Curriculum Hierarchical Knowledge Distillation for Bias-Free Survival Prediction

Chaozhuo Li[1] , **Zhihao Tang**[1] , **Mingji Zhang**[2] , **Zhiquan Liu**[3] , **Litian Zhang**[1] , **Xi Zhang**[1*]
[1]Key Laboratory of Trustworthy Distributed Computing and Service (MoE), Beijing University of Posts and Telecommunications, Beijing 100876, China
[2]Clinical Oncology School of Fujian Medical University, Fujian Cancer Hospital, Fuzhou 350014, China
[3]Jinan University, Guangzhou 510632, China
{lichaozhuo, innerone, zhangx}@bupt.edu.cn, zhangmj@fjzlhospital.com, zqliu@vip.qq.com, litianzhang@buaa.edu.cn

Abstract

Survival prediction is a pivotal task for estimating mortality risk within a given timeframe based on whole slide images (WSIs). Conventional models typically assume that WSIs across patients are independent and identically distributed, an assumption that may not hold due to inherent variability in WSI preparation and the uncertain condition of infected tissues. These uncontrollable external factors introduce significant variability in the numbers and resolutions of WSIs across patients, leading to bias and compromised performance, particularly for tail patients with limited data. In this paper, we propose a novel approach, PathoKD, based on knowledge distillation. Recognizing the hierarchical nature of disease progression and the data scarcity issues associated with vanilla knowledge distillation methods, PathoKD integrates a novel curriculum learning framework with hierarchical knowledge distillation. This integration effectively mitigates the performance gap between head and tail patients, thereby enhancing prediction accuracy across patient groups. Our proposal is extensively evaluated over popular datasets and experimental results demonstrate its superiority.

1 Introduction

Survival prediction, the estimation of mortality risk within a given timeframe, represents a cornerstone of clinical oncology [Shedden *et al.*, 2008]. This endeavor predominantly entails the analysis of whole slide images (WSIs), which encapsulate intricate spatial patterns and the complexities of the tumor microenvironment [Pantanowitz *et al.*, 2011]. Deep learning has emerged as a transformative paradigm, automating WSI analysis and offering unprecedented prospects for alleviating pathologists' burdens and empowering physicians in critical decision-making [Hanna *et al.*, 2020].

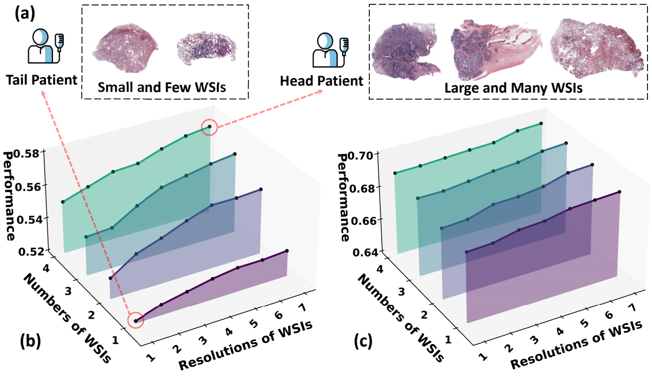

Figure 1: Illustration of the biases in the numbers and resolutions of WSIs across patients.

Existing WSI-based survival prediction models [Yao *et al.*, 2020a; Zhu *et al.*, 2017; Liu *et al.*, 2023; Shao *et al.*, 2023a; Tang *et al.*, 2019] follow a two-stage paradigm: WSI preparation, which encompasses collecting WSIs, segmenting them into patches, and encoding these patches to address gigapixel-level resolution; and patient-level prediction, wherein patient outcomes are forecasted by leveraging both the encoded patches and their inter-patch relationships. Most works [Shao *et al.*, 2023a; Liu *et al.*, 2023] focus on developing sophisticated prediction models for the patient-level prediction stage, such as Transformers [Chen *et al.*, 2022a], Graph Neural Networks [Gadiya *et al.*, 2020] and Capsule Networks [Tang *et al.*, 2019]. However, the substantial bias introduced during the WSI preparation stage is largely overlooked, which may limit the effectiveness of current approaches and their potential to serve as a panacea [Jaume *et al.*, 2021].

The efficacy of existing survival prediction models depends on the assumption that the data are independent and identically distributed (IID), implying that the WSIs of different patients are drawn from the same distribution. However, the inherent limitations in the WSI preparation process, coupled with the uncertain status of infected tissues and variability in pathologist expertise, often undermine the validity of the IID assumption in practical scenarios. For instance, biopsy strategies [Strassburg and Manns, 2006] and tissue sectioning practices [Fischer *et al.*, 2008] are influenced by the heuristic expertise of clinicians, leading to variation in the numbers of WSIs across patients. In the TCGA-LUAD dataset [Tomczak *et al.*, 2015], 20.1% of patients have more than 6 WSIs, while 30.3% have fewer than 3. This non-IID nature of WSI numbers may result in imbalanced training data and skewed per-

[*] Corresponding Author: Xi Zhang.

formance. Additionally, the variability in lesion morphology and slice positioning leads to significant differences in resolution across WSIs. As shown in Fig. 1(a), WSIs with higher resolution tend to be more informative (e.g., containing more tissue patches) than those with lower resolution.

To highlight the impact of the non-IID phenomenon on the survival prediction task, a preliminary study is conducted using the popular baseline DeepAttnMISL [Yao *et al.*, 2020a] on the NLST dataset [Team, 2011]. Patients within the testing set are divided into different groups based on two dimensions: (1) the average numbers of WSIs, as indicated by the vertical axis in Fig. 1(b), and (2) the average informative resolution of WSIs, as indicated by the horizontal axis. The average informative resolution is implemented as the average number of non-blank patches (500×500 pixels at a scaling factor of 20) for each patient's corresponding WSIs, calculated after applying the OTSU algorithm for filtering [Otsu, 1979]. Groups with larger labels correspond to a greater number of WSIs or patches. Fig. 1(b) demonstrates the performance of a well-trained model on different groups. One can clearly see that: (1) patients with more WSIs consistently outperform those with fewer WSIs; and (2) given the same number of WSIs, patients with WSIs of larger resolution (i.e., those with a larger number of patches) achieve better performance.

In this paper, we focus on the novel problem of mitigating bias within both the numbers and resolutions of WSIs across patients. A straightforward solution involves resampling techniques [Good, 2006], while undersampling may result in information loss, while oversampling can contribute to overfitting and noise [Zhang *et al.*, 2024b], particularly in the context of small survival prediction datasets [Zhang *et al.*, 2024a]. Inspired by the idea of transferring knowledge from head patients, who possess more informative WSIs, to tail patients with less informative WSIs, we adopt knowledge distillation [Hinton, 2015] as a core methodology. Specifically, a teacher model is trained on the head patients, and closeness minimization losses are employed to facilitate the transfer of knowledge [Zhao *et al.*, 2023].

However, directly applying vanilla knowledge distillation methods to survival prediction presents two challenges. (1) Hierarchical Structure. Survival prediction operates at the patient level [Shao *et al.*, 2023a], while biases in number and resolution are hierarchically structured at both the inter- and intra-WSI levels. Vanilla distillation models [Hinton, 2015] overlook the interdependencies across these levels, thus failing to capture the comprehensive hierarchical information. (2) Data Scarcity. Given the typically limited number of available patient cases, vanilla knowledge distillation methods often split datasets to train the teacher and student models, further exacerbating the scarcity challenge.

To address the aforementioned challenges, we propose a novel model, PathoKD, designed to enhance the robustness of survival prediction through hierarchical knowledge distillation within a triple curriculum framework. The learning process is organized into three curricula, each targeting progressively difficult stages determined by the number and resolution of each patient's WSIs. Specifically, patients are categorized into three groups: head patients, who exhibit a large number of extensive WSIs; virtual tail patients, which are synthesized by removing patches from the head patients to simulate fewer and smaller WSIs, bridging the gap between head and tail patients; and tail patients, who naturally possess smaller and fewer WSIs. Curriculum I initiates the process by training the teacher model to extract knowledge from head patients, employing a label-assistance strategy to mitigate data scarcity. Building upon this, Curriculum II facilitates hierarchical knowledge distillation across different levels, which transfers knowledge from the teacher model to the student models via virtual tail patients. To further alleviate the data scarcity and to leverage the full spectrum of available data, Curriculum III refines both the teacher and student models, incorporating retrieval and generative enhancement to integrate knowledge from tail patients. Extensive experiments across multiple datasets validate its superiority. The key contributions of this work are as follows:

- To the best of our knowledge, we are the first to examine the novel issue of bias stemming from the numbers and resolutions of WSIs across patients, and to investigate its impact on the task of survival prediction.
- A novel curriculum learning framework is proposed based on hierarchical knowledge distillation, which effectively bridges the gap between head and tail patients.
- The efficacy of our approach is demonstrated by showcasing its superior performance over existing methods across multiple real-world cancer datasets.

2 Problem Definition

Survival data refers to a set of patients' data, denoted as $\mathcal{P} = \{P_1, \ldots, P_{|\mathcal{P}|}\}$, where each patient's data P_i consists of a set of WSIs and a survival outcome label, $P_i = \{\mathcal{W}^i, Y_i\}$. Each WSI W_j^i represents a tissue sample from patient P_i, and the survival outcome label $Y_i = (t_i, c_i)$ includes the observation time t_i and survival status c_i. The binary status $c_i \in \{0, 1\}$ indicates whether t_i is a survival time ($c_i = 1$) or a right-censored time ($c_i = 0$). Censoring occurs when a patient is still alive at the end of the study or is lost to follow-up. For computational efficiency, each WSI is partitioned into fixed-size patches, $W_j^i = \{E_{j,1}^i, \ldots, E_{j,|W_j^i|}^i\}$, which represent the smallest unit in the hierarchical survival data structure.

WSI-based survival prediction forecasts a patient's survival outcome given a set of WSIs $\mathcal{W}^i$. For a patient P_i with a survival label comprising observed time t_i and status c_i, the deep survival model $f(\mathcal{W}^i)$ predicts a survival time $\hat{t}_i \in \mathcal{T}$:

$$\arg\max_{\hat{t}_i \in \mathcal{T}} \Pr\left(\hat{t}_i = t_i \mid \mathcal{W}^i\right), \tag{1}$$

3 Methodology

Our method is organized into four parts. We begin by quantifying biases in survival-prediction performance across patients with differing WSI coverage and quality. The next three sections describe the *PathoKD* framework (Fig. 2), structured as an easy-to-hard curriculum: (1) teacher training on head patients (green solid box), (2) student training on virtual tail patients (blue dotted box), and (3) progressive integration of real tail patients into both models (red dotted box).

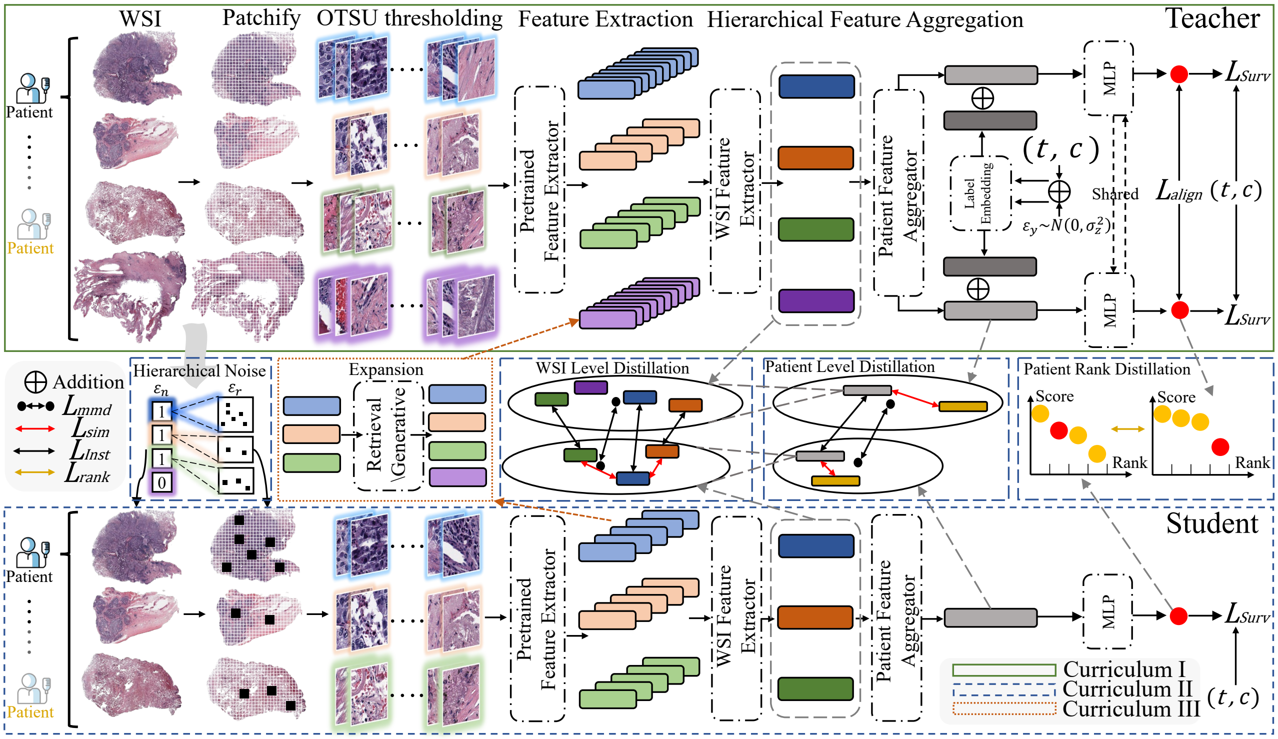

Figure 2: Overview of the PathoKD Framework.

3.1 Analysis of the Biases across Patients

In an ideal scenario, all patients would possess an equal number of uniformly-sized WSIs. However, real-world variability—arising from biopsy technique, scanner settings, and QC—introduces two noise terms: resolution noise $\epsilon_r^{i,j}$ (resolution fluctuations for the j-th WSI of patient i) and per-patient noise ϵ_n^i (variation in total WSI count for patient i). Modeling these terms allows us to quantify biases in WSI coverage and quality across the cohort.

To mitigate the influence of $\epsilon_r^{i,j}$ and ϵ_n^i, we assume that patients with large and more WSIs are less susceptible to noise, treating them as an ideal reference set. We aim to construct artificial noise $\hat{\epsilon}_\mathrm{n}^i$ and $\hat{\epsilon}_\mathrm{r}^{i,j}$, enabling a model to learn to predict outcomes $\hat{t}_i$ consistently, with or without noise:

$$\arg\max_{\hat{t}_i \in \mathcal{T}} \Pr(\hat{t}_i = t_i \mid \mathcal{W}^i) \approx \Pr(\hat{t}_i = t_i \,|\mathcal{W}^i + \hat{\epsilon}_\mathrm{n}^i + \sum_{j=1}^{|\mathcal{W}^i|} \hat{\epsilon}_\mathrm{r}^{i,j}), \tag{2}$$

This objective can be expressed as the minimization of the prediction discrepancy between the clean and noisy sets:

$$\min_f \mathbb{E}_{\mathcal{W}^i} \left[\left\| f(\mathcal{W}^i) - f(\mathcal{W}^i + \mathcal{A}(\epsilon_\mathrm{n}^i) + \sum_{j=1}^{|\mathcal{W}^i|} \mathcal{A}(\epsilon_\mathrm{r}^{i,j})) \right\| \right]. \tag{3}$$

By enforcing alignment between predictions in both settings, this framework results in more reliable survival predictions that remain invariant to noise artifacts.

3.2 Curriculum I: Teacher Model Training on Head Patient

Given a training set of patient survival data $\mathcal{P}_\text{train}$, head and tail patients, $\mathcal{P}_\text{head}$ and $\mathcal{P}_\text{tail}$, are identified based on average resolution and WSI counts via pre-defined thresholds. For each head patient P_i^head, each WSI W_j^i is segmented into patches $W_j^i = \{E_{j,1}^i, \ldots, E_{j,|W_j^i|}^i\}$, which are mapped to a representation space using a pre-trained histopathological encoder (e.g., HIPT [Chen *et al.*, 2022a]). The representation of the k-th patch of the j-th WSI for the i-th head patient is denoted $X_{i,j,k}^\text{head}$. WSI-level features are derived by modeling patch relationships within each WSI, which are then aggregated into patient-level features. These features are passed through a Multilayer Perceptron (MLP) to predict the survival score. The process is formulated as:

$$\begin{aligned}
\mathbf{S}_{i,j}^\text{head} &= \mathcal{A}_e\left(\mathbf{X}_{i,j,1}^\text{head}, ..., \mathbf{X}_{i,j,k}^\text{head}, ..., \mathbf{X}_{i,j,|W_j^i|}^\text{head}\right), \\
\mathbf{D}_i^\text{head} &= \mathcal{A}_w\left(\mathbf{S}_{i,1}^\text{head}, ..., \mathbf{S}_{i,j}^\text{head}, ..., \mathbf{S}_{i,|\mathcal{W}^i|}^\text{head}\right), \\
\mathbf{O}_i^\text{head} &= f_\text{MLP}\left(\mathbf{D}_i^\text{head}\right).
\end{aligned} \tag{4}$$

The feature $\mathbf{X}_{i,j,1}^\text{head}$ is input into an inter-WSI aggregator $\mathcal{A}_e$, yielding $\mathbf{S}_{i,j}^\text{head}$, the representation of the j-th WSI for the i-th head patient. $\mathbf{D}_i^\text{head}$ denotes the representation of the i-th head patient. The predicted survival score $\mathbf{O}_i^\text{head}$ is obtained by passing $\mathbf{D}_i^\text{head}$ through an MLP, f_MLP.

Teacher Training with Noisy Survival Labels. The limited number of head patients may hinder teacher model training. Therefore, the survival label is incorporated to provide additional knowledge. Unlike traditional training, where test set labels are unavailable, our teacher model uses survival labels to guide the student model's learning rather than predicting them. The survival label is encoded into an embedding to incorporate it as input for training the teacher model. Given that the survival label $Y_i^{\text{head}} = (t_i^{\text{head}}, c_i^{\text{head}})$ consists of continuous (t_i^{head}) and discrete (c_i^{head}) components, which differ in dimensionality from the pathological data, two distinct embedding functions are applied. The continuous observed time t_i^{head} is normalized and embedded into a learnable vector, while the discrete binary censored status $c_i^{\text{head}} \in \{0, 1\}$ is embedded using a learned binary vector:

$$\begin{aligned}&\mathbf{E}_t^{\text{head}} = \text{Embedding}(t_i^{\text{head}}), \mathbf{E}_c^{\text{head}} = \text{Embedding}(c_i^{\text{head}}),\\&\mathbf{E}_Y^{\text{head}} = [\mathbf{E}_t^{\text{head}}, \mathbf{E}_c^{\text{head}}].\end{aligned} \tag{5}$$

where $\mathbf{E}_Y^{\text{head}}$ represents the encoded survival label. Directly incorporating the label with the WSI may lead to overreliance, diminishing generalizability. To mitigate this, Gaussian noise is introduced into the label embedding, resulting in a noisy embedding: $\hat{\mathbf{E}}_Y^{\text{head}} = \mathbf{E}_Y^{\text{head}} + \epsilon_{\text{y}}$, where $\epsilon_{\text{y}} \sim \mathcal{N}(0, \sigma^2)$. The noisy embedding is added to the patient feature representation: $\hat{\mathbf{D}}_i^{\text{head}} = \mathbf{D}_i^{\text{head}} + \hat{\mathbf{E}}_Y^{\text{head}}$.

Contrastive learning is employed to align augmented versions of the same patient's data, thereby reducing the impact of noise. The alignment loss $\mathcal{L}_{\text{align}}$ is defined as:

$$\mathcal{L}_{\text{align}} = f_{\text{dist}}(\mathbf{O}_{\text{i,1}}^{\text{head}}, \mathbf{O}_{\text{i,2}}^{\text{head}}) = f_{\text{dist}}(f_{\text{MLP}}(\hat{\mathbf{D}}_{\text{i,1}}^{\text{head}}), f_{\text{MLP}}(\hat{\mathbf{D}}_{\text{i,2}}^{\text{head}})).$$

where $\hat{\mathbf{D}}_{\text{i,1}}^{\text{head}}$ and $\hat{\mathbf{D}}_{\text{i,2}}^{\text{head}}$ are two versions of the patient representation, each incorporating different noisy label embeddings. $\hat{\mathbf{O}}_{\text{i,1}}^{\text{head}}$ and $\hat{\mathbf{O}}_{\text{i,2}}^{\text{head}}$ are the predicted survival outcomes, and f_{dist} measures the dissimilarity between them.

The teacher model loss combines survival prediction losses and the alignment loss:

$$\mathcal{L}_{\text{Teacher}} = \mathcal{L}_{\text{sur}}(\hat{\mathbf{O}}_{\text{i,1}}^{\text{head}}, Y) + \mathcal{L}_{\text{sur}}(\hat{\mathbf{O}}_{\text{i,2}}^{\text{head}}, Y) + \lambda \mathcal{L}_{\text{align}}(\hat{\mathbf{O}}_{\text{i,1}}^{\text{head}}, \hat{\mathbf{O}}_{\text{i,2}}^{\text{head}}).$$

where $\mathcal{L}_{\text{sur}}$ is the survival loss and λ balances alignment. The Cox loss [Zhu *et al.*, 2016] is used to optimize hazard risk predictions, assigning higher risks to patients with shorter times:

$$\mathcal{L}_{\text{sur}}(O_i, Y) = \sum_{i:c_i=1} \left(-O_i + \log \sum_{j:t_j \geq t_i} \exp(O_j) \right) \tag{6}$$

where O_i denotes the hazard risk for patient p_i. This function encourages higher hazard risk predictions for patients with shorter survival times, improving overall accuracy.

3.3 Curriculum II: Hierarchical Knowledge Distillation On Virtual Tail Patient

Head patients with a greater number and larger size of WSIs exhibit increased robustness to noise (e.g., tissue information loss) and thus can be utilized as a reference group with relatively complete data [Hellmann *et al.*, 2014]. To improve predictions for tail patients, we introduce noise into the WSIs of head patients, thereby simulating conditions characteristic of patients with limited data. The teacher-student distillation framework is formalized based on Eq. (7) as follows:

$$\min_{f_{\text{stu}}} \mathbb{E}_{\mathcal{W}^i} \left[\left\| f_{\text{tch}}(\mathcal{W}^i) - f_{\text{stu}}(\hat{\mathcal{W}}^i) \right\| \right], \tag{7}$$

where f_{stu} and f_{tch} are the student and teacher models, and $\mathcal{W}^i$ and $\hat{\mathcal{W}}^i$ represent the clean and noisy WSIs of the head patient. $\hat{\mathcal{W}}^i$ is the virtual tail patient, aligning the head and tail patient distributions.

Virtual Tail Patient Construction. As shown in Eq. (3), the virtual tail patient is generated by introducing artificial noise into the data of the head patient. Given the WSIs of a head patient, W_i^{head}, noise is applied both to the number and the resolution of the WSIs. The parameter ϵ_{n} modifies the number of WSIs, yielding a modified set $|W_i^{\text{head}} + \epsilon_{\text{n}}|$, while ϵ_{r} alters the resolution of each individual WSI, represented as $|W_j^{i,\text{head}} + \epsilon_{\text{r}}|$. The noise addition follows a "simple-to-difficult" strategy, wherein patches with low attention values are initially masked, and progressively, patches with higher attention values are also masked. Upon the application of this strategy to the head patient's WSIs W_i^{head}, the resulting virtual tail patient $\hat{W}_i^{\text{head}}$ is generated.

Hierarchical Knowledge Distillation. The distributional differences between head and tail patients span WSI representations, aggregated patient embeddings, and prediction distributions [Guan and Liu, 2021; Zhang *et al.*, 2024c], highlighting variations in data quality, feature representation, and predictive outcomes, which challenge knowledge transfer and model generalization. Thus, hierarchical knowledge distillation is applied at each level. A student model is constructed to mirror the teacher model's architecture, excluding the label embeddings. The architecture is shown in Eq. (4).

WSI-level Distillation. For each head and virtual tail patient, WSIs are passed through both teacher and student models, producing feature embeddings $\mathcal{S}_i^{\text{h}}$ and $\mathcal{S}_i^{v}$, respectively. A feature-based distillation loss is applied to the WSI pairs:

$$\mathcal{L}_{\text{Inst}} = \frac{1}{|W_i^{\text{head}}|} \sum_{j=1}^{|W_i^{\text{head}}|} \left\| \mathbf{S}_{i,j}^{\text{h}} - \mathbf{S}_{i,j}^{v} \right\|_2^2 . \tag{8}$$

Here, $|W_i^{\text{head}}|$ denotes the number of WSIs for the head patient. Given the data heterogeneity, directly minimizing the pairwise WSI distance is inefficient. Instead, we minimize the distance between individual WSIs and the domain-level discrepancy, focusing on mean and variance alignment.

To align the mean of the distributions, the Maximum Mean Discrepancy (MMD) loss [Gretton *et al.*, 2006] is used:

$$\mathcal{L}_{\text{mmd}}(\mathbf{S}_i^h, \mathbf{S}_i^v) = ||E_{s_t \sim \mathbf{S}_i^h}[k(\cdot, s_t)] - E_{s_v \sim \mathbf{S}_i^v}[k(\cdot, s_v)]||_{\mathcal{H}_k}, \tag{9}$$

which measures the distance between the means of these two distributions in the reproducing kernel Hilbert space (RKHS) associated with the kernel function k.

To align the variance, the similarity matrices for WSIs within each model are computed as

$$\mathbf{M}_i^{\text{head}} = \left[\text{sim}(S_{i,j}^{\text{h}}, S_{i,k}^{\text{h}}) \right]_{j,k=1}^{|W_i^{\text{head}}|}, \mathbf{M}_i^{\text{virtual}} = \left[\text{sim}(S_{i,j}^{\text{v}}, S_{i,k}^{\text{v}}) \right]_{j,k=1}^{|W_i^{\text{head}}|}. \tag{10}$$

The difference between these matrices quantifies the gap:

$$\mathcal{L}_{\text{sim}} = \left\|\mathbf{M}_i^{\text{head}} - \mathbf{M}_i^{\text{virtual}}\right\|_2^2. \tag{11}$$

The total WSI-level loss is:

$$\mathcal{L}_{\text{WSI}} = \mathcal{L}_{\text{Inst}} + \mathcal{L}_{\text{mmd}} + \mathcal{L}_{\text{sim}}. \tag{12}$$

In our implementation, each loss function is assigned a corresponding weight to balance its contribution to the overall loss. However, to simplify the notation and avoid redundancy, we omit the explicit representation of these weights.

Patient-level Distillation: At the patient level, alignment of features, distribution mean, and dispersion is required. The patient feature alignment mirrors the WSI-level losses:

$$\mathcal{L}_{\text{Patient}} = \mathcal{L}'_{\text{Inst}} + \mathcal{L}'_{\text{mmd}} + \mathcal{L}_{\text{sim}'}. \tag{13}$$

Additionally, the predicted risk depends not only on the score but also on the ranking within the batch. Thus, we minimize the distance between predicted scores and ranks:

$$\mathcal{L}_{\text{ot}} = \left\|\mathbf{O}_i^{\text{h}} - \mathbf{O}_i^{\text{v}}\right\|_2^2, \mathcal{L}_{\text{rank}} = \left\|\text{Rank}\left(\mathbf{O}_i^{\text{h}}\right) - \text{Rank}\left(\mathbf{O}_i^{\text{v}}\right)\right\|_2^2. \tag{14}$$

where $\mathcal{L}_{\text{ot}}$ represents the distance between predicted risk scores, and $\mathcal{L}_{\text{rank}}$ is the distance between their rankings.

The total loss for the student model in Curriculum II is:

$$\mathcal{L}_{\text{Stu}} = \mathcal{L}_{\text{sur}} + \mathcal{L}_{\text{WSI}} + \mathcal{L}_{\text{Patient}} + \mathcal{L}_{\text{ot}} + \mathcal{L}_{\text{rank}}. \tag{15}$$

3.4 Curriculum III: Student-centered Progressive Training On Tail Patient

Curriculum I extracts knowledge from head patients for the teacher model, while Curriculum II transfers this knowledge to the student model via virtual tail patients, reducing the performance gap. However, both curricula depend solely on head patient data, limiting tail patient usage. The challenge lies in the lack of identity-matched knowledge for direct transfer to tail patients, complicating learning. Incorporating tail patients with insufficient data risks performance degradation, especially in limited survival datasets. To address this, we propose constructing virtual head patients by augmenting tail patient data with potential WSIs. The teacher model learns from these virtual head patients, improving knowledge transfer to the student model. This student-centered knowledge distillation approach enhances tail patient representation using augmented data before guiding the student model.

Virtual Head Patient Construction. To transform a tail patient into a virtual head patient, we propose two methods:

Retrieval-based Expansion: A retrieval-based approach is proposed to enhance the number and resolution of WSIs. For resolution expansion, representative patches from the tail patient's WSI are selected based on attention scores. The top-M patches $E_{i,j}^{\text{tail},k}$, which capture critical information, are identified. Similar patches are then retrieved from the training set using a similarity function $\text{sim}(\cdot,\cdot)$, and the top-N_1 patches with the highest average similarity are selected for data augmentation. The expanded WSI $\hat{W}_{i,j}^{\text{tail}}$ is defined as:

$$\hat{W}_{i,j}^{\text{tail}} = \arg\max_{E_{i,j}^{\text{train}}} \frac{1}{M} \sum_{k=1}^{M} \text{sim}(E_{i,j}^{\text{tail},k}, E_{i,j}^{\text{train}}),\ k = 1, \ldots, N_1.$$

Dataset	# Patient	# WSI	# WSI of a patient		Cancer Type
			Mode	Maximum	
NLST [Team, 2011]	449	1,224	3	6	ADC&SCC
TCGA-LUSC [Tomczak *et al.*, 2015]	504	1,612		13	SCC
TCGA-LUAD [Tomczak *et al.*, 2015]	514	1,608		14	ADC
TCGA-BRCA [Tomczak *et al.*, 2015]	1,098	3,111		9	BIC
TCGA-BLCA [Tomczak *et al.*, 2015]	412	926	2	10	BUC

Table 1: The statistics of five datasets.

For number expansion, the aggregated WSI representations are compared to identify the top-N_2 similar WSIs, which are incorporated into the tail patient's data:

$$\hat{W}_i^{\text{tail}} = \mathcal{D}_2(W_i^{\text{tail}}; \phi).$$

Generative-based Expansion: To further expand WSI resolution and number, a generative approach is employed leveraging the Denoising Diffusion Probabilistic Model (DDPM). For resolution expansion, the DDPM is trained by masking representative patches, identified using attention values, and using the remaining patches as conditions to reconstruct the masked regions. This enables the DDPM to learn the latent feature space of critical regions, allowing it to generate synthetic patch features. For each tail WSI $\hat{W}_{i,j}^{\text{tail}}$, the DDPM $\mathcal{D}_1$ generates synthetic patch features $\{\hat{E}_{i,j}^{\text{tail},k}\}$:

$$\{\hat{E}_{i,j}^{\text{tail},k}\} = \mathcal{D}_1(E_{i,j}^{\text{tail},k}; \theta),$$

where θ represents the trained parameters of the DDPM.

For number expansion, another DDPM is trained to generate synthetic WSIs by masking complete WSIs and using the remaining WSIs as conditions to reconstruct the masked WSIs. The synthetic WSIs are obtained as:

$$\hat{W}_i^{\text{tail}} = \mathcal{D}_2(W_i^{\text{tail}}; \phi).$$

Progressive Training. The expanded WSIs of tail patients $\hat{W}_i^{\text{tail}}$ are first used to train the teacher model, and a progressive curriculum is employed to incorporate orignal WSIs of tail patients W_i^{tail} into the student model. The proportion and difficulty of the tail patients introduced increase over time. The difficulty of a tail patient is quantified by the change in survival risk before and after training the teacher model:

$$\Delta R_i^{\text{tail}} = R_i^{\text{tail, post-teacher}} - R_i^{\text{tail, pre-teacher}}, \tag{16}$$

where ΔR_i^{tail} reflects the improvement in the model's understanding of the patient. Tail patients are ranked by difficulty, and the batch inclusion function $\gamma(t)$ determines the proportion of tail patients included at time t:

$$\hat{W}_i^{\text{tail}}(t) = \{\hat{W}_i^{\text{tail}} \mid \Delta R_i^{\text{tail}} \leq \mathcal{Q}_{\gamma(t)}\}, \tag{17}$$

where $\mathcal{Q}_{\gamma(t)}$ represents the difficulty threshold based on the quantile of the sorted tail patients. The proportion $\gamma(t)$ increases over time following a sigmoid schedule:

$$\gamma(t) = \frac{1}{1 + e^{-\lambda(t-t_0)}}. \tag{18}$$

This curriculum learning strategy enables a gradual introduction of simpler tail patients, followed by more challenging ones, facilitating effective learning progression from simple to complex cases.

Architecture	**Model**	**Dataset**									
		NLST		LUSC		LUAD		BRCA		BLCA	
		C-index	STAGE-5	C-index	STAGE-5	C-index	STAGE-5	C-index	STAGE-5	C-index	STAGE-5
CNN	WSISA	$0.662_{.033}$	$0.433_{.021}$	$0.608_{.048}$	$0.565_{.028}$	$0.582_{.011}$	$0.501_{.030}$	$0.591_{.035}$	$0.534_{.022}$	$0.504_{.041}$	$0.432_{.036}$
	DeepAttnMISL	$0.630_{.038}$	$0.427_{.028}$	$0.670_{.049}$	$0.569_{.031}$	$0.563_{.022}$	$0.522_{.019}$	$0.603_{.025}$	$0.531_{.018}$	$0.517_{.023}$	$0.420_{.020}$
Transformer	HIPT	$0.619_{.026}$	$0.450_{.029}$	$0.655_{.009}$	$0.571_{.021}$	$0.552_{.039}$	$0.516_{.030}$	$0.589_{.030}$	$0.549_{.032}$	$0.552_{.019}$	$0.451_{.025}$
	SeTranSurv	$0.677_{.032}$	$0.401_{.021}$	$0.688_{.047}$	$0.550_{.033}$	$0.580_{.009}$	$0.489_{.018}$	$0.612_{.020}$	$0.537_{.019}$	$0.554_{.030}$	$0.458_{.021}$
	ESAT	$0.730_{.039}$	$0.435_{.034}$	$0.681_{.050}$	$0.564_{.033}$	$0.593_{.019}$	$0.510_{.014}$	$0.625_{.031}$	$0.522_{.015}$	$0.568_{.017}$	$0.472_{.032}$
	LongViT	$0.677_{.018}$	$0.440_{.029}$	$0.665_{.010}$	$0.562_{.013}$	$0.618_{.024}$	$0.531_{.034}$	$0.628_{.027}$	$0.545_{.018}$	$0.556_{.025}$	$0.477_{.020}$
	Prov-GigaPath	$0.665_{.030}$	$0.421_{.039}$	$0.661_{.034}$	$0.560_{.032}$	$0.609_{.028}$	$0.537_{.017}$	$0.620_{.015}$	$0.522_{.010}$	$0.562_{.021}$	$0.489_{.023}$
Bias	AugDiff	$0.653_{.025}$	$0.441_{.016}$	$0.678_{.018}$	$0.562_{.024}$	$0.580_{.022}$	$0.507_{.018}$	$0.619_{.029}$	$0.538_{.020}$	$0.532_{.014}$	$0.462_{.015}$
	Resampling	$0.639_{.040}$	$0.425_{.035}$	$0.675_{.044}$	$0.573_{.045}$	$0.562_{.018}$	$0.490_{.018}$	$0.631_{.020}$	$0.522_{.041}$	$0.544_{.010}$	$0.470_{.055}$
	MoE	$0.650_{.011}$	$0.423_{.037}$	$0.685_{.010}$	$0.573_{.029}$	$0.592_{.009}$	$0.529_{.045}$	$0.630_{.044}$	$0.550_{.025}$	$0.541_{.022}$	$0.476_{.021}$
	Retrival	$0.657_{.023}$	$0.402_{.025}$	$0.689_{.033}$	$0.541_{.010}$	$0.573_{.055}$	$0.530_{.025}$	$0.602_{.035}$	$0.544_{.041}$	$0.505_{.039}$	$0.489_{.020}$
	Ours	$\mathbf{0.760_{.015}}$	$\mathbf{0.499_{.015}}$	$\mathbf{0.720_{.033}}$	$\mathbf{0.619_{.015}}$	$\mathbf{0.675_{.031}}$	$\mathbf{0.583_{.020}}$	$\mathbf{0.683_{.019}}$	$\mathbf{0.597_{.040}}$	$\mathbf{0.612_{.022}}$	$\mathbf{0.530_{.019}}$

Table 2: The results achieved by all competing methods on five datasets. The boldface indicates the best results.

4 Experiment

4.1 Experimental Settings

Datasets We evaluate our proposal on five real-world cancer datasets. The first is the National Lung Screening Trial (NLST) dataset [Team, 2011], which includes cases of adenocarcinoma (ADC) and squamous cell carcinoma (SCC). The remaining four datasets are from The Cancer Genome Atlas (TCGA) [Tomczak *et al.*, 2015], encompassing lung cancer (LUSC and LUAD), breast cancer (BRCA), and bladder cancer (BLCA). Each dataset contains WSIs stained with hematoxylin and eosin (H&E) and corresponding survival labels.

Baselines Three types of SOTA baselines are considered: (1) Traditional CNN-based methods that using partial WSI data: **WSISA** [Zhu *et al.*, 2017] and **DeepAttnMISL** [Yao *et al.*, 2020b]; (2) Transformer-based methods utilizing full WSI data: **HIPT** [Chen *et al.*, 2022b], **SeTranSurv** [Huang *et al.*, 2021], **ESAT** [Shen *et al.*, 2022], **LongViT** [Wang *et al.*, 2023a], and **Prov-GigaPath** [Xu *et al.*, 2024]; and (3) Methods addressing bias: **AugDiff** [Shao *et al.*, 2023b], **Resampling** [Good, 2006], **MoE** [Masoudnia and Ebrahimpour, 2014], and **Retrieval** [Wang *et al.*, 2023b].

Implementation Details All baseline survival models are re-implemented faithfully from their original publications and available open-source repositories. For a fair comparison, we adopt a 5-fold cross-validation scheme: in each fold, 10% of the training split is held out as a validation set for early stopping [Bai *et al.*, 2021]. Models are trained for up to 200 epochs (with patience of 10 epochs on validation loss) using the AdamW optimizer (weight decay $1\mathrm{e}{-4}$) and a fixed batch size of 64. All experiments are carried out in PyTorch 2.2.0 on NVIDIA V100 GPUs with 32 GB of memory. Performance metrics are averaged over the five folds.

4.2 Quantitative Evaluation

Table 2 presents survival prediction results using the C-index and STAGE-5 metrics. (1) PathoKD outperforms all state-of-the-art baselines, with improvements of over 4.28% in C-index and 4.58% in STAGE-5. (2) Transformer-based models outperform earlier CNN-based methods, which rely on partial WSI data, by processing the full WSI. (3) Our method further improves performance by addressing the challenges of limited WSI number and resolution through knowledge distillation. (4) Compared to other models that tackle number and resolution bias without distillation, our approach offers superior performance by transferring high-level, task-relevant knowledge rather than manipulating low-level model inputs.

4.3 Performance on Mitigating Bias

A controlled experiment is conducted to validate our approach in addressing resolution and number biases in WSIs and its impact on survival prediction. Patients are categorized into three groups based on WSI resolution and number: low (fewer than 3 WSIs and the bottom 20% in size), medium (exactly 3 WSIs and the middle 60%), and high (more than 3 WSIs and the top 20%). As shown in Table 3, applying knowledge distillation significantly reduces the performance gap between these groups, demonstrating that PathoKD effectively mitigates this challenge. Notably, even the high group benefits from knowledge distillation, underscoring the effectiveness of our hierarchical alignment loss.

4.4 Ablation Study

We conduct ablation studies on the NLST and LUAD datasets to evaluate the contribution of each module in PathoKD, with results presented in Table 4. The first row shows that using label embedding alone, without noise, leads to over-reliance on labels and a significant performance drop, emphasizing the need for noise incorporation. The second row reveals that excluding Curriculum I severely degrades performance, highlighting the importance of labels as external knowledge. Removing instance distillation $\mathcal{L}_{\text{Inst}}$ results in a notable performance decline, underlining its role in mitigating bias and facilitating knowledge transfer. Omitting MMD distillation $\mathcal{L}_{\text{mmd}}$ also decreases performance, stressing the importance of aligning distributions between head and tail patients. Excluding similarity distillation $\mathcal{L}_{\text{Sim}}$ further reduces performance, demonstrating its critical role in maintaining variance in the representation distribution. The removal of prediction score alignment—one of the highest-level features—has the most significant impact, emphasizing the importance of ranking in measuring prediction score distribution. Lastly, excluding Curriculum III and neglecting tail patients results in a performance decline, reinforcing the effectiveness of our progres-

KD	NLST				LUAD				LUSC			
	Low	Medium	High	Gap	Low	Medium	High	Gap	Low	Medium	High	Gap
w/o KD	$0.617_{.045}$	$0.651_{.042}$	$0.670_{.029}$	0.053	$0.552_{.050}$	$0.593_{.047}$	$0.617_{.039}$	0.051	$0.643_{.036}$	$0.680_{.039}$	$0.705_{.027}$	0.062
PathoKD	$0.683_{.022}$	$0.689_{.030}$	$0.701_{.011}$	0.018	$0.611_{.017}$	$0.619_{.029}$	$0.632_{.009}$	0.021	$0.699_{.020}$	$0.705_{.016}$	$0.712_{.010}$	0.013

Table 3: C-index performance across different groups of patients with and without knowledge distillation.

Models	***C-index*** ↑		***STAGE-5*** ↑	
	NLST	**LUAD**	**NLST**	**LUAD**
w/o Noise	$0.652_{.053}$	$0.600_{.042}$	$0.381_{.055}$	$0.501_{.028}$
w/o C I	$0.730_{.041}$	$0.642_{.035}$	$0.453_{.027}$	$0.548_{.020}$
w/o $\mathcal{L}_{\text{Inst}}$	$0.709_{.055}$	$0.602_{.049}$	$0.447_{.045}$	$0.531_{.030}$
w/o $\mathcal{L}_{\text{mmd}}$	$0.731_{.030}$	$0.635_{.022}$	$0.466_{.033}$	$0.554_{.047}$
w/o $\mathcal{L}_{\text{Sim}}$	$0.750_{.022}$	$0.649_{.009}$	$0.472_{.029}$	$0.577_{.040}$
w/o $\mathcal{L}_{\text{Rank}}$	$0.742_{.038}$	$0.650_{.027}$	$0.471_{.029}$	$0.556_{.035}$
w/o C III	$0.748_{.029}$	$0.659_{.037}$	$0.482_{.020}$	$0.565_{.041}$
Retrival	$0.755_{.020}$	$0.654_{.025}$	$0.493_{.021}$	$0.578_{.031}$
Generative	$\mathbf{0.760_{.015}}$	$\mathbf{0.675_{.031}}$	$\mathbf{0.499_{.015}}$	$\mathbf{0.583_{.020}}$

Table 4: Ablation study on two datasets.

sive learning approach. Between the two methods in Curriculum III, the generative approach outperforms retrieval, showcasing the DDPM's ability to capture latent features.

4.5 Efficiency Study

The efficiency of the proposed model is evaluated on the NLST dataset [Team, 2011] and compared with SOTA models. All configurations follow the original papers, using official implementations where available. For WSISA, the implementation was recreated due to the absence of official code. This study is the first to evaluate bias mitigation strategies (Resampling, MoE, and Retrieval) on survival prediction. All models were run on two Xeon E5-2690 v4 processors (2.60 GHz) and four NVIDIA V100 GPUs. Results show that our approach outperforms others in both accuracy and inference speed. Specifically, (1) Our method achieves an 11.4% higher C-index and runs 1.40 times faster than clustering-based models like WSISA and DeepAttnMISL, due to the transformer architecture's direct processing of patches without computationally expensive clustering. (2) Our approach also outperforms other bias mitigation methods, achieving 34.6% faster inference by directly using original data without the additional preprocessing required by other models.

4.6 Parameter Sensitivity Analysis

Influence of noise strength γ_1 and γ_2 of discarding WSIs and patches: The γ_1 and γ_2 specify the number and percentage of discarding WSIs and patches. As illustrated in the left side of Fig. 3, PathoKD performs best with a mask WSI number of 3 and a mask patch ratio of 50%. Lower mask ratios might necessitate masking a greater proportion of WSIs to align virtual tail patients with tail patients, while higher ratios may leave too few visible histopathological features to effectively capture prognosis.

Influence of noise strength γ_3 and γ_4 of expanding WSIs and patches: The γ_3 and γ_4 specify the number and percentage of expanding WSIs and patches. As illustrated in the right side of Fig. 3, PathoKD performs best with an expanding WSI number of 3 and a mask patch ratio of 30%. Lower expansion ratios may require masking a larger proportion of WSIs to align virtual head patients with actual head patients, while higher ratios could introduce excessive noise due to the reduced proportion of real WSIs.

Architecture	Methods	C-index	Throughput	Delay
			(WSIs/s)	(ms)
CNN	WSISA	$0.662_{.033}$	0.25	3528
	DeepAttnMISL	$0.630_{.038}$	0.14	8263
Transformer	ESAT	$0.730_{.039}$	0.28	4350
	Prov-GigaPath	$0.665_{.030}$	0.21	6187
Bias	AugDiff	$0.653_{.025}$	0.10	9818
	Resampling	$0.639_{.040}$	0.26	5010
	MoE	$0.650_{.011}$	0.19	6802
	Retrival	$0.657_{.023}$	0.15	7974
	Ours	$\mathbf{0.760_{.015}}$	**0.35**	**2270**

Table 5: Our performance and efficiency with comparisons to state-of-the-art models trained on the NLST dataset.

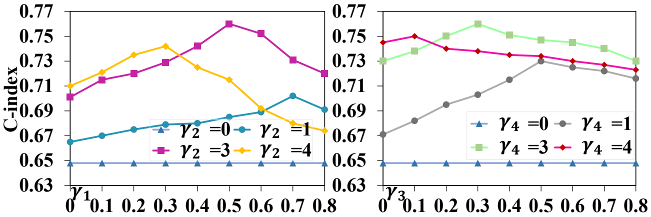

Figure 3: C-index results with varying strength of discarding (**Left**) or expanding (**Right**) WSIs and Patches.

5 Conclusion

In this paper, we propose PathoKD, a novel model for WSI-based survival prediction, designed to address the challenges of data variability and scarcity, particularly for tail patients. By integrating hierarchical knowledge distillation with curriculum learning, PathoKD effectively mitigates the performance gap between patients with limited and abundant data. Extensive experiments on multiple widely used datasets demonstrate its superiority in improving prediction accuracy across diverse patient groups, providing a promising approach for developing more robust and equitable survival prediction models.

Acknowledgments

Zhihao Tang is the first student author. This work was supported by the Natural Science Foundation of China (No. 62372057). This work was partially sponsored by the CAAI-Huawei MindSpore Open Fund. Compute services from Hebei Artificial Intelligence Computing Center.

References

[Bai *et al.*, 2021] Ying-Long Bai, Erkun Yang, Bo Han, Yanhua Yang, Jiatong Li, Yinian Mao, Gang Niu, and Tongliang Liu. Understanding and improving early stopping for learning with noisy labels. *ArXiv*, abs/2106.15853, 2021.

[Chen *et al.*, 2022a] Richard J Chen, Chengkuan Chen, Yicong Li, Tiffany Y Chen, Andrew D Trister, Rahul G Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16144–16155, 2022.

[Chen *et al.*, 2022b] Richard J. Chen, Chengkuan Chen, Yicong Li, Tiffany Y. Chen, Andrew D. Trister, Rahul G. Krishnan, and Faisal Mahmood. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16123–16134, 2022.

[Fischer *et al.*, 2008] Andrew H Fischer, Kenneth A Jacobson, Jack Rose, and Rolf Zeller. Paraffin embedding tissue samples for sectioning. *CSH protocols*, 2008:pdb–prot4989, 2008.

[Gadiya *et al.*, 2020] Shrey Gadiya, Deepak Anand, and Amit Sethi. Histographs: Graphs in histopathology. In *Medical Imaging: Digital Pathology*, 2020.

[Good, 2006] Phillip I Good. *Resampling methods*. Springer, 2006.

[Gretton *et al.*, 2006] Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex Smola. A kernel method for the two-sample-problem. *Advances in neural information processing systems*, 19, 2006.

[Guan and Liu, 2021] Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2021.

[Hanna *et al.*, 2020] Matthew G Hanna, Anil Parwani, and Sahussapont Joseph Sirintrapun. Whole slide imaging: technology and applications. *Advances in Anatomic Pathology*, 27(4):251–259, 2020.

[Hellmann *et al.*, 2014] Matthew D Hellmann, Jamie E Chaft, William N William, Valerie Rusch, Katherine MW Pisters, Neda Kalhor, Apar Pataer, William D Travis, Stephen G Swisher, and Mark G Kris. Pathological response after neoadjuvant chemotherapy in resectable non-small-cell lung cancers: proposal for the use of major pathological response as a surrogate endpoint. *The lancet oncology*, 15(1):e42–e50, 2014.

[Hinton, 2015] Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[Huang *et al.*, 2021] Ziwang Huang, Hua Chai, Ruoqi Wang, Haitao Wang, Yuedong Yang, and Hejun Wu. Integration of patch features through self-supervised learning and transformer for survival analysis on whole slide images. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part VIII 24*, pages 561–570. Springer, 2021.

[Jaume *et al.*, 2021] Guillaume Jaume, Pushpak Pati, Valentin Anklin, Antonio Foncubierta, and Maria Gabrani. Histocartography: A toolkit for graph analytics in digital pathology. *ArXiv*, abs/2107.10073, 2021.

[Liu *et al.*, 2023] Pei Liu, Luping Ji, Feng Ye, and Bo Fu. Graphlsurv: A scalable survival prediction network with adaptive and sparse structure learning for histopathological whole-slide images. *Computer methods and programs in biomedicine*, 231:107433, 2023.

[Masoudnia and Ebrahimpour, 2014] Saeed Masoudnia and Reza Ebrahimpour. Mixture of experts: a literature survey. *Artificial Intelligence Review*, 42:275–293, 2014.

[Otsu, 1979] Nobuyuki Otsu. A threshold selection method from gray-level histograms. *IEEE Trans. Syst. Man Cybern.*, 9:62–66, 1979.

[Pantanowitz *et al.*, 2011] Liron Pantanowitz, Paul N Valenstein, Andrew J Evans, Keith J Kaplan, John D Pfeifer, David C Wilbur, Laura C Collins, and Terence J Colgan. Review of the current state of whole slide imaging in pathology. *Journal of pathology informatics*, 2(1):36, 2011.

[Shao *et al.*, 2023a] Zhuchen Shao, Yang Chen, Hao Bian, Jian Zhang, Guojun Liu, and Yongbing Zhang. Hvtsurv: Hierarchical vision transformer for patient-level survival prediction from whole slide image. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 2209–2217, 2023.

[Shao *et al.*, 2023b] Zhuchen Shao, Liuxi Dai, Yifeng Wang, Haoqian Wang, and Yongbing Zhang. Augdiff: Diffusion based feature augmentation for multiple instance learning in whole slide image. *arXiv preprint arXiv:2303.06371*, 2023.

[Shedden *et al.*, 2008] Kerby Shedden, Jeremy M. G. Taylor, Steven Enkemann, Ming-Sound Tsao, Timothy Yeatman, William L. Gerald, Steven A. Eschrich, Igor Jurisica, Thomas J. Giordano, David E. Misek, Andrew C. Chang, Changyun Zhu, D. Strumpf, Samir M. Hanash, Frances A. Shepherd, Keyue Ding, Lesley Seymour, Katsuhiko Naoki, Nathan A. Pennell, Barbara A. Weir, Roel G. W. Verhaak, Christine Ladd-Acosta, Todd R. Golub, Mike Gruidl, Anupama Sharma, János Szőke, Maureen F. Zakowski, Valerie Rusch, Mark G Kris, Agnes Viale, Noriko Motoi, William D. Travis, Barbara A. Conley, Venkatraman E. Seshan, Matthew L. Meyerson, Rork Kuick, Kevin K. Dobbin, Tracy Lively, James W. Jacobson, and David G. Beer. Gene expression–based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nature Medicine*, 14:822–827, 2008.

[Shen *et al.*, 2022] Yifan Shen, Li Liu, Zhihao Tang, Zongyi Chen, Guixiang Ma, Jiyan Dong, Xi Zhang, Lin Yang, and Qingfeng Zheng. Explainable survival analysis with convolution-involved vision transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2207–2215, 2022.

[Strassburg and Manns, 2006] Christian P Strassburg and Michael P Manns. Approaches to liver biopsy techniques-revisited. In *Seminars in liver disease*, volume 26, pages 318–327. Copyright© 2006 by Thieme Medical Publishers, Inc., 333 Seventh Avenue, New . . . , 2006.

[Tang *et al.*, 2019] Bo Tang, Ao Li, Bin Li, and Minghui Wang. Capsurv: Capsule network for survival analysis with whole slide pathological images. *IEEE Access*, 7:26022–26030, 2019.

[Team, 2011] National Lung Screening Trial Research Team. The national lung screening trial: overview and study design. *Radiology*, 258(1):243–253, 2011.

[Tomczak *et al.*, 2015] Katarzyna Tomczak, Patrycja Czerwińska, and Maciej Wiznerowicz. Review the cancer genome atlas (tcga): an immeasurable source of knowledge. *Contemporary Oncology/Współczesna Onkologia*, 2015(1):68–77, 2015.

[Wang *et al.*, 2023a] Wenhui Wang, Shuming Ma, Hanwen Xu, Naoto Usuyama, Jiayu Ding, Hoifung Poon, and Furu Wei. When an image is worth 1, 024 x 1, 024 words: A case study in computational pathology. *ArXiv*, abs/2312.03558, 2023.

[Wang *et al.*, 2023b] Xiyue Wang, Yuexi Du, Sen Yang, Jun Zhang, Minghui Wang, Jing Zhang, Wei Yang, Junzhou Huang, and Xiao Han. Retccl: Clustering-guided contrastive learning for whole-slide image retrieval. *Medical image analysis*, 83:102645, 2023.

[Xu *et al.*, 2024] Hanwen Xu, Naoto Usuyama, Jaspreet Bagga, Sheng Zhang, Rajesh Rao, Tristan Naumann, Cliff Wong, Zelalem Gero, Javier González, Yu Gu, Yanbo Xu, Mu-Hsin Wei, Wenhui Wang, Shuming Ma, Furu Wei, Jianwei Yang, Chun yue Li, Jianfeng Gao, Jaylen Rosemon, Tucker Bower, Soohee Lee, Roshanthi K. Weerasinghe, Bill Wright, Ari Robicsek, Brian Piening, Carlo Bifulco, Sheng Wang, and Hoifung Poon. A whole-slide foundation model for digital pathology from real-world data. *Nature*, 630:181 – 188, 2024.

[Yao *et al.*, 2020a] Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical Image Analysis*, 65:101789, 2020.

[Yao *et al.*, 2020b] Jiawen Yao, Xinliang Zhu, Jitendra Jonnagaddala, Nicholas Hawkins, and Junzhou Huang. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. *Medical image analysis*, 65:101789, 2020.

[Zhang *et al.*, 2024a] Litian Zhang, Xiaoming Zhang, Chaozhuo Li, Ziyi Zhou, Jiacheng Liu, Feiran Huang, and Xi Zhang. Mitigating social hazards: Early detection of fake news via diffusion-guided propagation path generation. In *ACM Multimedia 2024*, 2024.

[Zhang *et al.*, 2024b] Litian Zhang, Xiaoming Zhang, Ziyi Zhou, Feiran Huang, and Chaozhuo Li. Reinforced adaptive knowledge learning for multimodal fake news detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 16777–16785, 2024.

[Zhang *et al.*, 2024c] Litian Zhang, Xiaoming Zhang, Ziyi Zhou, Xi Zhang, Senzhang Wang, S Yu Philip, and Chaozhuo Li. Early detection of multimodal fake news via reinforced propagation path generation. *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[Zhao *et al.*, 2023] Qihao Zhao, Chen Jiang, Wei Hu, Fan Zhang, and Jun Liu. Mdcs: More diverse experts with consistency self-distillation for long-tailed recognition. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11597–11608, 2023.

[Zhu *et al.*, 2016] Xinliang Zhu, Jiawen Yao, and Junzhou Huang. Deep convolutional neural network for survival analysis with pathological images. *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 544–547, 2016.

[Zhu *et al.*, 2017] Xinliang Zhu, Jiawen Yao, Feiyun Zhu, and Junzhou Huang. Wsisa: Making survival prediction from whole slide histopathological images. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6855–6863, 2017.