DOI: https://doi.org/10.53350/pjmhs2022161177

ORIGINAL ARTICLE

Prediction of the Outcome of Pakistani Heart Failure Patients by Various Supervised Machine Learning Methods

SANA SAEED[1], MAHAM FAHEEM[2], KANWAL SALEEM[3], NIMRA ISHAQ[4]
[1]Assistant Professor, College of Statistical and Actuarial Sciences, University of the Punjab
[2]Lecturer, College of Statistical and Actuarial Sciences, University of the Punjab
[3,4]College of Statistical and Actuarial Sciences, University of the Punjab
Correspondence to Sana Saeed, Email: sana.stat@pu.edu.pk

ABSTRACT

Aim: To foresee the outcome of heart failure(HF) in Pakistani patients with potential predictors and through various machine learning (ML) methods.
Study design:The secondary data of Pakistani patients is taken from the UCI repository in which a cross-sectional, analytical study was planned.
Place and duration: This data was collected in April-December, 2015 at the Institute of Cardiology and Allied hospital Faisalabad-Pakistan.
Methodology: The data set consisted of299 patients distributed among male (194) and female patients (105). Ages, serum sodium (SS), serum creatinine (SC), gender, smoking, high blood pressure (HBP), ejection fraction (EF), anemia, platelets, Creatinine Phosphokinase (CPK), and diabetes were considered as the potential predictors for predicting the outcome of HF.The data set was analyzed with the help of various machine learning (ML) predictive models including Logistic regression (LR), K-nearest neighbor (KNN), and Decision trees (DT).
Results: The ages of the patients were within 60.833±11.894 years. Out of 299 patients, 129 were anemic, 105 had high blood pressure (HBP), and 96 had a smoking history. A statistical model was estimated by applying LR which assisted us in identifying the significant predictors. The sensitivity of the LRwas observed to be 92.1%, whereas 85.6% of the outcome of HF patients was correctly predicted by this model (LR) and DT achieved89.6% prediction accuracy.
Conclusion:Since HF is a substantial reason for deaths in Pakistan. Therefore, the identification of its potential risk factors and its accurate prediction by some modern tools are highly demanded. This study applied ML tools for the said task and concluded that among all the fitted ML models, DT predicted the correct outcome for HF patients proficiently.
Keywords: Heart failure, machine learning, logistic regression, k-nearest neighbor, decision trees

INTRODUCTION

Heart attacks, strokes, and heart failure (HF) are types of cardiovascular diseases (CADs)[1]. HF/congestive HFis a condition that occurs when the muscles in the heart cannot pump adequate blood. Shortness of breath occurs during this condition due to filling the fluid in the lungs of the body. HF is not a condition when the heart completely stops working;rather, it is a condition of developing stiffness and thickness in it[2]. Mortality due to this ailment is approximately 25% in advanced nations and 80% in emerging nations[3]. People are extremely prone to CADs in the subcontinent, Asia which causesmany deaths[4]. However, in emerging nations, females are more at riskthan males[5].Several factors can add to the complications of HF such as aging, smoking, high blood pressure, etc.Research guided that certain characteristics of human beingssuch as gender, age, and spousal relationship may be allied with a higher risk for CHD[6]. The interaction of these changeable measures forcefully increases the hazard of HF with negative outcomes. In the UK, around £980 million per year is consumed on the administration of HF and the World Bank assessed that it cost globally $108 billion/per annum[7]. HF is assumed to be an ailment of senior persons. Though, the latest studies have specified that the HF strain among adults may be growing[8]. Thestarring role of behavioral hazardsin the expansion of heart diseases is also perceptible.For example, cigarette smoking, diabetes, hyperlipidemia, and hypertension played a vital part in the development of heart ailments[9]**.** The medical history of patients also played a momentous part in predicting the outcome of HF. Correspondingly, this medical ailment psychology affects patients if they developed HF. Bivol and Grib 2019 suggested that patients with renal damage faced this ailment differently than controls[10]. This could originatesadness, anxiety, and concentratedliveliness related to heart disorders and renal dysfunction as well. Clinically, the HF can be segmented into two groups depending on the EF value[11–15].

In most developing nations, low education, unemployment, and many other factors contribute to low quality of life. Hence, certain ailments including heart diseases are much more common among aged patients with poor mental health[11]. Pakistan Demographic survey publicized that cardiac ailments comprising heart attacks and HF alone were liable for 14.74 or 221,100 expiries in the country whereas strokes caused deaths in 6.45% or 96,750 population.However, according to WHO,29% of the entire expiries happened due to cardiovascular disease in Pakistan which is comprised of both heart diseases and strokes. WHO also discussed that the prevailing condition in Pakistan revealed a serious fact that preventable diseases are now causing more deaths than infectious diseases like Covid-19, pneumonia, and others[16].

A study conducted in Pakistan[17] exhibited that the wholeincidencelevel of cardiac disease was 6.2% and the elder women who were older than 30 years of age had a considerably increased hazard of heart attack than men. Also, the occurrence of stroke among women was higher than among men. These findings recommended that the prevalence of heart disease was higher in women than men in the country.

Quantitative analysis of HF data is done by numerous methods. Amongst them, the ML procedures are in foremost top positions because of the attractiveness and effectiveness of these approaches.

Knowing the significant role of HF in causing deaths nationwide, the accurate prediction of HF is highly looked-for by considering the significant predictors. Hence, significant predictors are first screened then the prediction will be done by ML methods comprised of LR, KNN, and DT.

METHODOLOGY

The data set for this study is tied up from UCI ML Source[18].

Received on 25-06-2022
Accepted on 15-10-2022

Place of Study: By using the data from[18] the current study is conducted at the College of Statistical and Actuarial Sciences, University of Punjab.

Duration of Study: The data collection was completed 9 months from April-December, 2015 through a cross-sectional study.

Sample Size: This data set is grounded on 299 HF clinical records of men (194) and women (105).

Inclusion criteria: All of the patients were above 40 years of age with confirmed left ventricular systolic dysfunction, and NYHA classes III and IV[18].

Exclusion: All the other patients not fulfilling the criteria were not considered for this study.

Data collection procedure: Since the patient's data were obtained through a cross-sectional study. Therefore, their follow-up time was between 285±4 days. Their ailment was detected from their reports of cardiac echo reports or records.

Their ages, gender, smoking status, and other clinical variables including SS, SC, BP, EF, anemia, platelets, Creatinine Phosphokinase (CPK), and diabetes were measured. A brief description of the previously mentioned variables is given below:

Age: ages of the patients are measured in years (continuous).

Anemia: shrinkage of red blood cells or hemoglobin (binary variable)

High blood pressure (HBP): presence and absence of hypertension (binary variable)

Creatinine phosphokinase (CPK): level of the CPK enzyme in the blood (mcg/L) (continuous variable)

Diabetes: Presence and absence of diabetes (binary variable)

Ejection fraction (EF): Percentage of blood leaving the heart at each contraction (percentages)

Platelets(P): platelets in the blood (kilo platelets/mL) (count)

Gender: woman or man (binary variable)

Serum creatinine (SC): Level of serum creatinine in the blood (mg/dL) (continuous)

Serum sodium (SS): Level of serum sodium in the blood (mEq/L) (continuous)

Smoking:Smoking status of the patients if the patient smokes or not (binary variable)

HF/Dependent variable: if the patient is deceased during the follow-up period(abinary variable)

Data Analysis: In this study, some variables are continuous and categorical, but the dependent variables are dichotomous. Therefore, we applied LR, KNN, and DT. A brief introduction to all these methods is provided below:

Logistic Regression: LR, a supervised ML model, is used where the dependent variable is binary; for example, pass/fail, yes/no, death/fails, etc. The likelihood is the ratio of the probability of the event of interest divided by the probability of the not-interesting event. Suppose that the dependent variable is represented by the Y . Then the LR model will be well-defined as the natural logarithm of this likelihood as a regression function of the P-predictors/explanatory variables ($X_1, X_2, ..., X_P$) (see Eq 1.)

$$\ln(\frac{Y}{1-Y}) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_P X_P \quad (1)$$

where β_0 is the intercept, and $\beta_1, \beta_2, ..., \beta_P$ are the regression coefficients. Logistic regression can also be used as the prediction technique, with the help of the obtained probabilities shown in Eq 2.

$$P = \frac{1}{1 + \exp(\beta_0 + \sum_{j=1}^{P} \beta_j X_j)^{-1}} \quad (2)$$

Thereafter, with the help of a predefined threshold allocation and prediction are done [19].

K nearest neighbor: Another powerful ML prediction method is recognized as KNN. The application of this technique is seen in various fields due to its simplicity. It isa powerful and less expensive technique. For the allocation and prediction, the following steps are adopted[20]:

1. Select K, the number of neighbors
2. Calculate the similarity measure of each data point from the unallocated data points of the training set.
3. Determine the number of labels of the data points having the shortest distance according to K.
4. Define the majority by seeing the comportment of the neighbors in the range of K.
5. Allocate the labels of the majority to the unclassified data points.

Decision tree: Another supervised ML processuseful for model building and prediction purposes is renowned as DTs. The chore is to shape a model which can positively regulate the class/label of the outcome variable. This process has universal recognition because of the easy preparation process. However, the over-complex arrangements discourage the exercise of enormous data [21].

Performance evaluation measures: Since the foremost objective of this research is to predict the outcome of HF. Therefore, a classification table is used to show the outcomes: observed and predicted [20].The following measures are used as performance evaluation measures.

1. Sensitivity
2. Specificity
3. Overall accuracy

RESULTS AND DISCUSSIONS

A statistical quantitative explanation of categorical variables is shown in Table 1. It can be observed from the table that 83 (40.9%) anemic patients having HF are alive whereas 46(47.9%) patients died after having HF. 40(41.7%) Patients with Diabetes having HF died. However, the majority (85) of diabetic patients are alive. Similar behavior can be seen for the other variables i.e., patients having HBP (66) and having a smoking history are alive after contracting HF.

A detailed description of the continuous variables is presented in Table 2 with a statistical summary including the minimum value (Min), maximum value (Max), their mean values (Mean), and standard deviation (SD) of the observations. The ages of most respondents fall within 60.833±11.894. The Creatinine phosphokinase (CPK) of HF patients falls between 581.84±970.288. Correspondingly, the ranges of Mean ±SD for the other variables can be observed in Table 2.

Stimated Model by Logistic Regression: After the detailed description of the data set, the statistical model is estimated to know about the significant variables affecting HF. Since the dependent variable (HF outcome) is binary. Therefore, the model is estimated by one of the renowned ML techniques i.e., LR. The estimated model is shown in Eq (3).

$$\begin{aligned}\ln\frac{\mathrm{y}}{1-\mathrm{y}} &= 9.63 + 0.74\mathrm{age} + 0.007\mathrm{Anemia} + 0.0001\mathrm{CPK}\\ &\quad - 0.14\mathrm{Diabetes} - 0.77\mathrm{EF} + 0.10\mathrm{HBP} + 0.0001\mathrm{P}\\ &\quad + 0.66\mathrm{SC} + 0.53\mathrm{Gender} - 0.06\mathrm{SS}\\ &\quad + 0.01\mathrm{Smoking} \qquad (3)\end{aligned}$$

and p-value are accessible from Table 4. It is perceived from Table 4 that only three variables (age (0.003), EF (0.000), SC (0.000)) are meaningfully affecting the HF and are affirmed to be highly significant at a 5% level of significance. However, the remaining variables could not express their statistical worth by ML. Another significant feature of LR is the interpretation of the variables in terms of odds ratio (OR). The last two columns of Table 4 are showing the ORs and confidence intervals of the variables. The OR for age i.e., exp (0.04) is 1.04 showing that increasing age will increase 1.04 times the risk of HF. Gender another variable showed a high risk of HF in males than in females. Anemic patients also showed an increased risk of HF (exp (0.007) =1.007). This value is close to one, indicating the insignificant role of this variable in the risk development of HF. Diabetes, EF, and SS showed a decreasing risk of HF by the presence or increasing level of these variables respectively. Contrary to this, the OR of HBP (1.10), SC (1.94), and smoking (1.01) showed an increased risk of HF with increasing the level of these variables.

Variation Explained: The total variation explained by the estimated model can be observed by the Nagelkerke R square which is R^2=0.568. This value showed that 56.8% variations in the response/dependent variable (HF outcome) areenlightened by the fitted LR model shown in Eq (3).

Table 1:Statistical quantitative description of the categorical predictors

Predictors		Outcome of HF					
		Alive		Death		Total	
		Count	%age	Count	%age	Count	%age
Anemia	**0**		59.1%	50	52.1%	170	57%
	1	83	40.9%	46	47.9%	129	43%
Diabetes	**0**	118	58.1%	56	58.3%	174	58%
	1	85	41.9%	40	41.7%	125	42%
HBP	**0**	137	67.5%	57	59.4%	194	65%
	1	66	32.5%	39	40.6%	105	35%
Gender	**0**	71	35.0%	34	35.4%	105	35%
	1	128	65.0%	62	64.6%	194	65%
Smoking	**0**	137	67.5%	66	68.8%	203	68%
	1	66	32.5%	30	31.3%	96	32%

Table 2: Statistical summary of the continuous variable

	N	Min	Max	Mean	S.D
Age	299	40	95	60.83	11.89
CPK	299	23	7861	581.84	970.28
EF	299	14	80	38.08	11.83
P	299	25100	850000	263358.02	97804.23
SC	299	0.50	9.40	1.39	1.03
SS	299	113	148	136.63	4.41

Table 3: Estimated model by Logistic Regression

Predictors	B	S.E.	Wald	df	P-value	Exp(B)	95%C.I.for EXP(B)	
							Lower	Upper
Age	0.04	0.01	9.00	1	**0.003**	**1.04**	1.01	1.08
Anemia	.007	0.36	0.00	1	0.98	**1.00**	0.49	2.04
CPK	0.00	0.00	1.56	1	0.21	1.00	1.00	1.00
Diabetes	-0.14	0.35	0.17	1	0.67	0.86	0.43	1.72
EF	-.077	0.01	22.04	1	**0.00**	0.92	0.89	0.95
HBP	0.10	0.35	0.08	1	0.77	**1.10**	0.54	2.23
P	0.00	0.00	0.40	1	0.52	1.00	1.00	1.00
SC	0.66	0.18	13.47	1	**0.00**	**1.94**	1.36	2.77
Gender	0.53	0.41	1.66	1	0.19	**1.70**	0.75	3.83
SS	-0.06	0.04	2.84	1	0.09	0.19	0.86	1.01
Smoking	0.01	0.41	0.001	1	0.97	**1.01**	0.45	2.27
Constant	9.67	5.60	2.97	1	0.08			

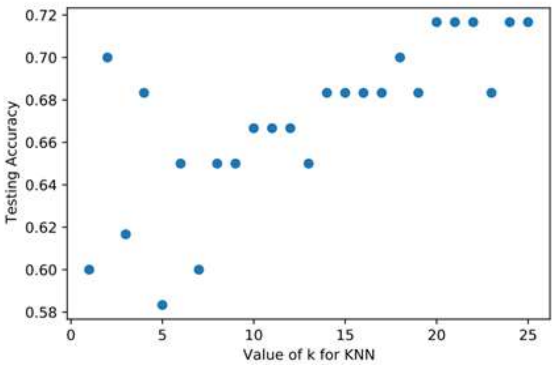
Prediction by Logistic Regression: As is already mentioned that LR is an influential ML procedure that has both kinds of capacities: modeling and prediction. Since the dependent variable (HF) has two possible outcomes: death and alive. Therefore, the prediction is done by the estimated model. A classification table is used (see Table 5) to show the prediction outcomes and actual outcomes of HF. The sensitivity of the model is 92.1 % which mentions the percentage of the participants perceived to fall in the target class and wasappropriatelyforeseen by the model to fall into that class. The specificity of the model is 71.9 %, which statesthe percentage of the participants/cases observed to fall in the non-target class and were appropriately predicted by the model to fall into that class and an overall correct prediction/accuracy by the model is 85.6%.

K Nearest Neighbor: KNN another most renowned ML procedure is applied for the outcome prediction of HF. Since its efficiency is entirely dependent on the values of K: the number of nearest neighbors. Therefore, the model is trained for the various values of K. Considering a compromise between the complexity and accuracy of the model, various models for a range of K values (2 to 20) are observed. It is noticed that the larger K values lead to a smoother decision boundary with a less complex model. However, smaller Kofferedrelatively complex models leading to overfitting (Figure 1). Therefore, the maximum accuracy achieved by this model is 70%.

Table 4: Classification table by LR

Perceived	Predicted HF		Corrected %
	Alive	Death	
Alive	187	16	92.1
Death	27	69	71.9
Overall			85.6

Figure 1: Accuracy of the KNN method



Decision Tree: The DT structure can be analyzed to gain further insight into the relationship between the features and the target to predict. So, apply the DT in python software. We get that the DT provides 89.6% accuracy.The sensitivity and specificity of the model are 81.6 % and 93.4% respectively (see Table).

Table 5: Classification table by DT

Perceived	Predicted HF		%age
	Alive	Death	
Alive	194	29	81.6
Death	5	71	93.4
Overall			89.6

Identification of the most efficient ML model: The prediction accuracies by all above mentioned three ML methods are shown in Table 7.We concluded that the maximum accuracy of KNN is 70% at K =20 which is not much noticeable, so we applied another method to predict our dependent variable. The accuracy of LR is 85.6% and the accuracy of the DT is 89.6%. Therefore, we concluded that the accuracy of the DT is more accurate for our data.

CONCLUSION

Owing to the growing percentage of heart diseases particularly HF in our country and the emerging usage of ML procedures, this research is directed to predict the outcome of HF patients by applying various ML techniques. For the said task, the secondary data of Pakistani HF patients was taken from a UCL repository. The data set had both demographic and clinical variables. A few of them were dichotomous and the remaining were continuous variables.

With the assistance of the LR model,the potential predictors were identified as the risk factors of HF. Inclusively; theestimated model could explain 56.8% variation in HF based on all the predictors. While observing the statistical significance of these predictors, it was noticed that among all, only three were found statistically significantly affectingthe variability in HF at a 5% level of significance. Whereas, all the remaining predictors were found to be insignificant in affecting HF. It was also noticed that the estimated model by LR 85.6% correctly predicted the outcome of HF among patients.

Thereafter, KNN and DT were also applied to the dataset. However, it was seen that the prediction ability of DT (89.6%) was much more appreciable than KNN (70%). Another worth-citing point was that among all three ML procedures, DT efficiently predicted the outcome of HF more than other procedures (LR and KNN).

Recommendation: It is recommended here that by casting various other ML procedures with multiple subsets of predictors, we can more efficiently perform prediction in the medical fields, particularly for heart diseases.

Conflict of interest: Nil

REFERENCES

1. Joshi SB. Exercise training in the management of cardiac failure and ischemic heart disease. Heart, Lung, and Circulation. 2007 Jan 1;16:S83-7.
2. Sun R, Liu M, Lu L, Zheng Y, Zhang P. Congenital heart disease: causes, diagnosis, symptoms, and treatments. Cell biochemistry and biophysics. 2015 Jul;72(3):857-60.
3. Sullivan K, Doumouras BS, Santema BT, Walsh MN, Douglas PS, Voors AA, Van Spall HG. Sex-specific differences in heart failure: pathophysiology, risk factors, management, and outcomes. Canadian Journal of Cardiology. 2021 Apr 1;37(4):560-71.
4. Mosca L, Benjamin EJ, Berra K, Bezanson JL, Dolor RJ, Lloyd-Jones DM, Newby LK, Pina IL, Roger VL, Shaw LJ, Zhao D. Effectiveness-based guidelines for the prevention of cardiovascular disease in women2011 update: a guideline from the American Heart Association. Circulation. 2011 Mar 22;123(11):1243-62.
5. Zile MR, Baicu CF, Gaasch WH. Diastolic heart failure—abnormalities in active relaxation and passive stiffness of the left ventricle. New England Journal of Medicine. 2004 May 6;350(19):1953-9.
6. Gaziano T, Reddy KS, Paccaud F, Horton S, Chaturvedi V. Cardiovascular disease. Disease control priorities in developing countries. The International Bank for Reconstruction and Development, Washington (DC). 2006.
7. Bivol E, Grib L. Psychosocial stress and quality of life in patients with type 2 cardiorenal syndrome. Arch Balkan Medical Union. 2019 Mar 1;54:147-54.
8. Yarmohammadian MH, Abdar-e-Esfahani M, Yoosefi AR, Shooshtarizadeh S. A study about the effects of health behavior on lifestyle changes of the people affected by cardiovascular diseases. Pakistan Heart Journal. 2005;38(1-2).
9. Christiansen MN, Køber L, Weeke P, Vasan RS, Jeppesen JL, Smith JG, Gislason GH, Torp-Pedersen C, Andersson C. Age-specific trends in incidence, mortality, and comorbidities of heart failure in Denmark, 1995 to 2012. Circulation. 2017 Mar 28;135(13):1214-23.
10. Virani SS, Alonso A, Benjamin EJ, Bittencourt MS, Callaway CW, Carson AP, Chamberlain AM, Chang AR, Cheng S, Delling FN, Djousse L. Heart disease and stroke statistics—2020 update: a report from the American Heart Association. Circulation. 2020 Mar 3;141(9):e139-596.
11. Al_Bairmani ZA, Ismael AA. Using Logistic Regression Model to Study the Most Important Factors Which Affects Diabetes for The Elderly in The City of Hilla/2019. InJournal of Physics: Conference Series 2021 Mar 1 (Vol. 1818, No. 1, p. 012016). IOP Publishing.
12. Sperandei S. Understanding logistic regression analysis. Biochemiamedica. 2014 Feb 15;24(1):12-8.
13. Nauta JF, Jin X, Hummel YM, Voors AA. Markers of left ventricular systolic dysfunction when LVEF is normal. Eur J Heart Fail. 2018 Dec;20(12):1636-8.
14. Pfeffer MA, Braunwald E. Treatment of heart failure with preserved ejection fraction: reflections on its treatment with an aldosterone antagonist. JAMA Cardiology. 2016 Apr 1;1(1):7-8.
15. Mesquita ET, Grion DC, Kubrusly MC, Silva BB, Santos ÉA. Phenotype mapping of heart failure with preserved ejection fraction. International Journal of Cardiovascular Sciences. 2018 Jul 19;31:652-61.
16. https://www.thenews.com.pk/print/951231-shocking-revelations-cardiovascular-diseases-top-killer-in-pakistan
17. Nanayakkara S, Kaye DM. Targets for heart failure with preserved ejection fraction. Clinical Pharmacology & Therapeutics. 2017 Aug;102(2):228-37.
18. https://archive.ics.uci.edu/ml/datasets/Heart+failure+clinical+records
19. Katz DH, Deo RC, Aguilar FG, Selvaraj S, Martinez EE, Beussink-Nelson L, Kim KY, Peng J, Irvin MR, Tiwari H, Rao DC. Phenomapping for the identification of hypertensive patients with the myocardial substrate for heart failure with preserved ejection fraction. Journal of cardiovascular translational research. 2017 Jun;10(3):275-84.
20. Jabbar MA. Prediction of heart disease using k-nearest neighbor and particle swarm optimization. Biomed. Res. 2017 Jan 1;28(9):4154-8.
21. Maji S, Arora S. Decision tree algorithms for prediction of heart disease. InInformation and communication technology for competitive strategies 2019 (pp. 447-454). Springer, Singapore.