Continual Domain Incremental Learning for Privacy-aware Digital Pathology

Pratibha Kumari[1*@][0000−0003−3681−3700], Daniel Reisenbüchler[1*][0000−0001−8820−0530], Lucas Luttner[1][0000−0003−1103−9077], Nadine S. Schaadt[2][0000−0001−7685−8087], Friedrich Feuerhake[2][0000−0002−1234−982X], and Dorit Merhof[1,3][0000−0002−1672−2185]

[1] Faculty of Informatics and Data Science, University of Regensburg, Regensburg, Germany
[2] Institute of Pathology, Hannover Medical School, Hannover, Germany
[3] Fraunhofer Institute for Digital Medicine MEVIS, Bremen, Germany
*Equal contribution, @Correspondance (Pratibha.Kumari@ur.de)

Abstract. In recent years, there has been remarkable progress in the field of digital pathology, driven by the ability to model complex tissue patterns using advanced deep-learning algorithms. However, the robustness of these models is often severely compromised in the presence of data shifts (e.g., different stains, organs, centers, etc.). Alternatively, continual learning (CL) techniques aim to reduce the forgetting of past data when learning new data with distributional shift conditions. Specifically, rehearsal-based CL techniques, which store some past data in a buffer and then replay it with new data, have proven effective in medical image analysis tasks. However, privacy concerns arise as these approaches store past data, prompting the development of our novel Generative Latent Replay-based CL (GLRCL) approach. GLRCL captures the previous distribution through Gaussian Mixture Models instead of storing past samples, which are then utilized to generate features and perform latent replay with new data. We systematically evaluate our proposed framework under different shift conditions in histopathology data, including stain and organ shift. Our approach significantly outperforms popular buffer-free CL approaches and performs similarly to rehearsal-based CL approaches that require large buffers causing serious privacy violations.

Keywords: Continual learning · Digital pathology · Domain shift

1 Introduction

Rapid advancements in deep learning models have revolutionized digital pathology. However, such models are mostly validated in stationary environments where data is presumed to follow a static distribution, which is usually not the case in clinical settings. Histopathology slides originating from different organs, staining protocols, centers, etc. possess various levels of covariance shifts [10] (see examples in Fig. 1). However, deep learning models would show a drastic performance degradation on datasets that do not follow the data distribution
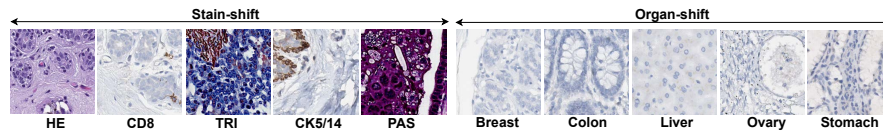
Fig. 1. Example of domain-shift in digital pathology. *Left*: different stainings of breast samples. *Right*: Hematoxylin samples of different organs for tumor detection.

initially used for training the model [13]. On the other hand, naively updating the model with training data of a new domain may cause disruption in the previously learned domain, also known as catastrophic forgetting [17]. One possible solution to prevent catastrophic forgetting is to retrain the deployed model on accumulated datasets of both past and new domains. However, sharing past datasets is often not feasible with medical data due to privacy concerns [24]. Also, past data may no longer be accessible, which would also hamper federated learning approaches. Further, such approaches demand considerable amounts of computation time and memory to store datasets and would require full model retraining each time distribution shifts in the dataset are encountered.

Alternatively, Continual Learning (CL) has emerged as a promising incremental learning paradigm to avoid catastrophic forgetting [7]. CL sequentially accumulates knowledge over a stream of datasets, frequently referred to as tasks, each with possible shifts, without truly revisiting the previous task. There are various CL strategies detailed in [18], all aiming to reduce catastrophic forgetting and tackle performance drops. Medical imaging researchers have also started exploring these CL strategies for various medical image analysis applications that are prone to encounter domain shifts and novel classes, including segmentation, disease classification, drug discovery, etc. [18]. In digital pathology, Bandi et al. [1] provide a comparative study of existing CL strategies on cancer detection datasets, reflecting shifts among three organs (breast, colon, and head-neck). Similarly, Kaustaban et al. [15] simulate five domain incremental datasets by changing stain composition in a colon cancer dataset (CRC) and benchmark existing CL approaches. Some other works [7,29] benchmark popular CL approaches on less complex datasets such as MedMNIST. Further CL research in digital pathology indicates that rehearsal-based strategies that require storing some past samples in a memory buffer tend to perform better compared to others [1,15]. However, storing past samples may cause privacy violations [12], which can be a major bottleneck of such CL approaches in medical applications. A less concerning direction, storing features instead of actual images, refereed as latent-replay (e.g., chest x-ray classification [26], Ultrasound cardiac view classification [24]) is also explored. Besides the promising results from current research, such a strategy is compromised by complexities in sharing features across hospitals and large buffer requirements.

To address these limitations related to the sharing of medical data in histopathology, we propose a novel, buffer-free CL approach that takes advantage of a Gaussian Mixture Model (GMM) to encapsulate the traits of past training
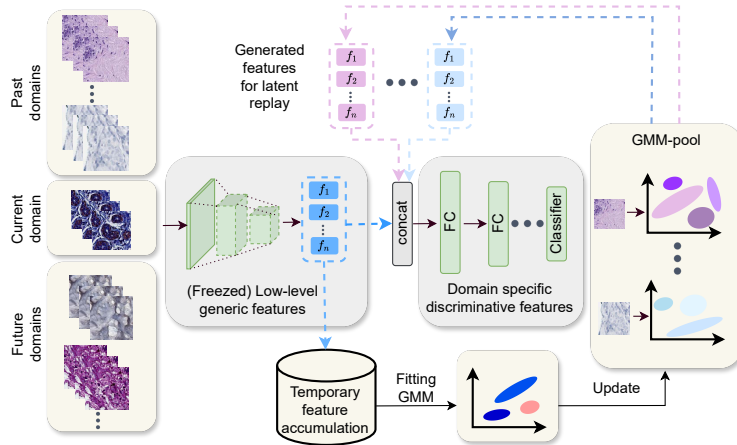
Fig. 2. Overview of proposed CL framework for digital pathology

domains, enabling feature generation for incorporation in subsequent training sessions. We acquire real-world histological slides from different environments and curate domain-incremental scenarios to show shifts in terms of stain, organ, center, or a mix of those. An extensive evaluation is performed on three different domain shift problems in digital pathology, i.e., (1) breast tumor detection across different histological stains, (2) tumor detection across different organs, and (3) tumor detection in the presence of heterogeneous types of shifts. For comparison, we consider established buffer-free and buffer-based CL approaches.

2   Method

In the following, we provide the problem formulation and introduce the two key novelties of our proposed CL framework: (a) generative latent replay and (b) domain specific generators. Fig. 2 serves as overview of our proposed framework.

Problem formulation: The goal of GLRCL is to learn sequentially from datasets containing domain shifts and without storing any samples from past datasets while maintaining performance over previous datasets. Such an incremental learning setting where the classification task is fixed, but domain shift is observed over time, is referred to as continual learning under domain incremental scenarios. Each dataset in the sequence is termed as a CL **task** here. Formally, there is a sequence of tasks $\{T_1, T_2, ...\}$, where each task T_t has training D_t and evaluation E_t set for tumor classification. Here, the training set $D_t = \{x_t, y_t\}$ contains x_t patches extracted from Whole Slide Images (WSIs) and corresponding y_t annotations for $\{1, ...C\}$ classes. The training data of tth task offers samples for learning in the current session (t^{th} domain) only and is not accessible in past or future training sessions. Meanwhile, the testing data from each task is available throughout the sessions. The CL classification model

$\mathbb{M}_t(\cdot)$ is initialized with model parameters θ_{t-1} learned during the immediate past session. $\mathbb{M}_t(\cdot)$ then aims to mitigate catastrophic forgetting, by learning the model parameters θ_t that minimize the loss ($\boldsymbol{L}$) over all the past domains. In the proposed GLRCL framework, $\mathbb{M}_t(\cdot)$ learns jointly on the training data from the current domain D_t and latent vectors produced from a pool of generators representing past domains.

a) Generative latent replay: The early layers of deep learning models are responsible for extracting low-level features from input images. After undergoing pre-training on large datasets, these layers' weights become stable and can be effectively reused in various applications, including medical image processing tasks. On the other hand, layers close to the classification head tend to extract discriminative features tailored to specific classes and domains, and fine-tuning them is often essential for maximizing accuracy. We propose to extract rich low-level features from an early layer, coined as generative-replay-layer, and learn generators from these activations. Thus, rather than storing input histopathology samples in an external memory in raw or latent form; we just store the generators and hence avoid possible privacy violation. At the arrival of any new domain, separate generators are learned on tumor and healthy images and then the pool containing generators is updated with these newly learned generators. These generators then can be used to produce latent representations mimicking past domains which serve as replay features. In each Adam-based mini-batch training, patterns coming from the input layer are concatenated with the replay features at the generative-replay-layer on the mini-batch dimension.

$$\arg\min_{\theta_t} \boldsymbol{L}\left(\mathbb{M}_{t(\bar{\theta},\theta_t)}((x_t, y_t) \oplus (\mathbb{G}_{t-1}(\mathbf{f}_{t-1}), y_{t-1}) \oplus \ldots), \forall t > 0\right)$$

where, $\bar{\theta}$ refers to freezed parameters, $\mathbb{G}_{t-1}(\cdot)$ represents $(t-1)^{th}$ domain generator, and $\oplus$ indicates concatenation operation. To uphold the stability of the generated features and the validity of stored generators, we propose freezing all layers beneath the generative-replay-layer and allowing the layers above to learn autonomously, i.e., the forward and backward passes are performed only on layers after generative-replay-layer. Our approach offers applicability to a wide range of encoder models. We use a pre-trained ResNet50 [11] as our feature extraction model. Then fully connected layers and a classification head are added.

b) Domain specific generators: For each of the encountered domains having a unique tuple of attributes in set {stain, organ, center}, we propose to maintain light-weight generators, separately for each class. The generators are learned from extracted image features and hence cannot be used to produce raw images, ensuring data privacy. Moreover, this opens an avenue for restricted data-sharing in terms of distribution parameter sharing. We utilize GMMs as the feature generators. GMMs are well known for their diverse range of data modeling capabilities, also evidenced by their popularity in encoding latent features in deep layers [2,27], encoding raw images [23], features [28,30], etc. Further,

they typically require fewer parameters (only mean and variance) compared to any deep neural network-based generators which would otherwise require storing the entire model weights. The cost-effective generators are particularly advantageous, especially in CL scenarios involving multiple distinct domains that require a large number of generators. Specifically, during training, we extract features from the last batch normalization layers of the ResNet50 model. Then, at the end of each training session, the accumulated features $\mathbf{f}_t \leftarrow BN(x_t)$ are used to learn domain-specific GMMs, for each class. For t^{th} domain and i^{th} class, the generator $\mathbb{G}_t^i(\cdot)$ posses K multivariate Gaussians $\{g_1, g_2....g_K\}$ in the mixture with probability density function (PDF) as:

$$p(\mathbf{f}_t) = \sum_{k=1}^{K} w_k \cdot \mathcal{N}(\mathbf{f}_t \mid \mu_k, \Sigma_k); \; \mu_k = \frac{1}{N_t}\sum_{n=1}^{N_t} \mathbf{f}_t^n; \; \Sigma_k = \frac{1}{N_t}\sum_{n=1}^{N_t}(\mathbf{f}_t^n - \mu_k)(\mathbf{f}_t^n - \mu_k)^T$$

where, N_t denotes the number of samples in t^{th} domain for a particular class, $\mathcal{N}(\mathbf{f} \mid \mu_k, \Sigma_k)$ represents PDF of each kth Gaussian component g_k with mean μ_k and covariance matrix Σ_k, and w_i is a positive weight associated with k^{th} Gaussian such that $\sum_{k=1}^{K} w_k = 1$. For a given value of K, the parameters μ_k, Σ_k for $k \in \{1, 2...K\}$ of mixture distribution are estimated via expectation maximization [6]. The fitted mixture model $\mathbb{G}_t^i$ allows to generate random samples reflecting the ith class and t^{th} domain. Here, the optimal value of K is learned using the Bayesian Information Criterion (BIC) [9]. The candidate values for K lie in range 1 and a given upper limit K_{max}, and the one that minimizes BIC is selected. Finally, the learnt generators from t^{th} domain are then added to GMM-pool containing all past generators $\mathbb{G}_t^i$ for $i \in \{1, \ldots, C\}$ and $t \in \{1, \ldots, t-1\}$.

3 Experimental Setup

(1) Datasets: Various datasets characterizing shifts in staining, organ, and center are curated. We outline these datasets in Table 1. Please note that train-test split does not follow slide-level split except for the organ shift datasets.

   Stain shift (SS): The data was acquired at center C1 and includes healthy and tumorous breast tissue. Experienced pathologists carefully annotated breast cancer regions in detail for Hematoxylin and eosin (HE), Cytokeratin (CK5/14), and Cluster of differentiation (CD8) stained WSIs. For preprocessing, we selected foreground tissue regions using the CLAM preprocessing pipeline [21]. As stainings highly vary in colorization, we manually adjusted tissue detection

Table 1. Details of shift types and available datasets used in experiments

Details of shift	Datasets	#Train patches	#Test patches
SS: Staining varies, organ (breast), center (C1)	HE, CD8, TRI, CK5/14, PAS	1510, 3370,4706, 4704, 3372	1000 each
OS: staining (H), organ varies, center (C2),	Breast, Colon, Liver, Lung, Ovary, Pancreas, Prostate, Stomach, Uterus	1704, 1982, 1694, 1472, 1338, 1308, 1916, 1836, 1900	600 each
HS: Heterogeneous	CD8/breast/C1, HE/breast/C3 (BRACS), H/colon/C2, HE/colon/C4 (CRC)	3370, 6096, 1982, 6600	600-1000

Table 2. Best performance result in buffer-based / buffer-based with low buffer / buffer-free categories indicated in red / blue / green, respectively. **Bold:** Upper bound.

	Exp.→	SS			OS			HS		
	Approach	BWT↑	Acc.↑	ILM↑	BWT↑	Acc.↑	ILM↑	BWT↑	Acc.↑	ILM↑
Non-CL	Naive	-31.09	71.72	77.54	-7.03	79.41	82.72	-38.80	66.92	73.88
	Joint	–	98.04	–		95.46	–	–	86.73	–
	Cumulative	**-0.30**	**98.00**	**98.18**	**8.12**	**95.80**	**94.99**	**-0.07**	**88.67**	**93.77**
Buffer-based	GEM	-0.83	86.64	83.87	-1.77	88.70	89.02	-19.63	69.11	80.93
	AGEM	5.63	91.64	89.83	-2.18	87.39	89.38	-20.07	69.31	80.53
	ER	-1.62	96.80	97.10	7.83	96.22	94.32	-2.60	90.01	93.53
	LR	-1.06	96.72	97.20	5.99	94.61	93.71	-2.17	87.54	92.89
	ER*	-3.42	94.12	95.87	2.65	95.50	93.64	-8.08	89.07	91.94
	LR*	-2.41	95.28	95.97	5.97	93.76	92.87	-5.02	85.97	91.54
Buffer-free	SI	-26.86	78.56	79.83	-6.77	82.31	84.51	-35.92	70.71	75.45
	LwF	-28.99	70.00	78.28	-4.15	83.44	84.45	-30.32	74.72	77.50
	EWC	-27.16	77.56	80.15	-9.12	81.00	82.69	-39.42	66.41	73.54
	Proposed	-1.65	96.20	96.76	6.43	94.48	93.23	-2.58	87.38	92.66

thresholds. Each annotated region was tiled into 512×512 px patches at $40\times$ magnification. Since we obtained $5\times$ more patches from CK5/14 and CD8 compared to HE and to include more inter-stain variation, we used half of both datasets and augmented them using Cycle-GAN (cGAN) into artificial trichome (TRI) and periodic acid-Schiff (PAS) stainings, respectively.

Organ shift (OS): This dataset from center C2 comprises 20 tissue microarray (TMA) cores, occupied with tissue of 9 organs in hematoxylin (H) staining. TMA cores were systematically sampled from a diverse cohort of patients and contained cancerous and healthy tissue. The same preprocessing strategy as in stain-variation is applied here. We split datasets spot-wise to prevent data leakage at both, patient and tissue level.

Heterogeneous shift (HS): In addition to homogeneous kinds of shift (i.e., either SS or OS), we also consider heterogeneous shifts involving a mix of different stains, organs, and centers, simultaneously. Specifically, we consider the "CD8" dataset from the SS sequence (stain:CD8, organ:breast, center:C1), the "Colon" dataset from the OS sequence (stain:H, organ:colon, center:C2), a subset from the BRACS dataset [4] (stain:HE, organ:breast, center:C3), and a subset from the CRC dataset [14] (stain:HE, organ:colon, center:C4). All these datasets are curated such that they comprise only normal and tumor classes.

(2) Comparable Methods: Both non-CL and CL approaches are considered for comparison. The non-CL approaches include (a) lower-bound performance obtained by a naive method (i.e. training dataset of the current task is used only) and (b) upper-bound performance by a joint method (i.e. training datasets from all tasks are used) and a cumulative method (training datasets are accumulated from seen tasks). For CL, we consider frequently used approaches in the two categories as (a) buffer-based (i.e. a few past samples are stored in raw or latent

form) and (b) buffer-free (no storing of past samples). In the buffer-based category, we consider GEM [20], A-GEM [5], ER [25], and LR [22]. In the buffer-free category, EWC [17], SI [31], and LwF [19] are considered.

(3) Implementation details: In the following, we describe the training strategies of our models and the included evaluation metrics.

cGAN pre-training: For cGAN pre-training we extracted patches in the scale of 512×512 px from TRI and PAS stained WSIs as provided by the KPMP database [16]. Regarding hyperparameter settings, we followed the literature [3] and performed a grid search around an order of magnitude around the base parameters. We selected the best-performing model by visually inspecting the derived stain augmentations, resulting in the number of iterations of $30 \in [20, 30, 40]$ epochs and a learning rate of $1.5e-4 \in [1.0e-3, 1.5e-3, \ldots, 1e-5]$. To prevent any data leakage, we used unlabeled excess tiles from CK5/14 and CD8 from our stain-variation WSIs to train our cGAN using target domains from the KPMP dataset. A total of 5000 images were used for training. The training and inference time was approximately 8 hours per stain.

CL training: We selected the hyperparameters according to common CL evaluation schemes. For all comparable methods, we used the CL benchmark library Avalanche[4]. We used AdamW as an optimizer and used learning of $1e-03$. Images were resized to 256×256 px and batched with a size of 64. All methods used a pre-trained ResNet50 model initialized with ImageNet weights. After the last batch normalization layer in ResNet50, 5 fully connected layers {512, 256, 128, 64, 32} are added. For all the experiments, we keep $K_{max} = 10$. All experiments were conducted using an Nvidia A100 GPU, with training sessions typically completed within a maximum duration of 4 hours across all approaches.
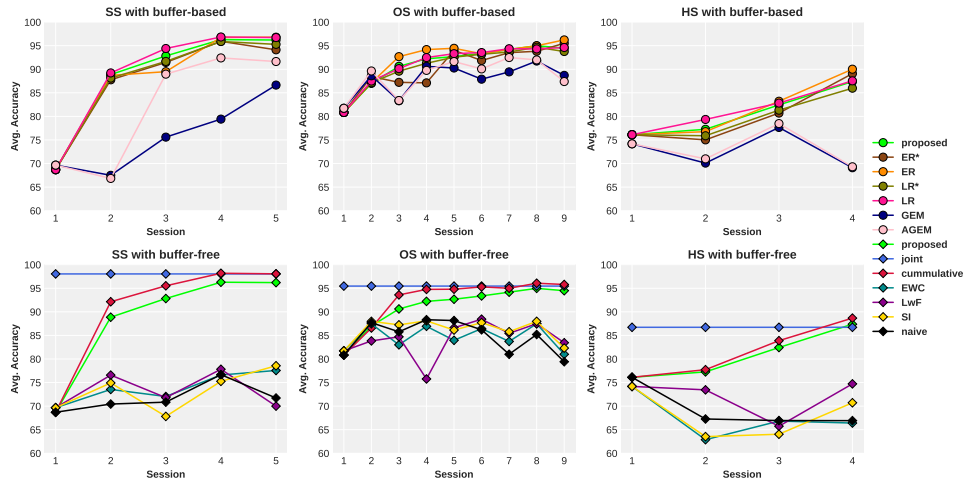
Evaluation metrics: The performance is assessed using the classification accuracy metric. For any experiment with T tasks in sequence, we get a train-test matrix A of size $T \times T$ where the cell value A_{ij} represents accuracy on j^{th} task after training up to i^{th} task. This matrix can be used to compute various metrics to compare approaches. We consider popularly used metrics in CL literature, specifically Backward Transfer (BWT) for measuring forgetting (Eq. 3 in [8]), Incremental Learning Metric (ILM) to measure incremental learning capability (Eq. 2 in [8]) and average accuracy across all the tasks after learning the last task (Eq. 2 in [20]). For all these metrics a higher positive value indicates superiority.

4   Results and Discussion

Within each shift scenario (SS, OS, HS), a random ordering of the available data is created and utilized for evaluating different approaches (see Table 2). Our supplementary file provides results on further random orderings (showing same trends) and further discussion on individual evaluation metrics. As expected, "cumulative" and "joint" represent the upper, and "naive" indicates the lower

---

[4] https://avalanche.continualai.org/

Fig. 3. Performance over time for different shifts (SS, OS, HS) and for buffer-based *(upper row)* versus buffer-free CL approaches + non-CL *(lower row)*.

bound. For all three evaluation metrics, there is a large performance gap between buffer-based approaches and buffer-free approaches across all experiments (SS, OS, and HS). This is in line with previous findings that rehearsal-based approaches handle forgetting in medical data better [18,7]. In the buffer-based category, ER and LR approaches largely offer the best results. Our buffer-free GLRCL approach (green) achieves performances slightly worse than the best-performing buffer-based approaches with a large buffer (red). However, when reducing the buffer size by 1/4 (see ER* and LR*), a significant performance drop is observed (best results in blue), resulting in performance values mostly below our GLRCL approach in all experiments. This shows that the performance of buffer-based approaches gets compromised when reducing the buffer size. When comparing with existing buffer-free CL approaches, we observe that our novel GLRCL method clearly outperforms these methods by a large margin across all three domain shift conditions. Further, in terms of accuracy (Acc. and ILM), HS performs generally worse than SS and OS, due to the multitude of variations (stain, organ, center shifts) encountered in HS.

To analyze performance over time, we report the average accuracy of the cancer classification tasks after each training session for each of our three domain shift experiments (see Fig 3). The first row shows a comparison of our GLRCL approach (green, "proposed") with buffer-based, and the second row with buffer-free and non-CL approaches, respectively. These graphs clearly show that GLRCL significantly outperforms the buffer-free approaches and performs similarly to the buffer-based approaches, creating a new CL benchmark and an alternative to buffer-based CL approaches for digital pathology.

5   Conclusion

Our novel privacy-aware CL approach for histopathology tumor detection in domain incremental scenarios outperforms existing buffer-free CL approaches by a large margin, and mitigates data privacy concerns compared to image-storing buffer-based CL methods with comparable performance. Our work significantly advances privacy-aware tumor detection from histology data, pushing the boundaries of current CL strategies under privacy preservation. For future work, we investigate for a better choice of latent representations so that the learned GMM better captures the domain and class differences. Additionally, we aim for a single generator with a dynamic update mechanism triggered with drift detection module which eventually deals with reoccurring and overlapping domains.

Disclosure of Interests The authors declare that they have no conflicts of interest related to this work.

References

1. Bándi, P., Balkenhol, M., van Dijk, M., Kok, M., van Ginneken, B., van der Laak, J., Litjens, G.: Continual learning strategies for cancer-independent detection of lymph node metastases. Medical Image Analysis **85**, 102755 (2023)
2. Bengio, Y., Courville, A., Vincent, P.: Representation learning: A review and new perspectives. IEEE transactions on pattern analysis and machine intelligence **35**(8), 1798–1828 (2013)
3. Bouteldja, N., Hölscher, D.L., Klinkhammer, B.M., Buelow, R.D., Lotz, J., Weiss, N., Daniel, C., Amann, K., Boor, P.: Stain-independent deep learning–based analysis of digital kidney histopathology. The American Journal of Pathology **193**(1), 73–83 (Jan 2023)
4. Brancati, N., Anniciello, A.M., Pati, P., Riccio, D., Scognamiglio, G., Jaume, G., De Pietro, G., Di Bonito, M., Foncubierta, A., Botti, G., et al.: Bracs: A dataset for breast carcinoma subtyping in h&e histology images. Database **2022**, baac093 (2022)
5. Chaudhry, A., Ranzato, M., Rohrbach, M., Elhoseiny, M.: Efficient lifelong learning with a-gem. arXiv preprint arXiv:1812.00420 (2018)
6. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the royal statistical society: series B (methodological) **39**(1), 1–22 (1977)
7. Derakhshani, M.M., Najdenkoska, I., van Sonsbeek, T., Zhen, X., Mahapatra, D., Worring, M., Snoek, C.G.: Lifelonger: A benchmark for continual disease classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 314–324. Springer (2022)
8. Díaz-Rodríguez, N., Lomonaco, V., Filliat, D., Maltoni, D.: Don't forget, there is more than forgetting: new metrics for continual learning. arXiv preprint arXiv:1810.13166 (2018)
9. Fraley, C., Raftery, A.E.: How many clusters? which clustering method? answers via model-based cluster analysis. The computer journal **41**(8), 578–588 (1998)
10. Guan, H., Liu, M.: Domain adaptation for medical image analysis: a survey. IEEE Transactions on Biomedical Engineering **69**(3), 1173–1185 (2021)
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Holub, P., Müller, H., Bíl, T., Pireddu, L., Plass, M., Prasser, F., Schlünder, I., Zatloukal, K., Nenutil, R., Brázdil, T.: Privacy risks of whole-slide image sharing in digital pathology. Nature Communications **14**(1), 2577 (2023)
13. Karani, N., Chaitanya, K., Baumgartner, C., Konukoglu, E.: A lifelong learning approach to brain mr segmentation across scanners and protocols. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 476–484. Springer (2018)
14. Kather, J.N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C.A., Gaiser, T., Marx, A., Valous, N.A., Ferber, D., et al.: Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLoS medicine **16**(1), e1002730 (2019)
15. Kaustaban, V., Ba, Q., Bhattacharya, I., Sobh, N., Mukherjee, S., Martin, J., Miri, M.S., Guetter, C., Chaturvedi, A.: Characterizing continual learning scenarios for tumor classification in histopathology images. In: International Workshop on Medical Optical Imaging and Virtual Microscopy Image Analysis. pp. 177–187. Springer (2022)

16. Kidney Precision Medicine Project: Kidney Precision Medicine Project Data. Accessed October 15, 2023. https://www.kpmp.org, the results here are in whole or part based upon data generated by the Kidney Precision Medicine Project. Funded by the National Institute of Diabetes and Digestive and Kidney Diseases

17. Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A.A., Milan, K., Quan, J., Ramalho, T., Grabska-Barwinska, A., et al.: Overcoming catastrophic forgetting in neural networks. Proceedings of the national academy of sciences **114**(13), 3521–3526 (2017)

18. Kumari, P., Chauhan, J., Bozorgpour, A., Azad, R., Merhof, D.: Continual learning in medical imaging analysis: A comprehensive review of recent advancements and future prospects. arXiv preprint arXiv:2312.17004 (2023)

19. Li, Z., Hoiem, D.: Learning without forgetting. IEEE Transactions on Pattern Analysis and Machine Intelligence **40**(12), 2935–2947 (2018)

20. Lopez-Paz, D., Ranzato, M.: Gradient episodic memory for continual learning. Advances in neural information processing systems **30** (2017)

21. Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F.: Data-efficient and weakly supervised computational pathology on whole-slide images. Nature Biomedical Engineering **5**(6), 555–570 (2021)

22. Pellegrini, L., Graffieti, G., Lomonaco, V., Maltoni, D.: Latent replay for real-time continual learning. In: 2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 10203–10209. IEEE (2020)

23. Pfülb, B., Gepperth, A.: Overcoming catastrophic forgetting with gaussian mixture replay. In: 2021 International Joint Conference on Neural Networks (IJCNN). pp. 1–9. IEEE (2021)

24. Ravishankar, H., Venkataramani, R., Anamandra, S., Sudhakar, P., Annangi, P.: Feature transformers: privacy preserving lifelong learners for medical imaging. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2019: 22nd International Conference, Shenzhen, China, October 13–17, 2019, Proceedings, Part IV 22. pp. 347–355. Springer (2019)

25. Rolnick, D., Ahuja, A., Schwarz, J., Lillicrap, T., Wayne, G.: Experience replay for continual learning. Advances in Neural Information Processing Systems **32** (2019)

26. Srivastava, S., Yaqub, M., Nandakumar, K., Ge, Z., Mahapatra, D.: Continual domain incremental learning for chest x-ray classification in low-resource clinical settings. In: MICCAI Workshop on Domain Adaptation and Representation Transfer. pp. 226–238. Springer (2021)

27. Viroli, C., McLachlan, G.J.: Deep gaussian mixture models. Statistics and Computing **29**, 43–51 (2019)

28. Yang, B., Lin, M., Zhang, Y., Liu, B., Liang, X., Ji, R., Ye, Q.: Dynamic support network for few-shot class incremental learning. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(3), 2945–2951 (2022)

29. Yang, H., Huang, W., Liu, J., Li, C., Wang, S.: Few-shot class-incremental learning for cross-domain disease classification. arXiv preprint arXiv:2304.05734 (2023)

30. Yang, Y., Cui, Z., Xu, J., Zhong, C., Wang, R., Zheng, W.S.: Continual learning with bayesian model based on a fixed pre-trained feature extractor. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 397–406 (2021)

31. Zenke, F., Poole, B., Ganguli, S.: Continual learning through synaptic intelligence. In: International conference on machine learning. pp. 3987–3995. PMLR (2017)