Mapana – Journal of Sciences
2024, Vol. 23, Special Issue 3, 55-64
ISSN 0975-3303 | https://doi.org/10.12723/mjs.sp3.3

Comparison of Traditional Cox Proportional Hazard Model with Machine Learning Models for Survival Analysis in Predicting Risk of Death due to Heart Failure

Anuska Mukherjee* and Hemlata Joshi†

Abstract

Among the various types of cardiovascular diseases that kill millions of people worldwide each year, heart failure is a significant contributor to people's death. Risk of death due to heart failure can be influenced by various biological and anatomical factors in patients. In this study, we utilized a dataset comprising medical records of 299 patients, monitored over a specific period. While Cox Proportional Hazard (CPH) Model is the most conventional approach for analyzing survival data, machine learning (ML) models are also being used recently. The drawback of these ML methods is their inability to account for censoring. To incorporate censoring, especially right censoring, here in this article we have used Random Survival Forest Model, Gradient Boosted Model and Survival Support Vector Machine to predict the risk of death due to heart failure and compared their performances with traditional CPH model by Harrell's Concordance index and time dependent AUC. At the end of the study, it is observed that traditional CPH model outperforms rest of the ML techniques in predicting risk of death due to heart failure.

Keywords— Survival Analysis, Heart Failure, Cox Proportional Hazard Model, Random Survival Forest Model, Gradient Boosted Model, Survival Support Vector Machine

* Department of Statistics and Data Science, CHRIST (Deemed to be University), Bengaluru - 560029, India; anuska.mukherjee@stat.christuniversity.in; hemlata.joshi@christuniversity.in

I. Introduction

Cardiovascular diseases (CVDs) encompass a diverse range of conditions affecting the heart and blood vessels. These ailments pose a significant global health challenge, contributing to a substantial number of deaths each year. From coronary heart disease and cerebrovascular disease to congenital heart issues and peripheral arterial diseases, CVDs manifest in various forms, impacting individuals across diverse age groups and demographics. The chance of having heart failure depends on a lot of factors – age of an individual, usage of tobacco, blood pressure, presence of anaemia, presence of diabetes and many more. Understanding the complexities of cardiovascular diseases is crucial for developing effective prevention and treatment strategies.

In this study, we employed the dataset named 'Heart Failure Clinical Records Data Set' which was made publicly available by Ahmad and collaborators in July 2017[2]. The dataset contains medical record of 299 patients who were monitored during their follow-up period (April-December,2015). The most popular statistical method for dealing with this kind of data is Cox Proportional Hazard (CPH) Model (Cox,1972) [1]. The dataset used in this article was initially analysed by Ahmad et al. (2017) [2] where they used Cox regression to model patient mortality. The general survival pattern was analyzed using Kaplan-Meier survival plots, and the functional form of variables was assessed using Martingale residuals. Zahid et al. (2019) [3] utilized the same dataset to identify risk factors for male and female patients separately, employing CPH model. They assessed the predictive power of the model using the Concordance Index. However, these articles leave room for machine learning approaches. In recent studies it has been found that ML models such as k-Nearest Neighbor, Decision Tree, Random Forest Model, Naïve Bayes Algorithm, Support Vector Machines can also perform well in predicting survival of patients. Chicco and Jurman (2020) [4] analysed this dataset by applying several biostatistical and machine learning concepts and eventually showed that traditional statistical techniques and modern machine learning algorithms identifies the same factors as driving factors behind heart failure. Similar approaches were also taken by Maini et al. (2021) [5]. Though they used a different dataset, the ML methods applied for predicting heart disease of patients were similar.

But there is one major drawback of these ML models. These models take the event of death as the only dependent variable and classify the dependent variable based on other covariates. Whereas the survival analysis considers the time to a particular event as the dependent variable. Generally, in survival data we deal with censored data, i.e., data which is partially known. Censoring is encountered when the time until an event is not observed or not accessible for certain study participants, often due to factors like loss to follow-up or the event not occurring before the study concludes. To incorporate censoring into ML models and to improve predictive power we need to modify the techniques such that they are appropriate for survival analysis.

For this purpose, in this article we used Machine learning models combined with Survival Analysis, such as Random Survival Forests, Gradient Boosted Models and Survival Support Vector Machine, which model time-to-event, and compared their performances in predicting risk scores of patients with classical CPH model using Harrell's Concordance index and time-dependent AUC (mean area under the ROC curve).

II. Methodology

A. Data Description

The dataset we analysed here was elaborated by David Chicco [4] in January 2020 and donated to the UCI Machine Learning Repository. The dataset can be found here. The dataset has 12 clinical features for each patients including their follow up period. The study included a total of 105 women and 194 men, with ages spanning from 40 to 95 years *(Table I)*. Description of all the variables is given in *Table I*. We can visualize the nature of some of the important variables from *Fig. 1* and *Fig. 2*.

B. Model

1) Cox Proportional Hazard Model: The Cox proportional-hazards model, introduced by Cox in 1972 [1] is a commonly employed statistical model for examining the relationship between survival time and predictor variables. The general form of the CPH model is expressed as follows:

$$H(t, X) = H_0(t) \exp \sum_{i=1}^{n} b_i(x_i)$$

where t represents the survival time, (x_1, x_2,...., x_n) are n covariates, the function H (t) is known as the hazard function at time t for a specific set of covariates X. It is also interpreted as the instantaneous risk of undergoing the event of interest at time t, $H_0(t)$ denotes the baseline hazard function, representing the hazard function when all covariates are zero, (b_1, b_2,....,b_n) are the regression coefficients. This model is considered semi-parametric because $H_0(t)$ can assume various forms while the covariates enter the model in a linear fashion.

2) Random Survival Forest Model: The principle of Random Survival Forest (Ishwaran et al., 2008) [6] Model is similar to original random forest model (Breiman, 2001) [7] except it considers censoring information while forming the splitting rules. The algorithm is as follows: (1) Survival trees are grown for each bootstrapped dataset. (2) While splitting the nodes the goal is to split the nodes in such a manner that the survival differences across daughter nodes are maximized. (3) After growing the trees, an ensemble cumulative hazard estimate is computed by aggregating information from all the survival trees.

3) ***Gradient Boosted Model:*** Gradient Boosted Models (GBM) are somewhat similar to Random Forest Model since they both use ensemble methods. The difference is that Random Forests use bagging method while GBMs use boosting. In GBM, features from one model are fed into the next model in a sequential manner i.e., one model is built to reduce the errors present in the previous model. It uses decision trees as weak learners and keeps adding trees to the model to reduce loss function at each step.

Table I Description of Each Variable of The Dataset

Variable	**Description**	**Variable type**	**Range**
Patient's age	Age of the individual in years	Quantitative	40-95
Anaemic status	Indicates whether the patient has anaemia	Factor	0(no), 1(yes)

High BP	Indicates whether the patient has high blood pressure	Factor	0(no), 1(yes)
CPK level	Concentration of the CPK (Creatinine phosphokinase) in blood (measured in mcg/L)	Quantitative	23-7861
Diabetic status	Indicates whether the patient has diabetes	Factor	0(no), 1(yes)
Heart ejection fraction	Percentage of blood ejected from the heart with each beat	Quantitative	14-80
Gender	Specifies the gender of the patient	Factor	0(woman), 1(man)
Platelet Count	Platelet count in the blood (kiloplatelets/mL)	Quantitative	25.01-850.00
Serum creatinine	Creatinine concentration in the blood (measured in mg/dL)	Quantitative	0.50-9.40
Serum sodium	Sodium level in blood (mEq/L)	Quantitative	114-148
Smoking status	Indicates whether the patient smokes	Factor	0(no), 1(yes)
Follow-up Duration	Duration of follow-up in days	Quantitative	4-285
Death event	Indicates whether the patient experienced death during the follow-up period	Factor	0(survived), 1(dead)

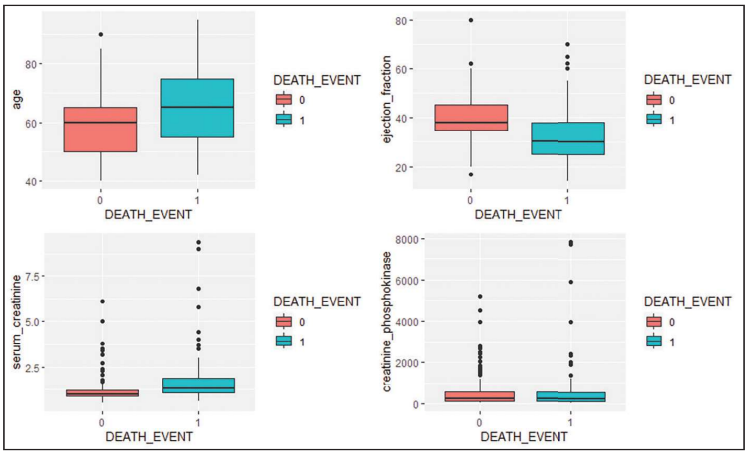

Fig. 1. Boxplots for numerical variables vs death event

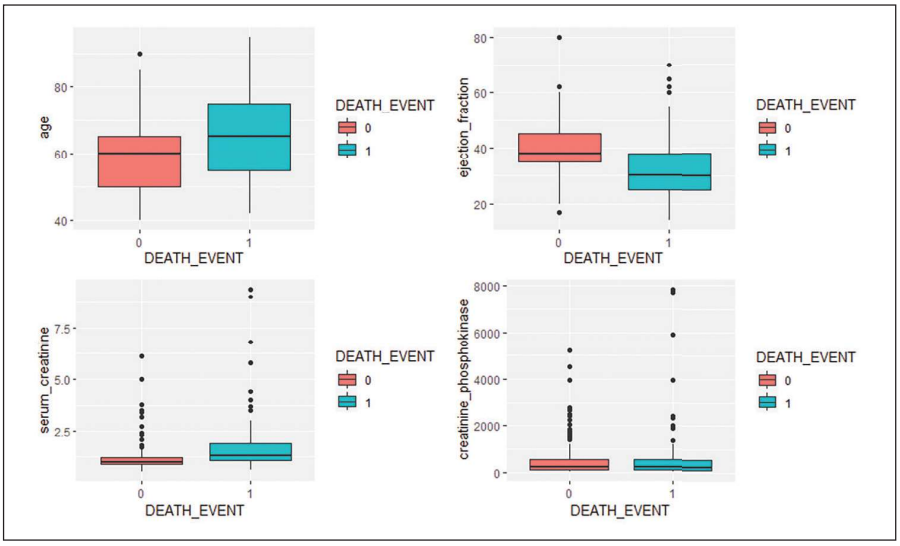

Fig. 2. Stacked Bar plots for categorical variables vs death event

4. ***Survival Support Vector Machine:*** To solve survival problems using support vector machine, three approaches have been proposed till now. The regression approach initially relied on the principles of original SVMs, aiming to identify a function that estimates observed survival times. Later Shivaswamy et al. [8] enhanced this approach by incorporating censoring considerations. In contrast, ranking approach [9],[10],[11] is focused on predicting the relative risk ranks among individuals rather than estimating specific survival times. The objective of this method is to maximize the concordance index (discussed later). The hybrid approach [12] is a combination of ranking and regression approach in survival SVM problems.

C. *Evaluation*

1) Concordance index: Harrell's Concordance Index, proposed by Harrell et al. in 1982 [13], is a statistical measure commonly used in survival analysis to assess the predictive accuracy of models. It evaluates how well a model distinguishes between higher and lower risk subjects, providing a valuable metric for the performance of survival prediction models. This index evaluates the rank correlation between the predicted risk scores generated by the model and the actual observed time points. It is calculated as

$$c = \frac{\#\ concordant\ pairs}{\#\ concordant\ pairs + \#\ discordant\ pairs}$$

Consider the i^{th} patient with time-to-event denoted as T_i and corresponding predicted risk score of η_i. Then for two patients i and j, the pair (i,j) is considered concordant if $\eta_i > \eta_j$ and $T_i < T_j$. Conversely, it is deemed discordant pair if $\eta_i > \eta_j$ and $T_i < T_j$. If the c-index is close to 0.5, it means that the predictions about which patient will live longer are not very accurate. However, if the c-index is close to 1, it indicates that the predictions are good at figuring out which of two patients is more likely to pass away sooner.

2) **Time-dependent AUC:** ROC stands for Receiver Operating Characteristic curve. It illustrates the relationship between the False Positive Rate (on the x-axis) and the True Positive Rate (on the y-axis) across various threshold values of risk, ranging from 0 to 1. This risk is assumed to be fixed over time. However, in medical studies, where patients are being monitored for a time period, it is natural that the risk of developing a disease or the risk of dying changes over time. It is possible that a person who has no risk of dying at the earlier stages of the study, may develop greater risk of death during the end of the study due to long follow-up period. Thus, in these cases, using ROC curve as a function of time is more appropriate. The mean area under the ROC curves (mean AUC) over different follow-up period will act as a measure of how good the model is.

III. Results and discussion

Table II shows the results of the CPH analysis. From the p-values we infer age, serum creatinine and heart ejection fraction are the three most statistically significant variables. Other significant variables are cpk and serum sodium. RSF model indicates serum creatinine, age, anaemic status, high BP and ejection fraction are the five most significant variables. Gradient boosted model suggests that serum creatinine, age, high BP and heart ejection fraction are the four most significant variables, whereas age, platelet count, CPK and heart ejection fraction are the significant variables according to survival SVM.

CPH model yielded a c-index value of 0.75 suggesting that it is a good model at predicting the ordering of patient's death. RSF model also produces a good value of c-index (0.7219) which is almost at par with traditional CPH model. Performance of Gradient boosted model and Survival SVM are not good enough in predicting risk scores as concordance index for these models are 0.684 and 0.542 respectively.

To examine how well the models performed at various time points we plotted AUC at different time point of the study (*Fig. 3*). As the plots suggest mean AUC is highest for CPH model.

From this study we see that traditional Cox proportional hazard model outperforms several machine learning models in predicting risk of death due to heart failure.

References

[1] Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, *34*(2), 187-202.

[2] Ahmad, T., Munir, A., Bhatti, S. H., Aftab, M., & Raza, M. A. (2017). Survival analysis of heart failure patients: A case study. *PloS one*, *12*(7), e0181001.

[3] Zahid, F. M., Ramzan, S., Faisal, S., & Hussain, I. (2019). Gender based survival prediction models for heart failure patients: A case study in Pakistan. *PloS one*, *14*(2), e0210602.

[4] Chicco, D., & Jurman, G. (2020). Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. *BMC medical informatics and decision making*, *20*(1), 1-16.

[5] Maini, E., Venkateswarlu, B., Maini, B., & Marwaha, D. (2021). Machine learning–based heart disease prediction system for Indian population: An exploratory study done in South India. *medical journal armed forces india*, *77*(3), 302-311.

[6] Ishwaran, H., Kogalur, U. B., Blackstone, E. H., & Lauer, M. S. (2008). Random survival forests. 841-860

[7] Breiman, L. (2001). Random forests. *Machine learning*, *45*, 5-32.

[8] Shivaswamy, P. K., Chu, W., & Jansche, M. (2007). A support vector approach to censored targets. In *Seventh IEEE international*

conference on data mining, 655-660.

[9] Van Belle, V., Pelckmans, K., Suykens, J. A., & Van Huffel, S. (2007). Support vector machines for survival analysis. In *Proceedings of the third international conference on computational intelligence in medicine and healthcare*, 1-8.

[10] Evers, L., & Messow, C. M. (2008). Sparse kernel methods for high-dimensional survival data. *Bioinformatics*, *24*(14), 1632-1638.

[11] Van Belle, V., Pelckmans, K., Suykens, J. A., & Van Huffel, S. (2008). Survival SVM: a practical scalable algorithm. In *ESANN*, 89-94.

[12] Van Belle, V., Pelckmans, K., Van Huffel, S., & Suykens, J. A. (2011). Support vector methods for survival analysis: a comparison between ranking and regression approaches. *Artificial intelligence in medicine*, *53*(2), 107-118.

[13] Harrell, F. E., Califf, R. M., Pryor, D. B., Lee, K. L., & Rosati, R. A. (1982). Evaluating the yield of medical tests. *Jama*, *247*(18), 2543-2546.

Appendix

Table II SUMMARY OF COX REGRESSION

Variable	Hazard Ratio	95% CI		z-value	p-value
		LL	UL		
Patient's age	1.04	1.02	1.06	4.00	<0.005
Anaemic status	1.49	0.92	2.41	1.62	0.11
CPK level	1.00	1.00	1.00	2.27	0.02
Diabetic status	1.05	0.63	1.74	0.18	0.85
Heart ejection fraction	0.95	0.92	0.97	-4.87	<0.005
High BP	1.46	0.90	2.37	1.54	0.12
Platelet count	1.00	1.00	1.00	-0.59	0.55
Serum creatinine	1.43	1.22	1.68	4.47	<0.005
Serum sodium	0.95	0.90	0.99	-2.32	0.02
Gender	0.76	0.43	1.33	-0.97	0.33
Smoking status	1.07	0.61	1.86	0.24	0.81

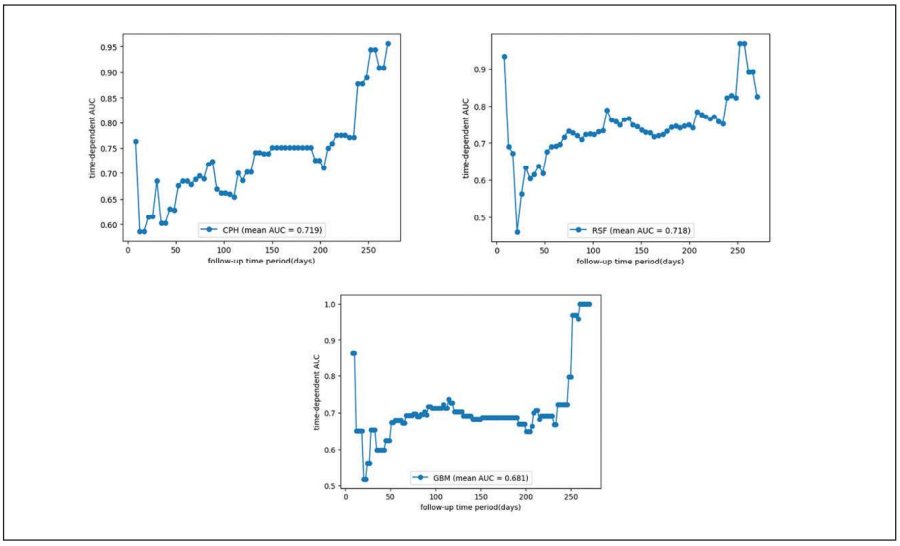

Fig. 3. Time Dependent AUC for CPH, RSF and Gradient Boosted Model