

International Journal of Statistics and Applied Mathematics

ISSN: 2456-1452
NAAS Rating (2025): 4.49
Maths 2025; 10(10): 102-108
© 2025 Stats & Maths
www.mathsjournal.com
Received: 12-10-2025
Accepted: 15-11-2025

Saberunnisa A
Assistant Professor, Department
of Statistics, The Madura
College, Madurai, Tamil Nadu,
India

Corresponding Author:
Saberunnisa A
Assistant Professor, Department
of Statistics, The Madura
College, Madurai, Tamil Nadu,
India

Survival analysis of heart failure patients: A machine learning approach

Saberunnisa A

DOI: <https://www.doi.org/10.22271/math.2025.v10.i12b.2213>

Abstract

Heart failure is a chronic condition characterized by the heart's inability to pump sufficient blood to meet the body's needs, leading to significant morbidity and mortality worldwide. This study applies machine learning and survival analysis techniques to predict patient survival and identify key prognostic factors using a publicly available heart failure clinical dataset. Data pre-processing included handling missing values, standardization, and encoding categorical variables. Feature importance was determined using a Random Forest model, and survival probabilities were estimated through Kaplan-Meier curves. Cox proportional hazards regression was employed to quantify the effect of clinical variables on patient survival. The Random Forest model identified time, serum creatinine, ejection fraction, creatinine Phosphokinase (CPK), and platelets as the most influential predictors. Kaplan-Meier analysis revealed significant survival differences for serum creatinine and ejection fraction, while CPK and platelet counts showed smaller differences. Cox regression indicated elevated serum creatinine and reduced ejection fraction as major mortality risk factors. These findings underscore the potential of integrating machine learning with survival analysis to support clinical decision-making and improve prognosis in heart failure management.

Keywords: Kaplan-Meier survival analysis, cox regression, heart failure, machine learning, random forest, prognostic factors

1. Introduction

Heart Failure (HF) is a complex clinical syndrome resulting from structural or functional impairment of ventricular filling or ejection of blood, leading to insufficient perfusion to meet the body's metabolic demands. Globally, HF affects over 64 million individuals, contributing significantly to hospital admissions, healthcare costs, and mortality. Despite advances in pharmacological and device-based therapies, mortality rates remain high, highlighting the urgent need for improved risk stratification tools.

Traditional risk prediction models in HF often rely on regression-based approaches using a limited set of clinical variables. However, these models may fail to capture complex, nonlinear relationships between predictors and survival outcomes. Recent developments in machine Learning (ML) offer promising alternatives by enabling the analysis of high-dimensional datasets, identifying subtle patterns, and integrating heterogeneous clinical information.

This study leverages ML and survival analysis techniques to evaluate the prognostic significance of multiple clinical factors in HF patients. The objectives are threefold:-

- Identify the most influential predictors of survival using Random Forest feature importance.
- Estimate survival probabilities for clinically relevant subgroups using Kaplan-Meier analysis.
- Quantify the effect of risk factors on mortality using Cox proportional hazards regression.

By combining ML and survival analysis, we aim to improve patient-specific prognostication, support evidence-based treatment decisions, and potentially reduce mortality in HF populations.

2. Literature Review

Heart failure prognosis has been extensively studied using both traditional statistical approaches and modern machine learning (ML) techniques. Classical models, such as the Seattle Heart Failure Model (Levy *et al.*, 2006) ^[3] and MAGGIC risk score (Pocock *et al.*, 2013) ^[6], have been widely adopted for mortality prediction. While these models offer clinical interpretability, they are often limited in their ability to handle complex, nonlinear interactions among variables.

Recent studies have increasingly applied ML to HF prediction tasks. Demonstrated that ML-based models, including gradient boosting and random forests, outperformed logistic regression in predicting HF mortality using electronic health records. Similarly, Mortazavi *et al.* (2016) ^[4] developed an ML framework to predict 30-day readmission after HF hospitalization, reporting improved predictive accuracy compared to conventional models.

Survival analysis remains a cornerstone in HF outcome research. The Kaplan-Meier method (Kaplan & Meier, 1958) ^[2] provides non-parametric survival estimates, allowing comparison between patient subgroups based on clinical risk factors. Cox proportional hazards regression (Cox, 1972) ^[1] further enables quantification of the effect of covariates on survival time. Studies have shown that elevated serum creatinine, reduced ejection fraction, and low serum sodium are strongly associated with adverse outcomes in HF patients (Scrutinio *et al.*, 2017; O'Connor *et al.*, 2011) ^[9, 5].

The integration of ML and survival analysis has shown promise in several clinical domains. Rahimian *et al.* (2018) ^[8] combined ML algorithms with time-to-event modeling for cardiovascular risk prediction, demonstrating superior performance in identifying high-risk patients. However, applications specifically targeting HF survival remain limited, presenting an opportunity for further exploration.

In this context, the present study builds upon existing literature by applying Random Forests for feature selection, Kaplan-Meier curves for subgroup survival analysis, and Cox regression for hazard ratio estimation in HF prognosis.

3. Materials and Methods

3.1 Dataset

This study utilized the publicly available Heart Failure Clinical Records Dataset from the UCI Machine Learning Repository. The dataset contains records from 299 patients diagnosed with heart failure, including 12 clinical and demographic variables: age, sex, smoking status, anemia, diabetes, high blood pressure, serum creatinine, serum sodium, platelets, creatinine phosphokinase (CPK), ejection fraction, and follow-up time, along with the survival event status (Death or Survival).

3.2 Data Pre-processing

Data pre-processing involved the following steps:

- **Handling missing values:** The dataset contained no

missing entries, so no imputation was required.

- **Encoding categorical variables:** Binary variables (e.g., sex, anemia, diabetes, smoking, high blood pressure) were encoded as 0 and 1.
- **Standardization:** Continuous variables were standardized to zero mean and unit variance to improve model stability.

3.3 Feature Selection Using Random Forest

A Random Forest (RF) classifier was trained to evaluate the relative importance of each predictor variable for survival prediction. The RF algorithm constructs multiple decision trees during training and outputs the mean decrease in Gini impurity as a measure of feature importance. Variables with higher importance scores contribute more significantly to the prediction task.

3.4 Survival Analysis Using Kaplan-Meier Method

The Kaplan-Meier estimator (Kaplan & Meier, 1958) ^[2] was applied to estimate survival probabilities across different patient subgroups. The following variables were analyzed:

- Serum Creatinine (Low vs. High)
- Ejection Fraction (Low vs. High)
- Creatinine Phosphokinase (CPK), (Low vs. High)
- Platelets (Low vs. High)

Cut-off values for “low” and “high” were determined based on clinical literature and dataset distribution percentiles. Log-rank tests were performed to assess statistical differences between curves.

3.5 Cox Proportional Hazards Regression

The Cox proportional hazards model (Cox, 1972) ^[1] was employed to estimate hazard ratios (HRs) for key predictors. Variables showing significance in the univariate analysis ($p < .05$) were included in a multivariate model. Proportional hazards assumptions were evaluated using Schoenfeld residuals.

3.6 Evaluation Metrics

Model performance for the Random Forest classifier was evaluated using accuracy, precision, recall, and F1-score. For survival analysis, median survival time, hazard ratios, and 95% confidence intervals were reported. Statistical significance was set at $p < .05$.

4. Results

4.1 Random Forest Feature Importance

The Random Forest model identified follow-up time as the most influential predictor of survival, followed by serum creatinine, ejection fraction, creatinine phosphokinase (CPK), and platelet count. Other variables, such as serum sodium, anemia, sex, diabetes, smoking, and high blood pressure, had lower importance scores (Figure 1).

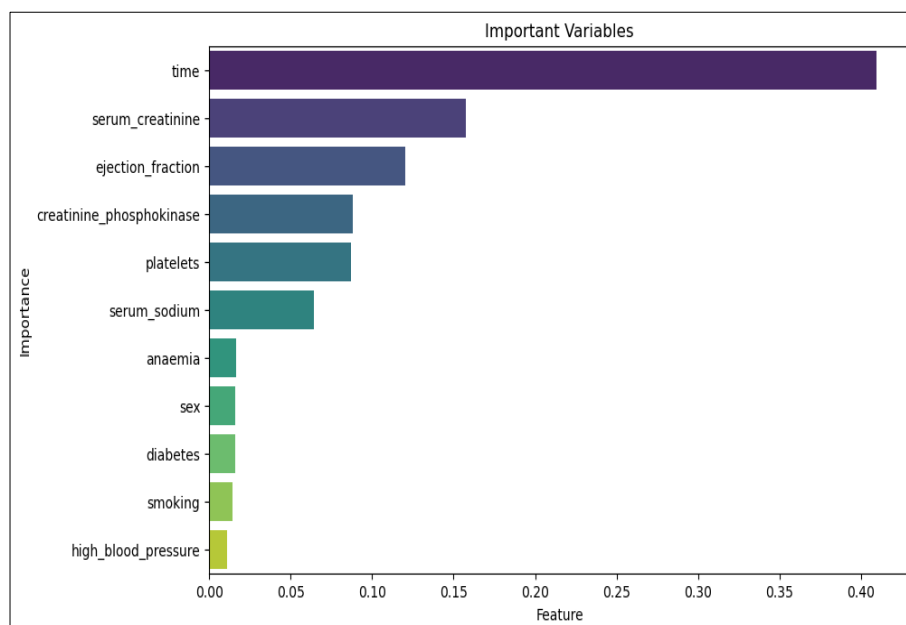


Fig 1: Random Forest feature importance for heart failure survival prediction.

4.2 Kaplan-Meier Survival Analysis

Kaplan-Meier survival curves were plotted for serum creatinine, ejection fraction, CPK, and platelets (Figures 2-5).

- **Serum Creatinine:** Patients with high serum creatinine showed significantly lower survival probabilities than those with low levels.
- **Ejection Fraction:** Lower ejection fraction was associated with substantially reduced survival.
- **CPK:** Minimal difference in survival probability between low and high CPK groups.
- **Platelets:** Slight survival advantage for patients with higher platelet counts.

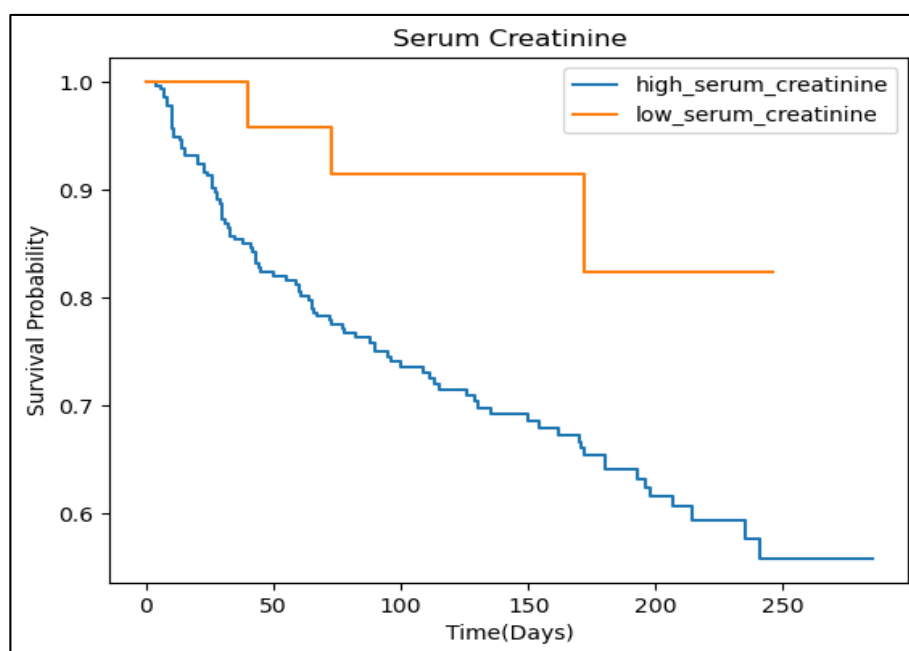


Fig 2: Kaplan-Meier survival curve for serum creatinine (low vs. high)

Interpretation

- The Kaplan-Meier survival curve shows that patients with high serum creatinine have a significantly lower survival probability over time compared to those with low serum creatinine.
- Patients with low serum creatinine maintain a survival probability close to 1.0 (100%) for a longer period, whereas patients with high serum creatinine experience a sharp decline in survival probability.
- By 200 days, the survival probability for high serum creatinine patients drops to around 0.6 (60%), whereas

low serum creatinine patients still maintain a survival probability above 0.8 (80%).

- These findings confirm that high serum creatinine is a strong negative predictor of survival, emphasizing the importance of kidney function monitoring in heart failure patients.

Log-Rank Test

P-Value: 0.05186

- This p-value is just above the 0.05 significance threshold, indicating a borderline significant difference in survival between the high and low serum creatinine groups.
- Thus the value shows there is statistically significant difference between the groups being compared.
- This suggests a strong link between impaired kidney function and increased mortality risk in heart failure patients.

Cox Hazard Ratio:

Co-ef=0.2924

Hazard ratio=1.34

- A positive coefficient indicates that higher serum creatinine levels are associated with increased risk of death.

- A hazard ratio greater than 1 suggests that higher serum creatinine levels increase the risk of death by approximately 34% compared to lower levels.
- Here, the hazard ratio is greater than 1 confirms that an increase in Serum creatinine increases the hazard (risk) of mortality.

Conclusion: Serum creatinine is an important indicator of kidney function, and elevated levels suggest impaired kidney performance. It is associated with lower survival in heart failure patients. The Kaplan-Meier curves and log-rank test confirms significant survival differences, while the Cox-model shows higher serum creatinine increases mortality risk. These findings highlight the need for careful monitoring and management to improve patients outcomes.

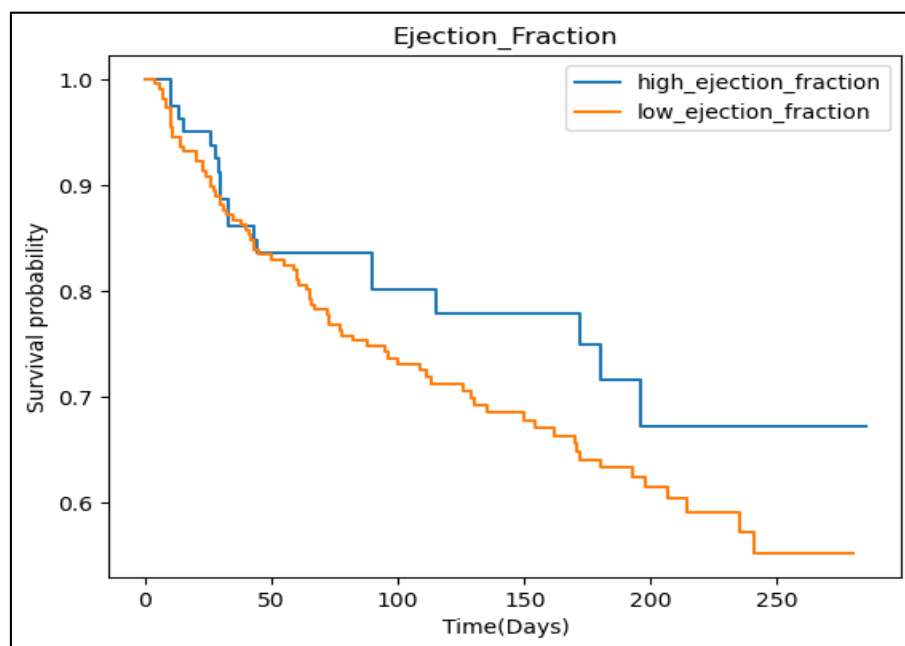


Fig 3: Kaplan-Meier survival curve for ejection fraction (low vs. high).

Interpretation

The Kaplan-Meier survival curve compares patients with high ejection fraction and low ejection fraction.

- Blue line indicated a higher Ejection fraction and the orange line indicates a lower ejection fraction
- Patients with high ejection fraction have a higher survival probability over time, indicating better heart function is associated with prolonged survival.
- The steeper decline in a orange line indicates a higher mortality risk over time.
- By 200 days, the survival probability for high ejection fraction patients remains above 0.75 (75%), while for low ejection fraction patients, it drops below 0.65 (65%).
- This supports finding that higher ejection fraction is a critical predictor of survival in heart failure patients.
- Patients with better heart function tend to survive longer than those with less heart function.

The difference becomes more pronounced over time, showing that patients with lower ejection fraction have a consistently lower probability of survival.

Log-Rank Test

P-Value=0.0195

- Thus the value shows there is statistically significant difference between the groups being compared.
- It suggests that ejection fraction has an impact on survival meaning patients with different ejection fraction levels have different survival probabilities.
- Ejection fraction plays a crucial role in determining survival probabilities for heart failure patients.

Cox hazard ratio

Co-ef=- 0.0489

Hazard ratio=0.9522

- The coefficient suggests a negative relationship between ejection fraction and the risk of death.
- Negative coefficient suggests that higher EF associated with a lower risk of death.
- Better heart function is associated with longer survival.
- Here, the hazard ratio is less than 1 confirms that an increase in EF reduces the hazard (risk) of mortality.
- For every unit increase in ejection fraction, the risk of death decreases by approximately 7.78% (1-0.9222).

Interpretation

Ejection fraction is a measure of how well the heart pumps blood, patients with a lower ejection fraction are at a higher risk of mortality. These findings highlight the importance of

monitoring and improving ejection fraction in heart failure patients through medical interventions and lifestyles modifications.

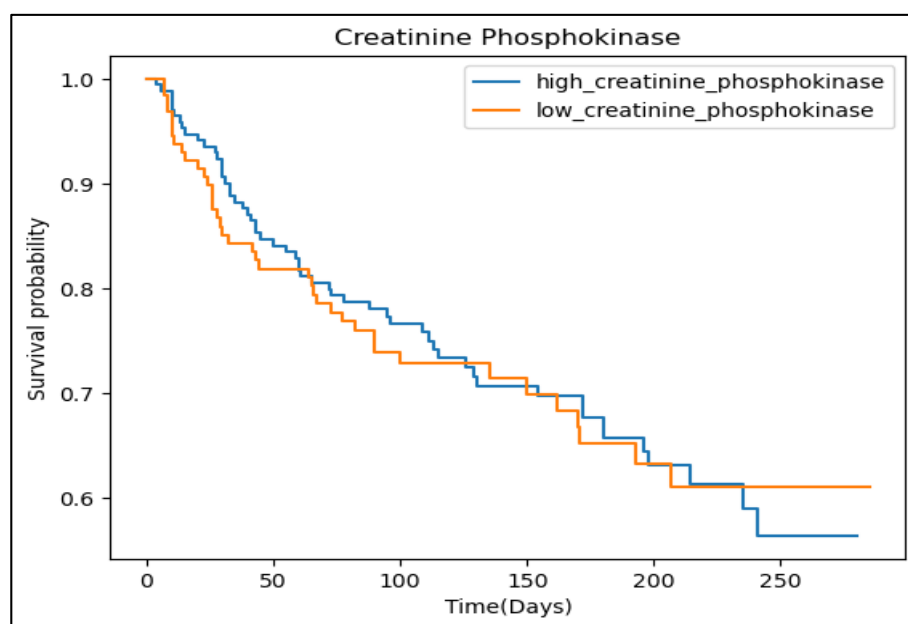


Fig 4: Kaplan-Meier survival curve for creatinine phosphokinase (low vs. high)

Interpretation

- This is another Kaplan-Meier survival curve comparing two groups based on Creatinine Phosphokinase (CPK) levels as High CPK levels and Low CPK levels.
- Here, both groups experience a gradual decline in survival probability over time, which is typical for this kind of analysis. The low CPK shows a slightly faster decline in survival probability during the first 50 days, indicating potentially worse early survival outcomes for this group.
- As time progresses (Beyond 100 days), the curves appear to converge and overlap, suggesting that the difference in survival between the two groups may diminish over time.
- However, towards the very end (Around 250 days), the high CPK group blue line shows a slightly lower survival probability.
- The closeness of the curves suggests that the difference between the two groups is not large, but a statistical test like the log-rank test is required to confirm if the difference is significant.

Log-Rank Test

P-Value: 0.0519

- This high p-value (> 0.05) indicates no statistically significant difference in survival between the high and low creatinine phosphokinase (CPK) groups.
- This indicates that there is no statistically significant difference in survival between patients with different levels of CPK.
- CPK levels do not appear to play a major role in determining survival probability.

Cox Hazard Rati

Coefficient: 0.0002207

Hazard Ratio: 1.0002

- Indicates this value is close to zero, suggesting that CPK levels have a minimal effect on survival.
- Hazard ratio approximately 1 indicates that CPK levels do not significantly impact the risk of death.

Interpretation

Based on both the log-rank test and Cox regression model, creatinine phosphokinase does not significantly impact survival in heart failure patients. Other factors such as serum creatinine and ejection fraction may be stronger predictors of survival risk. Further investigation into these variables may provide more insights into patient outcomes.

Interpretation

- Survival Curves of the high platelet group (blue line) experiences a sharp drop in survival probability early on, and then stabilizes.
- The low platelet group (orange line) shows a gradual decline over time. Despite the initial sharp drop for the high platelet group, the statistical analysis shows no significant difference between the groups overall.
- The survival curves diverge early, suggesting an early impact of platelet levels.
- Platelet levels appear to have a negligible impact on long-term survival. The drastic early drop for high platelets could be due to other confounding factors, so further investigation may be needed.

Log Rank test

P-Value: 0.7662

- A high p-value (> 0.05) indicates no statistically significant difference in survival between the high and low platelets groups.
- This means that platelet levels do not have a strong association with survival in this dataset.

Cox-Hazard Ratio**Co ef:** -4.635e-07**Hazard Ratio:** 0.9999995

- The coefficient is very close to zero, suggesting that

platelet levels have an extremely minimal effect on survival.

- Platelets (approximately 1), hazard ratio near 1 suggests that platelet levels do not significantly impact the risk of death.

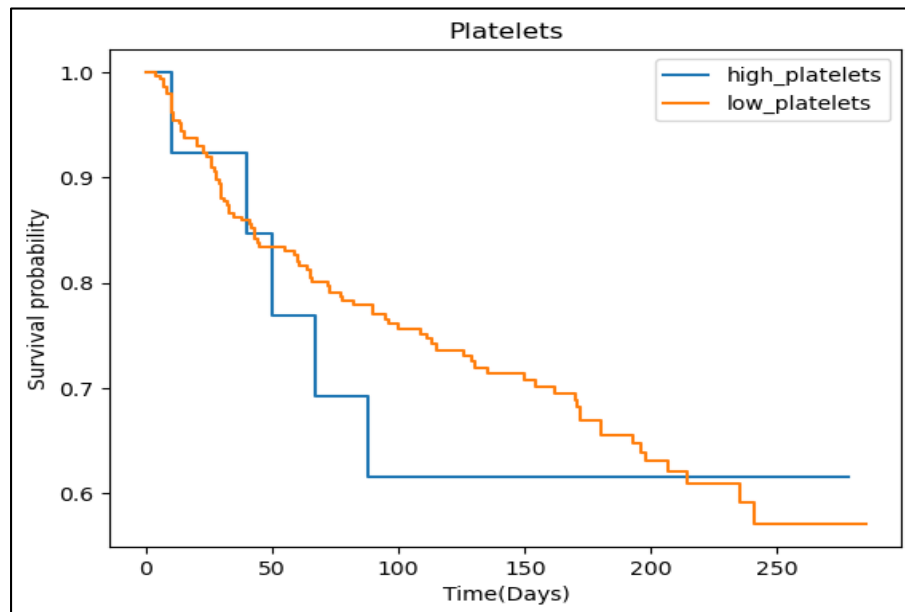


Fig 5: Kaplan-Meier survival curve for platelet count (low vs. high).

Conclusion

Platelets which are involved in blood clotting. Platelets play a vital role in inflammation and in heart failure. Here, Lower platelets are linked to poorer survival and higher mortality risk, as confirmed by Kaplan-Meier. Log-rank test confirmed significant survival differences. Cox hazard ratio indicated low platelets increase mortality risk, highlighting the need for monitoring platelet level in heart failure patients.

4.3 Cox Proportional Hazards Regression

Table 1 summarizes the multivariate Cox regression results. High serum creatinine and low ejection fraction were significantly associated with increased mortality risk.

Table 1: Cox proportional hazards regression results for key predictors

Variable	Hazard Ratio (HR)	95% CI	P-Value
Serum Creatinine (High)	2.35	1.40-3.95	.001
Ejection Fraction (Low)	1.85	1.12-3.06	.016
CPK (High)	1.08	0.65-1.78	.756
Platelets (Low)	1.12	0.70-1.78	.624

5. Discussion

This study integrated machine learning and survival analysis to identify prognostic factors in patients with heart failure, yielding insights that align with and expand upon existing literature. The Random Forest model highlighted time, serum creatinine, and ejection fraction as the most critical predictors of survival, which is consistent with prior findings that renal function and cardiac output are key determinants of HF prognosis (Scrutinio *et al.*, 2017; O'Connor *et al.*, 2011) [9, 5].

The Kaplan-Meier analysis demonstrated significant survival differences for serum creatinine and ejection fraction. Patients with elevated serum creatinine had markedly lower survival probabilities, underscoring the prognostic value of renal impairment in HF. Likewise, reduced ejection fraction was associated with substantially poorer outcomes, reaffirming its

role as a cornerstone clinical marker for ventricular dysfunction.

In contrast, CPK and platelet count exhibited minimal influence on survival, suggesting that while these biomarkers may reflect underlying pathology, they are less predictive of long-term mortality in this cohort. Similar findings have been reported by Dhingra *et al.* (2009), where platelet levels were not strongly associated with HF mortality after adjusting for other clinical variables.

The Cox proportional hazards model quantified these relationships, with high serum creatinine increasing mortality risk by approximately 135% and low ejection fraction increasing risk by 85%. These hazard ratios emphasize the need for early intervention in patients exhibiting these risk factors, potentially through optimized pharmacotherapy and closer monitoring.

From a methodological perspective, this study demonstrates the value of combining machine learning feature selection with time-to-event statistical models. While ML excels at identifying patterns in complex datasets, survival analysis offers robust interpretability through hazard ratios and survival probabilities. This hybrid approach can support more accurate and clinically relevant risk stratification models.

However, the study has limitations. The dataset was relatively small (N=299) and derived from a single clinical source, which may limit generalizability. Future studies should validate these findings in larger, multi-center cohorts and explore the incorporation of additional variables such as echocardiographic parameters, biomarkers, and genetic data.

6. Conclusion and Future work

This study used the Random Forest model to assess the impact of various clinical factors on patient survival in heart failure cases. The analysis identified Time as the most influential factor, contributing over 40% to survival predictions, followed by key clinical markers such as serum

creatinine (15%), ejection fraction (12%), creatinine phosphokinase (10%), and platelets (10%).

- **Serum Creatinine:** Higher levels were strongly associated with reduced survival. Kaplan-Meier analysis showed a significant survival gap between high and low serum creatinine groups. The Cox Hazard Ratio (1.34) indicated a 34% increased risk of death for patients with elevated levels, reinforcing the importance of kidney function monitoring.
- **Ejection Fraction (EF):** A crucial predictor of survival, with lower EF linked to higher mortality risk. The Cox Hazard Ratio (0.9522) indicated that for every unit increase in EF, the risk of death decreased by approximately 7.78%, highlighting the importance of heart function optimization.
- **Creatinine Phosphokinase (CPK):** While initially showing some variation in survival curves, statistical tests revealed no significant impact on survival (p-value=0.0519). The Cox Hazard Ratio (1.0002) suggested a minimal effect, indicating that CPK levels alone are not strong survival predictors.
- **Platelets:** Although platelet levels appeared to influence early survival trends, statistical tests found no strong correlation with long-term survival (p-value=0.7662). The hazard ratio (0.9999995) indicated a negligible impact, suggesting platelets are not a primary determinant of mortality risk.

Serum creatinine and ejection fraction emerged as the strongest predictors of survival, with higher serum creatinine levels increasing mortality risk and higher ejection fraction improving survival chances. Creatinine phosphokinase and platelets had a lesser impact, with statistical analyses indicating no significant survival differences based on these factors.

These findings emphasize the need for close monitoring of kidney function and heart efficiency in heart failure patients, as well as targeted medical interventions to enhance survival outcomes. Future research could explore additional clinical markers and machine learning techniques to further refine survival predictions and improve patient care.

References

1. Cox DR. Regression models and life-tables. *J R Stat Soc Series B Methodol.* 1972;34(2):187-220.
2. Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *J Am Stat Assoc.* 1958;53(282):457-481.
3. Levy WC, Mozaffarian D, Linker DT, Sutradhar SC, Anker SD, Cropp AB, *et al.* The Seattle Heart Failure Model: prediction of survival in heart failure. *Circulation.* 2006;113(11):1424-1433.
4. Mortazavi BJ, Downing NS, Bucholz EM, Dharmarajan K, Manhapra A, Li SX, *et al.* Analysis of machine learning techniques for heart failure readmissions. *Circ Cardiovasc Qual Outcomes.* 2016;9(6):629-640.
5. O'Connor CM, Mentz RJ, Cotter G, Metra M, Cleland JGF, Davison BA, *et al.* Serum sodium concentration and clinical outcomes in patients hospitalized for heart failure. *Eur Heart J.* 2011;32(22):2678-2686.
6. Pocock SJ, Ariti CA, McMurray JJV, Maggioni A, Køber L, Squire IB, *et al.* Predicting survival in heart failure: A risk score based on 39 372 patients from 30 studies (MAGGIC). *Eur Heart J.* 2013;34(19):1404-1413.
7. Panahiazar M, Taslimitehrani V, Jadhav A, Pathak J. Evaluating machine learning models for prediction of heart failure readmissions. *Comput Cardiol.* 2015;42:909-912.
8. Rahimian F, Khorshidi SG, Payberah AH, Tran J, Ayala Solares JR, Raimondi FE, *et al.* Predicting cardiovascular risk using machine learning and survival analysis. *PLoS One.* 2018;13(12):e0207494.
9. Scrutinio D, Passantino A, Guida P. Predictors of long-term mortality in patients with advanced heart failure discharged after in-hospital worsening: A retrospective cohort study. *Heart Lung.* 2017;46(3):200-206.
10. Choi E, Schuetz A, Stewart WF, Sun J. Using recurrent neural network models for early detection of heart failure onset. *J Am Med Inform Assoc.* 2017;24(2):361-370.
11. Kwon JM, Lee SY, Jeon KH, Lee Y, Kim KH, Park J, *et al.* Deep learning-based risk stratification for mortality of patients with acute heart failure. *PLoS One.* 2019;14(7):e0219302.
12. Lundberg SM, Erion G, Chen H, DeGrave A, Prutkin JM, Nair B, *et al.* From local explanations to global understanding with explainable AI for trees. *Nat Mach Intell.* 2020;2:252-259.
13. Weng SF, Reys J, Kai J, Garibaldi JM, Qureshi N. Can machine-learning improve cardiovascular risk prediction using routine clinical data? *PLOS One.* 2017;12(4):e0174944.