# Combining Cox Model and Tree-Based Algorithms to Boost Performance and Preserve Interpretability for Health Outcomes

Diana Shamsutdinova[1](✉) , Daniel Stamate[2,3] , Angus Roberts[1] , and Daniel Stahl[1]

[1] Biostatistics and Health Informatics Department, Institute of Psychiatry Psychology and Neuroscience, King's College London, London, UK
`diana.shamsutdinova@kcl.ac.uk`
[2] Data Science and Soft Computing Lab, Computing Department, Goldsmiths University of London, London, UK
[3] Division of Population Health, Health Services Research and Primary Care, School of Health Sciences, University of Manchester, Manchester, UK

**Abstract.** Predicting health outcomes such as a disease onset, recovery or mortality is an important part of medical research. Classical methods of survival analysis such as Cox proportionate hazards model have successfully been employed and proved robust and easy to interpret. Recent development of computational methods and digitalization of medical records brought new tools to survival analysis, which can handle large data with complex non-linear relationships. However, such methods often result in "black box" models hard to interpret. In this project we combine the Cox model with tree-based machine-learning algorithms to take advantage of both approaches' strength and to boost the overall predictive performance. Moreover, we aimed to preserve interpretability of the results, quantify the contribution of linear and non-linear and cross-term dependencies, and get insight into a potential non-linearity. The first method includes the Cox model, ensembled with the survival random forest. The second employs a survival tree algorithm to cluster the data, and then fits a separate Cox model in each cluster. The third uses the clusters obtained with a survival tree to identify interaction and non-linear terms and adds them as new terms to the Cox model. We tested the methods on simulated and real-life medical data and compared their internally validated discrimination and calibration. Our results show that classical models outperform combined methods in data with predominantly linear relationships. The proposed methods were more effective in predicting survival outcomes with strong non-linear and inter-dependent relationships and provided an insight into where the non-linearity is placed.

**Keywords:** Survival analysis · Health research · Cox model · Survival random forest · Machine learning · Ensemble methods

# 1   Introduction

Survival analysis is one of the main methods in health research for longitudinal data. The outcome of interest can be disease onset, recovery, hospital re-admission, mortality, and others. The aims could be either explanatory or prognostic. In explanatory analysis a researcher investigates relationships of various risk factors and incidence rate of an event of interest, their statistical significance and impact size, while an overall model fit, or total variance explained may not be of primary interest. Prognostic research focuses on accurate predictions of the future incidence rate, for which the model fit is crucial, while understanding the relationship between the risk factors and the outcome can be less important. For this task machine learning (ML) algorithms have proven effective and provide alternatives to classical statistical methods. ML methods are flexible and easily adapt to data with complex dependencies. However, flexibility is often achieved by optimizing many model parameters, and the resulting logic can be difficult to interpret. For health outcomes, however, interpretability is one of the key factors: transparency can greatly facilitate model implementation into clinical practice, where clinicians and patients should sufficiently trust predictions to act on them. At the same time, black-box decisions may be avoided for legal and compliance reasons.

This project tries to merge the two approaches and intertwine linear and ML models. We aim to benefit from the interpretability of the linear model while enhancing its predictive performance with the tree-based features. Our baseline linear model is the Cox proportional hazards model (CoxPH), the machine learning algorithms are survival trees [1] and survival random forest (SRF) [2]. CoxPH is a robust model whose regression coefficients represent a multiplicative impact of a risk factor on the baseline log-hazard function, where hazard is an instantaneous event risk [3]. Proportionality assumption can be viewed as another aid for interpretability, as it ignores a potential time-dependence of the risk factor, so the coefficients estimate an integrated impact over the observation time. Nonetheless, direct input of predictors only accounts for the linear effects, while introducing non-linear and interaction terms requires adding such terms explicitly to the equation. In contrast, tree-based algorithms have a built-in ability to capture the non-linearity and cross-dependence of the predictors. However, a final prediction function that relates risk factors and the outcome may considerably with minor data alterations. Therefore, we aim to test if certain combined methods can outperform the baseline algorithms. Our first method includes the Cox model, ensembled with the survival random forest. The second employs a survival tree algorithm to cluster the data and then fits a separate Cox model in each cluster. The third uses the clusters obtained with a survival tree to identify interaction and non-linear terms and adds them to the Cox model. We test the methods on simulated and real-life medical data and compare their discrimination and calibration performance using internally validated area under receiver-operating curve (AUC-ROC) and calibration slopes.

Previously, machine learning community has been proposing to combine various methods in a similar stepwise manner [4, 5]; other papers suggested using a decision tree to automatically cluster the data or select interaction terms [6, 7]. However, we did not find papers with a similar focus on taking advantage of different approaches to enhance interpretability and performance, or recommendations on how these models can be used for these purposes.

## 2 Methods

We tested the following three hybrid methods to predict event incidence over a specified time. Risk factors were assumed to be constant in time and measured at the baseline. First, we provide a brief description of the baseline Cox and tree-based models, then describe those are proposed to be combined.

### 2.1 Models

**Cox Model.** Using standard notations in survival analysis, where T is a random time-to-event, its survival function $S(t) = Prob(T > t)$. Hazard function is $h(t) = \frac{-S'(t)}{S(t)}$ and represents a current rate of failure for those event-free by t. In prediction modelling the aim is to estimate survival function (or a correspondent hazard) and its dependence on predictors. That is, to estimate a conditional distribution of T, $S(t \mid x)$, or $h(t \mid x)$, given a vector of predictors $x = (x^1, \ldots, x^K)$.

Cox proportionate hazard model [3] assumes that hazard rates of observations are proportionate, and the multiplier is exponentiated linear combination of the risk factors,

$$h(t \mid x) = h_0(t) \cdot \exp(\beta_1 \cdot x_1 + \ldots + \beta_K \cdot x_K) \tag{1}$$

Here, $h_0(t)$ is a non-parametric baseline hazard function. Coefficients are estimated by maximizing the partial log-likelihood function [3] and can be maximized without estimating the baseline hazard. Coefficients represent relative change in the hazard rate to a unit of the underlying factor, $h(t \mid x_1 = 1)/ h(t \mid x_1 = 0) = \exp(\beta_1)$. In the classical model, coefficients and their impact are time-invariant.

Baseline hazard function can be estimated in different ways [8, 9]. We will use the Kalbfleisch-Prentice estimator following recommendations by Xia and colleagues [10]. Kalbfleisch and Prentice introduced baseline conditional survival probabilities for each time between the failure times and expressed the likelihood function in terms of these probabilities and beta coefficients from the Cox regression. Maximizing such likelihood function by the baseline conditional probabilities, one can find the baseline function corresponding to the estimated Cox regression [9]. Finally, the individual risk of failure is estimated from the baseline survival and linear predictors, $LP = \sum_i \beta \cdot x_i$,

$$h(t \mid x) = h_0(t) \cdot \exp(L.P.), \ S(t \mid x) = S_0(t)^{\exp(LP)} \tag{2}$$

**Survival Decision Tree and Survival Random Forests.** Survival tree is a type of classification and regression tree algorithm [11], in which a sample is recursively partitioned by a condition in predictors space guided by a splitting rule. All conditions are comparisons of a predictor value with a threshold. The splitting rule aims to find more homogeneous sub-populations at the two subsequent nodes (or most heterogeneous between the nodes). Definition of homogeneity is not straightforward for survival data due to the presence of censoring and time dimension [12]. Many existing packages use splitting rules based on the log-rank test statistics measuring a statistical difference of the survival curves [2, 13]. LeBlanc and Crowley suggested using the local full likelihood function under the proportional hazards assumption [1]. Purity measures for censored

observations have also been proposed [14]. Shimokawa and colleagues [15] showed that different splitting rules could be preferred depending on the hazard function properties. Here we use Survival Random Forest [2] with the log-rank splitting rule from R package randomForestSRC [16] and a survival tree from the RPart package developed by Therneau and Atkinson [17]. RPart uses a rescaled Poisson process which makes the likelihood equivalent to that in the LeBlanc and Crowley's tree [17].

Any survival tree has final leaves, in which all observations are clustered depending on their risk factors and conditions at the nodes. Irrespective of the splitting rule, survival probabilities are estimated by non-parametric Kaplan-Meier curves $KM_L(t)$ in the final leaves, fitted to the observations in the training set falling into that leaf,

$$S(t \mid x) = S(t \mid \text{leaf } L : x \text{ is in } L) = KM_L(t). \tag{3}$$

**Survival Random Forest.** Survival random forest is an ensemble method that averages predictions over many survival trees grown on a bootstrapped version of the data [2].

**Ensemble Method 1 (Cox_SRF).** We develop the idea of employing the results of one model as an input to another. This approach was proposed by others [4, 5] and is similar to stacking in ML. First, we fit a standard CoxPH and compute linear predictors for the observations in the train set. Second, we add the linear predictors to the list of the risk factors and train a survival random forest. This extended SRF predicts the survival:

$$S(t \mid x) = S^{SRF}\Big(t \mid x' = (x^1, \ldots, x^K, \text{ Cox\_linear\_predictor}\Big) \tag{4}$$

In this method, we supplied an additional predictor to SRF that aggregated "linear information" captured by CoxPH. Hence, the difference in the SRF performance and baseline CoxPH quantifies the non-linear and interaction terms contribution, and CoxPH regression coefficients describe the linear impact of the predictors.

To minimize overfitting, we split the sample into two, train the Cox model separately on either half to predict survival probability for the other half. Out-of-sample predictions are used in the consequent SRF. The ensemble method 1b explores the same idea in the reverse order, fits SRF, and adds out-of-the-bag predictions to the CoxPH model (Fig. 1).

**Ensemble Method 2 (Tree_ClusterCox).** The second method uses a survival tree to partition the survival data into clusters. The tree depth is limited by 4, so the number clusters is not more than 16. We then fit a separate CoxPH model in each cluster. Predictions are made by first identifying into which leaf (cluster) an observation falls, then applying the corresponding CoxPH model:

$$S(t \mid x) = \sum_L I(x \in \text{leaf } L) \cdot S_0^L(t)^\wedge \exp\Big(\beta_1^L \cdot x_1 + \ldots + \beta_K^L \cdot x_K\Big) \tag{5}$$

where $\beta_i^L$ are coefficients of the CoxPH model fitted in the leaf L, $S_0^L$ is a corresponding baseline survival function. The method aims to enhance the Cox model in two ways. First, CoxPH assumes that there is a single baseline survival function for the entire population, and risk factors shift survival in log-hazard space. Aiming a better fit, we relax this assumption by allowing each cluster to have a separate baseline survival function.

Second, the survival tree may identify strong non-linearities or cross-dependencies in the top nodes, which could lead to an easier optimization of the linear models in the final leaves, and a better overall fit. The main challenge in this approach is the underlying tree instability in response to small data permutations. To address this problem, we select predictors with the highest variable importance index (VIMP) computed by a baseline SRF [18], then cross-validate to tune the number of predictors used in the tree.

The method results in a perfectly transparent model: one can display the tree from the 1st step and the CoxPH coefficients for each leaf. We test two versions of the method, 2A and 2B, using the packages RPart with LeBlanc and Crowley splitting rule, and randomForestSRC guided by the log-rank splitting (Fig. 1).

**Ensemble Method 3 (Tree_ModifiedCox).** We start by growing a tree to cluster the survival data as in Ensemble 2, using only the top VIMP variables. Then, we add cluster identifier as a categorical variable to the baseline Cox regression along with the original predictors. The underlying idea is that a survival tree can serve as a tool to identify non-linear and cross-terms, which we can add to CoxPH. For $c$ final leaves (clusters), $L_i(x) = I(x \in \text{leaf } i)$ as leaf identifiers, the final model is

$$S(t \,|\, x) = S_0^{Cox}(t)^{\wedge} \exp(\beta_1 \cdot x_1 + \ldots + \beta_K \cdot x_K + a_1 \cdot L_1(x) + \ldots + a_c \cdot L_c(x)), \quad (6)$$

where betas are the linear impact factors, and alphas are the impact factors of the cross-terms defined by the final tree leaves, that is a product of $I(x_i < k_i)$ or $I(x_j \geq k_j)$ terms. The method is very similar to the one developed by Su and Tsai [7]. The authors developed a tree-growing algorithm that considers the linear terms already in the Cox regression while searching for the new non-linear ones, however, the advantage of our method is its straightforward implementation employing the existing software (Fig. 1).

**Cox Model with Fractional Polynomial Regression (Cox_FP).** Additionally, we compare the results of the combined methods to the CoxPH with fractional polynomial terms (Cox-FP) [19] using R package mfp [20]. The function selects statistically significant fractional polynomial terms that enhance the model fit once added to the CoxPH regression. We used a default version of the mfp function, in which a prediction variable $x$ can be transformed into a linear combination

$$\xi_0 + \xi_1 x^{(P_1)} + \xi_2 x^{(P_2)}, \quad (7)$$

where $x^{(P)}$ is $x^p$ for $p \neq 0$, and $\ln(x)$ for $p = 0$; $p_1$, $p_2$ are from $\{-2, -1, -0.5, 0, 0.5, 1, 2\}$ [19, 20]. This method allows CoxPH to handle non-linearity (though not the cross-terms), so we included it as another baseline models.
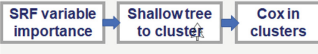
| Ensemble 1 | Ensemble 2 | Ensemble 3 |
|---|---|---|
| 1. **Fit Cox model on all data**<br>2. **Fit survival random forest with Cox survival probabilities as an additional factor** | 1. **VIMP SRF to find important predictors**<br>2. **Grow a single tree with these predictors, treat final nodes as clusters**<br>3. **Build Cox PH model in each cluster** | 1. **-Same as in Ensemble 2-**<br>2. **-Same as in Ensemble 2-**<br>3. **Build Cox model on all data with additional risk factors – tree clusters as non-linear and interaction terms** |
| **Rationale**: separate linear and non-linear steps<br><br>**Interpretability**: Cox coefficients represent linear terms. Difference in performance (Cox and Cox+SRF) quantifies non-linear effects<br><br>**Challenges**: individual impact factors not known | **Rationale**: CoxPH has single baseline survival, here baseline function by cluster<br><br>**Interpretability**: Tree can be displayed to see the clusters. Cox coefficients are available for each cluster<br><br>**Challenges**: overfitting, stability | **Rationale**: tree automatically identifies non-linear predictors<br><br>**Interpretability**: Tree can be displayed to see leaf definitions. Cox model shows hazard ratio of added predictors<br><br>**Challenges**: overfitting, stability |
| **Cox** → **Cox lp as risk factor** → **SRF** | **SRF variable importance** → **Shallow tree to cluster** → **Cox in clusters** | **VIMP from SRF** → **Shallow tree to cluster** → **Clusters to Cox** |

**Fig. 1.** Summary of the methods. Each method combines the Cox proportional hazards model with tree-based machine learning algorithms (survival tree or random forest).

## 2.2 Samples

**Simulated Samples.** We tested the models using four samples, three simulated and a health-related data. Simulated samples were described by the four predictors: $x_1$ is uniformly distributed between -1.73 and 1.73 (so the mean is 0 and standard deviation, SD, is 1.00), $x_2$ is sampled from a normal distribution (mean = 0, SD = 1), $x_3$ is a binomial random variable, positive outcome probability p = 0.2, $x_4$ - binomial with p = 0.5. Such variables can typically appear in health data. For example, $x_1$ can represent normalized age, $x_2$ – normalized body mass index, $x_3$ – the presence of hypertension, $x_4$ – gender, for the outcome of cardiovascular disease.

Further, we assume an exponential time-to-event, so the survival function is $S(t \mid x) = \exp(-h(x) \cdot t)$. The hazard rate $h(x)$ is constant in time and varies with risk factors.

We defined the hazard rate in the first sample such that it depends linearly on predictors:

$$h(x) = 0.10 \cdot \exp(0.4 \cdot x_1 + 1.0 \cdot x_2 + 0.7 \cdot x_3) \qquad (8)$$

From Eq. (1), the true Cox regression coefficients are $\beta_1 = 1$, $\beta_2 = 0.7$, $\beta_3 = 0.4$, $\beta_4 = 0$, baseline hazard is 0.10. We assumed that $x_4$ does not impact the outcome. In the second sample, we added non-linear dependencies:

$$h(x) = 0.08 \cdot \exp(0.2 \cdot x_1 + 1 \cdot I(x_1 \geq 1) + 1 \cdot I(1 < |x_2| \leq 1.5) + 2.0 \cdot I(|x_2| > 1.5) + 0.7 \cdot x_3) \quad (9)$$

So, one non-linearity is the jump in $x_1$'s impact after a certain threshold. This could describe risk acceleration after a certain age. Another non-linear relationship is the impact of $x_2$, which is null within one standard deviation from the mean, 1 for absolute values till 1.5, 2 otherwise. It may represent an effect when the weight in a normal range does not affect disease onset, while very low or high values increase the risk.

The third sample has the same non-linearity in $x_2$ as the second sample, a linear $x_1$ impact, but non-linearity in $x_1$ is replaced by an interaction between $x_1$ and $x_3$, in which a combination of high $x_1$ and positive $x_3$ constitutes an elevated disease risk.

$$h(x) = 0.07 \cdot \exp(0.2 \cdot x_1 + 1.0 \cdot I(x_1 \geq 1 \ \& \ x_3 = 1) + 1.0 \cdot I(1 < |x_2| \leq 1.5) + 2.0 \cdot I(|x_2| > 1.5) + 1.0 \cdot x_3) \tag{10}$$

Compared to the previous sample, not all individuals experience a jump in $x_1$ impact for $x_1 \geq 1$, but those with $x_3 = 1$. This could be a situation when acceleration in a disease risk after a certain age only affects those with pre-existing conditions (for example, hypertension).

Baseline hazards (0.10, 0.08, and 0.07) are set such that 50% of the population experienced the event by $t = 5$, which is the time point where the models' performance for the simulated samples was assessed. In the three simulated samples, coefficients for the Eqs. 8, 9, and 10 were meant to express large non-linear or interaction effects with no linkage to a certain health data. However, the equation forms were chosen to have plausible interpretations.

**ELSA Sample.** The fourth sample came from the English Longitudinal Study of Ageing (ELSA). We tested our models to predict the incidence of type two diabetes over 7.5 years of observation. ELSA is an ongoing multidisciplinary study with a core cohort of 11391 individuals aged $\geq$ 50, representative of the older U.K. population [21]. The participants are interviewed every two years since wave 2 (2002/3), and medical examinations occur every four years from wave 2. We included participants with available blood tests, genetic information, and diabetes status for at least one wave after the baseline. Diabetes was established by a self-reported medical diagnosis of diabetes, or HbA1C $\geq$ 48 mmol/mol (6.5%). The analytical sample had 5957 participants, mean observation time 8.9 years, mean time before diabetes incidence 4.9 years; 398(7%) developed diabetes before 7.5y. Risk factors were age, gender, body mass index, hypertension history, accumulated wealth (low/medium/high), level of education, exercise regime, smoking, depression, and blood cholesterol. As type 2 diabetes has a considerable genetic component [22], we included a polygenic risk score for type 2 diabetes into the analysis, which sums common genetic variants associated with the disease weighted by their impact size [23].

### 2.3  Performance Assessment

*Performance Measures.* The performance of a survival model varies in time, so we compare the models for the task of predicting an outcome by a pre-defined time. Examples in health research could be predicting 1-year mortality after a surgery or risk of heart failure in the next 5 years. We assessed model performance in three domains: discrimination, calibration, and interpretability. Discrimination is an ability to separate high-risk and low-risk individuals, which we measure with AUC-ROC. Censored observations should be accounted for while computing AUC-ROC, and we use timeROC function developed for survival data [24]. Calibration is how well an estimated event probability corresponds to the observed share of individuals with similar risk factors experiencing the event. It is measured by the calibration slope and intercept. Intercept, or calibration-in-the-large,

is the difference between the mean observed and predicted rates, so the ideal value is 0. Calibration slope is the correspondence of the predicted values to the observed across the probability scale. The ideal value is 1, a lower number means predictions are too extreme (too low for low-risk and too high for high-risk observations), and a sign of over-fitting; a slope $>1$ may indicate underfitting. To qualitatively assess interpretability, we question whether predictions are easy to explain and if they give an insight into the underlying relationships.

*Internal Validation.* We used 5-fold cross-validation to assess model performance. Further, an internal loop of the 3-fold cross-validation was used to tune model parameters: tree depth and a number of risk factors for a node split for SRF; minimum node size, maximum tree depth (2, 3 or 4), and the number of risk factors for tree clustering tree in Ensemble 2 and 3 (the factors were sorted by variable importance and method tuned for using the top 3, …, or 10). Combinations with the highest AUC-ROC at the selected time defined the final models.

## 3   Results

Tables 1, 2, 3 and 4 contain internally validated performance statistics of the described methods. The first table is for the sample with linear dependencies, the second for the sample with non-linear terms, the third with interaction terms, and the fourth is for the diabetes data from the ELSA study. Baseline methods were Cox model, Cox model with fractional polynomials, survival random forest, ensemble methods 1A (training Cox regression, then using probabilities to SRF), 1B (training SRF, then using probabilities in Cox), 2A and 2B (building a shallow tree for data clustering, then fitting Cox models in each cluster) and 3 (with clusters as additional parameters in the Cox model).

All methods performed well in the linear sample with similar performance metrics (Table 1). AUCs were between 0.809 (SD 0.011) and 0.816 (SD 0.011).

For the non-linear sample, AUC was considerably better in the ensemble methods compared to the baseline Cox regression results (Table 2). AUC-ROC was 0.715 (SD 0.023) for baseline Cox, and 0.769 (SD 0.010) for Ensemble 3. Worth noting that the Cox model with fractional polynomials performed as well as the ensembled methods, AUC 0.757 (SD 0.013).

The outperformance of the ensembled methods was even higher when cross-terms were present (Table 3). Baseline Cox had AUC-ROC 0.648 (SD 0.027), adding fractional polynomials got it to 0.694 (SD 0.022), Ensemble 1A reached AUC 0.715 (SD 0.017), Ensemble 2A 0.719 (SD 0.016), Ensemble 3 0.721 (SD 0.016).

However, in the real-life health data from the ELSA study, all the methods displayed similar results (Table 4). Figure 2 displays the RPart tree from methods 2 and 3, which captured non-linear dependencies.

**Table 1.** Internally validated performance of the methods in the first sample

| Linear sample | Cox | SRF | CoxMFP | 1A | 1B | 2A | 2B | 3 |
|---|---|---|---|---|---|---|---|---|
| AUC-ROC at t = 5 | 0.815 | 0.815 | 0.809 | 0.811 | 0.816 | 0.814 | 0.813 | 0.815 |
| AUC-ROC std | 0.011 | 0.014 | 0.011 | 0.012 | 0.011 | 0.012 | 0.002 | 0.011 |
| AUC diff to Cox | **0.000** | **0.000** | **−0.007** | **−0.005** | **0.000** | **−0.001** | **−0.002** | **0.000** |
| Slope | 0.991 | 0.991 | 1.007 | 0.989 | 0.994 | 0.990 | 0.979 | 0.990 |
| Alpha | −0.012 | −0.013 | 0.008 | 0.000 | −0.014 | 0.001 | 0.004 | 0.007 |

**Table 2.** Internally validated performance of the methods in the second sample

| Non Linear | Cox | SRF | CoxMFP | 1A | 1B | 2A | 2B | 3 |
|---|---|---|---|---|---|---|---|---|
| AUC-ROC | 0.715 | 0.773 | 0.757 | 0.767 | 0.758 | 0.762 | 0.755 | 0.769 |
| AUC-ROC std | 0.023 | 0.006 | 0.013 | 0.009 | 0.015 | 0.013 | 0.012 | 0.010 |
| AUC diff to Cox | **0.000** | **0.059** | **0.042** | **0.053** | **0.044** | **0.048** | **0.040** | **0.054** |
| Slope | 1.051 | 1.017 | 1.119 | 0.956 | 0.998 | 0.943 | 0.956 | 0.948 |
| Alpha | −0.024 | 0.000 | −0.006 | −0.002 | −0.003 | 0.007 | 0.006 | 0.006 |

**Table 3.** Internally validated performance of the methods in the third sample

| Cross-terms | Cox | SRF | CoxMFP | 1A | 1B | 2A | 2B | 3 |
|---|---|---|---|---|---|---|---|---|
| AUC-ROC at t = 5 | 0.648 | 0.715 | 0.694 | 0.715 | 0.705 | 0.719 | 0.720 | 0.721 |
| AUC-ROC std | 0.027 | 0.014 | 0.022 | 0.017 | 0.017 | 0.016 | 0.008 | 0.016 |
| AUC diff to Cox | **0.000** | **0.067** | **0.047** | **0.067** | **0.058** | **0.071** | **0.072** | **0.073** |
| Slope | 0.958 | 0.964 | 1.027 | 0.972 | 0.942 | 0.969 | 0.972 | 0.995 |
| Alpha | −0.012 | 0.018 | −0.005 | 0.018 | −0.014 | 0.002 | 0.003 | 0.001 |

**Table 4.** Internally validated performance of the methods in the fourth sample

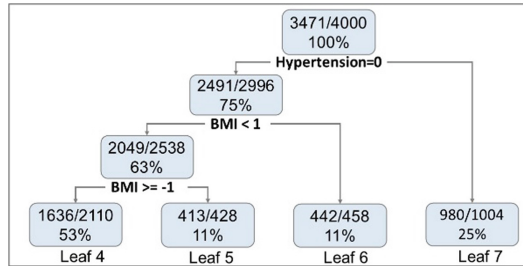| Cross-terms | Cox | SRF | CoxMFP | 1A | 1B | 2A | 2B | 3 |
|---|---|---|---|---|---|---|---|---|
| AUC-ROC at t = 5 | 0.750 | 0.753 | 0.753 | 0.758 | 0.750 | 0.731 | 0.722 | 0.752 |
| AUC-ROC std | 0.011 | 0.014 | 0.011 | 0.012 | 0.011 | 0.021 | 0.023 | 0.011 |
| AUC diff to Cox | **0.000** | **0.003** | **0.003** | **0.008** | **−0.001** | **−0.019** | **−0.028** | **0.001** |
| Slope | 0.977 | 0.948 | 1.229 | 1.087 | 0.999 | 0.755 | 0.758 | 0.850 |
| Alpha | 0.101 | 0.098 | 0.145 | 0.149 | 0.182 | 0.156 | 0.239 | 0.158 |

**Fig. 2.** Example of the RPart survival tree used as the first step for the Ensemble methods 2 and 3 for the simulated sample 3 with non-linear and interaction terms.

|  | coef | exp(coef) | se(coef) | Pr(>|z|) |  |
|---|---|---|---|---|---|
| age | 0.2851 | 1.3298 | 0.0182 | <0.001 | *** |
| BMI | -0.0046 | 0.9964 | 0.0278 | 0.867 |  |
| hypertension | 1.4731 | 4.3626 | 0.0439 | <0.001 | *** |
| gender | 0.0212 | 1.0214 | 0.0340 | 0.534 |  |
| as.factor(Leaf)5 | 0.8663 | 2.3781 | 0.0696 | <0.001 | *** |
| as.factor(Leaf)6 | 0.9127 | 2.4910 | 0.0692 | <0.001 | *** |
| as.factor(Leaf)7 | na | na | 0.0000 | na |  |

**Fig. 3.** Example of the modified Cox regression model which uses clusters of a survival tree as additional risk factors for simulated sample 3. The clusters are the final leaves of the tree in Fig. 2 (number 4–7, left to right): Cluster 4 (baseline) is (hypertension $= 0$) and (BMI $< 1$) and (BMI $> -1$), Cluster 5 is (hypertension $= 0$) and (BMI $< -1$), Cluster 6 is (hypertension $= 0$) and (BMI $> 1$), Cluster 7 is (hypertension $= 1$). Cluster 7 coincided with hypertension, so not estimated.

## 4  Discussion

We have proposed several methods to embed the survival tree algorithms and classical Cox regression into each other to explore their advantages and overcome their limitations. We aimed to employ Cox model interpretability and enhance it with the CART ability to capture non-linear and interaction relationships. The methods performed at par with the Cox model in linear data and outperformed complex data.

Ensemble 1 stacks two algorithms in a different order (first Cox, then SRF in Ensemble 1A, reverse in 1B). This is not a novel approach [4], but we reiterate its utility for health research, where classical regression models are often preferred. For example, one could train Ensemble 1a or 1b, and if a similar performance is achieved compared to the baseline Cox model, this could justify using a Cox regression.

Moreover, if the difference is considerable, it may represent the marginal contribution of the non-linear terms in the predictive performance. Indeed, AUC-ROC difference for the simulated samples 2 and 3 was 0.04–0.07 with a standard deviation of 0.02 (Table 2 and 3). Had we not known the underlying distribution, this would be a sign of non-linearity. Similarly, a comparison of the fractional polynomial model and ensemble methods results for sample 3 indicates cross-dependencies (Table 3).

Ensembles 2 and 3 can give an insight into where non-linearity lies. A clustering tree structure is the first source of such information. For example, the tree in Fig. 2 has "guessed" hypertension and BMI non-linearity. Further, Ensemble 2 results contain

estimated Cox parameters for each cluster; the difference in coefficients could reveal structural differences in the relationships across the clusters. A similar algorithm was used in [6], where the authors utilized SRF to identify clusters of different survival patterns. However, they used Kaplan-Meier curves to compute survival probabilities, and model interpretation was inferred by investigating the object properties by cluster.

Ensemble 3 gives yet another view of the data. The augmented Cox model has clusters as risk factors, and respective coefficients represent the risk of being in a cluster, in addition to the estimated linear effects. For example, Fig. 3 illustrates how Ensemble 3 has correctly identified that the linear BMI impact is negligible, while absolute BMI values above 1 (clusters 6 and 7) possess an elevated risk.

There are several ways to develop the methods further. First, combined methods 2 and 3 rely on a clustering tree. Methods behind the tree construction may affect predictive accuracy, and we may test other splitting rules. For example, an optimal tree proposed by Dunn and colleagues [25] may perform well, in which splitting minimizes the loss of the entire tree instead of optimizing a current node. Customizing the splitting criteria to the proposed methods may also enhance their performance: a partition maximizing Cox fit in the daughter leaves would realign how the tree is grown, and the predictions are made. An integrated Brier score [26] can be used as an alternative assessment measure, aggregating performance over time.

The strength of this work is a focus on targeting specific strengths and weaknesses of the classical Cox model and employing machine learning algorithms not to compete but enhance its performance. However, we acknowledge several weaknesses. First, qualitative insight into non-linearity still requires assessing the tree structures and differences between the Cox regressions in the Ensemble methods. The underlying tree instability is a challenge; further work to increase stability could be done. Second, we did not test our models on non-exponential time distribution. Finally, we primarily tested the models on the simulated data; future work to focus on real-life health data.

**Conclusion.** The proposed ensemble methods combining classical Cox proportionate hazard rate model and non-linear machine-learning algorithms can effectively build high-performing and interpretable prognostic models for health research, especially for the data where strong interaction dependencies are suspected.

## References

1. LeBlanc, M., Crowley, J.: Relative risk trees for censored survival data. Biometrics **48**, 411–425 (1992)
2. Ishwaran, H., Kogalur, U.B., Blackstone, E.H., Lauer, M.S.: Random survival forests. Ann. Appl. Stat. **2**, 841–860 (2008)
3. Cox, D.R.: Regression models and life-tables. J. Roy. Stat. Soc.: Ser. B (Methodol.) **34**, 187–202 (1972)
4. Amunategui, M.: Data Exploration & Machine Learning, Hands-on. https://amunategui.github.io/survival-ensembles/index.html
5. Marmerola, G.D.: Calibration of probabilities for tree-based models | Guilherme's Blog. https://gdmarmerola.github.io/probability-calibration/

6. Shi, T., Seligson, D., Belldegrun, A.S., Palotie, A., Horvath, S.: Tumor classification by tissue microarray profiling: random forest clustering applied to renal cell carcinoma. Mod. Pathol. **18**, 547–557 (2005)
7. Su, X., Tsai, C.-L.: Tree-augmented Cox proportional hazards models. Biostatistics **6**, 486–499 (2005)
8. Breslow, N.E.: Discussion of Professor Cox's paper. J. Roy. Stat. Soc. Ser. B (Methodol.) **34**, 202–220 (1972)
9. Kalbfleisch, J.D., Prentice, R.L.: The Statistical Analysis of Failure time Data. Wiley, Hoboken (2002)
10. Xia, F., Ning, J., Huang, X.: Empirical comparison of the Breslow estimator and the Kalbfleisch prentice estimator for survival functions. J. Biom. Biostat. **9**, 392 (2018)
11. Breiman, L., Friedman, J.H., Olshen, R.A., Stone, C.J.: Classification and Regression Trees. Wadsworth & Brooks/Cole Advanced Books & Software, Monterey, CA (1984)
12. Zhou, Y., McArdle, J.J.: Rationale and applications of survival tree and survival ensemble methods. Psychometrika **80**, 811–833 (2015)
13. Segal, M.R.: Regression trees for censored data. Biometrics **44**, 35 (1988)
14. Molinaro, A.M., Dudoit, S., van der Laan, M.J.: Tree-based multivariate regression and density estimation with right-censored data. J. Multivar. Anal. **90**, 154–177 (2004)
15. Shimokawa, A., Kawasaki, Y., Miyaoka, E.: Comparison of splitting methods on survival tree. Int. J. Biostat. **11**, 175–188 (2015)
16. Ishwaran, H., Lauer, M.S., Blackstone, E.H., Lu, M.: randomForestSRC: Random Survival Forests Vignette (2021)
17. Therneau, T., Atkinson, E.: An introduction to recursive partitioning using the RPART routines (2019)
18. Ishwaran, H.: Variable importance in binary regression trees and forests. Electron. J. Stat. **1**, 519–537 (2007)
19. Royston, P., Altman, D.G.: Regression using fractional polynomials of continuous covariates: parsimonious parametric modelling. Appl. Stat. **43**, 429 (1994)
20. Heinze, G., Ambler, G., Benner, A.: Package 'mfp.' CRAN (2022)
21. Steptoe, A., Breeze, E., Banks, J., Nazroo, J.: Cohort profile: the English longitudinal study of ageing. Int. J. Epidemiol. **42**, 1640–1648 (2013)
22. Hackinger, S., et al.: Evidence for genetic contribution to the increased risk of type 2 diabetes in schizophrenia. Transl. Psychiatry. **8**, 252 (2018)
23. Wray, N.R., Lee, S.H., Mehta, D., Vinkhuyzen, A.A.E., Dudbridge, F., Middeldorp, C.M.: Research review: polygenic methods and their application to psychiatric traits. J. Child Psychol. Psychiatry **55**, 1068–1087 (2014)
24. Blanche, P., Latouche, A., Viallon, V.: Time-dependent AUC with right-censored data: a survey. In: Lee, M.-L., Gail, M., Pfeiffer, R., Satten, G., Cai, T., Gandy, A. (eds.) Risk Assessment and Evaluation of Predictions. LNS, vol. 210, pp. 239–251. Springer, New York (2013). https://doi.org/10.1007/978-1-4614-8981-8_11
25. Dunn, J., Gibson, E., Orfanoudaki, A.: Optimal survival trees. arXiv preprint (2020)
26. Graf, E., Schmoor, C., Sauerbrei, W., Schumacher, M.: Assessment and comparison of prognostic classification schemes for survival data. Stat. Med. **18**, 2529–2545 (1999)