TCSurv:
Time-based Clustering for Reliable Survival Analysis

Anonymous authors
Paper under double-blind review

Abstract

Survival analysis is critical in healthcare for predicting time-to-event outcomes such as disease progression or patient survival. While deep learning excels at capturing meaningful representations from complex clinical data and has improved performance in deep survival models, it inherently struggles with reliability and robustness, challenges that are especially significant when deploying these models in real-world clinical practice. Out-of-distribution (OOD) detection, designed to identify or flag samples that deviate from the training distribution, has become a key method for evaluating AI reliability across fields. This capability is especially important in clinical applications, where noisy or heterogeneous patient data can lead to incorrect assessments; yet, OOD detection remains underexplored and challenging in deep survival analysis due to the need to handle both censored and observed samples, which are unique to this domain. In this study, we address this critical gap by introducing TCSurv, a novel time-base clustering approach for survival analysis that handles both observed and censored samples for robust OOD detection. TCSurv initializes cluster centers using in-distribution data, creating time-specific clusters that anchor model predictions for both observed and censored samples. Experiments in real-world clinical data, including Alzheimer's dementia progression, and benchmark medical imaging datasets demonstrate that TCSurv effectively distinguishes OOD samples without compromising survival performance compared to existing deep survival analysis frameworks. The full code is available at https://anonymous.4open.science/r/TCSurv-F585.

1 Introduction

Survival analysis is a critical statistical tool used to predict the time until the occurrence of a specific event, such as death, disease recurrence, or hardware failure Klein & Goel (2013); Clark et al. (2003); Allison (2010). In healthcare, it offers a trajectory for disease progression until a critical point (i.e. disease onset or death). This is particularly relevant in cases like Alzheimer's dementia, for which there is currently no known cure and brain damage is irreversible, making early prediction through survival analysis essential for proactive clinical intervention.

In recent years, deep learning (DL) approaches have demonstrated exceptional capabilities in estimating complex survival functions from high-dimensional signals such as medical images Wiegrebe et al. (2024). These advancements have enabled researchers to model disease progression from data sources like MR images Thrasher et al. (2024), CT scans Saeed et al. (2024), and whole-slide images Xu et al. (2025). While DL-based survival analysis has achieved notable quantitative success, ensuring the reliability of these models remains a critical challenge Zheng et al. (2023); Jafarzadeh et al. (2021). This concern is especially pressing in medical contexts, where, without reliability guarantees, clinicians may be unable to trust DL-based predictions.

One crucial aspect of ensuring reliability is the ability to detect out-of-distribution (OOD) inputs Jafarzadeh et al. (2021), which are data samples that deviate significantly from the training distribution. OOD detection is essential because deep learning models can produce confident yet incorrect predictions when exposed to unfamiliar, corrupted, or low quality inputs. For example, in survival analysis for Alzheimer's Disease (AD) progression, a model should be able to flag distorted or corrupted MR images for further clinical

review to prevent misleading prognostic outputs. Without such safeguards, incorrect predictions could delay necessary interventions or lead to inappropriate treatment decisions. Therefore, developing robust OOD detection mechanisms is critical for ensuring reliable survival predictions and facilitating the safe translation of DL based models into real world healthcare practice.

In recent years, DL has seen the emergence of various frameworks aimed at identifying and rejecting OOD samples to enhance model reliability. However, most of these algorithms have been developed for relatively static tasks such as classification or detection Yang et al. (2022); Cao et al. (2020); Shi & Lee (2024), overlooking the unique challenges in survival analysis—particularly its temporal dynamics and the presence of censored data. While the fundamental principle of ensuring reliable performance on in-distribution samples remains relevant, adapting OOD detection for survival analysis demands novel methodologies that account for its time-to-event structure. To the best of our knowledge, no prior work has explicitly addressed OOD detection in the context of survival analysis.

Toward this, we propose TCSurv, a novel framework that leverages time-based clustering for OOD detection in survival analysis. TCSurv forms compact, time-specific clusters by guiding sample representations based on their event status: uncensored samples are pulled toward designated time-based anchor points, while censored samples are grouped with clusters beyond their censorship time, ensuring meaningful grouping even under censoring constraints. This approach enables the model to account for temporal dynamics and censoring patterns inherent in survival data. By organizing representations around meaningful time-based clusters, TCSurv enhances the survival model's ability to flag anomalous or unfamiliar inputs as OOD, thereby improving the real-world applicability of survival models.

We evaluate TCSurv across a diverse set of survival analysis datasets, including ADNI (MRI imaging for Alzheimer's disease), FLCHAIN (clinical lab data), GBSG (breast cancer clinical data), and METABRIC (gene expression and clinical features). Additionally, we test on six medical imaging datasets from the MedMNIST collection to further assess performance across modalities and organs. Our results demonstrate that TCSurv effectively distinguishes OOD samples while maintaining strong survival analysis performance, consistently outperforming existing deep survival analysis models. Through comprehensive experiments on diverse datasets, TCSurv shows remarkable robustness and generalizability across multiple medical imaging modalities and clinical data types. These extensive analyses not only validate the efficacy of our time-based clustering approach for OOD detection but also highlight its potential for reliable real-world deployment in survival prediction tasks.

Overall, our contributions are as follows: (1) For the first time, we conduct an OOD detection study within the context of deep survival models, addressing a significant gap in current survival analysis research. (2) We introduce TCSurv, a novel time-based clustering approach specifically designed to improve OOD detection in survival analysis. TCSurv achieves superior OOD detection performance compared to existing methods, without compromising predictive accuracy in survival tasks. (3) We conduct a comprehensive evaluation of our approach across a diverse set of survival analysis healthcare datasets, including a real-world Alzheimer's disease progression dataset, demonstrating the applicability and effectiveness of our proposed method for OOD detection and survival analysis in clinical settings.

2 Related Work

2.1 Survival analysis

Nonparametric deep survival analysis approaches have emerged as promising methods to model survival distributions. Unlike parametric approaches Sheng & Henao (2025); Lee et al. (2023), which make strong assumptions about the underlying distributions, or semi-parametric methods Katzman et al. (2018), which follow a more restrictive proportional hazards assumption, nonparametric models seek to directly estimate the survival function without relying on such assumptions. This is achieved by decomposing the study window into T discrete time intervals and estimating the probability that the event of interest will occur at each time $t \in \{0, 1, ..., T\}$. Under this formulation, Lee & Whitmore (2006) extend standard negative log-likelihood (NLL) to support right-censored data, whereas Fotso (2018b) instead approach it as a multi-task logistic regression (MLTR) problem. However, a study by Kvamme & Ørnulf Borgan (2019) claims that MLTR is

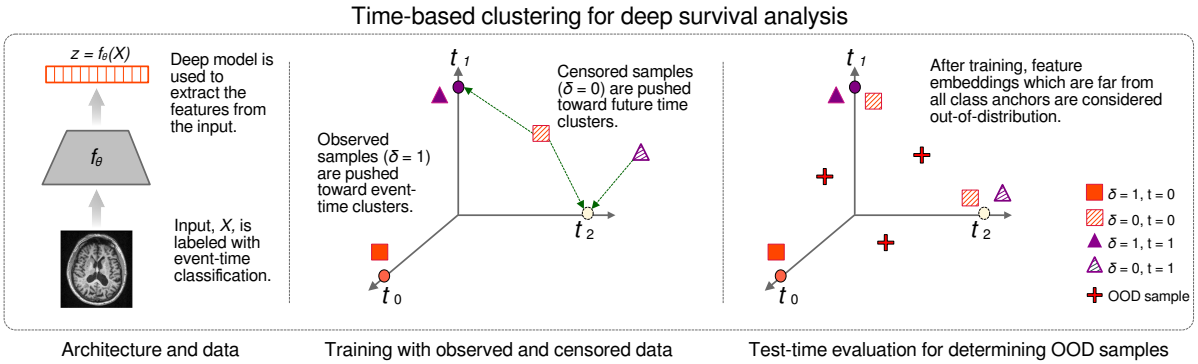

Figure 1: (Left) A deep model, $f_\theta(X)$, extracts features z from an input X. (Center) A distance based loss is applied to the extracted z. Samples where the event is observed ($\delta = 1$) are pulled toward time cluster t_k corresponding to the observation time, while the samples which are censored ($\delta = 0$) are pulled equally toward all time clusters after the time of censoring, reflecting the possibility of the event occurring at some (unknown) time in the future. (Right) Out-of-distribution samples are embedded further away in the feature space.

practically equivalent to NLL, but with an additional reversed cumulative sum. Additionally, other works have noted that utilizing an additional ranking loss component, which seeks to learn the proper ordering of uncensored individuals, can be utilized as a powerful regularizer to further improve the discriminative performance of nonparametric survival models Lee et al. (2018); Kamran & Wiens (2021b). These strategies are being applied across various domains, including medical imaging Nakagawa et al. (2020); Bello et al. (2019); however, no prior work has considered the reliability of these models in terms of OOD detection. Our study, for the first time, considers the problem of OOD detection for survival analysis.

2.2 Out-of-distribution detection

Given the reliability issues associated with DL models, out-of-distribution (OOD) detection has become essential for assessing model reliability Zheng et al. (2023); Jafarzadeh et al. (2021). Significant work has been done in this area, from post-processing techniques that analyze learned predictive distributions (e.g., MaxLogit Zhang & Xiang (2023)) to methods that normalize feature and logit spaces to distinguish in-distribution (ID) from OOD data (e.g., LogitNorm Wei et al. (2022)). Other approaches include open-set recognition and clustering-based methods Liu et al. (2022); Sinhamahapatra et al. (2022). Additionally, benchmark studies have been established to understand and evaluate model performance across various datasets Yang et al. (2022), including medical imaging applications Cao et al. (2020). However, these studies have primarily focused on classification and segmentation tasks, with alarmingly few exploring OOD detection in survival analysis. While some previous work have somewhat explored OOD detection within this domain Loya et al. (2020); Liu et al. (2025), their scope is limited to evaluating OOD through uncertainty quantification, which is often insufficient due to the tendency of DL models to provide overconfident predictions regardless of the input distribution. In contrast, our study targets OOD detection specifically within the survival analysis framework. While it may appear that standard OOD detection techniques developed for classification tasks, such as Max Softmax Probability (MSP) Hendrycks & Gimpel (2017), could be readily applied, this is not the case in survival analysis. Survival models produce time-dependent risk estimates and censored outcomes rather than normalized class probabilities, rendering confidence-based detectors ill-defined and often misleading in this setting.

3 Methods

3.1 Preliminary I: Survival analysis

The objective of survival analysis is to model the time until a specific event occurs, such as disease progression or patient mortality, which is particularly relevant for medical prognosis. Formally, consider a dataset $\mathcal{D} = \{X_i, \delta_i, t_i \mid i \in \mathbb{N}\}$ of size N, where X_i represents features extracted from medical data, $\delta_i \in \{1, 0\}$ is an indicator of whether the event of interest occurred within the observation window (1) or is right-censored (0), and t_i denotes the time of event occurrence or censoring. Importantly, censorship implies that the event was not observed during the study period but may still occur later. The goal is to estimate the survival function $S(t) = P(T > t)$, which expresses the probability that the event has *not* occurred by time t. Estimating $S(t)$ from image-derived features provides critical insights into patient prognosis and disease progression while handling both observed and censored outcomes. We provide more details in the Appendix.

3.2 Preliminary II: OOD detection

OOD detection aims to identify inputs different than the training data distribution, a crucial capability in applications like medical domain where unusual cases may indicate rare conditions, novel pathologies, or even noisy or faulty samples. Given a dataset $\mathcal{D} = \{X_i | i \in \mathbb{N}\}$ of size N, consisting of in-distribution samples, OOD detection seeks a score function $S_c(X)$ that quantifies how likely a new sample X_{new} belongs to the training distribution (also referred as in-distribution (ID) data). A threshold ϕ is then defined, where X is OOD if $S_c(X) > \phi$ and ID otherwise.

3.3 Time-based clustering

To enhance the robustness of survival analysis models, we present a novel approach that leverages time-based clustering to create well-defined feature spaces. We introduce a distance-based objective function specifically designed to learn representations that maintain clear separations between samples associated with different event times, ensuring that embeddings are closely clustered around time-specific anchors. This improves alignment with survival outcomes and naturally positions atypical samples farther from these clusters, resulting in a feature space that is inherently resilient to variations outside the training distribution.

Toward this, we begin by discretizing survival times into T equally spaced intervals, such that each time t maps to a discrete time label $t \in \{0, 1, ..., T\}$. For each sample $\mathbf{X}^*$ with an observed event at time t, we aim to learn a feature representation $\mathbf{z}^* = f_\theta(\mathbf{X}^*)$, where $f_\theta(\cdot)$ is a deep learning model parameterized by θ. The objective is to ensure that this embedding $\mathbf{z}^*$ is positioned close to a predefined anchor c_t. This setup encourages ID data to cluster around their respective time-based anchors, while OOD data is expected to lie farther from these anchors. To achieve this clustering effect, we propose the following objective function, which combines temporal separation and anchor proximity terms to guide embeddings toward their time-specific clusters:

$$\mathcal{L} = \mathcal{L}_T + \lambda \cdot \mathcal{L}_A \tag{1}$$

where $\mathcal{L}_T$ promotes separation between time clusters, and $\mathcal{L}_A$ ensures that embeddings remain close to their respective time anchors. Here, λ is a hyperparameter that controls the relative influence of anchor proximity.

While similar distance-based objective functions have been used in the computer vision literature to learn class-specific representations for tasks like classification and detection Miller et al. (2021), the survival analysis context presents unique challenges. Specifically, the presence of censored samples—instances where the exact event time is unknown—complicates the learning process, as these samples lack precise time labels yet still contribute valuable information. Consequently, our objective function must incorporate a specialized approach that distinguishes between censored and observed samples, ensuring that both types are effectively utilized in training.

As such, for *observed* data points, the calculation of $\mathcal{L}_T$ and $\mathcal{L}_A$ is straightforward, and is expressed as:

$$\mathcal{L}_\mathcal{T} = \log\left[1 + \sum_{j \neq t}^{T} \exp\left(d_t - d_j\right)\right], \tag{2}$$

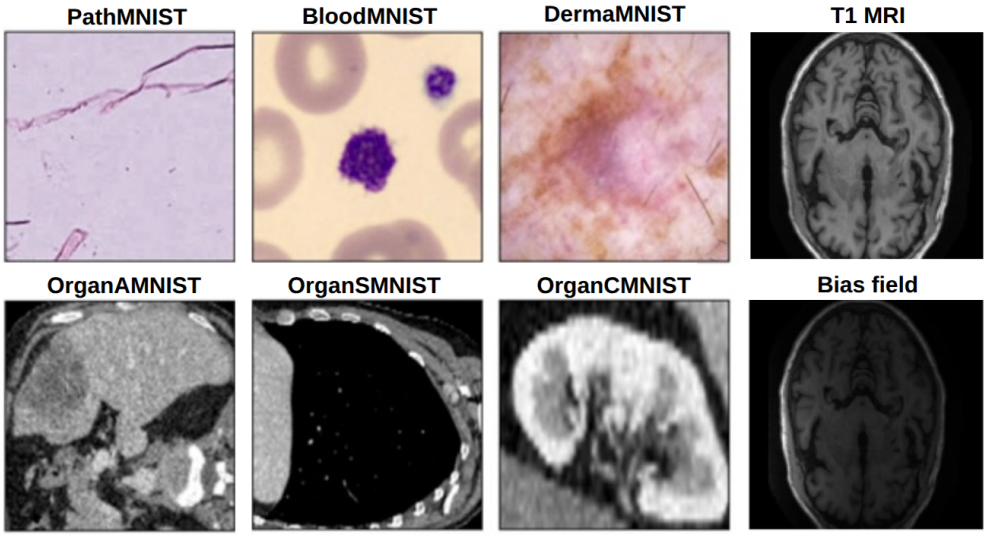

Figure 2: Examples from the ID datasets and their corresponding OOD samples.

$$\mathcal{L}_{\mathcal{A}} = ||f(X) - c_t||_2 \tag{3}$$

where, $d = \{||z - c_1||_2, ||z - c_2||_2, ..., ||z - c_n||_2\}$ is the distance vector which represents all of the distances between an embedding z and every anchor, n is the total number of clusters, and t corresponds to the ground truth time anchor.

Whereas, for *censored* samples, we lack precise event times, making it challenging to assign them directly to a specific time anchor. However, useful survival analysis-based insights can still guide their embedding. For instance, if a sample $\mathbf{X}^*$ is censored at time $t = 3$, we know that $\mathbf{X}^*$ cannot correspond to the time anchors for $\{0, 1, 2, 3\}$ but may belong to any of the anchors in $\{4, 5, ..., T\}$. Additionally, a sample censored at the maximum observation time $t = T$ suggests that the event occurs beyond the observation window. To account for such cases, we introduce an extra time anchor at $t = T + 1$ to capture this uncertainty. To incorporate censored samples effectively, we define modified loss terms, $\mathcal{L}_{\mathcal{T}_c}$ and $\mathcal{L}_{\mathcal{A}_c}$ which extend the original objective in equations Eqns 2 and 3 to support the unique requirements of censored data, as follows:

$$\mathcal{L}_{\mathcal{T}_c} = \frac{\sum_{k=t+1}^{T+1} \log\left[1 + \sum_{t}^{j=0} \exp\left(d_k - d_j\right)\right]}{T - t}, \tag{4}$$

$$\mathcal{L}_{\mathcal{A}_c} = \frac{\sum_{k=t+1}^{T+1} ||f(x) - c_k||_2}{T - t}. \tag{5}$$

Finally, our combined time-based clustering objective ($\mathcal{L}$) for both censored and uncensored samples in the training data is defined as:

$$\mathcal{L} = \underbrace{\mathcal{L}_{\mathcal{T}} + \lambda_1 \mathcal{L}_{\mathcal{A}}}_{\text{Observed samples}} + \underbrace{\mathcal{L}_{\mathcal{T}_c} + \lambda_2 \mathcal{L}_{\mathcal{A}_c}}_{\text{Censored samples}} \tag{6}$$

Prior to training, the time anchors are initialized as the scaled one-hot vector $e_i, \forall i \in \{0, 1, ..., T + 1\}$. The set of anchors is then defined as $C = \{c_1, ..., c_{T+1}\} = \{\alpha \cdot e_1, ..., \alpha \cdot e_{T+1}\}$, where α is a hyperparameter which defines the magnitude of the e_i. It is worth noting that λ_1 and λ_2 could be tuned individually, but we use a single λ by setting $\lambda_1 = \lambda_2$ for our experiments.

Models	C-td (Survival performance)	AUROC (OOD performance)	
		ADNI-T2	Bias
NLL	0.7786 ± .047	0.3361 ± .050	0.5361 ± .028
N-MLTR	0.8128 ± .015	0.4560 ± .069	0.4704 ± .051
RPS	0.8668 ± .025	0.4041 ± .017	0.4992 ± .037
RPS+Rank	0.8576 ± .023	0.3475 ± .036	0.4391 ± .029
DeepHit	0.7770 ± .026	0.4819 ± .067	0.4105 ± .044
SurvRNC	0.6528 ± .079	0.4806 ± .058	0.4118 ± .036
TE-SSL	0.6103 ± .018	0.4690 ± .047	0.3789 ± .032
TCSurv (Ours)	**0.8798 ± .017**	**0.7165 ± .080**	**0.5765± .014**

Table 1: (Left) Survival performance. (Right) OOD performance for ADNI data, averaged across three seeds.

3.4 TCSurv Evaluation

3.4.1 Survival evaluation

TCSurv follows a more recent trend of directly predicting the survival function, in contrast to the proportional hazard assumption used in prior methods such as DeepSurv Katzman et al. (2018). To achieve this, we begin by relocating the initial time anchors to the average locations of the correctly predicted elements. Specifically, for uncensored elements, a prediction is considered correct if it most closely aligns with the anchor corresponding to its ground truth event time. Since censored elements are more ambiguous, we define a prediction as correct if it is assigned to any time anchor *after* the time of censoring, allowing flexibility for model assignments of censored elements. We demonstrate that this action improves overall performance in Sec. 4.2.1. After updating the time anchors, the survival function is modeled as the relative distance between the feature representation and each of the time clusters, given by:

$$\hat{S}(t|X) = \sum_{i=0}^{t} 1 - \sigma(d)_i \tag{7}$$

where $\sigma(\cdot)$ is the soft*min* function applied to the distance vector. The softmin function is the inverse of the softmax function, emphasizing the smallest values in a vector by assigning them higher probabilities. We then quantify our model's discriminative survival performance based on the time-dependent concordance index (C-td) Antolini et al. (2005), enabling a fair comparison of our method against existing approaches.

3.4.2 OOD evaluation

For OOD evaluation, the central idea is that the further away any new input X_{new} is from the time-based anchors, the more likely it is to be an OOD sample. A straightforward way to achieve this would be to simply use the distance d (which represents distance to all the anchors) as the score function. However, prior works in the classification domain Miller et al. (2021) have considered the combination of d and softmin$(\cdot)$ function to achieve a more robust score function. Hence, we also consider a similar strategy for designing the score function for evaluating OOD detection with TCSurv as:

$$S_c(X) = \min\big(d \cdot (1 - \sigma(d))\big). \tag{8}$$

This score function will have minimum value when the inputs have both a low distance and high softmin score. We note that, since baseline methods do not have a notion of time anchors, we initialize the "learned anchors" as the average location of correctly predicted elements similar to the method described in Sec. 3.4.1. This allows us to fit our OOD detector to other survival methods for fairer comparison.

4 Experiments

This section presents our experimental results and ablation analyses. We begin with a description of the datasets, followed by experimental settings and baselines. Finally, we report survival prediction, OOD detection results, and ablation studies to validate the framework.

	C-td	ID:FLCHAIN	
Model	FLCHAIN	GBSG	METABRIC
NLL	**0.8328 ± .004**	0.4371 ± .032	0.4091 ± .030
N-MTLR	0.8321 ± .006	0.4003 ± .020	0.4189 ± .036
RPS	0.8126 ± .005	0.4810 ± .021	0.4565 ± .026
RPS+Rank	0.8197 ± .007	0.4632 ± .063	0.4514 ± .048
DeepHit	0.8389 ± .002	0.4488 ± .054	0.4610 ± .038
SurvRNC	0.7831 ± .009	0.4892 ± .003	0.4596 ± .008
TE-SSL	0.7816 ± .009	0.4772 ± .006	0.4453 ± .018
TCSurv	0.8229 ± .002	**0.7010 ± .041**	**0.7000 ± .021**

Table 2: (Left) Survival performance (Right) OOD performance for non-imaging data.

4.0.1 Alzheimer's Disease progression MRI dataset

We utilize MRI data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) Petersen et al. (2010) for our analysis on real world medical imaging data. ADNI provides follow up information on patients' progression towards Alzheimer's dementia. This dataset includes 1,978 T1-weighted 3D MRI scans used as our ID data, and 919 T2-weighted MRIs treated as OOD. Since survival models are usually trained and evaluated on high-quality MRIs, artifacts from equipment errors naturally represent OOD data. To this end, we also generate additional OOD samples which simulate intensity inhomogeneity, known as bias fields (shown in Fig. 2). These variations create a blurring effect on the MRI, which inhibits feature detection, weakening model performance Jindal et al. (2022). We generated these artifacts using the RandomBias augmentation pipeline from the Torchio library Pérez-García et al. (2021).

4.0.2 Non-imaging data

To reinforce our results, we include additional non-imaging data from *three* standardized survival datasets: 1) Assay of Serum Free Light Chain (**FLCHAIN**) Dispenzieri et al. (2012), 2) German Breast Cancer Study (**GBSG**) Schumacher et al. (1994), and 3) the Molecular Taxonomy of Breast Cancer International Consortium (**METABRIC**) Curtis et al. (2012). These continuous time datasets are then discretized with the pipeline laid out in Kvamme & Ørnulf Borgan (2019). For our experiments, we train on FLCHAIN and treat GBSG and METABRIC as OOD, truncating and padding features as necessary to align with model dimensions. We provide further details on these datasets in the Appendix.

4.0.3 Medical imaging benchmark dataset

We also include medical imaging datasets from the MedMNIST Yang et al. (2023) collection. In specific, we utilize six different datasets, which are divided into two groups. In the first group, we analyze the far-OOD performance on three different modalities: Colon Pathology (**PathMNIST**), Blood Cell Microscope (**BloodMNIST**), and Dermatoscope (**DermaMNIST**) imaging. The second group analyzes near-OOD detection performance when exposed to different 2D views of the same 3D abdominal CT images, namely the Axial (**OrganAMNIST**), Saggital (**OrganSMNIST**), and Coronal (**OrganCMNIST**) views. We provide some examples in Fig. 2 and more details on their statistics in the Appendix.

Synthesizing survival signals: Since benchmark medical datasets lack survival information, we synthesize survival signals by assigning event times and censoring indicators to each class. This overcomes the scarcity, noise, and heavy censoring of real-world survival data and provides a controlled environment with adjustable censorship and known ground truth. For each (X_i, y_i) image-label pair, the ground truth event time t_i^* is initially set to the class label y_i. We then uniformly censor a portion of the data by setting by setting $\delta_i = 0$. However, since, by definition of censorship, it is known that an element cannot be censored at its ground truth event time, we shift the time of censorship to a uniformly random $t \in \{0, 1, ..., t_i^* - 1\}$. Conversely, elements with $t_i^* = 0$ are always observed and are made to be uncensored. Finally, we completely censor all elements where $t_i^* = \max(T)$ to allow the model to provide elements which should be assigned to the "future" time cluster. We prefix these datasets with "Surv-" and drop the "MNIST" (e.g., SurvPath) to distinguish them from standard MedMNIST datasets. We also use "Surv*" as shorthand for the entire set of synthetic survival datasets for simplicity.

	C-td		ID:SurvPath		ID:SurvBlood	
	SurvPath	SurvBlood	SurvBlood	SurvDerma	SurvPath	SurvDerma
NLL	0.9773 ± .004	**0.9954 ± .001**	0.4985 ± .004	0.5004 ± .007	0.5452 ± .011	0.5197 ± .007
N-MTLR	0.9881 ± .002	0.9958 ± .001	0.5076 ± .021	0.5322 ± .025	0.5731 ± .018	0.5174 ± .016
RPS	0.9733 ± .002	0.9936 ± .001	0.5160 ± .004	0.4907 ± .004	0.5459 ± .009	0.5227 ± .008
RPS+Rank	0.9740 ± .002	0.9927 ± .001	0.5125 ± .011	0.4929 ± .008	0.5716 ± .015	0.5256 ± .016
DeepHit	**0.9861 ± .002**	0.9951 ± .001	0.5141 ± .023	0.5113 ± .011	0.5717 ± .008	0.5172 ± .011
SurvRNC	0.9722 ± .003	0.9899 ± .001	0.6131 ± .035	0.5108 ± .044	0.6123 ± .035	0.6335 ± .141
TE-SSL	0.9660 ± .001	0.9878 ± .001	0.5265 ± .034	0.4644 ± .020	0.6479 ± .048	0.5209 ± .012
TCSurv (Ours)	0.9813 ± .001	0.9809 ± .002	**0.7989 ± .042**	**0.7649 ± .016**	**0.8933 ± .020**	**0.9403 ± .016**
	C-td		ID:SurvOrganA		ID:SurvOrganS	
	SurvOrganA	SurvOrganS	SurvOrganS	SurvOrganC	SurvOrganC	SurvOrganC
NLL	**0.9927 ± .001**	0.9671 ± .004	0.5273 ± .005	0.5177 ± .004	0.5272 ± .000	0.5129 ± .001
N-MTLR	0.9921 ± .001	0.9678 ± .004	0.5617 ± .018	0.5406 ± .011	0.5358 ± .012	0.5175 ± .004
RPS	0.9892 ± .001	0.9682 ± .002	0.5515 ± .006	0.5463 ± .005	0.5280 ± .006	0.5197 ± .005
RPS+Rank	0.9920 ± .002	0.9708 ± .003	0.5425 ± .017	0.5359 ± .012	0.5267 ± .003	0.5153 ± .000
DeepHit	0.9918 ± .002	**0.9724 ± .002**	0.5276 ± .005	0.5214 ± .004	0.5364 ± .014	0.5137 ± .005
SurvRNC	0.9853 ± .003	0.9627 ± .004	0.6138 ± .045	0.5737 ± .031	0.5911 ± .016	0.5276 ± .004
TE-SSL	0.9878 ± .002	0.9582 ± .018	0.5525 ± .011	0.5316 ± .003	0.5853 ± .036	0.5229 ± .012
TCSurv (Ours)	0.9809 ± .002	0.9496 ± .002	**0.7453 ± .005**	**0.7032 ± .011**	**0.6098 ± .023**	**0.5572 ± .011**

Table 3: (Left) Survival performance (Right) OOD performance for medical imaging benchmark datasets

4.1 Experimental setup

We evaluate our method against Negative Log-Likelihood (**NLL**) Kvamme & Ørnulf Borgan (2019), Neural Multi-Task Logistic Regression (**N-MTLR**) Fotso (2018a), and Ranked Probability Scoring (**RPS**) Kamran & Wiens (2021a) models. Additionally, some works have explored the use of an additional ranking component, which seeks to learn the proper ordering of uncensored individuals. As such, we include **DeepHit** ($\mathcal{L}_{NLL} + \mathcal{L}_{Ranking}$) Lee et al. (2018) and **RPS+Rank** ($\mathcal{L}_{RPS} + \mathcal{L}_{Ranking}$) Kamran & Wiens (2021a) in our baseline suite. Finally, since TCSurv can be viewed as a representation learning method, we include two recent approaches: **SurvRNC** ($\mathcal{L}_{DeepHit} + \gamma\mathcal{L}_{SurvRNC}$) Saeed et al. (2024) and **TE-SSL** ($\mathcal{L}_{DeepHit} + \gamma\mathcal{L}_{TESSL}$) Thrasher et al. (2024), to ensure that our improved OOD detection is not merely due to better feature learning, but rather to learning a more optimal feature space.

All models were optimized using the Adam algorithm Kingma (2014) with a learning rate Φ. We use a 3D CNN Liu et al. (2020) as the backbone model for ADNI, ResNet18 for the Surv* scenario, and a feed forward neural network Kvamme & Ørnulf Borgan (2019) for the nonimaging datasets. ADNI were trained with $\alpha = 1$, $\lambda = 0.5$, $\Phi = 1 \times 10^{-4}$, Surv* with $\alpha, \lambda = 1$, $\Phi = 1 \times 10^{-4}$, and FLCHAIN with $\alpha, \lambda = 1$, $\Phi = 1^{-3}$. All experiments were conducted using NVIDIA A30, A40, and RTX 4500 Ada Generation GPUs. Results are reported as the mean and standard deviation across three random seeds. We provide more details on each baseline model in Appendix C.

4.2 Results

We present the primary results for the ADNI, non-imaging, and Surv* datasets in Tables 1, 2, and 3, respectively. In each table, we show the OOD results (AUROC) on the right and the corresponding survival results (C-td) on held-out in-distribution data on the left.

First, as shown across nearly all experiments, the OOD detection performance of TCSurv is significantly higher, achieving up to a 49% increase in the ADNI experiments, 43% with non-imaging data, and 48% with the Surv* suite compared to top performing baselines. This improvement in OOD detection can be attributed to TCSurv's explicit design that encourages in-distribution (ID) training data to form tight clusters around time-dependent anchor points. To empirically observe this behavior, we plot the distance to the nearest time-based center for our method in Figure 3 and compare it to the baselines. Here, we clearly observe that baseline methods fail to distinguish between ID and OOD data. In some cases (e.g., NLL), the logit space

		SurvPathMNIST			SurvOrganAMNIST		
		C-td	AUROC (SurvBloodMNIST)	AUROC (SurvDermaMNIST)	C-td	AUROC (SurvOrganSMNIST)	AUROC (SurvOrganCMNIST)
α	0.1	0.9725	0.7292	**0.8371**	0.9583	0.6743	0.6510
	1	**0.9836**	**0.8434**	0.7914	**0.9794**	0.7495	0.7132
	5	0.9671	0.7086	0.4976	0.9548	**0.7499**	**0.7163**
	10	0.9295	0.5430	0.3329	0.9021	0.7342	0.6919
λ	0	0.8557	0.5200	0.3724	0.9299	0.7403	0.6912
	.25	**0.9837**	0.8371	0.7834	**0.9860**	0.7838	0.7356
	.5	0.9836	0.7832	0.8082	0.9844	0.7659	0.7258
	.75	**0.9837**	**0.8751**	**0.8505**	0.9828	0.7536	0.7157
	1	0.9832	0.8112	0.8433	0.9809	**0.7909**	**0.7498**

Table 4: Ablation of α and λ for Surv{Path, OrganA}.

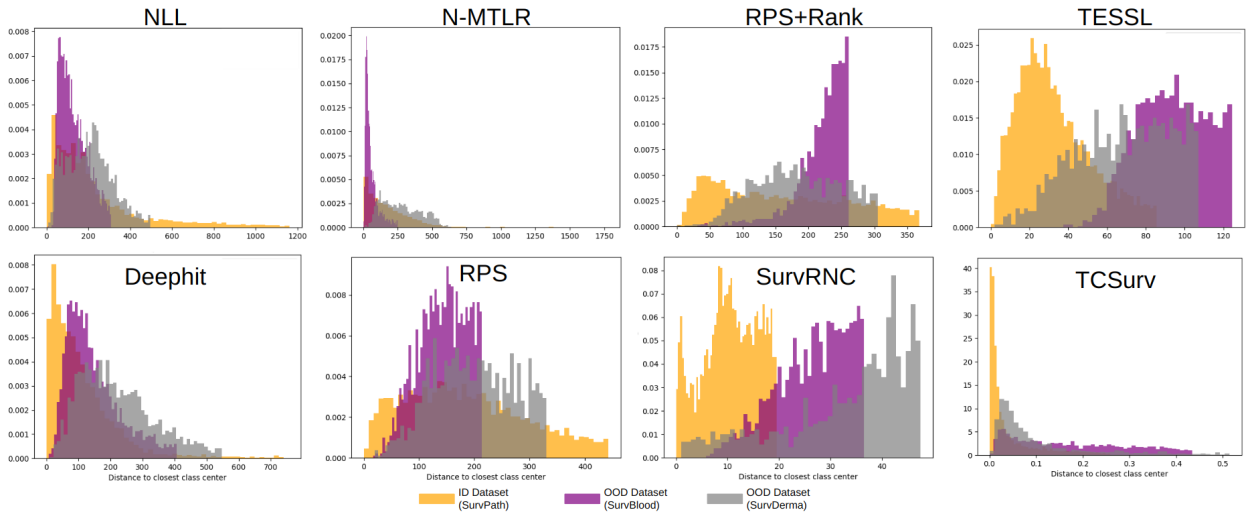

Figure 3: TCSurv allows the ID dataset to cluster more tightly around the time-based centers, while the distances of OOD samples to the nearest time-based centers are more spread out compared to the baselines.

even places ID data farther from the center than OOD data. In contrast, TCSurv effectively achieves tighter clusters for ID data, while OOD data remain more dispersed. We additionally provide distance plots for the representative non-imaging experiments in the Appendix D.

Second, in terms of survival performance, TCSurv demonstrates results that are comparable and competitive with existing baselines. For example, on ADNI, a real-world dataset for Alzheimer's dementia, TCSurv achieves the best performance. On other datasets, with the exception of SurvOrganS, its survival outcomes remain closely aligned with those of the strongest baseline methods, with only marginal differences. Importantly, as discussed earlier, TCSurv consistently outperforms baselines in OOD detection across all datasets. This is particularly critical in real-world applications, where the reliability and robustness of a model are just as important as its accuracy.

4.2.1 Ablation analysis

α & λ analysis: We examine the effect of α, which controls the magnitude of the class anchor, and λ, which balances the contributions of $\mathcal{L}_{\mathcal{T}}$ and $\mathcal{L}_{\mathcal{A}}$, on both survival performance and OOD detection. The results, presented in Table 4 (top), indicate that the model generally prefers smaller α values, suggesting that a compact feature space is well-suited for both survival analysis and OOD detection.

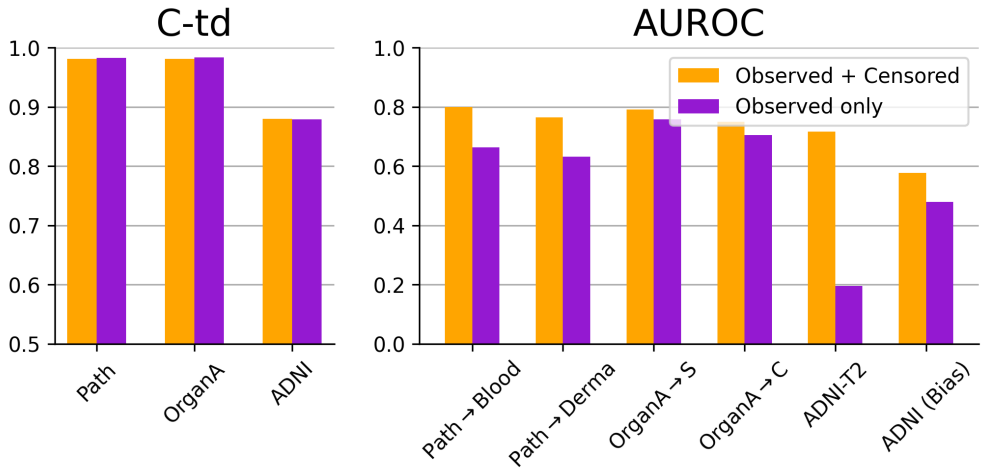

Figure 4: Analysis of the effect of including censored samples when updating anchor locations for both survival performance (left) and OOD detection (right).

In Table 4 (bottom), we analyze the effect of λ to understand the role of anchor loss. As seen for both datasets and in both survival analysis and OOD performance, when $\lambda = 0$ (corresponding to zero anchor loss), the performance is poor, demonstrating the importance of $\mathcal{L}_A$.

Censored samples in determining the anchor position: TCSurv repositions the time anchors based on correctly predicted samples. As discussed in Sec 3.4.1, we consider both observed and censored elements for relocation, rather than updating centers solely based on observed ones. As shown in Fig. 4, while including censored samples does not significantly impact survival performance on benchmark datasets, it proves crucial for all OOD evaluations, where censored data plays an important role in determining accurate anchor positions. This is particularly apparent in the evaluation of ADNI-T2 data, where the model is entirely unable to predict OOD samples when anchors are relocated based solely on observed data.

Censorship ratio: As discussed in Sec. 4.0.3, synthetic data allows us to conduct a more thorough analysis of our model's behavior as the level of censorship varies. As such, Figure 5 depicts TCSurv's performance across varying levels of censorship. From the four plots on the right, it can be seen that our method consistently achieves superior AUROC on the OOD detection task when compared to the baseline models. It can also be seen from the two leftmost plots that survival performance from TCSurv maintains steady and competitive performance. While it is slightly below some of the baseline methods, we argue that TCSurv still performs strongly overall on the survival task, and that the massively improved OOD performance justifies a slight loss in the survival task since reliability is equally crucial in real-world applications. But we also acknowledge this as a limitation of our method as Eqs. 4 and 5 handle censored data in a more restrictive manner than the unbounded approaches in the baseline models. Finding a better balance between OOD detection performance and a less restrictive mechanism for censored data could be a valuable area for future work.

5 Discussion

In this work, we address the problem of OOD detection head-on in survival analysis for the first time by introducing TCSurv, a novel time-based clustering framework. TCSurv organizes representations of both observed and censored samples into distinct, compact time-based clusters, enhancing the model's ability to distinguish ID images from those in unseen distributions, thus improving OOD detection in survival analysis. The approach incorporates an objective function that promotes separation between time clusters while ensuring that learned embeddings remain close to their respective time anchors. Through extensive evaluations across a real-world MRI dataset, three clinical lab feature datasets, and six medical imaging benchmark datasets, TCSurv consistently improves OOD detection without compromising survival performance. Further, uur experiments demonstrate that TCSurv consistently creates compact clusters around the designated

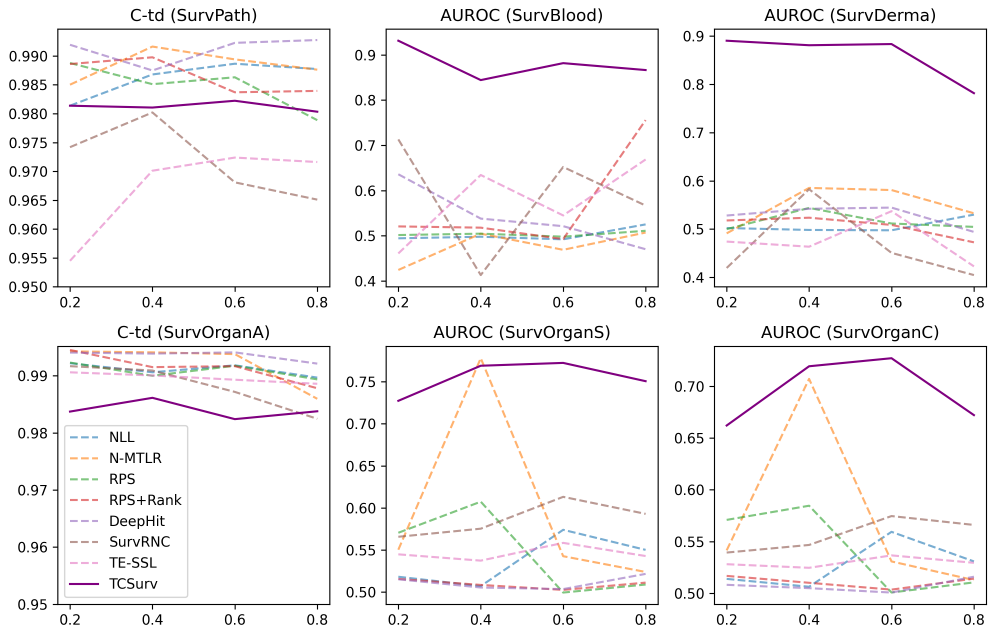

Figure 5: Analysis of model performance with varying levels of censorship. (Top) Trained on SurvPath as ID data and Surv{Blood, Derma} as OOD data. (Bottom) Trained on SurvOrganA as ID data and SurvOrgan{S, C} as OOD data.

time anchors and is robust the high degrees of censorship. Among the OOD detection tasks studied, the ability of our method to identify artifact-filled MRIs as OOD is particularly significant, as such artifacts, caused by equipment interference, can severely degrade image quality and diagnostic accuracy Noda et al. (2022).

However, this study has certain limitations that point to important directions for future research. While prior work has explored OOD detection in survival analysis using epistemic uncertainty quantification Loya et al. (2020); Liu et al. (2025), this work does not focus on uncertainty-based approaches. A key reason is that uncertainty estimates alone do not reliably capture distributional shifts in unseen data, as modern deep neural networks are often poorly calibrated and tend to exhibit overconfident predictions even under significant distributional mismatch Wang et al. (2021). In addition, our experimental formulation simplifies the clinical progression modeled in the ADNI dataset, which contains labels for Cognitively Normal (CN), Mild Cognitive Impairment (MCI), and Alzheimer's Dementia (AD) samples. To frame it as a survival task, we merged CN and MCI into a single category, predicting AD directly. Extending TCSurv to handle longitudinal survival data, thereby modeling transitions from CN to MCI to AD, could be a valuable future direction with significant translational implications. Since longitudinal methods follow a single subject over time, ensuring reliability across multiple visits is another important challenge to consider. Additionally, we note that the current anchor initialization method yields equidistant time centers, which may imply that all event times are equally spaced (i.e. $t = 1$ is as close to $t = 2$ as it is $t = 10$). While this formulation is effective, organizing anchors such that centroids are closer to temporally nearby anchors than distance ones could offer further improvement to the survival modeling process. We also note that since TCSurv is designed as a nonparametric approach, future work should explore OOD detection for parametric and semi-parametric methods. Lastly, our analysis using the Surv* datasets could be expanded to develop a standardized benchmark for survival-based OOD detection. This would likely attract broad interest and help drive progress in addressing this critical reliability challenge.

References

Paul D Allison. Survival analysis. *The reviewer's guide to quantitative methods in the social sciences*, pp. 413–425, 2010.

Laura Antolini, Patrizia Boracchi, and Elia M Biganzoli. A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24, 2005. URL https://api.semanticscholar.org/CorpusID:25663825.

Dara Bahri, Heinrich Jiang, Yi Tay, and Donald Metzler. Scarf: Self-supervised contrastive learning using random feature corruption, 2022. URL https://arxiv.org/abs/2106.15147.

Ghalib A Bello, Timothy JW Dawes, Jinming Duan, Carlo Biffi, Antonio De Marvao, Luke SGE Howard, J Simon R Gibbs, Martin R Wilkins, Stuart A Cook, Daniel Rueckert, et al. Deep-learning cardiac motion analysis for human survival prediction. *Nature machine intelligence*, 1(2):95–104, 2019.

Tianshi Cao, Chin-Wei Huang, David Yu-Tung Hui, and Joseph Paul Cohen. A benchmark of medical out of distribution detection. *arXiv preprint arXiv:2007.04250*, 2020.

Taane G Clark, Michael J Bradburn, Sharon B Love, and Douglas G Altman. Survival analysis part i: basic concepts and first analyses. *British journal of cancer*, 89(2):232–238, 2003.

Christina Curtis, Sohrab P Shah, Suet-Feung Chin, Gulisa Turashvili, Oscar M Rueda, Mark J Dunning, Doug Speed, Andy G Lynch, Shamith Samarajiwa, Yinyin Yuan, Stefan Gräf, Gavin Ha, Gholamreza Haffari, Ali Bashashati, Roslin Russell, Steven McKinney, METABRIC Group, Anita Langerød, Andrew Green, Elena Provenzano, Gordon Wishart, Sarah Pinder, Peter Watson, Florian Markowetz, Leigh Murphy, Ian Ellis, Arnie Purushotham, Anne-Lise Børresen-Dale, James D Brenton, Simon Tavaré, Carlos Caldas, and Samuel Aparicio. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, 486(7403):346–352, April 2012.

Angela Dispenzieri, Jerry A Katzmann, Robert A Kyle, Dirk R Larson, Terry M Therneau, Colin L Colby, Raynell J Clark, Graham P Mead, Shaji Kumar, L Joseph Melton, 3rd, and S Vincent Rajkumar. Use of nonclonal serum immunoglobulin free light chains to predict overall survival in the general population. *Mayo Clin Proc*, 87(6):517–523, June 2012.

Stephane Fotso. Deep neural networks for survival analysis based on a multi-task framework, 2018a. URL https://arxiv.org/abs/1801.05512.

Stephane Fotso. Deep neural networks for survival analysis based on a multi-task framework. *arXiv preprint arXiv:1801.05512*, 2018b.

Dan Hendrycks and Kevin Gimpel. A baseline for detecting misclassified and out-of-distribution examples in neural networks. *Proceedings of International Conference on Learning Representations*, 2017.

Mohsen Jafarzadeh, Touqeer Ahmad, Akshay Raj Dhamija, Chunchun Li, Steve Cruz, and Terrance E Boult. Automatic open-world reliability assessment. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 1984–1993, 2021.

Sumit Kumar Jindal, Sayak Banerjee, Ritayan Patra, and Arin Paul. 9 - deep learning-based brain malignant neoplasm classification using mri image segmentation assisted by bias field correction and histogram equalization. In Jyotismita Chaki (ed.), *Brain Tumor MRI Image Segmentation Using Deep Learning Techniques*, pp. 135–161. Academic Press, 2022. ISBN 978-0-323-91171-9. doi: https://doi.org/10.1016/B978-0-323-91171-9.00008-9. URL https://www.sciencedirect.com/science/article/pii/B9780323911719000089.

Fahad Kamran and Jenna Wiens. Estimating calibrated individualized survival curves with deep learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pp. 240–248, 2021a.

Fahad Kamran and Jenna Wiens. Estimating calibrated individualized survival curves with deep learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(1):240–248, May 2021b. doi: 10.1609/aaai.v35i1.16098. URL https://ojs.aaai.org/index.php/AAAI/article/view/16098.

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*, 18:1–12, 2018.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

John P Klein and Prem K Goel. Survival analysis: state of the art. 2013.

Håvard Kvamme and Ørnulf Borgan. Continuous and discrete-time survival prediction with neural networks, 2019. URL https://arxiv.org/abs/1910.06724.

Changhee Lee, William Zame, Jinsung Yoon, and Mihaela van der Schaar. Deephit: A deep learning approach to survival analysis with competing risks. *Proceedings of the AAAI Conference on Artificial Intelligence*, 32 (1), Apr. 2018. doi: 10.1609/aaai.v32i1.11842. URL https://ojs.aaai.org/index.php/AAAI/article/view/11842.

Hyunjun Lee, Junhyun Lee, Taehwa Choi, Jaewoo Kang, and Sangbum Choi. Towards flexible time-to-event modeling: Optimizing neural networks via rank regression, 2023. URL https://arxiv.org/abs/2307.08044.

Mei-Ling Ting Lee and G. A. Whitmore. Threshold regression for survival analysis: Modeling event times by a stochastic process reaching a boundary. *Statistical Science*, 21(4), November 2006. ISSN 0883-4237. doi: 10.1214/088342306000000330. URL http://dx.doi.org/10.1214/088342306000000330.

Sheng Liu, Chhavi Yadav, Carlos Fernandez-Granda, Narges Razavian, Adrian V Dalca, Matthew Mcdermott, Emily Alsentzer, Sam Finlayson, Michael Oberst, Fabian Falck, and Brett Beaulieu-Jones. On the design of convolutional neural networks for automatic detection of alzheimer's disease, 2020. URL https://github.com/NYUMedML/.

Yen-Cheng Liu, Chih-Yao Ma, Xiaoliang Dai, Junjiao Tian, Peter Vajda, Zijian He, and Zsolt Kira. Open-set semi-supervised object detection. In *European Conference on Computer Vision*, pp. 143–159. Springer, 2022.

Yu Liu, Weiyao Tao, Tong Xia, Simon Knight, and Tingting Zhu. Survunc: A meta-model based uncertainty quantification framework for survival analysis. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, pp. 1903–1914, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714542. doi: 10.1145/3711896.3737140. URL https://doi.org/10.1145/3711896.3737140.

Hrushikesh Loya, Pranav Poduval, Deepak Anand, Neeraj Kumar, and Amit Sethi. Uncertainty estimation in cancer survival prediction, 2020. URL https://arxiv.org/abs/2003.08573.

Dimity Miller, Niko Suenderhauf, Michael Milford, and Feras Dayoub. Class anchor clustering: A loss for distance-based open set recognition. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3570–3578, 2021.

Tomonori Nakagawa, Manabu Ishida, Junpei Naito, Atsushi Nagai, Shuhei Yamaguchi, Keiichi Onoda, and Alzheimer's Disease Neuroimaging Initiative. Prediction of conversion to alzheimer's disease using deep survival analysis of mri images. *Brain communications*, 2(1):fcaa057, 2020.

Chikara Noda, Bharath Ambale Venkatesh, Jennifer D Wagner, Yoko Kato, Jason M Ortman, and João AC Lima. Primer on commonly occurring mri artifacts and how to overcome them. *Radiographics*, 42(3): E102–E103, 2022.

Fernando Pérez-García, Rachel Sparks, and Sébastien Ourselin. TorchIO: a Python library for efficient loading, preprocessing, augmentation and patch-based sampling of medical images in deep learning. *Computer Methods and Programs in Biomedicine*, pp. 106236, 2021. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2021.106236. URL https://www.sciencedirect.com/science/article/pii/S0169260721003102.

R. C. Petersen, P. S. Aisen, L. A. Beckett, M. C. Donohue, A. C. Gamst, D. J. Harvey, C. R. Jack, W. J. Jagust, L. M. Shaw, A. W. Toga, and et al. Alzheimer's disease neuroimaging initiative (adni). *Neurology*, 74(3):201–209, Jan 2010. doi: 10.1212/wnl.0b013e3181cb3e25.

Numan Saeed, Muhammad Ridzuan, Fadillah Adamsyah Maani, Hussain Alasmawi, Karthik Nandakumar, and Mohammad Yaqub. Survrnc: Learning ordered representations for survival prediction using rank-n-contrast. *arXiv preprint arXiv:2403.10603*, 2024.

M Schumacher, G Bastert, H Bojar, K Hübner, M Olschewski, W Sauerbrei, C Schmoor, C Beyerle, R L Neumann, and H F Rauschecker. Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. german breast cancer study group. *J Clin Oncol*, 12(10):2086–2093, October 1994.

Deming Sheng and Ricardo Henao. Learning survival distributions with the asymmetric laplace distribution. *arXiv preprint arXiv:2505.03712*, 2025.

Xiangxi Shi and Stefan Lee. Benchmarking out-of-distribution detection in visual question answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5485–5495, 2024.

Poulami Sinhamahapatra, Rajat Koner, Karsten Roscher, and Stephan Günnemann. Is it all a cluster game?–exploring out-of-distribution detection based on clustering in the embedding space. *arXiv preprint arXiv:2203.08549*, 2022.

Jacob Thrasher, Alina Devkota, Ahmad P. Tafti, Binod Bhattarai, and Prashnna Gyawali. Te-ssl: Time and event-aware self supervised learning for alzheimer's disease progression analysis. In Marius George Linguraru, Qi Dou, Aasa Feragen, Stamatia Giannarou, Ben Glocker, Karim Lekadir, and Julia A. Schnabel (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2024*, pp. 324–333, Cham, 2024. Springer Nature Switzerland. ISBN 978-3-031-72390-2.

Deng-Bao Wang, Lei Feng, and Min-Ling Zhang. Rethinking calibration of deep neural networks: Do not be afraid of overconfidence. *Advances in Neural Information Processing Systems*, 34:11809–11820, 2021.

Hongxin Wei, Renchunzi Xie, Hao Cheng, Lei Feng, Bo An, and Yixuan Li. Mitigating neural network overconfidence with logit normalization. In *International conference on machine learning*, pp. 23631–23644. PMLR, 2022.

Simon Wiegrebe, Philipp Kopper, Raphael Sonabend, Bernd Bischl, and Andreas Bender. Deep learning for survival analysis: a review. *Artificial Intelligence Review*, 57(3):65, 2024.

Yingxue Xu, Fengtao Zhou, Chenyu Zhao, Yihui Wang, Can Yang, and Hao Chen. Distilled prompt learning for incomplete multimodal survival prediction, 2025. URL `https://arxiv.org/abs/2503.01653`.

Jiancheng Yang, Rui Shi, Donglai Wei, Zequan Liu, Lin Zhao, Bilian Ke, Hanspeter Pfister, and Bingbing Ni. Medmnist v2-a large-scale lightweight benchmark for 2d and 3d biomedical image classification. *Scientific Data*, 10(1):41, 2023.

Jingkang Yang, Pengyun Wang, Dejian Zou, Zitang Zhou, Kunyuan Ding, Wenxuan Peng, Haoqi Wang, Guangyao Chen, Bo Li, Yiyou Sun, et al. Openood: Benchmarking generalized out-of-distribution detection. *Advances in Neural Information Processing Systems*, 35:32598–32611, 2022.

Zihan Zhang and Xiang Xiang. Decoupling maxlogit for out-of-distribution detection. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3388–3397, 2023. doi: 10.1109/CVPR52729.2023.00330.

Haotian Zheng, Qizhou Wang, Zhen Fang, Xiaobo Xia, Feng Liu, Tongliang Liu, and Bo Han. Out-of-distribution detection learning with unreliable out-of-distribution sources. *Advances in Neural Information Processing Systems*, 36:72110–72123, 2023.

Appendix

A Survival analysis

Survival analysis is a statistical tool used to model the time it takes until a specific event occurs. Survival data are collected in follow-up studies, where a subject is monitored until the event of interest has occurred, at which point the time between the initial visit and event occurrence is recorded as the survival outcome. Importantly, not all survival outcomes can be recorded due to the study ending prior to event observation or a subject dropping out of the study for some other reason. These instances are called *right-censored* samples, and are a crucial aspect to survival analysis because it is unknown if/when the event will occur in the future. In these cases, we instead record the last visit as the time of censorship ($\delta = 0$) since it is at least known that this subject has not experienced an event by this time.

Formally, consider subject data X. The ground truth event time can be sampled from a distribution $t^* \in \mathcal{T}$. Similarly, the time of censorship can be sampled from $c \in \mathcal{C}$. We choose the subject's event indicator to be $\delta = 1$ if $t^* \leq c$ and 0 otherwise. Then, the corresponding event time is $t = \min(t^*, c)$. Note that it is typically assumed that $\mathcal{T} \perp\!\!\!\perp \mathcal{C}|X$ for right-censored data.

A.1 Survival evaluation

The Concordance Index (CI) evaluates the model's ability to correctly rank the event times of two samples, given the information. However, CI is only computed at the initial time of observation and does not reflect the change in risk over time Lee et al. (2018). Instead, the time-dependent concordance index (C-td) evaluates the average concordance across *each time interval* to provide a more holistic view of the risk across the study. To compute C-td, we first must define the Cumulative Incidence Function (CIF), which expresses the probability that the event will occur on or before some time t. This is essentially the cumulative sum of the $S(t|X)$, and is defined as:

$$F(t|X) = \sum_{i=0}^{t} \sigma(f_\theta(X))_i \tag{9}$$

where $\sigma(\cdot)$ is the Softmax (or softmin for TCSurv) function. We then define C-td as:

$$Ctd = \frac{A_{i,j} * 1(F(t_i|X_i) > F(t_i|X_j)}{\sum_{i \neq j} A_{i,j}} \tag{10}$$

Here, $A_{i,j}$ is a function which determines whether a pair of elements i, j are comparable, defined as $A_{i,j} = 1(\delta_i = 1, t_i < t_j)$. Higher is better for C-td, where 1 corresponds to perfect ranking, 0.5 is complete randomness, and 0 is inversely perfect.

B Detailed dataset descriptions

B.1 ADNI

The ANDI dataset includes a cohort of 493 unique patients from the Alzheimer's Disease Neuroimaging Initiative (ADNI) Petersen et al. (2010). Each participant has one or more visits that include 3D T1-weighted MRI scans, resulting in a total of 1,978 data points. A subset of patient visits additionally include a T2-weighted MRI, which we utilize as OOD data during evaluation. At each visit, patients are classified as cognitively normal (CN), having mild cognitive impairment (MCI), or diagnosed with Alzheimer's dementia (AD). We designate as converters those individuals who were initially CN or MCI but developed AD over the course of the study. Additionally, each visit records the number of months since the baseline observation, serving as a time-to-event indicator, with a 68.7% censor rate. The data underwent preprocessing following

Dataset	Modality (#classes)	# Train / Val / Test
Path	Colon Pathology (9)	89,996/10,004/7,180
Blood	Blood Cell Microscope (8)	11,959/1,712/3,421
Derma	Dermatoscope (7)	- / - /2,005
OrganA	Abdominal CT - Axial (11)	34,561/6,491/17,778
OrganS	Abdominal CT - Sagittal (11)	13,932 / 2,452 / 8,827
OrganC	Abdominal CT - Coronal (11)	- / - / 8,216

Table A1: Overview of medical imaging benchmark datasets Yang et al. (2023).

the pipeline outlined in Liu et al. (2020) and were split by unique participants to prevent data leakage. For patients with multiple visits, each visit is treated as an independent data point.

B.2 Nonimaging data

All nonimaging datasets included in this study are continuous time survival datasets which were binned into 10 time windows using LabTransDiscreteTime from the Pycox Kvamme & Ørnulf Borgan (2019) library. For OOD evaluation, we truncated/padded the features to align every dataset with the model, as needed.

B.2.1 FLCHAIN

This dataset contains 7874 individuals with 7 features over a time horizon of 5166 days and studies the relationship between serum free light chain (FLC) and mortality. FLCHAIN is approximately 69.9% right-censored.

B.2.2 GBSG

GBSG contains follow up information for 473 patients with a 56 month time horizon and 7 features. The event of interest for GBSG is death due to breast cancer. Finally, GBSG has is 43.2% censored.

B.2.3 METABRIC

METABRIC also contains follow up information for breast cancer mortality. It contains 9 gene expression features with a time horizon of 355 months and a censor rate of 42.1%.

B.3 Synthetic data

Table A1 provides more details on MedMNIST datasets that were used to generate the synthetic data. Additionally, Figure A1 provides a visualization of our synthetic data generation approach.

C Futher implementation details

Table A2 outlines all of the hyperparameter configurations used in our experiments. We now provide more details on the baseline methods, hyperparameters, and model architectures.

C.1 Baseline methods

We now provide more details on the baseline methods used for comparison.

C.2 Standard survival methods

C.2.1 Negative Log Likelihood (NLL)

This is an adaptation of log-likelihood proposed by Lee & Whitmore (2006) for right-censored survival data, defined as:

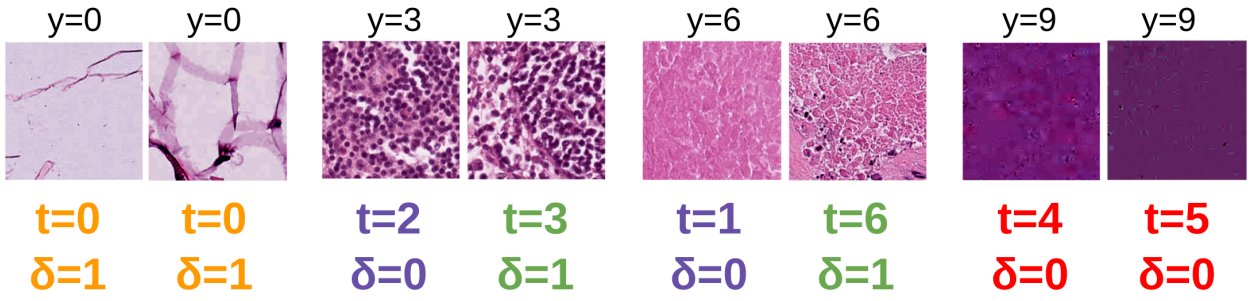

Figure A1: Schematic outlining synthetic data generation, where y is the class label, t is the time of event/censoring, t^* is the *ground truth* event time, and δ is the event indicator.

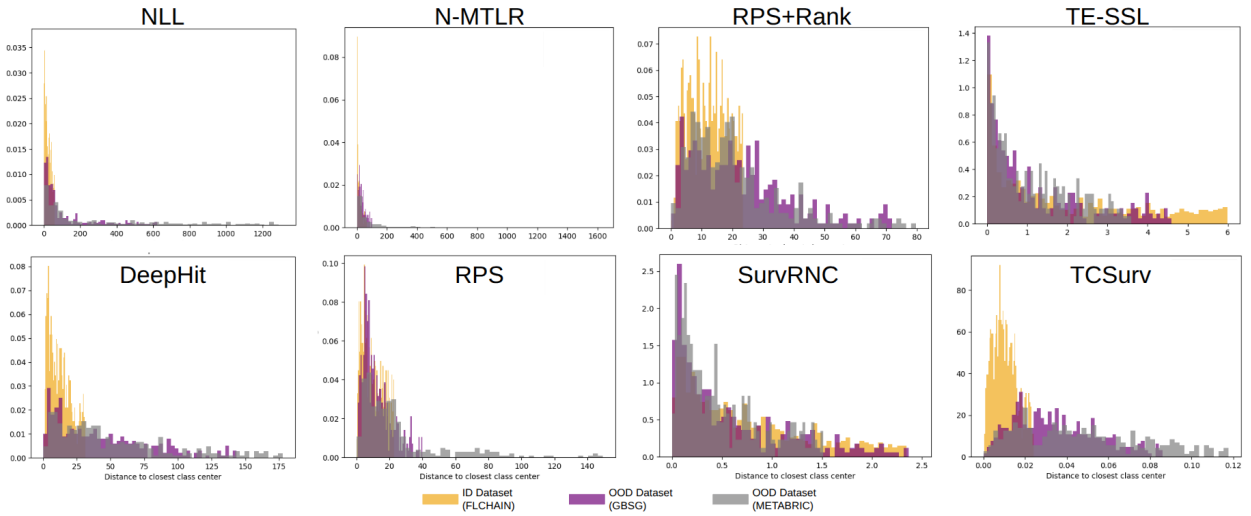

Figure A2: Histograms showing the distance of elements to their closest center for FLCHAIN experiment.

$$\mathcal{L}_{NLL} = -\sum_{i=1}^{B} \delta_i \log \hat{y}_i + (1 - \delta_i) \log \sum_{j=t_i+1}^{T} \hat{y}_j \tag{11}$$

	ADNI	FLCHAIN	Surv*
Optimizer	Adam	Adam	Adam
Learning rate	1×10^{-4}	1×10^{-3}	1×10^{-4}
Batch Size	32	32	128
α	1	1	1
λ	0.5	1	1
γ_1	0.5	0.25	0.5
γ_2	0.2	0.25	0.5
β	0.5	0.9	0.5

Table A2: Complete hyperparameter configuration for experiments.

where B is the batch size and $\hat{y}_i = \text{softmax}(f_\theta(X_i))$. Specifically, for observed elements, $\mathcal{L}_{NLL}$ seeks to maximize the predicted probability at the observed event time. Then, for censored elements, it instead maximizes the sum of all probabilities *after* the time of censorship.

C.2.2 Neural Multi-task Logistic Regression (N-MTLR)

According to Kvamme & Ørnulf Borgan (2019), N-MTLR is practically the equivalent to NLL loss, but with a reversed cumulative sum applied to the output of the neural network.

C.2.3 Ranked Probability Scoring (RPS)

Instead of focusing only on an element's event time, RPS loss instead applies to the entire time horizon, defined as:

$$\begin{aligned}\mathcal{L}_{RPS} = &\sum_{i=1}^{B} \delta_i \sum_{k=1}^{T} (\hat{S}(k|X_i) - 1_{k<t_i})^2 \\ &+ (1-\delta_i) \sum_{k=1}^{t_i} (\hat{S}(k|X_i) - 1)^2\end{aligned} \tag{12}$$

For uncensored individuals, the survival probability is pushed toward 1 at all times before the event time (rather than just *at* the event time) and is averaged over the entire time horizon. Then, for censored individuals, RPS loss still moves the survival probability towards 1 at all times before censorship, but then is only averaged up to the time of censorship (t_i) since the outcome is unknown after such time.

C.2.4 Ranking loss methods

Ranking loss is analogous to concordance index, in that it penalizes the model for an incorrect ordering of two concordant individuals. Formally, it is defined as:

$$\mathcal{L}_{ranking} = \sum_{i \neq j}^{B} A_{i,j} \exp \frac{-(\hat{F}(t_i|X_i) - \hat{F}(t_i|X_j))}{\zeta} \tag{13}$$

Some works Lee et al. (2018); Kamran & Wiens (2021b) have demonstrated that the inclusion on such loss can increase the discriminative performance of survival models.

C.2.5 Representation learning methods

There has been recent interest in applying contrastive learning-based losses to the survival context to improve the model's overall understanding of the training data, and therefore improve the survival performance. To evaluate the OOD performance of such models, we include SurvRNC Saeed et al. (2024) and TE-SSL Thrasher et al. (2024) in our baselines.

Since contrastive loss methods require a second, augmented view of the input batch, we apply the following to our data: For MRIs, we use random flip with a probability of 50%, and random affine to randomly rotate the input by 90 degrees in each direction. For the Surv* experiments, we apply color jitter and random crop. Finally, we use SCARF feature corruption Bahri et al. (2022) for the nonimaging datasets, with a corruption rate of 60%. SurvRNC and TE-SSL are then applied regularization functions with DeepHit loss as the prognosis function. Specifically $\mathcal{L}_{DeepHit} + \gamma_1 \mathcal{L}_{SurvRNC}$ and $\mathcal{L}_{DeepHit} + \gamma_2 \mathcal{L}_{TESSL}$, respectfully. We define the γ values in Table A2.

Layer	Details
Linear	(num features, 32)
ReLU	-
BatchNorm1D	-
Dropout	10%
Linear	(32, 32)
ReLU	-
BatchNorm1D	-
Dropout	10%
Linear	(32, 10)

Table A3: Details for non-imaging MLP network.

C.2.6 A note on the performance of SurvRNC and TE-SSL

We observe in our experiments that SurvRNC and TE-SSL both underperformed compared to the rest of the baselines. We believe that this could indicate that contrastive learning based survival methods may generally be sensitive to hyperparameter and environment configurations.

C.3 Architecture details

C.3.1 3D CNN for ADNI data

The 3D CNN used for the ADNI data takes $X \in \mathbb{R}^{B\times 96\times 96\times 96}$ and uses a series of four convolutional layers to create the embedding $z \in \mathbb{R}^{B\times 1024}$, where B is the batch size. This network utilizes kernel sizes that are unconventionally small compared to other models such as ResNet to preserve fine-grained details in the MRI data, which has been demonstrated to improve AD classification performance. Refer to Liu et al. (2020) for further details.

C.3.2 MLP for non-imaging data

We use a simple multi-layer perceptron (MLP) network from Kvamme & Ørnulf Borgan (2019) for the non-imaging data. Table A3 provides complete details.

D Additional Results

We present the additional distance histograms for experiments with FLCHAIN as the ID dataset in Figure A2. Here we can clearly see that TCSurv creates the tightest clusters around the time anchors and embeds OOD samples further away, thereby improving overall OOD detection. The baselines, on the other hand, tend to create a more spread representation space, making OOD detection more challenging.