generated using the stored prompt is also used. The results reported in Table 2 show a performance drop with WGR in comparison with the proposed method. This

Table 2: Performance comparison of the proposed method under three configurations: (1) without generative replay (WGR), (2) without freezing encoder weights (WFEW), and (3) with prompts initialized with random prompts, evaluated on the NIH, CXR, and MCXR datasets for incremental learning .

Configuration	NIH		CXP		MCXR	
	$A_{\mathbf{avg}} \uparrow$	$PD \downarrow$	$A_{\mathbf{avg}} \uparrow$	$PD \downarrow$	$A_{\mathbf{avg}} \uparrow$	$PD \downarrow$
WGR	72.92 ± 2.2	11.7 ± 1.5	66.73 ± 2.1	12.7 ± 1.1	66.42 ± 1.6	13.9 ± 1.6
WFEW	71.24 ± 4.1	12.7 ± 2.1	67.28 ± 1.3	13.1 ± 1.1	65.19 ± 1.9	13.9 ± 1.4
WRPI	73.42 ± 5.3	12.1 ± 1.7	68.17 ± 4.1	12.6 ± 1.2	67.15 ± 3.3	15.3 ± 2.3
Proposed	$\mathbf{74.70 \pm 2.3}$	$\mathbf{10.8 \pm 0.7}$	$\mathbf{69.48 \pm 1.2}$	$\mathbf{12.1 \pm 1.0}$	$\mathbf{68.50 \pm 2.1}$	$\mathbf{13.3 \pm 1.3}$

shows the utility of generative replay in alleviating catastrophic forgetting.

Next, we evaluate the importance of freezing the weights of the VAE encoder layers. For this, we allow weight update in LEmNet and learn the parameters μ and Σ in the incremental learning phase, keeping the image encoder E_I frozen (abbreviated as WFEW). From Table 2, notice that the performance with WFEW drops by more than 3% compared to the proposed method. Thus, we find that freezing the weights of the encoder helps in alleviating catastrophic forgetting by reducing the effect of distribution shift across different tasks.

Finally, we evaluate the importance of initializing prompts with semantic embeddings from the BioBERT model. For this, we initialize the prompts with random vectors for both the base and incremental classes. We ensure that these prompts are unique. The results (abbreviated as WRPI) are reported in Table 2. Notice that the performance with WRPI drops by about 1% compared to the proposed method, while the standard deviation increases by a factor of 2. Thus, by initializing the prompts with a semantically rich embedding, we can reduce the randomness in the model training.

5 Conclusion

We introduce a novel CIL method that combines distribution-guided generative replay with semantic prompts by integrating prompt tuning with a VAE-based framework. Our method effectively mitigates catastrophic forgetting while preserving privacy by operating in feature space. Experiments on multiple benchmark CXR datasets show consistent improvements over existing CIL approaches in both classification accuracy and knowledge retention. Ablation studies confirm the importance of distribution alignment, feature replay, and semantic prompts. While our approach is robust, it faces limitations common to CIL, such as performance dependency on the base class selection, which our experiments addressed by ensuring base class diversity. Similarly, freezing the image encoder after the initial session, while crucial for stabilizing feature representations and mitigating catastrophic forgetting, may limit the model's ability to adapt to new classes with vastly different visual features. Despite these limitations, our work presents a foundation for future research. We plan to explore a dynamic prompt growth mechanism to improve scalability as class counts increase. We will also investigate a conditional update of the encoder network to better capture evolving feature spaces while maintaining knowledge retention. Additionally, we aim to validate the generalizability of our learned prompts to other tasks and modalities. Finally, we will explore the use of more powerful backbones, such as Vision Transformers.

Acknowledgments

The authors thank the National Institute of Health Clinical Center, Stanford ML Group, and MIT Laboratory for Computational Physiology for providing the chest x-ray datasets.

References

[1] Prathyusha Akundi and Jayanthi Sivaswamy. Incremental learning for a flexible cad system design. In *2022 IEEE 19th International Symposium on Biomedical Imaging (ISBI)*, pages 1–4. IEEE, 2022.

[2] Pietro Buzzega, Matteo Boschini, Angelo Porrello, Davide Abati, and Simone Calderara. Dark experience for general continual learning: a strong, simple baseline. *Advances in neural information processing systems*, 33:15920–15930, 2020.

[3] Sandip Chakraborty, Deepak Chandran, Ranjan K Mohapatra, Md Aminul Islam, Mahmoud Alagawany, Manojit Bhattacharya, Chiranjib Chakraborty, and Kuldeep Dhama. Langya virus, a newly identified henipavirus in china - zoonotic pathogen causing febrile illness in humans, and its health concerns: Current knowledge and counteracting strategies - correspondence. *Int J Surg*, 105:106882, September 2022.

[4] Arthur Douillard, Alexandre Ramé, Guillaume Couairon, and Matthieu Cord. Dytox: Transformers for continual learning with dynamic token expansion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9285–9295, 2022.

[5] Yiduo Guo, Wenpeng Hu, Dongyan Zhao, and Bing Liu. Adaptive orthogonal projection for batch and online continual learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6783–6791, 2022.

[6] John R Hershey and Peder A Olsen. Approximating the kullback leibler divergence between gaussian mixture models. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pages IV–317. IEEE, 2007.

[7] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris 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*, volume 33, pages 590–597, 2019.

[8] Ahmet Iscen, Jeffrey Zhang, Svetlana Lazebnik, and Cordelia Schmid. Memory-efficient incremental learning through feature adaptation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVI 16*, pages 699–715. Springer, 2020.

[9] A Johnson, T Pollard, R Mark, S Berkowitz, and Steven Horng. Mimic-cxr database (version 2.0. 0). *PhysioNet*, 10:C2JT1Q, 2019.

[10] Alistair EW Johnson, Tom J Pollard, Seth J Berkowitz, Nathaniel R Greenbaum, Matthew P Lungren, Chih-ying Deng, Roger G Mark, and Steven Horng. Mimic-cxr,

a de-identified publicly available database of chest radiographs with free-text reports. *Scientific data*, 6(1):317, 2019.

[11] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013.

[12] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[13] Matthias Lenga, Heinrich Schulz, and Axel Saalbach. Continual learning for domain adaptation in chest x-ray classification. In *Medical Imaging with Deep Learning*, pages 413–423. PMLR, 2020.

[14] Qiwei Li and Jiahuan Zhou. Caprompt: Cyclic prompt aggregation for pre-trained model based class incremental learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 18421–18429, 2025.

[15] Zhizhong Li and Derek Hoiem. Learning without forgetting. *IEEE transactions on pattern analysis and machine intelligence*, 40(12):2935–2947, 2017.

[16] Chenxi Liu, Zhenyi Wang, Tianyi Xiong, Ruibo Chen, Yihan Wu, Junfeng Guo, and Heng Huang. Few-shot class incremental learning with attention-aware self-adaptive prompt. In *European Conference on Computer Vision*, pages 1–18. Springer, 2024.

[17] Yaoyao Liu, Bernt Schiele, and Qianru Sun. Adaptive aggregation networks for class-incremental learning. In *Proceedings of the IEEE/CVF conference on Computer Vision and Pattern Recognition*, pages 2544–2553, 2021.

[18] Devi Prasad Maharathy, Prabhala Sandhya Gayatri, and Angshuman Paul. Attention-based few-shot diagnosis of chest x-rays using semantic signatures. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 3199–3204. IEEE, 2024.

[19] Zichong Meng, Jie Zhang, Changdi Yang, Zheng Zhan, Pu Zhao, and Yanzhi Wang. Diffclass: Diffusion-based class incremental learning. In *European Conference on Computer Vision*, pages 142–159. Springer, 2024.

[20] Vuong D Nguyen, Samiha Mirza, Abdollah Zakeri, Ayush Gupta, Khadija Khaldi, Rahma Aloui, Pranav Mantini, Shishir K Shah, and Fatima Merchant. Tackling domain shifts in person re-identification: A survey and analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4149–4159, 2024.

[21] PhysioToolkit PhysioBank. Physionet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):e215–e220, 2000.

[22] Lucas Pinheiro Cinelli, Matheus Araújo Marins, Eduardo Antúnio Barros da Silva, and Sérgio Lima Netto. Variational autoencoder. In *Variational methods for machine learning with applications to deep networks*, pages 111–149. Springer, 2021.

[23] David Rolnick, Arun Ahuja, Jonathan Schwarz, Timothy Lillicrap, and Gregory Wayne. Experience replay for continual learning. *Advances in neural information processing systems*, 32, 2019.

[24] Hanul Shin, Jung Kwon Lee, Jaehong Kim, and Jiwon Kim. Continual learning with deep generative replay. *Advances in neural information processing systems*, 30, 2017.

[25] James Seale Smith, Leonid Karlinsky, Vyshnavi Gutta, Paola Cascante-Bonilla, Donghyun Kim, Assaf Arbelle, Rameswar Panda, Rogerio Feris, and Zsolt Kira. Coda-prompt: Continual decomposed attention-based prompting for rehearsal-free continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11909–11919, 2023.

[26] Shriyansh Srivastava, Dheeraj Sharma, Sathvik Belagodu Sridhar, Sachin Kumar, Rakesh Sahu, Sumel Ashique, Aroop Mohanty, Sanjit Sah, Vini Mehta, Rachana Mehta, and Vasso Apostolopoulos. HKU5-CoV-2: A silent threat looming in bats. *Travel Med Infect Dis*, 65:102838, March 2025.

[27] Yu-Ming Tang, Yi-Xing Peng, and Wei-Shi Zheng. When prompt-based incremental learning does not meet strong pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1706–1716, 2023.

[28] Kevin Thandiackal, Tiziano Portenier, Andrea Giovannini, Maria Gabrani, and Orcun Goksel. Generative feature-driven image replay for continual learning. *Image and Vision Computing*, 150:105187, 2024.

[29] Xiaosong Wang, Yifan Peng, Le Lu, Zhiyong Lu, Mohammadhadi Bagheri, and Ronald M Summers. 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*, pages 2097–2106, 2017.

[30] Yabin Wang, Zhiwu Huang, and Xiaopeng Hong. S-prompts learning with pre-trained transformers: An occam's razor for domain incremental learning. *Advances in Neural Information Processing Systems*, 35:5682–5695, 2022.

[31] Zifeng Wang, Zizhao Zhang, Sayna Ebrahimi, Ruoxi Sun, Han Zhang, Chen-Yu Lee, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, et al. Dualprompt: Complementary prompting for rehearsal-free continual learning. In *European conference on computer vision*, pages 631–648. Springer, 2022.

[32] Zifeng Wang, Zizhao Zhang, Chen-Yu Lee, Han Zhang, Ruoxi Sun, Xiaoqi Ren, Guolong Su, Vincent Perot, Jennifer Dy, and Tomas Pfister. Learning to prompt for continual learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 139–149, 2022.

[33] Guanxiong Zeng, Yang Chen, Bo Cui, and Shan Yu. Continual learning of context-dependent processing in neural networks. *Nature Machine Intelligence*, 1(8):364–372, 2019.

[34] Da-Wei Zhou, Qi-Wei Wang, Zhi-Hong Qi, Han-Jia Ye, De-Chuan Zhan, and Ziwei Liu. Class-incremental learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.