


Experience Replay and Zero-Shot Clustering for Continual Learning in Diabetic Retinopathy Detection

Gussepe Bravo-Rocca¹ ^a, Peini Liu¹ ^b, Jordi Guitart^{1,2} ^c, Ajay Dholakia³ ^d,
David Ellison³ ^e and Rodrigo M. Carrillo-Larco⁴ ^f

¹Barcelona Supercomputing Center, Barcelona, Spain

²Universitat Politècnica de Catalunya, Barcelona, Spain

³Lenovo Infrastructure Solutions Group, Morrisville, NC, U.S.A.

⁴Emory University, GA, U.S.A.

Keywords: Zero-Shot Clustering, Experience Replay, Diabetic Retinopathy Detection, Privacy-Preserving Learning, Medical Imaging.


Abstract: We present an approach to mitigate catastrophic forgetting in Continual Learning (CL), focusing on domain incremental scenarios in medical imaging. Our method leverages Large Language Models (LLMs) to generate task-agnostic descriptions from multimodal inputs, enabling zero-shot clustering of tasks without supervision. This clustering underpins an enhanced Experience Replay (ER) strategy, strategically sampling data points to refresh the model’s memory while preserving privacy. By incrementally updating a multi-head classifier using only data embeddings, our approach maintains both efficiency and data confidentiality. Evaluated on a challenging diabetic retinopathy dataset, our method demonstrates significant improvements over traditional CL techniques, including Elastic Weight Consolidation (EWC), Gradient Episodic Memory (GEM), and Learning Without Forgetting (LWF). Extensive experiments across Multi-Layer Perceptron (MLP), Residual, and Attention architectures show consistent performance gains (up to 3.1% in Average Mean Class Accuracy) and reduced forgetting, with only 6% computational overhead. These results highlight our approach’s potential for privacy-preserving, efficient CL in sensitive domains like healthcare, offering a promising direction for developing adaptive AI systems that can learn continuously while respecting data privacy constraints.


1 INTRODUCTION


Continual Learning (CL) aims to develop AI systems capable of acquiring and refining knowledge over time, mirroring human-like adaptive learning (Parisi et al., 2019). Unlike conventional Machine Learning approaches that separate training and inference phases, CL models must adapt to evolving real-world data and tasks (Wang et al., 2024). This adaptation is crucial in scenarios with blurred task boundaries (Koh et al., 2022) and in environments requiring continuous learning without catastrophic forgetting (De Lange et al., 2022; Kirkpatrick et al., 2017; Robins, 1993).


A significant challenge in CL is maintaining consistent performance across changing multimodal data distributions, particularly in domains like medical imaging where privacy concerns and data mutability are paramount. Domain Incremental Learning (DIL) faces acute challenges with domain shifts, such as variations in lighting, population characteristics, or noise in medical image classifiers. Traditional retraining approaches are often infeasible due to privacy constraints in healthcare (Kumar and Srivastava, 2018; Kumari and Singh, 2024; Lenga et al., 2020), leading to performance degradation on previously learned tasks (Khan et al., 2024; Kuang et al., 2018).


To address these challenges, we present a novel unsupervised learning framework that leverages Large Language Models (LLMs) for CL in privacy-sensitive domains. As shown in Figure 1, our approach uses LLMs to generate textual descriptions from multimodal inputs (images, labels), enabling


^a  <https://orcid.org/0000-0001-6824-1124>

^b  <https://orcid.org/0000-0003-0058-8732>

^c  <https://orcid.org/0000-0003-0751-3100>

^d  <https://orcid.org/0009-0007-8973-6063>

^e  <https://orcid.org/0000-0002-0752-5569>

^f  <https://orcid.org/0000-0002-2090-1856>

zero-shot clustering without predefined task boundaries. That is, once we get the LLM-generated descriptions, we map them to embeddings. On the same space, we compare these embeddings with the images' embeddings to perform the clustering. This method is particularly well-suited for medical imaging scenarios, such as Diabetic Retinopathy (DR) detection from fundus images, where data privacy and distribution shifts are critical concerns.

Our approach extends the concept of ER (Riemer et al., 2019) by integrating a strategic sampling methodology derived from zero-shot clusters. This new ER strategy refreshes the neural network's memory, mitigating knowledge degradation across tasks. The model architecture employs a multi-head classifier that expands incrementally with new tasks, each head containing a simple linear layer for adaptation.

Key features of our approach include:

- **Privacy Preservation.** By operating on embeddings rather than raw images, our method addresses critical privacy concerns in sensitive domains like healthcare.
- **Resource Efficiency.** Designed to function on CPUs using embeddings, our approach is computationally efficient and suitable for deployment in resource-constrained environments.
- **Adaptability.** The ER strategy, free from fixed task definitions, improves adaptability to evolving data distributions in medical imaging scenarios (Zhang et al., 2024; Serra et al., 2018).
- **Foundation Model Integration.** We leverage CLIP (Radford et al., 2021) to produce robust image embeddings, enhancing the zero-shot clustering process.

Our method seamlessly integrates with and enhances established CL strategies, including Elastic Weight Consolidation (EWC) (Kirkpatrick et al., 2017), Gradient Episodic Memory (GEM) (Lopez-Paz and Ranzato, 2017), and Learning Without Forgetting (LWF) (Li and Hoiem, 2017). We demonstrate its effectiveness on a challenging DR dataset (Karthik, 2019), showcasing improved robustness against forgetting and significant performance boosts over existing techniques.

The main contributions of our paper are:

1. A novel zero-shot clustering framework using LLM-generated descriptions for unsupervised image clustering, enhancing ER in CL.
2. A privacy-preserving, CPU-based ER strategy leveraging zero-shot clusters for efficient incremental learning in sensitive domains.

3. Comprehensive experiments demonstrating our method's efficacy in preventing catastrophic forgetting and enhancing existing CL performance across multiple model architectures (MLP, Residual, and Attention).
4. A generalizable approach to CL that addresses key challenges in medical imaging while showing potential applicability to other domains with similar privacy and distribution shift concerns.

2 RELATED WORK

LLMs for Zero-Shot Learning. LLMs have dramatically transformed machine capabilities for understanding and generating human-like text, notably enhancing zero-shot learning (Brown et al., 2020). Our research leverages these capabilities, using LLMs to create descriptive embeddings for images. These embeddings, when integrated with the visual embeddings from CLIP, facilitate effective zero-shot clustering. This method represents a departure from the usual applications of LLMs, which typically direct task execution. Instead, we use their generative power to enhance data organization for CL.

Experience Replay. ER is rooted in the aspiration to emulate aspects of human memory processes, where past experiences are occasionally revisited to solidify learning. The canonical form of ER (Riemer et al., 2019) involves interleaved training of new tasks with memory samples, seeking to approximate the joint distribution of tasks. Variants like Dark ER (Buzzega et al., 2020) have added layers of complexity, employing distillation loss to enforce output consistency. Recent trends in ER have seen the incorporation of dual-memory architectures, such as approaches mirroring the interplay between fast and slow learning processes by maintaining two semantic memories (Arani et al., 2022). While such architectures provide novel mechanisms to handle forgetting, the optimal way to structure and utilize these memories remains an open challenge. In our work, we use this idea to incorporate past data points to inform the replay, based on the data properties.

Privacy-Preserving Exemplars. ER is essential for mitigating catastrophic forgetting, typically involving raw data samples from previous tasks. Our method enhances privacy by storing only the embeddings of exemplars, not the raw images. This modification ensures data privacy while maintaining ER effectiveness. By fine-tuning zero-shot clustering on training datasets, we refine exemplar selection, ensuring the memory buffer contains the most representative embeddings (Rebuffi et al., 2017; Shin et al., 2017).

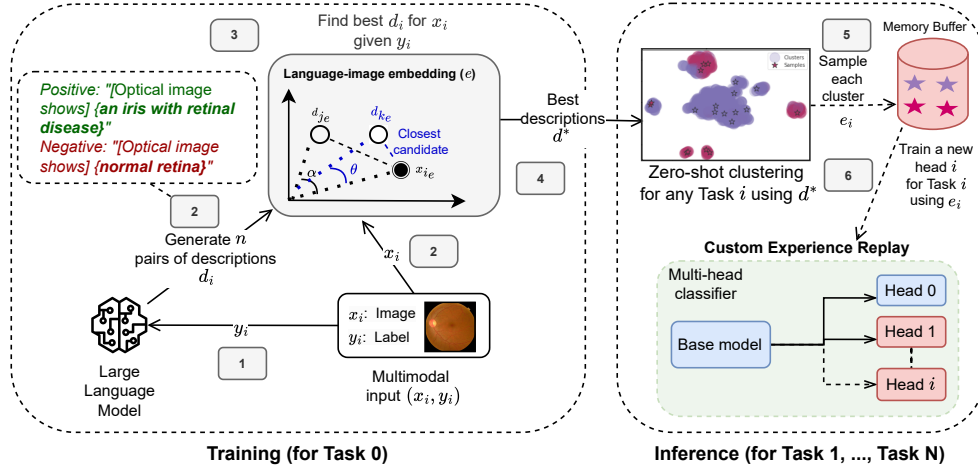


Figure 1: Our method uses a Large Language Model (LLM) to generate descriptions d_i for each image x , using its label y for initial domain learning in Task 0. These descriptions underpin unsupervised zero-shot clustering, forming clusters x_i . Key points from these clusters are buffered for replay. A multi-head classifier leverages this buffer in an Experience Replay (ER) strategy, learning the pertinent head i for predictions y , thus preserving knowledge across successive tasks.

This approach addresses privacy concerns in medical image analysis by storing embeddings instead of actual images, complying with privacy regulations and addressing security concerns (Shokri and Shmatikov, 2015). Our strategy meets the growing demand for privacy-preserving ML techniques.

CLIP Embeddings in CL. Our methodology repurposes CLIP as an embedder within a CL framework, eschewing the common practice of fine-tuning CLIP on downstream tasks. This strategy retains the model’s zero-shot learning capabilities while avoiding the pitfalls of catastrophic forgetting inherent in direct fine-tuning scenarios (Garg et al., 2023). By comparing LLM-generated descriptions, our zero-shot clustering fine-tuning process identifies optimal exemplars for memory storage, facilitating more effective learning across sequential tasks.

Continual Learning in Medical Imaging. CL in medical imaging presents unique challenges due to privacy concerns and data distribution shifts. Recent work has explored continuous domain adaptation for healthcare applications (Venkataramani et al., 2018), addressing the evolving nature of medical data. Additionally, domain adaptation techniques have been applied to medical image segmentation tasks (Valindria et al., 2018), demonstrating the potential of transfer learning in this field. Our work builds upon these foundations, introducing a novel approach that combines zero-shot learning with ER, specifically tailored to handle the privacy and distribution shift issues in medical imaging scenarios.

3 PROBLEM STATEMENT

Challenges and Requirements. The primary challenges in DIL include:

- **Catastrophic Forgetting.** New knowledge acquisition leads to the erosion of previously learned information.
- **Dynamic Data Distributions.** The data distribution \mathcal{D}_i changes over time, necessitating continual model adaptation.
- **Privacy Preservation.** Direct access to raw data is often restricted, especially in sensitive applications like healthcare.
- **Task Boundary Ambiguity.** In real-world scenarios, clear task boundaries may not exist, requiring models to adapt without explicit task delineation.

Formal Definition. DIL involves training a model \mathcal{H} on a sequence of tasks $\{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_n\}$, where the data distribution for each task may change over time. In our context, a task \mathcal{T}_i represents a specific domain or data distribution, such as fundus images with particular lighting conditions or noise levels. The model \mathcal{H} consists of a base network $b(\mathbf{x}; \theta_b)$ shared across tasks and a set of task-specific heads $\{g_k(\mathbf{z}; \theta_k)\}$, where $\mathbf{z} = b(\mathbf{x}; \theta_b)$ is the shared representation. The objective is to minimize the cumulative loss:

$$\min_{\theta_b, \{\theta_k\}} \sum_{i=1}^n \mathcal{L}(\mathcal{H}(\mathcal{D}_i; \theta_b, \theta_i), \mathcal{Y}_i), \quad (1)$$

where θ_b are the parameters of the base network, θ_i are the parameters of the head for task \mathcal{T}_i , \mathcal{D}_i and

\mathcal{D}_i are the data and labels for task \mathcal{T}_i , respectively, and \mathcal{L} denotes the loss function.

During training on a new task \mathcal{T}_n , a new head $g_n(\mathbf{z}; \theta_n)$ is added to the model:

$$\mathcal{H}(\mathbf{x}; \theta_b, \{\theta_k\}_{k=1}^n) = g_n(b(\mathbf{x}; \theta_b); \theta_n). \quad (2)$$

The goal is to optimize the parameters θ_b and $\{\theta_k\}$ such that the performance on all previously learned tasks $\{\mathcal{T}_1, \mathcal{T}_2, \dots, \mathcal{T}_{n-1}\}$ is maintained while learning the new task \mathcal{T}_n .

Task Iteration. In our approach, tasks are presented sequentially, with each task representing a distinct domain or data distribution. The model is trained on these tasks in order, without revisiting previous tasks except through the ER mechanism. This setup simulates real-world scenarios where data distributions evolve over time and previous data may not be fully accessible.

Significance and Impact. Addressing these challenges will enable the development of more robust and adaptive AI systems. In medical imaging, for instance, this means more accurate and timely diagnoses despite evolving data distributions, all while preserving patient privacy. Successfully solving this problem will also advance the broader field of CL, providing insights and techniques applicable to various dynamic environments where data distributions shift over time and privacy is a concern.

4 APPROACH

We propose a framework that synergizes LLMs, specifically GPT-4 (OpenAI, 2023), with vision-language models like CLIP to enhance CL through zero-shot clustering and ER. Our approach addresses key challenges in CL, particularly in privacy-sensitive domains like medical imaging, by leveraging GPT-4-generated descriptions and CLIP embeddings for zero-shot clustering. This method enables the identification of exemplars for ER without storing raw images, ensuring privacy preservation and efficient learning across sequential tasks.

4.1 Zero-Shot Clustering with LLM and CLIP

Our zero-shot clustering method harnesses the combined strengths of GPT-4 and CLIP to cluster images into predefined classes without explicit training. This approach is particularly valuable in CL scenarios where task boundaries are ambiguous and data distributions evolve over time.

Given a set of images $\{I_1, I_2, \dots, I_n\}$ and their associated labels ('Retinopathy' and 'No Retinopathy'), we employ GPT-4 to generate a set of textual descriptions $\{D_1, D_2, \dots, D_m\}$. These descriptions capture various aspects of the images, including potential class information (e.g., presence or absence of retinopathy) and other relevant features. We then utilize CLIP (ViT-L/14@336px configuration) to obtain embeddings for both images and textual descriptions, leveraging its ability to create a shared embedding space for multimodal data.

4.1.1 Embedding Generation

The embedding process consists of two key steps:

1. Textual Description Generation and Tokenization: GPT-4 generates descriptions based on image labels, which are then tokenized for CLIP input.
2. CLIP Encoding: Both tokenized descriptions and images are processed by CLIP to obtain normalized embeddings, denoted as \mathbf{X}_i for images and \mathbf{T}_j for text descriptions.

Formally, we represent this process as:

$$\mathbf{X}_i = \frac{\text{CLIP}_{\text{image}}(I_i)}{|\text{CLIP}_{\text{image}}(I_i)|}, \quad \mathbf{T}_j = \frac{\text{CLIP}_{\text{text}}(D_j)}{|\text{CLIP}_{\text{text}}(D_j)|} \quad (3)$$

where $\text{CLIP}_{\text{image}}(\cdot)$ and $\text{CLIP}_{\text{text}}(\cdot)$ are CLIP's image and text encoding functions, respectively. Normalization ensures all embeddings lie on the unit sphere, facilitating similarity computations.

4.1.2 Similarity Computation and Clustering

We perform zero-shot clustering by computing cosine similarities between image and text embeddings. For each image embedding \mathbf{X}_i , we calculate its similarity to all text embeddings \mathbf{T}_j :

$$S_{ij} = \cos(\mathbf{X}_i, \mathbf{T}_j) = \frac{\mathbf{X}_i \cdot \mathbf{T}_j}{|\mathbf{X}_i| |\mathbf{T}_j|} \quad (4)$$

Label assignment for each image is determined by the text description yielding the highest similarity score: $L_i = \text{argmax}_j S_{ij}$. This process effectively assigns each image to the class best represented by its most similar text description, achieving zero-shot classification without task-specific training.

4.1.3 Optimizing Description Selection

To maximize the effectiveness of zero-shot clustering, we optimize the selection of textual descriptions. We experiment with various templates and description sets (see Table 1), evaluating their clustering performance using F1-score (due to imbalanced data).

Algorithm 1 outlines this optimization process, and Figure 2 presents the results.

Data: Image set I , class labels $\mathcal{L} = \{0, 1\}$, set of description sets \mathcal{D} , set of templates \mathcal{T}
Result: Optimal template T^* , optimal description set D^*

```

foreach template  $t \in \mathcal{T}$  do
  foreach description set  $d \in \mathcal{D}$  do
    Generate text prompts using template  $t$  and description set  $d$ ;
    Obtain text embeddings  $\mathbf{T}_j$  using the CLIP model;
    foreach image  $I_i \in I$  do
      Compute similarity scores  $S_{ij}$  between image  $I_i$  and all embeddings  $\mathbf{T}_j$ ;
      Assign label  $L_i$  to image  $I_i$  based on the highest similarity score  $S_{ij}$ ;
    end
    Evaluate the F1-score for the current combination of  $t$  and  $d$ ;
    if current scores are higher than the best previous scores then
      Update  $T^*$  and  $D^*$  with the current template  $t$  and description set  $d$ ;
    end
  end
end

```

Algorithm 1: Optimizing Description Selection for Zero-shot Clustering.

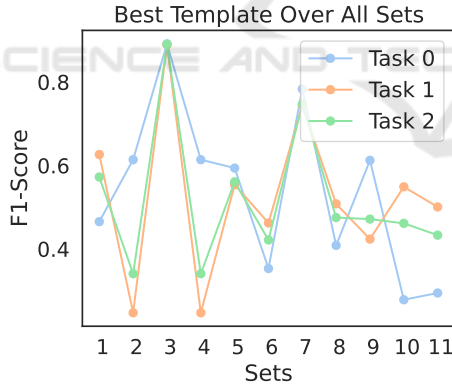


Figure 2: Optimal template against all the sets. We can see that 'Set 3' gets the highest score for Task 0 and also for Task 1 and Task2, indicating a useful description and template for zero-shot clustering at inference time.

We task GPT-4 to generate 11 description sets for the classes *No Retinopathy* and *Retinopathy*, paired with 10 templates for constructing prompts. These sets vary in detail, from technical terms (e.g., "Mild to Severe non-proliferative diabetic retinopathy") to simpler descriptions (e.g., "Signs of diabetic retinopathy"). Templates such as "An iris with {}", "A human eye with {}", and "An ocular image with {}" were used to form the final prompts.

Table 1: 10 Templates and 11 Description Sets optimized for Zero-shot Clustering.

#	Templates	Binary Description Sets
1	An iris with	Healthy / Diabetic damage
2	A human eye with	No damage / Diabetic signs
3	<i>no template</i>	Normal / Retinal disease
4	An ocular image with	No issues / Retinopathy
5	A retinal photo with	Clear fundus / Fundus changes
6	A fundus image displaying	No retinopathy / Mild-severe incl. laser
7	Visible symptoms suggest	Normal fundus / Retinopathy incl. laser
8	Retinal scan reveals	No abnormalities / Non-proliferative
9	Optical image shows	No pathology / Mild-severe incl. laser
10	The condition of the retina is	Healthy / Mild-severe incl. laser
11		No disease / Various stages incl. laser

This comprehensive generation enables our method to adapt to the nuances of DR detection.

4.2 Stratified Sampling for Experience Replay

After optimizing the textual descriptions and templates for zero-shot clustering, we employ stratified sampling to ensure balanced representation of each class within the ER buffer. This approach is crucial for constructing a diverse and representative collection of multimodal inputs, each containing an embedding, label, ensuring effective and privacy-preserving ER while promoting generalization across tasks and minimizing catastrophic forgetting.

4.2.1 Sampling Procedure

Given a collection of multimodal inputs \mathcal{M} , where each input $m_i \in \mathcal{M}$ is characterized by its embedding \mathbf{E}_i , zero-shot label z_i , we define a stratified sampling strategy to select a subset $\mathcal{S} \subseteq \mathcal{M}$ with proportional representation across the zero-shot classified labels.

For each distinct zero-shot label $l \in \mathcal{Z}$ derived from the clustering process, we define a subset $\mathcal{M}_l \subset \mathcal{M}$ containing inputs with label l . We then sample $n_{\text{neighbors}}$ inputs from each subset \mathcal{M}_l :

$$S_l = \begin{cases} \text{sample}(\mathcal{M}_l, n_{\text{neighbors}}), & \text{if } |\mathcal{M}_l| > n_{\text{neighbors}} \\ \mathcal{M}_l, & \text{otherwise} \end{cases} \quad (5)$$

The final sample set \mathcal{S} is the union of all samples across the labels:

$$\mathcal{S} = \bigcup_{l \in \mathcal{Z}} \mathcal{S}_l \quad (6)$$

This stratified sampling strategy ensures a balanced replay buffer, critical for maintaining diversity during ER and reducing the risk of catastrophic forgetting while reinforcing learning across tasks.

4.2.2 Clustering+Sampling Performance

We evaluate our zero-shot clustering and stratified sampling approach across three tasks of increasing complexity (described in Section 5.2.1). Some key insights can be derived from our results, which are shown in Table 2.

Table 2: Zero-shot clustering and sampling results across tasks 0, 1, and 2.

Task	Class 0	Class 1	Samples	F1-score
0	387	653	20 (1.92%)	0.892
1	1732	890	20 (0.76%)	0.890
2	2120	1542	20 (0.55%)	0.891

- **Class Distribution.** There is significant class imbalance across tasks, mirroring real-world medical imaging scenarios where pathological cases are less frequent.
- **Sampling Efficiency.** Consistent selection of 20 samples per task (1.92% to 0.55% of total data), demonstrating compact yet representative memory buffer maintenance as the dataset grows.
- **Performance Stability.** F1-scores remain consistent (0.89) across tasks despite increasing complexity and imbalance, highlighting the robustness of our approach in evolving data distributions.

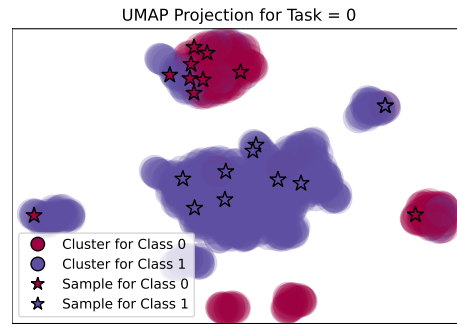
These quantitative results are further illustrated by the qualitative analysis shown in Figure 3, which presents UMAP projections of clusters and selected samples for each task.

4.2.3 Template and Description Set Performance

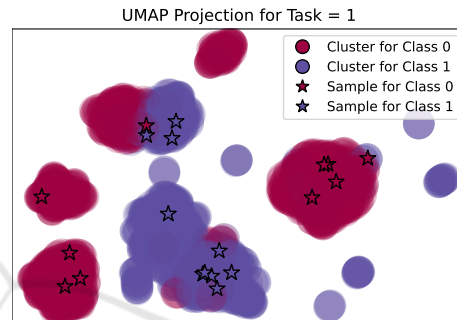
We evaluate the performance of various templates and description sets for zero-shot clustering across tasks. Figure 4 presents the top-performing combinations for each task.

The key findings from our template and description set analysis are as follows:

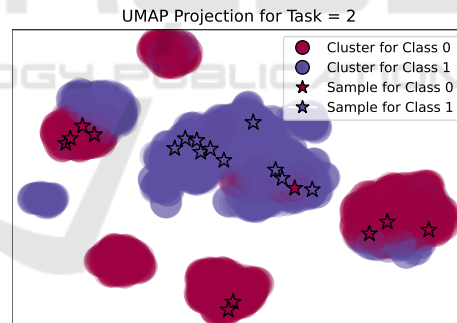
- Across all tasks, the template "Optical image shows" combined with description Set 3 ("Normal



(a) Task 0: Clusters with uniform image quality. The model successfully differentiates between Class 0 and Class 1, with well-separated cluster shapes.



(b) Task 1: Clusters impacted by lighting variation. Lighting variations cause more overlap between clusters, making classification harder, yet most sample points are still correctly clustered.

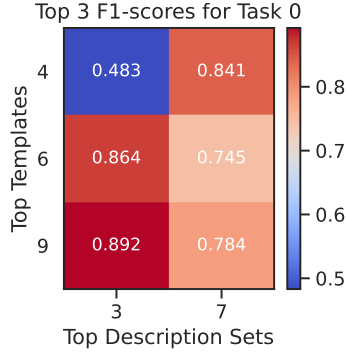


(c) Task 2: Clusters with Gaussian noise applied. Noise increases cluster overlap significantly. Despite the noise, some samples remain distinguishable, demonstrating moderate model robustness.

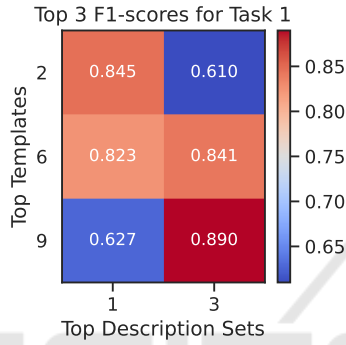
Figure 3: UMAP projections of clusters for embeddings across the three tasks. Each projection includes clusters and 10 samples from the memory buffer for both Class 0 and Class 1. These projections illustrate how the model adapts to progressively increasing task complexity.

/ Retinal disease") consistently achieves the highest F1-scores (0.892, 0.890, and 0.891 for Tasks 0, 1, and 2, respectively).

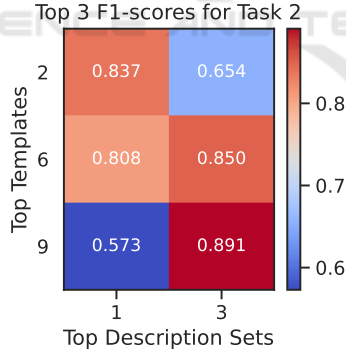
- Set 3 demonstrates robust performance across all tasks, indicating its effectiveness in zero-shot clustering for DR detection.



(a) Task 0: Top 3 templates and description sets. The best-performing pair is {Template 9, Set 3}, achieving an F1-score of 0.892.



(b) Task 1: Top 3 templates and description sets. The best-performing pair is {Template 9, Set 3}, achieving an F1-score of 0.890.



(c) Task 2: Top 3 templates and description sets. The best-performing pair is {Template 9, Set 3}, achieving an F1-score of 0.891.

Figure 4: Top 3 templates and description sets for each task, highlighting the highest F1-scores for zero-shot clustering across Tasks 0, 1, and 2. Set 3 consistently shows top performance in combination with different templates across all tasks.

- Sets 1 ("Healthy / Diabetic damage") and 7 ("Normal fundus / Retinopathy incl. laser") also show promising performance, highlighting the importance of carefully crafted descriptions in zero-shot learning scenarios.

These results highlight the robustness of our method, which consistently adapts to increasing task complexity (e.g., uniform quality, lighting variation, Gaussian noise) while maintaining high performance in zero-shot clustering and sample selection for ER. This demonstrates its effectiveness in real-world scenarios where image quality varies significantly.

4.3 Experience Replay Algorithm

We propose an enhanced ER algorithm that leverages zero-shot clustering and stratified sampling to address catastrophic forgetting in CL. Our method comprises two key components: a zero-shot exemplars buffer and an ER strategy.

4.3.1 Zero-shot Exemplars Buffer

The zero-shot exemplars buffer maintains a balanced set of exemplars based on zero-shot clustering outcomes. It is updated as shown in Algorithm 2.

Input: Max buffer size max_size , neighbors $n_neighbors$, strategy \mathcal{S}

Output: Updated replay buffer \mathcal{B}

$\mathcal{D} \leftarrow \{\text{CreateMultimodalInput}(d) \mid d \in \mathcal{S}.experience.dataset\};$

$\mathcal{C}, \mathcal{Z} \leftarrow$

$\text{ZEROSHOTCLUSTERING}(\mathcal{D}, text_embs_best);$

$\mathcal{S} \leftarrow \text{STRATIFIEDSAMPLING}(\mathcal{C}, \mathcal{Z}, n_neighbors);$

$\mathcal{B} \leftarrow \{(\mathbf{E}_s, y_s, t_s) \mid s \in \mathcal{S}, (\mathbf{E}_s, y_s, t_s) = \text{ExtractFeatures}(s)\};$

Update \mathcal{B} in strategy \mathcal{S} , respecting max_size ;

Algorithm 2: Updating Replay Buffer with Zeroshot Exemplars.

Key features of our zero-shot exemplars buffer include: multimodal input creation encapsulating embeddings and labels, unsupervised clustering using pre-computed text embeddings, stratified sampling for balanced cluster representation, privacy-preserving updates using embeddings instead of raw data and dynamic buffer group adjustment based on clustering outcomes.

This approach ensures diverse sample representation while maintaining privacy and efficiency. The use of zero-shot labels enhances applicability in scenarios with scarce ground truth.

4.3.2 Experience Replay Strategy

Our ER strategy integrates the zero-shot exemplars buffer into the training process as shown in Algorithm 3. This strategy addresses key CL challenges through several mechanisms. It employs adaptive sampling via zero-shot clustering, mitigating task boundary ambiguity, while ensuring balanced class and task repre-

sensation through stratified sampling. The approach maintains privacy preservation by operating on embeddings, and achieves computational efficiency with a compact, diverse replay buffer.

```

Input: Training strategy  $\mathcal{S}$ , Storage policy  $\mathcal{P}$  (with
zero-shot exemplars buffer  $\mathcal{B}$ )
Output: Updated training strategy with ER
Attach  $\mathcal{P}$  to  $\mathcal{S}$ ;
while training do
  if  $\mathcal{B} \in \mathcal{P}$  is not empty then
     $\mathcal{S}.dataloader \leftarrow$ 
    Combine( $\mathcal{S}.adapted\_dataset, \mathcal{B}$ );
  end
  ExecuteTrainingExperience();
   $\mathcal{B} \leftarrow \mathcal{P}.update(\mathcal{S})$ ;
end

```

Algorithm 3: Experience Replay Strategy.

5 EXPERIMENTAL EVALUATION

We rigorously evaluate our proposed CL approach on the challenging task of DR detection, assessing its efficacy under various conditions that simulate real-world scenarios.

5.1 Experimental Setup

5.1.1 Testbed

Our experiments were conducted on a CPU-based platform with comprehensive specifications. The hardware configuration consists of Dual Intel Xeon Platinum 8360Y CPUs operating at 2.40GHz with 256 GB RAM, running on Ubuntu 22.04 LTS. Our software stack includes the Intel AI Analytics Toolkit (Docker image: intel/oneapi-aikit:devel-ubuntu22.04)¹, Avalanche 0.3.1 for CL², Intel Extension for PyTorch 1.12.100+cpu³, and Intel Extension for Scikit-learn 2023.0.1⁴. This environment ensures reproducibility and leverages optimized libraries for enhanced performance on CPU architectures.

5.1.2 Dataset

We utilize the APTOS 2019 Blindness Detection dataset (Karthik, 2019), comprising 3,662 high-resolution retinal images. This dataset, developed in collaboration with Aravind Eye Hospital in India, captures real-world clinical complexities and image

quality variations, providing a robust testbed for our CL approach. Figure 5 presents sample images from this dataset, illustrating the diversity in image quality and pathological conditions.

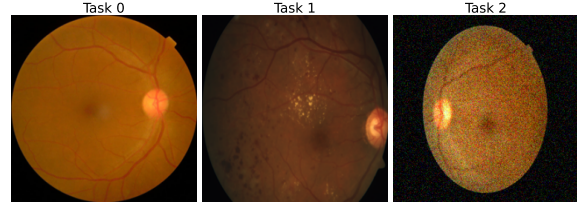


Figure 5: Representative fundus images from different tasks, showcasing varying image quality and conditions. Left: Task 0 - uniform image quality; Center: Task 1 - variation in lighting; Right: Task 2 - artificially added Gaussian noise to simulate challenging imaging conditions.

5.2 Experimental Methodology

5.2.1 Task Design

We construct three distinct tasks to assess our model’s performance under progressively challenging conditions:

- **Task 0 (Baseline).** Uniform image quality, representing ideal clinical conditions.
- **Task 1 (Lighting Variation).** Introduces variations in lighting, simulating different imaging environments.
- **Task 2 (Noise Addition).** Incorporates Gaussian noise, emulating low-quality or degraded images.

This task progression allows us to evaluate our model’s robustness to common real-world variations in medical imaging.

5.2.2 Model Architectures

To rigorously evaluate the generalizability and robustness of our approach, we employ three distinct neural network architectures, each chosen to address specific aspects of CL in medical imaging:

- **Multi-Layer Perceptron (MLP).** A baseline architecture with one hidden layer, selected for its simplicity and efficiency. This model serves as a litmus test for our method’s ability to enhance even basic architectures in CL scenarios.
- **Residual Network.** Incorporating skip connections, this architecture mitigates the vanishing gradient problem, crucial for maintaining performance across multiple tasks in CL. Its ability to learn residual functions is particularly relevant for detecting subtle changes in medical images across different domains.

¹<https://hub.docker.com/r/intel/oneapi-aikit>

²<https://avalanche.continualai.org/>

³<https://github.com/intel/intel-extension-for-pytorch>

⁴<https://github.com/intel/scikit-learn-intelx>

- **Attention-based Network.** Leveraging self-attention mechanisms, this model excels at capturing complex, long-range dependencies in data. In the context of medical imaging, it can focus on the most relevant features for diagnosis, potentially enhancing the model’s adaptability to new tasks.

All architectures are designed to process 768-dimensional CLIP embeddings as input, outputting binary classifications for DR. This unified input-output structure allows for fair comparison across architectures while leveraging the rich semantic information captured by CLIP embeddings.

5.2.3 Continual Learning Strategies

We benchmark our approach against a diverse set of state-of-the-art CL strategies, each addressing different aspects of the catastrophic forgetting problem:

- **Naive (fine-tuning).** Serves as a baseline, highlighting the severity of catastrophic forgetting in the absence of specialized CL techniques.
- **Elastic Weight Consolidation (EWC).** A regularization-based approach that selectively slows down learning on important parameters, crucial for preserving knowledge in medical imaging tasks where certain features may be universally important.
- **Learning without Forgetting (LwF).** Employs knowledge distillation to retain previous task information, potentially beneficial in scenarios where task boundaries in medical imaging are not clearly defined.
- **Gradient Episodic Memory (GEM).** Constrains gradient updates to maintain performance on previous tasks, offering insights into the trade-offs between stability and plasticity in medical AI models.

Each strategy is evaluated in its original form and enhanced with our proposed zero-shot clustering and stratified sampling approach. This comprehensive evaluation not only benchmarks our method against established techniques but also demonstrates its potential as a complementary enhancement to existing CL strategies in the challenging domain of medical image analysis.

5.2.4 Evaluation Metric

We employ the Average Mean Class Accuracy (AMCA) as our primary evaluation metric:

$$AMCA = \frac{1}{T} \sum_{t=1}^T \left(\frac{1}{C} \sum_{c=1}^C a_{c,t} \right) \quad (7)$$

where $T = 3$ (number of tasks) and $C = 2$ (number of classes: with and without DR). This metric ensures robustness against class imbalance and distribution shifts across tasks.

5.3 Experimental Protocol

To ensure robust and generalizable results, we implemented a rigorous experimental protocol. Our approach included extensive hyperparameter exploration, varying the number of neighbors (15, 20, 25, 30, 50) in our stratified sampling approach to assess sensitivity. For stochastic control, we utilized different random seeds for each run, ensuring statistical validity. We conducted a comprehensive evaluation across multiple architectures, including MLP, Residual, and Attention-based networks. These were tested with various CL strategies: Naive, EWC, LwF, and GEM, both in their original form and enhanced with our approach. For each configuration, we recorded key performance metrics, focusing on AMCA and Forgetting scores. This comprehensive evaluation framework enables us to draw statistically significant conclusions about our method’s efficacy across diverse scenarios in medical imaging CL.

5.4 Results and Discussion

We evaluate our approach across multiple architectures (MLP, Residual, Attention) and continual learning strategies (Naive, GEM, LwF, EWC), comparing performance and computational efficiency against baseline methods.

5.4.1 Performance Analysis

Table 3 presents Average Mean Class Accuracy (AMCA) scores and Forgetting metrics across different configurations. Our method consistently outperforms baselines across all architectures and strategies.

- **Architectural Robustness.** Performance improvements range from 0.8% to 3.1% in AMCA across all architectures, with the most significant gains observed in complex models (Residual: +2.8% for Naive, Attention: +3.1% for LwF).
- **Strategy Enhancement.** Our approach amplifies the strengths of existing CL strategies. For instance, it reduces Forgetting in Naive learning (8.5 to 5.2 in Residual models) and enhances knowledge distillation in LwF (3.1% AMCA increase in Attention models).

- **Forgetting Mitigation.** We observe consistent reductions in Forgetting metrics, particularly notable in the Naive strategy across all architectures, indicating improved knowledge retention.

These results suggest that our zero-shot clustering and stratified sampling approach provides a more diverse and representative set of samples, enhancing both learning and retention in CL scenarios.

5.4.2 Hyperparameter Sensitivity

The number of neighbors in KNN has a significant impact on both performance and forgetting. The Mean AMCA shows peak performance at around 25 neighbors, followed by a slight decline and stabilization between 30 and 50 neighbors. Similarly, forgetting is reduced most significantly at around 25 neighbors, after which it gradually decreases as the number of neighbors increases.

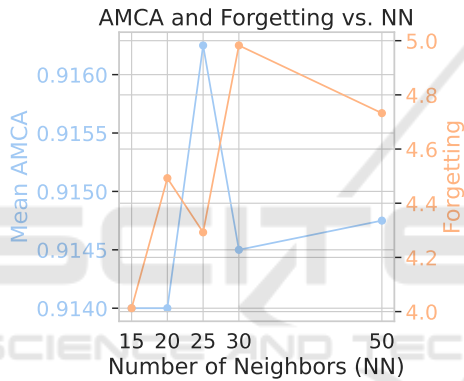


Figure 6: Both metrics reach optimal values around 25-30 number of neighbors (NN), illustrating the trade-off between performance and retention.

This highlights a sweet spot between 25 and 30 neighbors, where both performance (Mean AMCA) and retention (reduced forgetting) are optimized. Tuning within this range balances sample diversity and computational efficiency, ensuring high performance with minimal forgetting, as shown in Figure 6.

5.4.3 Computational Efficiency

Figure 7 compares the execution times between our method and the original strategy. Our approach incurs a minimal 6% increase in average execution time (5.81 s vs. 5.48 s). This negligible overhead is consistent across all architectures, with Attention models showing the highest variability due to their complexity.

The marginal increase in computational cost, coupled with significant performance gains, positions our method as an efficient in-place replacement for existing strategies. This balance is particularly valuable

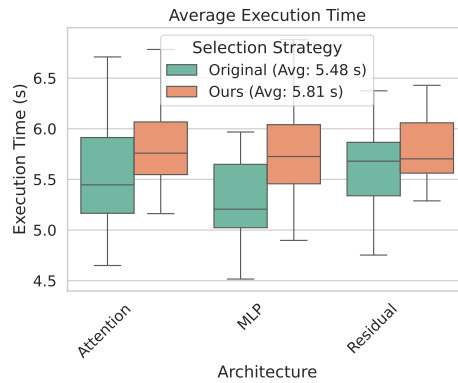


Figure 7: Average execution time comparison across architectures. Box plots show distribution of execution times, with mean values in the legend.

in time-sensitive applications like medical imaging, where improved accuracy without substantial computational overhead is crucial.

6 CONCLUSION

This study introduced a framework integrating zero-shot learning with Experience Replay for CL in medical imaging, with a focus on DR detection. Our approach, leveraging LLMs for DIL, demonstrated several key achievements. We observed consistent performance improvements across diverse model architectures and CL strategies, with AMCA increases up to 3.1%. The system showed effective mitigation of catastrophic forgetting, evidenced by reduced Forgetting metrics, particularly in naive learning scenarios. Additionally, we achieved this with negligible computational overhead (6% increase in execution time), enabling seamless integration into existing systems.

These results highlight the potential of our method to enhance the adaptability, efficiency, and privacy-preservation of AI systems in healthcare. The framework's ability to maintain performance across varying data distributions while operating on embeddings addresses critical challenges in medical AI deployment.

Future research directions encompass several key areas. We aim to scale to more complex, multi-modal medical datasets and develop adaptive clustering algorithms for dynamic medical imaging scenarios. Additionally, we plan to explore applicability in other domains with similar privacy and distribution shift concerns. A crucial component of future work involves conducting rigorous ethical analyses, particularly regarding data privacy and algorithmic bias in diverse patient populations.

While our work represents a step towards more robust and adaptable AI in healthcare, realizing its full

Table 3: Comparison of Mean AMCA scores and forgetting (parentheses) for different base models and strategies. Boldface indicates superior performance between *Ours* and *Original* strategies based on Mean AMCA. Averages are computed across different NN values for each base model.

Model	NN	Naive		GEM		LwF		EWC	
		Ours	Original	Ours	Original	Ours	Original	Ours	Original
Attention	15	0.936(3.7)	0.929(3.4)	0.967(3.5)	0.965(3.4)	0.823(4.8)	0.793(5.3)	0.930(4.1)	0.928(3.1)
	20	0.938(4.2)	0.926(4.3)	0.968(4.2)	0.965(3.2)	0.816(5.8)	0.793(5.3)	0.934(3.7)	0.919(4.4)
	25	0.939(4.7)	0.922(4.9)	0.967(3.3)	0.965(3.2)	0.826(5.5)	0.789(5.4)	0.933(3.6)	0.915(4.2)
	30	0.937(4.5)	0.930(2.5)	0.967(4.2)	0.961(3.4)	0.825(5.3)	0.792(5.5)	0.929(5.9)	0.918(4.1)
	50	0.938(4.3)	0.932(2.6)	0.965(4.1)	0.960(3.8)	0.824(5.7)	0.791(5.0)	0.932(4.8)	0.911(5.7)
	Avg	0.938(4.3)	0.928(3.5)	0.967(3.9)	0.963(3.4)	0.823(5.4)	0.792(5.3)	0.932(4.4)	0.918(4.3)
Residual	15	0.913(5.1)	0.880(9.7)	0.941(2.6)	0.934(2.8)	0.940(4.6)	0.928(4.5)	0.939(3.8)	0.933(3.7)
	20	0.912(5.0)	0.879(9.7)	0.942(2.8)	0.935(2.9)	0.934(4.5)	0.930(4.5)	0.938(4.3)	0.932(4.1)
	25	0.913(5.3)	0.879(9.7)	0.942(1.7)	0.938(2.8)	0.933(4.6)	0.928(4.4)	0.938(3.9)	0.926(4.5)
	30	0.912(5.7)	0.879(9.6)	0.940(3.9)	0.937(2.7)	0.938(4.4)	0.928(4.5)	0.938(4.9)	0.914(6.7)
	50	0.912(5.0)	0.903(3.8)	0.941(3.1)	0.935(2.6)	0.936(4.8)	0.929(4.5)	0.932(6.7)	0.901(10.0)
	Avg	0.912(5.2)	0.884(8.5)	0.941(2.8)	0.936(2.8)	0.936(4.6)	0.929(4.5)	0.937(4.7)	0.921(5.8)
MLP	15	0.915(5.1)	0.899(3.2)	0.931(5.1)	0.923(3.8)	0.942(4.8)	0.940(4.8)	0.930(6.9)	0.917(5.8)
	20	0.914(5.3)	0.897(4.3)	0.931(5.1)	0.923(3.7)	0.941(5.0)	0.940(4.5)	0.892(5.2)	0.874(5.4)
	25	0.914(5.2)	0.900(3.8)	0.931(5.1)	0.922(4.0)	0.944(5.0)	0.937(5.0)	0.891(5.6)	0.875(4.1)
	30	0.914(5.6)	0.896(4.2)	0.930(5.2)	0.923(3.8)	0.944(4.8)	0.937(5.3)	0.892(5.5)	0.879(3.2)
	50	0.911(5.5)	0.893(5.1)	0.927(5.3)	0.921(3.9)	0.948(4.9)	0.936(5.3)	0.890(5.9)	0.872(4.7)
	Avg	0.914(5.4)	0.897(4.1)	0.930(5.1)	0.922(3.9)	0.944(4.9)	0.938(5.0)	0.899(5.8)	0.883(4.6)

potential requires extensive clinical validation. As we progress towards real-world applications, addressing scalability, generalizability, and ethical considerations will be paramount.

ACKNOWLEDGEMENTS

We thank Lenovo for providing the technical infrastructure to run the experiments in this paper. This work was partially supported by Lenovo and Intel as part of the Lenovo AI Innovators University Research program, by the Spanish Ministry of Science (MICINN), the Research State Agency (AEI) and European Regional Development Funds (ERDF/FEDER) under grant agreements PID2019-107255GB-C22 and PID2021-126248OB-I00, MCIN/AEI/10.13039/501100011033/FEDER, UE, and by the Generalitat de Catalunya under contract 2021-SGR-00478.

REFERENCES

- Arani, E., Sarfraz, F. B., and Zonooz, B. (2022). Learning fast, learning slow: A general continual learning method based on complementary learning system. arXiv preprint abs/2201.12604.
- Brown, T. B. and et al., M. (2020). Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS’20. Curran Associates Inc.
- Buzzega, P., Boschini, M., Porrello, A., Abati, D., and Calderara, S. (2020). Dark experience for general continual learning: a strong, simple baseline. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS’20. Curran Associates Inc.
- De Lange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G., and Tuytelaars, T. (2022). A continual learning survey: Defying forgetting in classification tasks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3366–3385.
- Garg, S., Farajtabar, M., Pouransari, H., Vemulapalli, R., Mehta, S., Tuzel, O., Shankar, V., and Faghri, F. (2023). TiC-CLIP: Continual Training of CLIP Models. arXiv preprint abs/2310.16226.
- Karthik, Maggie, S. D. (2019). APTOS 2019 Blindness Detection. <https://kaggle.com/competitions/aptos2019-blindness-detection>.
- Khan, V., Cygert, S., Deja, K., Trzcinski, T., and Twardowski, B. (2024). Looking through the past: Better knowledge retention for generative replay in continual learning. *IEEE Access*, 12:45309–45317.
- Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

- Koh, H., Kim, D., Ha, J.-W., and Choi, J. (2022). Online continual learning on class incremental blurry task configuration with anytime inference. arXiv preprint abs/2110.10031.
- Kuang, K., Cui, P., Athey, S., Xiong, R., and Li, B. (2018). Stable prediction across unknown environments. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD'18*, pages 1617–1626, New York, NY, USA. Association for Computing Machinery.
- Kumar, P. and Srivastava, M. M. (2018). Example mining for incremental learning in medical imaging. In *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 48–51.
- Kumari, S. and Singh, P. (2024). Deep learning for unsupervised domain adaptation in medical imaging: Recent advancements and future perspectives. *Computers in Biology and Medicine*, 170:107912.
- Lenga, M., Schulz, H., and Saalbach, A. (2020). Continual learning for domain adaptation in chest x-ray classification. In *Proceedings of the Third Conference on Medical Imaging with Deep Learning, PMLR 121:413-423, 2020*.
- Li, Z. and Hoiem, D. (2017). Learning without forgetting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(12):2935–2947.
- Lopez-Paz, D. and Ranzato, M. (2017). Gradient episodic memory for continual learning. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, pages 6470–6479. Curran Associates Inc.
- OpenAI (2023). GPT-4 Technical Report. <https://openai.com/research/gpt-4>.
- Parisi, G. I., Kemker, R., Part, J. L., Kanan, C., and Wermter, S. (2019). Continual Lifelong Learning with Neural Networks: A Review. *Neural Networks*, 113:54–71.
- Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., and Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In Meila, M. and Zhang, T., editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.
- Rebuffi, S.-A., Kolesnikov, A., Sperl, G., and Lampert, C. H. (2017). iCaRL: Incremental Classifier and Representation Learning. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5533–5542.
- Riemer, M., Cases, I., Ajemian, R., Liu, M., Rish, I., Tu, Y., and Tesauero, G. (2019). Learning to learn without forgetting by maximizing transfer and minimizing interference. arXiv preprint abs/1810.11910.
- Robins, A. (1993). Catastrophic forgetting, rehearsal and pseudorehearsal. *Connection Science*, 5(2):123–146.
- Serra, J., Suris, D., Miron, M., and Karatzoglou, A. (2018). Overcoming catastrophic forgetting with hard attention to the task. In *International Conference on Machine Learning*, pages 4548–4557. PMLR.
- Shin, H., Lee, J. K., Kim, J., and Kim, J. (2017). Continual learning with deep generative replay. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17*, pages 2994–3003. Curran Associates Inc.
- Shokri, R. and Shmatikov, V. (2015). Privacy-preserving deep learning. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 909–910.
- Valindria, V. V., Lavdas, I., Bai, W., Kamnitsas, K., Aboagye, E. O., Rockall, A. G., Rueckert, D., and Glocker, B. (2018). Domain Adaptation for MRI Organ Segmentation using Reverse Classification Accuracy. arXiv preprint abs/1806.00363.
- Venkataramani, R., Ravishankar, H., and Anamandra, S. (2018). Towards continuous domain adaptation for healthcare. arXiv preprint abs/1812.01281.
- Wang, L., Zhang, X., Su, H., and Zhu, J. (2024). A Comprehensive Survey of Continual Learning: Theory, Method and Application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5362–5383.
- Zhang, J., Fu, Y., Peng, Z., Yao, D., and He, K. (2024). CORE: Mitigating Catastrophic Forgetting in Continual Learning through Cognitive Replay. arXiv preprint abs/2402.01348.