Learning Survival Distribution with Implicit Survival Function

Yu Ling, **Weimin Tan*** and **Bo Yan***

School of Computer Science, Shanghai Key Laboratory of Intelligent Information Processing, Shanghai Collaborative Innovation Center of Intelligent Visual Computing, Fudan University, Shanghai, China.

yling21@m.fudan.edu.cn, {wmtan, byan}@fudan.edu.cn

Abstract

Survival analysis aims at modeling the relationship between covariates and event occurrence with some untracked (censored) samples. In implementation, existing methods model the survival distribution with strong assumptions or in a discrete time space for likelihood estimation with censorship, which leads to weak generalization. In this paper, we propose Implicit Survival Function (ISF) based on Implicit Neural Representation for survival distribution estimation without strong assumptions, and employ numerical integration to approximate the cumulative distribution function for prediction and optimization. Experimental results show that ISF outperforms the state-of-the-art methods in three public datasets and has robustness to the hyperparameter controlling estimation precision.

1 Introduction

Survival analysis is a typical statistical task for tracking occurrence of the event of interest through modeling relationship between covariates and event occurrence. In some medical situations [Courtiol *et al.*, 2019; Zadeh Shirazi *et al.*, 2020], researchers model the death probability of some diseases using survival analysis to explore effects of prognostic factors. However, some samples lose tracking (censored) during observation. For example, some patients are still alive at the end of observation, whose survival times are unavailable. Such censored samples are valuable for analysis of favorable prognosis. Therefore, censorship is one key problem in survival analysis as well as survival distribution modeling.

The most widely-used survival analysis model Cox proportional hazard method [Cox, 1992] predicts a hazard rate, which assumes that the relationship between covariates and hazard is time-invariant. For optimization, Cox model and its extensions [Tibshirani, 1997; Li *et al.*, 2016; Katzman *et al.*, 2018; Zhu *et al.*, 2016] maximize the ranking accuracy of comparable pairs including comparison between uncensored samples and censored samples.
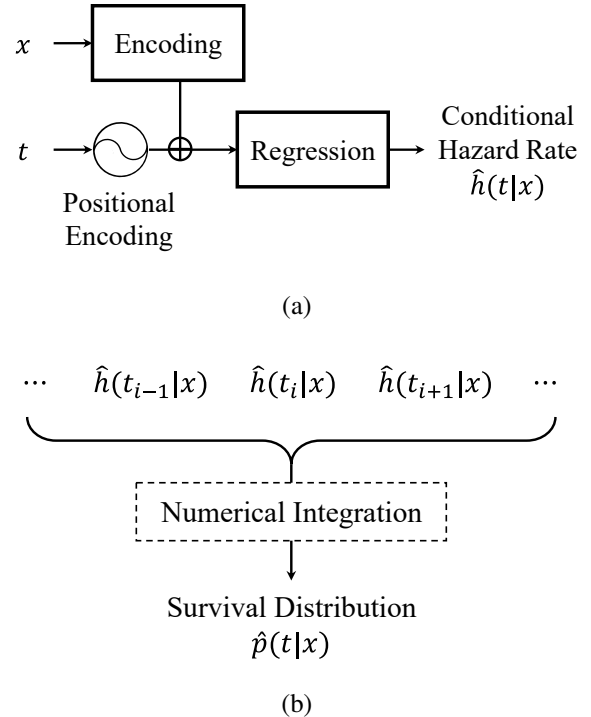
Figure 1: Brief framework of ISF. (a) ISF takes sample x and time t as input, and predicts conditional hazard rate $\hat{h}(t|x)$. (b) Based on estimated conditional hazard rates, we can derive survival distribution $\hat{p}(t|x)$ through numerical integration.

Lately, some works introduce deep neural networks to survival analysis. DeepSurv [Katzman *et al.*, 2018] and DeepConvSurv [Zhu *et al.*, 2016] simply replace the linear regression in the Cox model with neural networks for non-linear representations. These methods maintain the strong assumption of hazards' time-invariance in Cox model, leading to weak generalization of networks in real-world applications.

To avoid strong assumption on survival distribution, researchers try to estimate a distribution in a discrete time space instead of predicting a time-invariant risk. DeepHit [Lee *et al.*, 2018] is proposed to learn occurrence probabilities at preset time points directly without assumptions about underlying stochastic process. Deep Recurrent Survival Analysis

(DRSA) [Ren *et al.*, 2019] builds a recurrent network to capture the sequential patterns of the feature over time in survival analysis. Therefore, both DeepHit and DRSA learn a discrete survival distribution. Compared to the cross-entropy loss, the log-likelihood loss obtains better prediction for DeepHit and DRSA [Zadeh and Schmid, 2021]. On the basis of predicted occurrence probabilities in the discrete time space, the log-likelihood is naturally estimated in DeepHit and DRSA for both censored and uncensored samples.

Differing from discrete distribution estimation in DeepHit and DRSA, DSM [Nagpal *et al.*, 2021] estimates the average mixture of parametric distributions. In implementation, DSM employs Weibull and Log-Normal distributions for analytical solutions of the cumulative distribution functions (CDF) and support limited in the space of positive reals. Therefore, DSM includes censored samples during optimization through CDF estimation. However, DSM also introduces assumptions on survival distribution through parametric distribution selection.

In this paper, we propose Implicit Survival Function (ISF) based on Implicit Neural Representation which is widely-used in 2D and 3D image representation [Mildenhall *et al.*, 2020; Chen *et al.*, 2020]. As shown in Figure 1(a), ISF estimates a conditional hazard rate with the given sample and time. To capture time patterns, we embed the input time through Positional Encoding [Vaswani *et al.*, 2017]. The aggregated vector of encoded sample feature and time embedding is fed to a the regression module for conditional hazard rate estimation without strong assumptions on survival distribution. As shown in Figure 1(b), we employ numerical integration with predicted conditional hazard rates for survival distribution prediction.

For optimization, we maximize likelihood of both censored and uncensored samples on the basis of approximated CDF of survival in a discrete time space. And experimental results prove that ISF is robust to the hyperparameter setting of the discrete time space.

To summarize, the contributions of this paper can be listed as:

- The proposed Implicit Survival Function (ISF) directly models the conditional hazard rate without strong assumptions on survival distribution, and captures the effect of time through Positional Encoding.

- To estimate survival distribution with ISF, numerical integration is used to approximate the cumulative distribution function (CDF). Therefore, ISF can handle censorship common in survival analysis through maximum likelihood estimation based on approximated CDF.

- Though survival distribution estimation of ISF is based on a discrete time space, ISF has capability to represent a continuous survival distribution through Implicit Neural Representation. And experimental results show that ISF is robust to the setting of the discrete time space.

- To demonstrate performance of the proposed model compared with the state-of-the-art methods, experiments are built on several real-world datasets. Experimental results show that ISF outperforms the state-of-the-art methods.

2 Formulation

Survival analysis models aim at modeling the probabilistic density function (PDF) of tracked event defined as:

$$p(t|x) = Pr(t_x = t|x) \tag{1}$$

where t denotes time, and t_x denotes the true survival time.

Thus, the survival rate that the tracked event occurs after time t_i is defined as:

$$\begin{aligned} S(t_i|x) &= Pr(t_x > t_i|x) \\ &= \int_{t_i}^{\infty} p(t|x)dt \end{aligned} \tag{2}$$

Similarly, the event rate function of time t_i is defined as the cumulative distribution function (CDF):

$$\begin{aligned} W(t_i|x) &= Pr(t_x \leq t_i|x) = 1 - S(t_i|x) \\ &= \int_0^{t_i} p(t|x)dt \end{aligned} \tag{3}$$

The conditional hazard rate $h(t|x)$ is defined as:

$$h(t|x) = \lim_{\Delta t \to 0} \frac{Pr(t < t_x \leq t + \Delta t|t_x \geq t, x)}{\Delta t} \tag{4}$$

3 Related Work

In this section, we describe several related approaches. The previous methods are divided into three parts based on their target of estimation: proportional hazard rate, discrete survival distribution and distribution mixture.

3.1 Proportional Hazard Rate

The Cox proportional hazard method proposed in [Cox, 1992] is a widely-used method in survival analysis tasks. Cox model assumes that the hazard rate of occurrence of a certain event is constant with time and the log of hazard rate can be represented by a linear function. Thus, the basic form of Cox model is:

$$\hat{h}(t|x) = h_0(t)exp(w^T x) \tag{5}$$

where t denotes time, t_x denotes the true survival time, $x = (x_1, x_2, \ldots, x_p)^T$ denotes covariates of samples, $w = (w_1, w_2, \ldots, w_p)^T$ denotes parameters of the linear regression, and $h_0(t)$ denotes a fixed time-dependent baseline hazard function. Parameters w can be estimated by minimizing the negative log partial likelihood.

However, the time-invariance assumption of hazard in Cox model weakens its generalization. Other methods make different assumptions about the survival function such as Exponential distribution [Lee and Wang, 2003], Weibull distribution [Ranganath *et al.*, 2016], Wiener process [Doksum and Hóyland, 1992] and Markov Chain [Longini *et al.*, 1989]. These methods with strong assumptions about the underlying stochastic processes fix the form of survival functions, which suffers from generalization problem in real-world situations.

The outstanding capability of deep learning in non-linear regression achieve researchers' high attention. Therefore,

many approaches introduce deep learning to survival analysis. DeepSurv [Katzman *et al.*, 2018] replaces the linear regression of Cox model with a deep neural network for non-linear representation, but maintains the basic assumption of Cox model. Some works [Zhu *et al.*, 2016; Li *et al.*, 2019] extend DeepSurv with a deep convolutional neural network for unstructured data such as images.

3.2 Discrete Probability Distribution

To avoid strong assumptions about the survival time distribution, previous methods model the survival analysis problem in a discrete space with K time points $T = \{t_0^p, t_1^p, \cdots t_{k-1}^p\}$. DeepHit [Lee *et al.*, 2018] uses a fully-connected network to directly predict occurrence probability $\hat{p}(t_i^p|x)$ defined as:

$$\hat{p}(t_i^p|x) = Pr(t_x = t_i^p|x) \tag{6}$$

where t_i^p is a time point in the discrete time space $t_i^p \in T$.

DRSA [Ren *et al.*, 2019] employs standard LSTM units [Hochreiter and Schmidhuber, 1997] to capture sequential patterns of features over time, and predicts a conditional hazard rate defined as:

$$\hat{h}(t_i^p|x) = \lim_{\Delta t \to 0} \frac{Pr(t_{i-1}^p < t_x \leq t_i^p|t_x \geq t_{i-1}^p, x)}{\Delta t} \tag{7}$$

Hence, DRSA defines occurrence probability of event as:

$$\hat{p}(t_i^p|x) = \hat{h}(t_i^p|x) \prod_{j<i}(1 - \hat{h}(t_j^p|x)) \tag{8}$$

Although both DeepHit and DRSA predicts directly predict survival distribution without strong assumption, they only estimate probabilities at discrete time points.

3.3 Distribution Mixture

Discrete probability distribution estimation methods only estimate a fixed number of probabilities, which limits their applications. To generate a continuous probability distribution, DSM [Nagpal *et al.*, 2021] learns a mixture of K well-defined parametric distributions. Assuming that all survival times follows $t \geq 0$, DSM selects distributions which only have support in the space of positive reals. And for gradient based optimization, CDF of selected distributions require analytical solutions. In implementation, DSM employs Weibull and Log-Normal distributions, namely primitive distributions.

During inference, parameters of K primitive distributions $\{\beta_k, \eta_k\}_{k=1}^K$ and their weights $\{\alpha_k\}_{k=1}^K$ are estimated through MLP. Thus, the final individual survival distribution $\hat{p}(t|x)$ is defined as the weighted average of K primitive distributions:

$$\hat{p}(t|x) = \sum_{k=1}^K \alpha_k P_k^p(t|x, \beta_k, \eta_k) \tag{9}$$

However, DSM introduces assumptions of survival distributions since primitive distribution selection is taken as a hyperparameter.

4 Methodology

To model the survival distribution, we propose Implicit Survival Function (ISF) to estimate conditional hazard rate with positional encoding of time. In this section, we will demonstrate details of ISF as illustrated in Figure 2.
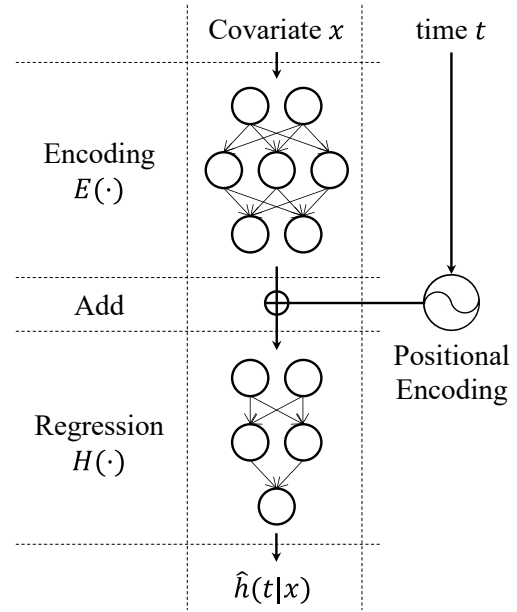

Figure 2: Pipeline of ISF. Time t is embedded through Positional Encoding (PE). Conditional hazard rate $\hat{h}(t|x)$ is estimated through $H(E(x) + PE(t))$, where $E(\cdot)$ and $H(\cdot)$ are implemented with MLP.

4.1 Implicit Survival Function

The proposed ISF aims at predicting $h(t|x)$ defined in Eq. 4. For a given sample x, ISF first generates a feature vector $z \in \mathbb{R}^d$ using a Multilayer Perceptron (MLP) denoted by encoder $E(\cdot)$:

$$z_x = E(x) \tag{10}$$

To capture the effect of time, Positional Encoding (PE) of time t is added to the feature vector z. Then, our hazard rate regression $\hat{h}(t|x)$ is defined as:

$$\begin{aligned}
\hat{h}(t|x) &= H(z_x + PE(t)) \\
&= H(E(x) + PE(t)) \tag{11}
\end{aligned}$$

where $H(\cdot)$ is implemented with a MLP.

Positional Encoding maps time t to a embedding of d dimensions using pre-defined sinusoidal functions [Vaswani *et al.*, 2017]:

$$\begin{cases}
PE(t, 2i) = sin(t/10000^{2i/d}) \\
PE(t, 2i+1) = cos(t/10000^{2i/d})
\end{cases} \tag{12}$$

The sinusoidal function based Positional Encoding provides shift-invariant representations, and let MLP learn high frequency functions [Tancik *et al.*, 2020]. Therefore, ISF employs Positional Encoding defined in Eq.12 for embedding of time in survival analysis.

4.2 Survival Distribution Estimation

For survival distribution estimation with ISF, we first estimate survival rate $S(t|x)$ defined in Eq. 2, and then approximate occurrence probability $p(t|x)$ defined in Eq. 1 through difference of survival rate.

From Eqs. 2 and 4, we can derive the log survival rate at time t_i as:

$$\ln S(t_i|x) = \ln Pr(t_x > t_i|x)$$

$$= \int_0^{t_i} \ln Pr(t_x > t|t_x \geq t, x)dt$$

$$= \int_0^{t_i} \ln (1 - h(t|x)) \, dt \qquad (13)$$

Therefore, the estimated survival rate $\hat{S}(t_i|x)$ is defined as:

$$\hat{S}(t_i|x) = \exp \int_0^{t_i} \ln \left(1 - \hat{h}(t|x)\right) dt$$

$$= \exp \int_0^{t_i} \ln (1 - H(E(x) + PE(t))) \, dt \quad (14)$$

The estimated occurrence probability $\hat{p}(t|x)$ is approximated through:

$$\hat{p}(t|x) \approx Pr(t < t_x \leq t + \epsilon|x)$$

$$\approx \hat{S}(t|x) - \hat{S}(t + \epsilon|x) \qquad (15)$$

where ϵ is a hyperparameter. The setting of ϵ depends on the precision of annotations in the dataset. Corresponding discussion is included in Section 5.5

For numerical stability, we manually set $\hat{S}(0|x) = 1$ and $\hat{S}(t_{max}|x) = 0$, where t_{max} is ensured to be larger than any possible survival time in the dataset.

4.3 Numerical Integration

Analytical solutions for integration in Eq. 14 is unavailable for ISF. To overcome such problem, we use numerical integration to approximate CDF in a discrete time space.

The duration of survival time $[0, t_{max})$ is split into K intervals $\{(t_i^p, t_{i+1}^p]\}_{i=0}^{K-1}$ with time points $T = \{t_i^p\}_{i=0}^K$, where $t_0^p = 0$ and $t_k^p = t_{max}$. In this paper, we set $t_{i+1}^p = t_i^p + \epsilon$ for convenience.

Let $g(t, x)$ denote $\ln(1 - \hat{h}(t|x))$. Therefore, the integration in Eq. 14 for $t_i^p \in T$ is calculated using Simpson Formula as:

$$\hat{S}(t_i^p|x) = \exp \int_0^{t_i^p} g(t, x)dt$$

$$\approx \exp \sum_{j<i} \frac{\epsilon}{6}[g(t_j^p, x) + 4g(t_j^p + \frac{\epsilon}{2}, x) + g(t_{j+1}^p, x)]$$

$$(16)$$

Thus, the event rate (CDF) is estimated as $\hat{W}(t_i^p|x) = 1 - \hat{S}(t_i^p|x)$.

4.4 Loss Function

Like existing approaches [Lee *et al.*, 2018; Ren *et al.*, 2019; Nagpal *et al.*, 2021], we construct loss functions on the basis of maximum likelihood estimation. Although ISF provides a conditional hazard rate in the continuous time space, the optimization is performed in the discrete time space for CDF approximation. In this section, for easily understanding, we describe the proposed loss function separately for censored and uncensored samples in the view of predicting $\hat{p}(t|x)$, though forms of loss functions for these two types of samples are the same.

Censored Samples
For a censored sample, the true survival time t_x is unknown but the latest observation time t_x^o is available, which indicates $t_x > t_x^o$. Thus, the loss function is expected to maximize $\hat{S}(t_x^o|x)$. For simplification, we maximize $\hat{S}(t_i^p|x)$ where $t_x^o \in (t_i^p, t_{i+1}^p]$.

Therefore, the loss function for censored samples is defined as:

$$L_{cs}(x) = -\ln \hat{S}(t_i^p|x)$$

$$= -\ln \sum_{j \geq i} \hat{p}(t_j^p|x) \qquad (17)$$

where the latest observation time $t_x^o \in (t_i^p, t_{i+1}^p]$.

Uncensored Samples
Given an uncensored sample (x, t_x^o), the observation time t_x^o is equal to the true survival time t_x. Thus, we maximize $\hat{p}(t_i^p|x)$ where the true survival time $t_x^o \in (t_{i-1}^p, t_i^p]$:

$$L_{ucs}(x) = -\ln \hat{p}(t_i^p|x) \qquad (18)$$

Unified Loss
According to L_{cs} in Eq. 17 and L_{ucs} in Eq. 18, loss for both uncensored and censored samples can be represented as sum of $\hat{p}(t_i^p|x)$ in the discrete time space. For unification, we first define an indicator vector $Y^x \in \mathbb{R}^K$ in the discrete time space including $K + 1$ time points as:

$$Y_i^x = \begin{cases} 1 & t_x^o \in (t_i^p, t_{i+1}^p] \\ 0 & t_x^o \notin (t_i^p, t_{i+1}^p] \end{cases} \qquad (19)$$

Thus, the proposed loss function can be unified as:

$$L(x) = -\ln (Y_i^x \hat{p}(t_i|x)) \qquad (20)$$

The unified loss function $L(\cdot)$ handles both censored and uncensored samples. We use indicator vector Y^x to control likelihood calculation. Hence, the proposed loss function is suitable for any type of censorship.

4.5 Computational Complexity

As discussed in Sections 4.2, 4.3 and 4.4, estimation and optimization of ISF is performed in a discrete time space with K time intervals. For N samples, ISF predicts $O(NK)$ occurrence probabilities for survival distribution estimation. However, such process can be accelerated in the parallel computation situation because of independent positional encoding of time points.

4.6 Difference from Existing Methods

In this section, we compare the proposed model ISF with deep-learning models DeepHit, DRSA and DSM whose survival distribution estimation is close to that of ISF. We illustrate brief frameworks of these models and ISF in Figure 3.

ISF vs DeepHit
As shown in Figure 3(a), DeepHit directly regresses occurrence probabilities at preset time points through MLP. Therefore, the number of parameters dependents on the number of time points in the discrete time space.

Since ISF takes positional encoding of time as input, the number of parameters in ISF is independent to the amount of time points. Therefore, ISF has better expansibility for time space variation.
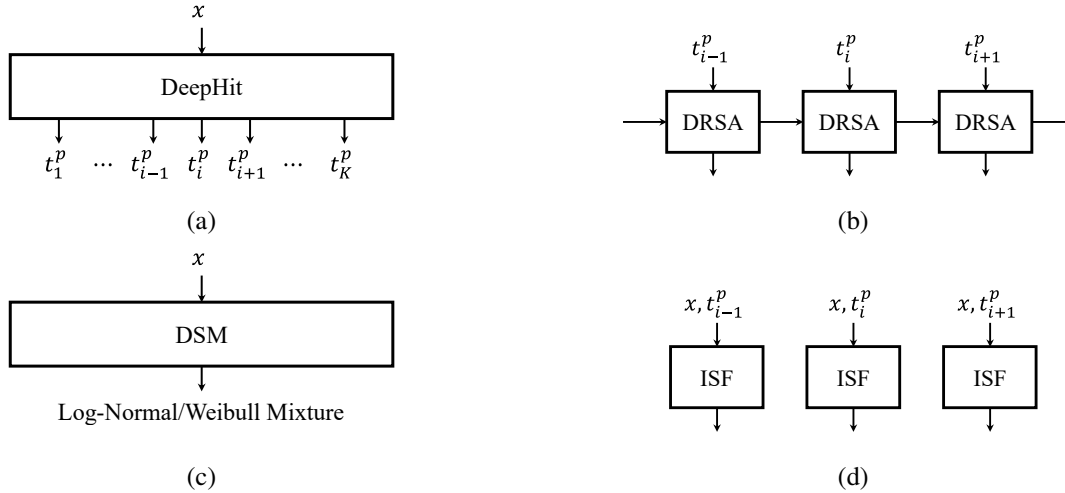
Figure 3: Framework comparison between existing methods and ISF. (a) DeepHit predicts occurrence probabilities at preset time points. (b) RNN based DRSA sequentially estimates conditional hazard rates over time. (c) DSM models survival distribution through estimates parameters of mixture of parametric distributions (Log-Normal/Weibull). (d) ISF takes sample x and time t_i^p as input, and generates independent estimation for time points.

ISF vs DRSA

According to Eqs. 7 and 11, the goal of both ISF and DRSA is conditional hazard rate estimation. With estimated hazard rate, occurrence probability can be easily derived as shown in Eqs. 8 and 15.

The main difference between ISF and DRSA is the method of capturing time effect. As shown in Figure 3(b), DRSA applies RNN to learn sequential patterns in a discrete time space and serially processes preset time points, while ISF uses positional encoding to exploit time information in the real field through parallel computation.

ISF vs DSM

DSM models continuous survival distribution with mixture of parametric distributions as shown in Figure 3(c). Instead of explicit distribution representation in Eq. 9, ISF learns a function $H(\cdot)$ taking time as input defined in Eq. 11 to directly estimate conditional hazard rate. Therefore, the implicit representation of survival distribution in ISF avoids strong assumptions on survival distribution.

With decrease of ϵ in Eq. 15, precision of occurrence probability approximation increase, and thus ISF can be regarded as approximation of a continuous survival distribution. Distribution mixture in DSM directly models a continuous survival distribution, but distribution selection is a hyperparameter with strong assumptions about the stochastic process.

5 Experiments

In this section, we compare the proposed method ISF with the state-of-the-art deep-learning survival distribution estimation methods including DeepHit, DRSA and DSM. DeepHit predicts the occurrence probability $\hat{p}(t|x)$ directly with a fully-connected neural network [Lee *et al.*, 2018]. DRSA estimates a conditional hazard rate $\hat{h}(t|x)$ with LSTM units to capture sequential patterns [Ren *et al.*, 2019]. Both Deep-Hit and DRSA perform survival analysis in the discrete time

space, while DSM estimates a continuous survival distribution through the mixture of parametric distributions [Nagpal *et al.*, 2021]. Besides, we also compare ISF with Cox [Cox, 1992], its deep-learning extension DeepSurv [Katzman *et al.*, 2018] and random forest based survival analysis method RSF [Ishwaran *et al.*, 2008].

5.1 Datasets

To demonstrate the performance of the proposed method, experiments are conducted on several public real-world dataset:

- **CLINIC** tracks patients' clinic status [Knaus *et al.*, 1995]. The tracked event is the biological death. Survival analysis in CLINIC is to estimate death probability with physiologic variables.

- **MUSIC** is a user lifetime analysis containing about 1000 users with entire listening history [Jing and Smola, 2017]. The tracked event is the user visit to the music service. The goal of survival analysis is to predict the time elapsed from the last visit of one user to the next visit.

- **METABRIC** dataset contains gene expression profiles and clinical features of the breast cancer from 1,981 patients [Curtis *et al.*, 2012]. Following the experimental setting of DeepHit, 21 clinical features are used during evaluation [Lee *et al.*, 2018].

The statistics of three datasets is shown in Table 1. The training and testing split of CLINIC and MUSIC follows the setting of DRSA [Ren *et al.*, 2019]. For METABRIC, 5-fold cross validation is applied following DeepHit [Lee *et al.*, 2018].

5.2 Metric

Concordance Index (C-index, CI) is a widely-used evaluation metric in survival analysis for measuring the probability of accurate pair-wise order of comparable samples' event

Dataset	#Total Data	#Censored Data	Censoring Rate	#Features	Max Time
CLINIC	6,036	797	0.132	14	82
MUSIC	3,296,328	1,157,572	0.351	6	300
METABRIC	1,981	1,093	0.552	21	356

Table 1: The statistics of CLINIC, MUSIC and METABRIC.

time. However, the ordinary CI [Harrell *et al.*, 1982] for proportional hazard models assumes the predicted value is time-invariant [Cox, 1992; Tibshirani, 1997; Katzman *et al.*, 2018], while distribution estimation based methods predict a time-dependent distribution of survival. Thus, following DeepHit and DSM, we perform time-dependent concordance index [Antolini *et al.*, 2005], which is defined as:

$$CI = Pr\left(W(t_{x_i}|x_i) > W(t_{x_j}|x_j)|t_{x_i} < t_{x_j}\right) \quad (21)$$

where t_{x_i} denotes the true survival time of x_i.

5.3 Implementation Details

For fair comparison, the discrete time space in experiments is set as $\{(0,1],(1,2],\dots,(K-1,K]\}$ following setting of DeepHit and DRSA. According to the maximum time shown in Table 1, t_{max} is set as 400, and $K = t_{max}$.

ISF is implemented with $PyTorch$. Number of hidden units of $E(\cdot)$ defined in Eq. 10 and $H(\cdot)$ defined in Eq. 11 are corresponding set as $\{256, 512, 256\}$ and $\{256, 256, 1\}$ for all experiments.

During training, we perform Adam optimizer. Models of the best CI is selected with variation in hyperparameters of learning rate $\{10^{-3}, 10^{-4}, 10^{-5}\}$, weight of decay $\{10^{-3}, 10^{-4}, 10^{-5}\}$ and batch size $\{8, 16, 32, 64, 128, 256\}$. The influence of ϵ will be discussed in the ablation study.

The reproduction of DeepHit and DRSA is based on the official code of DRSA[1]. And the reproduction of DSM refers to the official package $auto\_survival$[2].

5.4 Performance Comparison

To evaluate performance of ISF, we conduct experiments in three public datasets CLINIC, MUSIC and METABRIC compared with several existing methods. Since compared discrete time space methods DeepHit and DRSA set time points as $t_{i+1}^p = t_i^p + 1$, ϵ in Eq. 15 which controls precision of ISF is set as 1 during training and evaluation for fair comparison.

As shown in Table 2, ISF achieve the best CI in three datasets which censoring rates are 0.132, 0.351 and 0.552. Therefore, ISF is robust to censoring rate. Besides, the large number of samples in MUSIC dataset contributes to performance improvement of ISF, while ISF has relatively low improvement in METABRIC containing fewer samples.

5.5 Ablation Study

For further understanding of ISF, we conduct experiments on ISF with variation of ϵ in Eq. 15 which controls precision to study the effect of precision. As discussed in Section 4.5, ISF predicts $O(NK)$ occurrence probabilities for N samples with K time intervals where $K \propto 1/\epsilon$.

[1]https://github.com/rk2900/drsa

[2]https://autonlab.github.io/auton-survival/models/dsm

Method	CI ↑		
	CLINIC	MUSIC	METABRIC
Cox	$0.525^{\ddagger}$	$0.524^{\ddagger}$	$0.648^{\ddagger}$
	(0.512-0.538)	(0.523-0.525)	(0.634-0.662)
RSF	$0.598^{\ddagger}$	$0.566^{\ddagger}$	$0.672^{\ddagger}$
	(0.594-0.602)	(0.565-0.567)	(0.655-0.689)
DeepSurv	$0.532^{\ddagger}$	$0.578^{\ddagger}$	$0.648^{\ddagger}$
	(0.519-0.545)	(0.574-0.582)	(0.636-0.660)
DeepHit	$0.586^{\ddagger}$	$0.550^{\ddagger}$	$0.677^{\ddagger}$
	(0.567-0.605)	(0.549-0.551)	(0.665-0.688)
DRSA	$0.580^{\ddagger}$	$0.610^{\ddagger}$	$0.692^{\dagger}$
	(0.564-0.596)	(0.601-0.619)	(0.672-0.712)
DSM	$0.598^{\ddagger}$	$0.593^{\ddagger}$	0.697^{*}
	(0.582-0.613)	(0.579-0.606)	(0.677-0.718)
ISF	**0.612**	**0.701**	**0.704**
	(0.596-0.629)	(0.700-0.702)	(0.681-0.728)

∗: $p \geq 0.05$, †: $p < 0.05$, ‡: $p < 0.01$; unpaired t-test with respect to ISF.

Table 2: Comparison of CI (mean and 95% confidence interval) in four public datasets CLINIC, MUSIC and METABRIC.

Training ϵ						
1/10	1/5	1/2	1	2	5	10
0.613	0.614	0.613	0.612	0.613	0.611	0.600

Table 3: CI performance comparison with variation of ϵ during training in CLINIC. During inference, ϵ of all models is fixed to 1 for fair comparison and accurate evaluation.

Training Precision

Since survival time annotations in CLINIC are saved as integer, the ideal ϵ for CLINIC is $\epsilon = 1$. Therefore, we evaluate CI of ISF on CLINIC with variation of ϵ during training in this section. For fair comparison and accurate evaluation, ϵ in inference in this section is fixed to $\epsilon^{Inference} = 1$.

As defined in Eq. 15, ϵ determines precision of ISF. In CLINIC dataset, estimation precision of ISF is higher than annotation precision when $\epsilon^{Train} < 1$ during training. On the contrary, if $\epsilon^{Train} > 1$, annotation precision is higher than estimation precision. In such case, ISF predicts occurrence probabilities at unseen time points.

In Table 3, results of ϵ^{Train} from 0.1 to 10. For $\epsilon^{Train} \in [0.1, 1)$, ISF achieves close CI since estimation precision of these models is higher than annotation precision. For $\epsilon^{Train} \in \{2, 5\}$, the performance is also close to that of ISF with $\epsilon^{Train} = 1$, which indicates that ISF is capable of extrapolating in a certain range of time and robust to ϵ^{Train} variation. In the extreme case of $\epsilon^{Train} = 10$, CI of ISF

Dataset	Inference ϵ			
	1/10	1/5	1/2	1
CLINIC	0.609	0.610	0.612	0.612
MUSIC	0.695	0.696	0.698	0.701
METABRIC	0.703	0.703	0.704	0.704

Table 4: CI performance comparison with variation of ϵ during training in CLINIC, MUSIC and METABRIC. The evaluated ISF is trained with $\epsilon = 1$.

significantly decreases since the maximum survival time in CLINIC is 82.

Inference Precision

In this section, we study generalization ability of ISF with variation of $\epsilon^{Inference}$ during evaluation. Based on ISF trained with $\epsilon^{Train} = 1$, we adjust $\epsilon^{Inference}$ from 0.1 to 1 during inference, and evaluate corresponding CI performance in three public datasets. In $\epsilon^{Inference} < 1$ experiments, ISF predicts conditional hazard rates at time points unseen in training. Hence, results of CI demonstrate generalization ability of ISF.

As shown in Table 4, ISF performance has little decrease when $\epsilon^{Inference} < \epsilon^{Train}$. Hence, ISF has high generalization for occurrence probability prediction at time points beyond the preset discrete time space, which proves that ISF manages to capture patterns of time through representations from sinusoidal positional encoding.

6 Discussion

In this section, we discuss some features of ISF in details.

6.1 Estimation Precision

In this paper, We use a hyperparameter ϵ to control the sampling density of the discrete time space, which has impact on the estimation precision of ISF. Experimental results of the ablation study in Section 5.5 show that ISF with varied ϵ achieves close CI performance in a certain range, even if the estimation precision is lower than annotation precision.

ISF captures time patterns through positional encoding as defined in Eq. 12. Representation based on sinusoids is shift-variation and enables MLP learn high frequency functions [Tancik *et al.*, 2020]. Therefore, ISF manages to extrapolate occurrence probabilities unseen during training.

Although low ϵ leads to high computational complexity as discussed in Section 4.5, the generation ability of ISF enables models trained with relatively high ϵ to generate acceptable results of survival prediction.

6.2 Discrete Time Space

ISF estimates conditional hazard rates in a discrete uniform time space for optimization and inference. For N samples with K time intervals, ISF processes $O(NK)$ pairs of sample and time during training and inference. In this section, we discuss the necessity of uniform time sampling.

In Section 4.4, we maximize occurrence probabilities at time points t_i^p instead of observed time t_x^o. If ISF maximizes

$\hat{p}(t_x^o|x)$ or $\hat{S}(t_x^o|x)$ during optimization, the number and distribution of processed sample-time pairs depends on the training set. In the extreme case that the training set contains N samples with highly discrete survival time, ISF processes $O(N^2K)$ sample-time pairs with numerical integration in K intervals for optimization based on t_x^o. And the distribution of these sample-time pairs relies on the distribution of observed time, which perhaps introduces prior of the survival time distribution in the training set. Though ISF based on the discrete time space replaces the observed time with preset time points, the optimization process is based on adjustable uniform sampling of time. And the adjustment of the discrete time space is independent to the model architecture of ISF.

The ablation study of ϵ also proves that the preset discrete uniform time space based optimization and inference provides enough accuracy for survival analysis. Moreover, the estimation precision of ISF can be easily changed without model architecture modification through variation of hyperparameter ϵ. Hence, occurrence probabilities prediction in a discrete time space through ISF like previous works [Lee *et al.*, 2018; Ren *et al.*, 2019] is reasonable and robust.

6.3 Unified Loss Function

In real-world applications, right-censoring is most common in datasets, which indicates that the true survival time is larger than the observed time t_x^o. Therefore, existing discrete or continuous distribution prediction methods only considers right-censoring in loss functions [Lee *et al.*, 2018; Ren *et al.*, 2019; Nagpal *et al.*, 2021].

Instead of establishing two distinct loss functions for censored and uncensored samples, the proposed loss function uses indicator vector Y defined in Eq. 19 for likelihood calculation. Therefore, a unified loss function defined in Eq. 20 is proposed for both censored and uncensored samples and is easy to be extended for any type of censoring.

7 Conclusion

In this paper, we propose Implicit Survival Function (ISF) for conditional hazard rate estimation in survival analysis. ISF employs sinusoidal positional encoding to capture time patterns. Two MLP are used to encode input covariates and regress conditional hazard rates. For survival distribution estimation, ISF performs numerical integration to approximate CDF for survival rate prediction.

Compared with existing methods, ISF estimates survival distribution without strong assumptions about survival distribution and models a continuous distribution through Implicit Neural Representation. Therefore, ISF models based on different settings of the discrete time space share a common architecture of the network. Moreover, ISF has robustness to estimation precision controlled by the discrete time space whether the estimation precision is higher than the annotation precision or not. Experimental results show that ISF outperforms the state-of-the-art survival analysis models on Concordance Index performance in three public datasets with varied censoring rates.

References

[Antolini *et al.*, 2005] Laura Antolini, Patrizia Boracchi, and Elia Biganzoli. A time-dependent discrimination index for survival data. *Stats in Medicine*, 24(24):3927–3944, 2005.

[Chen *et al.*, 2020] Yinbo Chen, Sifei Liu, and Xiaolong Wang. Learning continuous image representation with local implicit image function. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8624–8634, 2020.

[Courtiol *et al.*, 2019] Pierre Courtiol, Charles Maussion, Matahi Moarii, Elodie Pronier, Samuel Pilcer, Meriem Sefta, Pierre Manceron, Sylvain Toldo, Mikhail Zaslavskiy, Nolwenn Le Stang, Nicolas Girard, Olivier Elemento, Andrew G. Nicholson, Jean-Yves Blay, Françoise Galateau-Sallé, Gilles Wainrib, and Thomas Clozel. Deep learning-based classification of mesothelioma improves prediction of patient outcome. *Nature Medicine*, 25(10):1519–1525, Oct 2019.

[Cox, 1992] David R. Cox. *Regression Models and Life-Tables*, pages 527–541. Springer New York, New York, NY, 1992.

[Curtis *et al.*, 2012] C. Curtis, Sohrab P. Shah, S. Chin, G. Turashvili, O. Rueda, M. Dunning, D. Speed, A. Lynch, Shamith A. Samarajiwa, Yinyin Yuan, S. Gräf, G. Ha, Gholamreza Haffari, A. Bashashati, R. Russell, S. McKinney, A. Langerød, A. Green, E. Provenzano, G. Wishart, S. Pinder, P. Watson, F. Markowetz, L. Murphy, I. Ellis, A. Purushotham, A. Børresen-Dale, J. Brenton, S. Tavaré, C. Caldas, and S. Aparicio. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486:346 – 352, 2012.

[Doksum and Hóyland, 1992] Kjell A. Doksum and Arnljot Hóyland. Models for variable-stress accelerated life testing experiments based on wiener processes and the inverse gaussian distribution. *Technometrics*, 34(1):74–82, 1992.

[Harrell *et al.*, 1982] Jr Harrell, Frank E., Robert M. Califf, David B. Pryor, Kerry L. Lee, and Robert A. Rosati. Evaluating the Yield of Medical Tests. *JAMA*, 247(18):2543–2546, 1982.

[Hochreiter and Schmidhuber, 1997] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Comput.*, 9(8):1735–1780, 1997.

[Ishwaran *et al.*, 2008] Hemant Ishwaran, Udaya B. Kogalur, Eugene H. Blackstone, and Michael S. Lauer. Random survival forests. *The Annals of Applied Statistics*, 2(3):841–860, 2008.

[Jing and Smola, 2017] How Jing and Alexander J. Smola. Neural survival recommender. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, WSDM '17, page 515–524, New York, NY, USA, 2017. Association for Computing Machinery.

[Katzman *et al.*, 2018] Jared L. Katzman, Uri Shaham, Alexander Cloninger, Jonathan Bates, Tingting Jiang, and Yuval Kluger. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18(1):24, Feb 2018.

[Knaus *et al.*, 1995] William A. Knaus, Frank Harrell, Joanne Lynn, Lee M. Goldman, Russell S. Phillips, Alfred F. Connors, Neal V. Dawson, William J. Fulkerson, Robert Califf, Norman A. Desbiens, Peter M. Layde, RobertK. Oye, Paul E. Bellamy, Rosemarie B. Hakim, and Douglas P. Wagner. The support prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Annals of Internal Medicine*, 122:191–203, 1995.

[Lee and Wang, 2003] Elisa T. Lee and John Wenyu Wang. *Statistical Methods for Survival Data Analysis*, volume 476. Wiley Publishing, 2003.

[Lee *et al.*, 2018] Changhee Lee, William R. Zame, Jinsung Yoon, and Mihaela van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. *AAAI*, pages 2314–2321, 2018.

[Li *et al.*, 2016] Yan Li, Jie Wang, Jieping Ye, and Chandan K. Reddy. A multi-task learning formulation for survival analysis. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1715–1724, New York, NY, USA, 2016. Association for Computing Machinery.

[Li *et al.*, 2019] Hongming Li, Pamela Boimel, James Janopaul-Naylor, Haoyu Zhong, Ying Xiao, Edgar Ben-Josef, and Yong Fan. Deep convolutional neural networks for imaging data based survival analysis of rectal cancer. *IEEE International Symposium on Biomedical Imaging*, pages 846–849, 2019.

[Longini *et al.*, 1989] Ira M. Longini, W. Scott Clark, Robert H. Byers, John W. Ward, William W. Darrow, George F. Lemp, and Herbert W. Hethcote. Statistical analysis of the stages of hiv infection using a markov model. *Statistics in Medicine*, 8(7):831–843, 1989.

[Mildenhall *et al.*, 2020] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In *European Conference on Computer Vision*, 2020.

[Nagpal *et al.*, 2021] Chirag Nagpal, Xinyu Li, and Artur Dubrawski. Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks. *IEEE Journal of Biomedical and Health Informatics*, 25(8):3163–3175, 2021.

[Ranganath *et al.*, 2016] Rajesh Ranganath, Adler Perotte, Noémie Elhadad, and David Blei. Deep survival analysis. *Machine Learning for Healthcare Conference*, 56:101–114, 2016.

[Ren *et al.*, 2019] Kan Ren, Jiarui Qin, Lei Zheng, Zhengyu Yang, Weinan Zhang, Lin Qiu, and Yong Yu. Deep recurrent survival analysis. *AAAI*, 33(1):4798–4805, 2019.

[Tancik *et al.*, 2020] Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron,

and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS'20, Red Hook, NY, USA, 2020. Curran Associates Inc.

[Tibshirani, 1997] Robert Tibshirani. The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16(4):385–395, 1997.

[Vaswani *et al.*, 2017] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6000–6010, Red Hook, NY, USA, 2017. Curran Associates Inc.

[Zadeh and Schmid, 2021] Shekoufeh Gorgi Zadeh and Matthias Schmid. Bias in cross-entropy-based training of deep survival networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(9):3126–3137, 2021.

[Zadeh Shirazi *et al.*, 2020] Amin Zadeh Shirazi, Eric Fornaciari, Narjes Sadat Bagherian, Lisa M. Ebert, Barbara Koszyca, and Guillermo A. Gomez. Deepsurvnet: deep survival convolutional network for brain cancer survival rate classification based on histopathological images. *Medical & Biological Engineering & Computing*, 58(5):1031–1045, May 2020.

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