

Machine Learning-Enhanced Survival Analysis: Identifying Significant Predictors of Mortality in Heart Failure

Heejeong Jasmine Lee¹, Sang-Sun Yoo^{2*}, and Kang-Yoon Lee^{3*}

¹ College of Information and Communication Engineering, Sungkyunkwan University, Suwon 16419, Korea
SKAICHips, 11 Hyowon-ro 266beon-gil, Gwonseon-gu, Suwon, Korea
[e-mail: heejl@skku.edu]

² SKAICHips, 11 Hyowon-ro 266beon-gil, Gwonseon-gu, Suwon, Korea
[e-mail: syoo@skaichips.co.kr]

³ College of Information and Communication Engineering, Sungkyunkwan University, Suwon 16419, Korea
SKAICHips, 11 Hyowon-ro 266beon-gil, Gwonseon-gu, Suwon, Korea
[e-mail: klee@skku.edu]

*Corresponding author: Sang-Sun Yoo, Kang-Yoon Lee

*Received March 21, 2023; revised April 3, 2024; revised August 25, 2024;
accepted September 8, 2024; published September 30, 2024*

Abstract

State of the art machine learning methods can enhance the analysis of clinical data and improve the ability to predict patient outcomes because data collected from clinical records, such as heart failure mortality studies, are often high dimensional, heterogeneous and give challenges to traditional statistical analysis techniques. To address this challenge, this study conducted a survival analysis based on a dataset of 299 patients with heart failure, using Python libraries. Cox regression was used to model and analyse mortality, and to find which features are strongly associated with this outcome. The Kaplan-Meier survival curve approach was used to show the patterns of patient survival over time. The analysis showed that age, ejection fraction, and serum creatinine level were significantly ($p \leq 0.001$) associated with mortality. Anaemia and creatinine phosphokinase also reached statistical significance (p -values 0.026 and 0.007, respectively). The Cox model showed good concordance (0.77) with the data, suggesting that the identified variables are useful for predicting mortality in patients with heart failure.

Keywords: machine learning, biomedical informatics, heart failure, cardiovascular heart diseases, survival analysis.

1. Introduction

Hear failure(HF) is a chronic and progressive condition in which the heart muscle is unable to meet the body's demands for oxygen and blood through adequate pumping [1].

Medical conditions such as diabetes, high blood pressure, and kidney disease can impair the prognosis of heart failure. The risk of heart failure increases with age, resulting in worse prognosis in older adults. In addition to these factors, several other variables may act as predictors of heart failure-related mortality. For example, reduced left ventricular ejection fraction (LVEF), which measures how efficiently the heart pumps blood, with a lower LVEF value indicates a higher risk of death [2]. Previous hospitalization due to heart failure, uncontrolled high blood pressure, anemia, decreased red blood cell count, sleep apnea (a disorder characterized by a brief respiratory arrest during sleep), and smoking are strong predictors of heart failure-related mortality [3].

Survival analysis is a statistical technique used to predict when an event (e.g., death) will occur in a given population [4]. Survival analysis uses censored data (i.e., time-to-event data), which are common in medical studies, to show that an observed event has not yet occurred during the observation window [5]. In patients with heart failure, survival analysis can be used to identify predictors of mortality, including age, comorbidities, heart function, and treatment variables [2, 6]. One of the most common factors that can predict mortality in patients with heart failure is old age, hypertension, kidney disease, diabetes, left ventricular dimensions, and ejection fraction [7]. Patients with heart failure can improve their survival by using medications such as aldosterone antagonists, angiotensin-converting enzyme inhibitors, and beta-blockers. In addition, elevated levels of B-type natriuretic peptides and brain natriuretic peptides have been identified as strong indicators of mortality in patients with heart failure [8]. By providing valuable insight into the predictors of mortality in patients with heart failure, survival analysis can help develop effective management programs aimed at reducing the risk of death.

The goals of this study were to model the survival probability of patients with heart failure over time and to identify the predictor variables that are strongly associated with mortality in this patient population. This study used a publicly available heart failure dataset from the UCI machine learning repository [9] to reproduce the findings of Ahmad et al. [10] using state-of-the-art Python libraries to implement survival analysis.

2. Related work

Since heart failure is a highly prevalent and deadly disease worldwide, there is a considerable amount of literature devoted to predicting the incidence and prognosis of heart failure. Various methodologies and datasets have been used by researchers in this field. Here, we briefly review the relevant studies implemented in the field.

Ahmad et al.[10] analyzed a dataset on 299 heart failure patients in Faisalabad, Pakistan to find significant risk factors for mortality using Cox regression and Kaplan-Meier plot. Factors considered in the analysis included age, ejection fraction, creatinine, anemia, blood pressure, and others. The results were validated via bootstrapping with a nomogram constructed for the graphical prediction of survival probability. The significant risk factors were age, renal impairment, blood pressure levels, anemia, and ejection fraction.

Subsequently, Zahid et al.[11] developed and evaluated survival prediction models for patients with left ventricular systolic dysfunction using gender-specific risk factors. A lasso approach was used to identify the best predictors, and separate models were built for all

patients, male patients, and female patients. The study found differences in the survival prediction models between male and female patients, with smoking, diabetes, and anemia being non-informative predictors for males and ejection fraction, sodium, and platelet count being non-informative for females. The results showed that the selected models performed as well as the overall models in terms of predictive performance. Further studies are required to confirm these differences.

The two previously mentioned studies produced results by using traditional biostatistical methods. Subsequently, Chicco and Jurman employed data mining and machine learning methods. Machine learning classifiers were utilized to analyze a dataset of 299 patients with heart failure to predict survival and identify important risk factors. As shown by the two feature ranking methods, the serum creatinine level and ejection fraction were identified as the two most relevant features. The results showed that using only these two factors allows for more accurate survival prediction than using the entire dataset. This finding may impact clinical practice by becoming a new tool for physicians to predict survival in heart failure patients [12, 13].

Machine learning methods are becoming increasingly popular in the medical field owing to their ability to handle complex multidimensional datasets that are now available in electronic medical records. For these reasons, more recent publications (2020-2023) on survival analysis for heart failure patients tend to favor machine learning over traditional biostatistics. Examples of such studies are discussed below to provide a background for this direction in medical research.

In 2020, Adler et al. [14] utilized a machine learning algorithm to determine the correlations between patient attributes and mortality in patients with heart failure. A boosted decision tree algorithm was used to train a model using a subset of patient data based on the high or low risk of mortality. This model generated a risk score with eight variables that effectively distinguished between low and high-risk mortality with an AUC value of 0.88. The score was validated in two separate HF populations, where the AUC outperformed other risk scores, and demonstrated its potential in evaluating patients with HF. Guo et al. [15] reviewed 335 related papers on machine learning and deep learning for heart failure prediction, identified through a search of the PubMed database. Machine learning models can be used to identify patients with HF and assess their risk of readmission and mortality. The authors suggested that novel techniques are needed for integrating diverse data and improving predictive accuracy. The various attributes of clinical electronic health record (HER) data pose challenges. However, machine learning models have the potential to revolutionize prediction accuracy for personalized prevention, treatment, and management of patients with HF.

A systematic literature review was conducted in 2020 to evaluate the resilience of prediction models in assessing the risk of heart failure (HF). Forty relevant publications were identified and assessed for statistical approach, validation, risk of bias (ROB) and common variables. A total of 58 models were examined and 55 results were evaluated and they included predictors such as blood urea nitrogen, brain natriuretic peptide, N-terminal prohormone, creatinine, and other variables [16].

Yazdani et al. (2021) introduced a method for predicting heart disease using Associative Rule Mining (ARM). This approach utilizes Weighted Associative Rule Mining to analyze the UCI machine learning dataset for heart disease. This study achieved a confidence score of 98% in predicting heart disease by demonstrating the potential of machine-learning techniques to enhance clinical decision-making processes [17].

Another study conducted by Ishaq et al. [18] in 2021 analyzed the same dataset to predict heart disease survivors using machine learning and data mining. A dataset of 299 heart failure patients admitted to a hospital was analyzed to identify significant features and effective techniques to enhance accuracy. Nine classifiers (DT, AdaBoost, LR, SGD, GBM, RF, ETC, SVM and G-NB) were employed and imbalanced class problems were handled using SMOTE. The findings indicate that ETC outperformed other models and achieved an accuracy of 0.9262 with SMOTE and learned about the top-rated features selected by RF.

In 2021, Newaz et al. [19] developed a decision support system using clinical records and laboratory tests to accurately predict heart failure survival. They utilized a heart failure dataset from Pakistan and machine learning techniques to identify risk factors and improve their accuracy. The authors utilized feature selection techniques to identify key risk factors and achieved a G-mean of 76.83% and sensitivity of 80.21%, which is higher than that reported in previous studies.

In 2021, Kavifah et al. [20] utilized data mining methods, such as regression and classification, to extract valuable insights from the Cleveland heart disease dataset. Random Forest, Decision Tree, & Hybrid models were applied for the prediction. The results showed an accuracy of 88.7% using the hybrid model. A hybrid model of the Decision Tree and Random Forest was used to forecast heart disease using user-input parameters.

In 2021, a study was conducted to compare the effectiveness of machine learning (ML) approaches and traditional statistical models (CSMs) in predicting readmission and mortality among patients with heart failure. A systematic literature search was performed and 20 articles comprising 686,842 patients were analyzed. ML methods include decision trees, random forests, support vector machines, regression trees, neural networks, and Bayesian techniques, whereas CSMs include Cox, logistic, or Poisson regression. In the majority of studies examining the prediction of readmission and mortality risk in patients with heart failure, ML demonstrated superior discrimination compared with CSMs. Nonetheless, one drawback of ML studies is that most of them lack external validation, and calibration is seldom evaluated. The study recommended that ML-based investigations be assessed based on clinical quality standards for prognosis research [21].

In 2022, Almazroi [22] used a standard dataset and benchmark algorithms to evaluate the performance and found that decision trees performed better than logistic regression, SVM, and artificial neural networks, with 14% higher accuracy. Unlike other studies, this study found that artificial neural networks are not as effective as decision trees or support vector machines.

In 2022, Sabor et al. [23] used nine classifiers to predict the occurrence of heart disease using the heart disease dataset, achieved results with hyperparameter tuning and data standardization. A study showed improvement in accuracy with hyperparameter tuning and data standardization using classifiers such as AB, CART, ET, LDA, LR, MNB, RF, SVM, and XGB. The accuracy of the prediction classifier was improved with hyperparameter tuning. The highest accuracy achieved was 96.72% using SVM.

In summary of this literature review, machine learning and artificial intelligence have great potential in medicine, including heart failure diagnosis and management. Current applications include new diagnostic approaches, patient classifications, and improved prediction capabilities. This paper provides an overview of machine learning for clinicians and evaluates current applications in heart failure. Many methods show potential but require further evaluation and validation before being incorporated into common practice. Despite challenges, machine learning has the potential to lead to more accurate diagnoses, precise treatments, and better patient outcomes [15, 24].

In contrast to our research, which focuses on identifying key predictors of cardiovascular disease, a concurrent study leveraged the University of California Irvine (UCI) machine learning dataset to explore the influence of cardiovascular disease on mortality rates and the complexities associated with its prediction. The simulation conducted in MATLAB 2020b revealed that their ensemble model achieved 96% accuracy [25].

3. Data Description

The dataset included 299 patients with HF comprising 194 men and 105 women, as collected by Ahmad et al. [10]. The dataset had 13 features: age, anaemia, creatinine phosphokinase (CPK), diabetes, ejection fraction (EF), high blood pressure, platelets, serum creatinine, smoking, serum sodium, sex, time and DEATH_EVENT. After loading the dataset, we checked for missing values and found missing values in any of the columns. The target feature is DEATH_EVENT. The “DEATH_EVENT” column has been renamed to “died” to simplify referencing in analysis.

For easier processing and visualization, we divided the columns into categorical and numeric columns.

- Categorical Columns: 'anaemia', 'diabetes', 'high_blood_pressure', 'sex', 'smoking', 'died'
- Numerical Columns: 'age', 'creatinine_phosphokinase', 'ejection_fraction', 'platelets', 'serum_creatinine', 'serum_sodium', 'time'

Count plots were generated for categorical columns, which were subdivided into “died” column (i.e., survival outcomes). Kaplan-Meier estimates were plotted for categorical and continuous variables to analyze the probability of survival over time.

The user-defined function `km_fits` was defined to perform Kaplan-Meier survival analysis based on categorical and numerical variables. Kaplan-Meier estimators were applied to both categorical and numerical data to determine the probability of survival based on different variables.

High blood pressure, anaemia, diabetes, smoking, sex, and DEATH_EVENT were categorical variables, while the rest were numerical variables. Patients were 40-94 years old. The follow-up duration ranged from 4 to 285 days, with a mean of 130.2 days. The patients who survived to the end of follow-up numbered 203 (68%), and 96 (32%) died.

The DEATH_EVENT column indicates the observation pertained to whether death had occurred. The variable DEATH_EVENT takes a value of 1 if the event is observed, indicating that the patient has passed away. Conversely, if data were censored, meaning the patient survived until the end of the follow-up period (which varied among patients), DEATH_EVENT took the value of 0. The time variable represents the time to event, that is, the duration of the patient's life prior to death or censorship. Censorship means that the observation ended without any observed event (i.e., death).

Fig. 1 shows a histogram of patient survival times before death or censoring (i.e., end of observation). This distribution is rather complicated, with several peaks.

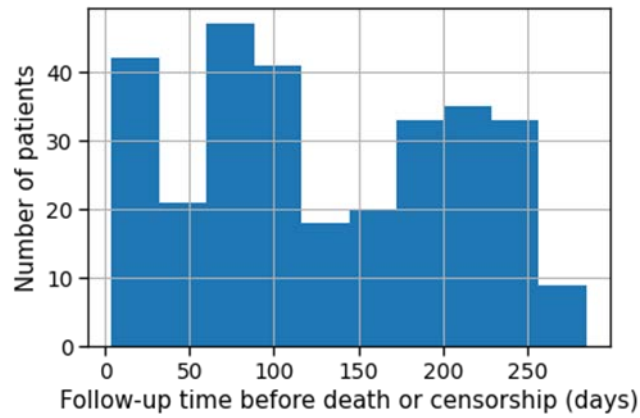


Fig. 1. Histogram of the survival time.

Fig. 2 shows the distributions of survival status (living or deceased) as a function of several relevant variables in the dataset. The variables selected for analysis were anemia, diabetes, high blood pressure, sex, and smoking. These variables were selected based on their known associations with heart failure outcomes. By analyzing these variables, we aimed to understand their contribution to mortality in patients with heart failure. The distributions help identify patterns that can inform clinical decision-making and patient management strategies.

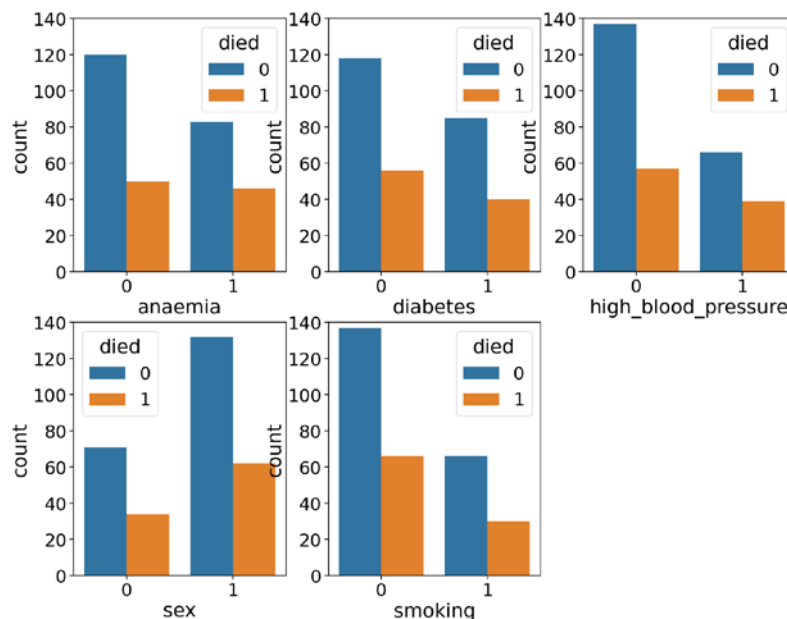


Fig. 2. Distribution of survival status as function of several relevant variables in the dataset.

In addition to these detailed views of the data, **Fig. 3** provides a more generalized picture using a heatmap of Pearson correlations between variables in the dataset. The heatmap included the following variables: age, anemia, creatinine phosphokinase, diabetes, ejection fraction, high blood pressure, platelets, serum creatinine, serum sodium, sex, smoking, time, and DEATH_EVENT. There was a moderate positive correlation between age, anaemia, creatinine phosphokinase, diabetes and DEATH_EVENT, suggesting these could be significant predictors of mortality in patients with heart failure.

These correlations help to understand the relationships between different clinical variables and their impact on patient outcomes. By identifying and discussing these correlations, we can better select and prioritize predictor variables for our survival analysis. This enhances the robustness of our predictive models and facilitates clinical decision making.

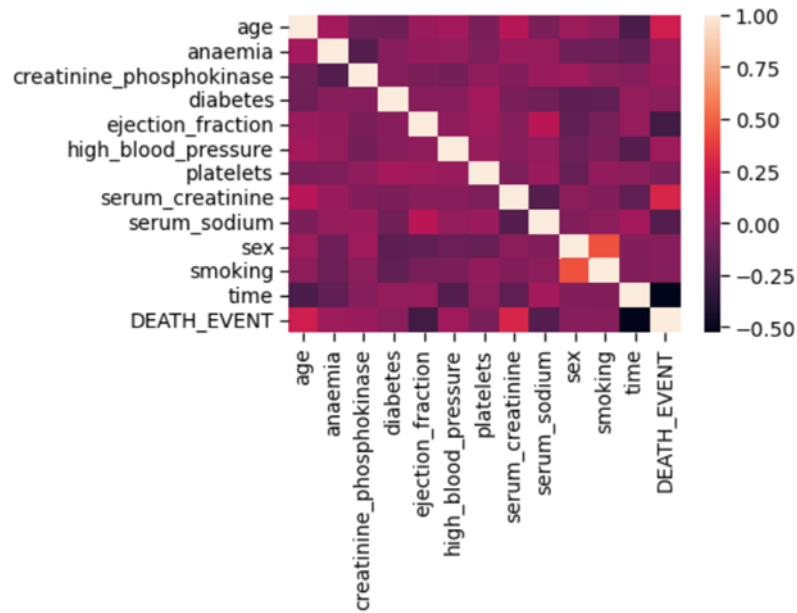


Fig. 3. Heatmap of Pearson correlations between variables in the dataset.

4. Methodology

There is a conceptual difference between survival analysis and classification models regarding the handling of time variables. For a classification model, the time variable is treated as a feature in the same manner as other covariates, whereas for survival analysis, it is used to estimate a hazard function for the outcome events (death in this case). The hazard function and the corresponding survival function mathematically describe the probability of survival longer than a particular time (see Eq. 1).

$$S(t)=P(T>t) \quad (1)$$

4.1 Cox proportional hazards

The Cox proportional hazards model (also called Cox regression, CoxPH, or Cox's model) has been the most commonly employed method for examining the association between a patient's survival and potential risk factors is known as survival analysis [5]. The h_i value depends on the predictor variables (x) and baseline hazard function h_0 . A convenient feature of this modelling method is that the baseline hazard function h_0 does not need to be explicitly modelled or estimated, and the modelling task involves only estimating the β parameters for the effects of predictor x . In other words, there are no assumptions regarding the form of the baseline hazard function, and predictors x have a multiplicative (proportional) effect on the hazard by the exponential function (for example, Eq. 2 shows an example of two predictors).

$$h_i(t) = h_0(t) e^{\beta_1 * x_1 + \beta_2 * x_2} \quad (2)$$

The Cox proportional hazards model, which assumes that hazard ratios are constant over time, was applied to the data to predict mortality. We tested this proportionality assumption using Schoenfeld residuals. The model considered 11 potential risk factors, including age, anemia, serum creatinine levels, and ejection fraction. One limitation of this model is that it does not consider potential time-varying covariates.

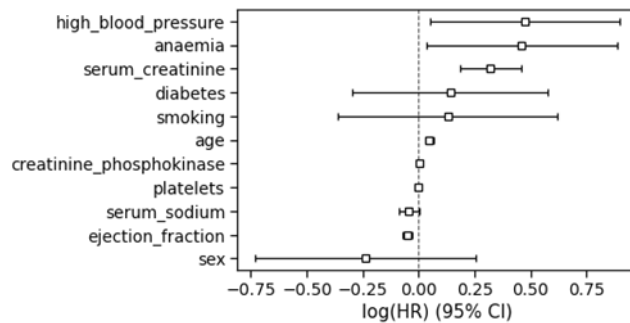


Fig. 4. Visualization of risk factors for mortality in HF patients.

Fig. 4 shows a visualization of the initial fit of the Cox model to the HF data. Diagnostics of this model showed that the variable `ejection_fraction` failed the test for the proportional hazards assumption (p -value = 0.0127). Consequently, we reparametrized this variable using splines (linear, quadratic and cubic polynomial terms) and the modified model passed the diagnostic procedure. The output is summarized in **Fig. 5**. The analysis showed that age, ejection fraction, and serum creatinine levels were significantly ($p \leq 0.001$) associated with mortality. The most significant term for the ejection fraction, was the linear polynomial, whereas the quadratic and cubic terms used to describe this variable did not reach statistical significance. However, these higher-order terms are helpful in satisfying the model's assumption of proportional hazards. Anaemia and creatinine phosphokinase levels also reached statistical significance (p -values 0.026 and 0.007, respectively).

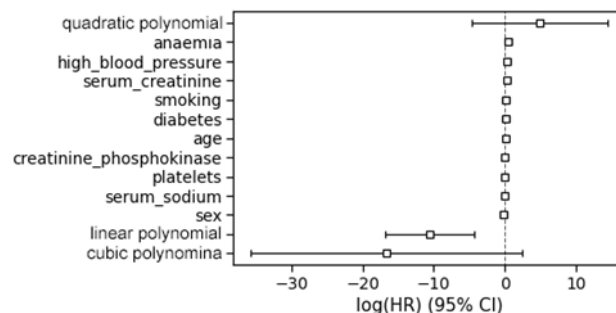


Fig. 5. Visualization of the risk factors for HF patient mortality in an improved Cox model where the ejection fraction was described by splines.

4.2 Kaplan-Meier model

The Kaplan–Meier model was used to estimate survival probability over time. In medical studies, it is often used to predict patient survival within a specific period after treatment. It is also commonly used by life insurance companies to plan life insurance products.

The Kaplan-Meier estimator is defined by the following formula (see Eq. 3). The Kaplan-Meier plot is a visual representation of the Kaplan-Meier model.

$$\hat{S}_t = \prod_{i=1}^t \left(1 - \frac{d_i}{n_i}\right) \quad (3)$$

4.3 Rationale for methodology used

In summary, combining the CoxPH and Kaplan-Meier methods provides both statistical accuracy and intuitive visualization to provide a comprehensive analysis of survival data in patients with heart failure. CoxPH identifies significant predictors of mortality, while Kaplan-Meier shows the survival experience of different groups over time. This combined approach provides detailed modeling with clear visual insights to enhance the analysis.

5. Experiment and Discussion

5.1 Cox proportional hazards

The prevailing approach for assessing the effectiveness of a CoxPH is to use the concordance index (i.e., C-index). The C-index for a survival model is equal to the AUC, which represents the area under the Receiver Operating Characteristic (ROC) curve. The model's discrimination ability was assessed using ROC curves, showing an AUC of 0.81 at 250 days and 0.77 at 50 days in the paper of Ahmad et al.[10]. This indicates that the model correctly identified the death event for 81% of patients within 250 days and 77% of patients within 50 days. Additionally, our study identified some variables, such as diabetes and smoking status, as non-significant predictors of mortality ($p=0.55$ and $p=0.567$, respectively). This is unexpected given the established links between these factors and cardiovascular health. One possible explanation is the limited sample size, which may have reduced the power to detect these associations. Additionally, the variability in treatment regimens and adherence among patients could have contributed to these non-significant results.

Table 1 shows the coefficients of the fitted CoxPH model, which represent the effect of each predictor variable on patient mortality. The presence of positive coefficients in the variables is indicative of a higher risk of mortality (i.e., decreased survival times) and vice versa for variables with negative coefficients (i.e., prolonged survival times). Age appears to play a significant role in increasing the mortality rate. The linear polynomial term for ejection fraction was also highly significant. As can be seen in the table, CoxPH can use multiple predictors (i.e., variables), whereas the Kaplan-Meier model can use one predictor.

Table 1. Summary of Cox proportional hazards model.

| variable name | coef | exp(coef) | exp(coef) lower 95% | exp(coef) upper 95% | z | p |
|---|---------|-----------|------------------------|------------------------|--------|---------|
| age | 0.049 | 1.05 | 1.03 | 1.07 | 5.1 | <0.0005 |
| anaemia | 0.492 | 1.636 | 1.061 | 2.523 | 2.228 | 0.026 |
| ejection_fraction: linear polynomial | -10.585 | 0 | 0 | 0.013 | -3.311 | 0.001 |
| ejection_fraction: quadratic polynomial | 4.845 | 127.116 | 0.009 | 1.76E+06 | 0.996 | 0.319 |
| ejection_fraction: cubic polynomial | -16.647 | 0 | 0 | 10.643 | -1.716 | 0.086 |
| creatinine_phosph okinase | 0 | 1 | 1 | 1 | 2.689 | 0.007 |
| diabetes | 0.134 | 1.144 | 0.736 | 1.778 | 0.597 | 0.55 |
| high_blood_press ure | 0.357 | 1.429 | 0.93 | 2.195 | 1.631 | 0.103 |
| platelets | 0 | 1 | 1 | 1 | -0.435 | 0.664 |
| serum_creatinine | 0.297 | 1.346 | 1.143 | 1.585 | 3.562 | <0.0005 |
| serum_sodium | -0.042 | 0.959 | 0.916 | 1.005 | -1.759 | 0.079 |
| sex | -0.274 | 0.761 | 0.458 | 1.264 | -1.055 | 0.291 |
| smoking | 0.148 | 1.159 | 0.699 | 1.922 | 0.573 | 0.567 |

The CoxPH model was used to assess the impact of various clinical features on heart failure survival. This model estimates the coefficient (β) parameter for each predictor variable to represent the relative risk of death. The proportional hazard assumption is fundamental, meaning that the relative risk between groups remains constant over time. In the CoxPH model, the key parameters are the coefficients for each predictor variable, rather than the hyperparameters of traditional machine learning.

5.2 Kaplan-Meier estimator

Fig. 6 shows the Kaplan-Meier curve showing the probability that patients will survive up to time t . The curve was created by plotting the survival function over time. It starts from 1 (i.e. 100% survival at the beginning) and decreases over time as an increasing number of patients die. The numbers beneath the figure show the accumulation of mortality events over time as well as censoring events when surviving patients are lost to follow-up.

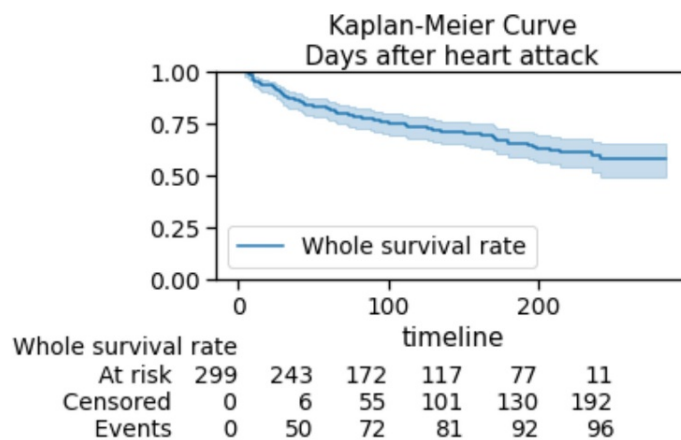
**Fig. 6.** Kaplan-Meier curve.

Fig. 7 shows the Kaplan-Meier estimates for the categorical variables. Compared to non-anaemic patient, anaemic patients had a higher likelihood of mortality.

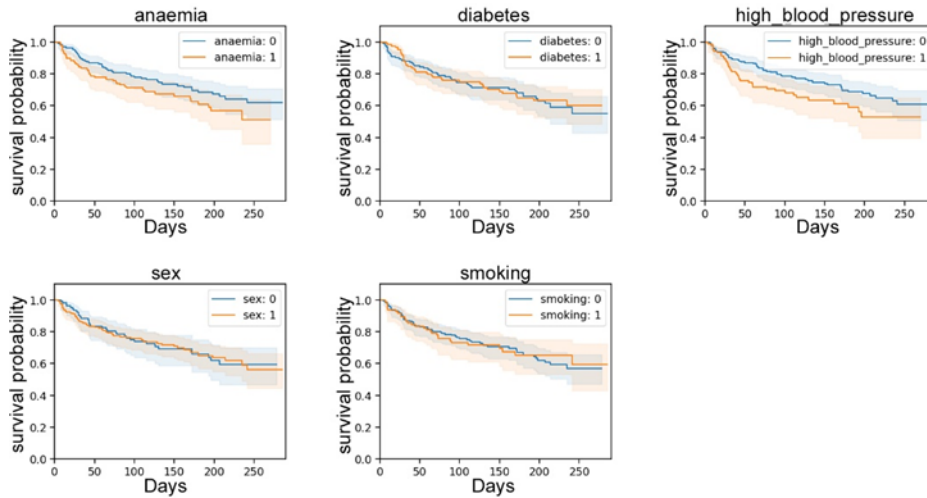


Fig. 7. Kaplan-Meier estimates for categorical variables.

Fig. 8 shows the predicted patient survival probability as a function of the continuous predictor variables. As mentioned previously, older age appears to play a role in increased death, with older patients tending to survive for shorter periods. This finding is consistent with intuitive expectations. Elevated serum creatinine level, which can indicate kidney function problems, appears to be associated with increased mortality as well. This result was not unexpected. CPK, platelets and serum sodium did not appear to have any significant influence on the risk of death.

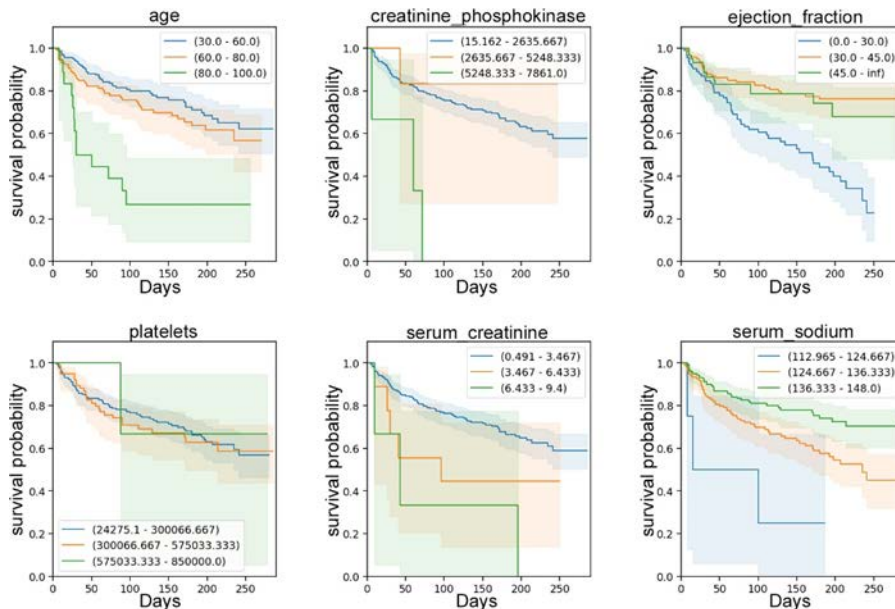


Fig. 8. Kaplan-Meier estimates for continuous variables.

In previous study [10], the ejection fraction levels were categorized into three groups (i.e., $EF \leq 30$, $30 < EF \leq 45$ and $EF > 45$) (Fig. 9).

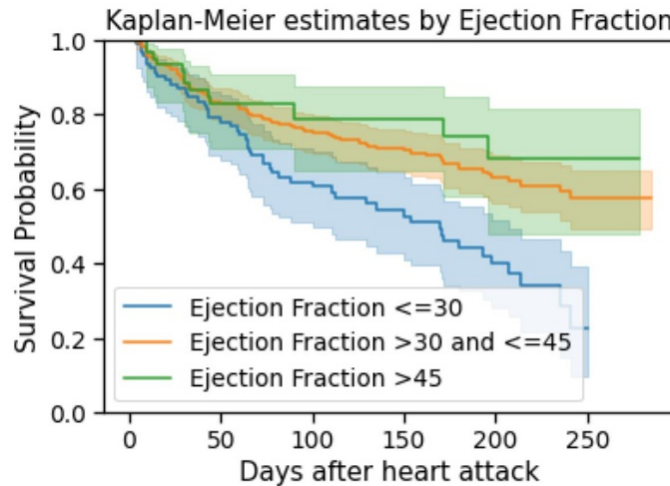


Fig. 9. Kaplan-Meier estimated by Ejection Fraction [10].

According to the American Heart Association (<http://heart.org>), a normal ejection fraction is about 50 - 75%. Therefore, this study set the borderline ejection fraction as between 41% and 50% (i.e. $EF \leq 41$, $41 < EF \leq 50$ and $EF > 50$) (Fig. 10).

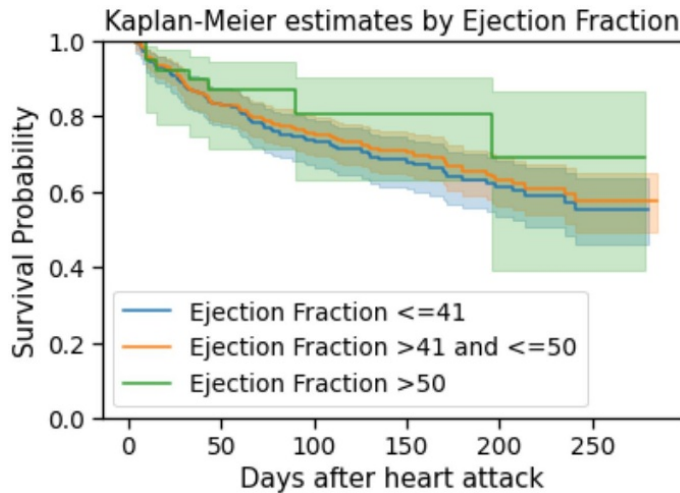


Fig. 10. Kaplan-Meier estimated by Ejection Fraction.

Lower ejection fraction seems to correlate with increased mortality. This seems reasonable, as the heart does not pump enough blood (i.e., low stroke volume) during heart failure.

5.3 Experiment

The experiment has been carried out in Google Colaboratory and encoded in Python using the Python library (e.g. lifelines). Some examples of source code have been modified from the open source under the Apache license, in which users are permitted to modify and distribute

the original open source code. The source code is publicly accessible on GitHub¹.

The objective of this study was to identify key predictors of mortality rather than to evaluate the contribution of individual components within a complex predictive model, negating the need for an additional ablation study to determine the importance of individual factors. Ablation studies are typically more relevant in scenarios where the impact of removing specific features or components from a model needs to be assessed, which does not align with the primary focus of this study.

5.4 Analysis

In this study, a comprehensive survival analysis was performed on a dataset consisting of 299 patients with HF using the Python library. In this study, mortality was analyzed using Cox regression modeling to identify the significant features strongly associated with patient outcomes. Furthermore, the Kaplan-Meier survival curve approach describes patient survival patterns over time. The results of this study showed that variables such as age, ejection fraction, and serum creatinine levels were significantly associated with mortality ($p \leq 0.001$), and anemia and creatinine phosphatases were statistically significance ($p = 0.026$ and $p = 0.007$, respectively). The Cox model shows praiseworthy concordance with the data (0.77) and highlights the usefulness of the identified variables in predicting mortality in patients with heart failure.

6. Conclusion

Using the Cox proportional hazards model, it was possible to model the survival functions for individuals with heart failure. Based on the C-index value, the predictive power of the model is good. The predictive impacts of the ejection fraction were adequately demonstrated using Kaplan-Meier curves. According to the Kaplan-Meier curves, age was strongly correlated with survival, and Cox proportional hazards modelling confirmed that age was the most significant variable.

Older age, anaemia, lower ejection fraction (EF), high BP, elevated serum creatinine, and decreased serum sodium levels appear to contribute to a higher risk of mortality among heart failure patients. Creatinine phosphokinase, diabetes, platelets, sex and smoking status were not significant. These findings generally agree with expectations regarding the main risk factors for mortality in the HF patient population. They illustrate the usefulness of the survival analysis methodology implemented in Python for modelling medical time-to-event data.

Future work includes 1) further refinement of the Cox proportional hazards model by incorporating additional covariates, such as medical history and lifestyle factors, to improve its predictive accuracy; 2) validation of the model's results by comparing them to results from other survival analysis techniques, such as the Weibull or log-normal models; 3) comparison of the results of the current study to similar studies on heart failure survival in other populations to assess the generalizability of the findings; 4) investigation of the impact of different treatment strategies, such as medication or surgery, on the survival of heart failure patients using the Cox proportional hazards model; and 5) development of a machine learning algorithm that can predict heart failure outcomes based on the results of the Cox proportional hazards model and other relevant data. This could lead to the creation of personalized treatment

¹ The source code is publicly available on <https://github.com/alwaysapril/survival-analysis-for-heart-failure-patients>

plans for patients with HF that consider their risk factors and predicted outcomes. Based on the results of these studies, the risk of heart failure-related mortality can be estimated using an internet-based tool that allows healthcare providers to input patient information.

The future directions of this study are as follows:

- Cox proportional risk model improvement: The addition of more variables, including medical history and lifestyle factors, might increase the predictive accuracy of the model.
- Validation with alternative methods: Validation of model performance compared to other survival analysis methods such as Weibull or log-normal models helps to establish the robustness and generalizability of our findings. This is possible with existing datasets and can enhance the reliability of our models.
- Comparative studies of diverse populations: A comparison of the findings with those of a similar heart failure survival study of diverse populations will allow us to evaluate the acceptance of our findings. This is essential for generalizing our model; however, it requires collaboration and access to various datasets.

This study conducted survival analysis of 299 patients with HF using Python libraries. Cox regression and Kaplan-Meier curves showed that age, ejection fraction, and serum creatinine levels were significantly associated with mortality. Good concordance (0.77) suggests that the identified variables are useful in predicting mortality in patients with heart failure. This study can help provide risk stratification categories for patients with HF that could be used by healthcare providers to prioritize patients for monitoring and intervention.

Limitations

Our study extends the study by Ahmad et al. by utilizing the same dataset and reaching different outcomes for the most critical factors influencing patient mortality. We explored these factors in greater depth and employed a hybrid approach combining both traditional statistical methods and advanced machine learning techniques such as Cox regression and Kaplan-Meier survival analysis with machine learning models to refine and extend previous research.

- Selection bias: Datasets may be limited to specific demographics or geographic regions and thus may not represent a broad population of heart failure patients.
- Measurement bias: Disagreements may arise owing to the variety of methods in which clinical measurements have been performed.
- Limitations of Cox regression (Proportional Risk Assumption): Cox regression assumes that the risk ratio is constant over time, which may not correspond to all variables.
- Omitted Variables: This study does not consider all relevant factors because we used the same dataset of Ahmad et al., which may result in biased estimates.

Future work could expand on this by exploring different patient datasets, introducing new variables, or applying state-of-the-art modeling techniques to further distinguish them from prior studies.

Acknowledgement

This work was supported in part by the National Research Foundation of Korea (NRF) funded by Korean Government through the Ministry of Science and ICT (MSIT) under Grant RS-2023-00250496, and in part by the National Research and Development Program through the National Research Foundation of Korea (NRF) funded by MSIT under Grant 2020M3H2A1076786.

References

- [1] V. Rudomanova and B. C. Blaxall, "Targeting GPCR-G β γ -GRK2 signaling as a novel strategy for treating cardiorenal pathologies," *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, vol.1863, no.8, pp.1883-1892, 2017. [Article \(CrossRef Link\)](#)
- [2] M. Shabani et al., "Pre-diagnostic predictors of mortality in patients with heart failure: The multi-ethnic study of atherosclerosis," *Frontiers in Cardiovascular Medicine*, vol.9, 2022. [Article \(CrossRef Link\)](#)
- [3] S. J. Pocock et al., "Predicting survival in heart failure: a risk score based on 39 372 patients from 30 studies," *European Heart Journal*, vol.34, no.19, pp.1404-1413, May 2013. [Article \(CrossRef Link\)](#)
- [4] J. B. Greenhouse, D. Stangl, and J. Bromberg, "An introduction to survival analysis: statistical methods for analysis of clinical trial data.," *Journal of Consulting and Clinical Psychology*, vol.57, no.4, pp.536-544, 1989. [Article \(CrossRef Link\)](#)
- [5] A. Spooner et al., "A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction," *Scientific Reports*, vol.10, no.1, pp.1-10, 2020. [Article \(CrossRef Link\)](#)
- [6] M. M. Alem, "Predictors of Mortality in Patients with Chronic Heart Failure: Is Hyponatremia a Useful Clinical Biomarker?," *International Journal of General Medicine*, vol.13, pp.407-417, 2020. [Article \(CrossRef Link\)](#)
- [7] J. A. Regan et al., "Impact of Age on Comorbidities and Outcomes in Heart Failure With Reduced Ejection Fraction," *JACC: Heart Failure*, vol.7, no.12, pp.1056-1065, 2019. [Article \(CrossRef Link\)](#)
- [8] M. Drozd et al., "Association of heart failure and its comorbidities with loss of life expectancy," *Heart*, vol.107, no.17, pp.1417-1421, 2021. [Article \(CrossRef Link\)](#)
- [9] A. Asuncion and D. J. Newman, "UCI machine learning repository," Irvine, CA, USA, 2007.
- [10] T. Ahmad, A. Munir, S. H. Bhatti, M. Aftab, and M. A. Raza, "Survival analysis of heart failure patients: A case study," *PLoS ONE*, vol.12, no.7, 2017. [Article \(CrossRef Link\)](#)
- [11] F. M. Zahid, S. Ramzan, S. Faisal, and I. Hussain, "Gender based survival prediction models for heart failure patients: A case study in Pakistan," *PLoS ONE*, vol.14, no.2, 2019. [Article \(CrossRef Link\)](#)
- [12] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," *BMC Medical Informatics and Decision Making*, vol.20, pp.1-16, 2020. [Article \(CrossRef Link\)](#)
- [13] A. Newaz, N. Ahmed, and F. Shahriyar Haq, "Survival prediction of heart failure patients using machine learning techniques," *Informatics in Medicine Unlocked*, vol.26, 2021. [Article \(CrossRef Link\)](#)
- [14] E. D. Adler et al., "Improving risk prediction in heart failure using machine learning," *European Journal of Heart Failure*, vol.22, no.1, pp.139-147, 2020. [Article \(CrossRef Link\)](#)
- [15] A. Guo, M. Pasque, F. Loh, D. L. Mann, and P. R. O. Payne, "Heart Failure Diagnosis, Readmission, and Mortality Prediction Using Machine Learning and Artificial Intelligence Models," *Current Epidemiology Reports*, vol.7, pp.212-219, 2020. [Article \(CrossRef Link\)](#)

- [16] G. L. Di Tanna, H. Wirtz, K. L. Burrows, and G. Globe, "Evaluating risk prediction models for adults with heart failure: A systematic literature review," *PLoS ONE*, vol.15, no.1, 2020. [Article \(CrossRef Link\)](#)
- [17] A. Yazdani, K. D. Varathan, Y. K. Chiam, A. W. Malik, and W. A. Wan Ahmad, "A novel approach for heart disease prediction using strength scores with significant predictors," *BMC Medical Informatics and Decision Making*, vol.21, no.1, 2021. [Article \(CrossRef Link\)](#)
- [18] A. Ishaq et al., "Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques," *IEEE Access*, vol.9, pp.39707-39716, 2021. [Article \(CrossRef Link\)](#)
- [19] A. Newaz, N. Ahmed, and F. S. Haq, "Survival prediction of heart failure patients using machine learning techniques," *Informatics in Medicine Unlocked*, vol.26, 2021. [Article \(CrossRef Link\)](#)
- [20] M. Kavitha, G. Gnaneswar, R. Dinesh, Y. R. Sai, and R. S. Suraj, "Heart Disease Prediction using Hybrid machine Learning Model," in *Proc. of 2021 6th international conference on inventive computation technologies (ICICT)*, pp.1329-1333, 2021. [Article \(CrossRef Link\)](#)
- [21] S. Shin et al., "Machine learning vs. conventional statistical models for predicting heart failure readmission and mortality," *ESC Heart Failure*, vol.8, no.1, pp.106-115, 2021. [Article \(CrossRef Link\)](#)
- [22] A. A. Almazroi, "Survival prediction among heart patients using machine learning techniques," *Mathematical Biosciences and Engineering*, vol.19, no.1, pp.134-145, 2022. [Article \(CrossRef Link\)](#)
- [23] A. Saboor, M. Usman, S. Ali, A. Samad, M. F. Abrar, and N. Ullah, "A Method for Improving Prediction of Human Heart Disease Using Machine Learning Algorithms," *Mobile Information Systems*, vol.2022, 2022. [Article \(CrossRef Link\)](#)
- [24] C. R. Olsen, R. J. Mentz, K. J. Anstrom, D. Page, and P. A. Patel, "Clinical applications of machine learning in the diagnosis, classification, and prediction of heart failure," *American Heart Journal*, vol.229, pp.1-17, 2020. [Article \(CrossRef Link\)](#)
- [25] S. K. Arunachalam, and R. Rekha, "A novel approach for cardiovascular disease prediction using machine learning algorithms," *Concurrency and Computation: Practice and Experience*, vol.34, no.19, 2022. [Article \(CrossRef Link\)](#)



HEEJEONG JASMINE LEE received the BSc. (Computer Science) from POSTECH, Korea, the MSc. (Computer Science) from the University of Edinburgh, UK, the MPhil. (Technology Policy) from Cambridge University, UK and the PhD. (Information Technology) from Monash University, Australia. She was with KT(Korea Telecom) R&D for 8 years; and has filed 11 patents on telecommunications applications. Currently, she is a research professor at the College of Information and Communication Engineering, Sungkyunkwan University, South Korea. Her move from industry to academia was motivated by a love of learning. Her research interests include AI in dentistry and integrating AI models into hardware.



Sang-Sun Yoo received his B.S. degree from Dong-guk University, Seoul, Korea, in 2004, and his M.S/Ph.D. degree received from the Korea Advanced Institute of Science and Technology (KAIST), Daejeon, in 2012. He worked for System LSI division in Samsung Electronics from 2012 to 2015 where he focused on ADPLL for 3/4G mobile applications as a senior design engineer. From 2015 to 2016, he was a Research Assistant Professor in KAIST and Sungkyunkwan University. Since 2017, he has been with the Department of Smart Automobile, Pyeongtaek University, where he was an Associate Professor. His research interests include RF systems for mobile communications, reconfigurable RFICs, automotive ICs, ADPLL, RFID, and sensor communications. Dr. Yoo was the recipient of the Best Paper Award of the IEEE TRANSACTIONS ON INDUSTRIAL ELECTRONICS in 2011.



KANG-YOON LEE (Senior Member, IEEE) received the B.S., M.S., and Ph.D. degrees from the School of Electrical Engineering, Seoul National University, Seoul, South Korea, in 1996, 1998, and 2003, respectively. From 2003 to 2005, he was with GCT Semiconductor Inc., San Jose, CA, USA, where he was the Manager of the Analog Division and worked on the design of CMOS frequency synthesizer for CDMA/PCS/PDC and single-chip CMOS RF chip sets for W-CDMA, WLAN, and PHS. From 2005 to 2011, he was with the Department of Electronics Engineering, Konkuk University, Seoul, as an Associate Professor. Since 2012, he has been with the College of Information and Communication Engineering, Sungkyunkwan University, Suwon, South Korea, where he is currently a Professor. His research interests include the implementation of power integrated circuits, CMOS RF transceiver, analog integrated circuits, and analog/digital mixed-mode VLSI system design.