Deep Learning for Survival Analysis: A Review

Simon Wiegrebe[1,3,4], Philipp Kopper[2,3], Raphael Sonabend[5], and Andreas Bender[2,3]

[1]*Statistical Consulting Unit StaBLab, Department of Statistics, LMU Munich, Munich, Germany.*
[2]*Department of Statistics, LMU Munich, Munich, Germany.*
[3]*Munich Center for Machine Learning, LMU Munich, Munich, Germany.*
[4]*Department of Genetic Epidemiology, University of Regensburg, Regensburg, Germany.*
[5]*MRC Centre for Global Infectious Disease Analysis, Jameel Institute, Imperial College London, School of Public Health, London, UK.*

Abstract

The influx of deep learning (DL) techniques into the field of survival analysis in recent years, coupled with the increasing availability of high-dimensional omics data and unstructured data like images or text, has led to substantial methodological progress; for instance, learning from such high-dimensional or unstructured data. Numerous modern DL-based survival methods have been developed since the mid-2010s; however, they often address only a small subset of scenarios in the time-to-event data setting - e.g., single-risk right-censored survival tasks - and neglect to incorporate more complex (and common) settings. Partially, this is due to a lack of exchange between experts in the respective fields.

In this work, we provide a comprehensive systematic review of DL-based methods for time-to-event analysis, characterizing them according to both survival- and DL-related attributes. In doing so, we hope to provide a helpful overview to practitioners who are interested in DL techniques applicable to their specific use case as well as to enable researchers from both fields to identify directions for future investigation. We provide a detailed characterization of the methods included in this review as an open-source, interactive table: https://survival-org.github.io/DL4Survival. As this research area is advancing rapidly, we encourage the research community to contribute to keeping the information up to date.

1 Introduction

The term *survival analysis* (SA), or equivalently *time-to-event analysis*, summarizes a set of techniques that enable the unbiased estimation of the (possibly improper) distribution of outcome variables that are partially censored, truncated, or both. Usually, the outcome is given by the time until the occurrence of an event such as death, system failure, or time to remission.

Non-parametric methods like the Kaplan-Meier estimator (Kaplan et al., 1958) provide a set of baseline tools still used today, yet semi-parametric methods received the most attention historically, in particular the Cox proportional hazards regression model (Cox, 1972) and its extensions. Both non-parametric and semi-parametric methods make no assumptions about the underlying distribution of event times and are popular baseline models in benchmark experiments as they are easy to fit and perform reasonably well. Since the early 2000s, Machine Learning (ML) methods have been successfully adapted to survival tasks: e.g., the Random Survival Forest (Ishwaran et al., 2008), boosting-based methods (Binder et al., 2008), and support vector machines (Van Belle et al., 2011); see Wang et al. (2019b) for an overview.

More recently, Deep Learning (DL) methods have expanded to the field of SA: The first neural networks (NNs) had already been applied to survival tasks in the 1990s (Faraggi et al., 1995; Brown et al., 1997), yet most modern deep survival models have been developed only since the late 2010s - fueled by the emergence of high-dimensional non-tabular input data, e.g., due to reduced costs for genotyping. Since then, many DL-based methods for a variety of different use cases and data settings have been developed; and while some papers do provide an overview of existing methods within specific contexts (e.g., Schwarzer et al., 2000; Deepa et al., 2022), to the best of our knowledge there is no general systematic review of DL-based survival methods and their properties. Such a systematic review should help channel methodological efforts

dimension	example: *DeepSurv*	section(s)
estimation	Cox-based with hazard rate parametrization	section 4.2.1
neural network architecture	FFNN	sections 4.1 and 4.2.2
outcome modalities	right-censoring	sections 2.2.1 and 4.3.1
feature modalities	-	sections 2.2.2 and 4.3.2
interpretability	-	section 4.4
reproducibility	code and data freely accessible	section 4.5
means of evaluation	synthetic and real-world data	section 4.5

Table 1: Overview of theoretical and practical dimensions reviewed, illustrated by the example of *DeepSurv* (Katzman et al., 2018).

toward the most promising fields, direct research to areas with gaps, and also equip practitioners with a comprehensive overview of suitable methods for any given survival task.

Motivated by the above, in this paper we aim to provide a comprehensive review of currently available DL-based survival methods, addressing both theoretical dimensions, such as model class and NN architecture, as well as data-related aspects, such as outcome and feature modalities supported (see Section 4 for definitions). Table 1 gives an overview of the dimensions we consider for one exemplary DL-based survival method.

This paper is structured as follows. Section 2 briefly introduces common SA notation and frequently encountered outcome and feature modalities of survival data. Section 3 summarizes historical developments within SA techniques, from simple non-parametric approaches to modern DL-based methods. Section 4 first outlines the review methodology, then explains general NN architecture choices in SA (Section 4.1), and subsequently provides a detailed, comprehensive overview of all methods reviewed, covering the dimensions estimation and network architecture (Section 4.2), supported survival tasks in terms of outcome and feature modalities (Section 4.3), interpretability (Section 4.4) as well as reproducibility and means of evaluation (Section 4.5); findings are summarized in the *Main Table* (https://survival-org.github.io/DL4Survival). Finally, Section 5 concludes.

2 Survival Analysis - Theoretical Concepts and Data Modalities

In this section, we first introduce some basic quantities that quantify the distribution of a non-negative variable $T > 0$, which are often the targets of estimation in SA (see Section 2.1). Later, we describe censoring and truncation, which need to be accounted for in order to estimate these quantities (see Section 2.2).

2.1 Targets of Estimation

Initially, assume that T is continuous. Let $f_T(t)$ and $F_T(t) := P(T \leq t)$ be its density and cumulative distribution function (CDF), respectively. Then, the survival function of T is defined as

$$S_T(t) := P(T > t) = 1 - F_T(t), \tag{1}$$

i.e., as the probability of surviving at least until time t. The hazard rate,

$$h_T(t) := \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq T < t + \Delta t | T \geq t) = \frac{f_T(t)}{S_T(t)}, \tag{2}$$

is the instantaneous risk of the event occurring given it has not yet occurred at time t. Both survival function and hazard rate are often targets of estimation in SA and fully characterize the distribution of T. Finally,

the cumulative hazard, defined as

$$H_T(t) := \int_0^t h_T(u)du = -\log(S_T(t)), \tag{3}$$

is the integral over the hazard rate, from 0 to t, and is often used as an intermediate step when calculating the survival probability.

In the above, T was assumed to be continuous. However, sometimes the time scale is discrete by nature (e.g., grade level at the time of school dropout) or a continuous time scale is discretized into intervals either because of substantive considerations (e.g., if only weekly or monthly information on event times is available or of interest) or for computational simplification. With discrete event times, the discrete hazard

$$h_T(t) := P(T = t|T \geq t),\ t = 1, 2, \ldots \tag{4}$$

is the conditional probability of the event occurring in the time interval t, conditional upon the individual still being alive at the beginning of t. This gives rise to the discrete-time survival probability

$$S_T(t) := P(T > t) = \prod_{j=1}^{t}(1 - h_T(j)), \tag{5}$$

i.e., the product of the complements of the discrete-time hazards across all time intervals (or time points) $j = 1, \ldots, t$. For further background on discrete time analysis from a statistical modeling point of view, see Tutz et al. (2016).

2.2 Data Modalities

We now define different data modalities, in terms of both outcomes and features, that are frequently encountered in real-world survival tasks. In Section 4.3, we provide detailed information regarding which of the reviewed methods can handle the respective modalities. We will see that many methods do not immediately support settings beyond right-censored data without further pre- or post-processing.

2.2.1 Outcome Modalities

Throughout this work, we consider a sample of size n and refer to a single $i \in \{1, \ldots, n\}$ as *individual* or *subject*. Let $T_i > 0$ be the non-negative random variable which represents the time until the event of interest for subject i occurs. We want to estimate the distribution of T_i given the p-dimensional feature vector $\mathbf{x}_i$. However, T_i often cannot be fully observed because the time-to-event is right-, left- or interval-censored. Let C_i^L and C_i^R be the left- and right-censoring times and L_i and R_i the endpoints of the censoring interval for subject i, respectively. For an interval-censored observation, we have $T_i \in (L_i, R_i]$ as we only know that the event occurs within the interval, but not the exact time. Right-censoring $T_i \in (L_i = C_i^R, \infty]$ and left-censoring $T_i \in (L_i = 0, R_i = C_i^L]$ can thus be considered special cases of interval-censoring.

Time-to-event data is also often subject to truncation. In SA truncation implies that subjects are either not part of the dataset at all or not part of the risk set for a specific event at certain time points. It can be a result of the selection criteria of a study or induced by the type of time-to-event process (primarily relevant in multi-state scenarios, including recurrent events). Formally, let T_i^L and T_i^R be the left- and right-truncation times, respectively. Left-truncation occurs when $T_i^R = \infty$, then subjects with $T_i < T_i^L$ never enter the study. Similarly, observations are right-truncated when $T_i^L = 0$ and subjects with $T_i > T_i^R$ are not observed. This is often the case when data is sampled from registry data.

Survival tasks are not restricted to single-risk scenarios. Another outcome modality is thus whether multiple, potentially competing events can occur. In a *competing risks* scenario each individual can experience only one of at least two distinct, mutually exclusive events, for example, death in hospital versus hospital discharge. More generally, in a *multi-state* setting multiple (transient and terminal) events (states) are possible, as well as certain transitions between them, e.g., transitions between different stages of an illness with death as their terminal event.

A final outcome-related modality is referred to as *recurrent events*. Typically, we record a single outcome (censoring or event) for each individual. However, when conditions such as epilepsy or sports injuries are

being modeled, subjects may experience events repeatedly. While recurrent events can be viewed as a special case of multi-state models, they can also be treated as standard survival data with subject-specific correlation (Box-Steffensmeier et al., 2006). Table 2 provides an overview and examples of the outcome modalities discussed in this section.

type of outcome modality	**example**
right-censoring	dropout in clinical trials: for some individuals, the exact event time is unobserved since they drop out of the study at some point
left-censoring	age at which children learn a certain task: some children already know the task at the beginning of the study but it is unknown at which age they learned it
interval-censoring	medical study with a periodic follow-up: exact event times are unknown, only the interval (between follow-up k and $k+1$) is known
right-truncation	transfusion-induced AIDS onset study (Klein et al., 1997): only patients developing AIDS from transfusion before the registry sampling date are included, while patients with onset after that date are right-truncated
left-truncation	coumarin abortion study (Meister et al., 2008): only women conscious of their pregnancy are included; women who had a spontaneous abortion before their pregnancy is recognized never enter the study
competing risks	study on dialysis mortality (Noordzij et al., 2013): the event of interest, death on dialysis, is precluded by the competing event kidney transplantation
multi-state	study on kidney failure: for all patients, transitions between the states *healthy*, *dialysis*, *kidney transplantation* and *death* are possible, sometimes even bidirectionally (e.g., between *dialysis* and *kidney transplantation*)
recurrent event	incidence of pneumonia in young children (Ramjith et al., 2021): the occurrence of multiple, recurrent pneumonia episodes is possible, with episodes within an individual child's history not being independent

Table 2: Overview of different outcome modalities.

2.2.2 Feature modalities

Features such as weight or lifestyle factors may vary over time. Similarly, feature effects on survival time may be time-dependent. *Time-varying features* (TVFs) and *time-varying effects* (TVEs) constitute deviations from the proportional hazards (PH) assumption (a very common assumption in SA, see Section 3) and, thus, need to be taken into account for survival modeling: with TVFs, the features themselves are observed at multiple time points and are hence time-dependent; with TVEs, the effect a feature has on the outcome (e.g., on the hazard rate or survival probability) varies over time.

The dimensionality of data input constitutes another important feature modality. Due to the prominence of SA in the life sciences, features derived from high-dimensional data - *omics* data in particular - are sometimes employed to predict and explain survival times. In order for a method to learn from such a high-dimensional feature space, the model architecture needs to be adapted, usually with appropriate penalization techniques.

Multimodality is the final feature-related data modality we consider. In the life sciences, in particular, we are oftentimes not restricted to structured tabular data (e.g., clinical patient data), but also have access to image data (e.g., CT scans) or unstructured text data (e.g., written doctor's notes); that is, the feature set is multimodal and special techniques are required to extract and combine information from it.

3 A Short History of Survival Methods

Featureless learners are particularly useful when approaching survival analysis. The Kaplan-Meier (KM) estimator (Kaplan et al., 1958) and the Nelson-Aalen (NA) estimator (Nelson, 1969; Nelson, 1972; Aalen,

1978) are popular non-parametric approaches to estimating even-time distributions, via estimation of the survival function (1) and the cumulative hazard (3), respectively. These are often used as baseline models for comparisons or as building blocks of more advanced techniques (Sonabend, 2021).

Instead of estimating the survival function directly, many popular statistical methods focus on parametrizing the hazard rate (2). Most prominently, the Cox PH regression (Cox, 1972) models the hazard rate at time t, conditional on features $\mathbf{x}$, as the product of a non-parametrically estimated baseline hazard $h_0(t)$ and the exponentiated log-risk $\eta = g(\mathbf{x})$ (with $g()$ being the link function):

$$h(t|\mathbf{x}) = h_0(t) \exp(\eta = g(\mathbf{x})). \tag{6}$$

In comparison studies, the Cox PH model is also often used as a baseline model for ML- and DL-based benchmark experiments usually in its most basic version with $\eta = \mathbf{x}^\top \boldsymbol{\beta}$ and PH assumption. However, the (DL-based) Cox model can be extended to also include interactions, non-linear effects, stratification, time-varying effects, and even unstructured components $d(\mathbf{z})$, yielding the generalized log-risk

$$\eta = g(\mathbf{x}, \mathbf{z}, t) = f(\mathbf{x}, t) + \gamma_1 d_1(\mathbf{z}_1) + \cdots + \gamma_G d_G(\mathbf{z}_G), \tag{7}$$

where $f(\mathbf{x}, t)$ denotes potentially non-linear, time-varying effects of tabular features $\mathbf{x}$ as well as their interactions. $d_g(\mathbf{z}_g)$ denotes embeddings learned in the deep part(s) of the model from unstructured data sources $\mathbf{z}_g$, $g \in \{1, \ldots, G\}$, such as images or text.

Piecewise Exponential Models (PEMs) also parametrize the hazard rate as in (6) or (7), however, the baseline hazard is parametrized and estimated alongside the feature-related coefficients. This is done by partitioning the time axis into J intervals and assuming piecewise constant hazards in each interval. Friedman (1982) showed that the likelihood of this model is proportional to a Poisson likelihood, which implies that, after appropriate data transformation, any method capable of minimizing a negative Poisson log-likelihood can also be used for various survival tasks (Bender et al., 2021). Note that while PEM-based approaches partition the follow-up into intervals, it is a method for continuous time-to-event data as it takes the full information about the event times into account.

Parametric survival models, such as the Accelerated Failure Time (AFT) model (Kalbfleisch et al., 2011), assume event times to follow a certain statistical distribution characterized by a set of parameters. Based on the distribution-specific likelihood, parametric survival models then estimate these distributional parameters as a function of features $\mathbf{x}$. We can write, e.g., the density for an event at time t as

$$f(t|\boldsymbol{\theta}(\mathbf{x})), \ t \geq 0, \tag{8}$$

with parameter vector $\boldsymbol{\theta}(\mathbf{x})$. That is, all distributional parameters (e.g., both shape and scale of a Weibull distribution) can be estimated as a function of $\mathbf{x}$; for instance, Campanella et al. (2022) use transformation models to do so. Once the parameters are estimated, any quantity of interest such as $h(t|\mathbf{x})$ or $S(t|\mathbf{x})$ can be calculated based on the distributional assumption.

Finally, discrete-time survival methods consider the time-to-event data to be a succession of binary outcomes. To do so, the time axis (if not discrete) is first partitioned into intervals, with $T = t$ implying the event occurred in interval $(a_{t-1}, a_t]$. Binary event indicators y_{it} are then defined for each time interval t and used as outcomes. For individual i, the discrete hazard $h_i(t|\mathbf{x}_i)$ in interval t is then

$$h_i(t|\mathbf{x}_i) = P(y_{it} = 1 | T \geq t, \mathbf{x}_i). \tag{9}$$

Analogously to PEM-based approaches, any ML algorithm that is applicable to binary outcomes can be used for discrete-time survival modeling after data transformation. The logit model, for instance, uses a logistic response function to model the probability of the event taking place in t (i.e., the discrete hazard (4)), conditional on feature values $\mathbf{x}$.

Since the late 2000s, ML-based survival methods increasingly started gaining traction, including tree-based

techniques like the Random Survival Forest (RSF) by Ishwaran et al. (2008), Support Vector Machines (Khan et al., 2008; Van Belle et al., 2011), and boosting methods (Binder et al., 2008). Through improved modeling of non-linear relationships and interactions, these methods often outperform traditional statistical models in terms of predictive power (Steele et al., 2018); on the downside, however, they are usually not interpretable directly, which makes them unattractive in fields where interpretability is crucial. Since this paper focuses on DL-based survival methods, we do not further go into detail here and instead refer the interested reader to the general overview of ML-based techniques by Wang et al. (2019b) and the comprehensive review in Sonabend (2021).

Many DL-based methods for SA have been developed in recent years. They usually build upon one of the aforementioned statistical survival approaches, while harnessing advantages of NNs such as non-linear and interaction modeling as well as (implicit) feature selection and engineering. Furthermore, recent advances in multimodal and multitask learning as well as interpretability have made DL-based survival methods even more attractive for many common survival tasks.

The earliest DL-based survival techniques date back to the mid-1990s (Liestbl et al., 1994; Faraggi et al., 1995; Brown et al., 1997) and are usually simple NN-based extensions of classical statistical survival methods discussed above. Likely the most prominent early NN-based survival model is the one by Faraggi et al. (1995), which builds upon the Cox PH model by replacing the linear predictor $g(\mathbf{x}) = \mathbf{x}^\top \boldsymbol{\beta}$ in (6) by an FFNN. The model employs only a shallow NN due to computational restrictions at the time. Liestbl et al. (1994) propose implementing the PEM as an NN, yet without any hidden layers. Going one step further, the *PEANN* model by Fornili et al. (2013) parametrizes the piecewise constant hazards by a shallow FFNN. Similarly, *PLANN* (Biganzoli et al., 1998) is an NN-based extension of the discrete-time logit model, parametrizing the discrete hazard by an FFNN.

4 Deep Learning in Survival Analysis

For this review, we initially conducted a formal literature screening using Web of Science. As the focus of this paper is to give an overview of distinct available methods and their capabilities, not to summarize (novel) applications of already existing techniques, we designed the following two-step literature screening process. In the first step (inclusion criteria), we searched Web of Science for the Topic

```
 ("survival analysis" or "time-to-event analysis" or "survival data" or
"time-to-event data") AND
("neural network" or "deep learning") AND
("model" or "method") AND
("performance" or "evaluation" or "comparison" or "benchmark")
```

with the cutoff date being December 31, 2022. Our inclusion criteria were thus chosen to be rather ample, resulting in a total of 211 articles. In the second step (exclusion criteria), we excluded all articles which did not satisfy all of the following four conditions:

(a) Development of a new method beyond the mere application of an already existing method to new data or contexts.

(b) Evaluation of performance results on at least one non-private benchmark dataset.

(c) Performance evaluation using metrics that are designed for time-to-event data, such as C-index, Integrated Brier Score (IBS), and others.

(d) Focus on estimation and prediction in the context of time-to-event data and learning the parameters of the model within the NN architecture.

Criteria (a), (b), and (c) aim to ensure that the paper in question develops a new method rather than applying a known method to new data or in a new context. Criterion (b) is a good complement to (a), as the predictive utility is often illustrated via benchmark experiments when new methods are proposed. Additionally, this

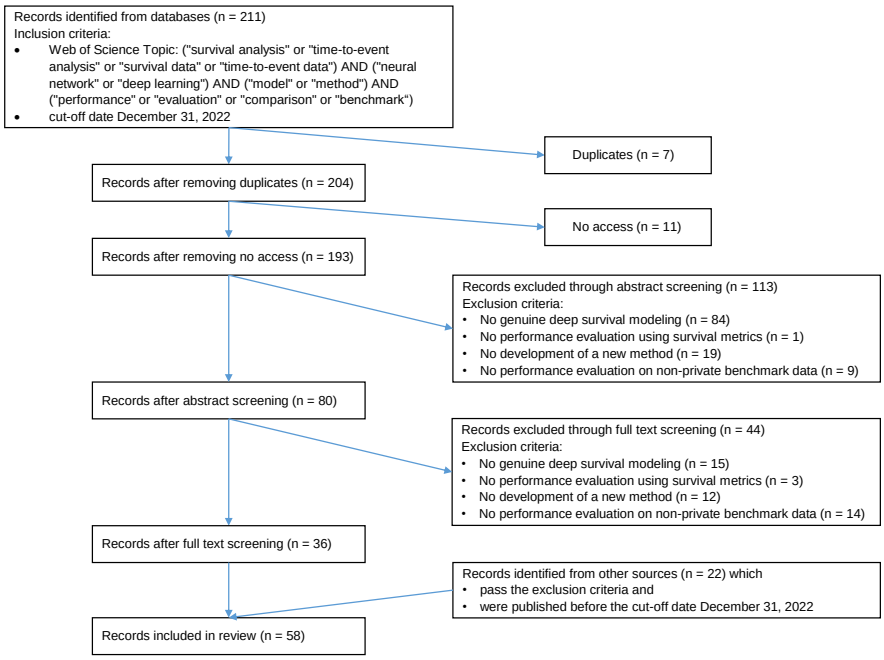

Figure 1: PRISMA diagram for literature screening of deep learning-based survival methods.

criterion introduces an open science aspect and ensures that at least one empirical comparison could be replicated in theory. Criterion (c) ensures that benchmark analyses actually focus on methods that model time-to-event data, as some papers that passed the initial screening eventually ignored the time-to-event nature of the data. Finally, criterion (d) aims to exclude two-step approaches where DL is used solely for feature extraction, while survival modeling is done using non-DL approaches with the extracted features outside of the NN.

Subsequently, we combined the resulting selection of articles with additional papers that had otherwise come to our attention and that fulfilled the above criteria, yielding a total of 58 articles - and thus, 58 distinct methods. The inclusion, exclusion, and screening process is visualized in the PRISMA diagram in Figure 1.

The following naming scheme is used in the remainder of the paper to reference individual methods/papers: the method name as specified in the publication, if provided and unique; if the method name is not unique, we append a suffix (the first three letters of the first author's last name followed by the year of publication) with an underscore; if no method name is provided, we use this suffix as a name. All methods are summarized in our *Main Table* (https://survival-org.github.io/DL4Survival).

Real-world survival tasks always come with their own idiosyncrasies, based on which the required capabilities of the survival methods are determined. For instance, a tabular medical dataset with multiple disease outcomes requires competing risk modeling but does not necessitate multimodal data techniques. In the following, we thus aim to provide a summary of the 58 methods based on a broad range of both theoretical and practical model characteristics.

4.1 Architectural Choices of Neural Networks in Survival Analysis

Feed-forward neural networks (FFNNs) were the earliest type of NN architecture, dating back to the 1940s (McCulloch et al., 1943). Within an FFNN, information passes from the input nodes through a user-specified number of hidden layers until the output nodes. The information only flows forward as there are no cyclical patterns or loops. The main contribution of FFNNs is their flexibility to model interaction effects and perform implicit feature selection and engineering. In the survival context, FFNNs naturally allow for a more flexible estimation of, e.g., (semi-)parametric hazard rates, as well as for the incorporation of TVEs and TVFs (in

theory); for instance, the hazard rate in (6) can be estimated much more flexibly by parametrizing $g(\mathbf{x})$ through an NN. At the same time, the FFNN architecture also contains multiple limitations: for instance, learning from multimodal data input - in particular, image data - is not possible when only employing an FFNN architecture. FFNNs constitute the main architecture of most early DL-based survival methods and still serve as a baseline block within most advanced architectures.

Convolutional neural networks (CNNs) were first introduced in the late 1980s (LeCun et al., 1989) and are most successfully employed in image analysis. Within the field of SA, CNNs are usually employed to extract information from images, which can then be combined with tabular data information in a multimodal fashion. Often, CNN-based methods use large pre-trained CNNs such as ResNet18 (He et al., 2016) and then fine-tune them on case-specific data. This transfer learning approach enables the adaptation of CNNs, which tend to require large training data, to smaller datasets.

Recurrent neural networks (RNNs), also invented in the 1980s (Rumelhart et al., 1986), distinguish themselves from FFNNs and CNNs by being able to memorize parts of the input through a short-term memory and are thus applicable to sequential data. In the survival context, RNNs are hence useful when TVFs are available. However, large data requirements and computational costs are drawbacks of RNNs.

The autoencoder (AE; Ballard, 1987) is another common NN architecture that learns how to reduce the dimensionality of input data and subsequently reconstructs the data from the learned latent representation; both stacked AEs (SAEs) and variational AEs (VAEs) are extensions. General Adversarial Networks (GANs; Goodfellow et al., 2014) consist of a generator that produces synthetic data of gradually improving quality as well as a discriminator that learns how to distinguish between true data input and generator-produced data points. Transformers (Vaswani et al., 2017) use an attention mechanism to learn a representation of context in sequential (e.g., language) data and can subsequently produce output (sequences) from it. Normalizing flows (NFs; Rezende et al., 2015) constitute a family of generative models which employ differentiable and invertible mappings to obtain complex distributions from a simple initial probability distribution for which sampling and density evaluation is easy. Neural Ordinary Differential Equations (nODEs; Chen et al., 2018) use NNs to parametrize the derivative of the hidden state, thus moving beyond the standard specification of a discrete sequence of hidden layers. Fuzzy neural networks (Lee et al., 1975) use fuzzy numbers as inputs and weights within the NN. Diffusion models (Sohl-Dickstein et al., 2015) employ a Markov chain to gradually add random noise to the input data and subsequently learn to undo this diffusion, this way learning to generate new data from noise.

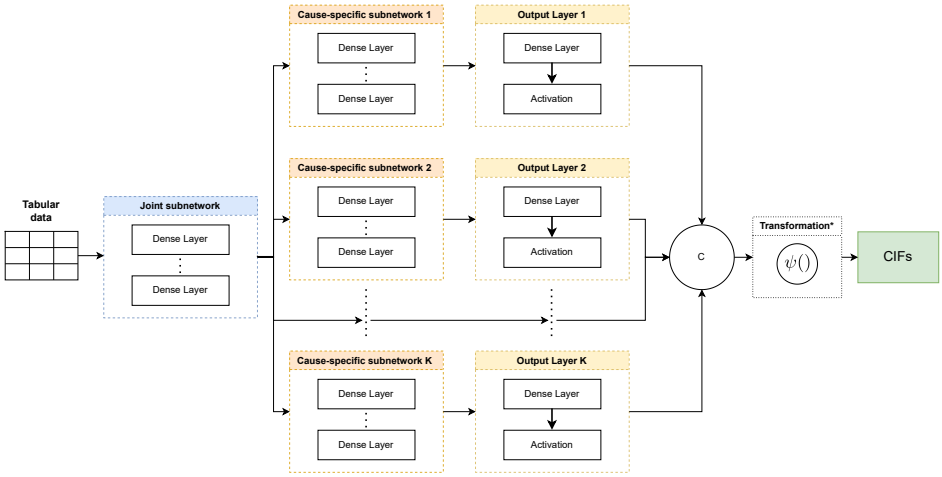

Figure 2: Schematic NN architecture for competing risks in survival analysis using shared and cause-specific subnetworks. As a final step, the network provides a suitable activation function to match the loss. Depending on the kind of output that has been obtained directly from the model, we may need to transform the model output (e.g., from hazards) to the cumulative incidence functions (CIFs) with the function $\psi()$. In the case of a direct survival function estimation $\psi()$ is the identity function.

Many adoptions of NNs for SA emphasize the replacement of the predictor in (7) through an NN. While this is associated with a more expressive and complex predictor, this approach does not tackle existing

problems with other feature or outcome modalities. However, more recently scholars have shown that NNs are particularly suited to facilitate specific modalities directly in their architecture. Some methods employ shared and cause-specific subnetworks for cause- or transition-specific hazards in competing risks and multi-state modeling analysis, which has become a common technique in DL-based survival modeling. Other methods represent the survival data as a longitudinal three-dimensional tensor format (see Kopper et al., 2022), which results in the ability to capture time-dependence (second dimension) and competing risks (third dimension). Additionally, many methods have shown how to integrate multimodal data, by using a separate subnetwork for each modality. For example, for image data one may use a CNN-based subnetwork while tabular data is modeled with an FFNN. The different modalities can be fused together in different ways in the network head. If interaction between different modalities is desired, vector representations of the data are concatenated and fed through another joint FFNN. Otherwise, separate scalars are learned and added onto each other. We illustrate two common architectures that tackle competing risks and multiple data modalities - and can also be combined - in Figures 2 and 3.

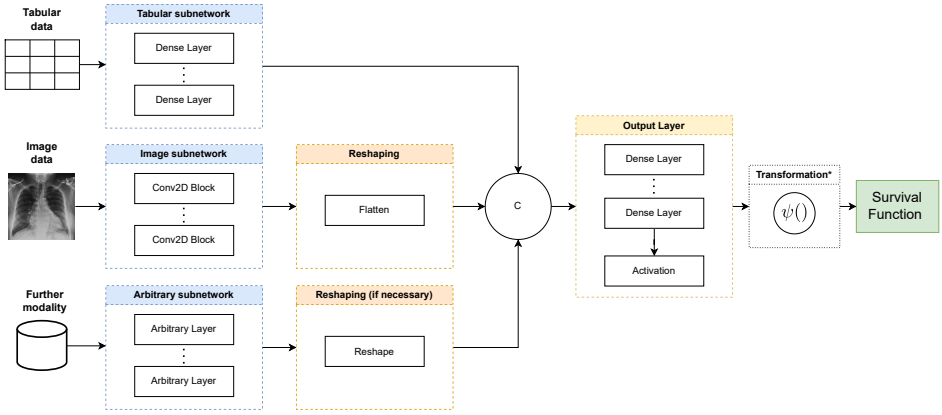

Figure 3: Schematic NN architecture for multimodal data input in survival analysis using separate subnetworks for all modalities. Their outputs are reshaped and concatenated so that their dimensions fit. After passing the flow through the network it is ultimately activated and possibly transformed via $\psi()$ to result in survival probabilities. The X-ray scan is obtained from Irvin et al. (2019).

4.2 Estimation and Network Architecture

We now review all 58 DL-based survival methods based on theoretical and technical aspects. In Section 4.2.1, we aim to categorize the methods in terms of concepts related to model estimation - model class, loss functions, and parametrization - and how these concepts correlate. In Section 4.2.2, we address the NN architecture choices of all methods reviewed.

4.2.1 Estimation

We classify DL-based survival methods in terms of three concepts related to model estimation. First, the *model class* describes which type of statistical survival technique forms the basis of the DL method - usually one of the approaches introduced in Section 3. Second, the *loss function* is often a direct consequence of the model class. However, as is common in DL, some methods employ multiple losses for improved performance or multi-task learning. For instance, some DL-based survival methods compute a ranking loss, in addition to a standard survival loss, for improvement of the C-index performance measure. The final loss is usually computed as the (weighted) average of all losses applied. Third, the *parametrization* determines which model component is being parametrized by an NN. The standard model parametrization is usually implied by the model class.

The left panel of Figure 4 at the end of this section depicts the absolute frequencies of model classes among all 58 methods reviewed here.

We now give a detailed description of the above estimation-related concepts as well as their interrelation for all methods reviewed. Ideally, all methods should be explicit about these concepts; unfortunately, this is not always the case.

Cox-based Methods

Out of the 58 methods included in this review, 25 methods are Cox-based; that is, these methods are essentially DL-based modifications and extensions of the Cox regression model. This is underlined by the fact that all of them parametrize the hazard rate - more precisely, the log-risk function $g(\mathbf{x})$ in (6) - by an NN and minimize the (sometimes slightly modified) Cox loss, i.e., the (negative logarithm of the) partial likelihood of the Cox model.

DeepSurv by Katzman et al. (2018) extends Faraggi et al. (1995) by using a deep FFNN as well as different non-linear hidden layer activation functions. In fact, the model by Faraggi et al. (1995) is a simple special case of *DeepSurv*, with a single hidden layer with logistic activation and identity output activation. Note that the PH assumption induced by the Cox PH regression model is maintained in *DeepSurv*, as $g(\mathbf{x})$ remains time-constant despite being parametrized by a (deep) NN. *DeepSurv* uses stochastic gradient descent (SGD) for optimization. To do so, *DeepSurv* uses a restricted risk set including only individuals in the current batch since the Cox loss originally sums over the *entire* risk set, which would impede batching. *Cox-Time* (Kvamme et al., 2019) is a more flexible extension of *DeepSurv* where a time-dependent predictor allows estimation of TVEs, i.e. $h(t|\mathbf{x}) = h_0(t)\exp(g(\mathbf{x}, t))$. However, this increased flexibility would render the batching strategy as applied by *DeepSurv* (and most other PH-restricted Cox-based methods) computationally expensive. Therefore, the *Cox-Time* loss function is modified to approximate the risk set by a sufficiently large subset of all individuals at risk, which enables mini-batching and thus scalability to large datasets. *NN-DeepSurv* (Tong et al., 2022) is another extension of *DeepSurv*, employing a nuclear norm for imputation of missing features.

More than half of all Cox-based methods (13) focus on the applicability to high-dimensional data, usually omics data. *MCAP* (Chai et al., 2022) and *VAECox* (Kim et al., 2020) both use multiple losses, the latter one within a transfer learning approach. *Cox-nnet* (Ching et al., 2018), *Cox-PASNet* (Hao et al., 2018), *GDP* (Xie et al., 2019), *DNNSurv_Sun2020* (Sun et al., 2020), *Qiu2020* (Qiu et al., 2020), *DeepOmix* (Zhao et al., 2021), and *CNT* (Fan et al., 2022) use simple FFNNs and only a single Cox loss, thus being very similar to *DeepSurv* and *Cox-Time*. *SALMON* (Huang et al., 2019) and *CNN-Cox* (Yin et al., 2022) distinguish themselves through their architecture (see Section 4.2.2), *Haa2019* (Haarburger et al., 2019) and *ConcatAE/CrossAE* (Tong et al., 2020) through additionally being multimodal (see below).

Seven Cox-based methods focus on unstructured or multimodal input (see also Section 4.3.2). *WideAndDeep* (Pölsterl et al., 2020) combines a linear predictor of tabular features (wide part) with a 1D embedding $d(\mathbf{z})$ learned from a point cloud, which is a latent representation learned from 3D shapes (deep part); subsequently both parts are fused by linearly aggregating the learned weights as in (7). The model uses the *DeepSurv* loss and thus preserves the PH assumption. *Haa2019* employs a pre-trained CNN of type ResNet18 for subsequent fine-tuning on CT scans, using a Cox loss. Both *DeepConvSurv* (Zhu et al., 2016) and *CapSurv* (Tang et al., 2019) can learn from image data - though without incorporating structured (tabular) data - by using CNN and CapsNet architectures (see Section 4.2.2), respectively. *DeepConvSurv* uses a single Cox loss, while *CapSurv* additionally employs the CapsNet margin and reconstruction losses. *ConcatAE/CrossAE* use classification and reconstruction losses in addition to the Cox loss to process both high-dimensional data and multimodal data. *Xie2021* (Xie et al., 2021) can learn from unstructured data for cure rate classification. *DAFT* (Wolf et al., 2022) employs CNNs and a single Cox loss to learn from both structured and unstructured data.

SurvNet (Wang et al., 2021) and *DCM* (Nagpal et al., 2021a) do not accommodate any of the outcome or feature modalities defined above (see also Section 4.3), yet they use multiple losses. In addition to a Cox regression module, *SurvNet* employs an input construction module and a survival classification module (with corresponding losses) for handling missing values and high- versus low-risk profile classification, respectively. *DCM* employs an approximate Monte Carlo Expectation Maximization (EM) algorithm for the estimation of a mixture of Cox models, the total loss also including an Evidence Lower Bound (ELBO) component. *ELMCoxBAR* (Wang et al., 2019a) and *San2020* (Sansaengtham et al., 2020) - are standard Cox-based methods in terms of estimation, their architectures being extensions of FFNNs (see Section 4.2.2).

Discrete-time Methods

Another 18 methods can be categorized as discrete-time approaches. They consider time to be discrete and usually employ classification techniques, with the outcome being binary event indicators for each discrete time point or interval. The standard loss function of discrete-time DL-based survival methods is the negative log-likelihood (NLL), while typically the discrete hazard (4) is parametrized by an NN - just like in the early *PLANN* model. However, as compared to the Cox-based methods which are rather homogeneous methodologically, discrete-time methods are much more heterogeneous in terms of loss functions and architecture.

DeepHit (Lee et al., 2018) is a very well known discrete-time DL-based survival method. It aims to learn first-hitting times directly by not making any assumptions about the underlying stochastic process and parametrizing the (discrete) PMF directly. *DeepHit* combines two loss functions: first, the log-likelihood derived from the joint distribution of first hitting time and the corresponding event, adjusted for right-censoring and taking into account competing risks; and second, a combination of ranking losses. *Dynamic-DeepHit* (Lee et al., 2019), an RNN-based extension of *DeepHit* which can handle longitudinal input data and thus TVFs, additionally employs a so-called prediction loss for the auxiliary task of step-ahead prediction of TVFs.

Nnet-survival (Gensheimer et al., 2019) parametrizes the discrete hazard (4) by an NN, using an NLL loss as well as mini-batch SGD for rapid convergence and scalability to large datasets. (Mini-batch SGD is easily applicable to discrete-time methods because the loss only depends on individuals in the current mini-batch - which is not the case for the Cox loss. This is why *Cox-Time* explicitly modifies its loss function to facilitate batching.) *Nnet-survival* can parametrize the discrete hazards, that is, the conditional probability for an event at each (discrete) time point, with different types of NNs, though not specifying whether information from more than one data modality can be used. The specific architecture - in particular, the number of neurons per hidden layer and the connectedness of layers - determines whether TVEs can be modeled or whether the PH restriction is upheld. Another four methods - *CNN-Survival* (Zhang et al., 2020), *MultiSurv* (Vale-Silva et al., 2021), *SurvCNN* (Kalakoti et al., 2021), and *Tho2022* (Thorsen-Meyer et al., 2022) - use the same loss and parametrization as *Nnet-survival*. *CNN-Survival* uses a CNN along with transfer learning to learn from CT data (without incorporating tabular data). The multimodal *MultiSurv* first extracts feature representations for each data modality separately, then fuses them, and finally outputs predictions of conditional survival probabilities. *SurvCNN* creates an image representation of multiple omics data types using CNNs and can combine this with clinical data for prediction. *Tho2022* can embed data from multiple modalities, such as electronic health records, and feeds these embedded representations into an RNN which in turn produces survival predictions.

N-MTLR (Fotso, 2018) builds upon Multi-Task Logistic Regression (MTLR) and parametrizes the corresponding logistic regression parameters. *RNN-SURV* (Giunchiglia et al., 2018), based on RNN architecture, employs both an NLL loss and a C-index-based loss. Another RNN-based approach, *DRSA* (Ren et al., 2019), uses multiple log-likelihood-based losses to predict the likelihood of uncensored events as well as survival rates for censored cases. *DCS* (Fuhlert et al., 2022) extends and modifies *DRSA*, using a combination of kernel loss and rank probability score (RPS) loss. The competing-risk and recurrent-event method *CRESA* (Gupta et al., 2019), again RNN-based, uses a loss based on recurrent cumulative incidence functions (RCIFs), parametrizing the discrete hazard. *DNNSurv_Zha2019* (Zhao et al., 2019) first computes individual-level pseudo (conditional) probabilities, defined as the difference between the estimated survival function with and without individual i and computed on a regular grid of time points, thus reducing the survival task to a regression task, and consequently uses a standard regression loss. *su-DeepBTS* (Lee et al., 2020) discretizes the time axis but then uses a Cox loss for each time interval, summing up the losses across intervals. *DeepComp* (Li et al., 2020) combines distinct losses for censored and uncensored observations with an additional penalty. *SSMTL* (Chi et al., 2021) transforms the survival task into a multi-task setting with binary outcome for all time points (or multi-class in case of competing risks), then predicting survival probabilities for each of the time points. *SSMTL* also employs a custom loss made up of a classification loss for uncensored data, a so-called semi-supervised loss for censored data, regularization losses (L1 and L2) as well as a ranking loss in order to ensure monotonicity of predicted survival probabilities.

Hu2021 (Hu et al., 2021), a transformer-based method, uses an entropy-based loss as well as a discordant-pair penalization loss, parametrizing the discrete hazard. *SurvTRACE* (Wang et al., 2022), another transformer-

based method, also parametrizes the discrete hazard, but additionally performs two auxiliary tasks on the survival data: classification and regression; accordingly, the final model loss is a combination of a *PC-Hazard* survival loss (see below), an entropy-based classification loss, as well as a Mean Squared Error (MSE)-based regression loss. Finally, the third transformer-based discrete-time survival method, *TransformerJM* (Lin et al., 2022), focuses on modeling survival data and longitudinal data jointly, parametrizing the PMF (similar to *DeepHit*) and training on a combination of NLL- and MSE-based losses.

PEM-based Methods

Three methods rely on the PEM framework to develop a deep survival approach. *PC-Hazard* (Kvamme et al., 2021) addresses the right-censored single-risk survival task by parametrizing the hazard rate through an FFNN and using the standard likelihood-based PEM loss. Support for other outcome or feature modalities, as introduced in Sections 2.2.1 and 2.2.2, is not discussed. Similarly, *DeepPAMM* (Kopper et al., 2022) uses a penalized Poisson NLL as a loss function and also parametrizes the hazard rate by an NN. This method combines a Piecewise Exponential Additive Mixed Model (PAMM; Bender et al., 2018) with Semi-structured Deep Distributional Regression (Rügamer et al., 2023), which embeds the structured predictor in an NN and further learns from other (unstructured) data types (see (7)).

Finally, *IDNetwork* (Cottin et al., 2022) implements an illness-death model, which uses a PEM-based approach to estimate probabilities for transitions between different states and utilizes FFNNs with shared and transition-specific subnetworks. *IDNetwork* then uses a penalized NLL loss based on the transition probabilities.

Parametric Methods

The two methods *DeepWeiSurv* (Bennis et al., 2020) and *DPWTE* (Bennis et al., 2021) - the latter one building on the former - are Weibull-based deep survival methods. Neither of them addresses any of the outcome or feature modalities presented in Section 2.2. *DeepWeiSurv* parametrizes a mixture of Weibull models, as well as both Weibull distribution parameters (see (8) with $\boldsymbol{\theta}(\mathbf{x}) = (\lambda(\mathbf{x}),\ \alpha(\mathbf{x}))^\top$), by an FFNN and uses an NLL-based loss function. *DPWTE* employs classification and regression subnetworks to learn an optimal mixture of Weibull distributions, using the same loss function as *DeepWeiSurv* with additional sparsity regularization with respect to the number of mixtures.

DSM (Nagpal et al., 2021b) is a hierarchical generative model based on a finite mixture of parametric primitive distributions similar to Ranganath et al. (2016), using a (mixture) likelihood-based loss as well as an additive loss based on ELBO for uncensored and censored observations; the choice of the parametric survival distribution - either Weibull or Log-Normal - is a hyperparameter and can thus be tuned. Its RNN-based extension *RDSM* (Nagpal et al., 2021c), is furthermore capable of handling TVFs.

Ranking-based Methods

RankDeepSurv (Jing et al., 2019) combines regression and ranking losses to augment the number of training samples, without advanced NN architecture or handling of non-standard survival data modalities. *SSCNN* (Agarwal et al., 2021) is a multimodal method that reduces histopathology images to whole slide feature maps and uses them, in addition to clinical features, as input of a Siamese Survival CNN (SSCNN); model training with a custom loss - a combination of a ranking loss with a loss to improve model convergence and pairwise differentiation between survival predictions - is built directly on the outputs of the Siamese NN.

ODE-based Methods

DeepCompete (Aastha et al., 2021) consists of an FFNN shared across all risks as well as an FFNN and a neural ordinary differential equation (ODE) block for each specific risk, using an NLL-based loss. *survNODE* (Groha et al., 2021) is based on a Markov process and aims to directly solve the Kolmogorov forward equations by using neural ODEs to achieve flexible multi-state survival modeling, with the transition rates parametrized by a nODE architecture (see Sections 4.1 and 4.2.2).

Other Methods

As for the remaining five methods, *DASA* (Nezhad et al., 2019) is a framework introducing a novel sampling strategy based on DL and active learning. The GAN-based *DATE* (Chapfuwa et al., 2018) seeks to learn the event time distribution non-parametrically by using adversarial learning and a loss function made up of an uncensored-data component, a censored-data component, as well as a distortion loss component. *Hua2018* (Huang et al., 2018) uses a CNN architecture and correlational layers for multimodal learning to produce person-specific risks, which are then directly fed into a smooth C-index loss function for model training. *Aus2021* (Ausset et al., 2021) employs normalizing flows in order to estimate the density of time-to-event data and predict individual survival curves via a transformation model, using an NLL-based loss augmented by an intermediary loss for regularization. Finally, *rcICQRNN* (Qin et al., 2022) is a deep survival method based on a quantile regression NN, parametrizing the quantile regression coefficients by means of an FFNN and using an inverse-probability-of-censoring weighted log-linear quantile regression loss.

Limitations

Out of all 26 Cox-based methods reviewed here, all but one (*Cox-Time*) assume proportional hazards and thus do not (directly) support TVFs or TVEs; the same applies to all PEM-based methods except *DeepPAMM*. Reliance on the PH assumption - unless explicit stratification is employed or TVEs are allowed for - is thus an apparent drawback of Cox- and PEM-based methods. PEM-based approaches furthermore require a pre-transformation of the data, which can slow down estimation depending on how the estimation routine handles it.

By comparison, discrete-time methods are much more flexible regarding usage of architectures and losses, with many not confined by the PH assumption. However, a disadvantage of assuming time to be non-continuous is that interpolation between the limited, fixed prediction time points is necessary in order to achieve some sort of continuous prediction. Alternatively, one can choose a quasi-continuous grid of discrete event times which, however, may imply much higher computational complexity.

As for parametric methods, their main disadvantage is of course their distributional assumption - i.e., the fact that they make an assumption about the distribution of event times.

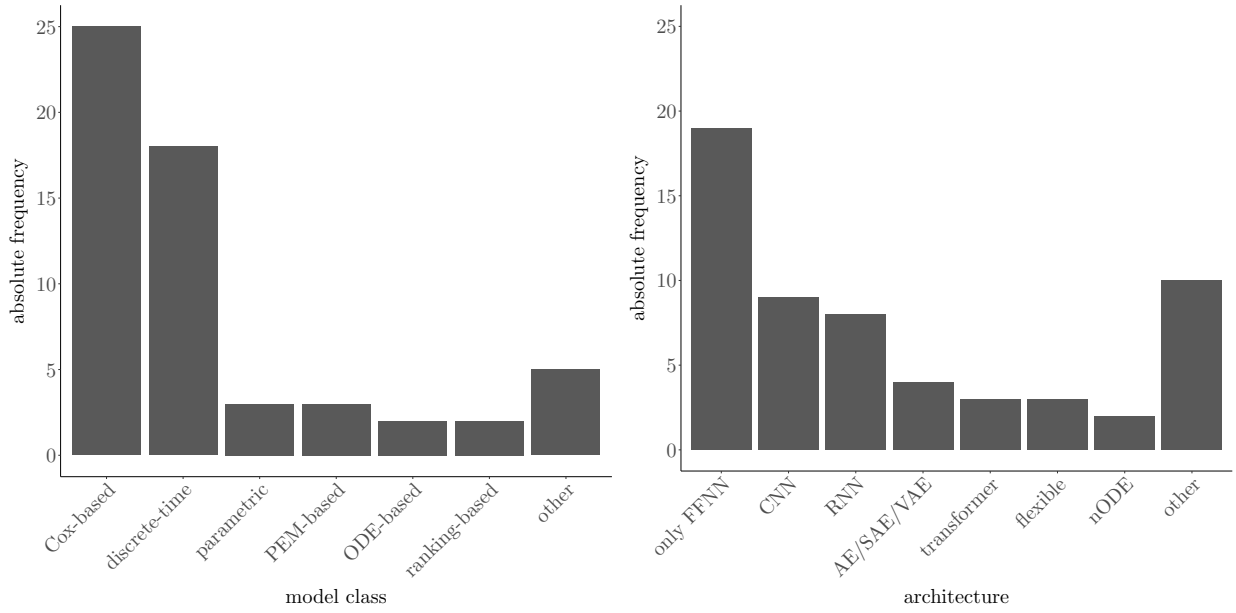

Figure 4: Bar plot showing absolute frequencies of model classes (left) and neural network architectures (right) among all 58 methods reviewed.

4.2.2 Network Architecture

Most DL-based survival methods in this review use FFNNs, often in combination with some other, more advanced architecture. Still, 20 methods - as well as all early DL-based methods such as the one by Faraggi et al. (1995) - exclusively rely on FFNNs.

Out of a total of nine CNN-based methods in this review, seven are multimodal methods that can work with image data: *DeepConvSurv*, *Hua2018*, *Haa2019*, *CNN-Survival*, *SSCNN*, *Xie2021*, and *DAFT*. The CNN-based method *SurvCNN* is not multimodal per se, but transforms high-dimensional omics data into an image representation in order to feed them into a CNN. *CNN-Cox* combines cascaded Wx (Shin et al., 2019), an NN-based algorithm selecting features based on how well they distinguish between high- and low-risk groups, with a 1D CNN architecture applied to gene expression data.

Eight methods reviewed here use RNN architectures. Five of them - *RNN-Surv*, *CRESA*, *DRSA*, *DCS*, and *Tho2022* - use a Long Short-Term Memory (LSTM), while the remaining one, *DeepComp*, does not state which specific RNN architecture it employs. Out of these methods, *DRSA*, *RNN-Surv*, and *DCS* do not go beyond the setting of single-risk, right-censored tabular data. *Tho2022* employs the RNN architecture for multimodal learning from text, medical history, and high-frequency data, while *DeepComp* uses it for competing risk modeling. *CRESA* models both recurrent events and competing risks by means of its RNN architecture. The final two RNN-based methods, *Dynamic-DeepHit* and *RDSM*, are actually extensions of the simpler FFNN-based methods *DeepHit* and *DSM*, respectively, enabling incorporation of TVFs.

Four methods - *DASA*, *DCM*, *ConcatAE/CrossAE*, and *VAECox* - use some form of AEs. Only a single method, *DATE*, uses a GAN architecture. Three recent methods, *Hu2021*, *SurvTRACE*, and *TransformerJM*, use a transformer architecture, while another two novel methods, *DeepCompete* and *survNode*, use a nODE architecture. Finally, *Nnet-survival*, *MultiSurv*, and *DeepPAMM* do not require a specific architecture, which can instead be flexibly chosen based on application requirements.

ElmCoxBAR uses an Extreme Learning Machine (ELM) architecture, which is similar to an FFNN but does not require backpropagation for optimization. *SALMON*, *San2020*, *DPWTE*, and *SurvNet* all use FFNNs, but in a modified manner. *SALMON* adds so-called eigengene modules, using eigengene matrices of gene co-expression modules (Zhang et al., 2014) instead of raw gene expression data as NN input. *San2020* uses a Stacked Generalization Ensemble Neural Network (Wolpert, 1992), which takes a combination of *DeepSurv* sub-models and concatenates them for improved hazard prediction. *DPWTE* adds a Sparse Weibull Mixture (SWM) layer to learn the optimal number of Weibull distributions for the mixture model, through an element-wise multiplication of its weights by the previous layer's output. *SurvNet* adds a context-gating mechanism, which is similar to the attention mechanism used in transformers, by adjusting log hazard ratios by survival probabilities from the survival classification module. *WideAndDeep* employs a PointNet (Qi et al., 2017) architecture to learn a latent representation of 3D shapes of the human brain while additionally learning from regular tabular data, subsequently fusing both parts. *CapSurv* modifies the CapsNet architecture (Sabour et al., 2017), developed for image classification, by adding a Cox loss and thus making it amenable to SA tasks.

The right panel of Figure 4 depicts absolute frequencies of NN architectures among all 58 methods included in this review.

4.3 Supported Survival Tasks

In Section 2.2, we introduced a wide range of different data modalities survival tasks can present. In this section, we now consider which methods can handle such modalities. We start by considering *outcome* modalities - i.e., event times and event indicators - and subsequently *feature* modalities. Finally, we summarize which methods offer (which kind of) interpretability of results.

4.3.1 Supported Outcome Modalities

Regarding censoring and truncation of event times, left-censoring and right-truncation are not addressed explicitly by any of the methods reviewed. *survNode* briefly addresses interval-censoring and left-truncation by stating how they would affect likelihood computations. *DSM* mentions that the modeling framework is amenable to these two output modalities. *DeepPAMM* can accommodate left-truncated data by adjusting loss contributions according to the left-truncated times at risk.

Nine methods are designed to deal with competing risks; interestingly, none of these methods is Cox-based, and four of them are discrete-time. *DeepHit*, *CRESA*, and *DeepComp* all assume time to be discrete and employ cause-specific subnetworks, with *DeepHit* using FFNNs to generate a final distribution over all competing causes for each individual; both *CRESA* and *DeepComp* use RNN architectures, yet while *CRESA* also generates a final distribution over all competing causes, *DeepComp* outputs cause-specific discrete hazards for each time interval. *SSMTL*, also discrete-time, uses an FFNN architecture, views competing risk SA as a multiclass problem, and creates a custom loss with separate components for non-censored and censored individuals. *DeepCompete* is a continuous-time method that employs nODE blocks within each of its cause-specific subnetworks in order to output a cumulative hazard function. *DSM* first learns a common representation of all competing risks by passing through a single FFNN. Based on this representation, and treating all other events as censoring, the event distribution for a single risk is then learned using cause-specific Maximum Likelihood Estimation (MLE); the ELBO loss is also adjusted to treat competing events as censoring. Both *survNode* and *IDNetwork* are based on Markov processes - illness-death process and Markov jump process, respectively - and thus naturally handle competing risks and even the more general case of multi-state outcomes. Being PEM-based, *DeepPAMM* parametrizes the hazard rate, which is a transition rate by definition; *DeepPAMM* can further specify multiple transitions and therefore model competing risks as well as multi-state outcomes, facilitated by three-dimensional output tensors. Finally, two methods are capable of handling recurrent events: *CRESA* employs an RNN architecture with time steps representing recurrent events, while *DeepPAMM* uses random effects inspired by statistical mixed models. Figure 5 summarizes which outcome modalities the methods reviewed here can handle.

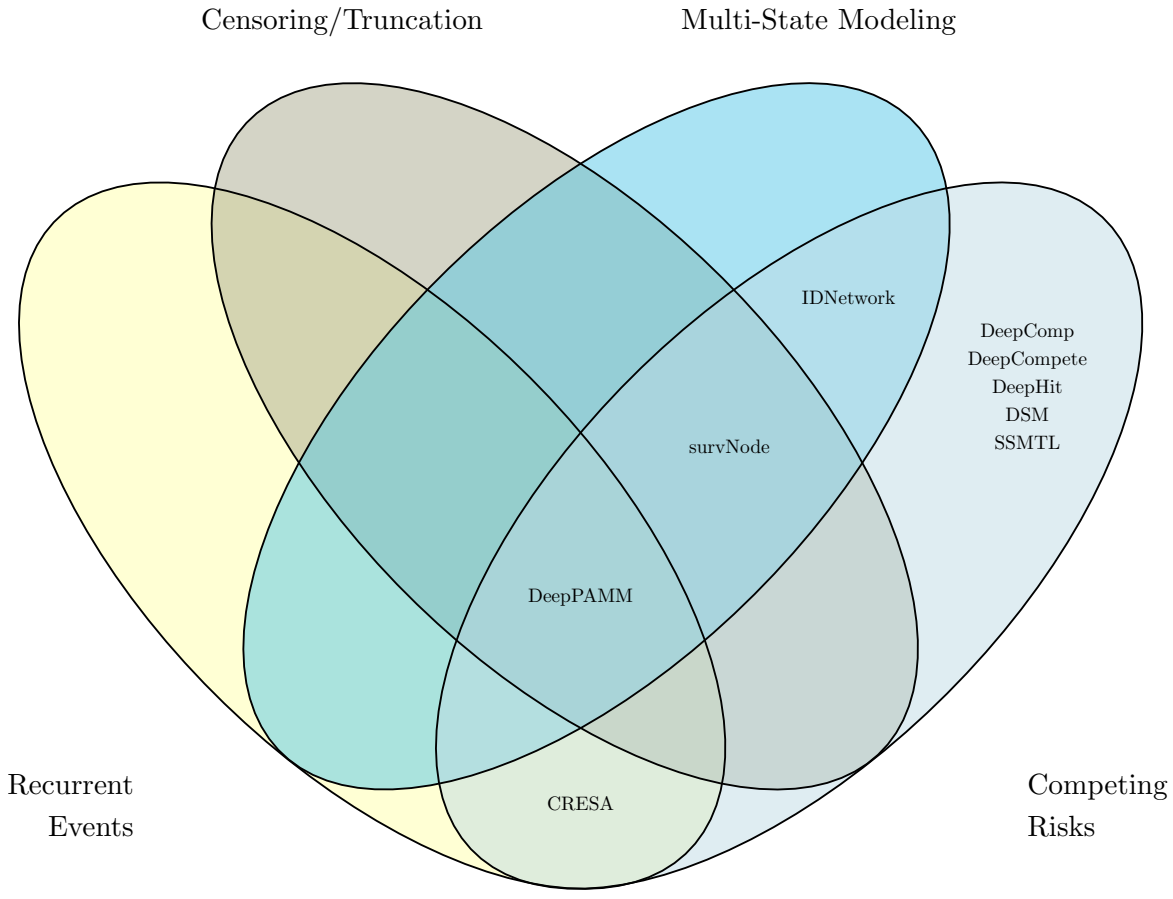

Figure 5: Venn diagram illustrating which methods can handle the distinct survival outcome modalities.

4.3.2 Supported Feature Modalities

One major feature modality is time dependence, a deviation from the PH assumption imposed by traditional survival models such as Cox regression or Weibull AFT. Only six methods can handle TVFs: *DeepHit*'s and *DSM*'s RNN-based extensions *Dynamic-DeepHit* and *RDSM*, as well as *CRESA*, *survNode*, *DeepPAMM*, and *TransformerJM*. The technical incorporation of TVFs is, for example, achieved by converting tabular time-varying feature input into long format (*DeepPAMM*) or by employing RNNs prior to each new feature measurement (*survNode*).

TVEs constitute another deviation from the PH assumption: Seven methods are capable of modeling effects that might not be constant over time, with four of them being time-discrete approaches. *Nnet-survival* and *MultiSurv* incorporate TVEs modeling by using a fully connected NN to connect the final hidden layer's neurons with the output nodes, while *RNN-Surv* captures TVEs through its RNN architecture. *Cox-Time* accommodates TVEs by making the Cox-style relative risk - which it parametrizes by an NN - time-dependent and *DeepPAMM* can address TVEs through the interaction of the follow-up time (represented as a feature) with other features. *DSM* and *SSMTL* do not provide further detail about how TVEs are being estimated.

Another feature modality is the integrability of high-dimensional (usually omics) data, which implies learning from a high-dimensional predictor space. While all DL-based methods are generally capable of handling high-dimensional feature inputs, here we focus on the 17 DL-based survival methods that are explicitly designed to work with high-dimensional data, usually by applying specialized regularization techniques. 13 of these methods - *Cox-nnet*, *Cox-PASNet*, *Haa2019*, *GDP*, *SALMON*, *ConcatAE/CrossAE*, *DNNSurv_Sun2020*, *Qiu2020*, *VAECox*, *DeepOmix*, *CNN-Cox*, *CNT*, and *MCAP* - are (partially) Cox-based. As for the remaining four methods, *CNN-Survival*, *MultiSurv*, and *SurvCNN* are discrete-time methods, while *rcICQRNN* is quantile regression-based.

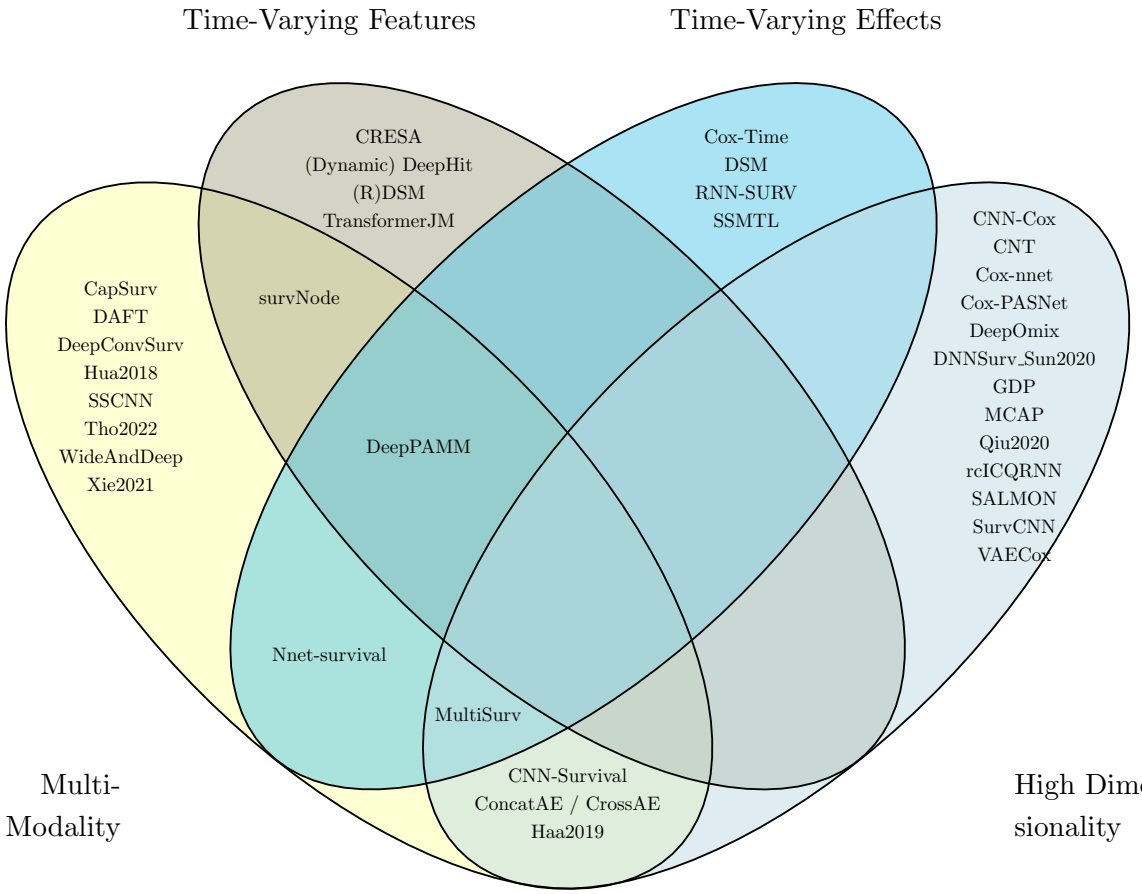

Figure 6: Venn diagram illustrating which methods can handle the distinct survival feature modalities.

Finally, a total of 15 methods can (hypothetically) extract information from unstructured or multimodal features. Eight of them are (partially) CNN-based, underlining the focus on processing mostly medical image data. *DeepConvSurv*, *CapSurv*, and *CNN-Survival* (the last one employing transfer learning) exclusively work with imaging data without incorporating any tabular information, which is why these methods are not truly *multi*modal. Similarly, *Nnet-survival*, being flexible in terms of NN architecture, can learn from image data by choosing a CNN, yet again at the cost of discarding any potentially available tabular data. *Hua2018* incorporates both image and molecular data yet without making any mention of tabular data. *Haa2019* fine-tunes a pre-trained ResNet18, optionally concatenating it with radiomics features, and additionally leverages clinical data. *ConcatAE/CrossAE* integrates information from multiple modalities, either through modality-specific autoencoders or cross-modality translation; the integration of tabular data is, however, not explicitly mentioned. *survNode* can conceptually account for multimodal features by encoding initial values with, e.g., CNN or NLP layers. *SSCNN* creates feature maps from whole slide images and employs a Siamese CNN to learn from both these feature maps as well as clinical features. *Xie2021*, being a cure rate model, only allows for (single-modality) unstructured data for determining the cure rate probability through a CNN. *DAFT* uses a ResNet CNN architecture as backbone, feeding tabular data into it through a novel Dynamic Affine Feature Map Transform (DAFT) module, which in turn enables a bidirectional information flow between image and tabular data. Finally, *Tho2022* employs an RNN architecture to create an embedding for electronic patient record data (such as medical history and free text) and further fuses tabular clinical features into the model before generating survival predictions. *WideAndDeep*, using a very specific Alzheimer's Disease (AD) dataset, learns a latent representation of 3D shapes of the human brain while additionally learning from regular tabular data, subsequently fusing both parts. *MultiSurv*, a multimodal extension of *Nnet-survival*, and *DeepPAMM* both provide flexibility in terms of architecture choice so that, for example, image data could be incorporated by employing CNNs for the NN part; they also fuse information from the different data modalities.

Figure 6 illustrates which of our methods can incorporate the different types of feature modalities.

4.4 Interpretability

By construction, DL methods are more complex and thus usually do not provide the same degree of interpretability as offered by statistical survival models. At the same time, in fields such as the life sciences results and model outputs are almost always required to be interpretable in order to provide a solid basis for oftentimes highly sensitive decision-making (Vellido, 2020). Here, we briefly summarize which of the methods provide (inherently) interpretable results.

Cox-nnet, *Cox-PASNet*, and *DeepOmix* provide some interpretability by assigning biological meaning to the nodes of their NNs. By fusing the output of an NN for image data with the output of a Cox PH model for tabular data, *WideAndDeep* retains the interpretability of a standard Cox regression for structured features. *Xie2021* also provides standard Cox model interpretability, owing to the fact that survival prediction is performed through non-deep Cox regression. *DeepPAMM* provides classical statistical interpretability of the structured effects, with identifiability ensured through orthogonalization to unstructured (image or text) input. *survNode* introduces a latent variable extension providing aspects of feature interpretability. The transformer-based *SurvTrace* method makes use of attention maps, comparing attention scores of different features across selected individuals, in order to provide some interpretability of feature effects.

It is worth noting that post-hoc methods from the field of Interpretable Machine Learning (IML), such as Local Interpretable Model-agnostic Explanations (LIME), Shapley Additive exPlanations (SHAP), and Permutation Feature Importance, are also applicable to DL-based survival methods. Among the papers reviewed here, *SALMON* explores feature importance of individual inputs, *DNNSurv_Sun2020* employs LIME, *Tho2022* uses SHAP, and *SSMTL* computes post-hoc feature importance and plots feature effects on cumulative incidence curves. *Qiu2020* uses a risk propagation technique called *SurvivalNet* (Yousefi et al., 2017), which is an explanation method specific to SA. Further survival-specific interpretability methods are *SurvNAM* (Utkin et al., 2022) and *SurvSHAP(t)* (Krzyziński et al., 2022).

4.5 Reproducibility and Means of Evaluation

Code and data accessibility are important in their own right, indicating the maturity of the respective method and its general applicability, as well as enhancing reproducibility. Therefore, here we briefly address to which degree code and data are accessible for the different methods under investigation.

All papers reviewed in this work use accessible (i.e., public or registered access) benchmark data by construction (see exclusion criteria), with many of them benchmarking their method on multiple datasets (see *Main Table*). The nature of the datasets underlines the special focus of survival methods on life science applications. For 21 out of all 58 methods, model performance is additionally evaluated on synthetic data (see *Main Table*).

Only 33 methods provide publicly accessible code of their algorithms and benchmark experiments. The fact that accompanying code for 25 recent DL-based survival methods is not publicly available is disappointing not only in terms of reproducibility but also because these methods are not at the disposal of the research community to help answer future questions. Furthermore, the corresponding codes of 25 methods are currently still one-shot implementations and have not yet been processed into easy-to-use packages; ideally, they should be integrated into existing software ecosystems such as `auton-survival` (Nagpal et al., 2022), `mlr3proba` (Sonabend et al., 2021), `pycox` (Kvamme et al., 2019), `scikit-survival` (Pölsterl, 2020), or similar. For each method, the *Main Table* provides links to the respective code repositories, if available.

5 Conclusion

Classical SA, also called time-to-event analysis, is concerned with modeling the time until an event of interest occurs while accounting for the fact that for some subjects the exact event times are not observed due to censoring. This standard scenario has numerous extensions, as detailed in section 2.2, and incorporating survival data with more complex modalities thus requires specialized techniques.

In this paper, we provide a structured, comprehensive review of existing DL-based survival methods, from a theoretical as well as practical perspective, targeting both survival experts and DL specialists. In doing so, we aim to facilitate access to the great wealth of already available methods for applied work as well as to help identify the most promising areas for future research. The main results are summarized in an open-source, interactive table (https://survival-org.github.io/DL4Survival) and all data, figures, and code scripts used for generating the table and figures are in the corresponding repository https://github.com/survival-org/DL4Survival.

We conclude that most methodologically innovative DL-based survival methods are survival-specific applications of novel methods developed in other areas of deep learning, such as computer vision or NLP. Regarding data modalities, little attention has been paid to outcome modalities beyond right-censoring, potentially due to a limited number of application cases. SA will certainly continue to benefit from advances in ML/DL, with big methodological advances being likely to swap over. In particular, generative DL techniques like diffusion are promising candidates for adaptation to survival tasks. However, a bottleneck regarding the application of advanced DL methods to SA is the lack of openly accessible, high-dimensional, potentially multimodal data sets.

Finally, a limitation of this overview is that it stops short of actively benchmarking the reviewed methods against each other. A unifying benchmarking and evaluation framework for (DL-based) SA methods would certainly be valuable; however, such benchmarking is hampered by non-disclosure of code and the abovementioned lack of data, particularly for modern techniques capable of, e.g., multimodal learning.

We believe that this review provides valuable insights into available DL-based survival methods from a variety of angles, for survival analysis scholars, practitioners, and DL experts, in order to improve future applied and methodological work.

References

Aalen, O. (July 1978). "Nonparametric Inference for a Family of Counting Processes". EN. In: *The Annals of Statistics* 6.4, pp. 701–726. ISSN: 0090-5364, 2168-8966. DOI: 10.1214/aos/1176344247. URL: http://projecteuclid.org/euclid.aos/1176344247 (visited on 10/31/2016).

Aastha, P. Huang, and Y. Liu (Jan. 2021). "DeepCompete : A deep learning approach to competing risks in continuous time domain". In: *AMIA Annual Symposium Proceedings* 2020, pp. 177–186. ISSN: 1942-597X. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8075516/ (visited on 11/17/2021).

Agarwal, S., M. Eltigani Osman Abaker, and O. Daescu (2021). "Survival prediction based on histopathology imaging and clinical data: A novel, whole slide cnn approach". In: *Medical Image Computing and Computer Assisted Intervention–MICCAI 2021: 24th International Conference, Strasbourg, France, September 27–October 1, 2021, Proceedings, Part V 24*. Springer, pp. 762–771.

Ausset, G., T. Ciffreo, F. Portier, S. Clémençon, and T. Papin (2021). "Individual Survival Curves with Conditional Normalizing Flows". In: *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 1–10. DOI: 10.1109/DSAA53316.2021.9564222.

Ballard, D. H. (1987). "Modular learning in neural networks." In: *Aaai*. Vol. 647, pp. 279–284.

Bender, A., A. Groll, and F. Scheipl (2018). "A generalized additive model approach to time-to-event analysis". In: *Statistical Modelling* 18.3-4, pp. 299–321.

Bender, A., D. Rügamer, F. Scheipl, and B. Bischl (Feb. 2021). "A General Machine Learning Framework for Survival Analysis". en. In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by F. Hutter, K. Kersting, J. Lijffijt, and I. Valera. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 158–173. ISBN: 978-3-030-67664-3. DOI: 10.1007/978-3-030-67664-3_10.

Bennis, A., S. Mouysset, and M. Serrurier (2020). "Estimation of conditional mixture Weibull distribution with right censored data using neural network for time-to-event analysis". In: *Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part I 24*. Springer, pp. 687–698.

— (2021). "DPWTE: A Deep Learning Approach to Survival Analysis Using a Parsimonious Mixture of Weibull Distributions". en. In: *Artificial Neural Networks and Machine Learning – ICANN 2021*. Ed. by I. Farkaš, P. Masulli, S. Otte, and S. Wermter. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 185–196. ISBN: 978-3-030-86340-1. DOI: 10.1007/978-3-030-86340-1_15.

Biganzoli, E., P. Boracchi, L. Mariani, and E. Marubini (1998). "Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach". In: *Statistics in medicine* 17.10, pp. 1169–1186.

Binder, H. and M. Schumacher (2008). "Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models". In: *BMC bioinformatics* 9.1, pp. 1–10.

Box-Steffensmeier, J. M. and S. De Boef (2006). "Repeated events survival models: the conditional frailty model". In: *Statistics in medicine* 25.20, pp. 3518–3533.

Brown, S. F., A. J. Branford, and W. Moran (1997). "On the use of artificial neural networks for the analysis of survival data". In: *IEEE transactions on neural networks* 8.5, pp. 1071–1077.

Campanella, G., L. Kook, I. Häggström, T. Hothorn, and T. J. Fuchs (2022). "Deep conditional transformation models for survival analysis". In: *arXiv preprint arXiv:2210.11366*.

Chai, H., L. Guo, M. He, Z. Zhang, and Y. Yang (2022). "A Multi-constraint Deep Semi-supervised Learning Method for Ovarian Cancer Prognosis Prediction". In: *Advances in Swarm Intelligence: 13th International Conference, ICSI 2022, Xi'an, China, July 15–19, 2022, Proceedings, Part II*. Springer, pp. 219–229.

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

Chen, R. T., Y. Rubanova, J. Bettencourt, and D. K. Duvenaud (2018). "Neural ordinary differential equations". In: *Advances in neural information processing systems* 31.

Chi, S., Y. Tian, F. Wang, Y. Wang, M. Chen, and J. Li (Aug. 2021). "Deep Semisupervised Multitask Learning Model and Its Interpretability for Survival Analysis". In: *IEEE Journal of Biomedical and Health Informatics* 25.8, pp. 3185–3196. ISSN: 2168-2194, 2168-2208. DOI: 10.1109/JBHI.2021.3064696. URL: https://ieeexplore.ieee.org/document/9373895/ (visited on 11/30/2021).

Ching, T., X. Zhu, and L. X. Garmire (2018). "Cox-nnet: an artificial neural network method for prognosis prediction of high-throughput omics data". In: *PLoS computational biology* 14.4, e1006076.

Cottin, A., N. Pecuchet, M. Zulian, A. Guilloux, and S. Katsahian (2022). "IDNetwork: A deep illness-death network based on multi-state event history process for disease prognostication". In: *Statistics in Medicine* 41.9, pp. 1573–1598.

Cox, D. R. (1972). "Regression models and life-tables". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 34.2, pp. 187–202.

Deepa, P. and C. Gunavathi (2022). "A systematic review on machine learning and deep learning techniques in cancer survival prediction". In: *Progress in Biophysics and Molecular Biology*.

Fan, Y., S. Zhang, and S. Ma (2022). "Survival Analysis with High-Dimensional Omics Data Using a Threshold Gradient Descent Regularization-Based Neural Network Approach". In: *Genes* 13.9, p. 1674.

Faraggi, D. and R. Simon (1995). "A neural network model for survival data". In: *Statistics in medicine* 14.1, pp. 73–82.

Fornili, M., F. Ambrogi, P. Boracchi, and E. Biganzoli (2013). "Piecewise exponential artificial neural networks (PEANN) for modeling hazard function with right censored data". In: *International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics*. Springer, pp. 125–136.

Fotso, S. (Jan. 2018). "Deep Neural Networks for Survival Analysis Based on a Multi-Task Framework". In: *arXiv:1801.05512 [cs, stat]*. arXiv: 1801.05512. URL: http://arxiv.org/abs/1801.05512 (visited on 03/21/2021).

Friedman, M. (1982). "Piecewise exponential models for survival data with covariates". In: *The Annals of Statistics* 10.1, pp. 101–113.

Fuhlert, P., A. Ernst, E. Dietrich, F. Westhaeusser, K. Kloiber, and S. Bonn (2022). "Deep Learning-Based Discrete Calibrated Survival Prediction". In: *2022 IEEE International Conference on Digital Health (ICDH)*. IEEE, pp. 169–174.

Gensheimer, M. F. and B. Narasimhan (2019). "A scalable discrete-time survival model for neural networks". In: *PeerJ* 7, e6257.

Giunchiglia, E., A. Nemchenko, and M. van der Schaar (2018). "RNN-SURV: A Deep Recurrent Model for Survival Analysis". en. In: *Artificial Neural Networks and Machine Learning – ICANN 2018*. Ed. by V. Kůrková, Y. Manolopoulos, B. Hammer, L. Iliadis, and I. Maglogiannis. Lecture Notes in Computer Science. Cham: Springer International Publishing, pp. 23–32. ISBN: 978-3-030-01424-7. DOI: 10.1007/978-3-030-01424-7_3.

Goodfellow, I. J., J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. C. Courville, and Y. Bengio (2014). "Generative Adversarial Nets". In: *NIPS*.

Groha, S., S. M. Schmon, and A. Gusev (Feb. 2021). "A General Framework for Survival Analysis and Multi-State Modelling". In: *arXiv:2006.04893 [cs, stat]*. arXiv: 2006.04893. URL: http://arxiv.org/abs/2006.04893 (visited on 03/20/2021).

Gupta, G., V. Sunder, R. Prasad, and G. Shroff (2019). "Cresa: a deep learning approach to competing risks, recurrent event survival analysis". In: *Advances in Knowledge Discovery and Data Mining: 23rd Pacific-Asia Conference, PAKDD 2019, Macau, China, April 14-17, 2019, Proceedings, Part II 23*. Springer, pp. 108–122.

Haarburger, C., P. Weitz, O. Rippel, and D. Merhof (Apr. 2019). "Image-Based Survival Prediction for Lung Cancer Patients Using CNNS". In: *2019 IEEE 16th International Symposium on Biomedical Imaging (ISBI 2019)*. ISSN: 1945-8452, pp. 1197–1201. DOI: 10.1109/ISBI.2019.8759499.

Hao, J., Y. Kim, T. Mallavarapu, J. H. Oh, and M. Kang (Dec. 2018). "Cox-PASNet: Pathway-based Sparse Deep Neural Network for Survival Analysis". In: *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 381–386. DOI: 10.1109/BIBM.2018.8621345.

He, K., X. Zhang, S. Ren, and J. Sun (2016). "Deep residual learning for image recognition". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778.

Hu, S., E. Fridgeirsson, G. v. Wingen, and M. Welling (May 2021). "Transformer-Based Deep Survival Analysis". en. In: *Proceedings of AAAI Spring Symposium on Survival Prediction - Algorithms, Challenges, and Applications 2021*. ISSN: 2640-3498. PMLR, pp. 132–148. URL: https://proceedings.mlr.press/v146/hu21a.html (visited on 10/31/2021).

Huang, C., A. Zhang, and G. Xiao (Jan. 2018). "Deep Integrative Analysis for Survival Prediction". en. In: *Biocomputing 2018*. Kohala Coast, Hawaii, USA: WORLD SCIENTIFIC, pp. 343–352. ISBN: 978-981-323-552-6 978-981-323-553-3. DOI: 10.1142/9789813235533_0032. URL: https://www.worldscientific.com/doi/abs/10.1142/9789813235533_0032 (visited on 12/16/2021).

Huang, Z., X. Zhan, S. Xiang, T. S. Johnson, B. Helm, C. Y. Yu, J. Zhang, P. Salama, M. Rizkalla, Z. Han, et al. (2019). "SALMON: survival analysis learning with multi-omics neural networks on breast cancer". In: *Frontiers in genetics* 10, p. 166.

Irvin, J., P. Rajpurkar, M. Ko, Y. Yu, S. Ciurea-Ilcus, C. Chute, H. Marklund, B. Haghgoo, R. Ball, K. Shpanskaya, et al. (2019). "Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 33. 01, pp. 590–597.

Ishwaran, H., U. B. Kogalur, E. H. Blackstone, M. S. Lauer, et al. (2008). "Random survival forests". In: *The annals of applied statistics* 2.3, pp. 841–860.

Jing, B., T. Zhang, Z. Wang, Y. Jin, K. Liu, W. Qiu, L. Ke, Y. Sun, C. He, D. Hou, et al. (2019). "A deep survival analysis method based on ranking". In: *Artificial intelligence in medicine* 98, pp. 1–9.

Kalakoti, Y., S. Yadav, and D. Sundar (2021). "SurvCNN: a discrete time-to-event cancer survival estimation framework using image representations of omics data". In: *Cancers* 13.13, p. 3106.

Kalbfleisch, J. D. and R. L. Prentice (2011). *The statistical analysis of failure time data*. John Wiley & Sons.

Kaplan, E. L. and P. Meier (1958). "Nonparametric estimation from incomplete observations". In: *Journal of the American statistical association* 53.282, pp. 457–481.

Katzman, J. L., U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger (2018). "DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network". In: *BMC medical research methodology* 18.1, p. 24.

Khan, F. M. and V. B. Zubek (2008). "Support vector regression for censored data (SVRc): a novel tool for survival analysis". In: *2008 Eighth IEEE International Conference on Data Mining*. IEEE, pp. 863–868.

Kim, S., K. Kim, J. Choe, I. Lee, and J. Kang (July 2020). "Improved survival analysis by learning shared genomic information from pan-cancer data". en. In: *Bioinformatics* 36.Supplement_1, pp. i389–i398. ISSN: 1367-4803, 1460-2059. DOI: 10.1093/bioinformatics/btaa462. URL: https://academic.oup.com/bioinformatics/article/36/Supplement_1/i389/5870509 (visited on 12/01/2021).

Klein, J. P. and M. L. Moeschberger (1997). *Survival analysis: Techniques for censored and truncated data*. New York: Springer. ISBN: 978-0-387-94829-4.

Kopper, P., S. Wiegrebe, B. Bischl, A. Bender, and D. Rügamer (2022). "DeepPAMM: Deep Piecewise Exponential Additive Mixed Models for Complex Hazard Structures in Survival Analysis". In: *Advances in Knowledge Discovery and Data Mining: 26th Pacific-Asia Conference, PAKDD 2022, Chengdu, China, May 16–19, 2022, Proceedings, Part II*. Springer, pp. 249–261.

Krzyziński, M., M. Spytek, H. Baniecki, and P. Biecek (2022). "SurvSHAP (t): Time-dependent explanations of machine learning survival models". In: *Knowledge-Based Systems*, p. 110234.

Kvamme, H. and Ø. Borgan (2021). "Continuous and discrete-time survival prediction with neural networks". In: *Lifetime data analysis* 27, pp. 710–736.

Kvamme, H., Ø. Borgan, and I. Scheel (2019). "Time-to-event prediction with neural networks and Cox regression". In: *Journal of machine learning research* 20.129, pp. 1–30.

LeCun, Y., B. Boser, J. S. Denker, D. Henderson, R. E. Howard, W. Hubbard, and L. D. Jackel (1989). "Backpropagation applied to handwritten zip code recognition". In: *Neural computation* 1.4, pp. 541–551.

Lee, B., S. H. Chun, J. H. Hong, I. S. Woo, S. Kim, J. W. Jeong, J. J. Kim, H. W. Lee, S. J. Na, K. S. Beck, B. Gil, S. Park, H. J. An, and Y. H. Ko (Feb. 2020). "DeepBTS: Prediction of Recurrence-free Survival of Non-small Cell Lung Cancer Using a Time-binned Deep Neural Network". en. In: *Scientific Reports* 10.1. Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Computational science;Non-small-cell lung cancer Subject_term_id: computational-science;non-small-cell-lung-cancer, p. 1952. ISSN: 2045-2322. DOI: 10.1038/s41598-020-58722-z. URL: http://www.nature.com/articles/s41598-020-58722-z (visited on 12/06/2021).

Lee, C., J. Yoon, and M. Van Der Schaar (2019). "Dynamic-Deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data". In: *IEEE Transactions on Biomedical Engineering* 67.1, pp. 122–133.

Lee, C., W. R. Zame, J. Yoon, and M. van der Schaar (2018). "DeepHit: A Deep Learning Approach to Survival Analysis With Competing Risks." In: *AAAI*, pp. 2314–2321.

Lee, S. C. and E. T. Lee (1975). "Fuzzy neural networks". In: *Mathematical Biosciences* 23.1-2, pp. 151–177.

Li, Y., W. Jia, Y. Kang, T. Chen, X. Li, X. Du, J. Dong, C. Ma, F. Wang, and G. Xie (Dec. 2020). "DeepComp: Which Competing Event Will Hit the Patient First?" In: *2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pp. 629–636. DOI: 10.1109/BIBM49941.2020.9313333.

Liestbl, K., P. K. Andersen, and U. Andersen (1994). "Survival analysis and neural nets". In: *Statistics in medicine* 13.12, pp. 1189–1200.

Lin, J. and S. Luo (2022). "Deep learning for the dynamic prediction of multivariate longitudinal and survival data". In: *Statistics in medicine* 41.15, pp. 2894–2907.

McCulloch, W. S. and W. Pitts (1943). "A logical calculus of the ideas immanent in nervous activity". In: *The bulletin of mathematical biophysics* 5.4, pp. 115–133.

Meister, R. and C. Schaefer (2008). "Statistical methods for estimating the probability of spontaneous abortion in observational studies—analyzing pregnancies exposed to coumarin derivatives". In: *Reproductive Toxicology* 26.1, pp. 31–35.

Nagpal, C., W. Potosnak, and A. Dubrawski (2022). "auton-survival: an Open-Source Package for Regression, Counterfactual Estimation, Evaluation and Phenotyping with Censored Time-to-Event Data". In.

Nagpal, C., S. Yadlowsky, N. Rostamzadeh, and K. Heller (2021a). "Deep Cox mixtures for survival regression". In: *Machine Learning for Healthcare Conference*. PMLR, pp. 674–708.

Nagpal, C., X. Li, and A. Dubrawski (2021b). "Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks". In: *IEEE Journal of Biomedical and Health Informatics* 25.8, pp. 3163–3175.

Nagpal, C., V. Jeanselme, and A. Dubrawski (2021c). "Deep parametric time-to-event regression with time-varying covariates". In: *Survival Prediction-Algorithms, Challenges and Applications*. PMLR, pp. 184–193.

Nelson, W. (1969). "Hazard plotting for incomplete failure data". In: *Journal of Quality Technology* 1.1, pp. 27–52.

— (1972). "Theory and applications of hazard plotting for censored failure data". In: *Technometrics* 14.4, pp. 945–966.

Nezhad, M. Z., N. Sadati, K. Yang, and D. Zhu (Jan. 2019). "A Deep Active Survival Analysis approach for precision treatment recommendations: Application of prostate cancer". en. In: *Expert Systems with Applications* 115, pp. 16–26. ISSN: 0957-4174. DOI: 10.1016/j.eswa.2018.07.070. URL: https://www.sciencedirect.com/science/article/pii/S0957417418304949 (visited on 07/23/2021).

Noordzij, M., K. Leffondré, K. J. van Stralen, C. Zoccali, F. W. Dekker, and K. J. Jager (2013). "When do we need competing risks methods for survival analysis in nephrology?" In: *Nephrology Dialysis Transplantation* 28.11, pp. 2670–2677.

Pölsterl, S. (2020). "scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn". In: *Journal of Machine Learning Research* 21.212, pp. 1–6. URL: http://jmlr.org/papers/v21/20-729.html.

Pölsterl, S., I. Sarasua, B. Gutiérrez-Becker, and C. Wachinger (2020). "A Wide and Deep Neural Network for Survival Analysis from Anatomical Shape and Tabular Clinical Data". en. In: *Machine Learning and Knowledge Discovery in Databases*. Ed. by P. Cellier and K. Driessens. Communications in Computer and Information Science. Cham: Springer International Publishing, pp. 453–464. ISBN: 978-3-030-43823-4. DOI: 10.1007/978-3-030-43823-4_37.

Qi, C. R., H. Su, K. Mo, and L. J. Guibas (2017). "Pointnet: Deep learning on point sets for 3d classification and segmentation". In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 652–660.

Qin, X., D. Yin, X. Dong, D. Chen, and S. Zhang (2022). "Survival prediction model for right-censored data based on improved composite quantile regression neural network." In: *Mathematical Biosciences and Engineering: MBE* 19.8, pp. 7521–7542.

Qiu, Y. L., H. Zheng, A. Devos, H. Selby, and O. Gevaert (2020). "A meta-learning approach for genomic survival analysis". In: *Nature communications* 11.1, p. 6350.

Ramjith, J., K. C. Roes, H. J. Zar, and M. A. Jonker (2021). "Flexible modelling of risk factors on the incidence of pneumonia in young children in South Africa using piece-wise exponential additive mixed modelling". In: *BMC Medical Research Methodology* 21.1, pp. 1–13.

Ranganath, R., A. Perotte, N. Elhadad, and D. Blei (2016). "Deep survival analysis". In: *Machine Learning for Healthcare Conference*. PMLR, pp. 101–114.

Ren, K., J. Qin, L. Zheng, Z. Yang, W. Zhang, L. Qiu, and Y. Yu (2019). "Deep recurrent survival analysis". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01, pp. 4798–4805.

Rezende, D. and S. Mohamed (2015). "Variational inference with normalizing flows". In: *International conference on machine learning*. PMLR, pp. 1530–1538.

Rügamer, D., C. Kolb, and N. Klein (2023). "Semi-Structured Distributional Regression". In: *The American Statistician* 0.0, pp. 1–12. DOI: 10.1080/00031305.2022.2164054. eprint: https://doi.org/10.1080/00031305.2022.2164054. URL: https://doi.org/10.1080/00031305.2022.2164054.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams (1986). "Learning representations by back-propagating errors". In: *nature* 323.6088, pp. 533–536.

Sabour, S., N. Frosst, and G. E. Hinton (2017). "Dynamic routing between capsules". In: *Advances in neural information processing systems* 30.

Sansaengtham, B., V. C. Barroso, and P. Phunchongharn (2020). "Survival Analysis For Computing Systems Using A Deep Ensemble Network". In: *2020 IEEE 6th International Conference on Control Science and Systems Engineering (ICCSSE)*. IEEE, pp. 57–62.

Schwarzer, G., W. Vach, and M. Schumacher (2000). "On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology". In: *Statistics in medicine* 19.4, pp. 541–561.

Shin, B., S. Park, J. H. Hong, H. J. An, S. H. Chun, K. Kang, Y.-H. Ahn, Y. H. Ko, and K. Kang (2019). "Cascaded Wx: A novel prognosis-related feature selection framework in human lung adenocarcinoma transcriptomes". In: *Frontiers in Genetics* 10, p. 662.

Sohl-Dickstein, J., E. A. Weiss, N. Maheswaranathan, and S. Ganguli (2015). "Deep Unsupervised Learning using Nonequilibrium Thermodynamics". In: *ICML*.

Sonabend, R., F. J. Király, A. Bender, B. Bischl, and M. Lang (Feb. 2021). "mlr3proba: An R Package for Machine Learning in Survival Analysis". In: *Bioinformatics*. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btab039.

Sonabend, R. E. B. (2021). "A Theoretical and Methodological Framework for Machine Learning in Survival Analysis: Enabling Transparent and Accessible Predictive Modelling on Right-Censored Time-to-Event Data". PhD. University College London (UCL), p. 345. URL: https://discovery.ucl.ac.uk/id/eprint/10129352/.

Steele, A. J., S. C. Denaxas, A. D. Shah, H. Hemingway, and N. M. Luscombe (2018). "Machine learning models in electronic health records can outperform conventional survival models for predicting patient mortality in coronary artery disease". In: *PloS one* 13.8, e0202344.

Sun, T., Y. Wei, W. Chen, and Y. Ding (2020). "Genome-wide association study-based deep learning for survival prediction". In: *Statistics in medicine* 39.30, pp. 4605–4620.

Tang, B., A. Li, B. Li, and M. Wang (2019). "CapSurv: Capsule Network for Survival Analysis With Whole Slide Pathological Images". In: *IEEE Access* 7. Conference Name: IEEE Access, pp. 26022–26030. ISSN: 2169-3536. DOI: 10.1109/ACCESS.2019.2901049.

Thorsen-Meyer, H.-C., D. Placido, B. S. Kaas-Hansen, A. P. Nielsen, T. Lange, A. B. Nielsen, P. Toft, J. Schierbeck, T. Strøm, P. J. Chmura, et al. (2022). "Discrete-time survival analysis in the critically ill: a deep learning approach using heterogeneous data". In: *NPJ digital medicine* 5.1, p. 142.

Tong, J. and X. Zhao (2022). "Deep survival algorithm based on nuclear norm". In: *Journal of Statistical Computation and Simulation* 92.9, pp. 1964–1976.

Tong, L., J. Mitchel, K. Chatlin, and M. D. Wang (Dec. 2020). "Deep learning based feature-level integration of multi-omics data for breast cancer patients survival analysis". en. In: *BMC Medical Informatics and Decision Making* 20.1, p. 225. ISSN: 1472-6947. DOI: 10.1186/s12911-020-01225-8. URL: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-01225-8 (visited on 12/02/2021).

Tutz, G., M. Schmid, et al. (2016). *Modeling discrete time-to-event data*. Springer.

Utkin, L. V., E. D. Satyukov, and A. V. Konstantinov (2022). "SurvNAM: The machine learning survival model explanation". In: *Neural Networks* 147, pp. 81–102.

Vale-Silva, L. A. and K. Rohr (2021). "Long-term cancer survival prediction using multimodal deep learning". In: *Scientific Reports* 11.1, p. 13505.

Van Belle, V., K. Pelckmans, S. Van Huffel, and J. A. Suykens (2011). "Support vector methods for survival analysis: a comparison between ranking and regression approaches". In: *Artificial intelligence in medicine* 53.2, pp. 107–118.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin (2017). "Attention is all you need". In: *Advances in neural information processing systems* 30.

Vellido, A. (2020). "The importance of interpretability and visualization in machine learning for applications in medicine and health care". In: *Neural computing and applications* 32.24, pp. 18069–18083.

Wang, H. and G. Li (2019a). "Extreme learning machine Cox model for high-dimensional survival analysis". In: *Statistics in medicine* 38.12, pp. 2139–2156.

Wang, J., N. Chen, J. Guo, X. Xu, L. Liu, and Z. Yi (2021). "SurvNet: A Novel Deep Neural Network for Lung Cancer Survival Analysis With Missing Values". In: *Frontiers in Oncology* 10, p. 3128. ISSN: 2234-943X. DOI: 10.3389/fonc.2020.588990. URL: https://www.frontiersin.org/article/10.3389/fonc.2020.588990 (visited on 12/02/2021).

Wang, P., Y. Li, and C. K. Reddy (2019b). "Machine learning for survival analysis: A survey". In: *ACM Computing Surveys (CSUR)* 51.6, pp. 1–36.

Wang, Z. and J. Sun (2022). "Survtrace: Transformers for survival analysis with competing events". In: *Proceedings of the 13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, pp. 1–9.

Wolf, T. N., S. Pölsterl, C. Wachinger, A. D. N. Initiative, et al. (2022). "DAFT: A universal module to interweave tabular data and 3D images in CNNs". In: *NeuroImage* 260, p. 119505.

Wolpert, D. H. (1992). "Stacked generalization". In: *Neural networks* 5.2, pp. 241–259.

Xie, G., C. Dong, Y. Kong, J. F. Zhong, M. Li, and K. Wang (2019). "Group lasso regularized deep learning for cancer prognosis from multi-omics and clinical features". In: *Genes* 10.3, p. 240.

Xie, Y. and Z. Yu (Dec. 2021). "Mixture cure rate models with neural network estimated nonparametric components". en. In: *Computational Statistics* 36.4, pp. 2467–2489. ISSN: 0943-4062, 1613-9658. DOI: 10.1007/s00180-021-01086-3. URL: https://link.springer.com/10.1007/s00180-021-01086-3 (visited on 12/02/2021).

Yin, Q., W. Chen, C. Zhang, and Z. Wei (2022). "A convolutional neural network model for survival prediction based on prognosis-related cascaded Wx feature selection". In: *Laboratory Investigation* 102.10, pp. 1064–1074.

Yousefi, S., F. Amrollahi, M. Amgad, C. Dong, J. E. Lewis, C. Song, D. A. Gutman, S. H. Halani, J. E. Velazquez Vega, D. J. Brat, et al. (2017). "Predicting clinical outcomes from large scale cancer genomic profiles with deep survival models". In: *Scientific reports* 7.1, p. 11707.

Zhang, J. and K. Huang (2014). "Normalized imqcm: An algorithm for detecting weak quasi-cliques in weighted graph with applications in gene co-expression module discovery in cancers". In: *Cancer informatics* 13, CIN–S14021.

Zhang, Y., E. M. Lobo-Mueller, P. Karanicolas, S. Gallinger, M. A. Haider, and F. Khalvati (2020). "CNN-based survival model for pancreatic ductal adenocarcinoma in medical imaging". In: *BMC medical imaging* 20, pp. 1–8.

Zhao, L., Q. Dong, C. Luo, Y. Wu, D. Bu, X. Qi, Y. Luo, and Y. Zhao (2021). "DeepOmix: A scalable and interpretable multi-omics deep learning framework and application in cancer survival analysis". In: *Computational and structural biotechnology journal* 19, pp. 2719–2725.

Zhao, L. and D. Feng (Aug. 2019). "DNNSurv: Deep Neural Networks for Survival Analysis Using Pseudo Values". In: *arXiv:1908.02337 [cs, stat]*. arXiv: 1908.02337 version: 1. URL: http://arxiv.org/abs/1908.02337 (visited on 03/20/2021).

Zhu, X., J. Yao, and J. Huang (Dec. 2016). "Deep convolutional neural network for survival analysis with pathological images". In: *2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. Shenzhen, China: IEEE, pp. 544–547. ISBN: 978-1-5090-1611-2. DOI: 10.1109/BIBM.2016.7822579. URL: http://ieeexplore.ieee.org/document/7822579/ (visited on 11/30/2021).