scientific reports

OPEN

Leveraging the variational Bayes autoencoder for survival analysis

Patricia A. Apellániz, Juan Parras & Santiago Zazo

Survival analysis in medical research has witnessed a growing interest in applying deep learning techniques to model complex, high-dimensional, heterogeneous, incomplete, and censored data. Current methods make assumptions about the relations between data that may not be valid in practice. Therefore, we introduce SAVAE (Survival Analysis Variational Autoencoder). SAVAE, based on Variational Autoencoders, contributes significantly to the field by introducing a tailored Evidence Lower BOund formulation, supporting various parametric distributions for covariates and survival time (if the log-likelihood is differentiable). It offers a general method that demonstrates robustness and stability through different experiments. Our proposal effectively estimates time-to-event, accounting for censoring, covariate interactions, and time-varying risk associations. We validate our model in diverse datasets, including genomic, clinical, and demographic tabular data, with varying levels of censoring. This approach demonstrates competitive performance compared to state-of-the-art techniques, as assessed by the Concordance Index and the Integrated Brier Score. SAVAE also offers an interpretable model that parametrically models covariates and time. Moreover, its generative architecture facilitates further applications such as clustering, data imputation, and synthetic patient data generation through latent space inference from survival data. This approach fosters data sharing and collaboration, improving medical research and personalized patient care.

Keywords Survival analysis, Time to event, Censored data, Deep learning, Variational autoencoders

In recent years, there has been a significant transformation in medical research methodologies towards the adoption of Deep Learning (DL) techniques for predicting critical events, such as disease development and patient mortality. Despite their potential to handle complex data, practical applications in this domain still need to be expanded, with most studies still relying on traditional statistical methods.

Survival Analysis (SA), or time-to-event analysis, is an essential tool for studying specific events in various disciplines, not only in medicine but also in fields such as recommendation systems[1], employee retention[2], market modeling[3], and financial risk assessment[4].

According to the existing literature, the Cox proportional hazards model (Cox-PH)[5] is the dominant SA method that offers a semiparametric regression solution to the non-parametric Kaplan-Meier estimator problem[6]. Unlike the Kaplan-Meier method, which uses a single covariate, Cox-PH incorporates multiple covariates to predict event times and assess their impact on the hazard rate at specific time points. However, it is crucial to acknowledge that the Cox-PH model is built on certain strong assumptions. One of these is the proportional hazards assumption, which posits that different individuals have hazard functions that remain constant over time. Also, the model assumes a linear relation between the natural logarithm of the relative hazard (the ratio of the hazard at time t to the baseline hazard) and the covariates. Although the standard Cox-PH model assumes the absence of interactions among these covariates, it can be extended by introducing interaction terms, such as quadratic or higher-order terms, allowing the modeling of more complex relations between covariates. However, even with these extensions, the model may struggle to capture non-linearities in real-world datasets, where intricate interactions between covariates and non-linear relationships might exist. Other traditional parametric statistical models for SA make specific assumptions about the distribution of event times. For instance, there are models that assume exponential and Weibull distributions, respectively, for event times[7,8]. However, one drawback of these models is that they lack flexibility when changing the assumed distribution for survival times, making them less adaptable to diverse datasets.

In response, researchers have explored Deep Neural Networks (DNNs) to effectively capture the intricate and non-linear relations between predictive variables and a patient's risk of failure. Significant emphasis has been placed on improving the Cox PH model, the standard SA approach. Recent approaches have introduced Neural Networks (NN) in various configurations, either enhancing the Cox-PH model with neural components or proposing entirely novel architectures. This exploration of NN applications for SA traces back to 1995[9], when

Information Processing and Telecommunications Center, ETSI Telecomunicación, Universidad Politécnica de Madrid, Avda. Complutense, 30, 28040 Madrid, Spain. email: patricia.alonsod@upm.es

a simple feed-forward NN was employed to replace linear interaction terms while incorporating non-linearities. Subsequently, the field saw the emergence of DeepSurv[10], a model designed to extract non-linearities from input data, albeit still assuming the proportional hazards assumption. This assumption persists in other related models[11]. Beyond addressing non-linearity, some researchers have sought to enhance prediction accuracy and model interpretability by combining Bayesian networks with the Cox-PH model[12]. Additionally, efforts have been made to introduce concepts that facilitate analysis when data availability is limited[13,14]. However, it is essential to note that all these models still depend on the assumption of proportional hazards. As a result, novel architectures such as DeepHit[15] have emerged as alternatives that do not rely on this assumption. While DeepHit has exhibited superior performance compared to other state-of-the-art models, it operates exclusively in the discrete-time domain, which comes with certain limitations, notably the requirement for a dataset with a substantial number of observations. This condition may not be feasible in real-world scenarios.

In light of the persistent limitations of existing approaches in the realm of SA, this paper introduces a novel, versatile algorithm grounded in DL advances named SAVAE (Survival Analysis Variational Autoencoder). SAVAE has been meticulously designed to predict the time distribution leading to a predefined event and adapts to application in various domains, explicitly focusing on the medical context. Then, our main contributions consist of:

- We introduce a generative approach that underpins the development of a flexible tool, SAVAE, based on Variational Autoencoders (VAEs). SAVAE can effectively reproduce the data by analytically modeling the discrete or continuous time to a specific event. This analytical approach enables the precision calculation of all necessary statistics, as the output provided by SAVAE are the estimated parameters of the predicted time distribution.
- SAVAE is a flexible tool that enables us to use various distributions to model the time-to-event and the covariates. This allows us not to assume proportional hazards. Using NN permits modeling complex, non-linear relations between the covariates and the time-to-event, as opposed to linearity assumptions in the state of the art. Also, the time-to-event is trained with standard likelihood techniques, unlike state-of-the-art models like DeepHit, which trains the Concordance Index (C-index). This makes our approach more general and flexible, as any differentiable distribution could be used to model the time and the covariates.
- Furthermore, our proposal can be trained on right-censored data, effectively leveraging information from patients who have not yet experienced the event of interest.
- We have conducted comprehensive time-to-event estimation experiments using datasets characterized by continuous and discrete time-to-event values and varying covariate natures, encompassing clinical and genomic data. These experiments involve a comparative analysis with the traditional Cox-PH model and other DL techniques. The results indicate that SAVAE is competitive with these models regarding the C-index and the Integrated Brier score (IBS).

Methods

Survival analysis

In a conventional time-to-event or SA setup, *N* observations are given. Each of these observations is described by $D=(x_i,t_i,d_i)_{i=1}^N$ triplets, where $x_i=(x_i^1,\dots,x_i^L)$ is an L-dimensional vector where $l=1,2,\dots,L$ indexes the covariates, t_i is the time-to-event, and $d_i\in\{0,1\}$ is the censor indicator. When $d_i=0$ (censored), the subject has not experienced an event up to time t_i, while $d_i=1$ indicates the observed events (ground truth). SA models are conditional on covariates: time probability density function $p(t|x)$, hazard rate function (the instantaneous rate of occurrence of the event at a specific time) $h(t|x)$, or survival function $S(t|x)=P(T>t)=1-F(t|x)$, also known as the probability of a failure occurring after time *t*, where $F(t|x)$ is the Cumulative Distribution Function (CDF) of the time. From standard definitions of the survival function, the relations between these three characterizations are formulated as follows:

$$p(t|x)=h(t|x)S(t|x). \tag{1}$$

Vanilla variational autoencoder

The original VAE was proposed in 2013[16], a robust approach employing DNNs for Bayesian inference. It addresses the problem of a dataset consisting of *N* i.i.d. samples x_i of a continuous or discrete variable, where $i\in 1,2,\dots,N$, x_i are generated by the following random process, which is depicted in Fig. 1:

1. A latent variable z_i is sampled from a given prior probability distribution *p*(*z*). The original research[16] assumes a form $p_\theta(z)$, i.e., the prior depends on some parameters θ, but its main result drops this dependence. Therefore, a simple prior *p*(*z*) is assumed in this paper.
2. A conditional distribution, $p_\theta(x|z)$, with parameters θ generates the observed values, x_i. A generative model governs this process. Certain assumptions are made, including the differentiability of probability density functions (pdfs), *p*(*z*), and $p_\theta(x|z)$, regarding θ and *z*.The latent variable *z* and the parameters θ are unknown. Without simplifying assumptions, evaluating the marginal likelihood $p_\theta(x)=\int p(z)p_\theta(x|z)dz$ is infeasible. The true posterior density $p_\theta(z|x)$, which we aim to approximate, can be defined as Eq. (2) using Bayes' theorem:

$$p_\theta(z|x)=\frac{p_\theta(x|z)p(z)}{p_\theta(x)}. \tag{2}$$

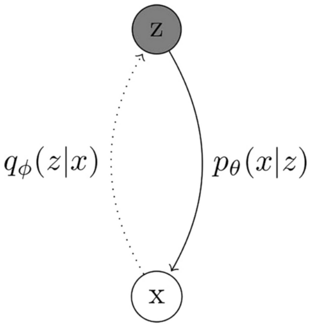

Figure 1. Bayesian VAE vanilla model. The shaded circle refers to the latent variable *z*, and the white circle refers to the observable *x*. Probabilities $p_\theta(x|z)$ and $q_\phi(z|x)$ denote, respectively, the generative model and the variational approximation to the posterior, since the true posterior $p_\theta(z|x)$ is unknown.

However, since the marginal likelihood $p_\theta(x)$ is often intractable, direct computation of the true posterior $p_\theta(z|x)$ is not practicable.

Variational methods offer a solution by introducing a variational approximation, $q_\phi(z|x)$, to the true posterior. This approximation involves optimizing the best parameters for a chosen family of distributions. The quality of the approximation depends on the expressiveness of this parametric family.

ELBO derivation

Since an optimization problem must be solved, the optimization target needs to be developed. Considering x_i are assumed to be i.i.d., the marginal likelihood of a set of points $\{x_i\}_{i=1}^N$ can be expressed as

$$\log p_\theta(x_1, x_2, \ldots, x_N) = \sum_{i=1}^{N} \log p_\theta(x_i), \tag{3}$$

where

$$p_\theta(x) = \int p_\theta(x,z)dz = \int p_\theta(x,z)\frac{q_\phi(z|x)}{q_\phi(z|x)}dz = \mathbb{E}_{q_\phi(z|x)}\left[\frac{p_\theta(x,z)}{q_\phi(z|x)}\right]. \tag{4}$$

Using Jensen's inequality, we can obtain:

$$\log p_\theta(x) = \log\left[\mathbb{E}_{q_\phi(z|x)}\left[\frac{p_\theta(x,z)}{q_\phi(z|x)}\right]\right] \geq \mathbb{E}_{q_\phi(z|x)}\left[\log\frac{p_\theta(x,z)}{q_\phi(z|x)}\right]. \tag{5}$$

Rearranging Eq. (5), we can express it as follows:

$$\begin{aligned}
&\mathbb{E}_{q_\phi(z|x)}\left[\log\left(\frac{p_\theta(x,z)}{q_\phi(z|x)}\right)\right]\\
&\quad= \int q_\phi(z|x)\log\frac{p_\theta(x|z)p(z)}{q_\phi(z|x)}dz\\
&\quad= \int q_\phi(z|x)\log\frac{p(z)}{q_\phi(z|x)}dz + \int q_\phi(z|x)\log p_\theta(x|z)dz\\
&\quad= -\int q_\phi(z|x)\log\frac{q_\phi(z|x)}{p(z)}dz + \int q_\phi(z|x)\log p_\theta(x|z)dz\\
&\quad= -D_{KL}(q_\phi(z|x)||p(z)) + \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right]\\
&\quad= \mathcal{L}(x,\theta,\phi),
\end{aligned} \tag{6}$$

where $D_{KL}(p||q)$ is the Kullback-Leibler divergence between distributions *p* and *q*, and $\mathcal{L}(x,\theta,\phi)$ is the Evidence Lower BOund (ELBO), whose name comes from Eq. (5):

$$\log p_\theta(x) \geq -D_{KL}(q_\phi(z|x)||p(z)) + \mathbb{E}_{q_\phi(z|x)}\left[\log p_\theta(x|z)\right] = \mathcal{L}(x,\theta,\phi), \tag{7}$$

the ELBO is a lower bound for the marginal log-likelihood of the relevant set of points. Thus, maximizing the ELBO maximizes the log-likelihood of the data. This would be the optimization problem to solve.

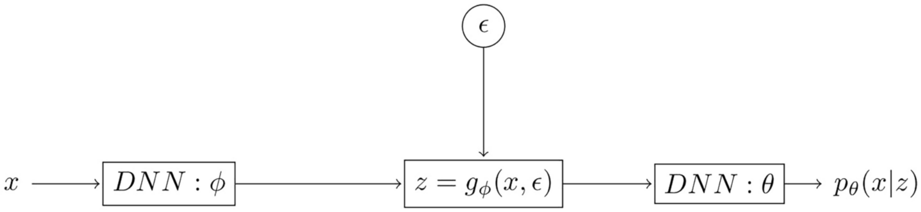

Figure 2. VAE vanilla model implementation using DNNs.

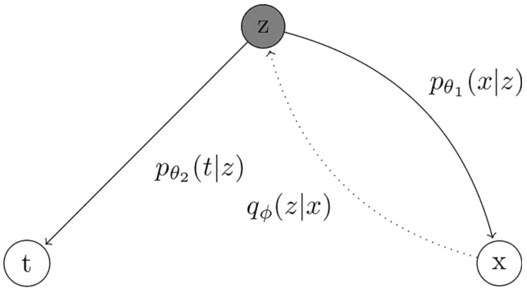

Figure 3. SAVAE Bayesian model. The shadowed circle refers to the latent variable, and the white circles refer to the observables. Note that the probabilities $p_{\theta_1}(x|z)$ and $p_{\theta_2}(t|z)$ denote the generative models, and $q_\phi(z|x)$ denotes the variational approximation to the posterior, since the true posterior $p_\theta(z|x)$ is unknown.

Implementation

The ELBO derived from Eq. (7) can be effectively implemented using a DNN-based architecture. However, computing the gradient of the ELBO concerning ϕ presents challenges due to the presence of ϕ in the expectation term (the second part of the ELBO in Eq. (7)). To address this issue, the original research[16] introduced the reparameterization trick. This method involves modifying the latent space sampling process to make it differentiable, enabling gradient-based optimization techniques. Rather than sampling directly from the latent space distribution, VAEs sample ϵ from a simple distribution, often a standard normal distribution. Subsequently, a deterministic transformation g_ϕ is applied to ϵ, producing $z = g_\phi(x, \epsilon)$ where $z \sim q_\phi(z|x)$ and $\epsilon \sim p(\epsilon)$. In this case, the ELBO can be estimated as follows.

$$\hat{\mathcal{L}}(x, \theta, \phi) = \frac{1}{N} \sum_{i=1}^{N} \Big(- D_{KL}(q_\phi(z|x_i)||p(z)) + \log p_\theta(x_i|g_\phi(x_i, \epsilon_i)) \Big). \tag{8}$$

This modification facilitates the calculation of the ELBO gradient concerning θ and ϕ, allowing the application of standard gradient optimization methods.

Equation (8) offers a solution using DNNs, with functions parameterized by ϕ and θ. Gradients can be conveniently computed using the Backpropagation algorithm, which various programming libraries automate. The term VAE derives from the fact that Eq. (8) resembles the architecture of an Autoencoder (AE)[17], as illustrated in Fig. 2. The variational distribution q_ϕ can be implemented using a DNN with weights ϕ, taking an input sample *x* and outputting parameters for the deterministic transformation g_ϕ. The VAE's latent space comprises the latent variable *z* distribution, a deterministic transformation g_ϕ of the encoder DNN output and random ancillary noise ϵ. A sampled value z_i is drawn from the latent distribution and used to generate an output sample, where another DNN with weights θ acts as a decoder, taking *z* as input and providing parameters of the distribution $p_\theta(x|z)$ as output.

Two key observations emerge.

1. The ELBO losses in Eq. (7) include a regularization term penalizing deviations from the prior in the latent space and a reconstruction error term that enforces similarity between generated samples from the latent space and inputs.
2. In contrast to standard AEs, VAEs incorporate intermediate sampling, rendering them non-deterministic. This dual sampling process is retained in applications where the distribution of output variables is of interest, facilitating the derivation of input value distribution parameters.

Our contribution

The interest lies in using VAEs to obtain the predictive distribution of time-to-event given covariates. The proposed approach termed Survival Analysis VAE (SAVAE), depicted in Fig. 3, extends the Vanilla VAE. SAVAE includes a continuous latent variable *z*, two vectors (an observable covariate vector *x* and the time-to-event *t*), and generative models $p_{\theta_1}(x|z)$ and $p_{\theta_2}(t|z)$, assuming conditional independence, which is a characteristic inherent to VAEs and their ability to model the joint distribution of variables effectively. This means that knowing *z*, the components of the vector *x* and *t* can be generated independently. A single variational distribution estimates the variational posterior $p(z|x)$ to define the predictive distribution based on covariates. While it is possible to include the effect of time ($p(z|t, x)$), this approach focuses on using only covariates to obtain the latent space, as the time *t* can be unknown to predict survival times for test patients and could be censored. SAVAE combines VAEs and survival analysis, offering a flexible framework for modeling complex event data.

Goal

To achieve the main objective, which is to obtain the predictive distribution for the time to event, variational methods will be used in the following way[18]:

$$p\left(t^*|x^*, \{x_i, t_i\}_{i=1}^N\right) = \int p\left(t^*|z, \{x_i, t_i\}_{i=1}^N\right) p\left(z|x^*, \{x_i, t_i\}_{i=1}^N\right) dz, \tag{9}$$

where x^* represents the covariates of a particular patient, and its survival time distribution $p\left(t^*|z, \{x_i, t_i\}_{i=1}^N\right)$ needs to be estimated.

ELBO derivation

Considering our main objective and the use of VAE as the architecture on which we base our approach, the previous ELBO development can be extended to apply to our case. SAVAE assumes that the two generative models $p_{\theta_1}(x|z)$ and $p_{\theta_2}(t|z)$ are conditionally independent. This implies that if *z* is known, generating *x* or *t* is possible. Furthermore, due to the VAE architecture, it is assumed that each component of the covariate vector *x* is also conditionally independent given *z*. Therefore,

$$p(x, t, z) = p_{\theta_1}(x|z)p_{\theta_2}(t|z)p(z) = p_\theta(x, t|z)p(z). \tag{10}$$

It also assumes that the distribution families of $p_{\theta_1}(x|z)$ and $p_{\theta_2}(t|z)$ are known, but not the parameters θ_1 and θ_2 . Taking into account these assumptions, the ELBO can be computed in a similar way to the Vanilla VAE. First, the conditional likelihood of a set of points $\{x_i, t_i\}_{i=1}^N$ can be expressed as follows:

$$\log p_\theta(x_1, x_2, \ldots, x_N, t_1, t_2, \ldots, t_N|z) = \sum_{i=1}^N \log p_\theta(x_i, t_i|z) = \sum_{i=1}^N \left(\log p_{\theta_2}(t_i|z) + \sum_{l=1}^L \log p_{\theta_1}(x_i^l|z)\right), \tag{11}$$

where the expected conditional likelihood can be expressed as:

$$\begin{aligned} &\mathbb{E}_z\left[p_\theta(x, t|z)\right] \\ &\quad= \int p_\theta(x, t|z)p(z)dz \\ &\quad= \int \frac{p_\theta(x, t, z)}{p(z)}p(z)dz \\ &\quad= \int p_\theta(x, t, z)dz \\ &\quad= p_\theta(x, t) = \int p_\theta(x, t, z)\frac{q_\phi(z|x)}{q_\phi(z|x)}dz \\ &\quad= \mathbb{E}_{q_\phi(z|x)}\left[\frac{p_\theta(x, t, z)}{q_\phi(z|x)}\right]. \end{aligned} \tag{12}$$

As the interest lies in computing the log-likelihood:

$$\log p_\theta(x, t) = \log\left[\mathbb{E}_{q_\phi(z|x)}\left[\frac{p_\theta(x, t, z)}{q_\phi(z|x)}\right]\right] \geq \mathbb{E}_{q_\phi(z|x)}\left[\log\frac{p_\theta(x, t, z)}{q_\phi(z|x)}\right], \tag{13}$$

where the inequality comes from applying Jensen's inequality. Then, this could be rearranged as:

$$
\begin{aligned}
&\mathbb{E}_{q_\phi(z|x)}\left[\log\left(\frac{p_\theta(x,t,z)}{q_\phi(z|x)}\right)\right] \\
&\quad = \int q_\phi(z|x)\log\frac{p_{\theta_1}(x|z)p_{\theta_2}(t|z)p(z)}{q_\phi(z|x)}dz \\
&\quad = -\int q_\phi(z|x)\log\frac{q_\phi(z|x)}{p(z)}dz + \int q_\phi(z|x)\left(\log p_{\theta_1}(x|z)+\log p_{\theta_2}(t|z)\right)dz \\
&\quad = -D_{KL}(q_\phi(z|x)||p(z)) + \mathbb{E}_{q_\phi(z|x)}\left[\log p_{\theta_1}(x|z)+\log p_{\theta_2}(t|z)\right] \\
&\quad = \mathcal{L}(x,\theta_1,\theta_2,\phi).
\end{aligned}
\tag{14}
$$

After computing this ELBO, it can be seen that it is similar to the Vanilla VAE's one (Eq. 8). The only difference lies in the reconstruction term, which is expressed differently to distinguish between the covariates and the time-to-event explicitly. By using Eq. (11) and the reparameterization trick, the ELBO estimator is obtained, explicitly accounting for each dimension of the covariates vector:

$$
\hat{\mathcal{L}}(x,\theta_1,\theta_2,\phi) = \frac{1}{N}\sum_{i=1}^{N}\left(-D_{KL}(q_\phi(z|x_i)||p(z)) + \log p_{\theta_2}(t_i|g_\phi(x_i,\epsilon_i)) + \sum_{l=1}^{L}\log p_{\theta_1}(x_i^l|g_\phi(x_i,\epsilon_i))\right). \tag{15}
$$

Three DNNs have been used in implementation, as specified in Fig. 4. Note that the decoder DNNs output the parameters of each distribution.

Divergence computation

SAVAE assumes that $q_\phi(z|x)$ follows a multidimensional Gaussian distribution defined by a vector of means μ, where each element is μ_j and by a diagonal covariance matrix C, where the main diagonal consists of variances σ_j^2. It can be stated that:

$$
-D_{KL}(q_\phi(z|x)||p(z)) = \frac{1}{2}\sum_{j=1}^{J}(1+\log(\sigma_j^2)-\mu_j^2-\sigma_j^2), \tag{16}
$$

where J is the dimension of the latent space z[16]. This means the Kullback-Leibler divergence from the ELBO equation 15 can be calculated analytically.

Time modeling

One significant challenge in handling survival data is the issue of censorship, which occurs when a patient has not yet experienced the event of interest. In such cases, the survival time remains unknown, resulting in partial or incomplete observations. Consequently, SA models must employ techniques capable of accommodating censored observations and uncensored ones to estimate relevant parameters reliably.

In our case, to account for censoring in survival data, we start from the time t reconstruction term from Eq. 15 for a single patient:

$$
\hat{\mathcal{L}}_{time}(x_i,\theta_2,\phi) = \log p_{\theta_2}(t_i|g_\phi(x_i,\epsilon_i)). \tag{17}
$$

Taking into account the censoring indicator d_i:

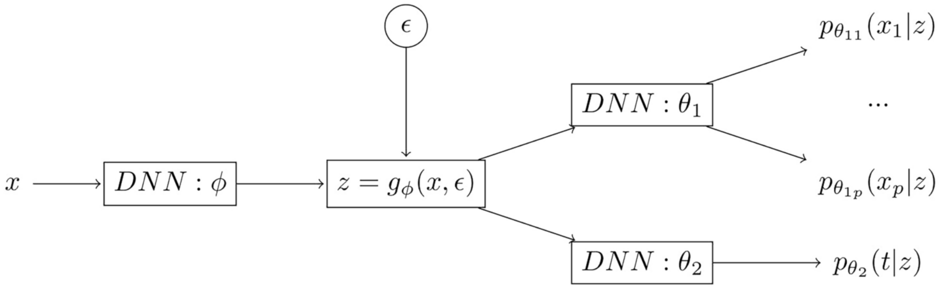

Figure 4. SAVAE implementation using DNNs. One of them acts as an encoder, which has the covariates vector as input. The other two act as decoders, one for the covariates and the other one for the time.

$$d_i = \begin{cases} 0 & \text{if censored} \\ 1 & \text{if event experienced} \end{cases}, \tag{18}$$

we could just use the information given by uncensored patients. However, we would waste information since we know that the censored patients have not experienced the event until time t_i. Hence, considering Eq. (1) and following[19], we model the time pdf as:

$$p_{\theta_2}(t_i|g_\phi(x_i,\epsilon_i)) = h(t_i|g_\phi(x_i,\epsilon_i))^{d_i} S(t_i|g_\phi(x_i,\epsilon_i)). \tag{19}$$

Therefore, the hazard function term is only considered when the event has been experienced, when the data are not censored. This way, SAVAE incorporates information from censored observations, providing consistent parameter estimates.

Regarding the distribution chosen for the time event, we have followed several publications such as[8], where the Weibull distribution model is used. This distribution is two-parameter, with positive support, that is, $p(t) = 0, \forall t < 0$. The two scalar parameters of the distribution are λ and α, where $\lambda > 0$ controls the scale and $\alpha > 0$ controls the shape as follows:

$$\begin{cases} p(t;\alpha,\lambda) = \frac{\alpha}{\lambda}\left(\frac{t}{\lambda}\right)^{\alpha-1} \exp\left(-\left(\frac{t}{\lambda}\right)^{\alpha}\right) \\ S(t;\alpha,\lambda) = \frac{\exp}{\left(-\left(\frac{t}{\lambda}\right)^{\alpha}\right)} \\ h(t;\alpha,\lambda) = \frac{p(t;\alpha,\lambda)}{S(t;\alpha,\lambda)} = \frac{\alpha}{\lambda}\left(\frac{t}{\lambda}\right)^{\alpha-1} \end{cases}. \tag{20}$$

Although the Weibull distribution is our primary choice for modeling time-to-event data in SAVAE, it is crucial to highlight that other distributions are feasible as long as their hazard functions and CDFs can be analytically calculated. This versatility distinguishes SAVAE from other models. For example, the exponential distribution, a particular case of Weibull with $\alpha = 1$, can represent constant hazard functions. Integrating alternative distributions, such as the exponential, into SAVAE is straightforward and only requires adjusting the terms in Eq. (19). The ability of SAVAE to predict the distribution parameters for each patient facilitates the calculation of various statistics, such as means, medians, and percentiles, providing flexibility beyond the models customized to a single distribution.

Marginal log-likelihood computation

Assigning distribution models to patient covariates in the reconstruction term is essential in SAVAE. This choice enables control over the resulting output variable distribution, but it also implies that the model approximates the chosen distribution even if the actual distribution differs. The third component of the ELBO (15) depends on the log-likelihood of the data, which for some representative distributions is:

- Gaussian distribution: Suitable for real-numbered variables ($x_i^l \in (-\infty, +\infty)$), it has parameters $\mu \in (-\infty, +\infty)$ and $\sigma \in (0, +\infty)$, known for its symmetric nature. Its log-likelihood function is:

$$\log(p(x_i^l;\mu,\sigma)) = -\log(\sigma\sqrt{2\pi}) - \frac{1}{2}\left(\frac{x_i^l - \mu}{\sigma}\right)^2. \tag{21}$$

-
- Bernoulli distribution: Applied to binary variables ($x_i^l \in \{0,1\}$), it has a single parameter $\beta \in [0,1]$, representing the probability of $x_i^l = 1$.Its log-likelihood function is:

$$\log(p(x_i^l;\beta)) = x_i^l \log(\beta) + (1 - x_i^l)\log(1-\beta). \tag{22}$$

-
- Categorical distribution: Models discrete variables with *K* possible values. We can think of x_i^l as a categorical scalar random variable with *K* different values. Each possible outcome is assigned a probability θ_k (note that $\sum_{k=1}^{K}\theta_k = 1$). The log-likelihood function can be computed based on the Probability Mass Function (PMF) following the expression:

$$\log(p(x_i^l|\theta_1,\theta_2,\ldots,\theta_k)) = \log\left(\prod_{k=1}^{K}\theta_k^{\mathbb{I}(x_i^l=k)}\right), \tag{23}$$

where the indicator function means:

$$\mathbb{I}(x_i^l = k) = \begin{cases} 1 & x_i^l = k \\ 0 & x_i^l \neq k \end{cases}. \tag{24}$$

- Recall that other desired distributions can be implemented in SAVAE if their log-likelihood is differentiable.

Results

This section proceeds with the experimental validation of SAVAE. First, we describe the survival data and the performance metrics used to validate the model. Then, we define the experimental setup (network architecture and training process). Finally, we analyze the different experiments carried out. The code can be found in https://github.com/Patricia-A-Apellaniz/savae.

Survival data

In SA datasets, each patient contributes information about whether events of interest occurred during a study period, categorizing them as censored or uncensored and indicating their respective follow-up times. To evaluate SAVAE, we trained it in nine diverse disease datasets, including WHAS, SUPPORT, GBSG, FLCHAIN, NWTCO, METABRIC, PBC, STD, and PNEUMON. We followed pre-processing procedures similar to state-of-the-art models, ensuring a fair evaluation of established benchmarks in SA.

The Worcester Heart Attack Study (WHAS)[20] focuses on patients with acute myocardial infarction (AMI), providing clinical and demographic data. The Study to Understand Prognoses Outcomes and Risks of Treatment (SUPPORT)[21] investigates seriously ill hospitalized adults and includes information on demographics, comorbidities, and physiological measurements. The Rotterdam & German Breast Cancer Study Group (GBSG)[22,23] combines data from node-positive breast cancer patients and a chemotherapy trial. The FLCHAIN[24] dataset studies the relationship between mortality and serum immunoglobulin-free Light Chains, which are essential in hematological disorders. NWTCO[25] studies Wilms tumor in children, Molecular Taxonomy of Breast Cancer International Consortium (METABRIC)[26] explores breast cancer, PBC[27] focuses on Primary Biliary Cholangitis, STD deals with sexually transmitted diseases, and PNEUMON examines infant pneumonia.

Table 1 offers a more comprehensive view of the temporal aspects and occurrences of events within the various datasets considered. It becomes evident that a deliberate selection of diverse disease datasets has been made, each characterized by distinct types and quantities of information. This diversity in the disease datasets showcases the model's versatility. Significantly, the evaluation of the model has been carried out systematically in datasets that show varying proportions of censored samples and differing time-to-event ranges. This strategic approach aims to provide a broader perspective on how the model might perform when applied to other real-world datasets.

Performance metrics

Recalling from the Survival Analysis Section, each dataset is described by $D = (x_i, t_i, d_i)_{i=1}^{N}$ triplets, where $x_i = x_i^1, \dots, x_i^L$ is an L-dimensional vector of covariates, t_i is the time to event and $d_i \in \{0, 1\}$ is the censoring indicator.

When evaluating an SA model, the literature shows that the most commonly used metric is the C-index, which is the generalization of the ROC curve for all data. It measures the rank correlation between predicted risk and observed times. The concept arises from the intuition that a higher risk of an event occurring has a complete relation with a short time to the event. Therefore, a high number of correlating pairs, i.e., pairs of samples that meet this expectation, is decisive to say that the model has good predictive quality.

In this case, the time-dependent C-index[28] will be used since the original one[29] cannot reflect the possible changes in risk over time being only computed at the initial observation time. This C-index is defined as follows:

$$C_{index} = P\Big(\hat{F}(t|x_i) > \hat{F}(t|x_j)|d_i = 1, t_i < t_j, t_i \leq t\Big), \tag{25}$$

where $\hat{F}(t|x_i)$ is the CDF estimated by the model at the time t given a set of covariates x_i. The probability is estimated by comparing the relative risks pairwise, as already mentioned.

Dataset	# Samples	# Censored	# Covariates	Event time (mean, (min - max))	Censoring time (mean, (min - max))
WHAS	1638	948 (57.88%)	5	1045.42 (1 - 1999) days	1298.92 (371 - 1999) days
SUPPORT	9104	2904 (31.89%)	14	478.45 (3 - 2029) days	1060.22 (344 - 2029) days
GBSG	1546	965 (43.23%)	7	44.49 (0.26 - 87.36) months	65.15 (0.26 - 87.36) months
FLCHAIN	6524	4562 (69.92%)	8	3647.5 (0 - 5166) days	4296.74 (1 - 5166) days
NWTCO	4028	3457 (85.82%)	6	2276.68 (4 - 6209) days	2588.23 (4 - 6209) days
METABRIC	1980	854 (56.18%)	21	2944.81 (3 - 9193) days	3424.81 (21 - 9193) days
PBC	418	257 (61.48%)	17	63.93 (1.37 - 159.8) months	75.22 (17.77 - 159.83) months
STD	877	530 (60.43%)	21	369 (1 - 1519) days	420 (1 - 1519) days
PNEUMON	3470	3397 (97.9%)	13	9.84 (0.5 - 12) months	9.98 (0.5 - 12) months

Table 1. Data information from datasets used to train SAVAE model. We have analyzed nine different disease datasets with different proportions of samples, censored data, and varying survival times. Additionally, each contains different patient information, be it genomic, clinical, or demographic data.

Based on the prediction index proposed[30,31], the second evaluation metric that has been used in this analysis: Brier Score (BS). It is essentially a square prediction error based on the Inverse Probability of Censoring Weighting (IPCW)[32], a technique designed to recreate an unbiased scenario compensating for censored samples by giving more weight to samples with similar features that are not censored. So, given a time *t* the BS can be calculated as follows, with $G(\cdot)$ being the survival function corresponding to censoring (1/*G*(*t*) is the IPCW):

$$BS(t) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{(S(t|x_i))^2}{G(t_i)}\cdot \mathbb{I}(t_i < t, d_i = 1) + \frac{(1-S(t|x_i))^2}{G(t)}\cdot \mathbb{I}(t_i \geq t)\right]. \tag{26}$$

Since the C-index does not take into account the actual values of the predicted risk scores, BS can be used to assess calibration, i.e., if a model predicts a 10% risk of experiencing an event at time *t*, the observed frequency in the data should match this percentage for a well-calibrated model. On the other hand, it is also a measure of discrimination: whether a model can predict risk scores that allow us to determine the order of events correctly.

In this case, the evaluation is made using the integral form of BS since it does not depend on the selection of a specific time *t*:

$$IBS(t_{max}) = \frac{1}{t_{max}}\int_0^{t_{max}} BS(t)dt. \tag{27}$$

To statistically assess each model's performance based on the global C-index, we propose the Mean Reciprocal Rank (MRR) as the third metric. It measures the effectiveness of a prediction by considering the rank of the first relevant C-index within a list composed of the C-indices obtained from each model. Formally, the Reciprocal Rank (RR) for a set of results for each model is the inverse of the position of the first pertinent result. For example, if the first relevant result is in position 1, its RR is 1; if it is in position 2, the RR is 0.5; if it is in position 3, the RR is approximately 0.33, and so on. Thus, the MRR is the average of the RRs for a set of models:

$$MRR = \frac{1}{Q}\sum_{i=1}^{Q}\frac{1}{rank_i}, \tag{28}$$

where *Q* is the total number of models being compared, and $rank_i$ is the position of the first relevant C-index for the $i - th$ model. Higher MRR values indicate that relevant results appear higher in the list.

Additionally, to add more statistical information on the performance of the models, we performed hypothesis testing to compare the mean C-index and IBS values of our model with those of the state-of-the-art models in multiple folds since we are using a five-fold cross-validation method. Specifically, we formulated a null hypothesis that assumes that the mean performance metrics of the state-of-the-art models are more significant than our model's mean performance metrics. To assess the validity of this null hypothesis, we used *p*-values as a statistical measure. We established a significance threshold of 0.05, a common practice in hypothesis testing. When the obtained *p*-value for each case fell below this threshold, we rejected the null hypothesis. In practical terms, this indicated that our model exhibited superior performance compared to the other models. On the contrary, if the *p*-value exceeded 0.05, we concluded that there were no statistically significant differences between our model and the others. It is important to note that this approach considered variations in results across different folds, providing a more comprehensive assessment of model performance beyond just the average results. Given the multiple hypothesis tests performed, we acknowledge that the Family-Wise Error Rate (FWER)[33] increases as the number of tests grows, as established in the literature[34,35]. This increase in FWER raises the risk of Type I errors, where false positives may occur due to the accumulation of multiple tests. To address this issue, we have applied an appropriate method to control this inflation, the Holm adjustment[36]. The results of these adjustments can be seen in the Supplementary Information, ensuring the robustness of our findings.

Finally, we performed a sensitivity analysis to assess the robustness of our model and to understand how variations in the input data influence its predictions. This analysis provides insights into the impact of individual features on the model's performance and contributes to a better understanding of the model's decision-making process. Furthermore, we analyzed the computational complexity of our model by comparing its runtime with state-of-the-art models. This analysis considers the time required for training and validating SAVAE across multiple datasets, providing insights into its efficiency relative to other methods. The findings illustrate that while SAVAE's computational demands are higher due to its complex architecture, they remain manageable, making it suitable for practical applications. The detailed results of these analyses can be found in the Supplementary Information.

Experimental setting

The implementation of SAVAE was executed using the PyTorch framework[37]. As defined in Section ELBO derivation , three different DNNs were trained, consisting of one encoder and two decoders. These decoders were designed to infer covariates and time parameters, respectively. The Gaussian encoder exhibits a straightforward architecture characterized by a single hidden linear layer featuring a Rectified Linear Unit (ReLU) activation function and an output linear layer with hyperbolic tangent activation. The input to this encoder consists of the covariate vectors from the training dataset, while the output generates a Gaussian latent space. The dimensionality of this latent space has been fixed to 5. The generated latent space is input for both decoders, each featuring two linear layers. The first layer employs a ReLU activation function and incorporates a dropout rate of 20%. However, the final layer of the decoders employs different activation functions based on the specified distribution, thereby tailoring the output to the parameters of the respective covariate distribution. Furthermore, the number of neurons in each hidden layer was also fixed at 50. The training process involved 3,000 epochs with a batch size of 64 samples while incorporating an Early Stop mechanism in case of an insufficient reduction in validation loss.

To better understand the behavior of the SAVAE model and to justify the selection of the defined hyperparameters, we conducted an ablation study. This study analyzed how changes in key hyperparameters, such as latent space dimensionality, number of neurons, and dropout rates, affect model performance. We could identify settings that balance performance and computational efficiency by systematically varying these parameters. The results of this ablation study are provided in the Supplementary Information, offering further insights into the rationale behind our chosen hyperparameter configuration.

We used a five-fold cross-validation technique to evaluate the results while ensuring their robustness against data partitioning. This method was applied to our model and the state-of-the-art models used for performance comparison and result evaluation, including Cox-PH, DeepHit, and DeepSurv. Moreover, due to the inherent sensitivity of VAE architectures to initial conditions, we conducted training using up to 10 different random seeds. Subsequently, the C-index was averaged among the three best-performing seeds. The average performance of the three seeds provides a representative and sufficient evaluation. Lastly, note that the three state-of-the-art models have been implemented using the Pycox package[38] and the different metrics used for validation, C-index, and IBS. The MRR has been calculated manually, while the *p*-value has been obtained using the SciPy[39] package.

Experiments and results

In this section, we present a comprehensive assessment of the performance of our proposed model, SAVAE, compared to three well-established state-of-the-art models. Cox-PH, DeepSurv, and DeepHit. Across multiple datasets encompassing a diverse range of medical and clinical scenarios, we conducted extensive experiments to assess the performance of these models. The key focus was evaluating their ability to predict survival outcomes, considering censored and uncensored data points.

As the initial set of results, we focus on comparing the performance and results in terms of the C-index. Table 2 provides a comprehensive view of how our model is completely comparable to the state-of-the-art models regarding the average C-index. Additionally, note that all intervals for the minimum and maximum values across various folds overlap, indicating consistent performance across different data subsets. The results displayed in the table reveal that our model consistently achieves a higher MRR compared to others across multiple datasets, showcasing its superiority in many cases regarding the average C-index. However, it is essential to acknowledge that the C-index results among the different models are generally similar, highlighting the competitiveness of our model within the field. Furthermore, it is important to note that the broad intervals are primarily attributed to the limited sample sizes commonly found in medical databases, a characteristic that poses challenges when assessing model performance. To address this issue, we employed cross-validation, as previously mentioned, ensuring that our model's performance is robust and reliable. In summary, while our model demonstrates its strength by outperforming other models in terms of MRR and achieving competitive average C-index scores, the overall similarity in C-index results underscores its robustness and suitability for various medical datasets.

In our validation process, we performed a statistical analysis using *p*-values to determine whether our model exhibited superior performance in terms of the C-index. To carry out this analysis, we compared the average C-index of our model with the mean C-index values obtained from multiple folds for each state-of-the-art model. The objective was to determine whether the performance of our model was statistically better than the alternative models. We established a significance threshold of 0.05, a common practice in hypothesis testing. Our findings in Table 3 reveal several instances in which our model outperformed the state-of-the-art models, as evidenced by *p*-values below the 0.05 threshold. These results highlight the effectiveness and competitiveness

	COXPH		DEEPSURV		DEEPHIT		SAVAE	
Dataset	Avg. C-index	(Min, max)	Avg. C-index	(Min, max)	Avg. C-index	(Min, max)	Avg. C-index	(Min, max)
WHAS	0.74	(0.66, 0.81)	0.78	(0.57, 0.88)	**0.89**	(0.82, 0.95)	0.74	(0.67, 0.80)
SUPPORT	0.58	(0.39, 0.78)	0.57	(0.37, 0.82)	0.55	(0.37, 0.73)	**0.61**	(0.40, 0.86)
GBSG	0.66	(0.61, 0.71)	**0.67**	(0.58, 0.73)	0.66	(0.58, 0.72)	**0.67**	(0.62, 0.72)
FLCHAIN	0.69	(0.50, 0.80)	0.67	(0.55, 0.80)	0.78	(0.73, 0.82)	**0.79**	(0.75, 0.83)
NWTCO	0.71	(0.64, 0.79)	0.70	(0.60, 0.79)	**0.72**	(0.66, 0.78)	0.71	(0.63, 0.79)
METABRIC	0.59	(0.52, 0.68)	**0.61**	(0.52, 0.69)	0.56	(0.46, 0.64)	**0.61**	(0.53, 0.70)
PBC	**0.81**	(0.64, 0.94)	0.80	(0.65, 0.92)	0.80	(0.62, 0.93)	**0.81**	(0.62, 0.95)
STD	**0.60**	(0.47, 0.72)	**0.60**	(0.49, 0.71)	0.59	(0.50, 0.68)	0.59	(0.46, 0.71)
PNEUMON	0.62	(0.54, 0.70)	0.65	(0.49, 0.80)	**0.67**	(0.57, 0.77)	0.65	(0.53, 0.77)
MRR	0.56		0.60		0.62		**0.76**	

Table 2. C-index average results across different folds for each state-of-the-art model. Average C-index results across the three best seeds for each fold in SAVAE performance. MRR values are given to rank each model attending only to the mean value. Bold highlights the best mean. For C-index and MRR, higher is better

Model	WHAS	SUPPORT	GBSG	FLCHAIN	NWTCO	METABRIC	PBC	STD	PNEUMON
COXPH	0.579	0.058	**0.000**	**0.000**	0.268	**0.003**	0.450	0.887	**0.003**
DEEPSURV	1.0	**0.020**	0.149	**0.000**	0.135	0.549	0.280	0.927	0.382
DEEPHIT	1.0	**0.000**	**0.000**	**0.001**	0.644	**0.000**	0.228	0.727	0.935

Table 3. *p*-values obtained to determine whether the mean of SAVAE is greater than the state-of-the-art folds C-indexes. Bold Implies a *p*-value below our threshold, 0.05. This means that SAVAE is significantly better than the other models

	COXPH		DEEPSURV		DEEPHIT		SAVAE	
Dataset	Avg. IBS	(Min, max)	Avg. IBS	(Min, max)	Avg. IBS	(Min, max)	Avg. IBS	(Min, max)
WHAS	0.171	(0.109, 0.279)	0.134	(0.067, 0.260)	**0.120**	(0.067, 0.175)	0.159	(0.114, 0.205)
SUPPORT	0.208	(0.074, 0.374)	**0.205**	(0.057, 0.363)	0.219	(0.086, 0.370)	0.208	(0.063, 0.385)
GBSG	0.182	(0.142, 0.223)	0.179	(0.137, 0.228)	0.208	(0.168, 0.248)	**0.179**	(0.139, 0.222)
FLCHAIN	0.137	(0.089, 0.185)	0.142	(0.088, 0.186)	0.121	(0.098, 0.145)	**0.102**	(0.078, 0.124)
NWTCO	**0.107**	(0.080, 0.138)	0.109	(0.082, 0.149)	0.111	(0.083, 0.147)	0.127	(0.101, 0.152)
METABRIC	0.186	(0.137, 0.233)	0.191	(0.143, 0.244)	0.214	(0.153, 0.275)	**0.180**	(0.127, 0.236)
PBC	0.147	(0.043, 0.281)	0.146	(0.046, 0.268)	0.195	(0.087, 0.340)	**0.138**	(0.034, 0.267)
STD	0.210	(0.121, 0.302)	0.212	(0.123, 0.305)	0.224	(0.142, 0.315)	**0.209**	(0.121, 0.307)
PNEUMON	**0.016**	(0.004, 0.031)	0.017	(0.004, 0.034)	**0.016**	(0.004, 0.031)	0.021	(0.007, 0.037)
MRR	0.55		0.55		0.47		**0.71**	

Table 4. IBS average results across different folds for each state-of-the-art model. Average IBS results in the three best seeds for each fold in SAVAE performance. MRR values are given to rank each model. Bold highlights the best mean. For IBS lower is better and for MRR, higher is better

Model	WHAS	SUPPORT	GBSG	FLCHAIN	NWTCO	METABRIC	PBC	STD	PNEUMON
COXPH	1.000	0.470	0.998	1.000	**0.000**	0.995	0.888	0.575	**0.000**
DEEPSURV	**0.000**	0.341	0.561	1.000	**0.000**	1.000	0.868	0.746	**0.000**
DEEPHIT	**0.000**	0.950	1.000	1.000	**0.000**	1.000	1.000	0.995	**0.000**

Table 5. *p*-values obtained to determine whether the mean of SAVAE is greater than the state-of-the-art folds IBS values. Bold Implies a *p*-value below our threshold, 0.05. This means that SAVAE is significantly better than the other models

of our proposed approach. This comprehensive analysis, which considers the diverse C-index values in multiple folds, provides a robust evaluation of the model's performance, extending beyond simple average comparisons.

Our validation through IBS values (Tables 4 and 5) yielded conclusions that closely parallel those derived from the C-index analysis. Overall, it is essential to note that our model's IBS results align closely with the state-of-the-art models, demonstrating comparable performance. However, our proposed model consistently demonstrated competitiveness and emerged as the top performer in the various datasets used in our study. This convergence of results across different evaluation metrics reinforces the robustness and effectiveness of our novel approach. While our model maintains a competitive edge within the context of the state-of-the-art models, further solidifying its potential and utility in the field of SA, it also stands out as a top-performing solution.

It is essential to recall that, like DeepSurv and Cox-PH, SAVAE is a parametric model. However, unlike these models, we do not limit ourselves to the exponential distribution to model survival time. Our approach allows for the use of any differentiable distribution. Unlike DeepHit, which trains the model using loss functions, our framework uses likelihood functions, providing considerable flexibility. We specifically assumed the Weibull distribution for these experiments, deriving the shape parameter α and the scale parameter λ for each patient, although any differentiable distribution could have been used. This ability enables us to extract vital statistical information for personalized patient treatments, offering a significant advantage in medical applications.

Conclusion

In this paper, we have successfully described an SA model (SAVAE), which stands out for its ability to avoid assumptions that can limit performance in real-world scenarios. It is a model based on VAEs in charge of estimating continuous or discrete survival times, first, modeling complex non-linear relations among covariates due to the use of highly expressive DNNs, and second, taking advantage of a combination of loss functions that capture the censoring inherent to survival data. Our model demonstrates efficiency compared to various state-of-the-art models, namely Cox-PH, DeepSurv, and DeepHit, because of its freedom from assumptions related to linearity and proportional hazards. In contrast to DeepHit, which directly learns the C-Index metric, we train using standard likelihood techniques. Note that this means that our approach is more flexible, as it allows using many different distributions to model the data, and the performance is competitive, as it performs well in C-Index and IBS, instilling confidence in its capabilities.

Furthermore, the adaptability of our model is a notable strength. While we have assumed specific distributions for both survival times and covariates in our experiments, SAVAE's versatility extends to accommodating any other parametric distribution, as long as their CDF and hazard function are differentiable, making it a scalable tool. Notably, our model can efficiently handle censoring to mitigate bias, introducing a novel improvement in results. However, it is essential to acknowledge that the model's reliance on specific parametric distributions could pose limitations. If the chosen distribution does not align well with the underlying data distribution, the model may perform suboptimally. This is a known challenge in parametric survival analysis models, and further research could explore more flexible non-parametric or semi-parametric approaches to address this limitation[40].

This work raises several attractive lines for the future. Since the parameters estimated by SAVAE are subject to statistical uncertainty, we propose as future work using Monte Carlo sampling from the latent space to derive confidence intervals for survival predictions, providing more robust patient-wise survival curves with associated margins of error. An additional advantage lies in our model's architecture, where time and covariates are reconstructed from latent space information. This feature opens opportunities for its utility to be expanded to various tasks that have been developed using VAEs, including clustering[41], imputation of missing data[42], and data augmentation[43] by the generation of synthetic patients. Thus, this tool has great potential and can be exploited in future work to have different functionalities even in the world of Federated Learning[44,45].

In summary, SAVAE emerges as a versatile and robust SA model, surpassing state-of-the-art methods while offering extensibility to a broader range of healthcare applications. It presents a compelling solution for healthcare professionals seeking enhanced performance and adaptability in SA tasks.

Data availability

All datasets used in this research are publicly available and can be found in the repository https://github.com/Patricia-A-Apellaniz/savae. Through the provided link, readers can also reproduce the results of the current study, ensuring transparency and facilitating further research in this area.

Received: 11 July 2024; Accepted: 10 October 2024
Published online: 19 October 2024

References

1. Jing, H. & Smola, A. J. Neural survival recommender. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*. 515–524 (2017).
2. Grob, C. M., Lerman, D. C., Langlinais, C. A. & Villante, N. K. Assessing and teaching job-related social skills to adults with autism spectrum disorder. *J. Appl. Behav. Anal.* **52**, 150–172 (2019).
3. Wang, R. et al. Estimation of global black carbon direct radiative forcing and its uncertainty constrained by observations. *J. Geophys. Res. Atmos.* **121**, 5948–5971 (2016).
4. Dellana, S. & West, D. Survival analysis of supply chain financial risk. *J. Risk Finance* **17**, 130–151 (2016).
5. Cox, D. R. Regression models and life-tables. *J. R. Stat. Soc. Ser. B (Methodological)* **34**, 187–202 (1972).
6. Kaplan, E. L. & Meier, P. Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.* **53**, 457–481 (1958).
7. Lee, E. T. & Wang, J. *Statistical Methods for Survival Data Analysis*. Vol. 476 (Wiley, 2003).
8. Ranganath, R., Tran, D., Altosaar, J. & Blei, D. Operator variational inference. *Adv. Neural Inf. Process. Syst.* **29** (2016).
9. Faraggi, D. & Simon, R. A neural network model for survival data. *Stat. Med.* **14**, 73–82 (1995).
10. Katzman, J. L. et al. Deepsurv: personalized treatment recommender system using a cox proportional hazards deep neural network. *BMC Med. Res. Methodol.* **18**, 1–12 (2018).
11. Luck, M., Sylvain, T., Cardinal, H., Lodi, A. & Bengio, Y. Deep learning for patient-specific kidney graft survival analysis. arXiv preprint arXiv:1705.10245 (2020).
12. Kraisangka, J. & Druzdzel, M. J. A bayesian network interpretation of the cox's proportional hazard model. *Int. J. Approx. Reas.* **103**, 195–211 (2018).
13. Vinzamuri, B. & Reddy, C. K. Cox regression with correlation based regularization for electronic health records. In *2013 IEEE 13th International Conference on Data Mining*. 757–766 (IEEE, 2013).
14. Vinzamuri, B., Li, Y. & Reddy, C. K. Active learning based survival regression for censored data. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. 241–250 (2014).
15. Lee, C., Zame, W., Yoon, J. & Van Der Schaar, M. Deephit: A deep learning approach to survival analysis with competing risks. In *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 32 (2018).
16. Kingma, D. P. & Welling, M. Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013).
17. Hinton, G. E. & Salakhutdinov, R. R. Reducing the dimensionality of data with neural networks. *Science* **313**, 504–507 (2006).
18. Ranganath, R., Perotte, A., Elhadad, N. & Blei, D. Deep survival analysis. In *Machine Learning for Healthcare Conference*. 101–114 (PMLR, 2016).
19. Liverani, S., Leigh, L., Hudson, I. L. & Byles, J. E. Clustering method for censored and collinear survival data. *Comput. Stat.* **36**, 35–60 (2021).
20. Hosmer Jr, D. W., Lemeshow, S. & May, S. *Applied Survival Analysis: Regression Modeling of Time-to-Event Data*. Vol. 618 (Wiley, 2008).
21. Knaus, W. A. et al. The support prognostic model: Objective estimates of survival for seriously ill hospitalized adults. *Ann. Intern. Med.* **122**, 191–203 (1995).
22. Foekens, J. A. et al. The urokinase system of plasminogen activation and prognosis in 2780 breast cancer patients. *Cancer Res.* **60**, 636–43 (2000).
23. Schumacher, M. et al. Randomized 2 x 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients. German breast cancer study group. *J. Clin. Oncol.* **12**, 2086–2093 (1994).
24. Dispenzieri, A. *et al.* Use of nonclonal serum immunoglobulin free light chains to predict overall survival in the general population. In *Mayo Clinic Proceedings*. Vol. 87. 517–523 (Elsevier, 2012).
25. Breslow, N. E. & Chatterjee, N. Design and analysis of two-phase studies with binary outcome applied to wilms tumour prognosis. *J. R. Stat. Soc. Ser. C (Appl. Stat.)* **48**, 457–468 (1999).

26. Pereira, B. *et al.* The somatic mutation profiles of 2,433 breast cancers refines their genomic and transcriptomic landscapes. *Nat. Commun.* **7**, 11479 (202).
27. Therneau, T. M. Extending the cox model. In *Proceedings of the First Seattle Symposium in Biostatistics: Survival Analysis*. 51–84 (Springer, 1997).
28. Antolini, L., Boracchi, P. & Biganzoli, E. A time-dependent discrimination index for survival data. *Stat. Med.* **24**, 3927–3944 (2005).
29. Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L. & Rosati, R. A. Evaluating the yield of medical tests. *Jama* **247**, 2543–2546 (1982).
30. Brier, G. W. Verification of forecasts expressed in terms of probability. *Mon. Weather Rev.* **78**, 1–3 (1950).
31. Graf, E., Schmoor, C., Sauerbrei, W. & Schumacher, M. Assessment and comparison of prognostic classification schemes for survival data. *Stat. Med.* **18**, 2529–2545 (1999).
32. Robins, J. M. *et al.* Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. In *Proceedings of the Biopharmaceutical Section*. Vol. 24 (American Statistical Association, 1993).
33. Tukey, J. W. The philosophy of multiple comparisons. *Stat. Sci.* 100–116 (1991).
34. Lehmann, E. L. & Romano, J. P. *Generalizations of the Familywise Error Rate* (Springer, 2012).
35. Van der Laan, M. J., Dudoit, S. & Pollard, K. S. Multiple testing. Part II. Step-down procedures for control of the family-wise error rate. In *Statistical Applications in Genetics and Molecular Biology*. Vol. 3 (2004).
36. Holm, S. A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* 65–70 (1979).
37. Paszke, A. *et al.* Pytorch: An imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32** (2019).
38. Kvamme, H., Borgan, Ø. & Scheel, I. Time-to-event prediction with neural networks and cox regression. *J. Mach. Learn. Res* **20**, 1–30 (2019).
39. Virtanen, P. et al. Scipy 1.0. fundamental algorithms for scientific computing in python. *Nat. Methods* **17**, 261–272 (2020).
40. Nelson, W. B. *Applied Life Data Analysis* (Wiley, 2005).
41. Lim, K.-L., Jiang, X. & Yi, C. Deep clustering with variational autoencoder. *IEEE Signal Process. Lett.* **27**, 231–235 (2020).
42. McCoy, J. T., Kroon, S. & Auret, L. Variational autoencoders for missing data imputation with application to a simulated milling circuit. *IFAC-PapersOnLine* **51**, 141–146 (2018).
43. Chadebec, C. & Allassonniere, S. Data augmentation with variational autoencoders and manifold sampling. arxiv:2103.13751 (2021).
44. Gu, Z. et al. Frepd: A robust federated learning framework on variational autoencoder. *Comput. Syst. Sci. Eng.* **39**, 307–320 (2021).
45. Polato, M. Federated variational autoencoder for collaborative filtering. In *2021 International Joint Conference on Neural Networks (IJCNN)*. 1–8 (IEEE, 2021).

Acknowledgements

This work was supported by the Genomics and Personalized Medicine for all though Artificial Intelligence in Haematological Diseases (GenoMed4All) and Synthetic generation of hematological data over federated computing frameworks (SYNTHEMA) Projects from European Union's Horizon 2020 Research and Innovation Program under Grant 101017549 and Grant 101095530. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the granting authority can be held responsible for them. The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Author contributions

P.A.A, J.P. and S.Z. conceived the study and designed the general architecture. P.A.A. and J.P. developed the code, conducted the experiments, and analyzed the results. P.A.A. wrote the original manuscript. All authors reviewed the manuscript.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-024-76047-z.

Correspondence and requests for materials should be addressed to P.A.A.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.