Classification Based on Event in Survival Machine Learning Analysis of Cardiovascular Disease Cohort

Shokh Mukhtar Ahmad (shokh.mukhtar@komar.edu.iq)
University of Sulaymaniyah

Nawzad Muhammed Ahmed
University of Sulaymaniyah

Research Article

Keywords: survival analysis, machine learning, logistic regression, SVM, tree descent, random forest.

Posted Date: April 18th, 2023

DOI: https://doi.org/10.21203/rs.3.rs-2681874/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.
Classification Based on Event in Survival Machine Learning Analysis of Cardiovascular Disease Cohort

Shokh Mukhtar Ahmad\textsuperscript{1,2*}, Nawzad Muhammed Ahmed\textsuperscript{1}

\textsuperscript{1} Department of Statistics and Informatics, College of Administration and Economics, Sulaymaniyah University, Sulaymaniyah, Kurdistan, Iraq

\textsuperscript{2} Department of Medical Laboratory, Komar University of Science and Technology Science, Sulaymaniyah, Kurdistan, Iraq

Abstract
The goal of this paper is to evaluate supervised learning classification models in a survival analysis problem of cardiovascular patients with a significant cured fraction. A sample include 919 patients referred to Sulaymaniyah Cardiac Hospital (including 365 females and 554 male) were followed up for a maximum of 650 days in 2021 to 2023. In the sample, 162 patients (17.6\%) during research time died. Cure fraction in this cohort was confirmed based on the Mahler and Zhu test (P<0.01), so for this case several machine learning classifications are applied to find the best patient status prediction procedure. Using several machine learning algorithms, the patients were classified into alive and dead. The results based on several indicators were almost the same. But in general, the random forest was identified as the best method in most of the indicators and of course with the Area under ROC = 0.934. The just weakness point of this method is that it did not work relatively well in the correct diagnosis of the deceased, and instead, the SVM with FP Rate = 0.263 performed better. After that, logistic and simple regression with Area under ROC = 0.911 and 0.909 showed better performance than others.

Keywords: survival analysis, machine learning, logistic regression, SVM, tree descent, random forest.

*Correspondence:
Shokh Mukhtar
shokh.mukhtar@komar.edu.iq
Introduction

Heart Cardiovascular disease (CVDs) is one of the leading causes of death worldwide. According to the World Health Organization, published in 2018, about 17.9 million people die annually due to this disease. This is while Middle East countries are experiencing a much worse situation than other parts of the world. The latest WHO report states about 19 percent of all deaths in Iraq are due to coronary heart disease. Based on these statistics, Iraq ranks 20th in the age-adjusted Death Rate 230.27 per 100,000 people.

In recent years, the treatment of many diseases, especially heart disease, has significantly improved, so the number of patients who will be alive has increased. The increase in the number of cured patients, which causes them to be censored from the data frame, requires new methods of survival analysis.

Survival analysis includes many methods to model and predict the probability of survival up to a certain time \( t \), \( P(T > t) \) where \( T \) is the survival time random variable;

\[
S(t) = P(T > t) = \int_t^\infty f(u)du = 1 - F(t) 
\]

(1)

To better estimate this probability, covariates variables such as \( (x_1, x_2, ..., x_k) \) are used in statistical models.

The most widely used survival analysis model is the Cox model. This pseudo-regression models and predicts the mentioned probability with the following function:

\[
\lambda(t) = \lambda_0(t) \exp(\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k) \\
= \lambda_0(t) \exp \left( \sum_{j=1}^k x_{ij} \beta_j \right) 
\]

(2)

In equation (2) the response variable is the hazard function \( \lambda(t) \), which assesses the probability that the event of interest (in this case, death) occurred at time of \( t \). The equation models this hazard as an exponential function of an arbitrary baseline hazard \( \lambda_0(t) \) when all covariates are null, and \( \beta \) is the regression coefficient of the covariates, \( (x_1, x_2, ..., x_k) \) [1].

On the other hand, hazard and survival function are related, so that:

\[
\lambda(t) = -\frac{d \log S(t)}{dt} = \frac{f(t)}{S(t)} 
\]

(3)

In equation 2, it can be seen that the logarithm of hazard function is a multiple regression on multi-dimensional covariates, but the very important different of this model and regression is due to especially of the data. In the survival analysis, the data frame consists of two groups of patients. One of the groups has experienced the event under study (which is death here), but the second group of patients were still alive at least during the studied time period, that’s why we call them sensors from the right. Therefore, the Cox proportional hazard (CPH) model is a special type of regression considering data of time-to-event data.

In the CPH model, partial likelihood is maximized for estimation and inference on the parameter \( \beta \):

\[
L(\beta) = \prod_i \left[ \frac{\lambda(y_i|x_i)}{\sum_{i'=1}^{n} \lambda(y_{i'}|x_{i'})} \exp \left( \sum_{j=1}^k x_{ij} \beta_j \right) \right] \\
= \prod_i \left[ \frac{\exp \left( \sum_{j=1}^k x_{ij} \beta_j \right)}{\sum_{i'=1}^{n} \exp \left( \sum_{j=1}^k x_{ij'} \beta_j \right)} \right] 
\]

(4)

After estimating the parameters in the CPH model, another important issue is choosing the variables to be included in the model. This topic has also been studied in many researches. For example, in [10], the lasso method for variable selection is proposed, in [5], smoothly clipped absolute deviation is presented; also in [8], an adaptive lasso method is also introduced.

Also [13], in their research, using a new method called “stacking”; they introduced the problem of survival analysis only as a classification problem. They also used several machine learning methods in addition to the Cox model in order to classify the subjects into two classes, alive and dead.

Although the most important issue in survival analysis is the probability of surviving until a particular time, predicting that a person belonging to the category of patients with their unique characteristics will survive or die during a certain time is also a very important issue in survival analysis. For this purpose, in this article, we have compared the results of different binary classification methods.

Since there are a wide range of classification methods, we have selected some of them in this research. Logistic regression is perhaps the most famous statistical method that has been frequently used in survival analysis. Also, machine learning methods such as random decision tree, J48 and random forest, have also been considered. In addition to them, the support vector machine (SVM) method is a very interesting method with the lowest risk in assigning subjects to groups, which is also one of the favorite techniques in survival analysis. In the section two, a brief introduction of each of these methods has been discussed. In section three, the data used in this research are introduced and practically each of the five classification methods are applied to them, and their results will be compared and discussed in section five.

Survival machine learning analysis

In clinical research, we deal often with high dimensional data, contain missing data and censored. Demographic status, physical conditions, and hospital interventions are all covariates that help us in predicting the patient’s condition during the study period. In addition to classical statistical methods such as regression, machine learning methods have attracted much attention from medical researcher due to their simplicity and sometimes more accurate predictions. Recently, many studies have compared machine learning methods in survival analysis [6,9].

Machine learning techniques as non-parametric and less complex are good alternatives to statistical methods. Users mostly like these methods because of their simplicity and the
results are often more accurate and close to reality.

The decision tree as one of them after introduce by [3], is a very flexible and easy-to-interpret model. Recently due to many researches, tree-based methods have improved significantly. Random forests technique [4] has become an excellent method in machine learning. Meanwhile, use of tree-based methods for survival analysis has drawn a lot interests. So that, much research has focused on tree building and dealing with censoring.

It is very important to remember that the purpose of survival analysis is to predict the survival time of patients in a cohort based on the available data. Although machine learning methods have been successful in achieving this goal in many ways due to the lack of complexities that exist in classical statistical models such as the Cox model, they also have some weaknesses. For example, in SVM survival analysis, predictions for survival time are made by ranking patients according to the probability of death. In other words, its results are obtained in the form of rank. This issue makes it difficult to compare its results in classic forms of survival analysis such as CPH [11,12]. Other techniques such as random forest have also been used in survival analysis. Random survival forests (RFS) landmarking as a nonparametric, machine learning alternative for obtaining dynamic predictions when there are complex or unknown relationships present is introduced. It requires little upfront decision-making and has comparable predictive performance and has preferable computational speed [7].

Of course, in this paper, several methods of machine learning will be used as binary classification methods in order to determine the survival or death of patients in during of treatment. This means that the problem of censoring will be just predicted variable. Their results will be compared use classification evaluation indices.

**Data analysis**

In this paper, a sample include 919 patients referred to Sulaymaniyyah Cardiac Hospital (including 365 females and 554 males) were followed up for a maximum of 650 days in 2021 to 2023. In the sample, 162 patients (17.6%) during research time died. Since the presence of cure fraction in these data was confirmed based on the Mahler and Zhu test (P<0.01), therefore, mixture cure models based on various probability distributions were used [1].

In this section, as a classification problem, two groups of survivors and dead during the follow-up period of the data have been discussed using of some variables. Covariates used in this research include demographic variables, selected blood sample markers and medical interventions presented in Figure 1.

<table>
<thead>
<tr>
<th>Covariates</th>
<th>Demographic variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gender</td>
<td>Age</td>
</tr>
<tr>
<td></td>
<td>Jobe</td>
</tr>
<tr>
<td></td>
<td>Location</td>
</tr>
<tr>
<td>Selected blood sample</td>
<td>Glucose</td>
</tr>
<tr>
<td>markers</td>
<td>Creatine</td>
</tr>
<tr>
<td></td>
<td>Urea</td>
</tr>
<tr>
<td></td>
<td>WBC</td>
</tr>
<tr>
<td></td>
<td>LYM</td>
</tr>
<tr>
<td></td>
<td>MID</td>
</tr>
<tr>
<td></td>
<td>GRA</td>
</tr>
<tr>
<td></td>
<td>HGB</td>
</tr>
<tr>
<td></td>
<td>RBC</td>
</tr>
<tr>
<td></td>
<td>MCV</td>
</tr>
<tr>
<td></td>
<td>PLT</td>
</tr>
<tr>
<td>Medical interventions</td>
<td>Doctor</td>
</tr>
<tr>
<td></td>
<td>Coronaryangio</td>
</tr>
<tr>
<td></td>
<td>Coronaryangiopci</td>
</tr>
<tr>
<td></td>
<td>CABG</td>
</tr>
</tbody>
</table>

Fig. 1 Covariates selected for patient classification
Classification results
After classifying the patients into two categories, the following results were obtained using different machine learning methods:

Table 1 Confusion matrices according to classification methods

<table>
<thead>
<tr>
<th>Classification</th>
<th>Random forest</th>
<th>Logistic</th>
<th>Regression</th>
<th>C4.5 tree</th>
<th>Random tree</th>
<th>SVM</th>
</tr>
</thead>
<tbody>
<tr>
<td>Patient Status</td>
<td>Alive</td>
<td>dead</td>
<td>Alive</td>
<td>dead</td>
<td>Alive</td>
<td>dead</td>
</tr>
<tr>
<td>Alive</td>
<td>748</td>
<td>9</td>
<td>729</td>
<td>28</td>
<td>739</td>
<td>18</td>
</tr>
<tr>
<td></td>
<td>740</td>
<td>17</td>
<td>691</td>
<td>66</td>
<td>742</td>
<td>15</td>
</tr>
<tr>
<td>dead</td>
<td>52</td>
<td>110</td>
<td>51</td>
<td>111</td>
<td>55</td>
<td>107</td>
</tr>
<tr>
<td></td>
<td>67</td>
<td>95</td>
<td>51</td>
<td>111</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Based on the results of the classifications in Table 1, the following indices have been obtained in order to evaluate different machine learning methods in classifying the final patient’s status.

Table 2 Classification indices according to classification methods

<table>
<thead>
<tr>
<th>Classifier</th>
<th>TP Rate</th>
<th>FP Rate</th>
<th>Precision</th>
<th>Recall</th>
<th>F-Measure</th>
<th>MCC</th>
<th>ROC Area</th>
<th>PRC Area</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random forest</td>
<td>0.934</td>
<td>0.267</td>
<td>0.933</td>
<td>0.934</td>
<td>0.929</td>
<td>0.757</td>
<td>0.934</td>
<td>0.961</td>
</tr>
<tr>
<td>Logistic</td>
<td>0.914</td>
<td>0.266</td>
<td>0.911</td>
<td>0.914</td>
<td>0.911</td>
<td>0.689</td>
<td>0.911</td>
<td>0.933</td>
</tr>
<tr>
<td>Regression</td>
<td>0.921</td>
<td>0.284</td>
<td>0.918</td>
<td>0.921</td>
<td>0.916</td>
<td>0.708</td>
<td>0.909</td>
<td>0.934</td>
</tr>
<tr>
<td>C4.5 tree</td>
<td>0.922</td>
<td>0.284</td>
<td>0.919</td>
<td>0.922</td>
<td>0.917</td>
<td>0.712</td>
<td>0.844</td>
<td>0.902</td>
</tr>
<tr>
<td>Random tree</td>
<td>0.855</td>
<td>0.356</td>
<td>0.855</td>
<td>0.855</td>
<td>0.855</td>
<td>0.500</td>
<td>0.843</td>
<td>0.878</td>
</tr>
<tr>
<td>SVM</td>
<td>0.928</td>
<td>0.263</td>
<td>0.926</td>
<td>0.928</td>
<td>0.925</td>
<td>0.737</td>
<td>0.833</td>
<td>0.885</td>
</tr>
</tbody>
</table>

As can be seen from the indices presented in Table 2, random forest has been better than other methods in all indices except for the FP Rate index. Contrary to that, SVM performed well in all indicators, especially in FP Rate, but the area under ROC was the lowest. Statistical methods such as logistic and simple regression are relatively balanced in terms of all indicators.
**Fig. 2** Receiver operating characteristic (ROC) plots according to classification methods

As can be seen in Figure 2, the ROC plots, the area under ROC curve, which shows the avoidance of false positive diagnosis and the tendency to correctly recognize positives, is greater than 0.5 for all the selected methods. The random forest has the greatest avoidance of false positives and the tendency to correctly recognize positives.
Conclusions
Although heart disease is one of the most widespread diseases and causes of death in the world and especially in the Middle East, the improvement of hospital and treatment services has led to the recovery of a significant part of these patients and their return to normal life. In a time-to-event problem, in order to predict the survival probability of each patient until a certain time in such conditions, it requires more complete models than Cox models, which are called Cure models. On the other hand, machine learning has interested the attention of researchers in this field as simpler methods with reality results. The output of survival machine learning is based on the rank of patient’s death. In this research, the survival problem is reduced just to prediction during the follow-up so that the results of several machine learning methods can be checked in faced with such a situation. In the results, we saw that random forest performed better based on all criteria except false positive rate. The reason for this is the high risk of this method in the problem of survival detection, which has led to misdiagnosis of some dead patients as cured. Contrary to that, since SVM is a minimum risk classification method in determining separation vectors, it has acted more conservatively, although this conservatism in detecting survival has the lowest false positive rate among other methods, but due to the problem with presence of a significant cured fraction of patients has caused this method to have the worst performance in the important indicator of the area under ROC. On the other hand, the presence of many variables related to death in medical issues has caused classical statistical methods such as logistic and simple regression to be in relatively ideal conditions in all indicators after random forest. In general, since the ROC curve indicates the avoidance of wrong diagnosis and the tendency towards the correct diagnosis of patients' lives, it was taken into consideration. Based on this criterion, random forest performed best and SVM performed worst. Therefore, conservative methods such as SVM are not recommended in problems like this, which has a significant survival expectation.

Declarations

Competing interests
The authors declare that they have no conflict of interest.

Ethics approval and consent to participate
This study was design based on Helsinki declaration for ethical principles of human research. All experiments were performed in accordance with relevant guidelines and regulations based on Helsinki declaration. A written informed consent was collected from each participant before data collection. Ethics Approval was obtained from the University of Sulaymaniyah Ethical approval committee.

Consent for publication
Not Applicable.

Availability of data and materials
The datasets of the current study are available from the corresponding author on reasonable request.

Funding
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.
Authors’ contributions

Shokh Mukhtar Ahmad is involved in research concept generation, designing the methods, responsible for collecting and analyzing the data from a large cohort of cardiovascular disease patients, using various statistical and machine learning techniques, writing manuscript and revising the manuscript. She conducted a comprehensive literature review on the topic of survival machine learning analysis in cardiovascular disease cohorts, which formed the basis for the study’s methodology. She also collaborated with other researchers to write up the findings in a clear and concise manner, and contributed significantly to drafting and revising the manuscript.

Nawzad Muhammed Ahmed is involved in writing proposal, responsible for collecting and analyzing the data from a large cohort of cardiovascular disease patients, using various statistical and machine learning techniques. He collaborated with other researchers to write up the findings in a clear and concise manner, and contributed significantly to drafting and revising the manuscript.

All the authors have read and approved the manuscript.

Acknowledgements
First and foremost, we would like to sincerely thank the respected doctors and staff of Sulaymaniyah Cardiac Hospital for their cordial guidance and constant supervision in enabling us with the access to the necessary information related to this research.

Authors' information
Corresponding Author:
My name is Shokh Mukhtar Ahmad, I am PhD student in (Statistics), I am also Lecturer at the Komar University for Science and Technology.

Second Author:
Professor Dr. Nawzad Muhammed Ahmed, Professor at the University of Sulaymaniyah Statistics and Informatics department. He also supervised the PhD thesis from the corresponding author.

References
12. Van Belle V, Peckmans K, Van Huffel S, Suykens JAK. Improved
Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- All.data.xlsx