Continual Domain Incremental Learning for Privacy-aware Digital Pathology (Supplementary)

Pratibha Kumari[1⋆@], Daniel Reisenbüchler[1⋆], Lucas Luttner[1], Nadine S. Schaadt[2], Friedrich Feuerhake[2], and Dorit Merhof[1,3]

[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)

1 Discussion for SS, OS, and HS

Here, we provide further discussions on the results reported in **Table 2 of the main manuscript**. Across all the approaches, we see a comparatively better value of BWT in OS as compared to SS and HS experiments. This is because in OS, A_{ii} was very low (58.83 in cumulative) for the "Uterus" dataset and it jumped to a very high value (92.67) with the introduction of other organs starting from the "Ovary" dataset, i.e., $A_{ji} > A_{ii} \forall j > i$. Further, similarity among the organs causes a positive reinforcement to each other, and hence overall a lower forgetting is observed. Further, in SS, the AGEM approach suffers from a plasticity issue (hard to learn new knowledge) while learning the CK5/14 staining, which was subsequently reinforced to a large extent by future stainings, and hence reports exceptionally high BWT as compared to other approaches. However, other metrics (Acc. and ILM) are low compared to ER and LR approaches. Therefore, analyzing all metrics is important for a detailed comparison.

Table 1. Best performance result in buffer-based / buffer-based with low buffer / buffer-free categories indicated in red / blue / green, respectively.

	Seq.→	SS-2			OS-2			HS-2		
	Approach	BWT	Acc.	ILM	BWT	Acc.	ILM	BWT	Acc.	ILM
Buffer-based	GEM	-8.96	88.70	88.09	0.36	90.80	91.18	-22.67	72.52	79.18
	AGEM	-11.57	83.62	88.37	-3.38	92.61	91.06	-20.28	78.66	82.07
	ER	-1.80	96.44	96.71	1.41	95.39	95.33	-7.43	86.10	83.66
	LR	-1.25	92.24	92.68	-0.04	94.24	94.84	-4.18	83.45	83.99
	ER*	-4.42	95.88	95.06	-0.32	95.13	94.98	-11.41	83.23	81.16
	LR*	-2.54	90.66	92.25	-0.73	94.13	94.08	-7.59	80.67	82.08
Buffer-free	SI	-24.92	78.36	81.34	-4.87	88.17	88.82	-31.27	66.45	76.93
	LwF	-24.80	78.02	80.93	-4.89	88.20	89.66	-27.24	67.41	76.87
	EWC	-29.77	70.14	78.19	-6.45	90.17	86.49	-33.93	65.67	75.37
	Proposed	-2.00	91.78	92.57	-1.17	94.24	94.06	-1.56	89.32	85.72

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

	Seq.→	SS-3			OS-3			HS-3		
	Approach	BWT	Acc.	ILM	BWT	Acc.	ILM	BWT	Acc.	ILM
Buffer-based	GEM	-26.21	79.60	79.58	-4.76	88.15	87.27	-13.80	75.55	77.76
	AGEM	-14.73	92.24	88.38	-4.13	85.81	87.10	-8.48	80.48	79.59
	ER	-1.25	97.10	97.86	5.05	96.07	93.14	-0.58	89.31	89.26
	LR	-1.28	93.04	94.86	0.03	93.61	93.16	-6.77	87.46	89.25
	ER*	-3.61	93.16	96.16	2.93	94.28	90.81	-6.19	81.03	83.76
	LR*	-3.35	90.40	93.59	-0.65	92.00	91.71	-10.24	83.88	87.37
Buffer-free	SI	-34.67	67.24	74.99	-4.43	88.24	86.50	-38.72	68.94	71.26
	LwF	-33.36	70.42	76.13	-1.46	86.11	86.64	-28.07	67.58	70.02
	EWC	-33.14	69.70	76.54	-8.13	80.43	83.66	-34.50	70.72	73.27
	Proposed	-1.05	92.18	94.45	0.03	92.98	91.99	-7.82	85.83	89.16

2 Experiments with other random ordering of tasks

We analyze performance on two other random ordering of datasets in three domain shift experiments, named: {SS-2, OS-2, HS-2} and {SS-3, OS-3, HS-3}.

SS-2, OS-2, HS-2 (Table 1): We can see that best performing approach in buffer-based category (red) outperforms buffer-free CL approaches. However, when the available buffer size is reduced their performance is greatly compromised. Whereas, our proposed buffer-free approach consistently perform well across various shifts. Interestingly, it surpass best performing buffer-based approach in HS-2 experiment. Upon deeply analyzing the train-test matrix of ER, LR, and cumulative, we found that the learning of "Colon" dataset (1st task) is hampered while learning others due to low training amount of this dataset as compared to others. However, this does not happen in our approach because we do not have any restriction in amount of data generated; we generate past domain samples in amount same as current learning domain. So latent vector for "Colon" were generated in large amount as available for current domain, leading to less forgetting of "Colon" (past domain).

SS-3, OS-3, HS-3 (Table 2): As found in other orderings, here also we can spot that BWT is better in OS-3 compared to SS-3 and HS-3. This is attributed to the fact that the organ datasets are positively reinforcing each other and hence performance on a particular organ is improved when a complementary organ is encountered in future tasks. We observe a better value of BWT with ER approach in HS-3 compared to others; this is attributed to inadequate learning of this dataset (T_j) and hence $A_{ij}, \forall i \geq j$ was mostly similar; wheareas with other approaches $A_{ij}, \forall i > j$ were reduced.

As expected, buffer-based approaches, except GEM and AGEM, work better than buffer-free approaches. However, under limited memory, their BWT, Acc., and ILM are largely reduced. Lastly, we can seen in buffer-free category, proposed approach consistently outperforms across all three domain shift experiments.