DOI: 10.54254/2753-8818/50/2024AU0142

Machine learning based risk factor analysis in the prognosis of heart failure

Mingke Li

School of Biopharmaceuticals, Xi’an Jiaotong-Liverpool University, Suzhou, China

Mingke.Li22@student.xjtlu.edu.cn

Abstract. Heart failure, one of the significant causes of death, poses challenges to healthcare systems due to its high hospitalization rates and substantial economic burden. This study explores the key factors predicting survival among heart failure patients by evaluating a variety of clinical, demographic, and lifestyle elements. Employing logistic regression, this analysis utilizes data from 299 patients at the Faisalabad Institute of Cardiology, each marked with 12 risk factors, and conducts a 10-fold cross-validation to ensure robustness. Results reveal that age, serum sodium, and serum creatinine levels are crucial predictors of mortality. Contrary to expectations, ejection fraction plays a lesser role. This study broadens our understanding of the complex risk factors linked to heart failure, helping to refine predictive models that could improve patient outcomes and reduce the economic pressures on healthcare systems, especially in regions facing financial difficulties. The study also highlights the potential role of lifestyle factors, such as dietary and physical activity patterns, in managing heart failure, suggesting a broader approach for future research and interventions.

Keywords: Heart failure, logistic regression, risk factor analysis.

1. Introduction

Heart failure, a fatal cardiovascular condition, is defined by the inability to pump enough blood to fulfill the body's needs. It continues to be a major cause of morbidity and mortality across the globe, heavily stressing healthcare systems and impacting the quality of life for a patient. Despite advances in medical treatments and interventions, the prognosis for individuals with heart failure remains poor, with significant rates of hospitalization and mortality [1]. Around 46 million people worldwide suffer from heart failure, and the number of newly diagnosed cases continues to rise each year [2]. Nevertheless, prior research indicates that the one-year survival rate for acute HF ranges from 55% to 65% and that only half of the chronic heart failure patients survive five years after diagnosis, and about 35% survive 10 years [3]. Moreover, the economic impact of heart failure is enormous. Globally, the total direct and indirect medical costs resulting from heart failure are estimated to exceed $108 billion annually through the costs of hospitalization, medications, care, and lost productivity. Direct costs are about 60% (about $65 billion); 40% are indirect (about $43 billion). The financial burden of heart failure is particularly heavy in developing countries, where healthcare resources are relatively limited to effectively respond to the increasing number of heart failure cases. Whereas high-income countries allocate a higher proportion of their expenditures on direct costs, low- and middle-income countries face a greater load of indirect costs [4]. It is, therefore, crucial to understand the factors that contribute to the survival or death of a patient with HF to enhance the targeted strategies that would improve clinical outcomes and reduce the economic burden of HF.

Previous studies reported certain factors. For example, a population-based control study of Olmsted County, Minnesota, found that in 66.2% of the subjects with heart failure, hypertension is the predominant risk factor, with smoking following at 51.2% [5]. Furthermore, a prior systematic review and pooled analysis from various global regions discovered that ischemic heart disease (IHD) is common among heart failure patients, exceeding 50% in Europe and North America, between 30% and 40% in East Asia and Latin America, and less than 10% in sub-Saharan Africa. Worldwide, hypertension remains a widespread risk factor, yet it is especially common in Eastern and Central Europe (35%) and Sub-Saharan Africa (32.6%) [6].

Of key importance in these findings is the complex and often multifactorial nature of the risk for heart failure. The relative importance and interplay among them, however, need to be further explained. In this regard, this study has a look at various kinds of risk factors associated with heart failure, stretching from clinical biomarkers to demographic data and lifestyle variables. This is expected to help explain the progression of heart failure by identifying the main predictors for survival and therefore assisting in making evidence-based recommendations for the management of such patients. This study analysed clinical biomarkers that include natriuretic peptides, renal function indicators, and inflammatory markers related to demographics: age, sex, and ethnicity, among others. Lifestyle variables will be diet, physical activity, and smoking status. This research attempts to decipher the main predictors for survival because they explain heart failure pathophysiology and would aid in applying evidence-based recommendations for the treatment of these patients.

2. Materials and method

2.1. Dataset

The study is conducted based on a comprehensive medical dataset that contains heart failure subjects collected from the Faisalabad Institute of Cardiology and at the Allied Hospital in Faisalabad [7, 8]. The dataset contains 299 individuals (105 females and 194 males) and incorporates 12 risk factors: age, anemia, high blood pressure, creatinine phosphokinase, diabetes, ejection fraction, sex, platelets, serum creatinine, serum sodium, smoking status, and survival time. Each individual is marked with a binary value that indicates whether a death event occurred. Survival time is recorded from the start of the monitoring period to either the individual's death or the end of the study. It's important to note that this study counts survival time as a risk factor in predicting the binary outcome of death.

2.2. Method

A machine learning approach is used to create a linear mapping between risk factors and mortality and to assess the relative importance of each feature. Specifically, logistic regression is applied as a statistical modeling technique to forecast the likelihood of event outcomes from the input variables. This method utilizes the sigmoid function to convert any real-valued number to a probability within the (0, 1) interval, effectively transforming the linear combination of input variables into a probability. The model is formally defined as follows:

$$P(Y = 1|X) = \frac{1}{1+e^{-z}} \tag{1}$$

Where z is the linear combination of the risk factors, represented as:

$$\mathrm{z} = \beta_0 + \beta_1 \mathrm{x}_1 + \beta_2 \mathrm{x}_2 + \cdots + \beta_n \mathrm{x}_n. \tag{2}$$

$\beta + \beta_1 + \cdots + \beta_n$ represent the trainable model parameters and $\mathrm{x} + \mathrm{x}_1 + \cdots + \mathrm{x}_n$ are the corresponding risk factors. The model outputs a probability of the subject being classified as $\mathrm{Y} = 1$, namely the occurrence of death in the context of this study. The final classification is based on a threshold value to distinguish the probabilistic predictions. In this study the threshold value is set to be 0.5, represented as:

$$if\ P(Y = 1|X) \geq 0.5, dead \quad (3)$$

$$if\ P(Y = 1|X) < 0.5, live \quad (4)$$

Logistic regression is based on a simple assumption that the feature space is linearly separable. Also, in addition to the category prediction, the model provides the corresponding probability, greatly enhancing the model's interpretability.

3. Experiments and results

3.1. Experimental setups

The model is implemented with the Sklearn package and trained on an Intel Core i7 CPU. Before the training, the Pandas package is utilized to clean the dataset by checking and filling the missing values. After that, the cleaned dataset is categorized into DataFrame format for modeling training. To avoid the randomness caused by data splitting, the 10-fold cross-validation is employed. The data is partitioned into ten equal or nearly equal portions. In each round, the model is trained using nine of these portions, while the last portion is reserved for testing. This procedure is repeated ten times and the average of the results across all ten rounds is then computed as the outcome. To evaluate the model’s predictive performance, two metrics are utilized: accuracy and area under the receiver operating characteristic curve (AUROC). Accuracy represents the ratio of correctly predicted outcomes relative to the total number of cases evaluated. AUROC quantifies the ability of a model to distinguish between classes under different threshold values.

3.2. Results

The accuracy of the cross-validation is 0.806±0.129, and the AUROC is 0.945±0.052. The 95% confidence interval for the parameter is calculated using the normal approximation method based on the standard deviation. The accuracy lies in [0.726, 0.886] while AUROC has a narrower confidence interval of [0.912, 0.977]. These metrics indicate the good performance of the logistic regression model in distinguishing between survivors and non-survivors in heart failure patients, with high discriminatory power and reliability. The feature importance analysis from the logistic regression model, as depicted in Figure 1, highlights several key factors associated with heart failure mortality. The X-axis indicates the intensity of feature importance ranging from 0 to 1. The Y-axis lists all 12 risk factors. In addition, standardized coefficients indicate the relative importance of each risk factor, with age emerging as the most significant predictor (coefficient = 1.000). Serum sodium and serum creatinine levels also stand out as critical factors (coefficients = 0.651 and 0.633, respectively), indicating the significance of these biomarkers in clinical practice.

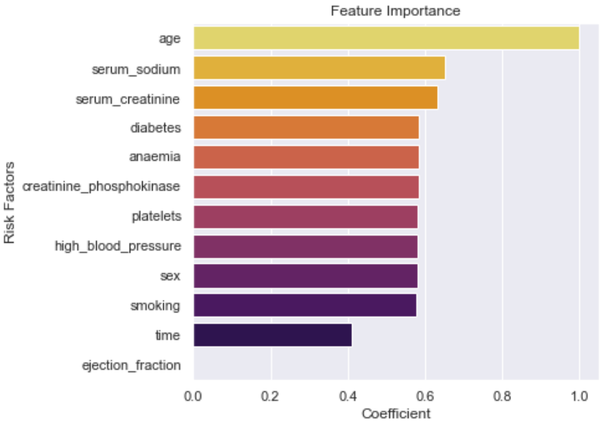

Figure 1. Feature Importance of Risk Factors in Predicting Mortality in Heart Failure Patients (Original).

Age's paramount importance aligns with clinical literature, which consistently identifies older age as a major risk factor for adverse outcomes in heart failure patients [9]. The heightened vulnerability of older individuals is possibly due to a combination of factors, including diminished physiological reserves, increased prevalence of comorbidities, and age-related decline in cardiac function [10]. Moreover, serum sodium and serum creatinine are well-established biomarkers used for the detection of heart failure. To be specific, serum sodium levels, indicative of electrolyte balance and fluid status, are critical in assessing the severity of heart failure and guiding treatment decisions. Hyponatremia, or low serum sodium levels, is related to worse outcomes and higher mortality rates in heart failure subjects. Serum creatinine, a marker of renal function, is equally important as renal impairment is common in heart failure and significantly affects prognosis. Elevated serum creatinine levels suggest renal dysfunction, which complicates heart failure management and is associated with increased mortality [11, 12].

In contrast, the ejection fraction (EF), which reflects the efficiency of the heart’s pumping action, is found to have a coefficient of 0.000, indicating no significant impact on mortality prediction in this model. Note that a coefficient of 0 does not mean EF is completely unrelated, but rather the least influential factor in this model. This finding is intriguing, as ejection fraction is traditionally considered a crucial factor in heart failure diagnosis [13]. Further analysis of the dataset is conducted to investigate the contribution of EF to the survival prediction. It is hypothesized that because all subjects included in the study had been diagnosed with heart failure, the differentiation of EF values between different risk groups is less pronounced. As shown in Figure 2, a higher proportion of deceased individuals had an EF of less than 40%, while the 40-50% EF range included more survivors. This distribution indicates that EF, while a critical diagnostic tool for heart failure, might not significantly influence mortality predictions within a population already diagnosed with the condition. This finding aligns with our model's results, where EF's contribution to mortality classification is minimal.

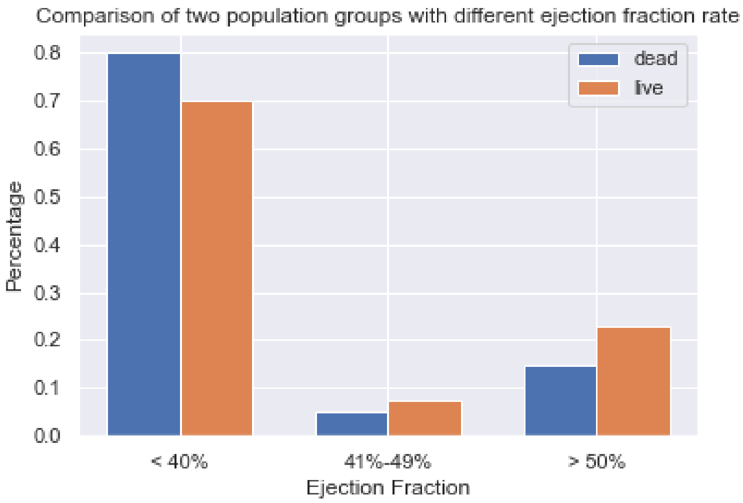

Figure 2. Comparison of Mortality Rates by Ejection Fraction Categories in Heart Failure Patients (Original).

One possible reason for the irrelevance of EF is that the logistic regression model, while effective, oversimplifies the complex interactions between various risk factors. More advanced deep learning algorithms such as neural networks could potentially establish the non-linear modeling of medical data more effectively, enhancing predictive accuracy [14]. The minimal impact of EF in the predictive model raises important questions. Given its clinical importance, further research is to explore the conditions under which EF might still hold significant predictive value, perhaps in conjunction with other advanced imaging biomarkers or through longitudinal studies assessing changes in EF over time [15].

Future research aims to expand datasets to include more diverse populations, apply advanced machine learning techniques to capture non-linear relationships, conduct longitudinal studies to track changes in EF and other biomarkers, and integrate additional biomarkers to enhance predictive accuracy. By addressing these areas, the aim is to refine predictive models and improve patient management strategies, ultimately enhancing outcomes for individuals with heart failure.

4. Conclusion

This study's findings emphasize the complexity of risk factors influencing heart failure mortality. Age, serum sodium, and serum creatinine levels were identified as the most significant predictors of mortality, aligning with existing clinical literature regarding their roles in HF management. The logistic regression model demonstrated robust performance, indicating its suitability for clinical application in risk prediction. The ejection fraction scored the lowest feature importance, suggesting that HF mortality prediction may benefit from incorporating more sophisticated machine learning techniques and additional biomarkers to capture the nuanced interplay of risk factors.

The plan should focus on including a more diverse patient population, applying advanced predictive models, and conducting longitudinal studies to further refine the predictive accuracy and enhance patient management strategies in heart failure.

References

[1] Ponikowski, P., Anker, S. D., AlHabib, K. F., Cowie, M. R., Force, T. L., Hu, S., Jaarsma, T., Krum, H., Rastogi, V., Rohde, L. E., Samal, U. C., Shimokawa, H., Siswanto, B. B., Sliwa, K., & Filippatos, G. (2014). Heart failure: preventing disease and death worldwide. ESC Heart Failure, 1(1), 4-25.

[2] Collaborators, G. B. D. (2018). Global, regional, and national incidence, prevalence, and years lived with disability for 354 diseases and injuries for 195 countries and territories, 1990-2017: a systematic analysis for the Global Burden of Disease Study 2017. The Lancet, 392(10159), 1789-1858.

[3] Jones, N. R., Roalfe, A. K., Adoki, I., Hobbs, F. D. R., & Taylor, C. J. (2019). Survival of patients with chronic heart failure in the community: a systematic review and meta‐analysis. European Journal of Heart Failure, 21(11), 1306-1325.

[4] Cook, C., Cole, G., Asaria, P., Jabbour, R., & Francis, D. P. (2014). The annual global economic burden of heart failure. International Journal of Cardiology, 171(3), 368-376.

[5] Dunlay, S. M., Weston, S. A., Jacobsen, S. J., & Roger, V. L. (2009). Risk factors for heart failure: a population-based case-control study. The American Journal of Medicine, 122(11), 1023-1028.

[6] Khatibzadeh, S., Farzadfar, F., Oliver, J., Ezzati, M., & Moran, A. (2013). Worldwide risk factors for heart failure: a systematic review and pooled analysis. International Journal of Cardiology, 168(2), 1186-1194.

[7] 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), 16.

[8] Ahmad, T., Munir, A., Khattak, F., Bhatti, S., & Muhammad, A. (2017). Survival analysis of heart failure patients: A case study. PLOS ONE, 12(7), e0181001.

[9] Stewart, S., MacIntyre, K., Capewell, S., & McMurray, J. J. (2003). Heart failure and the aging population: an increasing burden in the 21st century? Heart, 89(1), 49-53.

[10] Fulop, T., Larbi, A., Witkowski, J. M., McElhaney, J., Loeb, M., Mitnitski, A., & Pawelec, G. (2010). Aging, frailty, and age-related diseases. Biogerontology, 11(5), 547-563.

[11] Gheorghiade, M., Abraham, W. T., Albert, N. M., Greenberg, B. H., O'Connor, C. M., She, L., Stough, W. G., Yancy, C. W., & Young, J. B. (2007). Relationship between admission serum sodium concentration and clinical outcomes in patients hospitalized for heart failure: an analysis from the OPTIMIZE-HF registry. European Heart Journal, 28(8), 980-988.

[12] Metra, M., Cotter, G., Davison, B. A., Felker, G. M., Filippatos, G., Greenberg, B. H., Pang, P. S., Ponikowski, P., Teerlink, J. R., & Voors, A. A. (2012). The role of the kidney in heart failure. European Heart Journal, 33(17), 2135-2142.

[13] Sanderson, J. E. (2007). Heart failure with a normal ejection fraction. Heart, 93(2), 155-158.

[14] Biau, G. (2012). Analysis of a random forests model. The Journal of Machine Learning Research, 13(1), 1063-1095.

[15] Cheng, J. M., Akkerhuis, K. M., Meilhac, O., Oemrawsingh, R. M., Garcia-Garcia, H. M., Serruys, P. W., & Boersma, E. (2013). Biomarkers of heart failure with normal ejection fraction: a systematic review. European Journal of Heart Failure, 15(12), 1350-1362.