Scientific Journal of Informatics
Vol. 11, No. 3, Aug 2024
p-ISSN 2407-7658 https://journal.unnes.ac.id/journals/sji/index e-ISSN 2460-0040

Comparison of Extremely Randomized Survival Trees and Random Survival Forests: A Simulation Study

Mohamad Solehudin Zaenal[1*], Anwar Fitrianto[2], Hari Wijayanto[3]
[1,2,3]Department of Statistics, Faculty of Mathematics and Natural Science, IPB University, Indonesia

Abstract.
Purpose: This simulation study investigates the Extremely Randomized Survival Trees (EST) model, a machine learning technique expected to handle survival analysis, particularly in large survival datasets, effectively. The study compares the performance of the EST model with that of the Random Survival Forest (RSF) model, focusing on the C-index value to determine which model performs better.
Methods: The analysis begins with the generation of 540 simulated datasets, created by combining three levels of sample sizes, two levels of censoring proportions, three types of hazard functions, and 30 repetitions for each scenario. The simulation data were split into 80% training and 20% testing data. The training data were used to build the EST and RSF models, while the test data were used to evaluate their performance. The model with the highest C-index value was deemed the best performer, as a higher C-index indicates superior model performance.
Result: The results indicate that the sample size, type of hazard function, and the method used influence that model performance. The EST model significantly outperformed the RSF model when the sample size was large, though no significant difference was observed when the sample size was small or medium. Additionally, the EST model consistently demonstrated faster computation times across all simulation scenarios.
Novelty: This study provides a pioneering exploration into applying decision tree algorithms, specifically EST and RSF, in survival analysis. While these methods have been extensively studied in regression and classification contexts, their application in survival analysis remains relatively unexplored.

Keywords: Extremely randomized survival trees, Extra survival trees, Random survival forest
Received June 2024 / **Revised** July 2024 / **Accepted** August 2024

INTRODUCTION

Machine learning modeling advances aim for better model performance and faster computational capabilities. One rapidly evolving machine learning model is based on the decision tree algorithm, known for its effectiveness in handling complex data and providing a visual explanation of predictions [1]. While decision trees have been extensively studied and implemented in classification and regression cases, their application in survival case studies remains limited. In July 2022, data scraping was conducted to gather information on the number of publications related to algorithm-based models, and the following table summarizes the findings:

Table 1. Summary of publications for various tree algorithm-based models

No	Method	Number of Publications	Case of Study
1	*Classification and Regression Tree (CART)*	906	Classification/regression
2	*Bagging*	5,030	Classification/regression
3	*Random Forest*	20,900	Classification/regression
4	*Extremely Randomized Trees*	101	Classification/regression
5	*Double Random Forest*	10	Classification/regression
6	*Random Survival Forest (RSF)*	165	Survival
7	*Extremely Randomized Survival Trees (EST)*	Not Found	Survival

Source: Author's documentation, July 2022

[*]Corresponding author.
Email addresses: zmmohamad@apps.ipb.ac.id (Zaenal)[*], anwarstat@gmail.com (Fitrianto), hari@apps.ipb.ac.id (Wijayanto)
DOI: 10.15294/sji.v11i3.8464

Table 1 illustrates that the most frequently mentioned models in publications are the Random Forest model, with over 20,000 publications; Bagging, with over 5,000 publications; and the Classification and Regression Tree (CART), with over 900 publications. On the other hand, the models with the lowest number of publications are the Random Survival Forest (RSF), with over 100 publications; the Double Random Forest (DRF), with 10 publications; and the Extremely Randomized Survival Trees (EST) model, for which no publications were found. Another important observation from Table 1 is that models analyzing classification/regression case studies are more prevalent than those analyzing survival case studies. This finding reinforces that research on machine learning models for survival analysis remains limited. Consequently, this research focuses on one of the decision tree-based models for survival analysis, known as the survival tree model [2].

Survival analysis encompasses statistical methods or tools to analyze the time until an event occurs [3], [4]. The event experienced by an individual can be singular or multiple [5]. Survival analysis differs from general regression analysis because it involves censored data or data without complete information [6]. Censored data arises when the expected event does not occur within the observation period [7]. The presence of censored data renders ordinary statistical methods inadequate [8].

The EST model in survival analysis extends the Extremely Randomized Trees method introduced by Geurts [9] for classification and regression case studies. In various classification and regression cases, the Extra Trees model has outperformed other machine learning models, such as Random Forest (RF), Single CART, K-nearest Neighbor (KNN), Bagging, XGBoost (XGB), Quadratic Discriminant Analysis (QDA), and Bayesian models [10], [11], [12]. Dey introduced the EST model [13]. The RSF model was first introduced by Ishwaran et al. [14]. The RSF model is chosen as a comparison method because it shares characteristics suitable for comparison with EST, including the nature of the data used to build the model, the involvement of predictor variables in the splitting process, and the inclusion of candidate cut points during tree formation. This study aims to compare the performance of the Extremely Randomized Survival Trees (EST) and Random Survival Forest (RSF) models.

METHODS

The analysis commenced with generating 540 simulated datasets, created by combining three sample sizes, two levels of censoring proportions, three types of hazard functions, and 30 repetitions for each scenario. Each simulated dataset was partitioned into 80% training and 20% testing data [15]. The training data were used to build the Extremely Randomized Survival Trees (EST) and Random Survival Forest (RSF) models, while the testing data were employed to evaluate model performance. The best model was identified based on the C-index value, where a higher C-index indicates superior performance compared to models with lower C-index values [16]. The process flow for data generation and modeling is illustrated in Figure 1.

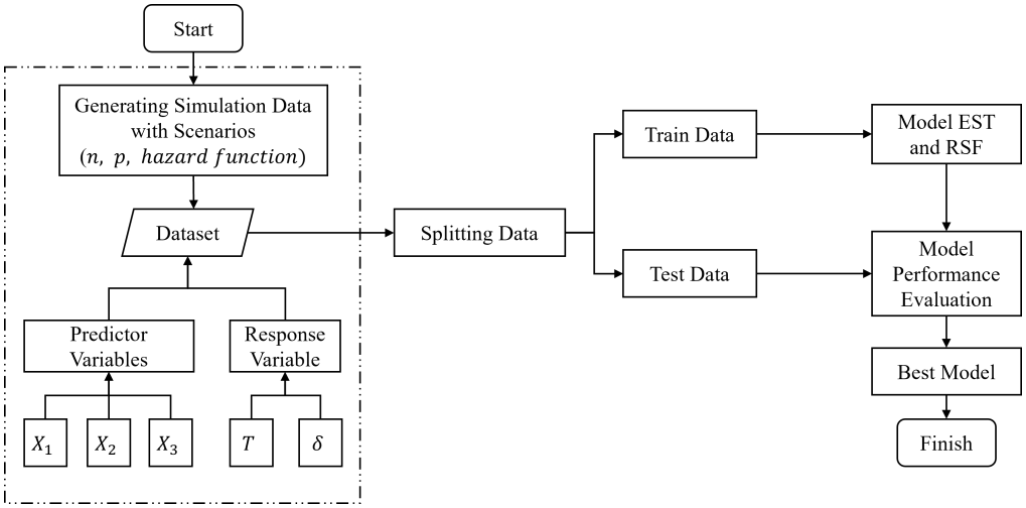


Figure 1. Analysis method

Scenario variables for simulation data

A simulation study with right-censored data was conducted using RStudio 2022.7.2 software. The data were generated by combining numerical and categorical predictor variables, three types of hazard functions, two levels of censored data proportions (p), and three levels of the number of observations (n). The hazard function represents the probability that an individual will experience a risk/event within a specific time interval [17].

Five variables were generated according to the survival data scenario, comprising one categorical predictor variable, two numerical predictor variables, one survival time variable, and one censored status variable. The predictor variables were generated following uniform and binomial distributions. Due to the data generation process, the appearance of predictor variable data is random. However, in the modeling context, the existence of predictor variables is considered constant. The survival time variable was generated following a Weibull distribution with shape parameter α and scale parameter λ in three scenario forms based on hazard aspects.

The form of the hazard function depends on the value of the parameter α [18]. First, the decreasing Weibull distribution, or decreasing hazard function, occurs when the shape parameter $\alpha < 1$. Secondly, the increasing Weibull distribution, or increasing hazard function, occurs when the shape parameter $\alpha > 1$. Third, the constant Weibull distribution, or constant hazard function, occurs when $\alpha = 1$.

Survival data simulation

The following are the steps for generating simulation data as conducted , [19], [20]:

1) Determine the number of observations (n) to be 300, 600, and 1.200, then determine the proportion of censored data (p) to be 20% and 50%.
2) Determine the β_0 value with three scenarios:
   a. $\beta_0 = -0{,}69$ for the decreasing hazard function;
   b. $\beta_0 = -2{,}08$ for the increasing hazard function;
   c. $\beta_0 = -1{,}39$ for the constant hazard function;
3) Determine $\beta_1, \beta_2, \ and \ \beta_3$ values with three scenarios:
   a. $\beta_1 = 0{,}1,\ \beta_2 = -0{,}1,\ \beta_3 = 0{,}5$ for the decreasing hazard function;
   b. $\beta_1 = 0{,}1,\ \beta_2 = -0{,}8,\ \beta_3 = 0{,}5$ for the increasing hazard function;
   c. $\beta_1 = 0{,}4,\ \beta_2 = -0{,}7,\ \beta_3 = 1$ for the constant hazard function;
4) Determine categorical predictor variable $X_1 \sim binomial(n; 0{,}5)$, and numerical predictor variables $X_2 \sim U(0{,}1)$ and $X_3 \sim U(0{,}1)$.
5) Generate survival time $T_i \sim Weibull(\alpha, \lambda)$ with three scenarios:
   a. $\alpha_t = 0{,}4$ for the decreasing hazard function;
   b. $\alpha_t = 1{,}5$ for the increasing hazard function;
   c. $\alpha_t = 1$ for the constant hazard function;

   The value of parameter λ is obtained as follows [21]:
   $$\lambda = \exp\left(-\frac{\beta_0}{\alpha} - \sum_{k=1}^{n} \frac{\beta_k}{\alpha} X_k\right), where\ k = 1{,}2{,}3. \quad (1)$$
6) Generate sensor time $C_i \sim Weibull(\alpha, \lambda)$ with three scenarios:
   a. $\alpha_t = 0{,}4$ for the decreasing hazard function;
   b. $\alpha_t = 1{,}5$ for the increasing hazard function;
   c. $\alpha_t = 1$ for the constant hazard function;

   The value of parameter λ is obtained from the uniroot algorithm [22].
7) Set the censored state variable δ_i, where:
   $$\delta_i = \begin{cases} 0, & censored\ if\ T_i > C_i \\ 1, & uncencored\ if\ T_i \leq C_i \end{cases} \quad (2)$$
8) Combine all variables into a data frame and repeat the process from 1 to 8 with 30 repetitions.

Random survival forest

1) Draw ***B*** bootstrap samples from the data by sampling with replacement. Each bootstrap sample is used to form a survival tree. Approximately 37% of the data are omitted from each bootstrap sample, known as out-of-bag (OOB) data.
2) For each terminal node in the tree, randomly select m predictor variables for splitting.

3) Split a node using the log-rank splitting rule based on the predictor variable that produces the most significant difference between the two survival functions of its child nodes.
4) Repeat steps (2) and (3) until a large number of trees are obtained, with the stopping rule criteria being that each terminal node has a minimum of $d_0 > 0$ unique failure data points.
5) Calculate the Cumulative Hazard Function (CHF) value for each terminal node in each tree using the Nelson-Alaen estimator:

$$\widehat{H}_h(t) = \sum t_{l,h} < \frac{t_{d_{l,h}}}{r_{l,h}} \quad (3)$$

where $t_{l,h}$ denotes the time to event – l in cluster h. $d_{l,h}$ is the number of events at $t_{l,h}$, and $r_{l,h}$ is the number of individuals at risk $t_{l,h}$.
6) Find the CHF ensemble value by averaging the CHF values of all trees to obtain the CHF ensemble bootstrap:

$$H_e^{**}(t|X_i) = \frac{1}{b} \sum_{b=1}^{B} H_b^*(t|X_i) \quad (4)$$

7) Use the OOB data to calculate the prediction error of the CHF ensemble:

$$H_e^{**}(t|X_i) = \frac{\sum_{b=1}^{B} I_i H_b^*(t|X_i)}{\sum_{b=1}^{B} I_{i,b}} \quad (5)$$

Extremely randomized survival trees

1) For each tree formed by the training data samples, do the following:
   a. For each tree node, randomly select $\sqrt{n}$ predictors for splitting.
   b. For each selected predictor, select a set of candidates for split points.
   c. Using the log-rank criterion, split a node by the predictor that maximizes the survival difference between child nodes.
   d. Repeat steps a, b, and c above until each terminal node contains no more than 0.632 times the number of events.
2) Calculate the CHF for each tree to obtain an ensemble of estimated values for the cumulative hazard.

Modeling

The modeling process for EST and RSF utilizes the *ExtraSurvivalTrees* and *RandomSurvivalForest* libraries within the *scikit-survival 0.22.2* module using the Python programming language. The following are the steps of analysis for the simulation data:

1) Split the data into training data (80%) and testing data (20%).
2) Construct the RSF and EST models based on the training data. The default model parameters are set to create 500 trees, with a minimum of 10 samples required for the splitting process and 15 samples per leaf node.
3) Evaluate the models using the testing data based on the C-index value.
4) Determine the best model, compare the average C-index values using a paired t-test, and assess influential factors using Analysis of Variance (ANOVA) [23]. The paired t-test is employed because the data are from the same source [24].

RESULTS AND DISCUSSIONS

Majority votes

Analysis was conducted on 180 simulations for each hazard function group: increasing, constant, and decreasing, resulting in 540 simulations. The frequency of model superiority based on C-index values for each simulation is presented in Table 2.

Table 2. EST and RSF comparisons based on c-index

No	Value of C-index	Frequency	Percentage
1	EST > RSF	365	67%
2	RSF > EST	175	33%
	Total	540	100%

Based on Table 2, out of 540 simulations, the C-index value of the EST model appeared 365 times (67%), surpassing the C-index value of the RSF model, which appeared 175 times (33%). According to the majority vote rule, the EST model outperforms the RSF model.

Model performance

Next, an analysis of model performance was conducted for the constant, increasing, and decreasing hazard function groups. This analysis was conducted to understand better the model's behavior within each hazard function group.

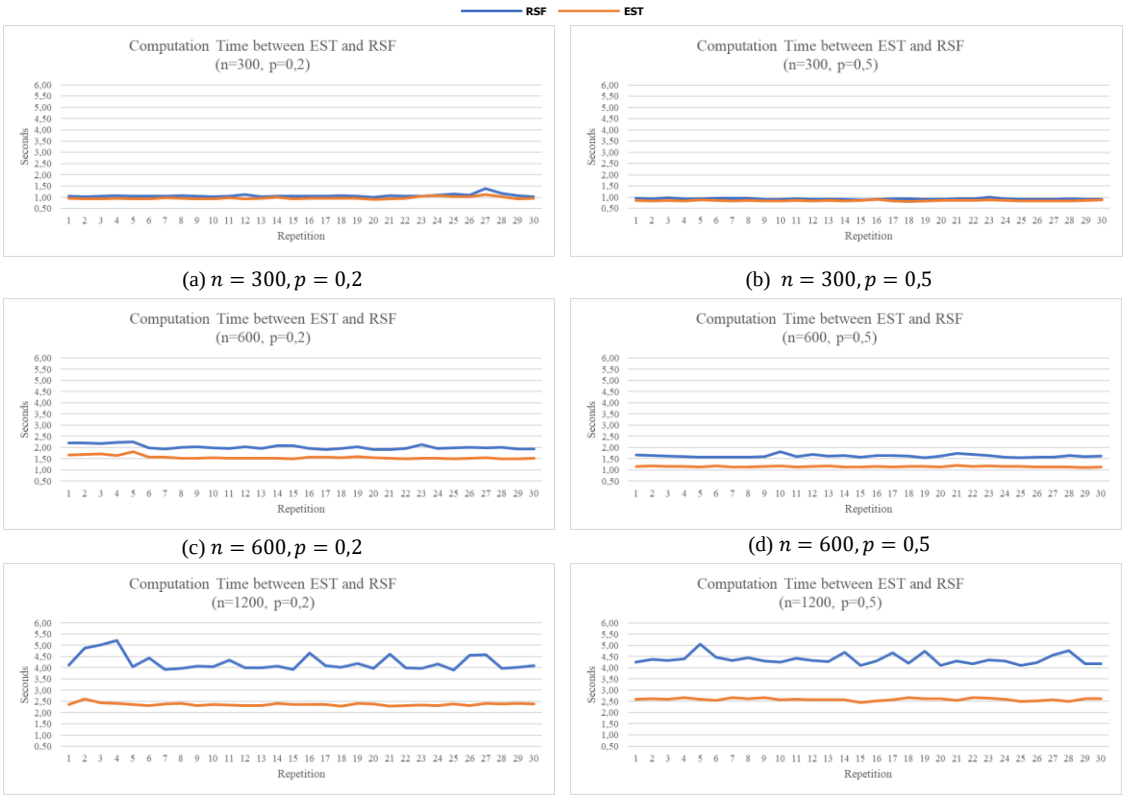
Table 3. Model performance based on hazard function groups

n	p	Constant Hazard Function				Increasing Hazard Function				Decreasing Hazard Function			
		Mean of C-index		Paired t-test		Mean of C-index		Paired t-test		Mean of C-index		Paired t-test	
		EST	RSF	t-hit	p-value	EST	RSF	t-hit	p-value	EST	RSF	t-hit	p-value
300	0.2	0.6060	0.6099	0.917	0.367	0.5778	0.5774	-0.074	0.941	0.5335	0.5348	0.238	0.814
	0.5	0.6236	0.6133	-1.508	0.142	0.5744	0.5740	0.000	0.983	0.5345	0.5352	0.123	0.903
600	0.2	0.6136	0.6108	-1.040	0.307	0.5670	0.5625	-1.888	0.069	0.5215	0.5229	0.371	0.713
	0.5	0.6131	0.6085	-1.083	0.288	0.5619	0.5567	-1.593	0.122	0.5208	0.5172	-0.720	0.477
1.200	0.2	0.6108	0.6005	-7.905	0.000	0.5754	0.5621	-7.199	0.000	0.5410	0.5256	-6.543	0.000
	0.5	0.6103	0.5940	-6.458	0.000	0.5729	0.5560	-5.980	0.000	0.5313	0.5156	-3.807	0.001

Analysis of Table 3 shows that in the constant hazard function group, when the number of observations is small ($n = 300$) or medium ($n = 600$), there is no significant difference between the average C-index values of the EST and RSF models, both for the 20% and 50% censored data proportions. However, when the number of observations is large ($n = 1{,}200$) across all censored data classes ($p = 20\%;\ 50\%$), the average C-index value of the EST model is significantly higher than that of the RSF model ($pvalue = 0{,}000 < \alpha = 5\%$). Specifically, with a large number of observations ($n = 1.200$) and a censored data proportion of $p = 20\%$, the average C-index value of the EST model is 0.6108, higher than the RSF model's average C-index value of 0.6005. Similarly, with many observations and a censored data proportion of $p = 50\%$, the EST model's average C-index value is 0.6103, compared to the RSF model's average C-index value of 0.5940. The comparison indicates that the EST model performs better due to having a higher C-index value than the RSF model.

Similar trends are observed in the increasing and decreasing hazard function groups. When the number of observations is small or medium, there is no significant difference between the average C-index values of the EST and RSF models. However, with many observations across all censored data classes ($p = 20\%;\ 50\%$), the average C-index value of the EST model is significantly higher than that of the RSF model. The average C-index values for the increasing hazard function group are 0.5754 and 0.5729 for the EST model, compared to 0.5621 and 0.5560 for the RSF model. The average C-index values for the decreasing hazard function group are 0.5410 and 0.5313 for the EST model, compared to 0.5256 and 0.5156 for the RSF model.

Computation time for the model



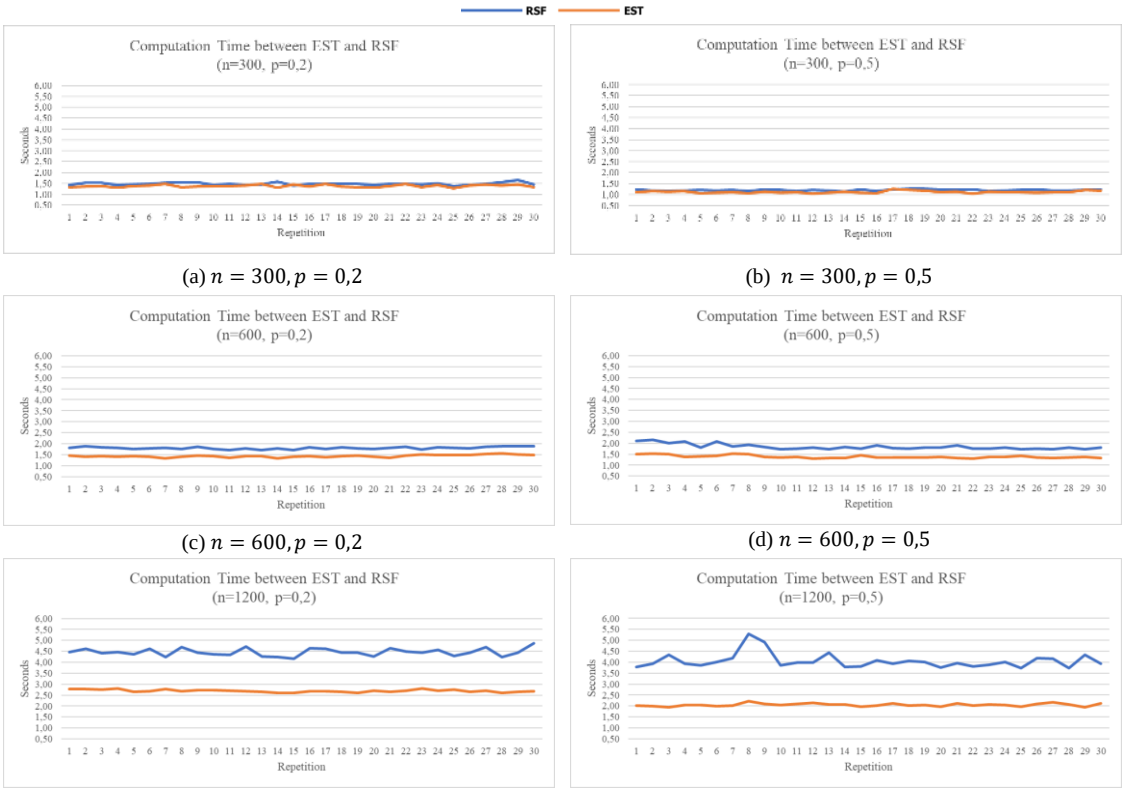
(a) $n = 300, p = 0{,}2$ (b) $n = 300, p = 0{,}5$

(c) $n = 600, p = 0{,}2$ (d) $n = 600, p = 0{,}5$

(e) $n = 1.200, p = 0{,}2$ (f) $n = 1.200, p = 0{,}5$

Figure 2. Computation time graph for the constant hazard function group

The analysis of Figure 2 indicates that within the constant hazard function group, the EST model consistently exhibits faster computation times than the RSF model. Specifically, the EST model's average computation time is 1.07 times faster than the RSF model for small observation classes (Figures 2a and 2b), 1.30 times faster for medium observation classes (Figures 2c and 2d), and 1.80 times faster for large observation classes (Figures 2e and 2f).



(a) $n = 300, p = 0{,}2$ (b) $n = 300, p = 0{,}5$

(c) $n = 600, p = 0{,}2$ (d) $n = 600, p = 0{,}5$

(e) $n = 1.200, p = 0{,}2$ (f) $n = 1.200, p = 0{,}5$

Figure 4. Computation time graph for the decreasing hazard function group

The analysis of Figure 4 shows that within the decreasing hazard function group, the EST model consistently has faster computation times than the RSF model. Specifically, the average computation time of the EST model is 1.08 times faster than the RSF model for small observation classes (Figures 4a and 4b), 1.32 times faster for medium observation classes (Figures 4c and 4d), and 1.74 times faster for large observation classes (Figures 4e and 4f).

Table 4. Computation of the model

n	p	Constant Hazard Function		Increasing Hazard Function		Decreasing Hazard Function	
		EST (s)	RSF (s)	EST (s)	RSF (s)	EST (s)	RSF (s)
300	0,2	0,969	1,075	1,382	1,483	1,039	1,135
	0,5	0,852	0,929	1,126	1,210	1,465	1,591
600	0,2	1,563	2,027	1,435	1,803	1,696	2,217
	0,5	1,152	1,621	1,385	1,841	1,584	2,123
1200	0,2	2,373	4,230	2,697	4,465	2,759	4,782
	0,5	2,593	4,376	2,051	4,055	2,672	4,662

Note: (s) indicates the unit of seconds.

The computation time of a model is directly proportional to the size of the observation class. Larger observation classes generally require longer computation times, while smaller observation classes lead to faster computation times. For example, in the RSF model with a constant hazard function group and a censored data proportion of 20%, the average computation time required is 1.075 seconds for n=300 observations. When the number of observations n increases to 600, the computation time is 2.207 seconds, and for n=1,200 observations, the computation time increases to 4.230 seconds.

Influencing Factors

Table 5. Influencing factors by analysis of variance (ANOVA)

Factor	Degrees of Freedom	p-value	Results
n	2	0,008	Significant
α	2	0,000	Significant
p	1	0,425	Not significant
$metode$	1	0,025	Significant
Interaction $n\&a$	4	0,334	Not significant
Interaction $n\&p$	2	0,437	Not significant
Interaction $n\&metode$	2	0,094	Not significant
Interaction $\alpha\&p$	2	0,584	Not significant
Interaction $\alpha\&metode$	2	0,967	Not significant
Interaction $p\&metode$	1	0,517	Not significant

Based on the results in Table 5, it is evident that the number of observations (n), hazard function type (α), and method significantly influence the model performance (C-index value), as indicated by $pvalues < 5\%$ [25]. This suggests that any changes in $n, \alpha,$ or $method$ will likely impact the model's performance. However, interactions involving n with $\alpha, p,$ and $method$ do not show significant effects, implying that the model's performance tends to remain similar even when interactions occur between $n, \alpha, p,$ and $method$ factors and other variables.

CONCLUSION

This study explores the performance differences between the EST and RSF models. The research involves several key stages, including generating survival data simulations, resulting in 540 simulations across various scenarios, model training, and model evaluation to determine the best performance based on the C-index metric. The simulation analysis reveals no significant performance difference between the EST and RSF models for small and medium sample sizes across all censoring proportions. However, the EST model significantly outperforms the RSF model for large sample sizes based on the C-index metric across all censoring proportions. Additionally, regarding computational time, the EST model consistently demonstrates faster computation times than the RSF model. The study recommends that future research include a comparative analysis of the EST and RSF models with more variables (e.g., more than 20), k-fold cross-validation, and parameter tuning to optimize model performance.

REFERENCES

[1] N. Romadoni, A. Mutoi Siregar, D. S. Kusumaningrum, and T. Rohana, "Classification Model of Public Sentiments About Electric Cars Using Machine Learning," *Scientific Journal of Informatics*, vol. 11, no. 2, 2024, doi: 10.15294/sji.v11i2.1309.

[2] Y. Zhou and J. J. McArdle, "Rationale and Applications of Survival Tree and Survival Ensemble Methods," *Psychometrika*, vol. 80, no. 3, pp. 811–833, Sep. 2015, doi: 10.1007/s11336-014-9413-1.

[3] D. Collett, *Modelling Survival Data in Medical Research*. Chapman and Hall/CRC, 2015. doi: 10.1201/b18041.

[4] N. R. N. Insyira, R. M. Atok, and I. S. Ahmad, "Pemodelan Waktu Survival Pekerja dengan Menggunakan Regresi Cox Proportional Hazard," *Jurnal Sains dan Seni ITS*, vol. 7, no. 2, Feb. 2019, doi: 10.12962/j23373520.v7i2.34697.

[5] C. Pancotti, G. Birolo, T. Sanavia, C. Rollo, and P. Fariselli, "Multi-Event Survival Prediction for Amyotrophic Lateral Sclerosis," 2022. [Online]. Available: http://ceur-ws.org

[6] N. A. Schuster, E. O. Hoogendijk, A. A. L. Kok, J. W. R. Twisk, and M. W. Heymans, "Ignoring competing events in the analysis of survival data may lead to biased results: a nonmathematical illustration of competing risk analysis," *J Clin Epidemiol*, vol. 122, pp. 42–48, Jun. 2020, doi: 10.1016/j.jclinepi.2020.03.004.

[7] A. J. Turkson, F. Ayiah-Mensah, and V. Nimoh, "Handling Censoring and Censored Data in Survival Analysis: A Standalone Systematic Literature Review," 2021, *Hindawi Limited*. doi: 10.1155/2021/9307475.

[8] J. B. Nasejje, "Application of survival analysis methods to study under-five child mortality in Uganda." [Online]. Available: https://www.researchgate.net/publication/281240839

[9] P. Geurts, D. Ernst, and L. Wehenkel, "Extremely randomized trees," *Mach Learn*, vol. 63, no. 1, pp. 3–42, Apr. 2006, doi: 10.1007/s10994-006-6226-1.

[10] X. Wang, L. Tan, and J. Fan, "Performance Evaluation of Mangrove Species Classification Based on Multi-Source Remote Sensing Data Using Extremely Randomized Trees in Fucheng Town, Leizhou City, Guangdong Province," *Remote Sens (Basel)*, vol. 15, no. 5, p. 1386, Mar. 2023, doi: 10.3390/rs15051386.

[11] A. Pagliaro, "Forecasting Significant Stock Market Price Changes Using Machine Learning: Extra Trees Classifier Leads," *Electronics (Basel)*, vol. 12, no. 21, p. 4551, Nov. 2023, doi: 10.3390/electronics12214551.

[12] N. P. Sarini and K. Dharmawan, "Estimation of Bali Cattle Body Weight Based on Morphological Measurements by Machine Learning Algorithms: Random Forest, Support Vector, K-Neighbors, and Extra Tree Regression," *Journal of Advanced Zoology*, vol. 44, no. 3, pp. 1–9, Oct. 2023, doi: 10.17762/jaz.v44i3.234.

[13] A. K. Dey, S. N., T. S. Teja, and A. Juneja, "Some variations on Ensembled Random Survival Forest with application to Cancer Research," Sep. 2017.

[14] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," *Ann Appl Stat*, vol. 2, no. 3, Sep. 2008, doi: 10.1214/08-AOAS169.

[15] J. M. Durden, B. Hosking, B. J. Bett, D. Cline, and H. A. Ruhl, "Automated classification of fauna in seabed photographs: The impact of training and validation dataset size, with considerations for the class imbalance," *Prog Oceanogr*, vol. 196, p. 102612, Aug. 2021, doi: 10.1016/j.pocean.2021.102612.

[16] A. Alabdallah, M. Ohlsson, S. Pashami, and T. Rögnvaldsson, "The Concordance Index decomposition: A measure for a deeper understanding of survival prediction models," *Artif Intell Med*, vol. 148, p. 102781, Feb. 2024, doi: 10.1016/j.artmed.2024.102781.

[17] I. Ise *et al.*, "Hazard function analysis of prognosis after recurrent colorectal cancer," *Langenbecks Arch Surg*, vol. 409, no. 1, p. 123, Apr. 2024, doi: 10.1007/s00423-024-03308-w.

[18] M. S. Shama *et al.*, "Modified generalized Weibull distribution: theory and applications," *Sci Rep*, vol. 13, no. 1, p. 12828, Aug. 2023, doi: 10.1038/s41598-023-38942-9.

[19] S. Nurhaliza, K. Sadik, and A. Saefuddin, "A Comparison of Cox Proportional Hazard and Random Survival Forest Models in Predicting Churn of the Telecommunication Industry Customer," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 16, no. 4, pp. 1433–1440, Dec. 2022, doi: 10.30598/barekengvol16iss4pp1433-1440.

[20] N. G. A. P. P. Suantari, A. Fitrianto, and B. Sartono, "Comparative Study of Survival Support Vector Machine and Random Survival Forest in Survival Data," *BAREKENG: Jurnal Ilmu Matematika dan Terapan*, vol. 17, no. 3, pp. 1495–1502, Sep. 2023, doi: 10.30598/barekengvol17iss3pp1495-1502.

[21] J. B. Nasejje, H. Mwambi, K. Dheda, and M. Lesosky, "A comparison of the conditional inference survival forest model to random survival forests based on a simulation study as well as on two applications with time-to-event data," *BMC Med Res Methodol*, vol. 17, no. 1, p. 115, Dec. 2017, doi: 10.1186/s12874-017-0383-8.

[22] F. Wan, "Simulating survival data with predefined censoring rates for proportional hazards models," *Stat Med*, vol. 36, no. 5, pp. 838–854, Feb. 2017, doi: 10.1002/sim.7178.

[23] Sevita Sari Dewi, Rizka Ermina, Veilla Anggoro Kasih, Fera Hefiana, Agus Sunarmo, and Rini Widianingsih, "Analisis Penerapan Metode One Way Anova Menggunakan Alat Statistik SPSS," *Jurnal Riset Akuntansi Soedirman*, vol. 2, no. 2, pp. 121–132, 2023, doi: 10.32424/1.jras.2023.2.2.10815.

[24] Nurhasan, Rico Septia B, and Syamsudin Baharsyah, "Efektivitas Pembukuan Terhadap Kinerja Keuangan Pada Toko Ritel Di Lingkungan Desa Cileungsi, Kecamatan Cileungsi, Bogor," *Journal Of Social Science Research*, vol. 4, no. 1, pp. 7029–7038, 2024.

[25] M. Albassam and M. Aslam, "Testing Internal Quality Control of Clinical Laboratory Data Using Paired t-Test under Uncertainty," *Biomed Res Int*, vol. 2021, pp. 1–6, Sep. 2021, doi: 10.1155/2021/5527845.