BAREKENG: Journal of Mathematics and Its Applications
September 2023 Volume 17 Issue 3 Page 1495–1502
P-ISSN: 1978-7227 E-ISSN: 2615-3017
https://doi.org/10.30598/barekengvol17iss3pp1495-1502

COMPARATIVE STUDY OF SURVIVAL SUPPORT VECTOR MACHINE AND RANDOM SURVIVAL FOREST IN SURVIVAL DATA

Ni Gusti Ayu Putu Puteri Suantari [1], Anwar Fitrianto[2*], Bagus Sartono[3]

[1,2,3]*Department of Statistics, Faculty of Mathematics and Natural Science, IPB University*
Raya Dramaga Street, Bogor, 16680, West Java, Indonesia

*Corresponding author's e-mail: * **anwarstat@gmail.com***

Article History:

Received: 01st March 2023
Revised: 4th August 2023
Accepted: 10th August 2023

Keywords:

Survival Support Vector Machine;
Random survival forest;
Survival analysis;
Machine learning

ABSTRACT

Survival analysis is a statistical procedure in analyzing data with the response variable is time until an event occurs (time-to-event). In the last few years, many classification approaches have been developed in machine learning, but only a few considered the presence of time-to-event variable. Random Survival Forest and Survival Support Vector Machine are machine learning approach which is a nonparametric classification method when dealing with large data and a response variable of survival time. Random Survival Forest is tree based method that using boostrapping algorithm, and Survival Support Vector Machine using hybrid approaches between regression and ranking constrain. The data used in this study is generated data in the form of right-censored survival data. This study uses the RandomForestSRC and SurvivalSVM packages on R software. This study aimed to compare the performance of the Survival Support Vector Machine and Random Survival Forest methods using simulation studies. Simulation results on right-censored survival data using binary predictor variables scenario indicate that the Survival Support Vector Machine (SSVM) method with Radial Basic Function Kernel (RBF Kernel) has the best model performance on data with small volumes, whereas when the data volume becomes larger, the method that has the best performance is Survival Support Vector Machine using Additive Kernel. Meanwhile, Random Survival Forest is a method that has the best performance for all conditions in mixed predictor variables scenario. Method, proportion of censored data and size of data are factors that affect the model performance.

How to cite this article:

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: J. Math. & App.*, vol. 17, iss. 3, pp. 1495-1502, September, 2023.

Journal homepage: *https://ojs3.unpatti.ac.id/index.php/barekeng/*
Journal e-mail: *barekeng.math@yahoo.com; barekeng.journal@mail.unpatti.ac.id*

Research Article • Open Access

1. INTRODUCTION

Survival analysis is a statistical procedure for analyzing data with a response variable: the time until an event occurs (time-to-event) [1] . Many things should be considerd in survival analysis. The first thing is types of censoring. There are three types of censoring in this analysis, such as right censoring, left censoring and interval censoring [2]. Right censoring is used for the object who is not have an event until the end of the research. Right censoring is also the most frequent type of censoring that used in the research [3]. Second, in survival analysis, there is a function that show the probability of an event occurring at a given time interval named hazard function. There are three kinds of hazard function: decreasing hazard function, increasing hazard function and constant hazard function [4]. Hazard function will affect the probability of the object to survive based on the time. Third, three kinds of survival analysis methods are parametric, semi-parametric and non-parametric methods [5].

One of the developments in survival analysis using non-parametric machine learning method was carried out by Fouodo et al who applied survival analysis with Support Vector Machine (SVM) approach [6]. Xia & Jin [7] developed and then Sudharsan [8] applied it in non-medical case for churn prediction. Ishwaran et al [9] created the development of Random Forest to analyze survival data and then Ullah *et al* [10] applied it non-medical case, especially in a telecommunication company.

Several subsequent studies have discussed the performance comparation between non-parametric methods using machine learning with the semi-parametric method. Van Belle *et al* [11] used Survival Support Vector Machine (SSVM). They compared it with Cox-Proportional Hazard (CPH), Accelerated Failure Time Model (AFT Model), cSVM-Linear (SVM with Linear Kernel) and cSVM-Gaussian Radial Basis Function Kernel (SVM with RBF Kernel and used 2 different dataset. The conclusion is the results using SSVM are better than using CPH or AFT for that case. Besides that, Hadanny *et al* [12] and Khotimah *et al* [13] compared CPH and Random Survival Forest (RSF) to build the best model that can find influential factor from survival time of the patient. They stated that RSF performs better based on concordance index (c-index) and log rank score. Then, Saadati & Bagheri [14] applied and compared RSF and CPH, CI-Forest, and Kaplan Meier to predict the distance between the first child birth and marriage and they found that RSF was the best method with the biggest c-index score.

Based on the explanation presented, this research aims to conduct a simulation study to compare the performance of Survival Support Vector Machine and Random Survival Forest on survival data, and determine the factors that influence model performance. This research follows the scenario of right censoring data and use package RandomForestSRC and SurvivalSVM in R.

2. RESEARCH METHODS

The simulations in this research follow the simulation scenarios carried out by Nesseije *et al* [15] and Wan *et al* [16]. Data in this simulation was generated from the right censoring survival scenario. The model used in this study is a proportional hazard model with a Weibull spread response variable. The general form of this model is as follow [17]:

$$h(y) = \lambda y^{\lambda-1} \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k), y \sim Weibull \quad (1)$$

which $\boldsymbol{\lambda}$ is the shape parameter from Weibull distribution.

The process includes 2 stages, such as, process of generating simulation data and process of analyzing simulated data.

2.1 Process of Generating Simulated Data

The data was generated using 2 scenarios of predictor variables, there are binary and secondary predictor variables. The steps taken in the process of generating data with binary or mixed predictor scenarios are as follows:

1. Set the amount of data, n = 200, 400, 600, 800, 1000, 1200, 1400, 1600, 1800 dan 2000.
2. Generate predictor variables X_1, X_2, X_3, X_4, for the simulation with binary predictor variables scenarios, all of the predictor variables spread $X_1 \sim Bernoulli$ (0.3), $X_2 \sim Bernoulli$ (0.4), $X_3 \sim Bernoulli$

(0.5), $X_4 \sim$ *Bernoulli* (0.6). For the simulation with mixed predictor variables, the predictor variables spread $X_1 \sim N(0,1)$, $X_2 \sim U(0,1)$, $X_3 \sim$ *Bernoulli*(0.5) dan $X_4 \sim$ *Poisson*(5).

3. Set the coefficient value $\beta_0, \beta_1, \beta_2, \beta_3, \beta_4$. The coefficient was only until β_4 because this study only used 4 predictor variables.
4. Generate data with 3 scenarios with parameter α = 0.8 for scenario 1 (decreasing hazard function), α = 1.2 for scenario 2 (increasing hazard function) and α = 1 for scenario 3 (constant hazard function). Parameter λ obtained based on formula:

$$\lambda = \exp\left(-\frac{\beta_0}{\alpha} - \sum_{i=i}^{k} \frac{\beta_k}{\alpha} X_k\right), k = 1,2,3,4 \quad (2)$$

5. Generate the survival time $t_i, i = 1,2, \ldots n$ use random value from $t_i \sim$ *Weibull*(*scale* = α, *shape* = λ).
6. Set value of parameter θ used this formula:

$$(\theta|p) = \mathbb{P}(\delta = 1|\alpha, \theta) - p \quad (3)$$
$$= \int_0^{+\infty} \mathbb{P}(\delta = 1|u, \alpha, \theta) f_{\lambda_i}(u_l) du - p,$$

with

$$\mathbb{P}(\delta = 1|\lambda_i, \alpha, \theta) = \frac{\lambda_i}{\alpha\theta} \Gamma\left(\frac{1}{\alpha}, \left(\frac{\theta}{\lambda_i}\right)^{\alpha}\right) \quad (4)$$

Then the solution of parameter θ is obtained by solving the equation $\gamma(\theta|p) = 0$ for all possible combinations of individual censor opportunities $\mathbb{P}(\delta = 1|\lambda_i, \alpha, \theta)$ and density function $f_{\lambda_i}(u_l)$ The proportion of censored data (p) used is 20%, 50% and 80%. The density functions $f_{\lambda_i}(u_l)$ in binary predictor variables scenario can be written as follows:

$$f_{\lambda_i}(u_l) = \prod_{j=1}^{k} p_j^{l_{(k,2)}^{(j)}} (1 - p_j)^{1 - l_{(k,2)}^{(j)}} \quad (5)$$

with

$$u_l = \exp\left(-\frac{\beta_0}{\alpha} - \sum_{j=1}^{k} \frac{\beta_j l_{(k,2)}^{(j)}}{\alpha}\right) \quad (6)$$

For $l \in \{0,1,2, \ldots, 2^k\}$ and $l_{(k,2)}^{(j)}$ is j^{th} element in $k \times 1$ vector from l^{th} realization from binary variables $< X_1, X_2, \ldots, X_k >$. While the probability density function $f_{\lambda_i}(u_l)$ in the mixed predictor variable scenario can be written as follows:

$$f_{\lambda_i}(u) = \frac{1}{u\sqrt{2\pi}\sqrt{\sum_{j=1}^{k} \frac{\beta_j^2}{\alpha^2}}} \exp\left(-\frac{\left(\ln(\mu) + \frac{\beta_0}{\alpha}\right)^2}{2\sum_{j=1}^{k}\left(\frac{\beta_j^2}{\alpha^2}\right)}\right) \quad (7)$$

7. Generate T_i censored time from a random variable $T_i \sim U(0, \theta)$, assuming that the data is right censored and δ_i is an uncensored indicator.

$$\delta_i = \begin{cases} 0, & censored\ if\ t_i > T_i \\ 1, & uncensored\ if\ t_i \leq T_i \end{cases} \quad (8)$$

8. Combining survival time t_i, uncensored indicator δ_i and independent variables into one dataset.
9. Repeat step 1-8 for 3 times using 3 different generation data for each combination of n, α and p.

2.2 Process of Simulation Data Analysis

After generating the data, the following stages of analysis are carried out to obtain the best method.

1. Divide the data into 2 parts, 70% training data and 30% testing data.
2. Using training data, build a Random Survival Forest and Survival Support Vector Machine model with Linear Kernel, Gaussian Radial Basic Kernel Function, and Additive Kernel.
3. Evaluate all models formed using test data by considering the c-index value.
4. Conduct an Analysis of Variance (ANOVA) test to determine the effect of each parameter and the interaction of the c-index value using the following test hypothesis **[18]** :

   H_0 : The factor being tested has no significant effect on model performance.

   H_1 : The factor being tested has a significant effect on model performance.

   The conclusion obtained is reject H_0 when the *p-value* < 0.05, meaning that the factor being tested has a significant effect on model performance.
5. Interpretation of result.

3. RESULTS AND DISCUSSION

There are 2 main scenarios used in this study, scenarios using binary predictor variables and scenarios using mixed predictor variables.

3.1 Simulation using Binary Predictor Variables

Three hazard function scenarios were represented by α, 3 censored data proportion scenarios represented by *p*, 10 data quantity scenarios described by *n* and modeled by 4 methods then evaluated with *c-index*. C-*index* is show the performance of model and its value is between 0-1. After that, the result will be tested with ANOVA (Analysis of Variance) to determine the factors influencing the model. The results of ANOVA test can be seen in **Table 1**.

Table 1. Result of ANOVA from Binary Predictor Variables Scenario

Tested Factors	DB	P-Value	Hypothesis Result
n	9	0.000	Significant
α	2	0.641	Not Significant
p	2	0.000	Significant
Method	3	0.000	Significant
Interaction *n* & α	18	1.000	Not Significant
Interaction *n* & *p*	18	0.000	Significant
Interaction *n* & method	27	0.000	Significant
Interaction α & *p*	4	0.771	Not Significant
Interaction α & method	6	0.999	Not Significant
Interaction *p* & method	6	0.071	Not Significant

Based on the results on **Table 1**, the interaction between α and the other factors are not significant with *c-index* of the model. The main effect of *alpha* is also not significant with *c-index*. That means that model has the same performance pattern in all hazard functions. The quantity of data (*n*) and censored data proportion (*p*) significantly impact the *c-index*, which means the pattern of model for every *n* are not same for all *p*. Whereas the proportion of censored data (*p*) does not have significant interaction with method. However, the *p-value* was 0.071 and it closed to the significant level 0.05. Based on the previous explanation, sample size (*n*), censored data proportion (*p*) and method significantly impact the model performance, both on the main efects or the interaction of each factor on other factors. The impact of all the 3 factors will be explained in more detail based on the results obtained in **Figure 1**.

In **Figure 1**, six line charts show the model performance. The figure was build based on Random Survival Forest (RFS) and Survival Support Vector Machine (SSVM) with Additive Kernel Function (ADD), Radian Basic Function Kernel (RBF) and Linear Kernel Function (LIN). It has been conducted in every *n* and 3 type of censored data proportion. The first condition is when the censored data is dominant over the uncensored data (p = 0.8), second condition is when the censored data have same proportion with the uncensored data (p = 0.5), and the last one is when the uncensored data is dominant over the censored data (p = 0.2). Then the model was smoothed by using the moving average method to make it easier to see the patterns from analysis results. **Figure 1** shows that when the smoothing was not used, the average value of

c-index from all models that have been developed based on a combination of 3 parameters in the scenarios with binary predictor variables are in range 0.4571 and 0.6123. When the proportion of censored data dominates the uncensored data, the performance of 4 methods became unstable for all n since the line charts tend to fluctuate compared to when p=0.5 and p=0.2 for both models with moving average treatment or not.

It can be seen in **Figure 1**, that the more observations in the sample, the better the model performance will slowly be, especially when using the SSVM(ADD) method when creating the model. SSVM(ADD) is a method whose performance highly depends on the number of observations used. The position of the line chart shows this for the SSVM(ADD) method which is in rank 3 (when p = 0.8 & p = 0.5) and rank 4 (when p = 0.2) based on *c-index* when the amount of data is less than 800 observations, but following is the method with the best performance, especially when n is 1600 observations. Random Survival Forest (RSF) is the method that has the most consistent performance (shown from the unfluacted line on the smoothed chart, especially when the p = 0.8) , but it is also the method with the lowest performance compared to the other 3 methods in almost all of n, it shown from the unsmoothed chart, RSF has the smallest *c-index* in 21 of 30 combination n, p, α. Otherways, when the predictor variables used are all binary, the performance of Survival Support Vector Machine is better than the Random Survival Forest method. SSVM(RBF) is the method with the best model performance compared to the other three methods when n is small, while its performance will then be followed by SSVM(ADD) when n is large.

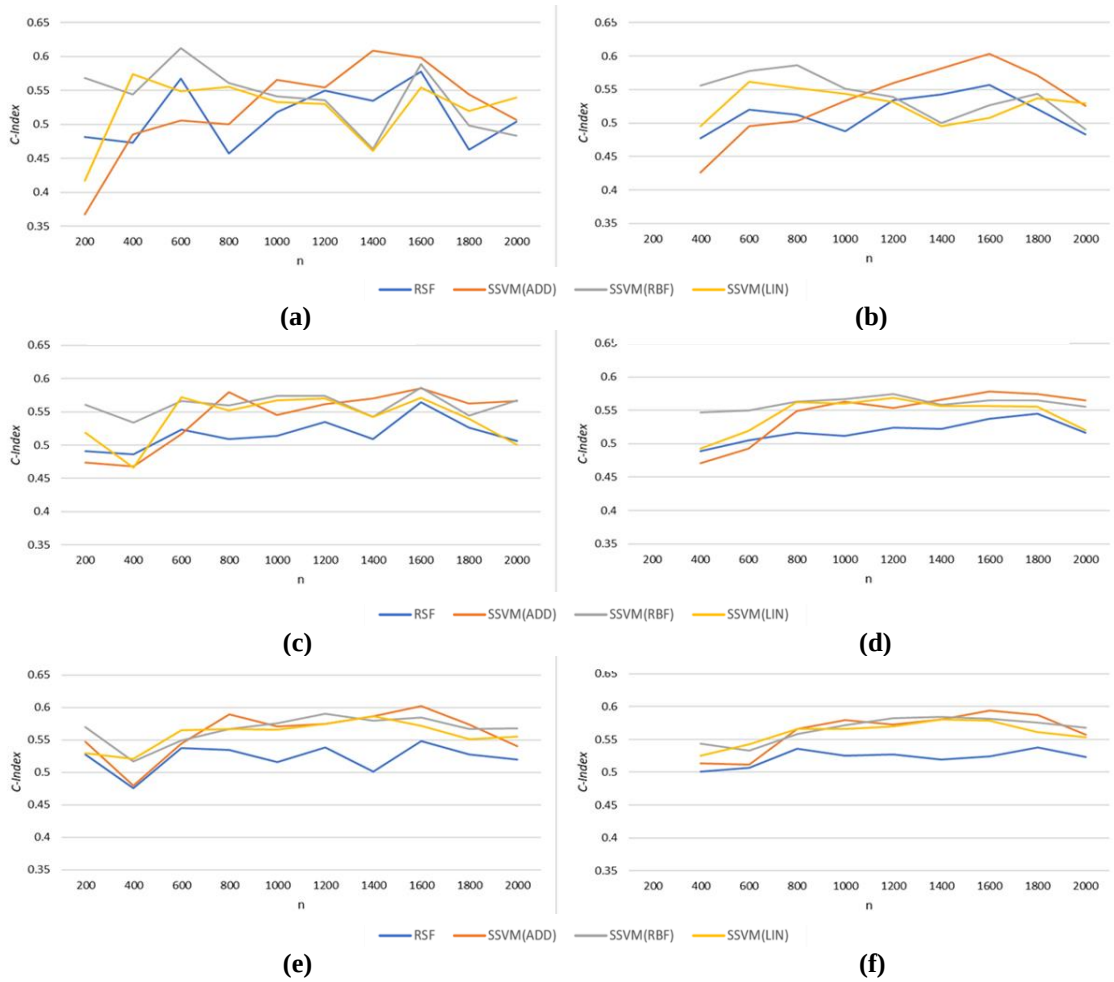

Figure 1. ***C-index*** **line chart of the model based on method and amount of data in the binary predictor variables scenario with (a) censored data proportion (p=0.8) without smoothing, (b) censored data proportion (p=0.5) using moving average, (c) censored data proportion (p=0.2) without smoothing, (d) censored data proportion (p=0.8) using moving average, (e) censored data proportion (p=0.5) without smoothing, (f) censored data proportion (p=0.2) using moving average.**

3.2Simulation using Mixed Predictor Variables

Three 3 hazard function scenarios are represented by α, 3 censored data proportion scenarios represented by p, 10 data quantity scenarios represented by n and modeled by 4 methods then evaluated with *c-index*. After that, the result will be tested with ANOVA (Analysis of Variance) to determine the factors that influence the model. The results of ANOVA test can be seen in **Table 2**.

Table 2. Result of ANOVA from Mixed Predictor Variables Scenario

Tested Factors	DB	p-Value	Hypothesis Result
n	9	0.175	Tidak Signifikan
α	2	0.046	Signifikan
p	2	0.002	Signifikan
Method	3	0.000	Signifikan
Interaction n & α	18	0.036	Signifikan
Interaction n & p	18	0.000	Signifikan
Interaction n & method	27	0.000	Signifikan
Interaction α & p	4	0.025	Signifikan
Interaction α & method	6	0.058	Tidak Signifikan
Interaction p & method	6	0.000	Signifikan

Based on the result in **Table 2**, the quantity of data (n) does not significantly impact model performance (based on *c-index*). That model performance tends to have same pathe Interaction between α and other factors have significant impact to the value of *c-index* of the model. Interaction α and method was insignificant, meaning the pattern of performance from all methods was the same for all hazard function. But, although the interaction α with the quantity of data (n) and censored data proportion (p) was significant, the *p-value* was so closed to 0.05. It can be concluded that α did not have a big impact on model performance. The method has significant interaction with the proportion of censored data (p). It means the pattern of the model was not same for all p. Based on *p-value*, the censored data proportion (p) and method have significant impact to model performance both as the main and the interaction effects. To see in more detail and compare the results with scenario of binary predictor variables, then a line diagram of 3 factors is formed, namely the method, the proportion of censored data (p), the amount of data (n) and the results can be seen in **Figure 2**.

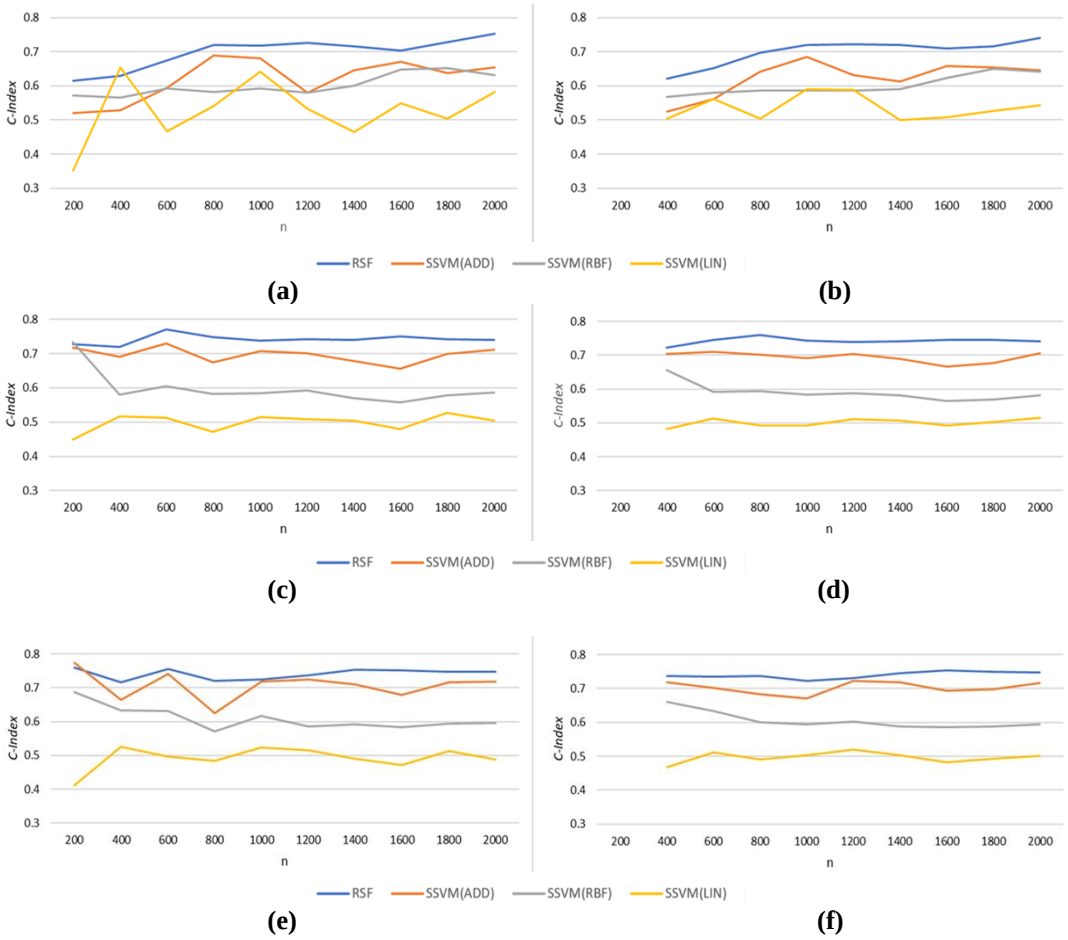

Figure 2. *C-index* line chart of the model based on method and amount of data in the mix predictor variables scenario with (a) censored data proportion (p=0.8) without smoothing, (b) censored data proportion (p=0.5) using moving average, (c) censored data proportion (p=0.2) without smoothing, (d) censored data proportion (p=0.8) using moving average, (e) censored data proportion (p=0.5) without smoothing, (f) censored data proportion (p=0.2) using moving average.

There is a line chart that shows the performance value of the model formed using Random Survival Forest (RSF), Survival Support Vector Machine (SSVM) with Additive Kernel Function (ADD), Radian Basic Function Kernel (RBF) and Linear Kernel Function (LIN) in every n with 3 condition of censored data proportion, such as when censored data proportion is bigger than uncensored data (p=0.8), censored data proportion is same with uncensored data (p=0.5) and when uncensored data proportion is bigger than censored data (p=0.2). Then the model was smoothed by using moving average method to make it easier to see the patterns from analysis results. Based on the results obtained from **Figure 2,** when smoothing is not carried out, the *c-index* values of all models formed in scenarios using mixed predictor variables range from 0.3512 to 0.7700. The range of performance values is wider than in scenarios using binary predictor variables.

For the binary predictor variable scenario, **Figure 2** show that when the proportion of censored data is dominant to uncensored data (p=0.8), the performance of all methods becomes more volatile compared to when the proportions are balanced (p=0.5) and the proportion of uncensored data is dominant to censored data (p=0.2). This volatility because uncensored data are observations with events, and in real conditions, this modelling is done to predict the time of an observation until an event occurs. In **Figure 2** can also seen that when the proportion of uncensored data is dominant (p = 0.2) and balanced (p = 0.5) against the censored data, the increase in the number of observations does not affect the performance of the model and applies equally to all methods, but when the proportion of censored data is dominant, the number of observations affects the model's performance.

In contrast to the results obtained in scenarios with binary predictor variables, in this case, the performance of Random Survival Forest is superior compared to the Survival Support Vector Machine with 3 kinds of Kernel Functions and applies to all n used in this study, then in terms of performance in proceed by SSVM(ADD), SSVM(RBF) and SSVM(LIN) methods. This is also in accordance with research that was conducted by **[19]** & **[20]** in real case with mixed predictor variables, the result was that RSF was better at predicting compared to CPH and SSVM.

4. CONCLUSIONS

The type of method used and the proportion of censored data affect the model's performance in both binary and mixed predictor scenarios. When the proportion of censored data is dominant, the model performance becomes less stable indicated by the level of diversity of the *c-index* value. When the predictor variables used are entirely binary, the number of observations affect the model performance especially in SSVM method, but when the predictor variables used are mixed, the number of observations will only have an effect when the proportion of censored data is dominant to uncensored data. The results of right-censored endurance data with scenarios using binary predictor variables show that the Survival Support Vector Machine (SSVM) using Radial Basic Function Kernel (RBF Kernel) performs better on small-size data. Meanwhile, when the data size becomes larger, the method that has the best performance is Survival Support Vector Machine using Additive Kernel (SSVM ADD). Random Survival Forest is a method that has the best performance for all conditions in mixed predictor scenarios.

REFERENCES

[1] J. In and D. K. Lee, "Survival analysis: Part I-analysis of time-to-event," *Korean J Anesthesiol*, vol. 71, no. 3, pp. 182–191, 2018.
[2] J. P. Klein and M. L. Moeschberger, *Survival analysis: techniques for censored and truncated data*, vol. 1230. Springer, 2003.
[3] P. Schober and T. R. Vetter, "Survival analysis and interpretation of time-to-event data: the tortoise and the hare," *Anesth Analg*, vol. 127, no. 3, p. 792, 2018.
[4] A. J. Turkson, "Perspectives on Hazard Rate Functions: Concepts; Properties; Theories; Methods; Computations; and Application to Real-Life Data," *Open Access Library Journal*, vol. 9, no. 1, pp. 1–23, 2022.
[5] R. Mokarram, M. Emadi, A. H. Rad, and M. J. Nooghabi, "A comparison of parametric and semi-parametric survival models with artificial neural networks," *Communications in Statistics-Simulation and Computation*, vol. 47, no. 3, pp. 738–746, 2018.
[6] C. J. K. Fouodo, I. R. König, C. Weihs, A. Ziegler, and M. N. Wright, "Support Vector Machines for Survival Analysis with R.," *R Journal*, vol. 10, no. 1, 2018.

[7] G. Xia and W. Jin, "Model of customer churn prediction on support vector machine," *Systems Engineering-Theory & Practice*, vol. 28, no. 1, pp. 71–77, 2008.
[8] R. Sudharsan, "SVM Based Churn Analysis for Telecommunication," *International Journal of Advanced Research in Engineering and Technology*, vol. 11, no. 6, 2020.
[9] H. Ishwaran, U. B. Kogalur, E. H. Blackstone, and M. S. Lauer, "Random survival forests," 2008.
[10] I. Ullah, B. Raza, A. K. Malik, M. Imran, S. U. Islam, and S. W. Kim, "A churn prediction model using random forest: analysis of machine learning techniques for churn prediction and factor identification in telecom sector," *IEEE access*, vol. 7, pp. 60134–60149, 2019.
[11] V. van Belle, K. Pelckmans, J. A. K. Suykens, and S. van Huffel, "Survival SVM: a practical scalable algorithm.," in *ESANN*, 2008, pp. 89–94.
[12] A. Hadanny *et al.*, "Machine learning-based prediction of 1-year mortality for acute coronary syndrome☆," *J Cardiol*, vol. 79, no. 3, pp. 342–351, 2022.
[13] C. Khotimah, S. W. Purnami, and D. D. Prastyo, "Additive survival least square support vector machines and feature selection on health data in Indonesia," in *2018 International Conference on Information and Communications Technology (ICOIACT)*, 2018, pp. 326–331.
[14] M. Saadati and A. Bagheri, "Comparison of Survival Forests in Analyzing First Birth Interval," *Jorjani Biomedicine Journal*, vol. 7, no. 3, pp. 11–23, 2019.
[15] 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, pp. 1–17, 2017.
[16] F. Wan, "Simulating survival data with predefined censoring rates for proportional hazards models," *Stat Med*, vol. 36, no. 5, pp. 838–854, 2017.
[17] H. H. Dukalang, "Analisis regresi Cox proportional hazard pada pemodelan waktu tunggu mendapatkan pekerjaan," *Jambura Journal of Mathematics*, vol. 1, no. 1, pp. 36–42, 2019.
[18] C. E. Smith and R. Cribbie, "Factorial ANOVA with unbalanced data: a fresh look at the types of sums of squares," *Journal of Data Science*, vol. 12, no. 3, pp. 385–403, 2014.
[19] S. Bai, X. Ji, B. Dai, Y. Pu, and W. Qin, "An Integrated Model for the Geohazard Accident Duration on a Regional Mountain Road Network Using Text Data," *Sustainability*, vol. 14, no. 19, p. 12429, 2022.
[20] S. Banerjee, S. Mitra, and L. O. Hall, "Analysis of MRI Biomarkers for Brain Cancer Survival Prediction," *arXiv preprint arXiv:2109.02785*, 2021.