Integrating Uncertainty and Latent Feature Diversity for Robust Memory Replay in Continual Medical Imaging Learning

Mathilde Dupouy[1 [0009−0000−6344−8329]⋆], Yamil Vindas Yassine[2 [0000−0003−3553−0071]⋆], Novia Putri Bahirah[1], Thibaut Dambry[3], Blaise Kévin Guépié[4 [0000−0002−0241−7315]], Douglas Teodoro[2 [0000−0001−6238−4503]], and Philippe Delachartre[1 [0000−0003−0492−1846]]

[1] INSA-Lyon, Université Claude Bernard Lyon 1, CNRS, Inserm, CREATIS UMR 5220, U1294, F-69XXX, LYON, France
[2] Data Science for Digital Health, Geneva University, Geneva, Switzerland
[3] Atys Medical, Soucieu-en-Jarrest 69510, France
[4] LIST3N, Université de Technologie de Troyes, Troyes, France
⋆ Equal contribution. Corresponding author: mathilde.dupouy@insa-lyon.fr

Abstract. Continual learning (CL) in medical imaging must address sequential hospital-level domain shifts under strict memory constraints. We propose an active rehearsal framework that optimizes memory buffer population through two complementary strategies: (1) an uncertainty-aware method using Monte Carlo Dropout to disentangle aleatoric and epistemic uncertainty into a robust informativeness score, and (2) an unsupervised dissimilarity strategy maximizing latent feature diversity without requiring selection labels. We evaluate our approach on three domain-incremental scenarios spanning distinct source (Task A) and target (Task B) domains: OrganMNIST, Camelyon17 (multi-center histopathology), and HITS-5K (a private multi-center transcranial Doppler ultrasound cohort). Results demonstrate that our targeted selection mechanisms achieve comparable or superior performance than regularization (e.g., EWC) and standard uniform replay baselines. Notably, on HITS-5K, all replay strategies exhibit strong positive backward transfer, improving Task B accuracy beyond 93% while reducing forgetting of Task A to −17%. At higher memory ratios, performance converges, suggesting that optimized minimal replay suffices for synergistic tasks. Moreover, qualitative analysis of memory coverage in the latent space validates the theoretical motivations behind our sampling mechanisms. Our work demonstrates that adaptive, semantics-aware replay CL improves model robustness for medical image analysis.

Keywords: Continual Learning · Uncertainty Quantification · Transcranial Doppler Ultrasound · Domain-Incremental Learning.

1 Introduction

In healthcare, severe data drift arises continuously from hardware updates, shifting acquisition protocols, and demographic variations [18,20]. Medical Artificial

Intelligence (AI) systems must adapt to these dynamic environments without sacrificing prior diagnostic knowledge, a challenge known as catastrophic forgetting. Furthermore, strict privacy regulations (e.g., HIPAA, GDPR) and data silos often restrict centralized model retraining, requiring other solutions for model adaptation. Continual learning (CL) addresses these challenges [18], enabling personalized, center-specific adaptation [21]. While federated learning facilitates privacy-preserving collaboration, its reliance on static data distributions highlights the need for CL mechanisms in lifelong medical AI deployments [7].

To mitigate catastrophic forgetting, CL strategies include regularization, dynamic architectures, and replay [5,6]. Replay methods [3,14,26], which store a constrained episodic memory of past samples, often demonstrate superior robustness and are highly advantageous for sequentially integrating heterogeneous clinical data [17]. However, clinical translation is hindered because standard replay relies on exhaustive, expert-level annotations, which are impractically expensive to obtain in medical imaging. While efforts exist to address label scarcity in medical [13,17,25] and broader domains [4,8,23], they often impose complex operational constraints. An underexplored gap remains in utilizing weakly-supervised or unsupervised mechanisms to optimize replay buffer curation and alleviate the annotation burden.

Active CL mitigates this burden by replacing sub-optimal uniform memory sampling with strategies that leverage predictive uncertainty to identify informative boundary samples [16,27,28] or dissimilarity metrics to maximize feature diversity [19]. Building upon these principles, we propose a tailored rehearsal-based CL framework optimizing memory population across sequential clinical domain shifts. Specifically, we introduce two complementary selection mechanisms: (1) an uncertainty-aware strategy using Monte-Carlo dropout (MCD) to distill aleatoric and epistemic uncertainty into a robust informativeness score, and (2) an unsupervised dissimilarity strategy maximizing latent domain coverage without requiring labels for selection. We demonstrate the clinical translation of this framework via the first CL application to transcranial Doppler (TCD) ultrasonography for cerebral emboli classification. Extensive evaluations on two other publicly available datasets reveal our targeted selection mechanisms often outperform memoryless and regularization baselines, particularly under strict memory constraints.

2 Methodology

2.1 Problem Description

We formalize domain-incremental CL in medical imaging under a sequential two-task classification setting with fixed labels, though the methods extend naturally to longer sequences. Let dataset A with distribution $\mathcal{D}_A$ originate from hospitals $\mathcal{H}_A$. Subsequently, dataset B arrives from disjoint centers $\mathcal{H}_B$ ($\mathcal{H}_A \cap \mathcal{H}_B = \varnothing$), at which point $\mathcal{D}_A$ is no longer (fully) accessible. All samples occupy a unified input-label space $(X_T, Y_T) \sim \mathcal{D}_\mathcal{T} \in \mathcal{X} \times \mathcal{Y}$ for $T \in \{A, B\}$.

The objective is to optimize a robust classifier $f^* : \mathcal{X} \to \mathcal{Y}$ across both domains. Let $f_{\theta_A} = g_{\psi_A} \circ h_{\gamma_A}$ denote the base model trained on $\mathcal{D}_A$, where $h_{\gamma_A} : \mathcal{X} \to \mathbb{R}^d$ is the feature encoder and $g_{\psi_A} : \mathbb{R}^d \to \mathcal{Y}$ is the classification head. Let $f_{\theta_{A,B}} = g_{\psi_{A,B}} \circ h_{\gamma_{A,B}}$ represent the sequentially updated model after learning from the target distribution $\mathcal{D}_B$. The goal is to maximize target accuracy while minimizing forgetting on $\mathcal{D}_A$.

2.2 Baselines

Sequential fine-tuning from f_{θ_A} directly onto $\mathcal{D}_B$ serves as our naive baseline. We compare our framework against Elastic Weight Consolidation (EWC) [2,10] and six replay-based strategies.

Replay methods relax the assumption that data from the source domain $\mathcal{D}_A$ is entirely inaccessible during adaptation to the target domain $\mathcal{D}_B$. Instead, they preserve a fixed-capacity episodic memory $\mathcal{M}$ of size C containing representative source samples. At the task boundary, memory updates occur iteratively at the batch level. Given an active data stream batch $\mathcal{B}$, a selection algorithm $\mathcal{A}$ evaluates the candidate pool $\mathcal{S} = \mathcal{M} \cup \mathcal{B}$. The algorithm maps this set to an updated buffer, $\mathcal{A}(\mathcal{S}, M) \mapsto \mathcal{M}'$, retaining $M = \min(C, |\mathcal{S}|)$ samples. The approaches differ in the selection criteria $\mathcal{A}$ used for sample retention.

We evaluate three established replay baselines adapted from [23]: **Uniform Replay**, **Loss Replay**, and **Dissimilarity Replay**. To adapt the originally self-supervised Dissimilarity Replay to our supervised context, we compute pairwise samples distances using latent feature representations extracted from the pre-trained feature encoder h_{γ_A} as a structural proxy.

2.3 Proposed Sample Selection Mechanisms

The remaining strategies constitute our proposed methodological contributions, leveraging predictive uncertainty (**Uncertainty-Based Replay**), class-conditional uncertainty, and a hybrid selection (**Hybrid Pool-Based Replay**) approach.

Combination with Uniform Sampling. To balance the exploitation of highly informative samples with a broader, unbiased coverage of the source domain $\mathcal{D}_A$, the **Loss**, **Uncertainty-Based**, and **Hybrid** replay methods are combined with **Uniform Replay**. Given a targeted selection strategy $\mathcal{A}_s$ and a fixed uniform allocation ratio $r_u \in [0, 1]$, the updated buffer $\mathcal{M}'$ is constructed sequentially from the candidate pool $\mathcal{S}$:

$$\begin{aligned} \mathcal{M}' = \mathcal{M}_u \cup \mathcal{M}_s \text{ with } \mathcal{M}_u &= \mathcal{A}_{\text{unif}}(\mathcal{S}, \lfloor r_u M \rfloor), \\ \mathcal{M}_s &= \mathcal{A}_s(\mathcal{S}_{\text{rem}}, M - \lfloor r_u M \rfloor), \end{aligned} \tag{1}$$

where $M = \min(C, |\mathcal{S}|)$, s is a sample selection strategy, $\mathcal{A}_{\text{unif}}$ is the uniform sampling algorithm and $\mathcal{S}_{\text{rem}} = \mathcal{S} \setminus \mathcal{M}_u$ are remaining samples from the pool.

Importantly, we propose two distinct ways of executing this uniform selection: an instance-level (global) approach that samples uniformly without accounting

for class labels, and a class-conditional (stratified) approach that enforces class-proportional representation.

Uncertainty-based Replay. While Loss Replay prioritizes samples with high loss—often interpreted as "hard" or highly informative examples—it risks retaining outliers or label-noisy samples from the source domain $\mathcal{D}_A$. Conversely, while Dissimilarity Replay promotes geometric diversity within the latent space, it may overlook semantically critical regions situated near decision boundaries. To balance these concerns, we propose **Uncertainty-Based Replay**, a strategy that selects samples based on an uncertainty score, reflecting either inherent data ambiguity or model ignorance.

Total predictive uncertainty (TU) is standardly decomposed into *aleatoric uncertainty* (AU), which captures irreducible data noise (e.g., ambiguous pathologies), and *epistemic uncertainty* (EU), which reflects model ignorance due to limited training data. We approximate these quantities using MCD [9,12]. Leveraging these decomposed metrics, we formulate a composite memory update score designed to: (1) be *proportional to EU*, prioritizing rare edge cases and structurally diverse samples; (2) be *proportional to TU*, emphasizing points near complex decision boundaries; (3) be *inversely proportional to AU*, mitigating the detrimental influence of irreducibly noisy or ambiguous samples.

Prior to score computation, all uncertainty metrics are min-max normalized to $[0, 1]$ using statistics from the source domain, yielding $\tilde{\mathrm{EU}}$, $\tilde{\mathrm{TU}}$, and $\tilde{\mathrm{AU}}$. This preserves informative, high-signal samples while filtering out noise-dominated ones. The scalar acquisition score for a sample x is defined as:

$$s_{\mathrm{UQ}}(x) = w_E \cdot \tilde{\mathrm{EU}}(x) + w_T \cdot \tilde{\mathrm{TU}}(x) - w_A \cdot \tilde{\mathrm{AU}}(x) \tag{2}$$

where w_E, w_T, and w_A are weighting hyperparameters. Note that, although TU is a linear combination of AU and EU, the min-max normalized quantity $\tilde{\mathrm{TU}}$ is not. Therefore, including all three normalized terms does not introduce a redundant linear dependency in the acquisition score and is designed to capture complementary aspects of the data with a finer control.

Moreover, to prevent the retention of outliers and noisy samples, we discard a fraction α of the candidate pool exhibiting the highest aleatoric uncertainty, obtaining a new candidate pool $\mathcal{S}_{\mathrm{clean}}$. Consequently, the memory is updated by selecting the $M' = \min(M_s, |\mathcal{S}_{\mathrm{clean}}|)$ samples from the candidate pool that maximize this uncertainty score:

$$\mathcal{A}_{\mathrm{UQ}}(\mathcal{S}, M') := \left\{ s_i \in \mathcal{S}_{\mathrm{clean}} \;\middle|\; i \in \mathrm{argtop}_{M'} \left(s_{\mathrm{UQ}}(\mathcal{S}_{\mathrm{clean}}) \right) \right\} \tag{3}$$

Hybrid Pool-Based Replay. While targeted strategies like Loss Replay or Uncertainty-Based Replay are designed to identify highly informative examples, they risk concentrating the memory buffer around narrow regions of the feature space, reducing overall geometric diversity. To mitigate this, we propose a **Hybrid Update** strategy that integrates informativeness, outlier mitigation, and latent space coverage into a unified selection approach.

As established, the hybrid algorithm $\mathcal{A}_{\mathrm{hybrid}}$ is subsequently applied to the remaining candidate pool after selecting uniformly a fraction of the capacity.

Information Maximization, Noise Mitigation, and Aleatoric Filtering: First, we compute the normalized uncertainties using the MCD framework previously defined and the min-max normalized network loss $\tilde{\ell}(x)$ for all candidates. We define a composite acquisition score that rewards informative samples while penalizing irreducible noise:

$$s_{\text{hybrid}}(x) = w_E \cdot \tilde{\text{EU}}(x) + w_T \cdot \tilde{\text{TU}}(x) + w_L \cdot \tilde{\ell}(x) - w_A \cdot \tilde{\text{AU}}(x) \tag{4}$$

As for the **Uncertainty-Based Replay**, aleatoric filtering is applied.
Strategic Pooling: From the filtered set, we isolate a concentrated high-signal pool $\mathcal{P}$ by selecting the top candidates according to the composite score. The size of this pool is scaled by a multiplier hyperparameter $\rho > 1$ (e.g., $\rho = 3$):

$$\mathcal{P} = \left\{ s_i \in \mathcal{S}_{\text{clean}} \;\middle|\; i \in \text{argtop}_{\min(|\mathcal{S}_{\text{clean}}|,\, \rho M_s)} \left(s_{\text{hybrid}}(\mathcal{S}_{\text{clean}}) \right) \right\} \tag{5}$$

Geometric Diversity via Greedy K-Center: Finally, to ensure the selected subset is geometrically diverse, we perform a greedy K-Center core-set selection on the embedded feature space $h_{\gamma_A}(\mathcal{P})$. If $|\mathcal{P}| \leq M_s$, then $\mathcal{M}_s = \mathcal{P}$. Otherwise, we initialize the target buffer $\mathcal{M}_s$ with a randomly selected sample from $\mathcal{P}$. We iteratively append the sample $u \in \mathcal{P} \setminus \mathcal{M}_s$ that maximizes the minimum distance to the already selected set:

$$\mathcal{M}_s \leftarrow \mathcal{M}_s \cup \left\{ \arg \max_{u \in \mathcal{P} \setminus \mathcal{M}_s} \min_{v \in \mathcal{M}_s} \left\| h_{\gamma_A}(u) - h_{\gamma_A}(v) \right\|_2 \right\} \tag{6}$$

until exactly M_s samples are acquired. The final buffer is the union of the uniform and strategically selected sets: $\mathcal{A}_{\text{hybrid}}(\mathcal{S}, M) = \mathcal{M}_u \cup \mathcal{M}_s$.

3 Experiments and Results

3.1 Datasets, Tasks Definitions, and Experiment Setup

Datasets and Tasks: To evaluate our framework under realistic clinical domain shifts, we define sequential tasks across three datasets (summarized in Table 1):

OrganMNIST [24] (organs from abdominal CTs): Task A comprises axial slices (OrganAMNIST); Task B introduces coronal slices (OrganCMNIST). The original training, validation, and test splits are used, corresponding to approximately 60%, 10%, and 30% of the data, respectively.

Camelyon17 [11] (metastasis detection from histopathology): Task A uses data from Hospitals 1 and 2; Task B uses Hospitals 3 and 4. To expedite experiments, this dataset was undersampled to 2,500 samples per hospital-class stratum. 70% and 10% of the samples are used for training and validation sets.

HITS-5K (Private): 4,771 verified transcranial Doppler ultrasound signals (Solid Emboli, Gaseous Emboli, or Artefact) from 51 patients across 11 clinical centers. To simulate institutional shifts, data is split at the hospital level into Task A, Task B, with disjoint subject-level train/test subsets within each task

Dataset	Classes	Task shift	Task A samples	Task B samples
OrganMNIST	11 organs	view plane	41,052/17,778	15,367/8,216
Camelyon17	2 (metastasis presence)	hospital	48,852/17,667	69,849/40,403
HITS-5k	3 HITS types	hospital	2817/947	868/139

Table 1: Description of the datasets with the train/test distribution of samples.

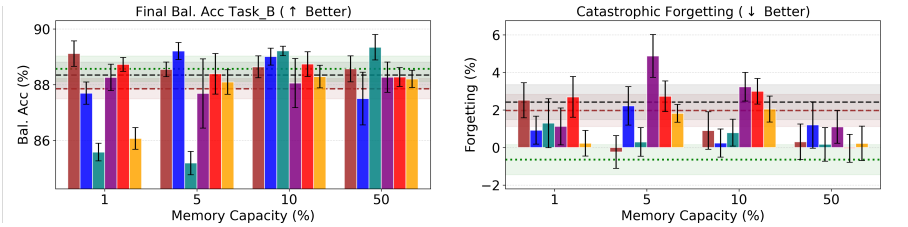
to prevent patient leakage. 15% of the training samples are randomly selected to form the validation set.

Architectures and optimization: We use dataset-specific models: a custom 3-block CNN for OrganMNIST, MobileNetV3-Small for Camelyon17, and a 2D Time-Frequency CNN [22] for HITS-5k. OrganMNIST and Camelyon17 networks expose their flattened penultimate representations for latent dissimilarity calculations and include classification-head dropout for MCD uncertainty estimation. To establish a rigorous baseline, we first optimize Task A hyperparameters (learning rate lr, weight decay wd) using Optuna [1] (30 trials) to maximize Task A balanced accuracy. Using the optimal configuration, we train n_{rep} independent base models ($n_{rep} = 10$ for OrganMNIST/HITS-5k; $n_{rep} = 5$ for Camelyon17) from scratch. Crucially, to isolate CL penalty effects, Task B hyperparameter search initializes directly from the best Task A model, explicitly maximizing the average balanced accuracy across Tasks A and B. We optimize a 3D space (lr, wd, $\lambda_{replay}/\lambda_{EWC}$) over 30 trials for standard baselines. For uncertainty-aware and hybrid methods, the search space expands (adding aleatoric drop fraction, pool multiplier, and w_a, w_e, w_h, w_l coefficients), so trials are scaled to 75 to ensure equitable convergence (except for hybrid strategy with Camelyon17, 30 trials kept). The ratio r_u , controlling the balance between uniform and active sampling, was fixed to 0.5.

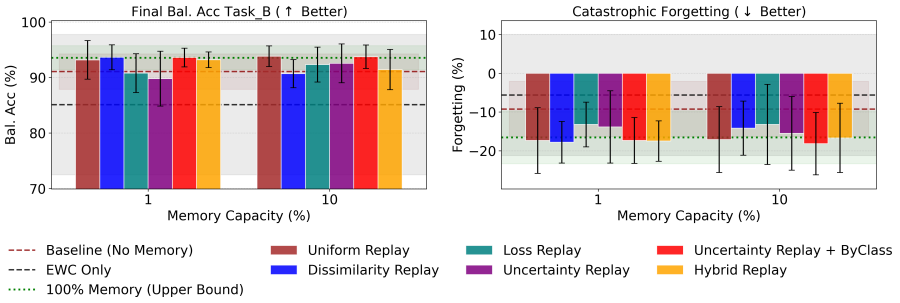
Evaluation & Statistics: Performance is quantified via standard classification metrics (Accuracy, Balanced Accuracy) alongside CL-specific metrics (Forgetting). Due to stochastic memory selection divergence, the n_{rep} runs act as independent samples. To control the family-wise Type I error rate when simultaneously evaluating three key objectives (Final Accuracy Task A, Final Accuracy Task B, Forgetting Task A), we assess significance using the non-parametric Mann-Whitney U test with a Bonferroni correction ($\alpha_{\text{adjusted}} = 0.05/3 \approx 0.0167$). All the codes used to run the different experiments can be found at: https://github.com/yamilvindas/uncertainty_based_CL.

3.2 Results

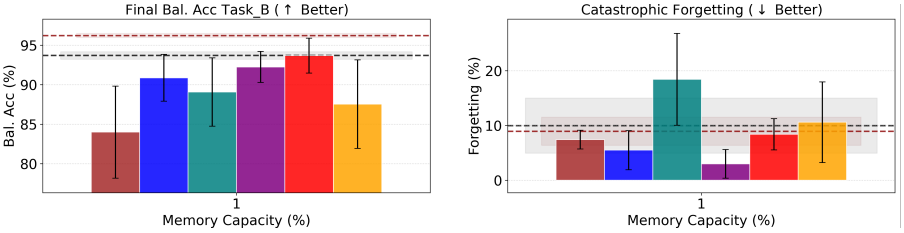
Catastrophic forgetting mitigation and memory size: Results can be found in Figures 1 (a)-(c). Baseline dynamics vary significantly by dataset: while OrganMNIST is intrinsically robust (baseline forgetting $1.96 \pm 0.85\%$, comparable to EWC, $p = 0.521$), Camelyon17 exhibits severe catastrophic forgetting ($8.93 \pm 2.56\%$). Conversely, the HITS-5k baseline exhibits a remarkable intrinsic positive backward transfer, yielding a negative forgetting rate of -9.25 $\pm$ 7.14%



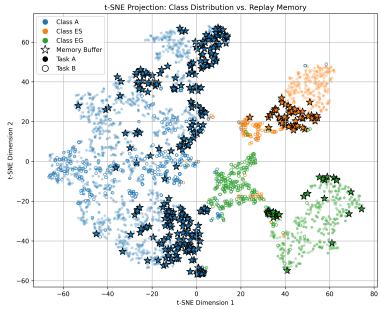
(a) OrganMNIST



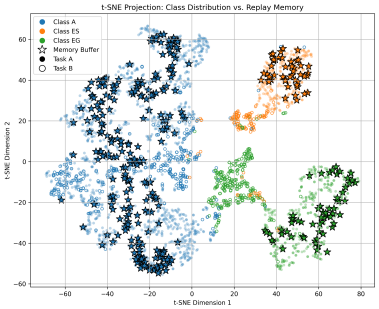
(b) HITS-5k



(c) Camelyon17



(d) Loss Replay



(e) Uncertainty ByClass

Fig. 1: **Continual learning performance scaling and latent space memory representation. (a)–(c)** Evaluation of final Task B balanced accuracy (↑) and task A catastrophic forgetting (↓) across varying memory capacities for OrganMNIST (top row), HITS-5k (second row), and Camelyon17 (third row). Horizontal reference markers denote the baseline performance configurations: memoryless training (dark red dashed line), EWC regularization (black dashed line), and the 100% full-memory gold standard (green dotted line), alongside their respective standard deviation error bands. **(d)-(e)** Qualitative representations of spatial memory coverage (bottom row) visualized on a representative model trained under a constrained 10% buffer capacity on the HITS-5k dataset.

(also statistically comparable to EWC, $p = 0.427$). Under severe memory constraints (1% capacity), active selection has a positive effect. For Camelyon17, all CL strategies except Loss Replay and Hybrid Replay reduce forgetting, at the expense of task B accuracy. Uncertainty Replay drastically cuts forgetting to $2.99 \pm 2.62\%$, vastly outperforming Uniform Replay ($7.42 \pm 1.70\%$). Similarly, for OrganMNIST, Dissimilarity, Hybrid, and Uncertainty Replay improve baseline forgetting ($p \leq 0.002$) while keeping similar Task B performance, whereas Uncertainty ByClass and Uniform Replay favor Task B accuracy ($p < 0.001$). For HITS-5k, results differ: Dissimilarity, Hybrid, Uncertainty ByClass, and Uniform Replay amplify the backward transfer to approximately -17% ($p < 0.002$) and simultaneously raise Task B accuracy from 91.06% to over 93%. At higher capacities (10% and 50%), OrganMNIST performances converge closely to the full-memory gold standard. Hybrid Replay and Uncertainty Replay ByClass outperform the global uncertainty strategy, improving both Task B accuracy and reducing forgetting. This suggests that balanced memory coverage across the feature space is beneficial for this dataset, which is further supported by the consistent performance of Uniform Replay across memory capacities. However, on HITS-5k at 10% capacity, all methods perform similarly due to substantial variance (standard deviations up to $\pm 10\%$). This indicates that for highly synergistic tasks like HITS-5k, maintaining a minimal episodic buffer is sufficient to unlock maximal cross-task facilitation, rendering complex sampling heuristics unnecessary at scale.

Memory coverage: To further analyze these results, we apply t-distributed stochastic neighbor embedding (t-SNE) [15] (perplexity 30) to the latent training data representations from $h_{\gamma_{A,B}}$. Figure 1(d)-(e) compares the resulting embeddings for the best- and worst-performing methods: Uncertainty ByClass and Loss Replay, respectively. While the overall manifold remains expectedly class-based (Solid Emboli vs. Gaseous Emboli), samples from Task A and Task B form distinct sub-clusters within each class, confirming the hospital-level domain shift. Regarding memory coverage, Loss Replay predominantly selects samples near cluster boundaries—likely the most challenging examples. In contrast, Uncertainty ByClass achieves a more balanced coverage across both central and boundary regions, likely explaining its superior performance. Furthermore, across multiple training runs, Loss Replay exhibits lower consistency in selected memory distributions, indicating higher sensitivity to initial Task A training.

4 Conclusion

We evaluated several replay-based continual learning strategies in a hospital-shift scenario across three real-world medical imaging datasets. Overall, replay methods effectively mitigate catastrophic forgetting, even under low-memory constraints. However, the best-performing strategy varies across datasets, highlighting that no single approach universally dominates. This underscores the value of our proposed methods, which leverage uncertainty estimates and parametrized hybrid criteria, for adapting to diverse clinical deployment settings. The per-

formance variability reflects real-world differences in data distributions, domain shifts, and class characteristics. Visual analysis of memory sample coverage via representation space projections provides valuable insight into how each strategy preserves knowledge, helping to align method choice with task semantics. This suggests that future continual learning systems in healthcare should support adaptive memory management, informed by both uncertainty and data representativeness.

Acknowledgments. This work was supported by a specific doctoral grant for Normaliens (CDSN) from the French Ministry of Higher Education and Research, awarded through ENS Paris-Saclay and by the Swiss National Science Foundation (SNSF) under grant number 229203 ("AIIDKIT: Artificial Intelligence for Improved Infectious Diseases Outcomes in Kidney Transplant Recipients"); the São Paulo Research Foundation (FAPESP – grants 2024/04761-0 and 19/07665-4); the National Research Council (CNPq 304805/2025-4); and the Coordination for the Improvement of Higher Education Personnel (CAPES – grant 001).

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Akiba, T., Sano, S., Yanase, T., Ohta, T., Koyama, M.: Optuna: A next-generation hyperparameter optimization framework. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. p. 2623–2631. Association for Computing Machinery, New York, USA (2019)
2. Baweja, C., Glocker, B., Kamnitsas, K.: Towards continual learning in medical imaging (2018). https://doi.org/10.48550/arXiv.1811.02496
3. Buzzega, P., Boschini, M., Porrello, A., Abati, D., Calderara, S.: 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. (2020)
4. Davari, M., Asadi, N., Mudur, S., Aljundi, R., Belilovsky, E.: Probing Representation Forgetting in Supervised and Unsupervised Continual Learning. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 16691–16700 (2022). https://doi.org/10.1109/CVPR52688.2022.01621
5. De Lange, M., Aljundi, R., Masana, M., Parisot, S., Jia, X., Leonardis, A., Slabaugh, G., Tuytelaars, T.: A continual learning survey: Defying forgetting in classification tasks. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(7), 3366–3385 (2022). https://doi.org/10.1109/TPAMI.2021.3057446
6. Farquhar, S., Gal, Y.: Towards Robust Evaluations of Continual Learning (2019). https://doi.org/10.48550/arXiv.1805.09733
7. Gholizade, M., Ruffini, F., Ducange, P., Marcelloni, F.: Federated continual learning: A comprehensive survey on lifelong and privacy-preserving learning over distributed and non-stationary data. Neurocomputing **694**, 133929 (2026)
8. Ju, J., Jung, H., Oh, Y., Kim, J.: Extending Contrastive Learning to Unsupervised Coreset Selection. IEEE Access **10**, 7704–7715 (2022). https://doi.org/10.1109/ACCESS.2022.3142758

9. Kendall, A., Gal, Y.: What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In: Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017)
10. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., Hassabis, D., Clopath, C., Kumaran, D., Hadsell, R.: Overcoming catastrophic forgetting in neural networks. Proceedings of the National Academy of Sciences **114**(13), 3521–3526 (2017). https://doi.org/10.1073/pnas.1611835114
11. Koh, P.W., Sagawa, S., Marklund, H., Xie, S.M., Zhang, M., Balsubramani, A., Hu, W., Yasunaga, M., Phillips, R.L., Gao, I., Lee, T., David, E., Stavness, I., Guo, W., Earnshaw, B., Haque, I., Beery, S.M., Leskovec, J., Kundaje, A., Pierson, E., Levine, S., Finn, C., Liang, P.: Wilds: A benchmark of in-the-wild distribution shifts. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. vol. 139, pp. 5637–5664. PMLR (2021)
12. Lambert, B., Forbes, F., Doyle, S., Dehaene, H., Dojat, M.: Trustworthy clinical AI solutions: A unified review of uncertainty quantification in Deep Learning models for medical image analysis. Artificial Intelligence in Medicine **150**, 102830 (2024). https://doi.org/10.1016/j.artmed.2024.102830
13. Liao, W., Xiong, H., Wang, Q., Mo, Y., Li, X., Liu, Y., Chen, Z., Huang, S., Dou, D.: Muscle: Multi-task self-supervised continual learning to pre-train deep models for x-ray images of multiple body parts. In: Wang, L., Dou, Q., Fletcher, P.T., Speidel, S., Li, S. (eds.) Medical Image Computing and Computer Assisted Intervention – MICCAI 2022. pp. 151–161. Springer Nature Switzerland, Cham (2022)
14. Lopez-Paz, D., Ranzato, M.: Gradient episodic memory for continual learning. In: Proceedings of the 31st International Conference on Neural Information Processing Systems. p. 6470–6479. NIPS'17, Curran Associates Inc. (2017)
15. Maaten, L.v.d., Hinton, G.: Visualizing Data using t-SNE. Journal of Machine Learning Research **9**(86), 2579–2605 (2008)
16. Park, J., Park, D., Lee, J.G.: Active learning for continual learning: Keeping the past alive in the present. In: The 13th International Conference on Learning Representations (2025)
17. Perkonigg, M., Hofmanninger, J., Herold, C.J., Brink, J.A., Pianykh, O., Prosch, H., Langs, G.: Dynamic memory to alleviate catastrophic forgetting in continual learning with medical imaging. Nature Communications **12**(1), 5678 (2021)
18. Qazi, M.A., Hashmi, A.U.R., Sanjeev, S., Almakky, I., Saeed, N., González, C., Yaqub, M.: Continual learning in medical imaging: a survey and practical analysis. ACM Computing Surveys **58**(8), 1–25 (2026). https://doi.org/10.1145/3785663
19. Quarta, A., Bruno, P., Calimeri, F.: Continual learning for medical image classification. In: Proceedings of the 1st AIxIA Workshop on Artificial Intelligence For Healthcare. CEUR Workshop Proceedings, vol. 3307, pp. 78–89 (2022)
20. Sahiner, B., Chen, W., Samala, R.K., Petrick, N.: Data drift in medical machine learning: implications and potential remedies. The British Journal of Radiology **96**(1150), 20220878 (2023). https://doi.org/10.1259/bjr.20220878
21. Verma, T., Jin, L., Zhou, J., Huang, J., Tan, M., Choong, B.C.M., Tan, T.F., Gao, F., Xu, X., Ting, D.S., Liu, Y.: Privacy-preserving continual learning methods for medical image classification: a comparative analysis. Frontiers in Medicine **10** (2023)
22. Vindas, Y., Guepie, B.K., Almar, M., Roux, E., Delachartre, P.: An hybrid cnn-transformer model based on multi-feature extraction and attention fusion mechanism for cerebral emboli classification. In: Lipton, Z., Ranganath, R., Sendak,

M., Sjoding, M., Yeung, S. (eds.) Proceedings of the 7th Machine Learning for Healthcare Conference. vol. 182, pp. 270–296. PMLR (2022)
23. Vindas, Y., Guillaud, M.: Dynamic Channel Charting: Integrating Online Sample Selection with Continual Learning for Streaming CSI Data. In: ICC 2025 - IEEE International Conference on Communications. pp. 3791–3796 (2025). https://doi.org/10.1109/ICC52391.2025.11161077
24. Yang, J., Shi, R., Wei, D., Liu, Z., Zhao, L., Ke, B., Pfister, H., Ni, B.: MedMNIST v2 - A large-scale lightweight benchmark for 2D and 3D biomedical image classification. Scientific Data **10**(1), 41 (2023). https://doi.org/10.1038/s41597-022-01721-8
25. Ye, Y., Xie, Y., Zhang, J., Chen, Z., Wu, Q., Xia, Y.: Continual Self-Supervised Learning: Towards Universal Multi-Modal Medical Data Representation Learning. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 11114–11124 (2024)
26. Zhao, B., Mopuri, K.R., Bilen, H.: Dataset condensation with gradient matching. In: International Conference on Learning Representations (2021)
27. Zheng, E., Yu, Q., Li, R., Shi, P., Haake, A.: A continual learning framework for uncertainty-aware interactive image segmentation. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 35, pp. 6030–6038 (2021). https://doi.org/10.1609/aaai.v35i7.16752
28. Zhou, Z., Shin, J.Y., Gurudu, S.R., Gotway, M.B., Liang, J.: Active, continual fine tuning of convolutional neural networks for reducing annotation efforts. Medical Image Analysis **71**, 101997 (2021)