Medical Image Analysis 91 (2024) 103020

Contents lists available at ScienceDirect

Medical Image Analysis

journal homepage: www.elsevier.com/locate/media

AdvMIL: Adversarial multiple instance learning for the survival analysis on whole-slide images[☆], [☆☆]

Pei Liu [a], Luping Ji [a,*], Feng Ye [b], Bo Fu [a]

[a] School of Computer Science and Engineering, University of Electronic Science and Technology of China, Xiyuan Ave, Chengdu, 611731, Sichuan, China
[b] Institute of Clinical Pathology, West China Hospital, Sichuan University, Guo Xue Xiang, Chengdu, 610041, Sichuan, China

ARTICLE INFO

Dataset link: https://github.com/liupei101/AdvMIL

Keywords:
Computational pathology
Whole-slide image
Survival analysis
Time-to-event modeling
Multiple instance learning
Generative adversarial network

ABSTRACT

The survival analysis on histological whole-slide images (WSIs) is one of the most important means to estimate patient prognosis. Although many weakly-supervised deep learning models have been developed for gigapixel WSIs, their potential is generally restricted by classical survival analysis rules and fully-supervised learning requirements. As a result, these models provide patients only with a completely-certain point estimation of time-to-event, and they could only learn from the labeled WSI data currently at a small scale. To tackle these problems, we propose a novel adversarial multiple instance learning (AdvMIL) framework. This framework is based on adversarial time-to-event modeling, and integrates the multiple instance learning (MIL) that is much necessary for WSI representation learning. It is a plug-and-play one, so that most existing MIL-based end-to-end methods can be easily upgraded by applying this framework, gaining the improved abilities of survival distribution estimation and semi-supervised learning. Our extensive experiments show that AdvMIL not only could often bring performance improvement to mainstream WSI survival analysis methods at a relatively low computational cost, but also enables these methods to effectively utilize unlabeled data via semi-supervised learning. Moreover, it is observed that AdvMIL could help improving the robustness of models against patch occlusion and two representative image noises. The proposed AdvMIL framework could promote the research of survival analysis in computational pathology with its novel adversarial MIL paradigm.

1. Introduction

Survival analysis, also known as time-to-event analysis, is one of the primary statistical approaches for analyzing data on time to event (Cox, 1975; Kalbfleisch and Prentice, 2011). It is usually adopted in medical fields to analyze clinical materials and assist doctors in understanding disease prognosis (Wulczyn et al., 2021). Histological whole-slide image (WSI) is one of these materials. It is produced by scanning tissue slides (millimeter scale) with a high-end microscope. Compared with other materials like demographics and genomics, digitized WSIs can present unique hierarchical views at a gigapixel-resolution (Zarella et al., 2018), *e.g.*, tissue phenotype, tumor microenvironment, and cellular morphology. These rich and diverse microscopic information could provide valuable cues for the prognosis of tumor diseases (Yu et al., 2016; Chen et al., 2022b), contributing to the improvement of patient management and disease outcomes (Nir et al., 2018; Kather et al., 2019; Skrede et al., 2020).

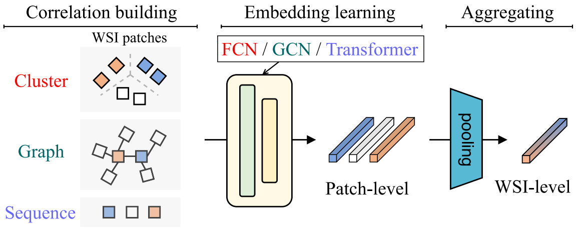


Fig. 1. Paradigm of embedding-level multiple-instance learning (MIL) for WSI representation learning. The structure of cluster, graph, or sequence is usually adopted to build patch correlations. These correlations are utilized by different networks to learn WSI-level representations.

Unlike regular natural images, histological WSIs are usually with an extremely-high resolution, *e.g.*, 40,000 × 40,000 pixels. This poses

[☆] This work is supported by the National Natural Science Foundation of China (NSFC) under Grant No. 61972072, partly by Sichuan Science and Technology Program under Grant No. 2021YFS0071 and 23QYCX0070, and the Fundamental Research Funds for the Central Universities under Grant No. ZYGX2022YGRH015.

[☆☆] Source code is available at https://github.com/liupei101/AdvMIL for facilitating further research.

[*] Corresponding author.
E-mail addresses: yuukilp@163.com (P. Liu), jiluping@uestc.edu.cn (L. Ji), fengye@scu.edu.cn (F. Ye), fubo@uestc.edu.cn (B. Fu).

https://doi.org/10.1016/j.media.2023.103020
Received 13 December 2022; Received in revised form 5 April 2023; Accepted 31 October 2023
Available online 2 November 2023

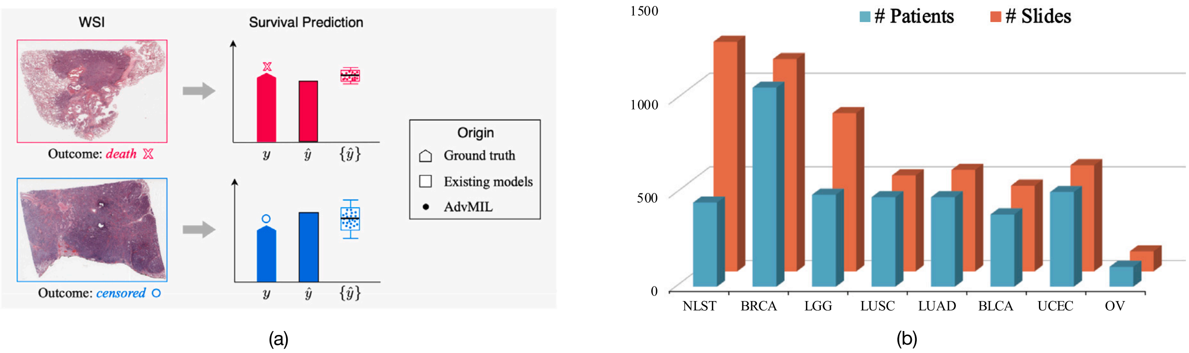


Fig. 2. The commonalities of existing WSI survival analysis models in terms of output and input: (a) model output, existing methods are limited to a point time-to-event estimation, whereas ours can provide an estimation of time-to-event distribution, believed to be more robust and interpretable; (b) model input, all the most frequently-used datasets for WSI survival analysis are at a very small scale, usually with around 500 patients or 1000 slides.

great challenges to WSI analysis and modeling, especially the global representation learning of WSIs. To tackle these challenges, many methods follow a weakly-supervised framework with three stages: (i) WSI patching, (ii) patch-level feature extracting, and (iii) slide-level representation learning (Chen et al., 2022a; Ghaffari Laleh et al., 2022). In procedure, this framework derives global representations through building patch correlations, learning patch-level embeddings, and aggregating patch-level embeddings, as shown in Fig. 1. It is often cast as embedding-level multiple instance learning (MIL) (Ilse et al., 2018; Carbonneau et al., 2018). According to the structure of patch correlation, the mainstream MIL methods for survival analysis can be roughly grouped into three categories: cluster-based (Yao et al., 2020; Shao et al., 2021), graph-based (Li et al., 2018; Chen et al., 2021), and sequence-based (Huang et al., 2021; Shen et al., 2022; Liu et al., 2023).

Although there are many kinds of methods for the survival analysis on gigapixel WSIs, these methods are generally restricted by the classical survival analysis rules that include a certain survival assumption regarding hazard function (Cox, 1975; Kalbfleisch and Prentice, 2011; Liu et al., 2021) and a likelihood estimation in discrete time domain Zadeh and Schmid (2021). This leads to a point survival estimation of them. However, the output of point estimation, compared with that of distribution estimation illustrated in Fig. 2(a), represents a single completely-certain result believed to be lacking in predictive robustness and interpretability (Lakshminarayanan et al., 2017; Kendall and Gal, 2017; Nazarovs et al., 2022; Linmans et al., 2023). In addition, in terms of input, all these methods train their own networks with slide-level labels, following a fully-supervised learning setting. This means that their training needs sufficient labeled data to make the models well-generalized to unseen samples. However, among current publicly-available WSI datasets, patient number is often limited to around 500, as shown in Fig. 2(b). By contrast, the standard datasets in deep learning, such as ImageNet (Deng et al., 2009), often contains more than 10,000 samples. This fact indicates that current WSI-based survival analysis is still in a small data regime that is believed to have adverse effects on the generalization ability of deep learning models (Chapelle et al., 2009; Zhou, 2021; Marini et al., 2021).

Generative adversarial network (GAN) (Goodfellow et al., 2014a) offers means to mitigate these problems. On one hand, GAN is a generative model capable of estimating complex data distribution via implicitly sampling, naturally fit for predictive distribution modeling. On the other hand, the generator–discriminator structure of GAN can take fake (or unlabeled) samples as input, just meeting the needs of semi-supervised learning (Goodfellow, 2016). In this way, GAN-based models are never limited to certain point estimation and fully-supervised learning; instead, they provide robust distribution estimations, and notably, they learn from unlabeled data to enhance their generalization ability (Springenberg, 2015; Salimans et al., 2016; Miyato et al., 2018; Li et al., 2021).

In the past few years, GAN has attracted great attention and inspired many interesting applications beyond image generation (Gui et al., 2021). Survival analysis is one of them. It is first combined with a conditional GAN (cGAN) (Mirza and Osindero, 2014), referred to as adversarial time-to-event modeling, in order to develop a general assumption-free survival model for analyzing clinical tabular data (Chapfuwa et al., 2018, 2020). Afterward, this general model is successfully extended to convolutional neural networks for the time-to-event analysis on CT images (Uemura et al., 2021). These models have shown promising results, especially their advantages of predictive accuracy and robustness. Further adoption of GAN for survival analysis in computational pathology is strongly anticipated, since the current survival analysis on WSI data still faces the problems posed by its conventional modeling paradigm aforementioned, *i.e.*, classical survival analysis rules and fully-supervised learning.

In this study, we propose a novel framework for the survival analysis on gigapixel whole-slide images, referred to as adversarial multiple instance learning (AdvMIL). This framework no longer relies on the classical paradigm of time-to-event modeling; instead, it is based on adversarial time-to-event modeling and integrates the multiple instance learning that is much necessary for WSI representation learning. As a result, most existing embedding-level MIL networks can be easily integrated into the proposed AdvMIL framework, thereby gaining the ability of survival distribution estimation and semi-supervised learning without introducing extra techniques. The results on three publicly-available WSI datasets verify that AdvMIL could often bring performance improvements to mainstream MIL networks at a relatively low computational cost; and most importantly, it could enable current models to effectively utilize unlabeled WSI data via semi-supervised learning.

The primary contributions of this work are listed as follows.

(I) We present an adversarial multiple instance learning (AdvMIL) framework for the survival analysis on gigapixel whole-slide images. Most existing MIL-based WSI survival analysis methods can be upgraded by applying this framework, thereby gaining the abilities of survival distribution estimation and semi-supervised learning. To our best knowledge, the proposed AdvMIL is the first one to adopt GAN in computational pathology for survival analysis.

(II) We demonstrate how GAN can be combined with MIL paradigm to perform survival analysis on gigapixel WSIs, by the two key components of AdvMIL: the MIL encoder in generator and the fusion network with region-level instance projection (RLIP) in discriminator.

(III) We further explore how to train existing WSI models with our AdvMIL framework in a semi-supervised manner. Moreover, we propose a k-fold training strategy to make the semi-supervised learning more effective.

(IV) We validate the effectiveness of AdvMIL in predictive performance, semi-supervised learning, and model robustness, through the extensive experiments on a total of 3101 WSIs from three publicly-available datasets. Empirical results suggest that AdvMIL could boost the development of survival analysis in computational pathology by its adversarial MIL paradigm.

2. Related work

2.1. Survival analysis of WSIs

Predicting time-to-event from digitized gigapixel WSIs is an active research topic in recent years. Here we review some representative works, mainly focusing on their networks and survival loss functions.

(1) *Multiple-instance learning network*

To learn global representations from gigapixel WSIs, MIL (see Fig. 1) is widely adopted in end-to-end deep learning models (Ghaffari Laleh et al., 2022). These models usually employ different backbones for different patch correlations, *e.g.*, fully-connected networks for clusters, graph convolution networks (Kipf and Welling, 2016) for graphs, and Attention-MIL (Ilse et al., 2018) or Transformers (Vaswani et al., 2017; Dosovitskiy et al., 2020) for sequences. From an unified perspective, these networks first transfer patch features along patch correlations to learn non-local patch-level embeddings, and then extract WSI representation by pooling these embeddings into a global feature vector.

(2) *Survival loss function*

Most models make some assumptions on hazard function (Kalbfleisch and Prentice, 2011) for survival analysis. Cox proportional hazard (Cox, 1975) and accelerated failure time (Wei, 1992) are the two most classical assumptions in them. The loss functions based on these two assumptions are very popular in WSI survival modeling, owing to their simplicity and interpretability. In addition, the survival loss function based on a maximum likelihood estimation, which has demonstrated good calibration and discrimination on tabular data (Zadeh and Schmid, 2021), is adopted by Chen et al. (2021, 2022a) and Liu et al. (2023) for WSI analysis. It requires a pre-discretization of survival time although it is assumption-free. These functions are derived from a class of *discriminative* models, often leading to a limitation—the point estimation of time-to-event.

2.2. Adversarial time-to-event analysis

(1) *Generative adversarial network*

GAN is a powerful *generative* model (Goodfellow et al., 2014a) with a wide variety of vision applications (Gui et al., 2021), *e.g.*, image generation, denoising, and editing. Additionally, it has also been extensively studied to perform semi-supervised learning (Springenberg, 2015; Carmon et al., 2019) or enhance the robustness of deep learning networks (Goodfellow et al., 2014b; Miyato et al., 2018). Following GAN, a conditional GAN (cGAN) (Mirza and Osindero, 2014) is first proposed to improve the quality of image generation by incorporating conditional labels. Based on cGAN, many efforts have been made to better exploit conditional labels, such as the projection discriminator (Miyato and Koyama, 2018), originally developed for a general application of regular natural images. As a representative work of cGAN, the projection discriminator projects a conditional vector on input's final embedding.

(2) *Time-to-event modeling via conditional GAN*

It is seen in DATE (Chapfuwa et al., 2018) for the first time, motivated by the fact that cGAN can capture complex data distribution (Goodfellow et al., 2014a). Like cGAN, DATE also uses a generator–discriminator architecture. Specifically, the generator, denoted as G, estimates a conditional distribution of time-to-event by

$$\hat{t} = G(x, \mathcal{N}) \sim P_{t|x}, \tag{1}$$

where x denotes a conditional input and $\mathcal{N}$ denotes a noise input. The discriminator, denoted as D, recognizes the real input (x, t) or the fake input $(x, \hat{t})$, where t is the label of time-to-event w.r.t. x. Such an adversarial generative way enables DATE to estimate the distribution of time-to-event via implicitly sampling from G, rather than via learning and optimizing the parameters of a priori distribution.

DATE is a general assumption-free model of time-to-event analysis, originally applied to clinical tabular data. Following the framework of DATE, pix2surv (Uemura et al., 2021) is further developed with convolutional networks for CT images. This research of time-to-event modeling on tabular data and CT images has benefited from the generative modeling paradigm, and has demonstrated better predictive accuracy and robustness. However, on pathological WSIs it still remains open. Moreover, how to perform semi-supervised learning and whether semi-supervised learning is effective for survival analysis, have not been studied yet in both DATE and pix2surv.

In Section 3, we show how the framework of adversarial time-to-event modeling can be generalized to MIL so as to perform survival prediction tasks on gigapixel WSIs.

3. Methodology

We show our Adversarial Multiple Instance Learning (AdvMIL) framework in Fig. 3. In a nutshell, AdvMIL generalizes adversarial time-to-event modeling to MIL by its two cores: the MIL encoder in generator and the fusion network with region-level instance projection (RLIP) in discriminator. At the end of this section, we describe the semi-supervised learning with AdvMIL, as well as a k-fold training strategy.

3.1. Preliminary

(1) *Bag construction for WSIs*

As illustrated in Fig. 3, we prepare two types of patches in bag construction: (i) the individual patches without any special structure for our generator, and (ii) the patches with equal region partition for our discriminator. These two types are *only* different in the way of organizing patches.

Specifically, we design a big-to-small continuous-patching scheme. With this scheme, we can not only obtain the two types of patches described above, but also filter some patches in background to save computation cost. Namely, for one WSI, we first fix its magnification at $f\times$ (20× is a typical setting) and slice it into *big* regions, each with the size of $\eta a \times \eta a$ pixels. In the meanwhile, some background regions are discarded. Then, we further slice each big region into η^2 *small* patches (let $s = \eta^2$), each with the size of $a \times a$ pixels. Finally, we apply a feature extractor to all patches, leading to a bag of patch features.

After the preprocessing steps given above, we denote the final bag of patch features from one patient by

$$X \in \mathbb{R}^{m\times c} : \{x_j \in \mathbb{R}^c\}_{j=1}^{j=m}, \tag{2}$$

where m is the patch number in X, c is the dimensionality of patch features, and x_j denotes the jth patch feature. On one hand, this given X can be taken as individual patch features for our generator. On the other hand, it also can be viewed as the patch features with region partition for our discriminator, formulated as

$$X_r \in \mathbb{R}^{m\times c} : \{X_\tau \in \mathbb{R}^{s\times c}\}_{\tau=r_1}^{\tau=r_{m/s}}, \tag{3}$$

where $\{r_1, r_2, \dots, r_{m/s}\}$ means the m/s regions in X, τ is a region notation, and X_τ denotes the patch features of region τ. Note that m may be different across patients. Its common value is usually larger than 1000 when $f = 20$.

(2) *Notation convention*

We denote survival data by $\mathcal{D} = \left\{(X_i, t_i, \delta_i)\right\}_{i=1}^{i=N}$, where X_i, t_i, and δ_i represent the bag features, follow-up time, and censorship status of the ith patient, respectively. For the patients without censoring (*i.e.*, with event occurrence), we denote their data by $\mathcal{D}_e = \left\{(X_i, t_i) \mid \delta_i = 0 \text{ for } i = 1, 2, \dots, N\right\}$. Similarly, the data of other patients with censoring (*i.e.*, no event occurrence) are denoted as $\mathcal{D}_{ne} = \left\{(X_i, t_i) \mid \delta_i = 1 \text{ for } i = 1, 2, \dots, N\right\}$. For the ith patient with $\delta_i = 1$, we only know that its real time-to-event is strictly *later* than t_i.

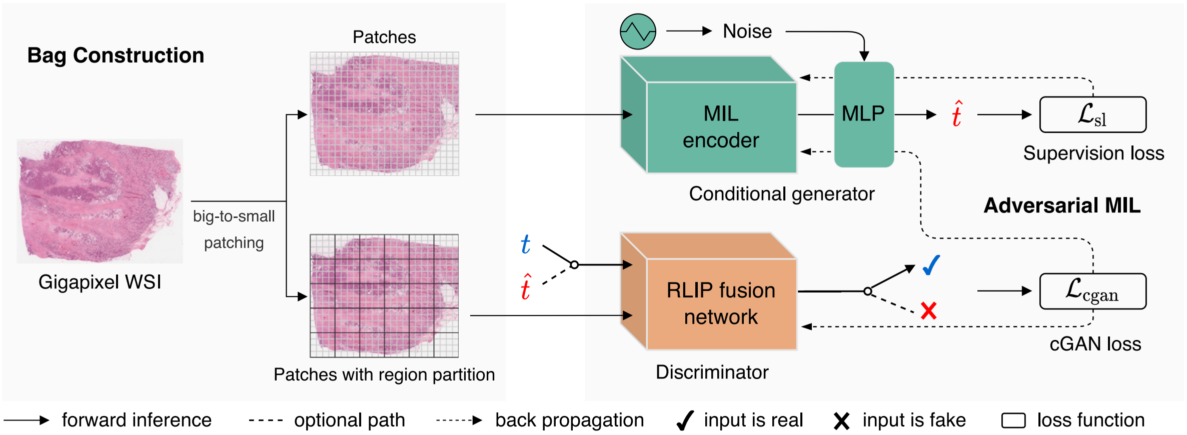


Fig. 3. AdvMIL overview. A gigapixel WSI is transformed into individual patches and the patches with equal region partition by big-to-small patching. A general MIL encoder is adopted as the backbone of conditional generator. An RLIP (region-level instance projection) fusion network is proposed to implement discriminator, elaborated in Fig. 4. Conditional generator is optimized by both $\mathcal{L}_{\text{cgan}}$ and $\mathcal{L}_{\text{sl}}$. Noise is given by a high-dimensional vector with certain distribution.

3.2. Adversarial multiple-instance learning

Next, we formulate our AdvMIL framework in the form of multiple instance learning.

(1) *Generator*

To make our framework adapt to MIL and compatible with the existing networks for WSI survival analysis, we implement a new generator with a general MIL encoder and an MLP (multi-layer perceptron) layer. This encoder is adopted to extract global bag representations. It could be any embedding-level MIL networks that output bag-level feature vectors, *e.g.*, cluster-, graph-, or sequence-based ones. The MLP layer is utilized to process both bag-level vectors and noise inputs, producing time-to-event estimations.

For a given X (the bag from one patient), we denote its time-to-event estimation by

$$\hat{t} = G(X, \mathcal{N}), \quad \mathcal{N} \sim P_n = \mathbf{U}(0, 1), \tag{4}$$

where $\mathcal{N}$ is a variable that represents the noise with uniform distribution, and G denotes a generator. In this way, G implicitly estimates the distribution of time-to-event via sampling from $\mathcal{N}$. The distribution output from G is expected to match the realistic conditional distribution of time-to-event on given X, *i.e.*, $P_{t|X}$. By the noise-outsourcing lemma (Kallenberg, 2002) from probability theory under minimal conditions, the existence of G can be guaranteed (Zhou et al., 2022), namely,

$$\exists\, G, s.t.\ \hat{t} = G(X, \mathcal{N}) \sim P_{t|X}. \tag{5}$$

Since $\mathcal{N}$ is an independent variable, Eq. (5) can be further written into its joint distribution form,

$$\big(X, G(X, \mathcal{N})\big) \sim P_{(X,t)}. \tag{6}$$

G can be determined by optimizing a cGAN (Mirza and Osindero, 2014) in which X is taken as a conditional input to G.

(2) *Discriminator*

The discriminator, denoted as D, aims to distinguish between the fake pair $(X, \hat{t})$, sampled from G, and the real pair (X, t), sampled from realistic data distribution. However, existing discriminators cannot be directly adopted to achieve this purpose, since D is required to efficiently process the fusion of the given X with an extremely-large matrix and the time t with a single real value. To tackle this problem, we propose a novel fusion network with region-level instance projection (RLIP), as shown in Fig. 4. It can deal with the fusion of a big matrix and a scalar value, inspired by projection discriminator (Miyato and Koyama, 2018). Moreover, our implementation of D would not bring too much additional computation costs to existing mainstream MIL models, as shown in .

Our RLIP fusion network is illustrated in Fig. 4. After big-to-small patching (described in Section 3.1 and shown in Fig. 3), the patch features with region partition, $\{X_\tau \in \mathbb{R}^{s \times c}\}_{\tau=r_1}^{\tau=r_{m/s}}$, are pooled within each region to form the region-level embeddings $X_{emb} : \{v_\tau \in \mathbb{R}^d\}_{\tau=r_1}^{\tau=r_{m/s}}$. This is implemented by a region embedding layer $\phi : \mathbb{R}^{s \times c} \to \mathbb{R}^d$. Then, a time embedding, denoted as $t_{emb} \in \mathbb{R}^d$, is computed from survival time $t \in \mathbb{R}$ by an MLP $\varphi : \mathbb{R} \to \mathbb{R}^d$. This time embedding and the region-level embeddings are fused through a *region-wise* inner product and a mean operation, represented by

$$y_{fusion} = \frac{1}{m/s} \sum_{v_\tau \in X_{emb}} v_\tau \cdot t_{emb}, \tag{7}$$

where $y_{fusion} \in \mathbb{R}$. Afterward, the region-level embeddings are fed into an output layer to yield a scalar output $y_{region} = \psi(\mathrm{gap}(X_{emb})) \in \mathbb{R}$, where $\mathrm{gap}(\cdot)$ and $\psi(\cdot)$ denote a global attention pooling function and a fully-connected layer, respectively. The final output of D is $y_D = \mathrm{sigmoid}(y_{fusion} + y_{region})$.

(3) *Network training*

As shown in Fig. 3, we use the two loss functions, cGAN loss and supervision loss, to optimize our network. cGAN loss is a general adversarial loss (Goodfellow et al., 2014a) that involves D and G, written as

$$\begin{aligned} \mathcal{L}_{\text{cgan}} = \min_G \max_D\ & \mathbb{E}_{(X,t)\sim P_{D_e}} \log D(X, t) \\ & + \mathbb{E}_{X\sim P_X,\ \mathcal{N}\sim P_n} \log\big[1 - D(X, G(X, \mathcal{N}))\big], \end{aligned} \tag{8}$$

where D is optimized to accurately recognize real or fake pairs, while G fools D by generating the time-to-event that is expected to better match realistic data distribution. Note that (X, t) is drawn from P_{D_e}, instead of P_D, in the first term of Eq. (8), because only the patients with event occurrence have the ground truth labels of time-to-event.

Like most cGANs, the convergence of $\mathcal{L}_{\text{cgan}}$ is guaranteed theoretically (Zhou et al., 2022). Nevertheless, in network training, we observe that such an adversarial network is very difficult to be optimized when only using $\mathcal{L}_{\text{cgan}}$. This problem is still open in adversarial learning (Goodfellow, 2016; Gui et al., 2021). To alleviate it, we additionally use an auxiliary supervision loss that fully utilizes time-to-event labels (Chapfuwa et al., 2018; Uemura et al., 2021), to train the network and speed up its convergence. This auxiliary loss is defined by

$$\begin{aligned} \mathcal{L}_{\text{sl}} = \min_G\ & \mathbb{E}_{(X,t)\sim P_{D_e},\ \mathcal{N}\sim P_n} \big|G(X, \mathcal{N}) - t\big| + \\ & \mathbb{E}_{(X,t)\sim P_{D_{ne}},\ \mathcal{N}\sim P_n} \mathrm{ReLU}\big(t - G(X, \mathcal{N})\big), \end{aligned} \tag{9}$$

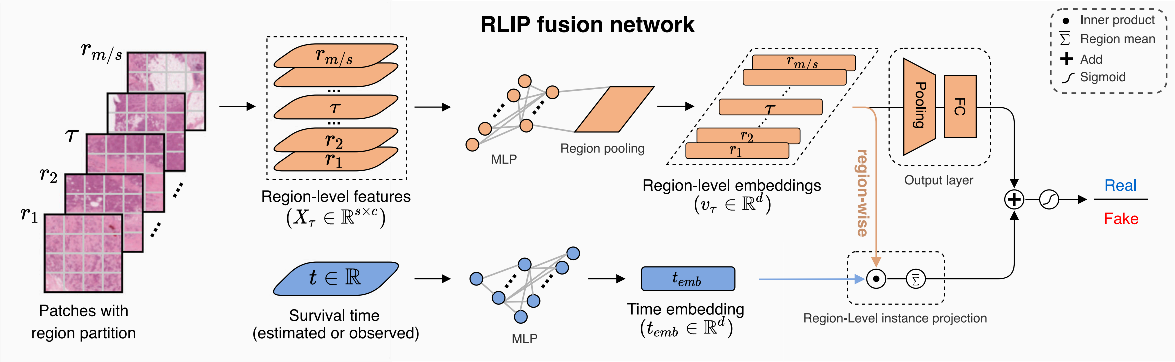

RLIP fusion network
$r_{m/s}$
τ
r_2
r_1
Patches with region partition
$r_{m/s}$
τ
r_2
r_1
Region-level features ($X_\tau \in \mathbb{R}^{s\times c}$)
MLP
Region pooling
$r_{m/s}$
τ
r_2
r_1
Region-level embeddings ($v_\tau \in \mathbb{R}^d$)
Pooling
FC
Output layer
Inner product
Region mean
Add
Sigmoid
region-wise
Real
Fake
$t \in \mathbb{R}$
Survival time (estimated or observed)
MLP
t_{emb}
Time embedding ($t_{emb} \in \mathbb{R}^d$)
Region-Level instance projection

Fig. 4. Our fusion network with region-level instance projection (RLIP). As the discriminator of AdvMIL, it comprises two key parts: WSI region embedding and region-level instance projection. The patches with region partition are obtained by big-to-small patching (refer to Section 3.1 and Fig. 3). For any region $\tau \in \{r_1, r_2, \dots, r_{m/s}\}$, X_τ represents its region-level features. s indicates the patch number in each region. v_τ is the region-level embedding of τ. MLP and FC mean multi-layer perceptron and fully-connected layer, respectively.

Algorithm 1: Mini-batch training with AdvMIL (one epoch).

```
Input: generator $G$, discriminator $D$, noise $\mathcal{N}$, batch size $b$,
       cGAN loss $\mathcal{L}_{\text{cgan}}$, supervision loss $\mathcal{L}_{\text{sl}}$, and training data
       $\mathcal{D}_{\text{train}} = \{(X_i, t_i, \delta_i)\}_{i=1}^{i=N_{\text{train}}}$.
load mini-batch samples, $B \leftarrow \{(X_i, t_i, \delta_i)\}_{i=1}^{i=b}$
// fix $G$, update $D$
$\mathcal{R} \leftarrow \{\}$
foreach $(X_i, t_i, \delta_i)$ in $B$ do
    if $\delta_i = 0$ then
        $\hat{y}_{real} \leftarrow D(X_i, t_i)$ // real pairs
        append $\hat{y}_{real}$ to $\mathcal{R}$
    end
    $\hat{y}_{fake} \leftarrow D(X_i, G(X_i, \mathcal{N}))$
    append $\hat{y}_{fake}$ to $\mathcal{R}$ // fake pairs
end
update $D$ by optimizing the $\mathcal{L}_{\text{cgan}}$ on $\mathcal{R}$
// fix $D$, update $G$
$\mathcal{R} \leftarrow \{\}$, $\mathcal{R}_{sl} \leftarrow \{\}$
foreach $(X_i, t_i, \delta_i)$ in $B$ do
    $\hat{t}_i \leftarrow G(X_i, \mathcal{N})$
    append $(\hat{t}_i, t_i)$ to $\mathcal{R}_{sl}$
    $\hat{y}_{fake} \leftarrow D(X_i, \hat{t}_i)$ // fake pairs
    append $\hat{y}_{fake}$ to $\mathcal{R}$
end
update $G$ by optimizing the $\mathcal{L}_{\text{cgan}}$ on $\mathcal{R}$ and the $\mathcal{L}_{\text{sl}}$ on $\mathcal{R}_{sl}$
```

where $\mathcal{L}_{\text{sl}}$ takes both censored and uncensored patients into account. Thus it could enable G to narrow its empirical errors simultaneously on $P_{\mathcal{D}_e}$ and $P_{\mathcal{D}_{ne}}$. Unlike other loss functions that are widely adopted in WSI survival analysis, $\mathcal{L}_{\text{sl}}$ no longer relies on any hazard function assumption or time pre-discretization, so our AdvMIL with $\mathcal{L}_{\text{sl}}$ can estimate time-to-event distribution via implicitly sampling from G.

The training procedure of AdvMIL, as shown in Algorithm 1, is very similar to that of vanilla GANs. Specifically, it can also be viewed as an adversarial min–max game between G and D. The key difference between AdvMIL and vanilla GANs lies in that AdvMIL training would involve censored patients ($i.e$, $\mathcal{D}_{ne}$) and auxiliary supervision loss ($i.e.$, $\mathcal{L}_{\text{sl}}$).

3.3. *k-fold semi-supervised learning*

Here we present how to perform semi-supervised learning with AdvMIL for WSI survival analysis, and then introduce our k-fold semi-supervised training strategy.

Algorithm 2: k-fold semi-supervised training with AdvMIL.

```
Input: labeled data $\mathcal{D}_{\text{l}} = \{(X_i, t_i, \delta_i)\}_{i=1}^{i=N_{\text{l}}}$, unlabeled data
       $\mathcal{D}_{\text{ul}} = \{(X_i)\}_{i=1}^{i=N_{\text{ul}}}$, the number of training epochs $\mathbb{T}$, the
       number of folds $k$.
randomly split $\mathcal{D}_{\text{ul}}$ into $k$ folds: $\mathcal{D}_{\text{ul}}^0, \mathcal{D}_{\text{ul}}^1, \dots, \mathcal{D}_{\text{ul}}^{k-1}$
for $T \leftarrow 0$ to $\mathbb{T} - 1$ do
    $\mathcal{D}_{\text{train}} \leftarrow \mathcal{D}_{\text{l}} + \mathcal{D}_{\text{ul}}^{T \bmod k}$
    mini-batch training on $\mathcal{D}_{\text{train}}$ using Algorithm 1
    // Algorithm 1 will skip the steps involving $t_i$
       or $\delta_i$ for unlabeled data
end
```

We feed both labeled data (denoted as $\mathcal{D}_{\text{l}}$) and unlabeled data (denoted as $\mathcal{D}_{\text{ul}}$) into AdvMIL for semi-supervised learning. Specifically, for the unlabeled samples in $\mathcal{D}_{\text{ul}}$, they will only be used by the second term of Eq. (8), *i.e.*, $\underset{G}{\min}\underset{D}{\max}\mathbb{E}_{X\sim P_X,\ \mathcal{N}\sim P_n} \log\left[1 - D(X, G(X, \mathcal{N}))\right]$, due to the absence of labels. Namely, unlabeled samples are utilized to infer time-to-event estimations through G, and then are further combined with these estimations to form fake pairs for optimizing the adversarial loss, as well as optimizing G and D.

Traditional semi-supervised learning usually directly utilizes all the unlabeled samples in $\mathcal{D}_{\text{ul}}$ for mini-batch training. However, when the ratio of $\mathcal{D}_{\text{ul}}$ is very high in training data, semi-supervised training would be dominated by those unlabeled samples, while often rarely focusing on the useful supervision signals from labeled samples. This problem could impair the efficiency of semi-supervised training, especially in a small data regime, *e.g.*, WSI data; because a small data regime implies fewer labeled samples and weaker supervision signals.

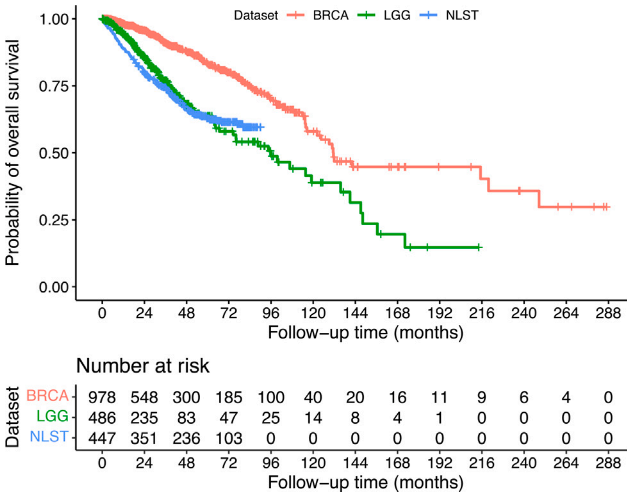
To decrease the influence of the problem above, we propose a new k-fold semi-supervised training strategy, shown in Algorithm 2. It splits unlabeled data $\mathcal{D}_{\text{ul}}$ into k folds with the same size, $\mathcal{D}_{\text{ul}}^0, \mathcal{D}_{\text{ul}}^1, \dots, \mathcal{D}_{\text{ul}}^{k-1}$. At the Tth training epoch, only $\mathcal{D}_{\text{ul}}^{\alpha}$ and labeled data $\mathcal{D}_{\text{l}}$ are chosen as training samples, where $\alpha = (T \bmod k)$. We could see that if $k = 1$, this k-fold semi-supervised training strategy exactly degenerates to traditional semi-supervised training.

4. Experiments and results

4.1. *Experimental settings*

(1) *Dataset description*

There are three publicly-available WSI datasets used in this study. They are National Lung Screening Trial (NLST) (National Lung Screening Trial Research Team, 2011), BReast CAncer (BRCA), and Low-Grade



Dataset	0	24	48	72	96	120	144	168	192	216	240	264	288
BRCA	978	548	300	185	100	40	20	16	11	9	6	4	0
LGG	486	235	83	47	25	14	8	4	1	0	0	0	0
NLST	447	351	236	103	0	0	0	0	0	0	0	0	0

Fig. 5. Survival curves of patient cohorts. It is obvious that breast cancer patients have better overall survival and longer follow-up time.

Table 1
Statistical details of three chosen WSI datasets.

Items	NLST	BRCA	LGG
Death ratio	35.9%	13.5%	23.4%
# Patients	447	978	486
# WSIs	1,222	1,043	836
# Patches	3,955,344	3,228,480	2,637,456
# Patches/WSI	3,236.8	3,095.4	3,154.9

Glioma (LGG), where both BRCA and LGG are from The Cancer Genome Atlas (TCGA) (Kandoth et al., 2013). Our criterion of dataset selection mainly considers: (i) the number of available patients, (ii) whether having been adopted in previous publications, and (iii) image quality. Finally, we collect a total of 3101 WSIs from 1911 patients without any subjective data curation. The overall survival curves of three datasets are shown in Fig. 5, and dataset details are presented in Table 1.

(2) *Chosen baselines*

The four methods, ABMIL (Ilse et al., 2018), DeepAttnMISL (Yao et al., 2020), PatchGCN (Chen et al., 2021), and ESAT (Shen et al., 2022), are chosen as our baselines, because (i) they are derived from three mainstream categories (*i.e.* , cluster, graph, and sequence) for end-to-end WSI modeling and (ii) they are the most representative ones of each category. We denote a baseline model by $\mathcal{M}$, and it could be one of ABMIL, DeepAttnMISL, PatchGCN, and ESAT. To validate the effectiveness of AdvMIL on four baselines, we compare three different models, as follows:

- $\mathcal{M}_{\text{origin}}$, which follows the implementation of $\mathcal{M}$, with original MIL encoder and survival loss function.
- $\mathcal{M}_{\text{base}}$, which employs the original MIL encoder of $\mathcal{M}$, with the new $\mathcal{L}_{\text{sl}}$ defined by Eq. (9).
- $\mathcal{M}$ + AdvMIL, which adopts a new AdvMIL scheme of a discriminator and a generator with the original MIL encoder of $\mathcal{M}$, and specially utilizes two new $\mathcal{L}_{\text{sl}}$ and $\mathcal{L}_{\text{cgan}}$ in adversarial learning.

(3) *Implementation details*

Bag construction. The magnification of each WSI is set to 20× (*i.e.*, $f = 20$). For big-to-small patching, we set the patch number in each region to 16 ($s = \eta^2 = 16$), and patch size to 256 × 256 pixels (*i.e.*, $a = 256$), so as to well balance computation efficiency and microscopic details. In addition, following CLAM (Lu et al., 2021), the feature extractor (with $c = 1024$) applied to patches is the official truncated ResNet-50 model (He et al., 2016), pre-trained on the ImageNet (Deng et al., 2009).

Table 2
Full training hyper-parameter settings.

Hyper-parameters	Values
Epoch number	300
Batch size	1
Gradient accumulation step	16
Early-stopping patience	30
Early-stopping warm-up	5
Learning rate for G	0.00008
Learning rate for D	0.00008
Optimizer	adam
Weight decay rate	0.0005

Generator. We empirically set the output dimensionality of MIL encoder to 384 for both sequence-based models and cluster-based ones, and 128 for graph-based ones (as it is efficient to train dense and large graphs). The MLP at the end of generator G only contains two layers. In each layer, when random noise $\mathcal{N}$ is added, its dimensionality is set to the same as input's. For simplicity, we use a binary code to represent whether random noise is added to a layer of MLP. Codes 0 and 1 indicate not adding and adding random noise to the corresponding layer of MLP, respectively. Thus, for adversarial learning, all the possible settings of noise adding are 0–1, 1–0, and 1–1 in our two-layer MLP. For example, 1–0 indicates that noise is added to the first layer but not to the second one, and so on.

Discriminator. For the RLIP fusion network D, we use an MLP and an average pooling layer to implement the region embedding layer ϕ. This ϕ can transform the region features with any number of patches into a single region-level feature vector, *i.e.*, ϕ is capable of working under various settings of patch regions (even if these regions have different patch numbers). This property indicates that region partition can be flexible and scalable. The feature dimensionality of X_{emb} and t_{emb} is $d = 128$. The time embedding layer φ, which implements $\mathbb{R} \rightarrow \mathbb{R}^{128}$, is an MLP with two layers. The fully-connected layer ψ implements $\mathbb{R}^{128} \rightarrow \mathbb{R}$.

Network training. The generator G is updated by a simplified version of the second term of $\mathcal{L}_{\text{cgan}}$, as follows:

$$\mathcal{L}_{\text{cgan}}(G) = \max_{G} \mathbb{E}_{X\sim P_X,\ \mathcal{N}\sim P_n} D\big(X, G(X,\mathcal{N})\big). \tag{10}$$

It often helps to train adversarial networks (Salimans et al., 2016). Our full training hyper-parameter settings are shown in Table 2. The learning rate of G decays by a factor of 0.5 if validation loss does not decrease within 10 epochs. Moreover, all these hyper-parameters always keep unchanged on three used WSI datasets. More experimental details can be seen in our publicly-available source codes.

(4) *Evaluation metrics*

We report the metrics frequently-used in survival analysis (Harrell et al., 1984), Concordance Index (C-Index). It measures the model ability of risk discrimination or ranking, just like the AUC in classification evaluation. We also evaluate the Mean Absolute Error (MAE) of estimated $\hat{t}$ on ground truth labels, as calculated in Eq. (9). During evaluation, for any patient we get its $\hat{t}$ by randomly sampling from generator for 30 times (as it is relatively time-consuming to infer once on gigapixel WSIs). Given limited sampling times, we use the median of $\hat{t}$ as the final estimation for performance evaluation, because it is often more robust to outliers than a mean.

Moreover, we use 5-fold Cross-Validation (CV) to evaluate each model. In training, the 20% samples of training set are randomly picked out to form a validation set for early stopping and model selection. Our data splitting is conducted at patient-level. All experiments run on a workstation with 2 × NVIDIA V100s (32G) GPU.

4.2. Results and analysis

(1) *Overall performance*

The sequence-based ESAT with AdvMIL. We first apply AdvMIL to sequence-based ESAT and test its performance, as ESAT is built

Table 3
Performance comparisons of different models. The binary code in bracket represents one specific setting of the random noise added into the two-layer MLP designed at the end of G. Codes 0 and 1 indicate not adding and adding random noises to the corresponding layer of MLP, respectively. Bold number indicates the best metric on a given dataset, and subscript number is a standard deviation of performance metrics.

	Model	C-Index ↑			MAE ↓		
		NLST	BRCA	LGG	NLST	BRCA	LGG
Sequence-based	ESAT$_{\text{origin}}$ [a]	$0.661_{0.039}$	$0.544_{0.034}$	$0.448_{0.037}$	$0.2054_{0.0263}$	$0.0529_{0.0108}$	$0.0663_{0.0087}$
	ESAT$_{\text{base}}$ [b]	$0.653_{0.047}$	$0.542_{0.063}$	$0.638_{0.091}$	$0.1920_{0.0261}$	$\mathbf{0.0344}_{0.0056}$	$\mathbf{0.0518}_{0.0060}$
	ESAT + AdvMIL (0–1)	$\mathbf{0.672}_{0.048}$	$0.545_{0.065}$	$0.621_{0.063}$	$0.1871_{0.0203}$	$0.0383_{0.0081}$	$0.0526_{0.0058}$
	ESAT + AdvMIL (1–0)	$0.649_{0.039}$	$\mathbf{0.562}_{0.067}$	$\mathbf{0.642}_{0.076}$	$0.1995_{0.0240}$	$0.0349_{0.0065}$	$0.0522_{0.0049}$
	ESAT + AdvMIL (1–1)	$0.660_{0.042}$	$0.545_{0.075}$	$0.634_{0.086}$	$\mathbf{0.1849}_{0.0153}$	$0.0366_{0.0065}$	$0.0523_{0.0051}$
	ABMIL$_{\text{base}}$ [b]	$0.525_{0.095}$	$0.502_{0.051}$	$0.493_{0.046}$	$0.2280_{0.0333}$	$\mathbf{0.0387}_{0.0051}$	$0.0651_{0.0048}$
	ABMIL + AdvMIL (0–1)	$0.522_{0.069}$	$\mathbf{0.566}_{0.035}$	$0.494_{0.023}$	$0.2285_{0.0375}$	$0.0433_{0.0070}$	$0.0692_{0.0057}$
	ABMIL + AdvMIL (1–0)	$0.531_{0.090}$	$0.523_{0.057}$	$\mathbf{0.535}_{0.029}$	$\mathbf{0.2261}_{0.0361}$	$0.0397_{0.0060}$	$\mathbf{0.0596}_{0.0095}$
	ABMIL + AdvMIL (1–1)	$\mathbf{0.544}_{0.084}$	$0.550_{0.073}$	$0.511_{0.065}$	$0.2277_{0.0390}$	$0.0419_{0.0059}$	$0.0655_{0.0049}$
Cluster-based	DeepAttnMISL$_{\text{origin}}$ [a]	$\mathbf{0.548}_{0.081}$	$0.496_{0.048}$	$\mathbf{0.593}_{0.062}$	–	–	–
	DeepAttnMISL$_{\text{base}}$ [b]	$0.543_{0.072}$	$0.508_{0.049}$	$0.517_{0.080}$	$0.2262_{0.0343}$	$\mathbf{0.0398}_{0.0060}$	$0.0703_{0.0069}$
	DeepAttnMISL + AdvMIL (0–1)	$0.528_{0.051}$	$0.498_{0.036}$	$0.452_{0.072}$	$0.2456_{0.0356}$	$0.0450_{0.0066}$	$0.0707_{0.0069}$
	DeepAttnMISL + AdvMIL (1–0)	$0.531_{0.068}$	$0.510_{0.075}$	$0.510_{0.062}$	$0.2260_{0.0334}$	$0.0414_{0.0055}$	$\mathbf{0.0640}_{0.0028}$
	DeepAttnMISL + AdvMIL (1–1)	$0.538_{0.048}$	$\mathbf{0.552}_{0.029}$	$0.506_{0.054}$	$\mathbf{0.2258}_{0.0325}$	$0.0421_{0.0053}$	$0.0664_{0.0042}$
Graph-based	PatchGCN$_{\text{origin}}$ [a]	$0.605_{0.024}$	$0.542_{0.072}$	$0.585_{0.055}$	–	–	–
	PatchGCN$_{\text{base}}$ [b]	$0.579_{0.080}$	$0.518_{0.049}$	$0.592_{0.071}$	$0.2131_{0.0375}$	$\mathbf{0.0405}_{0.0046}$	$0.0634_{0.0052}$
	PatchGCN + AdvMIL (0–1)	$0.613_{0.047}$	$\mathbf{0.562}_{0.094}$	$0.560_{0.073}$	$0.1944_{0.0266}$	$0.0408_{0.0040}$	$0.0590_{0.0056}$
	PatchGCN + AdvMIL (1–0)	$0.580_{0.060}$	$0.556_{0.063}$	$0.561_{0.101}$	$0.2080_{0.0327}$	$0.0478_{0.0081}$	$0.0561_{0.0054}$
	PatchGCN + AdvMIL (1–1)	$\mathbf{0.644}_{0.050}$	$0.535_{0.091}$	$\mathbf{0.593}_{0.065}$	$\mathbf{0.1895}_{0.0175}$	$0.0427_{0.0085}$	$\mathbf{0.0529}_{0.0043}$

[a] We adopt their original survival loss functions to report these results. The ABMIL with its original survival loss function is not reported, because ABMIL is originally proposed for classification. MAE is not reported for original DeepAttnMISL and PatchGCN since they do not provide the prediction of continuous survival time.

[b] We adopt the survival loss function given by Eq. (9) to report these results.

upon the Transformer (Vaswani et al., 2017) that has demonstrated remarkable success in various applications. Compared with ESAT$_{\text{base}}$, the model combining ESAT and AdvMIL is additionally optimized by the adversarial loss $\mathcal{L}_{\text{cgan}}$.

As shown in Table 3, we can see that (1) ESAT + AdvMIL could often obtain better performances than original ESAT in both C-Index and MAE, and its improvement on LGG is evident; (2) the model combining ESAT and AdvMIL outperforms its counterpart ESAT$_{\text{base}}$, by a C-Index improvement of 2.91%, 3.69%, and 0.63% on NLST, BRCA, and LGG, respectively; (3) Applying AdvMIL to ESAT decreases MAE by 3.70% on NLST, and slightly increases MAE by around +1% on the other two datasets. These observations suggest that the proposed AdvMIL has competitive advantages over sequence-based ESAT, especially in terms of C-Index. Intuitively, by adversarial learning $\mathcal{L}_{\text{cgan}}$ could help to optimize G, and then make G become more robust to the input space extended and smoothed by noise $\mathcal{N}$ (Miyato et al., 2018). In addition, it is observed that noise setting also affects model performance. This implies that a proper noise setting (the focus in adversarial training (Goodfellow et al., 2014b)) may help achieving better performance. More discussions on this can be seen in Section 4.3.

Other mainstream MIL networks with AdvMIL. We test the adaptability of AdvMIL to other mainstream MIL networks (with distinct backbones), and experimental results are shown in Table 3. From these results, we can see that (1) ABMIL + AdvMIL almost always outperforms its counterpart ABMIL$_{\text{base}}$ in terms of both C-Index and MAE, except the MAE on BRCA; (2) DeepAttnMISL + AdvMIL is obviously better in C-Index than its two baselines on BRCA, while performs worse on NLST and LGG; (3) After applying AdvMIL to PatchGCN, the combined model is comparable or even better than PatchGCN$_{\text{origin}}$ and PatchGCN$_{\text{base}}$ in most cases.

These empirical results from Table 3 suggest that AdvMIL could often help boosting the performance of mainstream MIL networks. Moreover, by comparing the results of four representative MIL networks, we notice that ESAT + AdvMIL achieves the best overall performance across three datasets.

It is worth noting that MAE is more sensitive to the value of $\hat{t}$ than C-Index since C-Index measures the quality of estimation ranking—it only considers relative errors in measurement. Given the fact that limited estimations are drawn from G for evaluation, the final estimation of $\hat{t}$ may be more likely to stray from the expectation of distribution, *i.e.*, $\mathbb{E}_{\mathcal{N}\sim P_n} P(t|X,\mathcal{N})$, resulting in worse MAEs, as shown in Table 3.

Table 4
Model efficiency in terms of # Params and # MACs.

Network		# Params	# MACs
D (ours)		**206.34K**	**452.08M**
G	ABMIL	985.54K	2.31G
	DeepAttnMISL	985.54K	1.33G
	PatchGCN	280.51K	662.77M
	ESAT	1.73M	1.61G

(2) *Computational overhead analysis*

We evaluate the computation efficiency of discriminator in model size and inference cost, to verify if the proposed discriminator will bring substantial computational overheads to mainstream MIL networks. Specifically, we measure the number of trainable parameters (# Params) for all used models. Moreover, we assess the theoretical amount of Multiply–Accumulate operations (# MACs) in models, by randomly selecting a patient from NLST to infer the model once. This patient has a total of 3360 patches in its WSI bag. A Python toolkit, *ptflops*, is adopted to calculate # Params and # MACs.

As shown in Table 4, our discriminator only increases extra 206.34K trainable parameters and 452.08M MACs, obviously smaller than four generators. These results imply that it is acceptable to introduce such a discriminator into existing mainstream MIL networks, in terms of computational overhead.

(3) *k-fold semi-supervised training*

We further investigate the potential of adopting AdvMIL in semi-supervised learning. This is not studied in both of the two classical works of adversarial survival analysis, *i.e.*, DATE (Chapfuwa et al., 2018) and pix2surv (Uemura et al., 2021). For simplicity, we denote training set and test set by $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$, respectively. To conduct this experiment, we prepare unlabeled data at first. Specifically, we randomly select some samples from $\mathcal{D}_{\text{train}}$ and mask their labels. These selected samples are label-unavailable in training, called unlabeled data and denoted as $\mathcal{D}_{\text{ul}}$. The remaining samples are labeled data, denoted as $\mathcal{D}_{\text{l}} = \mathcal{D}_{\text{train}} \setminus \mathcal{D}_{\text{ul}}$. Their labels are available in training. We set the ratio of labeled data $\frac{\mathcal{D}_{\text{l}}}{\mathcal{D}_{\text{train}}}$ to 0.2, 0.4, 0.6, 0.8 and 0.9. Then, we test and

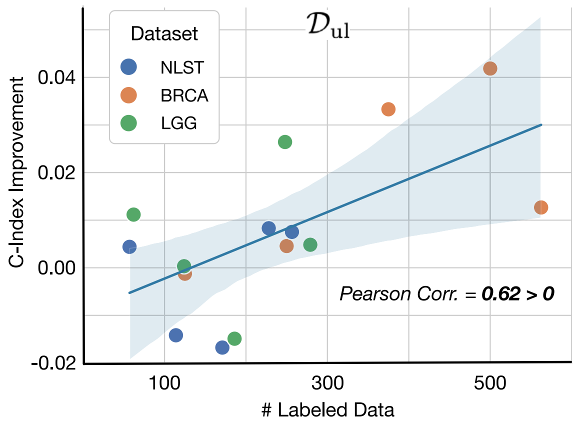


Fig. 6. Performance improvement of k-fold semi-supervised models over fully-supervised ones on $\mathcal{D}_{\mathrm{ul}}$. A linear regression of all points is plotted, with a 95% confidence region. "Corr". means "Correlation".

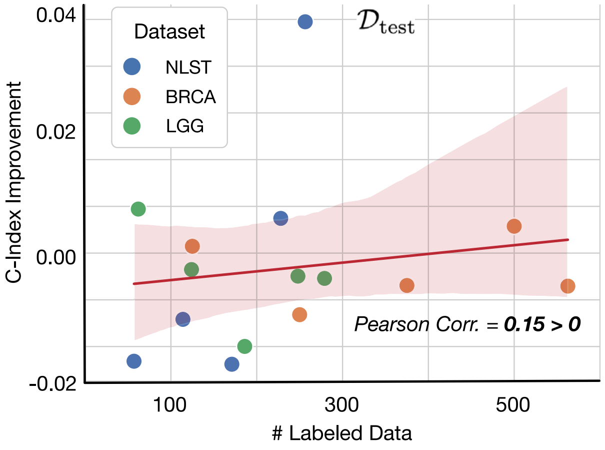


Fig. 7. Performance improvement of k-fold semi-supervised models over fully-supervised ones on $\mathcal{D}_{\mathrm{test}}$.

compare semi-supervised AdvMIL models and fully-supervised ones. (I) On one hand, we feed both $\mathcal{D}_{\mathrm{l}}$ and $\mathcal{D}_{\mathrm{ul}}$ into AdvMIL for *k-fold semi-supervised* training (refer to Algorithm 2), and test the performance of semi-supervised AdvMIL models on $\mathcal{D}_{\mathrm{ul}}$ and $\mathcal{D}_{\mathrm{test}}$. (II) On the other hand, we only feed $\mathcal{D}_{\mathrm{l}}$ into AdvMIL for *fully-supervised* training, as the baseline for comparisons. We also test the performance of fully-supervised AdvMIL models on $\mathcal{D}_{\mathrm{ul}}$ and $\mathcal{D}_{\mathrm{test}}$. Except training data, other settings keep the same. The network ESAT + AdvMIL is adopted in this experiment, since it can achieve the best overall performance across three datasets (see Table 3).

We calculate the performance improvement of semi-supervised AdvMIL models over fully-supervised ones. Then we visualize this improvement on the vertical axis of 2D plane, along with the size of $\mathcal{D}_{\mathrm{l}}$ on the horizontal axis. A total of 15 points are obtained. The results on $\mathcal{D}_{\mathrm{ul}}$ and $\mathcal{D}_{\mathrm{test}}$ are shown in Fig. 6 and Fig. 7, respectively. Full numerical results can be found in Table 5.

From these results, we can see that (1) on $\mathcal{D}_{\mathrm{ul}}$, the improvements of C-Index are often positive (10 out of 15 points) and they tend to become greater with increasing the size of $\mathcal{D}_{\mathrm{l}}$ (Pearson Corr. = 0.62); (2) on $\mathcal{D}_{\mathrm{test}}$, the improvements of C-Index are positive in a few cases (5 out of 15 points), and meanwhile they have a weakly-positive correlation with the size of $\mathcal{D}_{\mathrm{l}}$ (Pearson Corr. = 0.15). These findings mean that AdvMIL could be effective for predicting the unlabeled data *used* in training, while often losing its effectiveness for predicting those *unused* in training. Namely, AdvMIL could effectively utilize unlabeled WSIs in semi-supervised learning and often obtain superior performance on these unlabeled data; whereas the previous studies on WSI survival analysis have not paid enough attention on this. Moreover, increasing the size of labeled training data set would help to enlarge the C-Index improvement on unlabeled data $\mathcal{D}_{\mathrm{ul}}$.

In addition, from Table 5, we empirically find that our k-fold strategy ($k > 1$) could often lead to better results than the traditional one without fold splitting ($i.e, k = 1$), especially when the number of labeled data is small, namely when $\frac{\mathcal{D}_{\mathrm{l}}}{\mathcal{D}_{\mathrm{train}}}$ is 0.2, 0.4 or 0.6. These results confirm the effectiveness of our k-fold semi-supervised training strategy.

There are many studies that demonstrate GAN to be a promising and effective means for semi-supervised learning (Springenberg, 2015; Salimans et al., 2016; Miyato et al., 2018; Li et al., 2021; Gui et al., 2021). Our experiments also verify this, even when it is a small data regime (only around 500 patients in our task) in the field of WSI survival analysis.

4.3. Ablation study and analysis

(1) *Study on region-level instance projection*

We validate the region-level instance projection (RLIP) strategy that is proposed for the feature fusion in D, and compare it with a regular strategy, WSI-level projection, named *Projection*. This regular strategy directly projects a conditional vector on the global feature of WSI. It skips the smooth transition (used in RLIP) from patch to region and region to WSI. Note that the *Projection*, which makes a simple and direct fusion on WSI and survival time, is a strategy having not appeared in previous models. It is just used as a baseline for comparisons.

From the results in Table 6, we can summarize that (1) RLIP almost always gains C-Index rises (the highest one is +5.36%) on three datasets; (2) in most cases RLIP could decrease MAEs by large margins (2.61%–9.93%), and only in a few cases, the MAEs of RLIP increase by no larger than 2.13%; (3) when setting noise $\mathcal{N}$ to 1–1, RLIP can completely surpass *Projection* on three datasets in terms of both C-Index and MAE. These empirical results demonstrate the effectiveness of our region-level strategy, and suggest that an early feature fusion, *i.e.*, region-level projection rather than a direct WSI-level projection, may be more likely to better incorporate conditional information and help adversarial multiple instance learning.

(2) *Study on noise types*

We further test the effect of noise type on model performance. Two widely-used distributions, *Uniform* and *Gaussian*, are tested. From the results shown in Table 7, we can see that Uniform distribution could often obtain better performances. Especially in terms of MAE, Uniform ones always have obvious advantages over Gaussian ones. One possible reason is that when sampling $\mathcal{N}$ from Gaussian distribution, the estimation of time-to-event distribution via implicitly sampling would be more concentrated and it thereby would tend to enlarge the bias of predictions.

In addition, as mentioned in Section 4.2, we argue that the noise, which is combined together with conditional input for prediction, may be a critical factor affecting model performance. Some techniques like adversarial training (Goodfellow et al., 2014b; Goodfellow, 2016) may be a promising means to validate this argument. We leave this in the future work, since adversarial training is yet another research topic and it is not the focus of this paper.

4.4. Robustness analysis

We carry out experiments on three used WSI datasets to validate the robustness of AdvMIL-based models. Next, we show how we set up experiments for robustness analysis, and then present experimental results and findings.

Specifically, we evaluate the model's robustness to patch-level transformations, including image occlusion, Gaussian blurring and HED

Table 5
k-fold semi-supervised training with AdvMIL. Only using $\mathcal{D}_\mathrm{l}$ as training data indicates fully-supervised training. Using both $\mathcal{D}_\mathrm{l}$ and $\mathcal{D}_\mathrm{ul}$ as training data indicates semi-supervised training. $k = 1$ means the traditional semi-supervised training without the fold splitting on $\mathcal{D}_\mathrm{ul}$. We train ESAT + AdvMIL on training data and measure its C-Index performances on $\mathcal{D}_\mathrm{ul}$ and $\mathcal{D}_\mathrm{test}$.

$\mathcal{D}_\mathrm{l}$ / $\mathcal{D}_\mathrm{train}$	Training data	$\mathcal{D}_\mathrm{ul}$ (C-Index)			$\mathcal{D}_\mathrm{test}$ (C-Index)		
		NLST	BRCA	LGG	NLST	BRCA	LGG
0.2	$\mathcal{D}_\mathrm{l}$	$0.616_{0.047}$	$\mathbf{0.553}_{0.029}$	$0.619_{0.034}$	$\mathbf{0.623}_{0.082}$	$0.530_{0.062}$	$0.602_{0.055}$
	$\mathcal{D}_\mathrm{l} + \mathcal{D}_\mathrm{ul}$ $(k = 1)$	$0.578_{0.053}$	$0.545_{0.043}$	$0.622_{0.022}$	$0.576_{0.055}$	$\mathbf{0.540}_{0.093}$	$0.608_{0.041}$
	$\mathcal{D}_\mathrm{l} + \mathcal{D}_\mathrm{ul}$ $(k = 3)$	$0.597_{0.049}$	$0.542_{0.047}$	$0.621_{0.031}$	$0.616_{0.077}$	$0.518_{0.097}$	$0.592_{0.059}$
	$\mathcal{D}_\mathrm{l} + \mathcal{D}_\mathrm{ul}$ $(k = 5)$	$\mathbf{0.620}_{0.047}$	$0.552_{0.049}$	$\mathbf{0.630}_{0.019}$	$0.610_{0.065}$	$0.528_{0.080}$	$\mathbf{0.610}_{0.048}$
0.4	$\mathcal{D}_\mathrm{l}$	$\mathbf{0.632}_{0.047}$	$0.574_{0.042}$	$0.603_{0.053}$	$\mathbf{0.622}_{0.060}$	$\mathbf{0.579}_{0.135}$	$\mathbf{0.597}_{0.066}$
	$\mathcal{D}_\mathrm{l} + \mathcal{D}_\mathrm{ul}$ $(k = 1)$	$0.602_{0.027}$	$0.524_{0.052}$	$0.595_{0.090}$	$0.585_{0.065}$	$0.527_{0.085}$	$0.580_{0.099}$
	$\mathcal{D}_\mathrm{l} + \mathcal{D}_\mathrm{ul}$ $(k = 3)$	$0.618_{0.051}$	$0.549_{0.080}$	$\mathbf{0.604}_{0.062}$	$0.603_{0.054}$	$0.536_{0.069}$	$\mathbf{0.597}_{0.076}$
	$\mathcal{D}_\mathrm{l} + \mathcal{D}_\mathrm{ul}$ $(k = 5)$	$0.606_{0.050}$	$\mathbf{0.578}_{0.046}$	$0.592_{0.085}$	$0.586_{0.077}$	$0.576_{0.037}$	$0.595_{0.079}$
0.6	$\mathcal{D}_\mathrm{l}$	$\mathbf{0.637}_{0.054}$	$0.532_{0.058}$	$\mathbf{0.623}_{0.046}$	$\mathbf{0.646}_{0.043}$	$0.565_{0.097}$	$\mathbf{0.649}_{0.045}$
	$\mathcal{D}_\mathrm{l} + \mathcal{D}_\mathrm{ul}$ $(k = 1)$	$0.612_{0.032}$	$0.556_{0.030}$	$0.606_{0.027}$	$0.610_{0.083}$	$0.563_{0.092}$	$0.639_{0.045}$
	$\mathcal{D}_\mathrm{l} + \mathcal{D}_\mathrm{ul}$ $(k = 3)$	$0.618_{0.059}$	$0.535_{0.060}$	$0.601_{0.055}$	$0.645_{0.040}$	$0.533_{0.063}$	$0.637_{0.063}$
	$\mathcal{D}_\mathrm{l} + \mathcal{D}_\mathrm{ul}$ $(k = 5)$	$0.620_{0.057}$	$\mathbf{0.566}_{0.060}$	$0.608_{0.031}$	$0.632_{0.059}$	$\mathbf{0.569}_{0.097}$	$0.625_{0.046}$
0.8	$\mathcal{D}_\mathrm{l}$	$0.629_{0.044}$	$0.517_{0.062}$	$0.609_{0.048}$	$0.628_{0.062}$	$0.538_{0.089}$	$0.630_{0.082}$
	$\mathcal{D}_\mathrm{l} + \mathcal{D}_\mathrm{ul}$ $(k = 1)$	$\mathbf{0.638}_{0.049}$	$0.524_{0.065}$	$0.590_{0.039}$	$0.643_{0.048}$	$0.552_{0.072}$	$0.614_{0.080}$
	$\mathcal{D}_\mathrm{l} + \mathcal{D}_\mathrm{ul}$ $(k = 3)$	$0.623_{0.075}$	$0.511_{0.086}$	$\mathbf{0.635}_{0.092}$	$0.642_{0.048}$	$\mathbf{0.567}_{0.093}$	$0.630_{0.085}$
	$\mathcal{D}_\mathrm{l} + \mathcal{D}_\mathrm{ul}$ $(k = 5)$	$0.622_{0.032}$	$\mathbf{0.559}_{0.069}$	$0.615_{0.038}$	$\mathbf{0.646}_{0.046}$	$0.554_{0.085}$	$\mathbf{0.634}_{0.093}$
0.9	$\mathcal{D}_\mathrm{l}$	$0.713_{0.047}$	$0.470_{0.139}$	$0.552_{0.134}$	$0.661_{0.038}$	$0.556_{0.055}$	$0.637_{0.085}$
	$\mathcal{D}_\mathrm{l} + \mathcal{D}_\mathrm{ul}$ $(k = 1)$	$0.716_{0.037}$	$\mathbf{0.482}_{0.187}$	$0.550_{0.115}$	$0.650_{0.062}$	$\mathbf{0.559}_{0.070}$	$\mathbf{0.641}_{0.066}$
	$\mathcal{D}_\mathrm{l} + \mathcal{D}_\mathrm{ul}$ $(k = 3)$	$\mathbf{0.720}_{0.069}$	$0.449_{0.202}$	$\mathbf{0.557}_{0.117}$	$0.654_{0.055}$	$0.555_{0.095}$	$0.619_{0.050}$
	$\mathcal{D}_\mathrm{l} + \mathcal{D}_\mathrm{ul}$ $(k = 5)$	$0.699_{0.048}$	$0.380_{0.154}$	$0.545_{0.104}$	$\mathbf{0.664}_{0.051}$	$0.521_{0.074}$	$0.637_{0.060}$

Table 6
Ablation study on region-level instance projection (RLIP). Δ denotes the improvement of RLIP over regular projection. ESAT + AdvMIL is used.

Noise	Fusion operation	C-Index ↑			MAE ↓		
		NLST	BRCA	LGG	NLST	BRCA	LGG
0–1	Projection	$0.658_{0.047}$	$0.543_{0.072}$	$\mathbf{0.624}_{0.084}$	$0.1983_{0.0228}$	$\mathbf{0.0375}_{0.0082}$	$\mathbf{0.0519}_{0.0064}$
	RLIP	$\mathbf{0.672}_{0.048}$	$\mathbf{0.545}_{0.065}$	$0.621_{0.063}$	$\mathbf{0.1871}_{0.0203}$	$0.0383_{0.0081}$	$0.0526_{0.0058}$
	Δ	+ 2.13%	+ 0.37%	−0.48%	−5.65%	+ 2.13%	+ 1.35%
1–0	Projection	$0.616_{0.052}$	$0.548_{0.084}$	$0.641_{0.079}$	$0.2015_{0.0243}$	$\mathbf{0.0344}_{0.0063}$	$\mathbf{0.0518}_{0.0053}$
	RLIP	$\mathbf{0.649}_{0.039}$	$\mathbf{0.562}_{0.067}$	$\mathbf{0.642}_{0.076}$	$\mathbf{0.1995}_{0.0240}$	$0.0349_{0.0065}$	$0.0522_{0.0049}$
	Δ	+ 5.36%	+ 2.55%	+ 0.16%	−9.93%	+ 1.45%	+ 0.77%
1–1	Projection	$0.652_{0.032}$	$0.531_{0.080}$	$0.616_{0.082}$	$0.1996_{0.0213}$	$0.0389_{0.0054}$	$0.0537_{0.0049}$
	RLIP	$\mathbf{0.660}_{0.042}$	$\mathbf{0.545}_{0.075}$	$\mathbf{0.634}_{0.086}$	$\mathbf{0.1849}_{0.0153}$	$\mathbf{0.0366}_{0.0065}$	$\mathbf{0.0523}_{0.0051}$
	Δ	+ 1.23%	+ 2.64%	+ 2.92%	−7.36%	−5.91%	−2.61%

Table 7
Noise effect on the performance of ESAT + AdvMIL. The noise with *Uniform* distribution could often perform better than that with *Gaussian* distribution.

$\mathcal{N}$		C-Index ↑			MAE ↓		
		NLST	BRCA	LGG	NLST	BRCA	LGG
Uniform (0, 1)	0–1	$\mathbf{0.672}_{0.048}$	$0.545_{0.065}$	$0.621_{0.063}$	$0.1871_{0.0203}$	$0.0383_{0.0081}$	$0.0526_{0.0058}$
	1–0	$0.649_{0.039}$	$\mathbf{0.562}_{0.067}$	$\mathbf{0.642}_{0.076}$	$0.1995_{0.0240}$	$\mathbf{0.0349}_{0.0065}$	$\mathbf{0.0522}_{0.0049}$
	1–1	$0.660_{0.042}$	$0.545_{0.075}$	$0.634_{0.086}$	$\mathbf{0.1849}_{0.0153}$	$0.0366_{0.0065}$	$0.0523_{0.0051}$
Gaussian (0, 1)	0–1	$0.632_{0.039}$	$0.508_{0.102}$	$0.631_{0.090}$	$0.1928_{0.0211}$	$0.0392_{0.0063}$	$0.0567_{0.0080}$
	1–0	$\mathbf{0.634}_{0.042}$	$0.461_{0.078}$	$0.624_{0.050}$	$0.2011_{0.0264}$	$\mathbf{0.0363}_{0.0060}$	$\mathbf{0.0529}_{0.0033}$
	1–1	$0.590_{0.072}$	$\mathbf{0.554}_{0.047}$	$\mathbf{0.646}_{0.051}$	$\mathbf{0.2019}_{0.0223}$	$0.0403_{0.0078}$	$0.0556_{0.0071}$

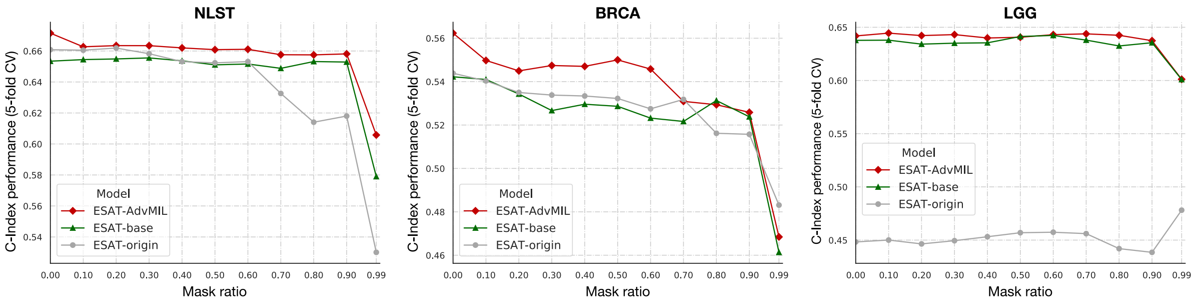


Fig. 8. Robustness against patch occlusion. A mask ratio of 0.99 indicates that there is only one region (with a size of 4096 × 4096 pixels at the highest magnification) reserved in each test WSI.

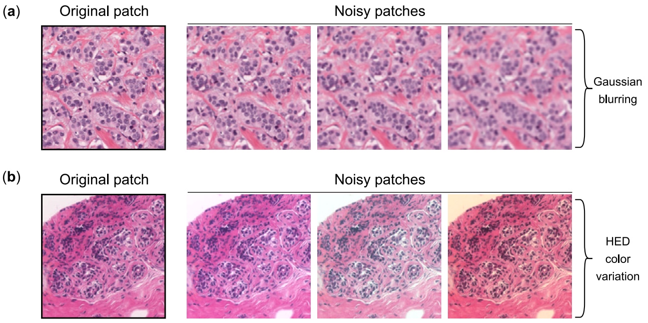


Fig. 9. Examples of (a) Gaussian blurring and (b) HED color variation.

Table 8
Robustness against patch image noises. C-Index is reported on the test set of the first fold.

Dataset	Model	Image noise		
		–	Gau. blur	HED var.
NLST	$\text{ESAT}_{\text{origin}}$	0.642	0.661	0.667
	$\text{ESAT}_{\text{base}}$	0.663	0.672	0.675
	ESAT + AdvMIL	**0.682**	**0.680**	**0.684**
BRCA	$\text{ESAT}_{\text{origin}}$	0.543	0.589	0.588
	$\text{ESAT}_{\text{base}}$	0.601	0.546	0.557
	ESAT + AdvMIL	**0.602**	**0.620**	**0.632**
LGG	$\text{ESAT}_{\text{origin}}$	0.468	0.492	0.498
	$\text{ESAT}_{\text{base}}$	0.554	0.579	0.582
	ESAT + AdvMIL	**0.604**	**0.584**	**0.597**

(Hematoxylin-Eosin-DAB) color space perturbation, in view of the following observations. (I) It is really challenging to assess model robustness by directly transforming an entire gigapixel WSI into various views, unlike the straightforward operations on most natural images with common sizes. (II) Patch occlusion has been utilized as an experimental scenario by previous works (Naseer et al., 2021; Chen et al., 2022c) to study the robustness of prevalent computer vision models. (III) Gaussian blurring is a classical transformation for natural images to simulate out-of-focus artifacts, also widely utilized for histological WSIs (Tellez et al., 2019; Cheng et al., 2021). (IV) HED color perturbation is specifically designed for H&E images (Tellez et al., 2018) to simulate a color variation routine, which is frequently adopted in computational pathology (Tellez et al., 2019) and could be more suitable for our applications.

(1) *Robustness to patch occlusion*

We randomly mask the patches of each WSI (in test sets), according to various mask ratios. A mask ratio of 0.99 indicates that there is only one region (with a size of 4096 × 4096 pixels at the highest magnification) reserved in each test WSI. The models, which have been fitted on complete training data previously, are adopted for test. The C-Index averaged on 5 folds is reported and shown in Fig. 8.

From Fig. 8, we can observe that AdvMIL-based models almost always outperform two baselines on various occlusion levels across three datasets. Moreover, it is found that ESAT + AdvMIL has more stable C-Index performance than both $\text{ESAT}_{\text{origin}}$ and $\text{ESAT}_{\text{base}}$ on NLST and LGG, when mask ratio is larger than 0.5. These experimental results demonstrate that our adversarial survival analysis models could often be more robust against occlusion than traditional ones on WSIs. In other words, the occlusion robustness of traditional WSI survival analysis models could be further enhanced via our adversarial MIL scheme.

Besides, we interestingly see that some meaningful results are still achieved by AdvMIL-based models (C-Index ≈ 0.6) on NLST and LGG, even when mask ratio is 0.99, *i.e.*, only reserving one region in each test WSI. The underlying reasons behind this phenomenon could have three-fold: (i) the information abundance and sparsity inherent in histological WSIs (Zarella et al., 2018); (ii) the intrinsic robustness of Transformer-based models, demonstrated in Naseer et al. (2021) and Laleh et al. (2022); (iii) the robustness further improved by our AdvMIL, seen from the improvement of AdvMIL-based ESAT over its non-adversarial counterpart $\text{ESAT}_{\text{base}}$.

(2) *Robustness to image noises*

We apply two representative noises to patch images, *i.e.*, Gaussian blurring (Gau. blur) and HED color variation (HED var.). Some examples of them are shown in Fig. 9. Similarly, those trained models are adopted for testing on noisy patch images. We choose the first fold in each dataset for experiments, as there is a huge total number of patch images (nearly 10,000,000 images of 256 × 256 pixels) to be processed in the 3101 gigapixel WSIs used by us.

From the experimental results shown in Table 8, we clearly see that AdvMIL-based models always obtain the best C-Index across three datasets, after applying image noises. Moreover, we also see a more stable performance on AdvMIL-based models in this experiment. For example, the biggest changes in C-Index of ESAT + AdvMIL are 0.004, 0.03, and 0.02 on NLST, BRCA, and LGG, respectively; while those of $\text{ESAT}_{\text{base}}$ are 0.012, 0.055, and 0.028, apparently larger than those of AdvMIL-based models. $\text{ESAT}_{\text{origin}}$ also shows similar results, except its nonsense results (C-Index ≤ 0.5) on LGG. And most notably, on BRCA, the C-Index of $\text{ESAT}_{\text{base}}$ decreases by 0.055 and 0.044 after applying Gaussian blurring and HED color variation, respectively; while that of ESAT + AdvMIL still keeps its superiority without any decline.

These empirical observations indicate the potential advantage of adversarial MIL models in robustness over traditional ones. One explicit and intuitive explanation for this is the mechanism of adversarial learning, *i.e.*, adversarial loss could help to optimize generator, and make generator become more robust to the input space extended and smoothed by noises, leading to better performance than non-adversarial ones in noisy environments, as stated in Miyato et al. (2018).

4.5. Case study and analysis

We randomly select 3 uncensored patients ($\delta = 0$, with event occurrence) and 3 censored patients ($\delta = 1$, without event occurrence) from the test set of NLST. We show their time-to-event estimations by $\text{ESAT}_{\text{base}}$ and ESAT + AdvMIL, as well as their last follow-up times, in Fig. 10.

From these results, we can see that (1) AdvMIL enables ESAT to provide many time-to-event estimations for one patient, while original ESAT only gives one single completely-certain result; (2) the distribution estimation given by AdvMIL often covers the single point estimation given by ESAT; (3) the median of distribution estimation is often closer to the ground truth than point estimation. However, we cannot quantitatively assess the goodness of distribution coverage, because the ground truth of time-to-event distribution is unknown.

5. Discussion

Histological WSIs contain rich microenvironmental cues that are vital for disease prognosis. An accurate assessment of WSI-based prognosis could help to improve patient management and disease outcomes. Although many end-to-end weakly-supervised models have been developed to estimate patient prognosis from WSIs, their potential is generally restricted by the classical paradigm of survival analysis and fully-supervised learning. Inspired by adversarial time-to-event modeling, we propose a novel adversarial MIL framework to exploit their potential. This framework could bring new vigor and vitality into the survival analysis in computational pathology, by enabling existing MIL networks to estimate a more robust time-to-event distribution and learn from unlabeled WSI data via semi-supervised training, at a relatively low computational cost.

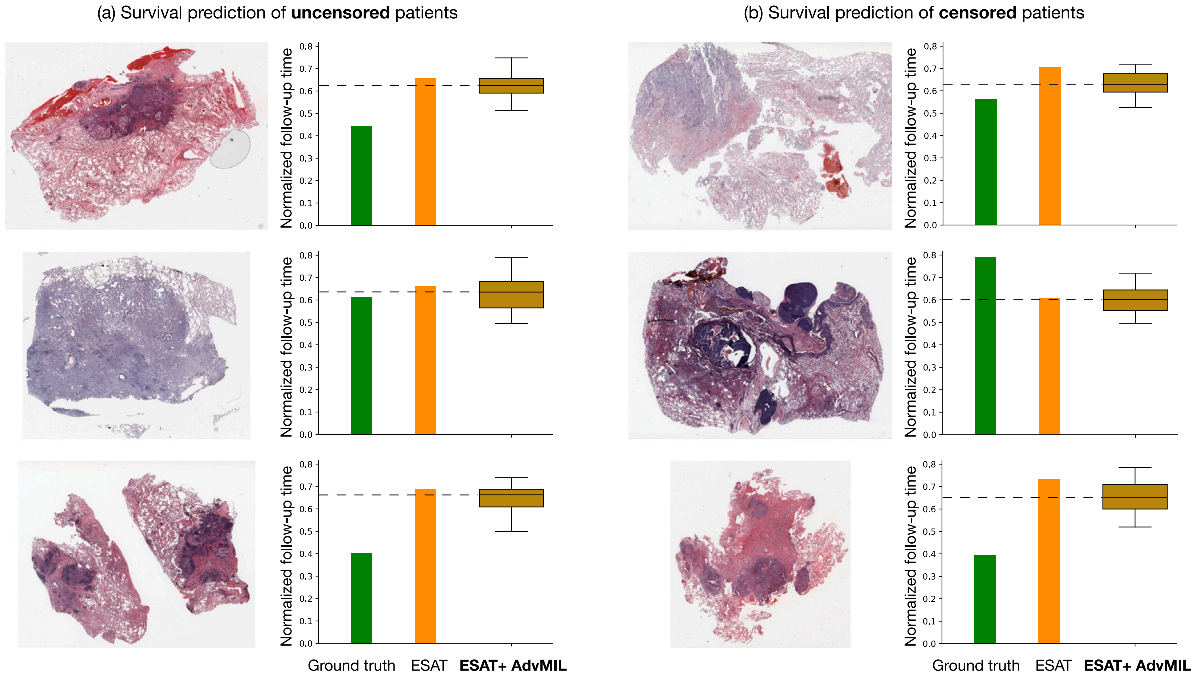


Fig. 10. Case study of the time-to-event estimations given by ESAT (a base version) and its AdvMIL-improved version. Three uncensored patients (left) and three censored patients (right) are randomly selected from the test set of NLST as study cases.

We emphasize that this study stands on the shoulders of GANs (Goodfellow et al., 2014a; Mirza and Osindero, 2014) and DATE (Chapfuwa et al., 2018), and shows for the first time how to generalize them to the MIL that is much necessary for WSI representation learning. Last but not least, apart from the technical contributions made by AdvMIL, this study also would like to highlight its practical contribution (*e.g.*, semi-supervised applications) to computational pathology community.

The main limitations of this study include that the coverage goodness of distribution estimation cannot be quantitatively evaluated since the ground truth of patient survival distribution is unavailable. In addition, there are also some constraints in our experiments: (1) limited sampling times are adopted to return distribution estimation because it is relatively time-consuming to infer once on gigapixel images, and (2) our datasets are limited to three cancer types, not covering more variety and clinical comparisons.

6. Conclusions

This paper proposes a novel framework, adversarial multiple instance learning (AdvMIL), for the survival analysis on gigapixel WSIs. This framework generalizes adversarial time-to-event modeling to MIL mainly by its two cores: the MIL encoder in generator and the fusion network with region-level instance projection in discriminator. It is a plug-and-play framework; namely it can be easily and efficiently applied to most end-to-end MIL models. The extensive experiments on three WSI datasets demonstrate that the proposed AdvMIL framework could often bring performance improvement to existing mainstream MIL models at a relatively low computational cost. Most importantly, AdvMIL could assist these existing models to fulfill a more effective estimation of time-to-event distribution and effectively learn from unlabeled WSIs via semi-supervised learning. Moreover, it is also observed that AdvMIL could help improving the robustness of models against patch occlusion and two representative image noises.

CRediT authorship contribution statement

Pei Liu: Conceptualization, Methodology, Software, Writing – original draft. **Luping Ji:** Conceptualization, Methodology, Funding acquisition, Supervision, Writing – review & editing. **Feng Ye:** Funding acquisition, Writing – review & editing. **Bo Fu:** Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All used datasets are publicly-available. Our code has been made publicly-available at https://github.com/liupei101/AdvMIL.

References

Carbonneau, M.-A., Cheplygina, V., Granger, E., Gagnon, G., 2018. Multiple instance learning: A survey of problem characteristics and applications. Pattern Recognit. 77, 329–353.

Carmon, Y., Raghunathan, A., Schmidt, L., Duchi, J.C., Liang, P.S., 2019. Unlabeled data improves adversarial robustness. Adv. Neural Inf. Process. Syst. 32.

Chapelle, O., Scholkopf, B., Zien, Eds., A., 2009. Semi-supervised learning (Chapelle, O. et al, Eds.; 2006) [book reviews]. IEEE Trans. Neural Netw. 20 (3), 542. http://dx.doi.org/10.1109/TNN.2009.2015974.

Chapfuwa, P., Tao, C., Li, C., Khan, I., Chandross, K.J., Pencina, M.J., Carin, L., Henao, R., 2020. Calibration and uncertainty in neural time-to-event modeling. IEEE Trans. Neural Netw. Learn. Syst..

Chapfuwa, P., Tao, C., Li, C., Page, C., Goldstein, B., Duke, L.C., Henao, R., 2018. Adversarial time-to-event modeling. In: International Conference on Machine Learning. PMLR, pp. 735–744.

Chen, R.J., Chen, C., Li, Y., Chen, T.Y., Trister, A.D., Krishnan, R.G., Mahmood, F., 2022a. Scaling vision transformers to gigapixel images via hierarchical self-supervised learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16144–16155.

Chen, R.J., Lu, M.Y., Shaban, M., Chen, C., Chen, T.Y., Williamson, D.F., Mahmood, F., 2021. Whole slide images are 2D point clouds: Context-aware survival prediction using patch-based graph convolutional networks. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 339–349.

Chen, R.J., Lu, M.Y., Williamson, D.F.K., Chen, T.Y., Lipkova, J., Noor, Z., Shaban, M., Shady, M., Williams, M., Joo, B., Mahmood, F., 2022b. Pan-cancer integrative histology-genomic analysis via multimodal deep learning. Cancer Cell 40 (8), 865–878.e6. http://dx.doi.org/10.1016/j.ccell.2022.07.004.

Chen, J.-N., Sun, S., He, J., Torr, P., Yuille, A., Bai, S., 2022c. Transmix: Attend to mix for vision transformers. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition. CVPR, IEEE, pp. 12125–12134. http://dx.doi.org/10.1109/CVPR52688.2022.01182.

Cheng, S., Liu, S., Yu, J., Rao, G., Xiao, Y., Han, W., Zhu, W., Lv, X., Li, N., Cai, J., Wang, Z., Feng, X., Yang, F., Geng, X., Ma, J., Li, X., Wei, Z., Zhang, X., Quan, T., Zeng, S., Chen, L., Hu, J., Liu, X., 2021. Robust whole slide image analysis for cervical cancer screening using deep learning. Nature Commun. 12, 5639. http://dx.doi.org/10.1038/s41467-021-25296-x.

Cox, D.R., 1975. Partial likelihood. Biometrika 62 (2), 269–276.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, pp. 248–255. http://dx.doi.org/10.1109/CVPR.2009.5206848.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations.

Ghaffari Laleh, N., Muti, H.S., Loeffler, C.M.L., Echle, A., Saldanha, O.L., Mahmood, F., Lu, M.Y., Trautwein, C., Langer, R., Dislich, B., Buelow, R.D., Grabsch, H.I., Brenner, H., Chang-Claude, J., Alwers, E., Brinker, T.J., Khader, F., Truhn, D., Gaisa, N.T., Boor, P., Hoffmeister, M., Schulz, V., Kather, J.N., 2022. Benchmarking weakly-supervised deep learning pipelines for whole slide classification in computational pathology. Med. Image Anal. 79, 102474. http://dx.doi.org/10.1016/j.media.2022.102474.

Goodfellow, I., 2016. Nips 2016 tutorial: Generative adversarial networks. arXiv preprint arXiv:1701.00160.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y., 2014a. Generative adversarial nets. In: Advances in Neural Information Processing Systems. Vol. 27.

Goodfellow, I.J., Shlens, J., Szegedy, C., 2014b. Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.

Gui, J., Sun, Z., Wen, Y., Tao, D., Ye, J., 2021. A review on generative adversarial networks: Algorithms, theory, and applications. IEEE Trans. Knowl. Data Eng. 1. http://dx.doi.org/10.1109/TKDE.2021.3130191.

Harrell, Jr., F.E., Lee, K.L., Califf, R.M., Pryor, D.B., Rosati, R.A., 1984. Regression modelling strategies for improved prognostic prediction. Stat. Med. 3 (2), 143–152. http://dx.doi.org/10.1002/sim.4780030207.

He, K., Zhang, X., Ren, S., Sun, J., 2016. Deep residual learning for image recognition. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition. CVPR, IEEE, pp. 770–778. http://dx.doi.org/10.1109/CVPR.2016.90.

Huang, Z., Chai, H., Wang, R., Wang, H., Yang, Y., Wu, H., 2021. Integration of patch features through self-supervised learning and transformer for survival analysis on whole slide images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 561–570.

Ilse, M., Tomczak, J., Welling, M., 2018. Attention-based deep multiple instance learning. In: International Conference on Machine Learning. PMLR, pp. 2127–2136.

Kalbfleisch, J.D., Prentice, R.L., 2011. The Statistical Analysis of Failure Time Data. John Wiley & Sons.

Kallenberg, O., 2002. Foundations of Modern Probability. Springer Science & Business Media.

Kandoth, C., McLellan, M.D., Vandin, F., Ye, K., Niu, B., Lu, C., Xie, M., Zhang, Q., McMichael, J.F., Wyczalkowski, M.A., Leiserson, M.D.M., Miller, C.A., Welch, J.S., Walter, M.J., Wendl, M.C., Ley, T.J., Wilson, R.K., Raphael, B.J., Ding, L., 2013. Mutational landscape and significance across 12 major cancer types. Nature 502, 333–339.

Kather, J.N., Krisam, J., Charoentong, P., Luedde, T., Herpel, E., Weis, C.-A., Gaiser, T., Marx, A., Valous, N.A., Ferber, D., Jansen, L., Reyes-Aldasoro, C.C., Zörnig, I., Jäger, D., Brenner, H., Chang-Claude, J., Hoffmeister, M., Halama, N., 2019. Predicting survival from colorectal cancer histology slides using deep learning: A retrospective multicenter study. PLOS Med. 16 (1), e1002730. http://dx.doi.org/10.1371/journal.pmed.1002730.

Kendall, A., Gal, Y., 2017. What uncertainties do we need in bayesian deep learning for computer vision? In: Advances in Neural Information Processing Systems. Vol. 30.

Kipf, T.N., Welling, M., 2016. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:1609.02907.

Lakshminarayanan, B., Pritzel, A., Blundell, C., 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In: Advances in Neural Information Processing Systems. Vol. 30.

Laleh, N.G., Truhn, D., Veldhuizen, G.P., Han, T., van Treeck, M., Buelow, R.D., Langer, R., Dislich, B., Boor, P., Schulz, V., Kather, J.N., 2022. Adversarial attacks and adversarial robustness in computational pathology. Nature Commun. 13, 5711. http://dx.doi.org/10.1038/s41467-022-33266-0.

Li, C., Xu, K., Zhu, J., Liu, J., Zhang, B., 2021. Triple generative adversarial networks. IEEE Trans. Pattern Anal. Mach. Intell..

Li, R., Yao, J., Zhu, X., Li, Y., Huang, J., 2018. Graph CNN for survival analysis on whole slide pathological images. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. Springer, pp. 174–182.

Linmans, J., Elfwing, S., van der Laak, J., Litjens, G., 2023. Predictive uncertainty estimation for out-of-distribution detection in digital pathology. Med. Image Anal. 83, 102655. http://dx.doi.org/10.1016/j.media.2022.102655.

Liu, P., Fu, B., Yang, S.X., Deng, L., Zhong, X., Zheng, H., 2021. Optimizing survival analysis of XGBoost for ties to predict disease progression of breast cancer. IEEE Trans. Biomed. Eng. 68 (1), 148–160. http://dx.doi.org/10.1109/TBME.2020.2993278.

Liu, P., Fu, B., Ye, F., Yang, R., Ji, L., 2023. DSCA: A dual-stream network with cross-attention on whole-slide image pyramids for cancer prognosis. Expert Systems With Applications 227, 120280. http://dx.doi.org/10.1016/j.eswa.2023.120280.

Lu, M.Y., Williamson, D.F., Chen, T.Y., Chen, R.J., Barbieri, M., Mahmood, F., 2021. Data-efficient and weakly supervised computational pathology on whole-slide images. Nat. Biomed. Eng. 5 (6), 555–570.

Marini, N., Otálora, S., Müller, H., Atzori, M., 2021. Semi-supervised training of deep convolutional neural networks with heterogeneous data and few local annotations: An experiment on prostate histopathology image classification. Med. Image Anal. 73, 102165. http://dx.doi.org/10.1016/j.media.2021.102165.

Mirza, M., Osindero, S., 2014. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784.

Miyato, T., Koyama, M., 2018. Cgans with projection discriminator. In: International Conference on Learning Representations.

Miyato, T., Maeda, S.-i., Koyama, M., Ishii, S., 2018. Virtual adversarial training: A regularization method for supervised and semi-supervised learning. IEEE Trans. Pattern Anal. Mach. Intell. 41 (8), 1979–1993.

Naseer, M.M., Ranasinghe, K., Khan, S.H., Hayat, M., Khan, F.S., Yang, M.-H., 2021. Intriguing properties of vision transformers. In: Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P.S., Vaughan, J.W. (Eds.), Advances in Neural Information Processing Systems. Vol. 34. pp. 23296–23308.

National Lung Screening Trial Research Team, 2011. The national lung screening trial: Overview and study design. Radiology 258 (1), 243–253. http://dx.doi.org/10.1148/radiol.10091808.

Nazarovs, J., Huang, Z., Tasneeyapant, S., Chakraborty, R., Singh, V., 2022. Understanding uncertainty maps in vision with statistical testing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 406–416.

Nir, G., Hor, S., Karimi, D., Fazli, L., Skinnider, B.F., Tavassoli, P., Turbin, D., Villamil, C.F., Wang, G., Wilson, R.S., Iczkowski, K.A., Lucia, M.S., Black, P.C., Abolmaesumi, P., Goldenberg, S.L., Salcudean, S.E., 2018. Automatic grading of prostate cancer in digitized histopathology images: Learning from multiple experts. Med. Image Anal. 50, 167–180. http://dx.doi.org/10.1016/j.media.2018.09.005.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., 2016. Improved techniques for training gans. In: Advances in Neural Information Processing Systems. Vol. 29.

Shao, W., Wang, T., Huang, Z., Han, Z., Zhang, J., Huang, K., 2021. Weakly supervised deep ordinal cox model for survival prediction from whole-slide pathological images. IEEE Trans. Med. Imaging 40 (12), 3739–3747.

Shen, Y., Liu, L., Tang, Z., Chen, Z., Ma, G., Dong, J., Zhang, X., Yang, L., Zheng, Q., 2022. Explainable survival analysis with convolution-involved vision transformer. In: Proceedings of the AAAI Conference on Artificial Intelligence. AAAI Press, pp. 2207–2215.

Skrede, O.-J., De Raedt, S., Kleppe, A., Hveem, T.S., Liestøl, K., Maddison, J., Askautrud, H.A., Pradhan, M., Nesheim, J.A., Albregtsen, F., Farstad, I.N., Domingo, E., Church, D.N., Nesbakken, A., Shepherd, N.A., Tomlinson, I., Kerr, R., Novelli, M., Kerr, D.J., Danielsen, H.E., 2020. Deep learning for prediction of colorectal cancer outcome: A discovery and validation study. Lancet 395 (10221), 350–360. http://dx.doi.org/10.1016/S0140-6736(19)32998-8.

Springenberg, J.T., 2015. Unsupervised and semi-supervised learning with categorical generative adversarial networks. arXiv preprint arXiv:1511.06390.

Tellez, D., Balkenhol, M., Otte-Holler, I., van de Loo, R., Vogels, R., Bult, P., Wauters, C., Vreuls, W., Mol, S., Karssemeijer, N., Litjens, G., van der Laak, J., Ciompi, F., 2018. Whole-slide mitosis detection in H&E breast histology using PHH3 as a reference to train distilled stain-invariant convolutional networks. IEEE Trans. Med. Imaging 37, 2126–2136. http://dx.doi.org/10.1109/TMI.2018.2820199.

Tellez, D., Litjens, G., Bándi, P., Bulten, W., Bokhorst, J.-M., Ciompi, F., van der Laak, J., 2019. Quantifying the effects of data augmentation and stain color normalization in convolutional neural networks for computational pathology. Med. Image Anal. 58, 101544. http://dx.doi.org/10.1016/j.media.2019.101544.

Uemura, T., Näppi, J.J., Watari, C., Hironaka, T., Kamiya, T., Yoshida, H., 2021. Weakly unsupervised conditional generative adversarial network for image-based prognostic prediction for COVID-19 patients based on chest CT. Med. Image Anal. 73, 102159.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. In: Advances in Neural Information Processing Systems. pp. 5999–6009.

Wei, L.J., 1992. The accelerated failure time model: A useful alternative to the cox regression model in survival analysis. Stat. Med. 11 (14–15), 1871–1879. http://dx.doi.org/10.1002/sim.4780111409.

Wulczyn, E., Steiner, D.F., Moran, M., Plass, M., Reihs, R., Tan, F., Flament-Auvigne, I., Brown, T., Regitnig, P., Chen, P.-H.C., Hegde, N., Sadhwani, A., MacDonald, R., Ayalew, B., Corrado, G.S., Peng, L.H., Tse, D., Müller, H., Xu, Z., Liu, Y., Stumpe, M.C., Zatloukal, K., Mermel, C.H., 2021. Interpretable survival prediction for colorectal cancer using deep learning. npj Digit. Med. 4 (1), 71. http://dx.doi.org/10.1038/s41746-021-00427-2.

Yao, J., Zhu, X., Jonnagaddala, J., Hawkins, N., Huang, J., 2020. Whole slide images based cancer survival prediction using attention guided deep multiple instance learning networks. Med. Image Anal. 65, 101789.

Yu, K.-H., Zhang, C., Berry, G.J., Altman, R.B., Ré, C., Rubin, D.L., Snyder, M., 2016. Predicting non-small cell lung cancer prognosis by fully automated microscopic pathology image features. Nature Commun. 7 (1), 12474. http://dx.doi.org/10.1038/ncomms12474.

Zadeh, S.G., Schmid, M., 2021. Bias in cross-entropy-based training of deep survival networks. IEEE Trans. Pattern Anal. Mach. Intell. 43 (9), 3126–3137. http://dx.doi.org/10.1109/TPAMI.2020.2979450.

Zarella, M.D., Bowman;, D., Aeffner, F., Farahani, N., Xthona;, A., Absar, S.F., Parwani, A., Bui, M., Hartman, D.J., 2018. A practical guide to whole slide imaging: A white paper from the digital pathology association. Arch. Pathol. Lab. Med. 143 (2), 222–234. http://dx.doi.org/10.5858/arpa.2018-0343-RA.

Zhou, Z.-H., 2021. Semi-supervised learning. In: Machine Learning. Springer Singapore, Singapore, pp. 315–341. http://dx.doi.org/10.1007/978-981-15-1967-3_13.

Zhou, X., Jiao, Y., Liu, J., Huang, J., 2022. A deep generative approach to conditional sampling. J. Amer. Statist. Assoc. 1–12.