Cross-cancer Expert-routing Knowledge Transfer in Federated Prognosis Prediction

Shu Wang[1,†], Junjian Li[1,†], and Jianxin Wang[1,*]

Hunan Provincial Key Lab on Bioinformatics, School of Computer Science and Engineering, Central South University, Changsha 410083, Hunan, China
jxwang@mail.csu.edu.cn

Abstract. Whole-Slide Image (WSI)-based prognosis prediction often struggles to scale to rare cancers due to limited patient cohorts, creating a critical generalization bottleneck. Furthermore, directly sharing knowledge across different cancers via multi-task learning frequently causes negative transfer due to significant prognostic heterogeneity. To address these challenges, we propose a novel Federated Expert-Routing Framework (FedERF) tailored for cross-cancer knowledge transfer. During pretraining, we utilize Federated LoRA across diverse source cancer clients to construct a shared cross-cancer expert pool. This extracts generalizable morphological priors from diverse multi-cancer cohorts, providing robust initialization for data-scarce clients. In the target-adaptive finetuning phase, we introduce a dynamic routing mechanism. Instead of applying a rigid global model, this mechanism adaptively selects and aggregates the most relevant experts from the pool based on the target domain's specific semantics. This routing strategy precisely injects complementary knowledge while isolating conflicting features, effectively mitigating negative transfer. Extensive experiments demonstrate that our FedERF successfully overcomes the data scarcity dilemma and achieves superior prognostic performance on target cancers. Code is available at https://github.com/ddayzzz/MICCAI2026-FedERF-public.

Keywords: Multiple Instance Learning · Whole Slide Images · Knowledge Transfer · Federated Learning

1 Introduction

Whole-Slide Image (WSI)-based survival analysis has emerged as a cornerstone of computational pathology, providing critical insights for personalized cancer treatment [3,20,8,10,11]. These WSIs serve as the clinical gold standard for cancer diagnosis. They encapsulate highly detailed visual information regarding tumor progression, encompassing critical features such as cellular morphology and tissue infiltration patterns. Consequently, accurate prognostic prediction is essential for guiding personalized therapeutic decisions, optimizing treatment plans, and ultimately improving patient survival rates [12,1].

* Corresponding author.
† These authors contributed equally to this work.

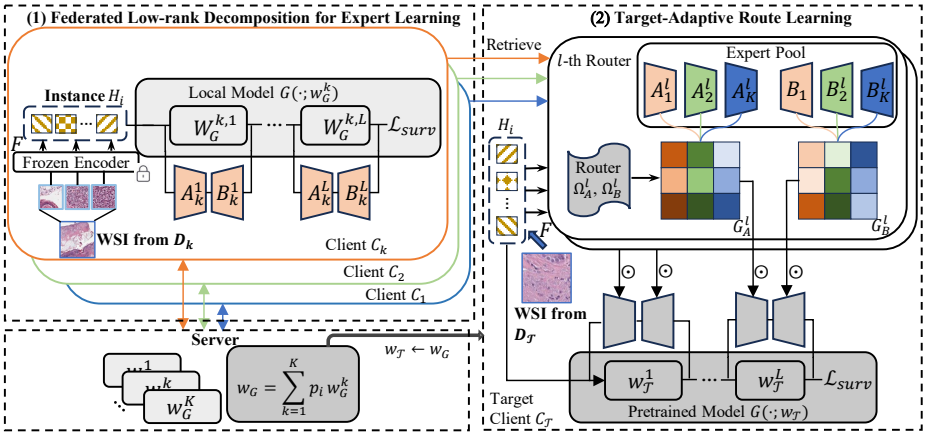


Fig. 1: Overview of FedERF. FedERF consists of two main phases: (1) Federated Low-rank Decomposition for Expert Learning, which constructs a diverse expert pool by collaboratively optimizing LoRA modules across multiple source cancers $\mathcal{S}$; and (2) Target-Adaptive Route Learning, which employs a dynamic routing mechanism to selectively recombine these experts for the target cancer $\mathcal{T}$, effectively mitigating negative transfer.

However, due to the extremely high resolution of WSIs, direct processing is computationally intractable. Consequently, recent works have widely adopted Multiple Instance Learning (MIL) frameworks [3,20,24,22,9]. In this paradigm, a WSI is divided into a bag of patches, and the model is optimized using only slide-level ground truth labels. Nevertheless, standard MIL methods exhibit inherent limitations that hinder their practical clinical application, specifically: **(1) Generalization Bottleneck.** Due to significant variations in data size and quality among different cancer types, directly transferring random initializations or off-the-shelf pre-trained weights to a target cancer fails to provide reliable performance guarantees. This issue is particularly severe in data-scarce scenarios, where weakly-supervised MIL models heavily rely on robust initial representations to discover meaningful patterns [17]. **(2) Fail to exploit robust prognostic knowledge from other cancers.** Existing methods struggle to fully exploit the rich and heterogeneous information embedded within the WSI patches from other cancers, which might contains underlying relevant cancer-oriented regions [15,19]. As a result, when training model based on target cancer type with limited data, they suffer from severe weak performance improvement. To address the two aforementioned issues, Federated Learning Multiple Instance Learning (FL-MIL) frameworks have emerged as a promising paradigm, enabling the construction of cross-cancer prognostic knowledge without centralizing data [17]. Recent developments in FL-MIL have significantly advanced WSI-based applications; for instance, Jin *et al.* [7] facilitated cross-model transfer via public samples, Lu *et al.* [17] enhanced security through noisy knowledge aggregation,

and Jiang *et al.* [5] improved generalization using local weight perturbation. However, existing FL-MIL methods encapsulate generalized knowledge into a single global model, thereby neglecting the learning of cancer-specific prognostic features during training. Consequently, this unified global representation fails to provide a robust parameter initialization for the target data-scarce cancer types.

Considering the lack of cancer-specific prognostic knowledge and negative transfer mitigation in conventional FL-MIL methods, we propose a novel Federated Expert-Routing Framework (FedERF), as shown in Fig. 1. Rather than relying on a rigid global model, FedERF provides a robust model parameter initialization for the target cancers by decomposing the learning process into two consecutive phases: **(1) Federated Low-rank Decomposition for Expert Learning:** We inject LoRA-based experts into the federated process to collaboratively build a cross-cancer expert pool, capturing distinct pathological and prognostic features. **(2) Target-Adaptive Route Learning:** To adapt the model for the target cancers, we introduce a layer-wise dynamic routing mechanism that selectively aggregates complementary expert knowledge based on target cancers. This adaptive approach effectively guides the model to leverage relevant prognostic knowledge while mitigating negative transfer from irrelevant cancer types, ensuring rapid convergence and robust performance. Extensive experiments validate that FedERF achieves superior prognostic performance.

2 Methodology

2.1 Preliminary

We formulate this setup as Federated Out-of-Distribution (OOD) Generalization [6,15]. Unlike conventional FL [18,14,23], decentralized source clients collaboratively learn generalized prognostic priors without sharing raw WSIs, aiming to transfer this knowledge to a completely unseen target cancer. Specifically, let $\mathcal{S} = \{C_1, \ldots, C_K\}$ be K source institutions and $C_\mathcal{T}$ be a target institution holding a cancer dataset $\mathcal{D}_\mathcal{T}$. We aim to transfer generalized prognostic knowledge from $\mathcal{S}$ to $C_\mathcal{T}$. Each source client $C_k \in \mathcal{S}$ holds a local dataset $\mathcal{D}_k = \{(X_i, t_i, \delta_i)\}_{i=1}^{N_k}$, where X_i, t_i, and δ_i represent the WSI, survival time, and censoring status, respectively. A feature extractor F encodes each X_i into patch embeddings $H_i \in \mathbb{R}^{m_i \times d}$, which are then processed by an L-layer MIL model G (parameterized by w_G) to yield the slide-level survival prediction.

During the Federated Low-rank Decomposition for Expert Learning phase (spanning T_p rounds), we attach cancer-specific expert parameters Θ_k to the local model of client C_k. The federated objective minimizes the empirical survival loss across all source clients:

$$\min_{w_G, \{\Theta_k\}_{k=1}^K} \frac{1}{N} \sum_{k=1}^{K} \sum_{i=1}^{N_k} \mathcal{L}_{surv}\big(G(F(X_i); w_G, \Theta_k), t_i, \delta_i\big) \tag{1}$$

where $N = \sum_{k=1}^{K} N_k$ is the total WSI number, and $\mathcal{L}_{surv}$ is the survival loss.

Specifically, w_G and $\{\Theta_k\}_{k=1}^K$ are jointly optimized for T_p rounds. The server aggregates the backbone via $w_G = \sum_{k=1}^K \frac{N_k}{N} w_G^k$ per round, while local experts are retained to form the cross-cancer expert pool.

In the Target-Adaptive Route Learning phase, the target client $C_\mathcal{T}$ initializes its model $w_\mathcal{T}$ with w_G and retrieves the expert pool. To avoid negative transfer caused by prognostic heterogeneity irrelevant source cancers, we introduce a dynamic routing policy Ω to selectively aggregate relevant morphological priors. To further align the model with the target domain's specific characteristics, it is refined over T_f epochs by optimizing:

$$\min_{\Phi_\mathcal{T} \in \mathbf{\Psi}} \left\{ \frac{1}{N_\mathcal{T}} \sum_{i=1}^{N_\mathcal{T}} \mathcal{L}_{surv}\big(G(F(X_i); \Phi_\mathcal{T}), t_i, \delta_i\big) \right\} \tag{2}$$

where $N_\mathcal{T}$ is the target dataset size. The target parameters $\Phi_\mathcal{T}$ are optimized within the routing space $\mathbf{\Psi}$ defined by $\{w_\mathcal{T}, \Omega, \{\Theta_k\}_{k=1}^K\}$.

2.2 Federated Low-rank Decomposition for Expert Learning

Upon completion of the federated training, the global model parameters w_G are distributed to the set of source clients $\mathcal{S}$. To enhance inter-cancer collaboration for accurate prognosis prediction while capturing cancer-specific knowledge, we adopt a layer-wise Low-rank Decomposition for Expert Learning strategy by employing Low-Rank Adaptation (LoRA) [2] to construct the expert modules. By introducing only a minimal set of trainable parameters, LoRA allows the model to effectively learn and store the distinct characteristics of each cancer type. This lightweight design ensures high parameter efficiency while preserving the general prognostic knowledge in the global model.

Specifically, for a source client $C_k \in \mathcal{S}$, we inject a total of L distinct LoRA expert modules into specific layers of the network (e.g., linear and convolutional layers) , while the global backbone w_G subsequently serves as the initialization for the second phase. This design is motivated by the insight that deep neural networks encode distinct and transferable cancer knowledge at semantic levels. By distributing these experts across multiple layers, the model captures cancer-specific nuances, providing relevant prognostic and feature knowledge from source clients for the subsequent fine-tuning phase of the target client initialized with w_G.

Formally, let $w_G^{k,l} \in \mathbb{R}^{O^l \times I^l}$ denote the weight matrix on client C_k of the l-th layer ($l \in \{1, \ldots, L\}$), which encodes the input $H_i^{l-1} \in \mathbb{R}^{I^l \times m}$ from the $(l-1)$-th layer, where O^l, I^l, and m denote the output dimension, input dimension, and bag sequence length, respectively. For notational simplicity, we omit the client index k in the subsequent formulations. We introduce a corresponding rank-decomposition expert parameterized by $\Theta_k^l = \{A_k^l, B_k^l\}$. Taking a bias-free linear layer as an example, the encoded cancer-specific representation at the l-th layer is formulated as follows:

$$H_i^l = \left(w_G^l + \frac{\alpha}{R^l} B_k^l A_k^l\right) H_i^{l-1} \tag{3}$$

where $A_k^l \in \mathbb{R}^{R^l \times I^l}$ and $B_k^l \in \mathbb{R}^{O^l \times R^l}$ capture the layer-specific adaptation.

Here, the rank R^l serves as an information bottleneck that determines the capacity of the expert module to absorb cancer-specific knowledge. The rank R^l depends on the layer's input and output dimensions, and is set to $R^l = \gamma \min(O^l, I^l)$. Finally, each source client $C_k \in \mathcal{S}$ maintains a set of cancer-specific experts, denoted as $\Theta_k = \{\Theta_k^l\}_{l=1}^L$.

2.3 Target-Adaptive Route Learning

In the route learning phase, our objective is to perform target-adaptive fine-tuning for a target cancer type $\mathcal{T}$. Through the federated collaborative expert learning in Section 2.2, the model backbone parameters $w_\mathcal{T}$ are initialized based on w_G, thus acquiring generalized prognostic classification knowledge. However, as indicated in [15,19], irrelevant cancers inevitably impose negative transfer on the target cancer, and the general global model is ill-adapted to the target cancer types. Accompanied by a comprehensive repository of source cancer-specific experts $\{\Theta_k\}_{k=1}^K$, it is imperative to establish an adaptive routing mechanism that guides the target cancers to selectively focus on the most relevant source cancer types while rigorously rejecting irrelevant or conflicting knowledge.

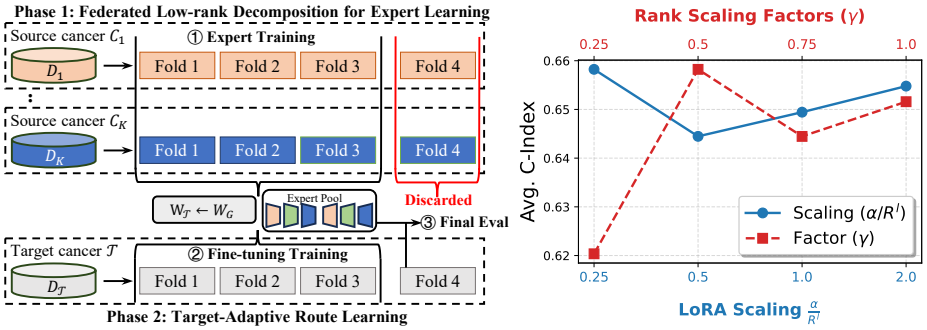
To this end, recognizing that semantic abstractions and morphological features vary significantly not only across different model layers but also across distinct local regions within a WSI, we propose a fine-grained Layer-wise Dynamic Routing Mechanism. Instead of utilizing a single global router for the entire slide, we construct independent routing modules for each layer l to guide relevant expert knowledge at the patch level. We introduce two sets of learnable routing parameter matrices: $\Omega_A^l, \Omega_B^l \in \mathbb{R}^{I^l \times K}$. Given the outputs from the previous layer H_i^{l-1}, the routers in l-th layer first compute the patch-level gating probability matrices $G_A^l, G_B^l \in \mathbb{R}^{K \times m}$ over the expert pool:

$$G_A^l = \text{Softmax}\left((\Omega_A^l)^\top H_i^{l-1}\right), \quad G_B^l = \text{Softmax}\left((\Omega_B^l)^\top H_i^{l-1}\right) \tag{4}$$

where the k-th row of the gating matrix represents the activation probability of the k-th expert across all m patches. The Softmax is applied along the expert dimension (K) to ensure the routing probabilities for each patch sum to 1.

To ensure computational efficiency, we formulate the decoupled expert routing directly in the activation space. Specifically, the intermediate low-rank projection is obtained by aggregating the features processed by all A experts, weighted by their routing distributions. Subsequently, the complementary expert matrices B project these features back to the output dimension. Combined with the general global model $w_\mathcal{T}^l$, the final enhanced representation $H_i^l \in \mathbb{R}^{O^l \times m}$ is computed directly as:

$$H_i^l = w_\mathcal{T}^l H_i^{l-1} + \frac{\alpha}{R^l} \sum_{k=1}^{K} g_{B,k}^l \odot \left(B_k^l \sum_{j=1}^{K} g_{A,j}^l \odot (A_j^l H_i^{l-1}) \right) \tag{5}$$



(a) (b)

Fig. 2: (a) Illustration of the evaluation protocol (taking fold=4 as an example) (b) Sensitivity analysis of γ and $\frac{\alpha}{R^l}$ with ABMIL backbone. Averaged C-Index values are reported.

where $g^l_{A,j}, g^l_{B,k} \in \mathbb{R}^{1\times m}$ denote the row vectors of the routing matrices G^l_A and G^l_B, and $\odot$ represents the element-wise Hadamard product with broadcasting along the channel dimension.

By jointly optimizing the decoupled routing parameters $\Omega = \{\Omega^l_A, \Omega^l_B\}^L_{l=1}$ alongside the target cancer data, this mechanism explicitly filters out conflicting features. By ensuring that every single patch obtains the optimal expert knowledge, this patch-level routing endows the model with the representational plasticity to exploit cross-cancer priors for the target cancers.

3 Experiments

3.1 Datasets and Implementation Details

Datasets and Baselines. We evaluate FedERF on five TCGA cohorts (LIHC, SARC, STAD, SKCM, and UCEC) using three representative MIL backbones: ABMIL [3], TransMIL [20], and WiKG [8]. We adopt a leave-one-out cross-cancer transfer setting: four cohorts act as source domains ($\mathcal{S}$) to collaboratively construct the expert pool, while the remaining one serves as the data-scarce target domain ($\mathcal{T}$). Baselines include three paradigms: (1) *Scratch*, where the target model parameters $w_\mathcal{T}$ are trained entirely from scratch using only $D_\mathcal{T}$; (2) *Standard FL* (FedAvg [18], FedProx [14], MOON [13], FedHisto [17]), which collaboratively learns w_G on $\mathcal{S}$ to initialize the target model ($w_\mathcal{T} \leftarrow w_G$); and (3) *MoE-based Transfer* (ROUPKT [15]), which constructs feature experts from $\mathcal{S}$ and trains a slide-level routing mechanism to optimize $w_\mathcal{T}$ during target adaptation. For a fair comparison, we run T_p additional epochs for ROUPKT and FedERF prior to the $\mathcal{S} \rightarrow \mathcal{T}$ transfer.

Evaluation Protocol and Implementation. WSIs are preprocessed into 448×448 patches at $20\times$ magnification via CONCH [16]. We employ a rigorous 4-fold cross-validation. To explicitly prevent data leakage, source test folds

Table 1: Performance comparison across different models and cancer datasets.

G	$\mathcal{T}$	Scratch	FedAvg	FedHisto	MOON	FedProx	ROUPKT	FedERF(Ours)
ABMIL	LIHC	**0.685**$_{0.058}$	0.669$_{0.071}$	0.671$_{0.069}$	0.675$_{0.071}$	0.669$_{0.073}$	0.673$_{0.072}$	<u>0.684</u>$_{0.069}$
	SARC	0.563$_{0.051}$	0.575$_{0.059}$	0.564$_{0.074}$	0.570$_{0.070}$	0.573$_{0.053}$	<u>0.579</u>$_{0.065}$	**0.636**$_{0.042}$
	SKCM	0.560$_{0.021}$	0.571$_{0.030}$	0.567$_{0.028}$	0.570$_{0.049}$	0.573$_{0.031}$	<u>0.577</u>$_{0.025}$	**0.598**$_{0.032}$
	STAD	0.553$_{0.029}$	0.540$_{0.040}$	0.535$_{0.030}$	0.543$_{0.032}$	0.536$_{0.039}$	<u>0.588</u>$_{0.039}$	**0.611**$_{0.031}$
	UCEC	0.675$_{0.041}$	0.731$_{0.014}$	0.731$_{0.011}$	0.735$_{0.016}$	0.730$_{0.012}$	<u>0.748</u>$_{0.033}$	**0.761**$_{0.043}$
	Avg.	0.607$_{0.060}$	0.617$_{0.071}$	0.614$_{0.075}$	0.619$_{0.074}$	0.616$_{0.072}$	<u>0.633</u>$_{0.068}$	**0.658**$_{0.059}$
TransMIL	LIHC	<u>0.662</u>$_{0.041}$	0.595$_{0.070}$	0.599$_{0.038}$	0.627$_{0.052}$	0.576$_{0.031}$	0.634$_{0.070}$	**0.683**$_{0.055}$
	SARC	0.558$_{0.084}$	0.569$_{0.089}$	0.524$_{0.083}$	<u>0.570</u>$_{0.047}$	0.548$_{0.082}$	0.481$_{0.032}$	**0.645**$_{0.026}$
	SKCM	<u>0.606</u>$_{0.010}$	0.556$_{0.045}$	0.570$_{0.039}$	0.532$_{0.050}$	0.580$_{0.018}$	0.553$_{0.052}$	**0.617**$_{0.037}$
	STAD	0.579$_{0.054}$	0.532$_{0.039}$	0.578$_{0.036}$	<u>0.589</u>$_{0.029}$	0.568$_{0.043}$	0.581$_{0.022}$	**0.600**$_{0.031}$
	UCEC	0.683$_{0.079}$	0.681$_{0.057}$	0.651$_{0.030}$	0.693$_{0.050}$	0.680$_{0.035}$	<u>0.710</u>$_{0.008}$	**0.726**$_{0.036}$
	Avg.	<u>0.618</u>$_{0.048}$	0.587$_{0.051}$	0.584$_{0.041}$	0.602$_{0.055}$	0.590$_{0.046}$	0.592$_{0.077}$	**0.654**$_{0.046}$
WiKG	LIHC	<u>0.699</u>$_{0.060}$	0.637$_{0.069}$	0.644$_{0.059}$	0.621$_{0.066}$	0.623$_{0.067}$	0.635$_{0.105}$	**0.715**$_{0.047}$
	SARC	0.567$_{0.070}$	0.549$_{0.059}$	<u>0.627</u>$_{0.079}$	0.598$_{0.067}$	0.626$_{0.043}$	0.612$_{0.045}$	**0.664**$_{0.031}$
	SKCM	0.564$_{0.052}$	0.579$_{0.029}$	0.545$_{0.035}$	<u>0.583</u>$_{0.029}$	0.570$_{0.024}$	0.543$_{0.028}$	**0.632**$_{0.027}$
	STAD	0.553$_{0.012}$	0.570$_{0.029}$	0.538$_{0.032}$	0.516$_{0.031}$	0.534$_{0.032}$	<u>0.581</u>$_{0.035}$	**0.617**$_{0.021}$
	UCEC	0.625$_{0.045}$	0.699$_{0.044}$	0.711$_{0.032}$	0.697$_{0.022}$	0.659$_{0.032}$	**0.719**$_{0.048}$	<u>0.714</u>$_{0.046}$
	Avg.	0.602$_{0.055}$	0.607$_{0.055}$	0.613$_{0.065}$	0.603$_{0.059}$	0.602$_{0.044}$	<u>0.618</u>$_{0.059}$	**0.668**$_{0.041}$

are completely discarded. As illustrated in Fig. 2a, the evaluation proceeds in three steps: ① Pre-training the expert pool using only the training folds of $\mathcal{S}$ ($T_p = 30$ rounds, AdamW, LR=2×10^{-4}); ② Fine-tuning on the training folds of $\mathcal{T}$ ($T_f = 20$ epochs, LR decayed to 1×10^{-4}); and ③ Measuring the Concordance Index (C-Index) on the held-out test fold of $\mathcal{D}_{\mathcal{T}}$. For all experiments, the batch size is 1, the expert rank scaling factor γ is 0.5, and the LoRA scaling $\frac{\alpha}{R^l}$ is bounded to 0.25.

3.2 Results and Discussion

Comparison with State-of-the-Art Methods. Table 1 reports the prognostic performance on five data-scarce target cancers. While standard FL-MIL methods generally improve upon training from scratch, their monolithic global models often suffer from severe negative transfer due to conflicting pan-cancer priors (e.g., performing worse than random initialization on the LIHC cohort). FedERF effectively overcomes this limitation. By dynamically routing decoupled structural priors, our framework explicitly isolates conflicting survival signals. Consequently, FedERF demonstrates highly competitive and generally superior performance compared to traditional FL-MIL baselines and the MoE-based ROUPKT [4,21,15], achieving the highest average C-Index across all three backbones: 0.658 (ABMIL), 0.654 (TransMIL), and 0.668 (WiKG). This establishes a highly robust and superior paradigm for cross-cancer knowledge transfer.

Few-shot Results on Target Cancers. To evaluate data efficiency, we conduct $\{8, 16, 32\}$-shot experiments on the target cohorts. As shown in Fig. 3, while most baselines suffer severe degradation under 8 shots, FedERF maintains strong robustness. Remarkably, at just 16 shots, our approach achieves highly competitive results that even rival fully supervised baselines. This confirms that

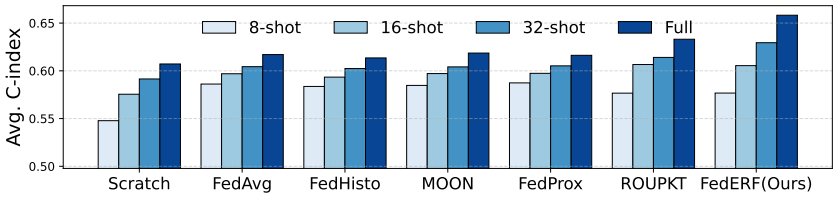


Fig. 3: Performance comparison across different few-shot settings on ABMIL.

Table 2: Ablation study on the components of FedERF.

G	Components	$\mathcal{T}$					
		LIHC	SARC	SKCM	STAD	UCEC	Avg.
ABMIL	w/o Exp	$\mathbf{0.685_{0.058}}$	$0.563_{0.051}$	$0.560_{0.021}$	$0.553_{0.029}$	$0.675_{0.041}$	$0.607_{0.060}$
	AverageExp	$0.668_{0.069}$	$0.624_{0.017}$	$0.582_{0.054}$	$\mathbf{0.619_{0.047}}$	$0.745_{0.022}$	$0.648_{0.056}$
	RandomExp	$0.678_{0.070}$	$0.623_{0.030}$	$0.577_{0.050}$	$0.616_{0.028}$	$0.736_{0.040}$	$0.646_{0.055}$
	FedERF(Ours)	$0.684_{0.069}$	$\mathbf{0.636_{0.042}}$	$\mathbf{0.598_{0.032}}$	$0.611_{0.031}$	$\mathbf{0.761_{0.043}}$	$\mathbf{0.658_{0.059}}$
WiKG	w/o Exp	$0.699_{0.060}$	$0.567_{0.070}$	$0.564_{0.052}$	$0.553_{0.012}$	$0.625_{0.045}$	$0.602_{0.055}$
	AverageExp	$0.654_{0.033}$	$0.623_{0.066}$	$0.578_{0.038}$	$0.599_{0.036}$	$0.666_{0.047}$	$0.624_{0.033}$
	RandomExp	$0.652_{0.054}$	$0.571_{0.022}$	$\mathbf{0.644_{0.014}}$	$0.590_{0.050}$	$0.714_{0.035}$	$0.634_{0.051}$
	FedERF(Ours)	$\mathbf{0.715_{0.047}}$	$\mathbf{0.664_{0.031}}$	$0.632_{0.027}$	$\mathbf{0.617_{0.021}}$	$\mathbf{0.714_{0.012}}$	$\mathbf{0.668_{0.041}}$

our dynamic expert routing effectively mitigates negative transfer and overcomes the generalization bottleneck in extreme data-scarce scenarios.

Ablation Study. We evaluate the specific contributions of the federated expert pool and the dynamic routing mechanism by comparing FedERF against three variants: 1) w/o Exp (removes all experts, relying solely on target data); 2) AverageExp (naively averages source experts, disabling routing); and 3) RandomExp (uses randomly initialized experts, skipping federated pre-training).

As shown in Table 2, FedERF achieves the highest average C-Index across both backbones. Notably, AverageExp yields sub-optimal results, indicating that indiscriminately aggregating pan-cancer experts reintroduces conflicting signals and negative transfer. Similarly, the performance drop in RandomExp highlights the indispensable value of the morphological priors extracted during pre-training. Overall, constructing a cross-cancer expert pool and adaptively routing its knowledge are both essential for robust survival prediction.

Hyperparameter Sensitivity. We analyze two key hyperparameters in FedERF: the expert rank scaling factor $\gamma \in \{0.25, 0.5, 0.75, 1.0\}$ and the LoRA knowledge integration scaling $\frac{\alpha}{R^l} \in \{0.25, 0.5, 1.0, 2.0\}$ (evaluated with fixed $\gamma = 0.5$). As shown in Fig. 2b, FedERF is sensitive to the rank scaling factor γ, suggesting that the bottleneck capacity heavily influences the experts' representational power. Conversely, it remains relatively robust to the integration scaling $\frac{\alpha}{R^l}$. Since $\frac{\alpha}{R^l}$ primarily weights the injected features, this stability indicates our dynamic routing effectively balances cross-cancer priors to mitigate negative transfer, alleviating the burden of extensive hyperparameter tuning.

4 Conclusion

In this paper, we present FedERF, a novel Federated Expert-Routing Framework designed to overcome the generalization bottleneck in cancer prognosis. By decoupling federated optimization into global low-rank expert learning and local target-adaptive routing, FedERF constructs a cross-cancer knowledge pool and dynamically recombines relevant structural priors. Extensive experiments validate that FedERF explicitly isolates conflicting survival signals and mitigates negative transfer. Compared to FL-MIL and MoE-based baselines, it achieves superior knowledge transfer efficiency and predictive robustness, providing a highly reliable paradigm for pan-cancer histopathological analysis.

Acknowledgments. This work was supported in part by Xinjiang Uygur Autonomous Region Key R & D program under Grant 2024B03039-3; the National Natural Science Foundation of China (No. U25A20449). This work was also carried out in part using computing resources at the High Performance Computing Center of Central South University.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Arzikulov, F., Komiljonov, A., et al.: The role of artificial intelligence in personalized oncology: predictive models and treatment optimization. Academic Journal of Science, Technology and Education **1**(6), 24–33 (2025)
2. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
3. Ilse, M., Tomczak, J., Welling, M.: Attention-based deep multiple instance learning. In: International conference on machine learning. pp. 2127–2136 (2018)
4. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. Neural computation **3**(1), 79–87 (1991)
5. Jiang, M., Wang, Z., Dou, Q.: Harmofl: Harmonizing local and global drifts in federated learning on heterogeneous medical images. In: Proceedings of the AAAI conference on artificial intelligence. vol. 36, pp. 1087–1095 (2022)
6. Jiang, M., Yang, H., Cheng, C., Dou, Q.: Iop-fl: Inside-outside personalization for federated medical image segmentation. IEEE Transactions on Medical Imaging **42**(7), 2106–2117 (2023)
7. Jin, H., Liu, S., Cong, C., Feng, Q., Liu, Y., Huang, L., Hu, Y.: Fedwsidd: Federated whole slide image classification via dataset distillation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 178–188. Springer (2025)
8. Li, J., Chen, Y., Chu, H., Sun, Q., Guan, T., Han, A., He, Y.: Dynamic graph representation with knowledge-aware attention for histopathology whole slide image analysis. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11323–11332 (2024)
9. Li, J., Kuang, H., Liu, J., Yue, H., Wang, J.: Ca2cl: Cluster-aware adversarial contrastive learning for pathological image analysis. IEEE Journal of Biomedical and Health Informatics pp. 5095–5108 (2025)

10. Li, J., Kuang, H., Liu, J., Yue, H., Wang, J.: Domain-specific self-supervised contrastive learning with contrast-aware pair refinement for pathological image analysis. Big Data Mining and Analytics (2025)
11. Li, J., Liu, J., Kuang, H., Yue, H., He, M., Wang, J.: Mico: Multiple instance learning with context-aware clustering for whole slide image analysis. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 376–385. Springer (2025)
12. Li, J., Liu, J., Yue, H., Cheng, J., Kuang, H., Bai, H., Wang, Y., Wang, J.: Darc: Deep adaptive regularized clustering for histopathological image classification. Medical image analysis **80**, 102521 (2022)
13. Li, Q., He, B., Song, D.: Model-contrastive federated learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10713–10722 (2021)
14. Li, T., Sahu, A.K., Zaheer, M., Sanjabi, M., Talwalkar, A., Smith, V.: Federated optimization in heterogeneous networks. Proceedings of Machine learning and systems **2**, 429–450 (2020)
15. Liu, P., Ji, L., Gou, J., Zeng, X.: Cross-cancer knowledge transfer in wsi-based prognosis prediction. arXiv preprint arXiv:2508.13482 (2025)
16. Lu, M.Y., Chen, B., Williamson, D.F., Chen, R.J., Liang, I., Ding, T., Jaume, G., Odintsov, I., Le, L.P., Gerber, G., et al.: A visual-language foundation model for computational pathology. Nature medicine **30**(3), 863–874 (2024)
17. Lu, M.Y., Chen, R.J., Kong, D., Lipkova, J., Singh, R., Williamson, D.F., Chen, T.Y., Mahmood, F.: Federated learning for computational pathology on gigapixel whole slide images. Medical image analysis **76**, 102298 (2022)
18. McMahan, B., Moore, E., Ramage, D., Hampson, S., y Arcas, B.A.: Communication-efficient learning of deep networks from decentralized data. In: Artificial intelligence and statistics. pp. 1273–1282 (2017)
19. Shao, D., Chen, R.J., Song, A.H., Runevic, J., Lu, M.Y., Ding, T., , Mahmood, F.: Do multiple instance learning models transfer? In: International conference on machine learning (2025)
20. Shao, Z., Bian, H., Chen, Y., Wang, Y., Zhang, J., Ji, X., et al.: Transmil: Transformer based correlated multiple instance learning for whole slide image classification. Advances in neural information processing systems **34**, 2136–2147 (2021)
21. Shazeer, N., Mirhoseini, A., Maziarz, K., Davis, A., Le, Q., Hinton, G., Dean, J.: Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In: International Conference on Learning Representations (2017)
22. Sun, S., van Midden, D., Litjens, G., Baumgartner, C.F.: Prototype-based multiple instance learning for gigapixel whole slide image classification. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 507–517. Springer (2025)
23. Wang, S., Qu, Z., Liu, Y., Kan, S., Liang, Y., Wang, J.: Fedmmr: Multi-modal federated learning via missing modality reconstruction. In: 2024 IEEE International Conference on Multimedia and Expo (ICME). pp. 1–6. IEEE (2024)
24. Xiang, J., Zhang, J.: Exploring low-rank property in multiple instance learning for whole slide image classification. In: International Conference on Learning Representations (2023)