Continual Model Merging with Test-Time Adaptation for Whole-Slide Image Analysis

No Author Given

No Institute Given

Abstract. Model merging offers a practical alternative to conventional continual learning by integrating independently fine-tuned models without retaining previous training data. Recent *state-of-the-art model merging methods employ test-time adaptation (TTA-guided merging)* to address distribution shifts by adjusting merging-related variables using unlabeled target data. However, these methods have primarily been studied in multi-task or single-target settings, and their behavior under sequential continual learning remains insufficiently understood. We present a benchmark study that maps this family of methods to rehearsal-free continual Whole Slide Image classification and evaluates them against traditional continual-learning approaches. Experiments on six TCGA cancer-subtyping cohorts cover CLASS-IL and TASK-IL scenarios, in-domain and out-of-domain evaluation, and different task orders. The results show that adapting model merging at test time can provide strong task-specific performance and improve retention of previously acquired knowledge without storing historical WSIs. Nevertheless, performance remains sensitive to task order and to the interaction between adaptation on the current distribution and accumulated knowledge. This benchmark identifies model merging with test-time adaptation as a promising direction for continual computational pathology and motivates future methods that balance adaptation to domain shift with explicit preservation of historical knowledge.

Keywords: lifelong learning · whole slide image analysis · pathology vision-language foundation model · test-time adaptation

1 Introduction

Cancer classification from Whole Slide Images (WSIs) must accommodate continuously emerging cohorts and diagnostic tasks, while the gigapixel scale of WSIs and institutional data restrictions make centralised retraining and long-term storage of previous patient data impractical [1], [2]. This motivates continual learning methods that incorporate new knowledge without revisiting historical slides, patches, or feature bags.

Sequential fine-tuning is susceptible to catastrophic forgetting, whereas rehearsal-based methods require storing previous samples [3], [4]. Model merging offers an exemplar-free alternative by consolidating independently fine-tuned task models directly in parameter space [5], [6]. MergeSlide [7] demonstrates the effectiveness

of continual model merging for WSI classification. However, its merging rule is fixed once a new task model is incorporated and is not explicitly adapted to the incoming test distribution, potentially limiting its flexibility under distribution shift.

Test-time adaptive merging provides a natural extension to this static paradigm. AdaMerging [8] adapts task-wise or layer-wise merging coefficients, AdaRank [9] learns masks over singular components of task vectors, and Hi-Vec [10] adapts hierarchical prediction layers and propagates target-specific updates through task-vector-based weight merging. Despite their different formulations, these methods follow a common pattern of adapting merging-related variables from unlabeled test samples before constructing the final inference weights. Crucially, they were originally developed for multi-task merging or standard test-time adaptation rather than continual learning. When only the current task distribution is available, such adaptation may bias the merged model toward recent data, interfere with previously accumulated knowledge, and introduce repeated update to merge to inference overhead.

In this work, we systematically transfer AdaMerging, AdaRank, and Hi-Vec to continual WSI classification over six TCGA cancer-subtyping cohorts. We compare them with exemplar-free continual learning baselines including EWC [11] and LwF [12], and DER++ [4] as a rehearsal-based reference. Our study examines whether test-time adaptive merging improves current-task adaptation without increasing catastrophic forgetting, and whether repeatedly updating and reconstructing merging variables provides sufficient gains to justify its computational cost.

2 Methodology

2.1 Preliminaries

Problem formulation. Given a sequential stream $\{\mathcal{D}_t\}_{t=1}^{T}$ of T cancer-subtyping tasks, where each $\mathcal{D}_t$ contains WSIs from c_t subtypes, the goal is to fine-tune a slide aggregator f_A on each $\mathcal{D}_t$ and incrementally merge the resulting weights into a unified model $\tilde{\theta}_{1:t} = \mathcal{M}(\theta_0, \theta_1, \ldots, \theta_t)$, where θ_0 is the shared pre-trained backbone. The task vector encoding task-specific knowledge is defined as $\tau_t = \theta_t - \theta_0$. Two constraints are strictly enforced: (i) no access to $\mathcal{D}_{1:t-1}$ after task $t-1$ completes, and (ii) no exemplar memory. We evaluate under CLASS-IL (task identity unknown at inference) and TASK-IL (task identity provided).

WSI processing. Inspired by MergeSlide [7], each WSI $X_i \in \mathcal{D}_t$ is first preprocessed via the CLAM pipeline [1], which segments tissue regions using Otsu thresholding and tiles them into patches. Each patch $x_{i,j}$ is then passed through the frozen vision encoder of TITAN [13], a state-of-the-art pathology VLM, to extract a patch embedding $\mathbf{v}_{i,j} \in \mathbb{R}^{d_{\text{vis}}}$. We obtain a bag of patch embeddings$\mathcal{B}_i = \{\mathbf{v}_{i,j}\}_{j=1}^{|\mathcal{B}_i|}$. Following MergeSlide, K patches are randomly sampled using an index set $\mathcal{I}_i$ satisfying $|\mathcal{I}_i| = K$, forming $\mathcal{B}'_i = \{\mathbf{v}_{i,j} \mid j \in \mathcal{I}_i\} \subset \mathcal{B}_i$. The sampled

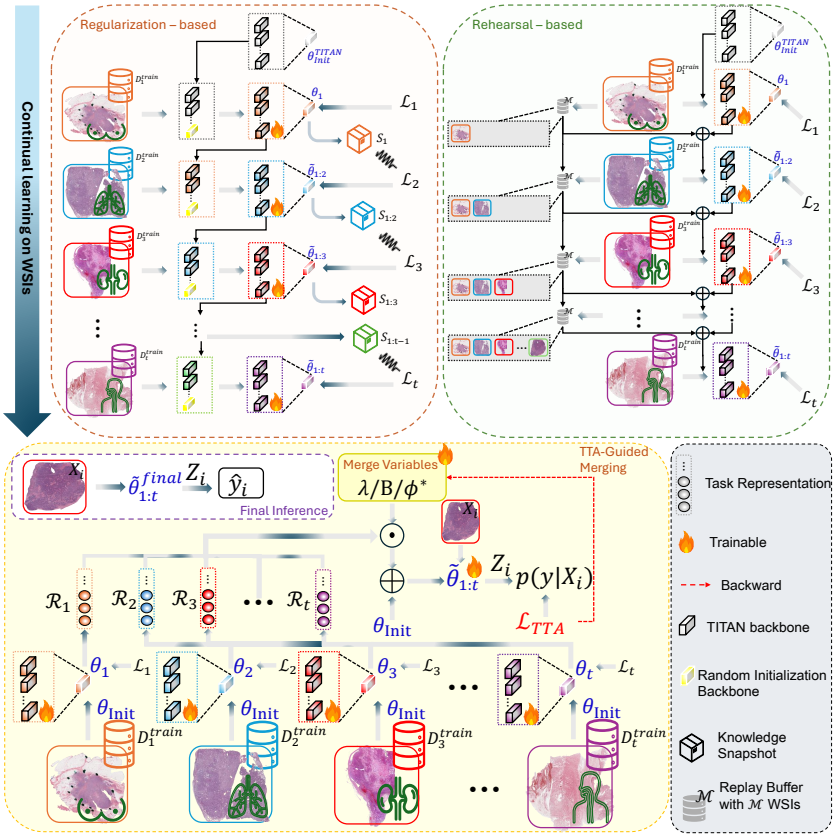


Fig. 1: Comparison of traditional continual learning and TTA-guided model merging for WSIs.

bag is then passed through the MLP-free slide aggregator $f_A(\cdot;\theta_t)$ to obtain the slide-level embedding$\mathbf{z}_i = f_A(\mathcal{B}'_i;\theta_t)$. No patch, slide, or feature bag from previous tasks is retained.

2.2 Continual Model Merging with Test-Time Adaptation

As illustrated in Fig. 1, each task is first fine-tuned independently from the same initialization θ_0, producing a task-specific model θ_k. Each method then converts this model into a method-dependent task representation $\mathcal{R}_k$: AdaMerging [8] uses task vectors, AdaRank [9] uses their singular-component decomposition, and Hi-Vec [10] uses hierarchical prediction layers with task-vector-based updates. At continual step t, the available representations $\mathcal{R}_{1:t}$ are combined with method-specific merge variables, namely coefficients $\boldsymbol{\lambda}_t$, binary masks $\mathbf{B}_t$, or a selected hierarchical layer ϕ_t^*, to construct the current model $\tilde{\theta}_{1:t} = \mathcal{M}(\theta_0, \mathcal{R}_{1:t}; \boldsymbol{\xi}_t)$. The

current unlabeled test stream $\mathcal{X}_t^{\text{test}}$ is then used to optimize $\boldsymbol{\xi}_t$ through the TTA objective, after which the model is reconstructed or updated for final inference. Since the original methods are not continual, we adapt them to a rehearsal-free sequential protocol in which no previous WSIs, patches, or feature bags are retained.

AdaMerging. AdaMerging learns task-wise coefficients $\lambda_{k,t}$, or layer-wise coefficients $\lambda_{k,t}^{(\ell)}$, by minimizing prediction entropy on the current test stream. The task-wise merged model is

$$\tilde{\theta}_{1:t}(\boldsymbol{\lambda}_t) = \theta_0 + \sum_{k=1}^{t} \lambda_{k,t}\tau_k, \qquad \tau_k = \theta_k - \theta_0. \tag{1}$$

For layer-wise merging, $\tilde{\theta}_{1:t}^{(\ell)} = \theta_0^{(\ell)} + \sum_{k=1}^{t} \lambda_{k,t}^{(\ell)}\tau_k^{(\ell)}$. The coefficients are optimized as $\boldsymbol{\lambda}_t^* = \arg\min_{\boldsymbol{\lambda}_t} \mathbb{E}_{X\sim\mathcal{X}_t^{\text{test}}}[H(p_A(\cdot \mid X; \tilde{\theta}_{1:t}(\boldsymbol{\lambda}_t)))]$.

AdaRank. AdaRank replaces fixed top-k SVD truncation with binary masks that select useful singular components. For task k and layer ℓ, let $\tau_k^{(\ell)} = U_k^{(\ell)}\Sigma_k^{(\ell)}V_k^{(\ell)\top}$ and $B_{k,t}^{(\ell)} \in \{0,1\}^{1\times m_\ell}$. The merged layer is

$$\tilde{\theta}_{1:t}^{(\ell)}(\mathbf{B}_t^{(\ell)}) = \theta_0^{(\ell)} + \lambda_t^{(\ell)} \sum_{k=1}^{t} U_k^{(\ell)} \left(\operatorname{diag}(B_{k,t}^{(\ell)}) \odot \Sigma_k^{(\ell)}\right) V_k^{(\ell)\top}. \tag{2}$$

The masks are optimized by entropy minimization on $\mathcal{X}_t^{\text{test}}$, using the Straight-Through Estimator to retain binary values in the forward pass while enabling gradient-based optimization.

Hi-Vec. Hi-Vec replaces a single classification head with hierarchical linear layers $\Phi = \{\phi_1, \ldots, \phi_R\}$ of increasing dimensionality. For each test batch, dynamic layer selection chooses $\phi_t^* = \arg\min_{\phi\in\Phi} \|\nabla_{W_\phi}\mathcal{L}_{\text{TTA},t}\|_2$. Hierarchical agreement is measured by $I(p_{\phi_t^*}; p_\phi) = H(p_{\phi_t^*}) - H(p_{\phi_t^*} \mid p_\phi)$; adaptation is suppressed when the average agreement is below δ_{OOD}. Finally, target information is shared with layers satisfying $\cos(\mathbf{v}_{\phi_t^*,t}, \mathbf{v}_{\phi,t}) > \delta_{\text{sim}}$, where $\mathbf{v}_{\phi,t} = \text{vec}(W_{\phi,t})$, through $W_{\phi,t}[\phi_t^*] \leftarrow W_{\phi_t^*,t} + \alpha W_{\phi,t}[\phi_t^*]$.

Test-Time Adaptation Objective. Despite their distinct architectural mechanisms, AdaMerging, AdaRank, and Hi-Vec share a common surrogate objective for unsupervised test-time adaptation: Shannon entropy minimization over the unlabeled test stream. Given a WSI $X \in \mathcal{X}_t^{\text{test}}$ and the predictive distribution $p_A(y = c \mid X; \theta)$ over C classes, the TTA objective at continual step t is defined as

$$\mathcal{L}_{\text{TTA},t}(\theta) = \mathbb{E}_{X\sim\mathcal{X}_t^{\text{test}}}\left[-\sum_{c=1}^{C} p_A(y = c \mid X; \theta) \log p_A(y = c \mid X; \theta)\right]. \tag{3}$$

Although the objective is shared, each method applies $\mathcal{L}_{\text{TTA},t}$ to a different set of adaptation variables:

- **AdaMerging** optimizes the task-wise or layer-wise merging coefficients by computing $\nabla_{\boldsymbol{\lambda}_t}\mathcal{L}_{\mathrm{TTA},t}$.
- **AdaRank** backpropagates $\mathcal{L}_{\mathrm{TTA},t}$ through the continuous proxy variables associated with the binary masks $\mathbf{B}_t$, using the Straight-Through Estimator to select singular components.
- **Hi-Vec** first uses the layer-wise gradient norm $\|\nabla_{W_\phi}\mathcal{L}_{\mathrm{TTA},t}\|_2$ for dynamic layer selection, and then minimizes $\mathcal{L}_{\mathrm{TTA},t}$ with respect to the selected layer ϕ_t^* and the shared encoder parameters.

3 Experiment

Datasets. We evaluate on six cancer subtyping cohorts sourced from the Genomic Data Commons (GDC) portal of The Cancer Genome Atlas (TCGA) [1]: TCGA-BRCA (breast), TCGA-RCC (kidney), TCGA-NSCLC (lung), TCGA-ESCA (esophagus), TCGA-TGCT (testis), and TCGA-CESC (cervix uteri). Each cohort is partitioned into 10 folds; 10-fold cross-validation is applied uniformly across all methods. Dataset statistics are summarized in Tab. 1

Table 1: Dataset statistics for the six TCGA cohorts.

Task	Dataset	Subtype	#WSIs	Ratio
1	TCGA-BRCA (B)	invasive ductal carcinoma (IDC)	726	4.9
		invasive lobular carcinoma (ILC)	149	1
2	TCGA-RCC (R)	clear cell renal cell carcinoma (CC)	498	4.2
		papillary renal cell carcinoma (P)	289	2.4
		chromophobe renal cell carcinoma (ChRCC)	118	1
3	TCGA-NSCLC (N)	squamous cell carcinoma (SCC)	845	7.8
		adenocarcinoma (A)	109	1
4	TCGA-ESCA (E)	squamous cell carcinoma (SCC)	114	1.3
		adenocarcinoma (A)	86	1
5	TCGA-TGCT (T)	seminoma (S)	66	2.3
		mixed germ cell tumor (MGCT)	29	1
6	TCGA-CESC (C)	squamous cell carcinoma (SCC)	270	5.5
		adenocarcinoma (A)	49	1

Implementation Details. WSIs are partitioned into non-overlapping 256×256 patches at $10\times$ magnification, and 768-dimensional patch embeddings are extracted using the frozen TITAN vision encoder. All methods share the same TITAN-initialized Transformer slide aggregator $f_{\mathcal{A}}$ and task-specific linear heads. CLASS-IL inference is performed over the concatenated outputs of all task heads, whereas TASK-IL uses only the head associated with the given task. Following MergeSlide [7], the continual-learning and TTA-guided merging methods represent each slide using K randomly sampled patch embeddings under a common

[1] https://portal.gdc.cancer.gov/

sampling protocol. In contrast, naive fine-tuning and fully supervised training use all available patch embeddings during both training and inference. Models are trained for $N_e = 10$ epochs per task using AdamW [14]. DER++ is evaluated with replay buffers of 10 and 30 WSIs. All experiments are conducted on a single NVIDIA A100-SXM4 GPU with 80 GB of memory.

Evaluation Metrics. We report bACC for CLASS-IL, where task identity is unavailable, and masked bACC for TASK-IL, where the task label is provided. Both use mean per-class recall to account for class imbalance. Final mean accuracy (Mean ACC), forgetting (FGT), and backward transfer (BWT) are also reported. CLASS-IL bACC remains the primary metric because it represents the more challenging deployment setting.

In/Out-of-Domain Evaluation. We evaluate all methods under in-domain (IND) and out-of-domain (OOD) protocols. IND permits site overlap between training and test sets, whereas OOD restricts test slides to institutions unseen during training [15].

Table 2: **Comparison of TTA-guided merging and conventional continual learning methods under the IND and OOD settings on six TCGA datasets with the task sequence B→R→N→E→T→C**. Results are reported as mean ± standard deviation in %. Within each setting, **bold** and underlined values denote the best and second-best results, respectively.

Setting	Method	Buffer	bACC ↑ (CLASS-IL)	Masked bACC ↑ (TASK-IL)	Mean ACC ↑ (CLASS-IL)	FGT ↓	BWT ↑
IND	Naive Finetuning	–	67.1044 (±3.2976)	90.5364 (±2.0347)	89.1803 (±1.2534)	32.1120 (±4.2359)	−31.6681 (±4.3679)
	Fully Supervised	–	**85.0110 (±2.3417)**	90.8313 (±1.8376)	–	–	–
	LwF [12]	–	65.1804 (±4.4238)	90.5335 (±1.8113)	88.7447 (±1.7930)	33.1937 (±5.7010)	−33.1457 (±5.6773)
	EWC [11]	–	64.4093 (±2.4041)	88.9731 (±2.1901)	87.5168 (±1.4701)	34.7054 (±2.6241)	−34.4042 (±2.7596)
	DER++ [4]	10 WSIs	69.2483 (±4.4018)	89.6904 (±2.2743)	89.9486 (±1.6597)	27.7758 (±4.9862)	−27.2555 (±5.0663)
	DER++ [4]	30 WSIs	70.8106 (±3.7026)	90.0952 (±2.3719)	**90.6375 (±1.5345)**	26.8298 (±4.7636)	−26.4299 (±4.6563)
	AdaMerging [8]	–	78.3230 (±2.0489)	89.8991 (±1.8120)	86.9496 (±1.8086)	**2.2073 (±0.7298)**	**−1.4375 (±0.9317)**
	AdaRank [9]	–	67.1714 (±3.0644)	92.8223 (±2.3648)	86.6967 (±1.6308)	22.1014 (±3.7348)	−21.8679 (±3.7708)
	Hi-Vec [10]	–	78.6298 (±2.1341)	**94.6061 (±0.7345)**	87.4145 (±1.8542)	2.6654 (±0.9230)	−2.3088 (±1.1964)
OOD	Naive Finetuning	–	62.9423 (±4.7075)	86.2913 (±3.1959)	88.8731 (±2.1782)	32.6840 (±6.6490)	−32.3411 (±6.8363)
	Fully Supervised	–	**79.3905 (±5.0927)**	86.8297 (±3.2221)	–	–	–
	LwF [12]	–	63.4133 (±5.8392)	86.8093 (±2.6748)	87.5284 (±2.5118)	31.5872 (±7.9211)	−31.5276 (±7.9176)
	EWC [11]	–	60.8898 (±4.0825)	84.8140 (±3.1537)	87.4747 (±2.0428)	34.2521 (±5.4305)	−34.0708 (±5.5881)
	DER++ [4]	10 WSIs	65.7675 (±5.8657)	87.1824 (±3.2936)	90.0000 (±1.6328)	30.2324 (±7.6581)	−29.8480 (±7.7467)
	DER++ [4]	30 WSIs	71.2885 (±4.0662)	88.1483 (±3.1624)	**90.5644 (±1.3354)**	21.6273 (±3.3084)	−21.2918 (±3.1684)
	AdaMerging [8]	–	76.9206 (±1.4339)	88.5119 (±1.6805)	88.6469 (±0.5660)	**2.2404 (±0.7536)**	**−1.9990 (±0.8746)**
	AdaRank [9]	–	67.9055 (±5.6059)	86.7569 (±3.6485)	87.1590 (±1.4127)	16.2504 (±3.9810)	−15.8924 (±4.3479)
	Hi-Vec [10]	–	77.2384 (±1.5898)	**94.1407 (±0.8933)**	88.9163 (±0.6548)	2.6285 (±0.6480)	−2.2167 (±0.9649)

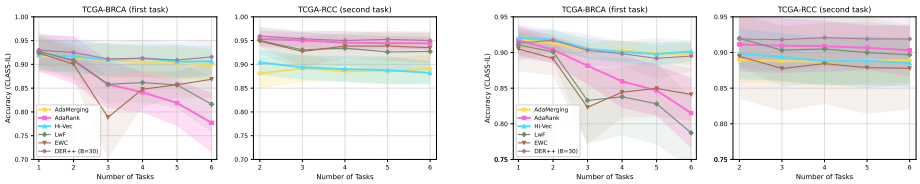
Main Results. As shown in Tabs. 2, TTA-guided merging is particularly effective in TASK-IL. Hi-Vec exceeds the fully supervised reference by 3.77% on IND and 7.31% on OOD, while AdaRank provides a −1.99% gain on IND. In CLASS-IL, AdaMerging and Hi-Vec outperform the strongest continual-learning baseline in bACC by 7.51% and 7.82% on IND, and by 5.63% and 5.95% on OOD, respectively. Although DER++ with 30 replay WSIs attains higher Mean ACC, AdaMerging and Hi-Vec preserve previous tasks more effectively: relative

to DER++, they reduce FGT by approximately 24.2%–24.6% and improve BWT by 24.1%–25.0% on IND, with corresponding gains of about 19% under OOD shift. AdaRank remains competitive in TASK-IL but exhibits greater forgetting and stronger task dependence in CLASS-IL.

Table 3: **Comparison of TTA-guided merging methods mapped to the continual learning setting against conventional continual learning methods on a benchmark of six TCGA datasets in the IND setting under the reversed task sequence E→T→C→B→R→N**. All results are reported as mean ± standard deviation in %. **Bold** denotes the best result, while underlined denotes the second-best result.

Method	Buffer	bACC ↑ (CLASS-IL)	Masked bACC ↑ (TASK-IL)	Mean ACC ↑ (CLASS-IL)	FGT ↓	#BWT ↑
Naive Finetuning	–	61.9742 (±4.1518)	90.5923 (±1.6786)	61.5699 (±2.2538)	41.3195 (±5.7577)	−41.1791 (±5.8597)
Fully Supervised	–	**85.0261 (±3.4855)**	90.2535 (±2.4831)	–	–	–
LwF [12]	–	55.5284 (±4.1972)	90.3635 (±1.8555)	61.7891 (±1.4478)	48.6558 (±4.1489)	−48.4978 (±4.2601)
EWC [11]	–	60.4054 (±4.9796)	90.6885 (±1.3657)	61.3409 (±2.0593)	42.4943 (±6.5684)	−42.3889 (±6.6177)
DER++ [4]	10 WSIs	73.8973 (±5.7857)	90.6087 (±1.5475)	80.0842 (±2.9199)	21.9437 (±8.1397)	−21.7424 (±8.1899)
DER++ [4]	30 WSIs	80.3150 (±4.8132)	91.1414 (±1.3826)	**86.3063 (±1.8397)**	**11.8343 (±5.6468)**	**−11.5218 (±5.8428)**
AdaMerging [8]	–	78.6677 (±2.5413)	90.0052 (±2.1859)	75.1242 (±0.9795)	14.3921 (±0.7467)	−13.9168 (±0.9786)
AdaRank [9]	–	65.8651 (±2.8355)	93.5975 (±1.7599)	74.3150 (±1.3113)	26.7375 (±3.6149)	−26.2774 (±3.6550)
Hi-Vec [10]	–	78.6292 (±1.8153)	**94.2190 (±1.1353)**	75.1268 (±1.0787)	14.6877 (±0.9015)	−14.1682 (±1.0023)

Sensitivity to Task Order. Under the reversed sequence, Hi-Vec and AdaRank retain strong TASK-IL performance, exceeding the fully supervised reference by 3.97% and 3.34%, respectively (Tab. 3). In CLASS-IL, AdaMerging and Hi-Vec outperform DER++ with 10 replay WSIs in bACC by approximately 4.8% while requiring no exemplar storage. They also reduce FGT by 7.3%–7.6% and improve BWT by 7.6%–7.8% relative to this rehearsal baseline. DER++ with 30 WSIs remains stronger in final accuracy and forgetting. Compared with the original task order, the higher FGT of AdaMerging and Hi-Vec indicates sensitivity to task sequence, although both continue to preserve prior knowledge more effectively than naive fine-tuning, LwF, EWC, and limited replay.



(a) Performance drop under the IND setting.

(b) Performance drop under the OOD setting.

Fig. 2: Performance degradation of all methods under the (a) in-distribution (IND) and (b) out-of-distribution (OOD) evaluation settings.

Performance retention. We further compare the performance degradation of different methods on TCGA-BRCA and TCGA-RCC as additional tasks are incrementally introduced, as shown in Fig. 2. The performance trajectories further support the effectiveness of TTA-guided merging for preserving previously acquired knowledge. For the earliest task, TCGA-BRCA, AdaMerging and Hi-Vec maintain nearly flat accuracy curves under both IND and OOD evaluation, with only a small reduction after five additional tasks are incorporated. Their retention is comparable to DER++ with a 30-WSI replay buffer, while clearly exceeding LwF, EWC, and AdaRank, which exhibit substantially larger degradation on the same task. A similarly stable trend is observed on TCGA-RCC, although the relative ranking varies across methods. In particular, AdaRank preserves RCC well but degrades markedly on BRCA, indicating greater task-dependent sensitivity.

4 Discussion

Applying TTA-guided merging to continual learning reveals a strong stability–plasticity trade-off, as adaptation relies only on the current unlabeled test distribution. AdaMerging [8] may shift earlier task coefficients toward recent tasks, AdaRank [9] may suppress singular components important to prior tasks, and Hi-Vec [10] must balance rejecting novel samples against excessive updates and representation drift. Without explicit protection of previously learned subspaces, these methods remain effective but sensitive to task order and current-task adaptation, as reflected by increased forgetting in the reversed order and stronger task dependence in AdaRank. Future work should combine test-time plasticity with explicit preservation of historical parameter subspaces.

5 Conclusion

This benchmark study examined state-of-the-art model merging methods that use test-time adaptation to address domain shift when transferred to rehearsal-free continual WSI classification. Across six TCGA cohorts, these approaches showed strong task-specific performance and, in several settings, improved knowledge retention over traditional continual-learning methods. However, their behavior remained sensitive to task order and to adaptation driven only by the current test distribution. These findings position model merging with test-time adaptation as a promising direction for continual computational pathology and motivate future methods that improve adaptability while explicitly protecting previously acquired knowledge.

Disclosure of Interests

The authors declare that they have no conflict of interest.

References

[1] M. Y. Lu, D. F. Williamson, T. Y. Chen, R. J. Chen, M. Barbieri, and F. Mahmood, "Data-efficient and weakly supervised computational pathology on whole-slide images," *Nature biomedical engineering*, vol. 5, no. 6, pp. 555–570, 2021.

[2] Y. Shen, A. Sowmya, Y. Luo, X. Liang, D. Shen, and J. Ke, "A federated learning system for histopathology image analysis with an orchestral stain-normalization gan," *IEEE Transactions on Medical Imaging*, vol. 42, no. 7, pp. 1969–1981, 2022.

[3] M. Boschini et al., "Transfer without forgetting," in *European conference on computer vision*, Springer, 2022, pp. 692–709.

[4] P. Buzzega, M. Boschini, A. Porrello, D. Abati, and S. Calderara, "Dark experience for general continual learning: A strong, simple baseline," *Advances in neural information processing systems*, vol. 33, pp. 15 920–15 930, 2020.

[5] G. Ilharco et al., "Editing models with task arithmetic," *arXiv preprint arXiv:2212.04089*, 2022.

[6] P. Yadav, D. Tam, L. Choshen, C. A. Raffel, and M. Bansal, "Ties-merging: Resolving interference when merging models," *Advances in neural information processing systems*, vol. 36, pp. 7093–7115, 2023.

[7] D. C. Bui, B. H. Ngo, H. L. Pham, K. Nguyen, M. K. Nguyen, and Y. Nakashima, "Mergeslide: Continual model merging and task-to-class prompt-aligned inference for lifelong learning on whole slide images," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026, pp. 4859–4868.

[8] E. Yang et al., "Adamerging: Adaptive model merging for multi-task learning," in *International Conference on Learning Representations*, vol. 2024, 2024, pp. 22 743–22 763.

[9] C. Lee, J. Choi, C. Lee, D. Kim, and S. Hong, "Adarank: Adaptive rank pruning for enhanced model merging," *arXiv preprint arXiv:2503.22178*, 2025.

[10] S. Ambekar, M. Hasny, L. A. Daza, D. M. Lang, and J. Schnabel, "Hierarchical adaptive networks with task vectors for test-time adaptation," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2026, pp. 4661–4672.

[11] J. Kirkpatrick et al., "Overcoming catastrophic forgetting in neural networks," *Proceedings of the national academy of sciences*, vol. 114, no. 13, pp. 3521–3526, 2017.

[12] Z. Li and D. Hoiem, "Learning without forgetting," *IEEE transactions on pattern analysis and machine intelligence*, vol. 40, no. 12, pp. 2935–2947, 2017.

[13] T. Ding et al., "A multimodal whole-slide foundation model for pathology," *Nature medicine*, pp. 1–13, 2025.

[14] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[15] S. Cheng et al., "Robust whole slide image analysis for cervical cancer screening using deep learning," *Nature communications*, vol. 12, no. 1, p. 5639, 2021.