MergeSurv: Merging-Based Continual Learning for Survival Analysis on Whole-Slide Images

Anonymous Submission

No Institute Given

Abstract. Survival analysis on Whole Slide Images (WSIs) is important in computational pathology for prognosis estimation and treatment planning. However, existing survival models are typically trained independently for each cancer cohort, making continual adaptation computationally expensive for gigapixel-scale WSIs. In this study, we propose MergeSurv, a merging-based continual learning framework for WSI survival analysis. A pathology vision-language foundation model is independently fine-tuned on each task, and the learned parameters are sequentially merged into a unified model without storing previous training data. We further investigate two inference strategies: One-for-All (OFA) and Voting-Expert Aggregation (VEA). Experiments on four TCGA cohorts demonstrate that MergeSurv outperforms naive fine-tuning as well as representative regularization-based and rehearsal-based continual learning methods, while effectively reducing catastrophic forgetting. The results suggest that model merging is a promising direction for scalable and privacy-preserving continual learning in computational pathology.

Keywords: lifelong learning · whole slide image analysis · pathology vision-language foundation model · survival analysis

1 Introduction

Survival risk prediction is an important task in computational pathology because it supports prognosis estimation and treatment planning for cancer patients [1, 2]. Whole Slide Images (WSIs), which provide detailed tissue information at the cellular level, are widely used for this purpose [2, 3]. However, existing survival prediction models are still largely task-specific, where a new standalone model must be trained whenever a new organ or cohort is introduced [2, 4]. This limitation is particularly problematic for WSI analysis because WSIs are gigapixel-scale images requiring substantial storage, preprocessing, and training resources [3]. Therefore, continual learning (CL), which enables models to learn sequential tasks while mitigating catastrophic forgetting [5, 6], has become a promising direction for WSI analysis.

Existing CL approaches mainly include regularization-based methods, such as EWC [5] and LwF [7], and rehearsal-based methods, such as ER [8], DER, and DER++ [9]. However, rehearsal-based methods require storing previous WSIs, leading to high memory costs and potential privacy concerns in medical applications. Moreover, most existing WSI CL studies focus on classification tasks

[10–12], while continual survival prediction remains underexplored [4]. Recent studies have explored model merging [13, 14], which integrates independently trained task-specific models into a unified model without storing original training data. This property makes model merging particularly suitable for lifelong WSI learning, where repeated training and cross-institutional data sharing are often expensive. *Although MergeSlide [10] recently introduced merging-based continual learning for WSI classification, it does not address survival-specific challenges such as censored data and continuous risk prediction.*

In this study, we propose MergeSurv, a merging-based continual learning framework for survival analysis on WSIs. Specifically, a pre-trained pathology foundation model (TITAN [15]) is independently fine-tuned on each survival task to obtain task-specific parameters, which are subsequently merged into a unified model. During inference, we investigate two prediction strategies: **1) One-for-All (OFA)**, which uses a single unified survival head, and **2) Voting-Expert Aggregation (VEA)**, which combines predictions from multiple task-specific heads. Experimental results on four TCGA cohorts demonstrate that the proposed approach consistently outperforms naive fine-tuning as well as regularization-based and rehearsal-based continual learning methods.

2 Methodology

2.1 An Overview of MergeSurv

MergeSurv is a CL framework for survival analysis using WSIs. Given a stream of N datasets $\mathcal{D} = \{D_t\}_{t=1}^{N}$ corresponding to N tasks, MergeSurv trains each survival task D_t using the WSI model $f_{\mathcal{S}}(\cdot;\theta_{\mathcal{S}})$ with a Cox regression head $H(\cdot;\theta_H)$ to obtain the task-specific optimal parameters $\theta^*_{\mathcal{S},t}$ and $\theta^*_{H,t}$. Subsequently, all parameters $\{\theta^*_{\mathcal{S},t}\}_{t=1}^{N}$ are sequentially merged into unified parameters, resulting in the merged parameters $\tilde{\Theta}_{\mathcal{S},1:N}$. Finally, only $f_{\mathcal{S}}(\cdot;\tilde{\Theta}_{\mathcal{S},1:N})$ and two risk prediction strategies to estimate survival risk scores are used for inference on all N tasks learned so far.

2.2 *Ad hoc* fine-tuning

Fine-tuning Process First, whenever a new WSI survival task t is defined, the model $f_{\mathcal{S}}$ is fine-tuned from the prior parameters θ_{prior} using the training samples of D_t to optimize the objective function $\mathcal{L}$ over E epochs. θ_{prior} is initialized from the pretrained weights of the TITAN foundation model [15]. As TITAN is a pathology-specific foundation model trained on large-scale unlabeled pathology images in a self-supervised manner, it effectively captures tissue structures in the embedding space. The fine-tuning process is formulated as follows.

$$\{\theta^*_{\mathcal{S},t}, \theta^*_{H,t}\} \approx \arg\min_{\theta_{\mathcal{S},t},\theta_{H,t}} \mathcal{L}\left(\hat{r}_i, Y_i\right), \quad (X_i, Y_i) \sim D_t, \quad \theta^{(0)}_{S,t} = \theta_{\text{prior}}, \quad e = 1 \rightarrow E,$$
$$\text{where} \quad Y_i = (T_i, \delta_i), \quad \hat{r_i} = H(f_{\mathcal{S}}(X_i;\theta_{\mathcal{S},t});\theta_{H,t}) \tag{1}$$

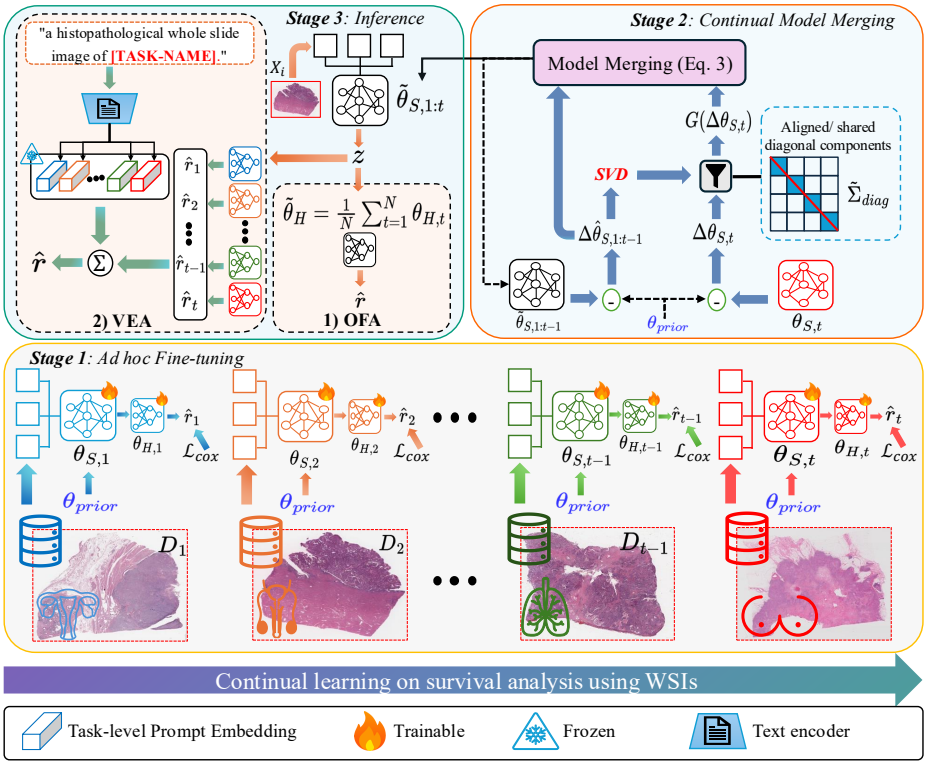

Fig. 1: Overall framework of MergeSurv. Each survival task is fine-tuned from TITAN pretrained weights θ_{prior} using the WSI model $f_{\mathcal{S}}$ with a Cox regressor. Task-specific WSI aggregator parameters are sequentially merged using Eq. 3. Two inference strategies are introduced: One-for-All (OFA) with a single Cox head, and Voting-Expert Aggregation (VEA) with multiple task-specific heads.

where $\{\theta^*_{\mathcal{S},t}, \theta^*_{H,t}\}$ denotes the optimal parameters of the WSI model and the regression head for task t and $X_i = \{x_k\}_{k=1}^{K}$ denotes a bag of K patch embeddings extracted from an input WSI, where $x_k \in \mathbb{R}^{1\times d_{\text{vis}}}$. H is a Cox regression head [1, 2] that predicts a scalar risk score $\hat{r_i} \in \mathbb{R}$, where a larger value indicates a higher event risk. Therefore, the fine-tuning stage jointly learns the WSI aggregator and the task-specific Cox head by optimizing the loss function.

Loss Function We optimize the model using the Cox partial likelihood loss $\mathcal{L}_{Cox}$, which is defined as follows:

$$\mathcal{L}_{Cox} = -\frac{1}{N_e} \sum_{i:\delta_i=1} \left(\hat{r_i} - \log \sum_{j:T_j \geq T_i} \exp(\hat{r_j}) \right), \tag{2}$$

where $N_e = \sum_i \delta_i$ is the number of uncensored event samples, T_i denotes the survival time, δ_i indicates the event status, and $\hat{r}_i$ is the predicted risk score for i-th sample. The Cox loss learns the relative risk ranking among patients, assigning higher risk scores to patients with earlier events.

2.3 Model Merging Operation

We follow OPCM [16] for the model merging operations of the WSI model parameters $f_{\mathcal{S}}$. Three sets of parameters are considered: 1) the prior parameters $\theta_{\text{prior}} \in \mathbb{R}^{m\times n}$, 2) the previously merged parameters of the WSI model $\tilde{\Theta}_{\mathcal{S},1:t-1} \in \mathbb{R}^{m\times n}$, and 3) the newly fine-tuned parameters $\theta^*_{\mathcal{S},t} \in \mathbb{R}^{m\times n}$. The merged parameters are updated as follows:

$$\tilde{\Theta}_{\mathcal{S},1:t} = \theta_{\text{prior}} + \frac{\lambda_{t-1}\widetilde{\Delta\theta}_{\mathcal{S},1:t-1} + G(\Delta\theta_{\mathcal{S},t}, \widetilde{\Delta\theta}_{\mathcal{S},1:t-1})}{\lambda_t}, \tag{3}$$

where $\Delta\theta_{\mathcal{S},t} = \theta^*_{\mathcal{S},t} - \theta_{\text{prior}}$ and $\widetilde{\Delta\theta}_{\mathcal{S},1:t-1} = \tilde{\Theta}_{\mathcal{S},1:t-1} - \theta_{\text{prior}}$ denote the current and accumulated task vectors, respectively [14]. $G(\cdot)$ is a projection function, and λ_t [10, 16] controls the balance between old and new knowledge. To preserve task separability, OPCM projects the new task vector $\Delta\theta_{\mathcal{S},t}$ onto the orthogonal space of $\widetilde{\Delta\theta}_{\mathcal{S},1:t-1}$. Specifically, singular value decomposition (SVD) is applied to $\widetilde{\Delta\theta}_{\mathcal{S},1:t-1}$ to obtain orthogonal subspaces, and the components of $\Delta\theta_{\mathcal{S},t}$ aligned with previous task directions are removed through a zero-diagonal masking operation. This projection satisfies $\langle G(\Delta\theta_{\mathcal{S},t}), \widetilde{\Delta\theta}_{\mathcal{S},1:t-1}\rangle_F = 0$, ensuring that the newly merged knowledge does not interfere with previously learned tasks.

To further prevent the merged model from drifting excessively from the base model, the balancing factor λ_t is adaptively computed based on the norm ratio between the accumulated merged task vectors and the average norm of all task-specific parameters. As a result, the distance between the merged model and the base model remains bounded by the average task norm, stabilizing the continual merging process across sequential survival tasks.

2.4 Risk Prediction

One-for-All. In the one-for-all (OFA) strategy, the WSI aggregator is merged using OPCM to obtain the merged parameters $\tilde{\Theta}_{\mathcal{S},1:N}$, while the survival head is constructed by averaging the weights of the learned task-specific Cox heads. Let $\theta^*_{H,t}$ denote the parameters of the Cox head H learned for task t. After observing N tasks, the merged Cox head parameters are computed by averaging all task-specific heads, and the final risk score is computed as:

$$\hat{r} = H(f_{\mathcal{S}}(X; \tilde{\Theta}_{\mathcal{S},1:N}); \tilde{\theta}_H), \quad \text{where } \tilde{\theta}_H = \frac{1}{N}\sum_{t=1}^{N} \theta^*_{H,t}. \tag{4}$$

This strategy creates a unified model consisting of one merged backbone and a single Cox head; therefore, *OFA is simple and memory-efficient during inference.*

Voting-Expert Aggregation. In the voting-expert aggregation (VEA) strategy, the WSI aggregator is still merged using OPCM, while the task-specific Cox head parameters $\{\theta^*_{H,t}\}_{t=1}^N$ are retained instead of being merged into a single head. In this manner, each head acts as an expert and produces an individual risk score:

$$\hat{r}_t = H(f_{\mathcal{S}}(X; \tilde{\Theta}_{\mathcal{S},1:N}); \theta^*_{H,t}), \quad t = 1, \ldots, N. \tag{5}$$

When the task identity is not provided during inference, MergeSurv estimates the relevance of each expert using task-level text prompts. Let $E_{\text{task},t}$ denote the text embedding of the task-level prompt corresponding to task t. The routing weight is computed using the cosine similarity between the WSI representation obtained by WSI model $f_{\mathcal{S}}$ and the task prompt embedding: $\pi_t = \frac{\exp(a_t/\gamma)}{\sum_{j=1}^N \exp(a_j/\gamma)}$, where $a_t = \cos\big(f_{\mathcal{S}}(X; \tilde{\Theta}_{\mathcal{S},1:N}), E_{\text{task},t}\big)$.

The final risk score is computed as the weighted sum of the predictions from all expert heads:

$$\hat{r} = \sum_{t=1}^{N} \pi_t \hat{r}_t. \tag{6}$$

This strategy helps preserve the task-specific knowledge of each survival head *while still using a single merged WSI aggregator to extract WSI representations.*

3 Experiment

Datasets We use four cancer cohorts from The Cancer Genome Atlas Program (TCGA) to evaluate WSI-based continual survival prediction: Bladder Urothelial Carcinoma (BLCA, n = 373), Uterine Corpus Endometrial Carcinoma (UCEC, n = 480), Lung Adenocarcinoma (LUAD, n = 453), and Breast Invasive Carcinoma (BRCA, n = 955). Each cohort is regarded as a separate task in the continual learning sequence. Patient-level overall survival time and censoring status are used to define the survival prediction targets. To evaluate MergeSurv and other continual learning methods, each cohort is split into five folds, and 5-fold cross-validation is performed. The task order used in this study is BLCA→UCEC→LUAD→BRCA.

Experimental Settings. We evaluate MergeSurv on a continual learning sequence of four TCGA cohorts: BLCA, UCEC, LUAD, and BRCA. *Joint training* is included as an upper-bound setting, while *Naive fine-tuning* serves as a sequential learning baseline without forgetting mitigation. We further compare with representative continual learning methods, including EWC [5], LwF [7], ER [8], DER [9], and DER++ [9], where replay-based methods use a buffer size of 32 WSIs. All experiments are conducted on an NVIDIA A100 (80GB) GPU.

Implementation Details. WSIs are tiled into 256×256 patches at $10\times$ magnification. Patch and prompt embeddings are extracted using TITAN encoders

[15]. Under the MIL [3] setting, each WSI is represented as a bag of instances with $K = 1000$ randomly sampled patches. MergeSurv employs a Transformer-based WSI aggregator $f_{\mathcal{S}}$ initialized from TITAN pretrained weights θ_{prior}. All models are trained for $E = 20$ epochs using AdamW with learning rates of 1×10^{-5} for the backbone and 1×10^{-4} for the head.

Main Results. We use the C-Index [17] and C-Index IPCW [18] to evaluate survival risk ranking, while Forgetting [19] and Backward Transfer (BWT) [20] are used for continual learning evaluation. As shown in Tab. 1, MergeSurv achieves strong continual survival prediction performance while effectively mitigating catastrophic forgetting across sequential TCGA cohorts. For the standard C-Index in Tab. 1(a), MergeSurv (VEA) achieves the best average C-Index (0.6773) and the lowest forgetting (0.0138), while MergeSurv (OFA) obtains the best BWT (-0.0054). Notably, MergeSurv (VEA) slightly outperforms Joint training (0.6773 vs. 0.6739), suggesting that parameter merging effectively preserves and integrates survival knowledge without joint training. Compared with replay-based methods, MergeSurv substantially reduces forgetting without requiring memory buffers. For the IPCW-adjusted C-Index in Tab. 1(b), EWC achieves the highest average score (0.6353), while MergeSurv (VEA) obtains the second-best score (0.6294) together with the lowest forgetting (0.0159) and competitive BWT (0.0011).

Table 1: Continual survival prediction results on four TCGA cohorts. Best results are highlighted in **bold**, and the second-best results are underlined.

(a) **C-Index**

Method	Buffer	Average (↑)	Forget (↓)	BWT (↑)
Joint training	–	0.6739 (±0.0125)	–	–
Naive fine-tuning	–	0.6618 (±0.0274)	0.0303 (±0.0133)	-0.0161 (±0.0260)
EWC	–	0.6596 (±0.0326)	0.0489 (±0.0449)	-0.0363 (±0.0590)
LwF	–	0.6531 (±0.0276)	0.0236 (±0.0196)	-0.0123 (±0.0255)
ER	32 WSIs	0.6535 (±0.0289)	0.0549 (±0.0217)	-0.0468 (±0.0194)
DER	32 WSIs	0.6630 (±0.0259)	0.0386 (±0.0266)	-0.0196 (±0.0356)
DER++	32 WSIs	0.6620 (±0.0334)	0.0500 (±0.0396)	-0.0445 (±0.0444)
MergeSurv (OFA)	–	0.6693 (±0.0245)	0.0262 (±0.0167)	**-0.0054 (±0.0227)**
MergeSurv (VEA)	–	**0.6773 (±0.0133)**	**0.0138 (±0.0082)**	-0.0097 (±0.0115)

(b) **C-Index IPCW**

Method	Buffer	Average (↑)	Forget (↓)	BWT (↑)
Joint training	–	0.6184 (±0.0386)	–	–
Naive fine-tuning	–	0.6099 (±0.0621)	0.0282 (±0.0210)	-0.0068 (±0.0339)
EWC	–	**0.6353 (±0.0474)**	0.0411 (±0.0150)	0.0040 (±0.0754)
LwF	–	0.6045 (±0.0663)	0.0264 (±0.0273)	**0.0164 (±0.0745)**
ER	32 WSIs	0.6082 (±0.0477)	0.0570 (±0.0454)	-0.0072 (±0.0452)
DER	32 WSIs	0.6246 (±0.0771)	0.0300 (±0.0116)	-0.0056 (±0.0201)
DER++	32 WSIs	0.6014 (±0.0698)	0.0511 (±0.0413)	-0.0288 (±0.0419)
MergeSurv (OFA)	–	0.6133 (±0.0571)	0.0639 (±0.0718)	-0.0275 (±0.0678)
MergeSurv (VEA)	–	0.6294 (±0.0497)	**0.0159 (±0.0099)**	0.0011 (±0.0346)

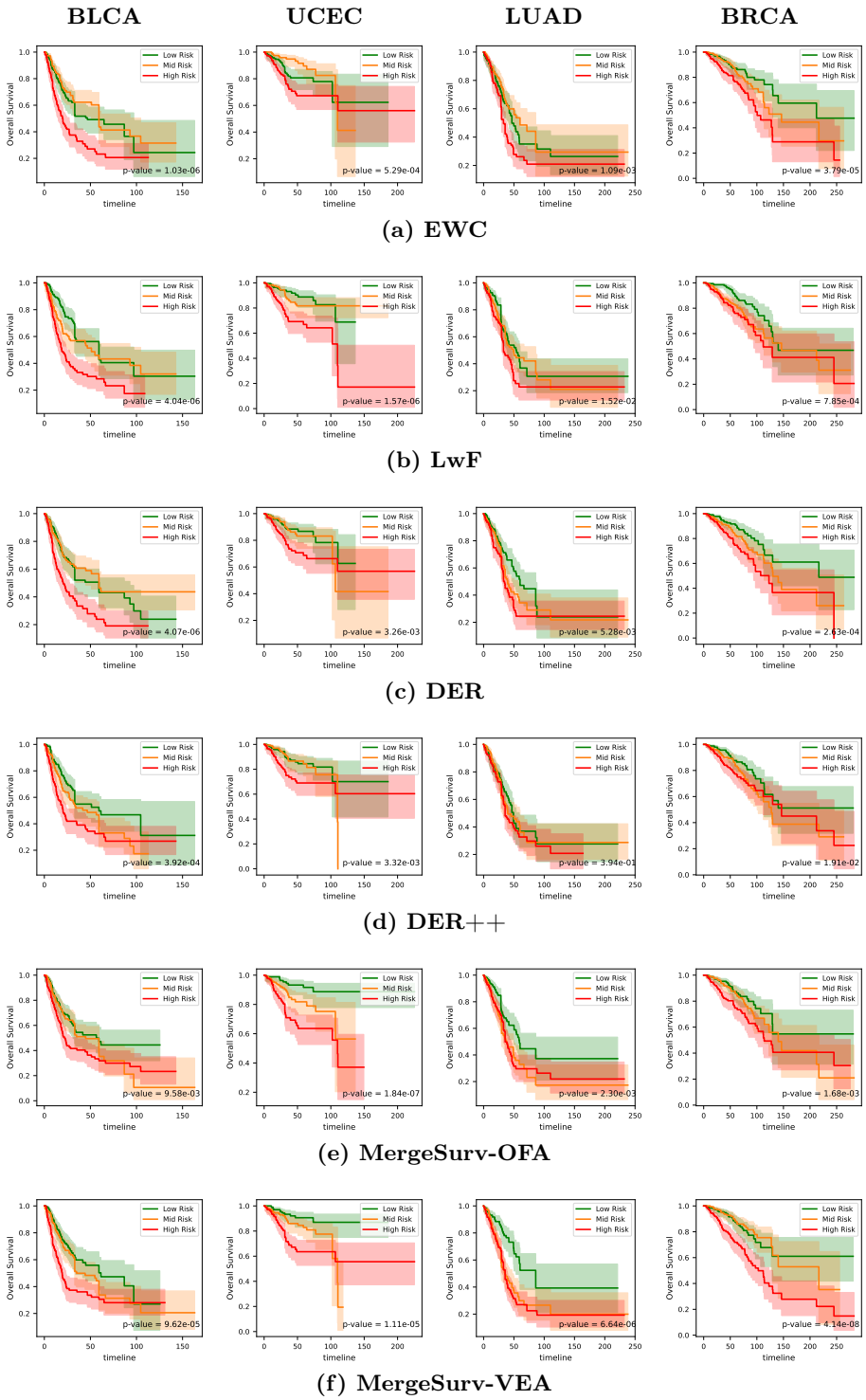

Fig. 2: Kaplan–Meier survival curves of selected continual learning methods across four cancer cohorts (Zooming 4× for viewing ease).

4 Kaplan-Meier Analyses

To provide a more comprehensive evaluation of survival stratification performance, we present Kaplan–Meier curves [21] for low-, mid-, and high-risk groups, which are defined according to task-wise tertiles of the out-of-fold predicted Cox risk scores for each method. On BLCA, most methods fail to maintain consistent risk ordering, where the low-risk curve occasionally falls below the high-risk curve, indicating poor stratification. On UCEC, MergeSurv-OFA shows clearer separation among risk groups, with the high-risk group consistently exhibiting the lowest survival probability. On LUAD, most baselines exhibit substantial overlap between the mid- and high-risk groups, whereas MergeSurv-VEA preserves a more consistent risk ordering. On BRCA, MergeSurv-VEA achieves the clearest separation between risk groups, while other methods produce overlapping survival curves. Furthermore, MergeSurv obtains the lowest log-rank p-values [22] on three out of four datasets, including UCEC (1.84×10^{-7} with OFA), LUAD (6.64×10^{-6} with VEA), and BRCA (4.14×10^{-8} with VEA), suggesting stronger statistical separation between predicted risk groups.

5 Conclusion

In this study, we presented MergeSurv, a merging-based continual learning framework for survival analysis on Whole Slide Images (WSIs). Unlike existing rehearsal-based continual learning approaches, MergeSurv avoids storing previous WSIs by independently fine-tuning task-specific models and sequentially merging their parameters into a unified model. We further investigated two inference strategies, namely One-for-All (OFA) and Voting-Expert Aggregation (VEA), for continual survival risk prediction across multiple cancer cohorts. Experimental results on four TCGA datasets demonstrated that MergeSurv effectively mitigates catastrophic forgetting while achieving competitive survival prediction performance compared with regularization-based and replay-based continual learning methods. In particular, MergeSurv-VEA achieved strong risk stratification performance and consistently low forgetting across sequential survival tasks. Overall, the results suggest that model merging is a promising direction for scalable and privacy-preserving continual learning in computational pathology.

Disclosure of Interests

The authors declare that they have no conflict of interest.

References

[1] David R. Cox. “Regression Models and Life-Tables”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2 (1972), pp. 187–202.

[2] Jiawen Yao et al. "Whole Slide Images Based Cancer Survival Prediction Using Attention Guided Deep Multiple Instance Learning Networks". In: *Medical Image Analysis* 65 (2020), p. 101789.

[3] Ming Y Lu et al. "Data-efficient and weakly supervised computational pathology on whole-slide images". In: *Nature Biomedical Engineering* 5.6 (2021), pp. 555–570.

[4] Dianzhi Yu et al. "ConSurv: Multimodal Continual Learning for Survival Analysis". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 40.33 (Mar. 2026), pp. 27899–27907.

[5] James Kirkpatrick et al. "Overcoming catastrophic forgetting in neural networks". In: *Proceedings of the National Academy of Sciences* 114.13 (2017), pp. 3521–3526.

[6] Liyuan Wang et al. "A Comprehensive Survey of Continual Learning: Theory, Method and Application". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 46.8 (2024), pp. 5362–5383.

[7] Zhizhong Li and Derek Hoiem. "Learning without Forgetting". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 40.12 (2018), pp. 2935–2947.

[8] David Rolnick et al. "Experience Replay for Continual Learning". In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc., 2019.

[9] Pietro Buzzega et al. "Dark experience for general continual learning: a strong, simple baseline". In: *Advances in Neural Information Processing Systems* 33 (2020), pp. 15920–15930.

[10] Doanh C. Bui et al. "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.

[11] Byung Hyun Lee et al. "Continual Multiple Instance Learning with Enhanced Localization for Histopathological Whole Slide Image Analysis". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2025, pp. 23232–23242.

[12] Yanyan Huang et al. "ConSlide: Asynchronous Hierarchical Interaction Transformer with Breakup-Reorganize Rehearsal for Continual Whole Slide Image Analysis". In: *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 2023, pp. 21349–21360.

[13] Mitchell Wortsman et al. "Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time". In: *Proceedings of the 39th International Conference on Machine Learning*. Vol. 162. Proceedings of Machine Learning Research. PMLR, 2022, pp. 23965–23998.

[14] Gabriel Ilharco et al. "Editing Models with Task Arithmetic". In: *International Conference on Learning Representations*. 2023.

[15] Tong Ding et al. "A Multimodal Whole-Slide Foundation Model for Pathology". In: *Nature Medicine* 31 (2025), pp. 3749–3761.

[16] Anke Tang et al. "Merging on the Fly Without Retraining: A Sequential Approach to Scalable Continual Model Merging". In: *Advances in Neural Information Processing Systems*. Vol. 38. 2025.

[17] Frank E. Harrell, Kerry L. Lee, and Daniel B. Mark. "Multivariable Prognostic Models: Issues in Developing Models, Evaluating Assumptions and Adequacy, and Measuring and Reducing Errors". In: *Statistics in Medicine* 15.4 (1996), pp. 361–387.

[18] Hajime Uno et al. "On the C-statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data". In: *Statistics in Medicine* 30.10 (2011), pp. 1105–1117.

[19] Arslan Chaudhry et al. "Riemannian Walk for Incremental Learning: Understanding Forgetting and Intransigence". In: *Proceedings of the European Conference on Computer Vision (ECCV)*. 2018, pp. 556–572.

[20] David Lopez-Paz and Marc'Aurelio Ranzato. "Gradient Episodic Memory for Continual Learning". In: *Advances in Neural Information Processing Systems*. Vol. 30. 2017.

[21] Edward L. Kaplan and Paul Meier. "Nonparametric Estimation from Incomplete Observations". In: *Journal of the American Statistical Association* 53.282 (1958), pp. 457–481.

[22] Nathan Mantel. "Evaluation of Survival Data and Two New Rank Order Statistics Arising in Its Consideration". In: *Cancer Chemotherapy Reports* 50.3 (1966), pp. 163–170.