ConSurv: Multimodal Continual Learning for Survival Analysis

Dianzhi Yu[1], Conghao Xiong[1], Yankai Chen[2], Wenqian Cui[1], Xinni Zhang[1], Yifei Zhang[3], Hao Chen[4], Joseph J.Y. Sung[3], Irwin King[1]

[1]The Chinese University of Hong Kong
[2]University of Illinois Chicago
[3]Nanyang Technological University
[4]The Hong Kong University of Science and Technology
{dzyu23, chxiong21, wqcui23, xnzhang23, king}@cse.cuhk.edu.hk, ychen588@uic.edu,
{yifei.zhang, josephsung}@ntu.edu.sg, jhc@cse.ust.hk

Abstract

Survival prediction of cancers is crucial for clinical practice, as it informs mortality risks and influences treatment plans. However, a *static* model trained on a single dataset fails to adapt to the *dynamically evolving* clinical environment and continuous data streams, limiting its practical utility. While continual learning (CL) offers a solution to learn dynamically from new datasets, existing CL methods primarily focus on unimodal inputs and suffer from severe catastrophic forgetting in survival prediction. In real-world scenarios, multimodal inputs often provide comprehensive and complementary information, such as whole slide images and genomics; and neglecting inter-modal correlations negatively impacts the performance. To address the two challenges of *catastrophic forgetting* and *complex inter-modal interactions* between gigapixel whole slide images and genomics, we propose **ConSurv**, the **first** multimodal continual learning (MMCL) method for survival analysis. ConSurv incorporates two key components: Multi-staged Mixture of Experts (MS-MoE) and Feature Constrained Replay (FCR). MS-MoE captures both task-shared and task-specific knowledge at different learning stages of the network, including two modality encoders and the modality fusion component, learning inter-modal relationships. FCR further enhances learned knowledge and mitigates forgetting by restricting feature deviation of previous data at different levels, including encoder-level features of two modalities and the fusion-level representations. Additionally, we introduce a new benchmark integrating four datasets, Multimodal Survival Analysis Incremental Learning (MSAIL), for comprehensive evaluation in the CL setting. Extensive experiments demonstrate that ConSurv outperforms competing methods across multiple metrics.

Code — https://github.com/LucyDYu/ConSurv
Extended version — https://arxiv.org/abs/2511.09853

1 Introduction

Survival prediction plays an important role in clinical practice, as it quantifies mortality risks and informs therapeutic decision-making. Recent advances in deep learning have empowered researchers to make substantial progress in developing effective survival prediction models (Zadeh and Schmid 2020). Such efforts originally started with unimodal data sources like Whole Slide Images (WSIs) (Ilse, Tomczak, and Welling 2018; Lu et al. 2021; Shao et al. 2021; Zhang et al. 2022) or genomics (Klambauer et al. 2017; Vaswani et al. 2017; Chaudhary et al. 2018). More recently, researchers have incorporated multimodal inputs, including both WSIs and genomics (Chen et al. 2021, 2022; Li et al. 2022; Zhou and Chen 2023; Xu and Chen 2023; Xiong et al. 2024a), as they provide comprehensive and complementary information. For comprehensiveness, genomic features responsible for tumor formation correlate strongly with image patches of WSIs containing tumor cells (Chen et al. 2021). Regarding complementarity, WSIs are particularly valuable in late-stage cancer diagnosis, where survival outcomes are more predictable with morphological patterns. Meanwhile, genomic data provides critical insights in early-stage cancer, where genetic factors drive tumor progression.

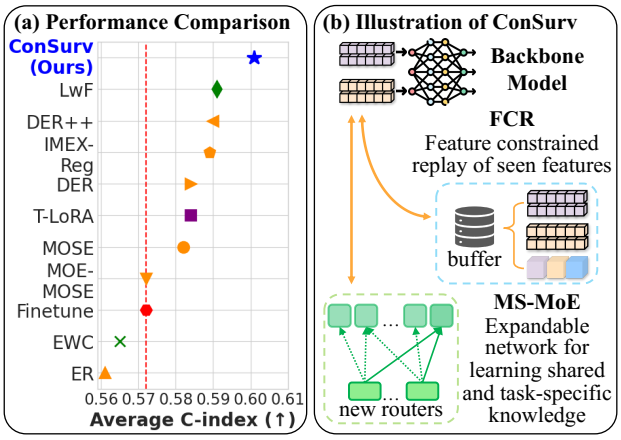


Figure 1: (a) Comparison of ConSurv against various CL methods on the MSAIL benchmark. (b) Illustration of ConSurv. It learns task-shared and task-specific multimodal knowledge during training with expandable MS-MoE, while consolidating previously acquired knowledge through FCR. A detailed architecture of ConSurv is in Figure 2.

However, given the *dynamically changing* clinical environment, the *static learning* paradigm, where a model is

trained on a single data source, struggles to adapt and generalize to new data (Pianykh et al. 2020). Such static models have limited sustainability and become outdated due to ongoing medical data collection, variations in staining protocols, and technological advancements that enhance the quality of WSI imaging and genomic data (Perkonigg et al. 2021; Shen et al. 2022; Huang et al. 2023). Moreover, data from different cancers often exhibit similar patterns (Baba and Câtoi 2010), and leveraging them appropriately enhances model performance (Xiong et al. 2024b). Consequently, it is both necessary and beneficial for models to learn from multiple datasets. Although retraining a model on new and existing datasets together is a possible solution, it incurs significant computational time and high resource cost.

Continual learning (CL) offers a viable framework to overcome the limitations of static models and repeated retraining (Huang et al. 2023; Yu et al. 2024a), thus enhancing the practical utility and effectiveness of survival prediction models. In the CL setting, a model is trained sequentially on multiple datasets, allowing it to adapt to new information without requiring explicit retraining on prior datasets. A direct approach is finetuning the well-trained model on new datasets, but this strategy leads to *catastrophic forgetting*, which means the model will suffer from severe performance decline on previously learned datasets (McCloskey and Cohen 1989; Ratcliff 1990). This phenomenon occurs because parameters are updated to accommodate new knowledge and thereby deviate from the old optimal state for the previous datasets (Hassabis et al. 2017). CL methods design mechanisms to mitigate catastrophic forgetting throughout the learning process. CL aims to effectively balance *plasticity*, the ability to acquire new knowledge continuously, and *stability*, the capacity to retain previously learned information. This is referred to as the *stability-plasticity dilemma* (Mermillod, Bugaiska, and Bonin 2013; Masana et al. 2022; Yu et al. 2024a), where prioritizing the retention of previously learned knowledge can compromise the acquisition of new knowledge, as this process inevitably diminishes the specific knowledge essential for previous datasets.

While extensive research has explored CL for unimodal data, work on multimodal data is limited (Yu et al. 2024a). Specifically, no CL research focuses on WSIs and genomics modalities. We find that directly applying unimodal CL methods to multimodal continual learning (MMCL) for survival prediction can lead to suboptimal performance. We identify two challenges: (C1) *Severe catastrophic forgetting*. As shown in Figure 1a, simple adaptations in this setting are not ideal, and sometimes even yield worse results than vanilla sequential finetuning, highlighting the severe catastrophic forgetting that current CL methods still face. (C2) *Complex inter-modal interactions*. While WSIs and genomic data offer complementary information, the model needs to effectively learn the complex correlations between these modalities (Xu, Zhu, and Clifton 2023), especially when these correlations vary across different datasets, which previous CL methods have neglected. Notably, these two challenges are intertwined, influencing and exacerbating each other. Catastrophic forgetting can be more severe when distinct modalities of WSIs and genomics are involved, due to different data sizes, inconsistent distributions and representations that arise from data heterogeneity (Baltrusaitis, Ahuja, and Morency 2019; Peng et al. 2021; Yu et al. 2024a). The learned multimodal interactions diminish due to catastrophic forgetting when the model concentrates on learning new cancer datasets.

To tackle the challenges above, we propose **Con**tinual **Surv**ival Analysis (**ConSurv**), a novel MMCL method designed to learn complex correlations from WSIs and genomics data and preserve previously acquired information throughout the CL process. We propose an expandable Multi-staged Mixture of Experts (MS-MoE) module (see Figure 1b). It facilitates the modeling of *complex intermodal relationships* during continual training by flexibly combining different experts. It captures both shared and task-specific knowledge from datasets at different multimodal learning stages within the model, specifically WSI and genomic encoders, and the fusion component. As new datasets are introduced, corresponding new routers are added to select relevant experts, which aids in mitigating catastrophic forgetting.

To further *enhance the retention of previous knowledge and mitigate forgetting*, we introduce Feature Constrained Replay (FCR). Since directly storing large WSIs and genomic data induces large storage costs, we store only the processed instance feature representations for replay. During training, FCR constrains the deviation of individual features of both modalities after their respective encoders, and the final fused representations of previous datasets to alleviate forgetting through knowledge distillation (Gou et al. 2021).

Acknowledging the absence of benchmarks in MMCL for survival prediction using WSIs and genomics, we propose a new challenging **M**ultimodal **S**urvival **A**nalysis **I**ncremental **L**earning (**MSAIL**) benchmark to evaluate various CL methods. This benchmark utilizes four publicly available survival prediction datasets from The Cancer Genome Atlas Program (TCGA), namely BLCA, UCEC, LUAD, and BRCA. We evaluate our ConSurv on the MSAIL benchmark, and it outperforms other methods on multiple metrics through extensive experiments. We emphasize that ConSurv successfully achieves an effective balance in the stability-plasticity trade-off, effectively acquiring new knowledge while retaining previously learned information.

The contributions of our paper are summarized as follows:

1. We propose ConSurv, the **first** MMCL method for survival analysis across multiple cancers using WSI and genomic data.
2. We design MS-MoE to handle complex, dynamic intermodal interactions, and FCR to alleviate catastrophic forgetting.
3. We propose a new MSAIL benchmark for evaluation, covering datasets of four cancers from TCGA.
4. The extensive experiments on MSAIL demonstrate not only the superiority of ConSurv over competing methods, but also the effectiveness of the proposed modules.

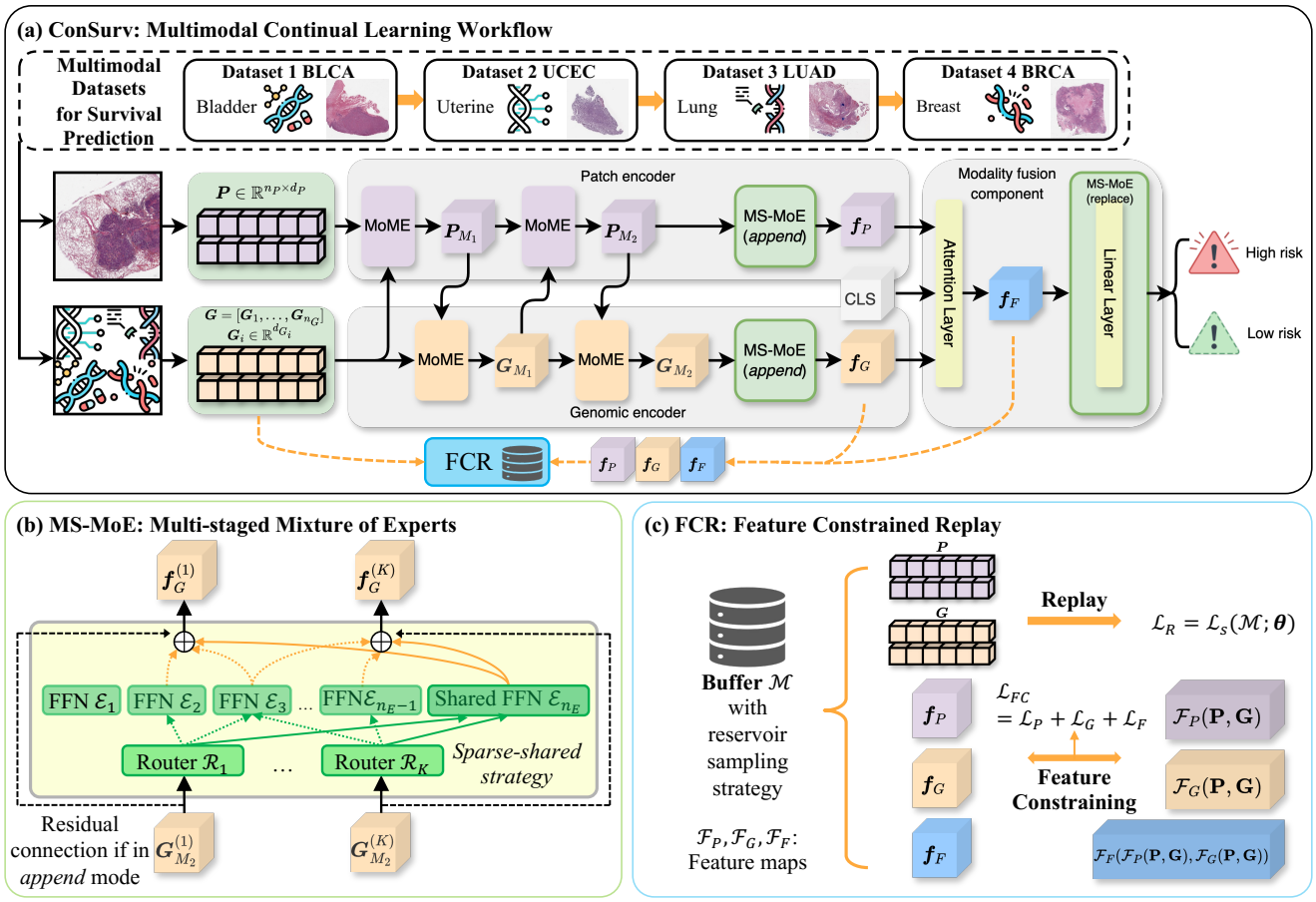


Figure 2: Overall architecture of ConSurv. (a) The MMCL workflow for continual survival prediction across different cancer datasets. We employ a recent SOTA model, MoME (Xiong et al. 2024a), in survival prediction as our backbone model. (b) MS-MoE learns both shared and task-specific knowledge at different learning stages of the network, including WSI and genomic encoders and the modality fusion component. (c) FCR preserves previously learned knowledge through additional loss terms on the replay buffer.

2 ConSurv Methodology

2.1 MMCL Workflow

MMCL for survival prediction is a *CL* setting where a model sequentially learns to predict the survival time on *multimodal* datasets. We provide detailed discussions on related work and preliminaries for survival prediction in Appendix A and B, respectively. We define a *multimodal dataset sequence* $\mathcal{D} = \{\mathcal{D}_1, \mathcal{D}_2, \ldots, \mathcal{D}_K\}$, where $\mathcal{D}_i$ are different multimodal cancer datasets and K is the number of datasets. In the MMCL setting, the model is trained on the current dataset $\mathcal{D}_k$ (optionally with limited access to previous datasets). During training, the model parameters $\boldsymbol{\theta}$ are updated in a controlled manner, facilitating learning on $\mathcal{D}_k$ while mitigating the forgetting of knowledge learned from $\{\mathcal{D}_1, \mathcal{D}_2, \cdots, \mathcal{D}_{k-1}\}$. The CL loss of dataset k has the following general form (Wang and Huang 2024):

$$\mathcal{L}_{CL}^{(k)}(\boldsymbol{\theta}) = \mathcal{L}_s(\mathcal{D}_k; \boldsymbol{\theta}) + \zeta \mathcal{L}_f^{(k)}(\boldsymbol{\theta}), \tag{1}$$

where $\mathcal{L}_s$ is the loss for survival prediction, applying to the current dataset $\mathcal{D}_k$; $\mathcal{L}_f$ is a forgetting-mitigation term, such as memory-replay and parameter-regularization loss; ζ denotes a constant that balances knowledge acquisition and forgetting avoidance in the stability-plasticity trade-off (Wang and Huang 2024). Model minimizes $\mathcal{L}_{CL}^{(k)}$ when training on each respective dataset $\mathcal{D}_k$, i.e., $\boldsymbol{\theta}^{*(k)} = \text{argmin}_{\boldsymbol{\theta}}\ \mathcal{L}_{CL}^{(k)}(\boldsymbol{\theta})$. In terms of model performance, CL aims to obtain a model which achieves high performance on all trained datasets, i.e., $\boldsymbol{\theta}^* = \text{argmax}_{\boldsymbol{\theta}} \sum_{i=1}^{K} P(\boldsymbol{\theta}, \mathcal{D}_i)$, where $P(\boldsymbol{\theta}, \mathcal{D}_i)$ is the performance evaluation function based on different types of datasets and tasks. In the case of survival prediction, the function is typically the Concordance index (C-index) and C-index IPCW (inverse probability of censoring weights) (Uno et al. 2011).

We focus on the task-incremental learning (TIL) setting in this work (van de Ven, Tuytelaars, and Tolias 2022). In the CL literature, the term "task" directly corresponds to its dataset, and therefore, terms "task" and "dataset" are often used interchangeably. In TIL, for any two distinct datasets $i \neq j$, $\mathcal{D}_i$ and $\mathcal{D}_j$ exhibit different input distributions and

label spaces (Wang et al. 2024). This aligns with our setting, where multimodal datasets correspond to different cancer types and exhibit different survival time distributions. Although we partition the survival time of each dataset into the same number of bins, the time intervals of bins are all different, hence different label spaces. Task identities are available at inference. We employ MoME (Xiong et al. 2024a) as the multimodal backbone and apply our proposed modules MS-MoE and FCR in the MMCL training workflow (Sections 2.2 and 2.3).

2.2 MS-MoE: Multi-staged Mixture of Experts

To address the challenge of learning dynamic inter-modal interactions between WSIs and genomics data in cancer datasets, we introduce Multi-staged Mixture of Experts (MS-MoE). It is an expandable module designed for integration with the multimodal backbone while preserving the backbone's core architecture (Figure 2a and 2b). We adopt and modify the Mixture of Experts (MoE) (Shazeer et al. 2017) as the base component of MS-MoE.

Routing Mechanism. The original MoE comprises a set of "expert" subnetworks and a "router" that selects them. Instead of adding new experts for each new cancer dataset $\mathcal{D}_k$, we keep a fixed number n_E of experts $\{\mathcal{E}_i\}_{i=1}^{n_E}$ and only introduce a new linear-layer router $\mathcal{R}_k$, to limit parameter growth, following (Yu et al. 2024b). Using task-specific routers helps to *mitigate catastrophic forgetting* when learning a new cancer type. We employ Sparse MoE (Jiang et al. 2024; Fedus, Dean, and Zoph 2022) to selectively utilize experts instead of always using all experts. This strategy reduces computational costs and encourages experts to learn task-specific knowledge. Additionally, when the same expert is selected for inputs from a subset of tasks, this strategy facilitates inter-task collaboration and the learning of shared knowledge. We visualize and support the above claims in Section 3.6. To enable the module to acquire knowledge across all datasets, we designate one expert as a shared expert, ensuring it remains consistently active, inspired by (Rajbhandari et al. 2022). The gating weights $W^{(k)}$ employing this *sparse-shared strategy* are defined as:

$$W^{(k)} = \text{Softmax}\left(\text{TopK-S}\left(\mathcal{R}_k(\mathbf{x}^{(k)})\right)\right), \tag{2}$$

where TopK-S$(\cdot)$ selects the shared expert and top k experts among the rest experts, while setting non-selected ones to be $-\infty$. Softmax$(\cdot)$ normalizes the weights. For input $\mathbf{x}^{(k)}$ when training on $\mathcal{D}_k$, the output $\mathbf{y}^{(k)}$ of the MoE module $\mathcal{MS}$ is defined as:

$$\mathbf{y}^{(k)} = \mathcal{MS}(\mathbf{x}^{(k)}) = \sum_{i=1}^{n_E} W_i^{(k)} \mathcal{E}_i(\mathbf{x}^{(k)}), \tag{3}$$

where $W_i^{(k)}$ denotes the i-th entry of $W^{(k)}$.

Integration into Backbone. Consistent with our goal of integrating the above MS-MoE modules into the backbone model while maintaining its core structure, we introduce two integration modes: *replace* and *append*. If the target insertion point within the MoME architecture contains a standard Feed-Forward Network (FFN) or linear layer, we *replace* this component entirely with an MS-MoE module. The experts $\{\mathcal{E}_i\}$ within this module are configured to have an architecture identical to that of the replaced component. In this mode, if the learned gating weights strongly favor the shared expert (i.e., weight close to 1), the computation approximates that of the original layer, and the modified structure effectively reduces to the original backbone. If the target insertion point lacks such layers to replace, we *append* the MS-MoE module residually. We employ two-layer FFNs as the experts $\{\mathcal{E}_i\}$. The output $\mathbf{y}^{(k)}$ incorporates a residual connection (He et al. 2016) from the input: $\mathbf{y}^{(k)} = \mathbf{x}^{(k)} + \mathcal{MS}(\mathbf{x}^{(k)})$. In this configuration, if the activated experts learn negligible transformations (effectively mapping inputs close to zero), the output $\mathbf{y}^{(k)}$ approximates the original input $\mathbf{x}^{(k)}$, again preserving the backbone's information flow. Consequently, these integration strategies allow MS-MoE to introduce additional capacity and flexibility for CL while maintaining the integrity and performance baseline of the original MoME architecture.

Capturing Inter-modal Interactions. To effectively capture and adapt the *complex inter-modal interactions* between WSI and genomic data, MS-MoE modules are strategically placed at the end of the WSI encoder, the genomic encoder, and the modality fusion component of the MoME backbone. As noted in (Xiong et al. 2024a), MoME's encoders already perform cross-modal processing (e.g., the patch encoder uses genomic information, and vice-versa). By adding MS-MoE at all three key stages, our approach aims to enhance the learning of these multimodal interactions while maintaining CL capabilities, adapting expert knowledge and task-specific routing as new cancer datasets are encountered.

2.3 FCR: Feature Constrained Replay

To further *alleviate catastrophic forgetting*, we introduce Feature Constrained Replay (FCR), designed to retain previously acquired knowledge when training on multimodal datasets, as depicted in Figure 2c. The FCR module maintains features of a small fixed number of seen instances. During training, it constrains the deviation of features from WSIs and genomics data after their respective encoders and the final fused representations of previous datasets to mitigate forgetting. Let $\mathcal{F}_P$, $\mathcal{F}_G$, and $\mathcal{F}_F$ denote the feature maps of the WSI patch encoder, genomic encoder, and the final fusion component, respectively. Given WSI patches and genomic data $(\mathbf{P}, \mathbf{G})$, we obtain patch feature representation $\mathbf{f}_P = \mathcal{F}_P(\mathbf{P}, \mathbf{G})$, genomic feature representation $\mathbf{f}_G = \mathcal{F}_G(\mathbf{P}, \mathbf{G})$, and the final fusion representation $\mathbf{f}_F = \mathcal{F}_F(\mathbf{f}_P, \mathbf{f}_G)$.

We introduce a fixed-size replay buffer $\mathcal{M}$ and utilize a reservoir sampling strategy (Vitter 1985) to randomly select sample features from the input data stream and update the buffer, ensuring equal retention probability for all seen instances. The buffer maintains three types of features: $\mathbf{f}_P, \mathbf{f}_G$, and $\mathbf{f}_F$. For the final fusion representations, an $\mathcal{L}_2$ loss is utilized as a feature distillation technique. It minimizes the

Train \ Val	BLCA	UCEC	LUAD	BRCA	Average (on trained)	Random
BLCA	**0.607±0.026**	0.545±0.025	-	-	0.607±0.026	0.519±0.052
UCEC	0.500±0.055	**0.648±0.080**	0.481±0.045	-	0.574±0.061	0.420±0.024
LUAD	0.512±0.046	0.607±0.076	**0.642±0.046**	0.518±0.044	0.587±0.036	0.475±0.073
BRCA	0.531±0.035	0.588±0.062	0.519±0.041	**0.650±0.059**	0.572±0.024	0.473±0.089

Table 1: Performance (C-index) of each dataset under sequential finetuning.

distance between the representation $\mathbf{f}_F$ from the buffer $\mathcal{M}$ and that from the current model by following (Gou et al. 2021; Bai, Islam, and Ren 2023):

$$\mathcal{L}_F(\mathcal{M};\boldsymbol{\theta}) = \mathbb{E}_{(\mathbf{P},\mathbf{G},\mathbf{f}_F)\sim\mathcal{M}}[\|\mathbf{f}_F - \mathcal{F}_F(\mathcal{F}_P(\mathbf{P},\mathbf{G}),\mathcal{F}_G(\mathbf{P},\mathbf{G}))\|_2^2]. \quad (4)$$

We define the loss for patch feature representations as:

$$\mathcal{L}_P(\mathcal{M};\boldsymbol{\theta}) = \mathbb{E}_{(\mathbf{P},\mathbf{G},\mathbf{f}_P)\sim\mathcal{M}}[\|\mathbf{f}_P - \mathcal{F}_P(\mathbf{P},\mathbf{G})\|_2^2], \quad (5)$$

and similarly for $\mathcal{L}_G(\mathcal{M};\boldsymbol{\theta})$. The total feature constraint loss is then given by:

$$\mathcal{L}_{FC}(\mathcal{M};\boldsymbol{\theta}) = \mathcal{L}_P(\mathcal{M};\boldsymbol{\theta}) + \mathcal{L}_G(\mathcal{M};\boldsymbol{\theta}) + \mathcal{L}_F(\mathcal{M};\boldsymbol{\theta}). \quad (6)$$

Furthermore, to leverage the replay buffer, the model is trained on the data points within the buffer using their ground truth labels, following ER (Chaudhry et al. 2019) and DER++ (Buzzega et al. 2020). This additional replay loss is denoted as $\mathcal{L}_R(\mathcal{M};\boldsymbol{\theta}) = \mathcal{L}_s(\mathcal{M};\boldsymbol{\theta})$. Consequently, during training on dataset $\mathcal{D}_k$, the overall loss for the model incorporating FCR is:

$$\mathcal{L}_{CL}^{(k)}(\boldsymbol{\theta}) = \mathcal{L}_s(\mathcal{D}_k;\boldsymbol{\theta}) + \alpha\mathcal{L}_{FC}(\mathcal{M};\boldsymbol{\theta}) + \beta\mathcal{L}_R(\mathcal{M};\boldsymbol{\theta}), \quad (7)$$

where α and β are hyperparameters that control the relative weights of the losses, enabling a balance between learning from the current dataset $\mathcal{D}_k$ and preserving previous knowledge. Note that here $\mathcal{L}_{FC}$ and $\mathcal{L}_R$ are not superscripted by k because their values depend on the update of buffer $\mathcal{M}$ throughout the training process.

3 Experiments

3.1 MSAIL Benchmark

We propose a new benchmark, namely **M**ultimodal **S**urvival **A**nalysis **I**ncremental **L**earning (MSAIL) to evaluate different CL methods. Experimental settings, evaluation protocol and implementation details are provided in Appendix C.

Data. Our MSAIL benchmark consists of four multimodal survival analysis datasets, which we collectively refer to as **Cancer4** for brevity in the context of continual training. These datasets are from The Cancer Genome Atlas Program (TCGA). Specifically, they are 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). The task order used for this benchmark is BLCA, UCEC, LUAD, and BRCA. We present an alternative task order and the results in Appendix C.4.

Metrics. To evaluate performance on individual datasets, we employ the Concordance index (C-index) as our evaluation metric. We moreover utilize C-index IPCW (inverse probability of censoring weights) (Uno et al. 2011) as another metric, which adjusts the bias introduced by censoring. To evaluate the model in the MMCL setting, we compute *Average Performance* of the above two metrics as the main metrics. We additionally report *Forgetting* (Chaudhry et al. 2018), *Backward Transfer* (BWT) and *Forward Transfer* (FWT) (Lopez-Paz and Ranzato 2017) for reference.

3.2 Research Questions

We aim to answer the following research questions:

- **RQ1**: Motivation for CL. **(1)** Does CL offer a performance benefit over a static model on new datasets? **(2)** What is the severity of catastrophic forgetting with direct sequential finetuning?
- **RQ2**: Performance comparison. How effective is our proposed ConSurv compared with other CL methods?
- **RQ3**: In-depth Analysis of ConSurv. **(1)** How effectively does ConSurv stratify patients into risk groups on each cancer dataset? **(2)** How does each component of ConSurv impact the performance? **(3)** How effective is the routing mechanism of MS-MoE for expert selection?

3.3 Experimental Results

Necessity of Continual Learning (RQ1.1). We first explore whether a static model can generalize to new datasets. We directly evaluate the initial model on all tasks before training, as the random performance baseline. As shown in Table 1, a model trained on BLCA achieves a C-index of 0.607. We then evaluate it on the next dataset, UCEC, before training, and the performance is 0.545, which exceeds the random performance. This positive forward knowledge transfer is consistently observed when evaluating on each subsequent task, indicating the acquisition of shared knowledge. However, this transferred performance is significantly lower than the performance after training on the respective dataset (e.g., 0.648 for UCEC), thus highlighting the necessity for CL approaches to achieve optimal performance on new datasets. Importantly, the existence of positive forward knowledge transfer enhances the efficacy of CL by providing an informed starting point for subsequent training.

Quantification of Catastrophic Forgetting (RQ1.2). A common baseline in CL is to finetuning the model on new datasets. We next examine the presence of catastrophic forgetting under this new setting. As evidenced in Table 1, the performance on the initially learned task drops dramatically

Type	Method	C-index				C-index IPCW			
		Average (↑)	Forget (↓)	BWT (↑)	FWT (↑)	**Average** (↑)	Forget (↓)	BWT (↑)	FWT (↑)
Base-line	Joint	0.611±0.037	-	-	-	0.545±0.045	-	-	-
	Finetune	0.572±0.024	0.094±0.041	-0.086±0.047	0.058±0.007	0.528±0.055	0.136±0.091	-0.101±0.108	0.022±0.160
Reg.	EWC	0.565±0.055	0.115±0.068	-0.111±0.072	0.050±0.021	0.510±0.067	0.105±0.093	-0.084±0.117	0.007±0.130
	LwF	0.591±0.034	0.072±0.039	-0.065±0.047	0.040±0.043	0.589±0.029	0.065±0.056	-0.019±0.093	0.020±0.156
Arch.	T-LoRA	0.584±0.022	**0.000±0.004**	**0.002±0.005**	0.048±0.047	0.553±0.046	0.053±0.042	-0.022±0.052	0.056±0.141
Re-play	ER	0.561±0.025	0.110±0.052	-0.110±0.052	0.014±0.031	0.511±0.057	0.107±0.056	-0.088±0.051	-0.046±0.111
	DER	0.584±0.017	0.100±0.041	-0.098±0.043	0.074±0.042	0.544±0.024	0.107±0.031	-0.077±0.082	0.037±0.165
	DER++	0.590±0.030	0.082±0.019	-0.074±0.009	**0.084±0.026**	0.541±0.074	0.102±0.097	-0.074±0.114	0.055±0.083
	MOSE	0.582±0.028	0.090±0.024	-0.090±0.024	-0.008±0.071	0.548±0.065	0.091±0.061	-0.074±0.041	-0.027±0.133
	MOE-MOSE	0.572±0.023	0.116±0.040	-0.113±0.044	-0.037±0.087	0.547±0.054	0.109±0.075	-0.077±0.052	-0.011±0.163
	IMEX-Reg	0.589±0.042	0.092±0.036	-0.087±0.037	0.048±0.049	0.589±0.032	**0.037±0.070**	-0.009±0.068	0.054±0.139
	ConSurv	**0.601±0.045**	0.088±0.052	-0.081±0.060	0.067±0.059	**0.597±0.039**	0.049±0.048	**0.002±0.080**	**0.083±0.103**

Table 2: Comparison results among different CL methods. The best performances are highlighted in bold. The main metrics are average C-index and average C-index IPCW. Forgetting, BWT, and FWT are reported for reference.

FCR	MS-MoE	C-index				C-index IPCW			
		Average (↑)	Forget (↓)	BWT (↑)	FWT (↑)	**Average** (↑)	Forget (↓)	BWT (↑)	FWT (↑)
		0.572±0.024	0.094±0.041	-0.086±0.047	0.058±0.007	0.528±0.055	0.136±0.091	-0.101±0.108	0.022±0.160
✓		0.581±0.043	0.121±0.061	-0.119±0.061	**0.079±0.059**	0.545±0.045	0.121±0.096	-0.112±0.104	0.042±0.138
	✓	0.585±0.022	**0.079±0.034**	**-0.079±0.034**	0.041±0.022	0.575±0.015	0.078±0.059	-0.065±0.054	0.067±0.087
✓(f)	✓	0.599±0.026	0.083±0.030	-0.082±0.030	0.044±0.051	0.554±0.045	0.086±0.104	-0.060±0.093	-0.022±0.111
✓	✓	**0.601±0.045**	0.088±0.052	-0.081±0.060	0.067±0.059	**0.597±0.039**	**0.049±0.048**	**0.002±0.080**	**0.083±0.103**

Table 3: Ablation study of FCR and MS-MoE in ConSurv. "✓(f)" denotes FCR with only the final fusion representation, as opposed to all three feature levels.

from 0.607 down to 0.531 when the BLCA-trained model is subsequently trained on UCEC, LUAD, and BRCA. This phenomenon of forgetting previously learned tasks is consistently observed throughout the training sequence, demonstrating severe catastrophic forgetting.

Comparison with Other CL Methods (RQ2). We compare ConSurv with finetuning and other SOTA unimodal CL methods in Table 2. Notably, several of them exhibit inferior performance compared to finetuning, suggesting that neglecting the complex inter-modal interactions during continual training negatively impacts the performance. Our ConSurv method outperforms all other methods in the main metrics: average C-index and average C-index IPCW. Furthermore, it achieves the highest BWT and FWT for C-index IPCW. Other metrics of ConSurv are not the highest, since there is a trade-off between absolute performance and resistance to forgetting, thus they cannot comprehensively assess the effectiveness of ConSurv (Huang et al. 2023). We list those metrics for reference, following previous works (Huang et al. 2023; Lopez-Paz and Ranzato 2017; Chaudhry et al. 2018).

3.4 Kaplan–Meier Analysis (RQ3.1)

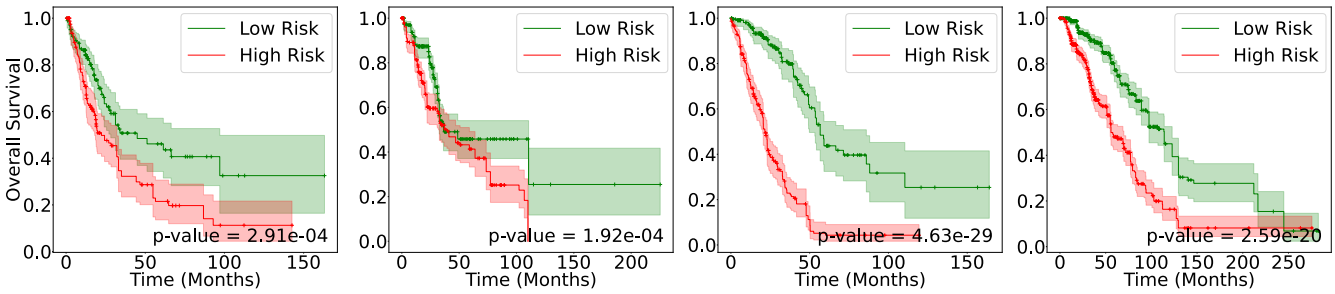
To further validate the differentiability of our ConSurv on each dataset in Cancer4, we perform a Kaplan–Meier analysis with the final model trained under the MMCL setting. Based on the mean risk value of a dataset, we partition patients into low-risk and high-risk groups (Xiong et al. 2024c). The survival outcomes for all patients are visualized in Figure 3. To assess the statistical significance of the difference between the two risk groups, we conduct a log-rank test, following (Xiong et al. 2024c), with a p-value less than 0.05 considered statistically significant by convention. As illustrated in Figure 3, ConSurv successfully stratifies patients into low-risk and high-risk groups with high statistical significance, thus demonstrating its ability to learn and retain knowledge from multimodal data throughout the CL process while effectively mitigating catastrophic forgetting.

3.5 Ablation Study (RQ3.2)

We conduct an ablation study on our proposed FCR and MS-MoE modules to investigate their individual effects. The results are presented in Table 3.

The Effects of MS-MoE. As shown in Table 3, employing MS-MoE improves the average C-index and average C-index IPCW, compared to the finetuning baseline (the first row). Note that MS-MoE operates without the need for data replay. Thus, the buffer is not used. This observation suggests that MS-MoE effectively facilitates learning of both shared and task-specific knowledge across datasets, while alleviating forgetting.

The Effects of FCR. The results in Table 3 indicate that utilizing FCR in isolation increases both the average C-index



(a) BLCA (b) UCEC (c) LUAD (d) BRCA

Figure 3: Kaplan-Meier curves of our ConSurv on Cancer4.

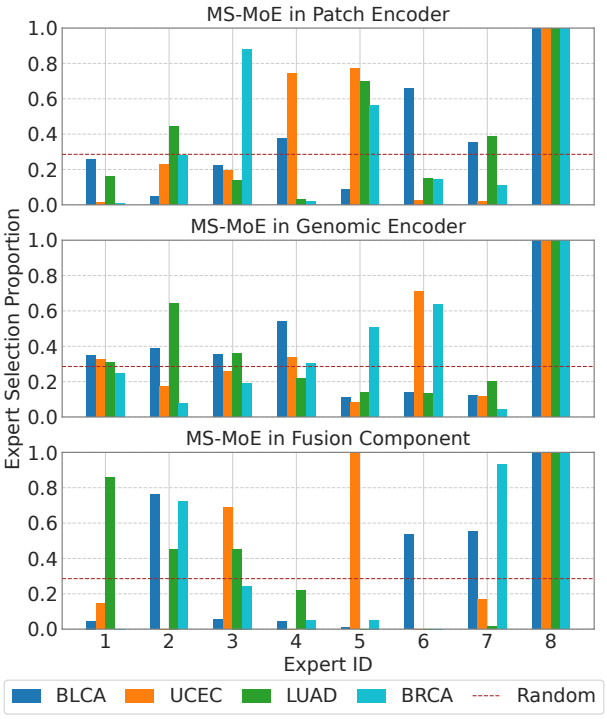


Figure 4: Proportion of each expert within the MS-MoE modules selected on inputs from different datasets. The brown dashed line represents the expected selection proportion under random sampling, which is 2/7. The last expert $\mathcal{E}_8$ functions as a shared expert and is always selected.

and the average C-index with IPCW, demonstrating its effectiveness. Furthermore, based on ConSurv with two modules, we investigate the performance when only the final fusion representation is constrained within FCR (the "✓(f)" row in Table 3). We discover that it already achieves the highest average C-index, compared with other CL methods, highlighting the efficacy of feature constraints. Constraining features at the patch, genomics, and fusion levels collectively results in further improvements across most evaluation metrics (comparing the last two rows). These findings suggest that the two modules collaborate effectively, ultimately leading to the best overall performance.

3.6 MS-MoE Routing Analysis (RQ3.3)

This section presents an analysis of MS-MoE expert selection performed by the routers $\mathcal{R}_k$ for each dataset $\mathcal{D}_k$. Our investigation aims to determine whether experts can acquire task-specific and shared knowledge across subsets of the datasets. As illustrated in Figure 4, we quantify the selection proportion for each expert on the validation datasets. The results reveal a diversity in expert preferences across different datasets. Some experts specialize in learning knowledge important to a single dataset; for example, expert $\mathcal{E}_3$ within the patch encoder's MS-MoE focuses on BRCA. Conversely, some experts are selected across multiple tasks, suggesting the acquisition of shared knowledge; for instance, expert $\mathcal{E}_6$ within the genomic encoder's MS-MoE has learned knowledge relevant to UCEC and BRCA. This demonstrates MS-MoE's capacity for appropriate expert selection, which facilitates ConSurv's learning of multimodal knowledge throughout the CL process, providing further evidence for the effectiveness of MS-MoE.

4 Conclusion

In this work, we first explore the necessity of CL in multimodal survival prediction and quantify severe catastrophic forgetting in this new setting. We propose **ConSurv**, the **first** MMCL method for survival analysis, to tackle the challenges of forgetting and complex inter-modal interactions between gigapixel WSIs and genomics in different cancers. The proposed MS-MoE effectively learn shared and task-specific knowledge at different learning stages of the network, including WSI and genomic encoders and the modality fusion component. We design FCR to enhance learned knowledge by limiting feature deviation at multiple levels, including encoder-level features of two modalities and the fusion-level representations. In addition, we establish the new MSAIL benchmark by integrating TCGA datasets and utilize it for evaluation. Extensive experiments demonstrate that ConSurv surpasses other methods across multiple metrics, with a better trade-off between acquiring new knowledge and retaining previously learned information. A detailed analysis of computational costs, limitations, and future work is provided in Appendix D.

Acknowledgments

The work described in this paper was partially supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (CUHK 2300246, RGC C1043-24G). We sincerely thank Professor Irwin King, for his unwavering support and expert guidance throughout every stage of this work. We sincerely thank Professor Joseph J. Y. Sung, for his invaluable and insightful suggestions that enhanced this paper, especially regarding the contrast between static and dynamic models, and the clinical importance of a dynamic model in survival prediction.

References

Baba, A. I.; and Câtoi, C. 2010. Comparative Oncology. *Publishing House of the Romanian Academy*.

Bai, L.; Islam, M.; and Ren, H. 2023. Revisiting Distillation for Continual Learning on Visual Question Localized-Answering in Robotic Surgery. In Greenspan, H.; Madabhushi, A.; Mousavi, P.; Salcudean, S.; Duncan, J.; Syeda-Mahmood, T.; and Taylor, R., eds., *Medical Image Computing and Computer Assisted Intervention – MICCAI 2023*, 68–78. Cham: Springer Nature Switzerland. ISBN 978-3-031-43996-4.

Baltrusaitis, T.; Ahuja, C.; and Morency, L.-P. 2019. Multimodal Machine Learning: A Survey and Taxonomy. *IEEE Trans. Pattern Anal. Mach. Intell.*, 41(2): 423–443.

Buzzega, P.; Boschini, M.; Porrello, A.; Abati, D.; and Calderara, S. 2020. Dark Experience for General Continual Learning: A Strong, Simple Baseline. *Advances in neural information processing systems*, 33: 15920–15930.

Chaudhary, K.; Poirion, O. B.; Lu, L.; and Garmire, L. X. 2018. Deep Learning–Based Multi-Omics Integration Robustly Predicts Survival in Liver Cancer. *Clinical cancer research*, 24(6): 1248–1259.

Chaudhry, A.; Dokania, P. K.; Ajanthan, T.; and Torr, P. H. 2018. Riemannian Walk for Incremental Learning: Understanding Forgetting and Intransigence. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 532–547.

Chaudhry, A.; Rohrbach, M.; Elhoseiny, M.; Ajanthan, T.; Dokania, P. K.; Torr, P. H. S.; and Ranzato, M. 2019. On Tiny Episodic Memories in Continual Learning. arxiv:1902.10486.

Chen, R. J.; Lu, M. Y.; Weng, W.-H.; Chen, T. Y.; Williamson, D. F.; Manz, T.; Shady, M.; and Mahmood, F. 2021. Multimodal Co-Attention Transformer for Survival Prediction in Gigapixel Whole Slide Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4025.

Chen, R. J.; Lu, M. Y.; Williamson, D. F.; Chen, T. Y.; Lipkova, J.; Noor, Z.; Shaban, M.; Shady, M.; Williams, M.; and Joo, B. 2022. Pan-Cancer Integrative Histology-Genomic Analysis via Multimodal Deep Learning. *Cancer cell*, 40(8): 865–878.

Fedus, W.; Dean, J.; and Zoph, B. 2022. A Review of Sparse Expert Models in Deep Learning. arxiv:2209.01667.

Gou, J.; Yu, B.; Maybank, S. J.; and Tao, D. 2021. Knowledge Distillation: A Survey. *International Journal of Computer Vision*, 129(6): 1789–1819.

Hassabis, D.; Kumaran, D.; Summerfield, C.; and Botvinick, M. 2017. Neuroscience-Inspired Artificial Intelligence. *Neuron*, 95(2): 245–258.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 770–778. Las Vegas, NV, USA: IEEE. ISBN 978-1-4673-8851-1.

Huang, Y.; Zhao, W.; Wang, S.; Fu, Y.; Jiang, Y.; and Yu, L. 2023. 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*, 21349–21360.

Ilse, M.; Tomczak, J.; and Welling, M. 2018. Attention-Based Deep Multiple Instance Learning. In *International Conference on Machine Learning*, 2127–2136. PMLR.

Jiang, A. Q.; Sablayrolles, A.; Roux, A.; Mensch, A.; Savary, B.; Bamford, C.; Chaplot, D. S.; de las Casas, D.; Hanna, E. B.; Bressand, F.; Lengyel, G.; Bour, G.; Lample, G.; Lavaud, L. R.; Saulnier, L.; Lachaux, M.-A.; Stock, P.; Subramanian, S.; Yang, S.; Antoniak, S.; Scao, T. L.; Gervet, T.; Lavril, T.; Wang, T.; Lacroix, T.; and Sayed, W. E. 2024. Mixtral of Experts. arxiv:2401.04088.

Klambauer, G.; Unterthiner, T.; Mayr, A.; and Hochreiter, S. 2017. Self-Normalizing Neural Networks. *Advances in neural information processing systems*, 30.

Li, R.; Wu, X.; Li, A.; and Wang, M. 2022. HFBSurv: Hierarchical Multimodal Fusion with Factorized Bilinear Models for Cancer Survival Prediction. *Bioinformatics*, 38(9): 2587–2594.

Lopez-Paz, D.; and Ranzato, M. 2017. Gradient Episodic Memory for Continual Learning. *Advances in neural information processing systems*, 30.

Lu, M. Y.; Williamson, D. F.; Chen, T. Y.; Chen, R. J.; Barbieri, M.; and Mahmood, F. 2021. Data-Efficient and Weakly Supervised Computational Pathology on Whole-Slide Images. *Nature biomedical engineering*, 5(6): 555–570.

Masana, M.; Liu, X.; Twardowski, B.; Menta, M.; Bagdanov, A. D.; and Van De Weijer, J. 2022. Class-Incremental Learning: Survey and Performance Evaluation on Image Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–20.

McCloskey, M.; and Cohen, N. J. 1989. Catastrophic Interference in Connectionist Networks: The Sequential Learning Problem. In *Psychology of Learning and Motivation*, volume 24, 109–165. Elsevier. ISBN 978-0-12-543324-2.

Mermillod, M.; Bugaiska, A.; and Bonin, P. 2013. The Stability-Plasticity Dilemma: Investigating the Continuum from Catastrophic Forgetting to Age-Limited Learning Effects. *Frontiers in Psychology*, 4.

Peng, Y.; Qi, J.; Ye, Z.; and Zhuo, Y. 2021. Hierarchical Visual-Textual Knowledge Distillation for Life-Long Correlation Learning. *International Journal of Computer Vision*, 129(4): 921–941.

Perkonigg, M.; Hofmanninger, J.; Herold, C. J.; Brink, J. A.; Pianykh, O.; Prosch, H.; and Langs, G. 2021. Dynamic Memory to Alleviate Catastrophic Forgetting in Continual Learning with Medical Imaging. *Nature communications*, 12(1): 5678.

Pianykh, O. S.; Langs, G.; Dewey, M.; Enzmann, D. R.; Herold, C. J.; Schoenberg, S. O.; and Brink, J. A. 2020. Continuous Learning AI in Radiology: Implementation Principles and Early Applications. *Radiology*, 297(1): 6–14.

Rajbhandari, S.; Li, C.; Yao, Z.; Zhang, M.; Aminabadi, R. Y.; Awan, A. A.; Rasley, J.; and He, Y. 2022. DeepSpeed-MoE: Advancing Mixture-of-Experts Inference and Training to Power next-Generation AI Scale. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; and Sabato, S., eds., *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 18332–18346. PMLR.

Ratcliff, R. 1990. Connectionist Models of Recognition Memory: Constraints Imposed by Learning and Forgetting Functions. *Psychological Review*, 97(2): 285–308.

Shao, Z.; Bian, H.; Chen, Y.; Wang, Y.; Zhang, J.; and Ji, X. 2021. Transmil: Transformer Based Correlated Multiple Instance Learning for Whole Slide Image Classification. *Advances in neural information processing systems*, 34: 2136–2147.

Shazeer, N.; Mirhoseini, A.; Maziarz, K.; Davis, A.; Le, Q.; Hinton, G.; and Dean, J. 2017. Outrageously Large Neural Networks: The Sparsely-Gated Mixture-of-Experts Layer. In *International Conference on Learning Representations*.

Shen, Y.; Sowmya, A.; Luo, Y.; Liang, X.; Shen, D.; and Ke, J. 2022. A Federated Learning System for Histopathology Image Analysis with an Orchestral Stain-Normalization GAN. *IEEE Transactions on Medical Imaging*, 42(7): 1969–1981.

Uno, H.; Cai, T.; Pencina, M. J.; D'Agostino, R. B.; and Wei, L. J. 2011. On the C-statistics for Evaluating Overall Adequacy of Risk Prediction Procedures with Censored Survival Data. *Statistics in Medicine*, 30(10): 1105–1117.

van de Ven, G. M.; Tuytelaars, T.; and Tolias, A. S. 2022. Three Types of Incremental Learning. *Nature Machine Intelligence*, 4(12): 1185–1197.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; ukasz Kaiser, Ł.; and Polosukhin, I. 2017. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Vitter, J. S. 1985. Random Sampling with a Reservoir. *ACM Transactions on Mathematical Software*, 11(1): 37–57.

Wang, L.; Zhang, X.; Su, H.; and Zhu, J. 2024. A Comprehensive Survey of Continual Learning: Theory, Method and Application. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1–20.

Wang, Z.; and Huang, H. 2024. Model Sensitivity Aware Continual Learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Xiong, C.; Chen, H.; Zheng, H.; Wei, D.; Zheng, Y.; Sung, J. J. Y.; and King, I. 2024a. MoME: Mixture of Multimodal Experts for Cancer Survival Prediction. In Linguraru, M. G.; Dou, Q.; Feragen, A.; Giannarou, S.; Glocker, B.; Lekadir, K.; and Schnabel, J. A., eds., *Medical Image Computing and Computer Assisted Intervention - MICCAI 2024 - 27th International Conference, Marrakesh, Morocco, October 6-10, 2024, Proceedings, Part IV*, volume 15004 of *Lecture Notes in Computer Science*, 318–328. Springer.

Xiong, C.; Lin, Y.; Chen, H.; Zheng, H.; Wei, D.; Zheng, Y.; Sung, J. J. Y.; and King, I. 2024b. TAKT: Target-Aware Knowledge Transfer for Whole Slide Image Classification. In Linguraru, M. G.; Dou, Q.; Feragen, A.; Giannarou, S.; Glocker, B.; Lekadir, K.; and Schnabel, J. A., eds., *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, volume 15004, 503–513. Cham: Springer Nature Switzerland. ISBN 978-3-031-72082-6 978-3-031-72083-3.

Xiong, C.; Liu, J.; Chen, H.; Zheng, H.; Wu, X.; Zheng, Y.; Sung, J. J.; and King, I. 2024c. Enhancing Multimodal Survival Prediction with Pathology Reports in Hyperbolic Space.

Xu, P.; Zhu, X.; and Clifton, D. A. 2023. Multimodal Learning With Transformers: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10): 12113–12132.

Xu, Y.; and Chen, H. 2023. Multimodal Optimal Transport-Based Co-Attention Transformer with Global Structure Consistency for Survival Prediction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21241–21251.

Yu, D.; Zhang, X.; Chen, Y.; Liu, A.; Zhang, Y.; Yu, P. S.; and King, I. 2024a. Recent Advances of Multimodal Continual Learning: A Comprehensive Survey. arxiv:2410.05352.

Yu, J.; Zhuge, Y.; Zhang, L.; Hu, P.; Wang, D.; Lu, H.; and He, Y. 2024b. Boosting Continual Learning of Vision-Language Models via Mixture-of-Experts Adapters. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, 23219–23230. IEEE.

Zadeh, S. G.; and Schmid, M. 2020. Bias in Cross-Entropy-Based Training of Deep Survival Networks. *IEEE transactions on pattern analysis and machine intelligence*, 43(9): 3126–3137.

Zhang, H.; Meng, Y.; Zhao, Y.; Qiao, Y.; Yang, X.; Coupland, S. E.; and Zheng, Y. 2022. Dtfd-Mil: Double-tier Feature Distillation Multiple Instance Learning for Histopathology Whole Slide Image Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 18802–18812.

Zhou, F.; and Chen, H. 2023. Cross-Modal Translation and Alignment for Survival Analysis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 21485–21494.