Survival prediction in heart failure using machine learning algorithms.

Amit Tak
RVRS Medical College, Bhilwara, Rajasthan, India https://orcid.org/0000-0003-2509-2311
Puran Mal Parihar
Geetanjali Medical College, Udaipur, India
Shikha Mathur
Mahatma Gandhi Medical College and Hospital, Jaipur, Rajasthan, India
Bhaskar Das (drbhaskar23@gmail.com)
5, Air Force Hospital, Jorhat, Assam, India
Divyanshu Sawal
Accord Solutions, Pune, India

Research Article

Keywords: classifiers, ejection fraction, heart failure, machine learning, prediction

Posted Date: August 15th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1960150/v1

Abstract

Background: Heart failure is the final stage of various cardiovascular diseases. Statistical models and machine learning (ML) algorithms have been proposed to predict heart failure. However, the present study used ML classifiers to predict survival in heart failure patients.

Materials and Methods: The study dataset consists of a random sample of medical records of 299 heart failure patients. The dataset is publicly available on the Machine Learning Repository website of the University of California Irvine (UCI ML). Thirteen predictors and one response variable ('Event') were present in the database. Except for 'Time', other predictors were used in predicting survival in heart failure patients. Oversampling methods were employed to balance the dataset using the ROSE package in R. Predictors differed and contributed significantly to prediction and were used to train ML classifiers on MATLAB classifier application with 5-fold cross-validation. The performance metrics of the machine learning classifier were expressed as accuracy, the area under the receiver operator characteristic (AU-ROC) curve, sensitivity, and specificity.

Results: The predictors used to train machine learning classifiers were hypertension, age, creatinine concentration, CPK, ejection fraction, and sodium concentration. The best model was the ensemble-based Subspace K-nearest neighbor model. The accuracy, AU-ROC, sensitivity, and specificity were 89.5%, 93%, 87%, and 92%, respectively.

Conclusion: The present study used biostatistical tests and a logistic regression model to optimize feature selection. The features that contributed significantly to the logistic regression model were used to train machine learning classifiers. The study showed better performance metrics in predicting survival in heart failure patients.

Introduction

Cardiovascular diseases (CVDs) are the leading cause of mortality worldwide, with an estimated 17.9 million deaths yearly. CVDs are a group of heart and blood vessel disorders, including coronary heart disease, cerebrovascular disease, rheumatic heart disease, and other conditions. (1) Heart failure (HF) is a clinical syndrome characterized by shortness of breath, fatigue, and signs of edema and pulmonary crackles. (2) Heart failure is a risk factor for diseases such as atrial fibrillation, stroke, and coronary artery disease. Further, heart failure is a consequence of other diseases, including rheumatic, hypertensive, or coronary artery disease. (3) Ischemic heart disease is the leading cause of mortality. Global Burden of Disease (GBD) Study 2019 ranked risk factors for ischemic heart disease including high systolic blood pressure, dietary risks, high LDL cholesterol, air pollution, high body mass index, tobacco, high plasma blood glucose, kidney dysfunction, non-optimal temperature, other environmental risks, alcohol use, and low physical activity. (4) Heart failure is characterized by inadequate pumping of blood by the heart. Approximately one-half of the patients with heart failure have preserved ejection fraction (HFpEF) rather than reduced ejection fraction (HFrEF). (5) Statistical and machine learning models were used to predict

disease models in medicine. (6)(7)(8)(9)(10)(11)(12) Many models predict heart failure using various risk factors. (13)(14) However, the present study predicts the survival of heart failure patients using machine learning classification.

Materials And Methods

The present study predicts survival in heart failure patients using machine learning (ML) classifiers. The study dataset consists of a random sample of medical records of 299 heart failure patients collected at the Faisalabad Institute of Cardiology and the Allied Hospital in Faisalabad (Punjab, Pakistan) during April–December 2015. The dataset is publicly available on the Machine Learning Repository website of the University of California Irvine (UCI ML). Thirteen predictors and one response variable ('Event') were present in the database. Except for 'Time', other predictors were used in predicting survival in heart failure patients. The predictors included were gender, smoking, diabetes mellitus, hypertension, anemia, age, ejection fraction, serum creatinine, sodium concentration, platelets, and creatine phosphate kinase concentration. The response variable 'Event' has two levels: death (N= 203) and alive (N= 96). As the dataset was imbalanced, oversampling methods were employed using the ROSE package in R. The balanced data has 197 alive and 203 death cases. The features differed significantly between the two response levels using biostatistical tests. A logistic regression model using the stepwise method was fitted using significantly differed features to find the contribution of each predictor in survival prediction. The features contributing significantly were chosen for ML training and classification. The ML classifier application on MATLAB 2019a was used for classification with 5-fold cross-validation.

The classifiers used in this application include Decision Trees, Support Vector Machine (SVM), K-Nearest Neighbors (KNN), and ensemble learning classifiers. The decision trees include complex, medium, and simple tree classifiers. Similarly, the SVMs include linear, quadratic, cubic, fine Gaussian, medium Gaussian, and coarse Gaussian classifiers. The ensemble classifiers have boosted trees, bagged trees, and RUS boosted tree classifiers.

Statistical analysis

The quantitative data were expressed in median (IQR) and compared using the non-parametric Mann-Whitney's U test. The categorical data were expressed in percentage, and the relationship between discrete variables was found using a chi-squared test. The performance metrics of the machine learning classifier were expressed as accuracy, the area under the receiver operator characteristic (AU-ROC) curve, sensitivity, and specificity. R version 4.1.2software was used for balancing data between two groups, and JASP version 0.16.2 was used for statistical analysis. MATLAB Classification Learner application 2019a was used for training and prediction. The significance level was considered at 5%.

Results

The discrete variables including gender [χ^2 = 0.175 , p = 0.676], smoking [χ^2 = 0.044 , p=0.833], diabetes mellitus [χ^2 = 0.023, p = 0.878], anemia [χ^2 = 0.21 , p = 0.647] did not differ significantly except hypertension [χ^2 = 10.913, p < 0.001] between the two groups of heart failure. The patients who died due to heart failure had a higher percentage of hypertensives than alive patients. (Table 1) In case of continuous variables, the patients died of heart failure showed higher age [W=15415 , p < 0.001], creatinine levels [W= 11481 , p <0.001], and creatine phosphate kinase concentration [W=17609.5 , p = 0.038]. However, the same patients had a lower levels of ejection fraction [W= 28476 , p <0.001], and sodium concentration [W=25622.5 , p <0.001] compared to alive patients. However, platelets showed no significant differences between the two groups [W = 21337, p = 0.246]. (Table 2)

The variables significantly differed between the two groups, including hypertension, age, creatinine concentration, CPK, ejection fraction, and sodium concentration, were selected to fit a logistic regression model using the stepwise method. The model showed that all the variables contributed significantly to the prediction of the heart failure group. (Table 3) The variables mentioned above were used as predictors to train various Machine learning classifiers using the Classification learner app on MATLAB. The best model was the ensemble-based Subspace K-nearest neighbor model. The accuracy, AU-ROC, sensitivity, and specificity were 89.5%, 93%, 87%, and 92%, respectively. (Figures 1 and 2)

Discussion

Heart failure is a clinical syndrome seen in the terminal stage of many heart diseases. The heart's reduced pumping ability leads to the inadequate blood supply to the body. (2) However, nearly half of the patients with heart failure have preserved ejection fraction. (6) The present study generates results that differ from the original dataset curators study. The present study used more predictors to increase performance metrics. Ahmad et al. did a survival analysis of heart failure patients using the same dataset. Cox regression was used to model the mortality. Researchers found age, renal dysfunction, blood pressure, ejection fraction, and anemia were significant risk factors governing the mortality risk. (15) The present study used an ensemble-based Subspace K-nearest neighbor model with accuracy, AU-ROC, sensitivity, and specificity were 89.5%, 93%, 87%, and 92%, respectively.

Many studies were focused on the prediction of heart failure. (16) (17) (18) However, fewer studies focused on the prediction of adverse outcomes in heart failure patients. The present study used machine learning classifiers to predict survival in heart failure patients.

Smith et al., with an echocardiogram of 4696 patients, developed a risk model to predict the 5-year mortality risk or hospitalization in heart failure patients from 1999 to 2004. Researchers observed a 56% five-year risk of hospitalization for heart failure or death (95% confidence interval, 54% to 58%). The hazard ratios for echocardiogram data contributed statistically significantly to the model. However, echocardiogram findings did not improve prediction risk once demographic and clinical data were used. (19) Chicco et al. predict the survival of heart failure patients using machine learning algorithms. They develop a two-feature model using ejection fraction and serum creatinine. The Random Forest performed

best in the survival prediction, obtaining an accuracy of 74% and AU-ROC of 80%. (20). Newaj et al. used a Random Forest classifier to predict survival on the same dataset. Researchers found a maximum G-mean score of 76.83% with a sensitivity score of 80.21%. (21) Zeman et al. used two unsupervised models (K-Means and Fuzzy C-Means clustering) and three supervised classifiers (Random Forest, XGBoost, and Decision Tree) to demonstrate a superior performance of the supervised ML algorithms over unsupervised models. The proposed supervised stacked ensemble learning model can achieve an accuracy, precision, recall, and F1 score of 99.98%. (22) The performance metrics might depend on the number of independent predictors used to train ML classifiers. The present study used six predictors, including hypertension, age, serum creatinine, CPK, ejection fraction, and serum sodium concentration, to train machine learning classifiers.

Conclusion

The present study used biostatistical and logistic regression models to optimize feature selection. The features that contributed significantly to the logistic regression model were used to train machine learning classifiers. The study showed better performance metrics in predicting survival in heart failure patients.

Declarations

Research Quality and Ethics Statement- Research Quality and Ethics Statement- The present study with report quality, formatting, and reproducibility guidelines set forth by the EQUATOR Network. The data used was acquired from the UCL repository, so exempted from an Institutional Review Board / Ethics Committee review.

Conflicts of interest: Nil

References

1. Cardiovascular diseases [Internet]. [cited 2022 Aug 12]. Available from: https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1

2. HEART FAILURE (HF) | Harrison's Manual of Medicine [Internet]. [cited 2022 Aug 12]. Available from: https://harrisons.unboundmedicine.com/harrisons/view/Harrisons-Manual-of-Medicine/623723/all/HEART_FAILURE__HF_

3. Banerjee A, Mendis S. Heart Failure: The Need for Global Health Perspective. Curr Cardiol Rev [Internet]. 2013 May 17 [cited 2022 Aug 12];9(2):97. Available from: /pmc/articles/PMC3682401/

4. Roth GA, Mensah GA, Johnson CO, Addolorato G, Ammirati E, Baddour LM, et al. Global Burden of Cardiovascular Diseases and Risk Factors, 1990–2019: Update From the GBD 2019 Study. J Am Coll Cardiol. 2020 Dec 22;76(25):2982–3021.

5. Ho JE, Lyass A, Lee DS, Vasan RS, Kannel WB, Larson MG, et al. Predictors of new-onset heart failure differences in preserved versus reduced ejection fraction. Circ Hear Fail. 2013 Mar;6(2):279–86.

6. Ho JE, Enserro D, Brouwers FP, Kizer JR, Shah SJ, Psaty BM, et al. Predicting Heart Failure with Preserved and Reduced Ejection Fraction: The International Collaboration on Heart Failure Inftypes. Circ Hear Fail [Internet]. 2016 Jun 1 [cited 2022 Aug 8];9(6). Available from: https://www.ahajournals.org/doi/abs/10.1161/circheartfailure.115.003116

7. Bhandari S, Tak A, Singhal S, Shukla J, Shaktawat AS, Gupta J, et al. Patient Flow Dynamics in Hospital Systems During Times of COVID-19: Cox Proportional Hazard Regression Analysis. Front Public Heal. 2020 Dec 8;8:820.

8. Bhandari S, Shaktawat AS, Tak A, Patel B, Shukla J, Singhal S, et al. Logistic regression analysis to predict mortality risk in COVID-19 patients from routine hematologic parameters. Ibnosina J Med Biomed Sci [Internet]. 2020 [cited 2022 Mar 26];12(2):123. Available from: http://www.ijmbs.org/article.asp?issn=1947-489X;year=2020;volume=12;issue=2;spage=123;epage=129;aulast=Bhandari

9. Bhandari S, Singh Shaktawat A, Tak A, Patel B, Gupta J, Gupta K, et al. Independent Role of CT Chest Scan in COVID-19 Prognosis: Evidence From the Machine Learning Classification (1) (2) (3) (4) (5) (6) (7). Scr Med. 2021;52(4):273–81.

10. Tak A, Dia S, Dia M, Wehner TC. Indian COVID-19 Dynamics: Prediction Using Autoregressive Integrated Moving Average Modelling ARTICLE INFO (1) (2). Scr Med [Internet]. 2021 [cited 2022 Apr 13];52(1):6–14. Available from: https://github.com/CSSEGISand-

11. Tak A, Punjabi P, Yadav A, Ankhla M, Mathur S, Dave HS, et al. Prediction of Type 2 Diabetes Mellitus Using Soft Computing. Mod Med [Internet]. 2022 Jun 22 [cited 2022 Aug 12];29(2):135–43. Available from: https://medicinamoderna.ro/prediction-of-type-2-diabetes-mellitus-using-soft-computing/

12. Darshan Shah K, Pancharia A, Bamaniya H, Sharma A, Somani S, Tak A. Evaluation of Risk Factors for Ten-Year Coronary Heart Disease using Logistic Regression Modeling International Journal of Pharmaceutical and Clinical Research. Int J Pharm Clin Res [Internet]. 2022 [cited 2022 Aug 12];14(6):764–71. Available from: www.ijpcr.com

13. Yin T, Shi S, Zhu X, Cheang I, Lu X, Gao R, et al. A Survival Prediction for Acute Heart Failure Patients via Web-Based Dynamic Nomogram with Internal Validation: A Prospective Cohort Study . J Inflamm Res [Internet]. 2022 Mar 20 [cited 2022 Aug 12];15:1953–67. Available from: https://www.dovepress.com/a-survival-prediction-for-acute-heart-failure-patients-via-web-based-d-peer-reviewed-fulltext-article-JIR

14. Florea VG, Anand IS. Predicting survival in heart failure. Curr Cardiol Reports 2007 93 [Internet]. 2007 May [cited 2022 Aug 12];9(3):209–17. Available from: https://link.springer.com/article/10.1007/BF02938352

15. Ahmad T, Munir A, Bhatti SH, Aftab M, Raza MA. Survival analysis of heart failure patients: A case study. PLoS One [Internet]. 2017 Jul 1 [cited 2022 Aug 6];12(7):e0181001. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0181001

16. Bharti R, Khamparia A, Shabaz M, Dhiman G, Pande S, Singh P. Prediction of Heart Disease Using a Combination of Machine Learning and Deep Learning. Comput Intell Neurosci. 2021;2021.

17. Tripoliti EE, Papadopoulos TG, Karanasiou GS, Naka KK, Fotiadis DI. Heart Failure: Diagnosis, Severity Estimation and Prediction of Adverse Events Through Machine Learning Techniques. Comput Struct Biotechnol J. 2017 Jan 1;15:26–47.

18. Di Tanna GL, Wirtz H, Burrows KL, Globe G. Evaluating risk prediction models for adults with heart failure: A systematic literature review. PLoS One [Internet]. 2020 Jan 1 [cited 2022 Aug 12];15(1):e0224135. Available from: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0224135

19. Smith DH, Johnson ES, Thorp ML, Yang X, Petrik A, Platt RW, et al. Predicting Poor Outcomes in Heart Failure. Perm J [Internet]. 2011 Dec 1 [cited 2022 Aug 8];15(4):4. Available from: /pmc/articles/PMC3267558/

20. Chicco D, Jurman G. Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone. BMC Med Inform Decis Mak [Internet]. 2020 Feb 3 [cited 2022 Aug 6];20(1):1–16. Available from: https://bmcmedinformdecismak.biomedcentral.com/articles/10.1186/s12911-020-1023-5

21. Newaz A, Ahmed N, Shahriyar Haq F. Survival prediction of heart failure patients using machine learning techniques. Informatics Med Unlocked. 2021 Jan 1;26:100772.

22. Zaman SMM, Qureshi WM, Raihan MMS, Monjur O, Shams A Bin. Survival Prediction of Heart Failure Patients using Stacked Ensemble Machine Learning Algorithm. 2021 Aug 30 [cited 2022 Aug 12]; Available from: https://arxiv.org/abs/2108.13367v1

Tables

Table 1. Comparison of discrete features in dead and alive patients of heart failure.

Category	Levels	Alive	Death	χ^2 value	p-value
Gender	Female	71(34.975)	65(32.995)	0.175	0.676
	Male	132(65.025)	132(67.005)		
Smoking	Non-smoker	137(67.488)	131(66.497)	0.044	0.833
	Smoker	66(32.512)	66(33.503)		
Diabetes Mellitus	Absent	118(58.128)	116(58.883)	0.023	0.878
	Present	85(41.872)	81(41.117)		
Hypertension	Absent	137(67.488)	101(51.269)	10.913	< .001
	Present	66(32.512)	96(48.731)		
Anemia	Absent	120(59.113)	112(56.853)	0.21	0.647
	Present	83(40.887)	85(43.147)		

Table 2. Comparison of continuous features in dead and alive patients of heart failure

		Median	IQR*	W	p-value
Age	Alive	60	15	15415	< .001
	Death	65	22		
Ejection fraction	Alive	38	10	28476	< .001
	Death	30	18		
Sodium	Alive	137	4.5	25622.5	< .001
	Death	136	5		
Creatinine	Alive	1	0.3	11481	< .001
	Death	1.3	0.83		
Platelets	Alive	263000	82500	21337	0.246
	Death	255000	122000		
CPK	Alive	245	473	17609.5	0.038
	Death	418	439		

*IQR : Interquartile range; W: test statistics

Table 3. Estimated parameters of the Logistic regression model using predictors with response variable as survival in heart failure patients.

				Wald Test		
Parameter	**Estimate**	**Standard Error**	**z**	**Wald Statistic**	**df***	**p-value**
(Intercept)	8.89	3.791	2.345	5.5	1	0.019
Ejection.Fraction	-0.063	0.011	-5.723	32.758	1	< .001
Serum creatinine	0.609	0.163	3.733	13.938	1	< .001
Age	0.048	0.01	4.617	21.321	1	< .001
CPK	0	0	3.002	9.01	1	0.003
Blood pressure	0.761	0.242	3.147	9.903	1	0.002
Serum sodium	-0.081	0.028	-2.898	8.397	1	0.004

*df: degrees of freedom

Table 4. Performance metrics of the ensemble-based Subspace KNN model used for predicting survival in heart failure patients.

Performance metrics	Percentage
Accuracy	89.5%
AUC	93%
Sensitivity	87%
Specificity	92%

Figures

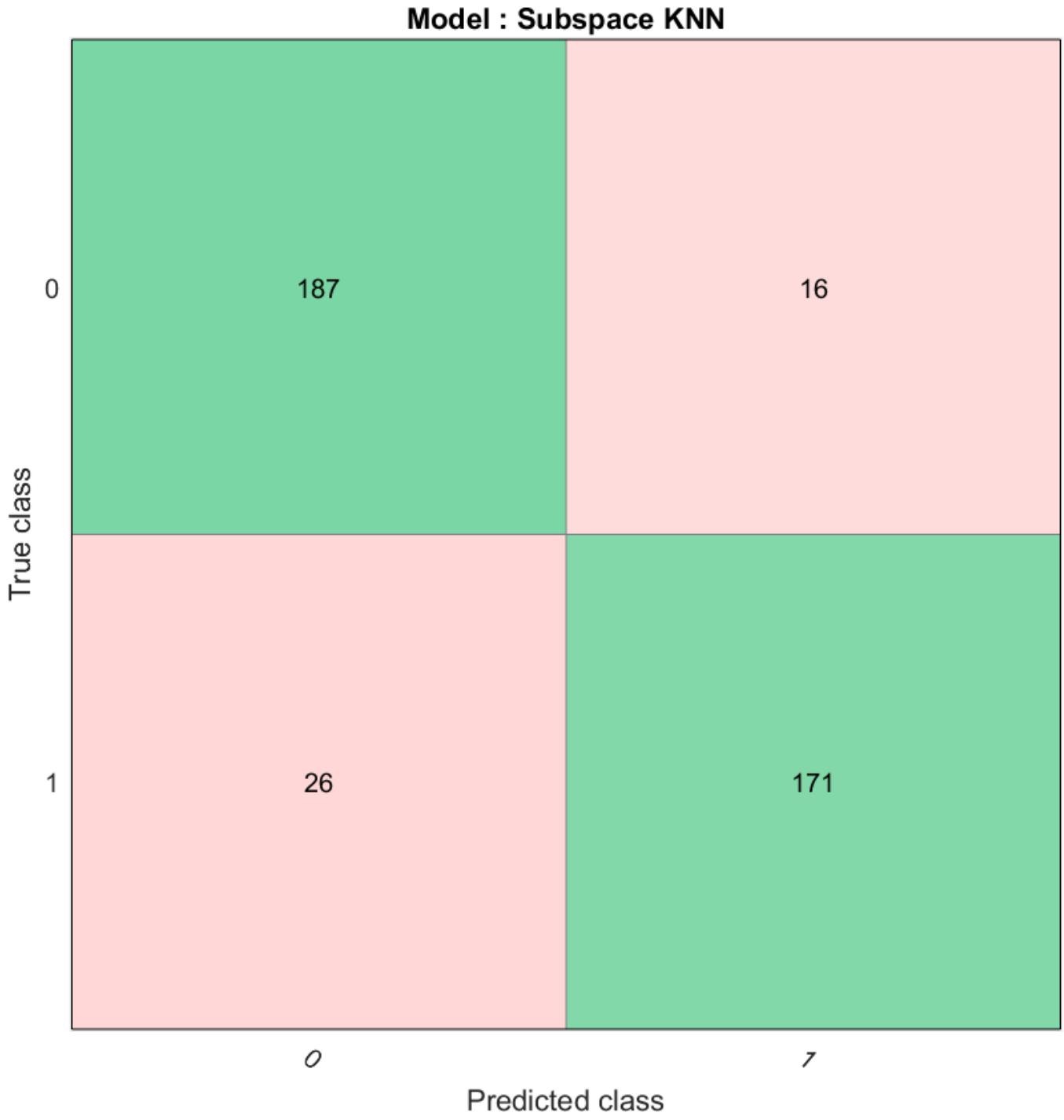

Figure 1

Confusion matrix for the ensemble-based Subspace KNN model.

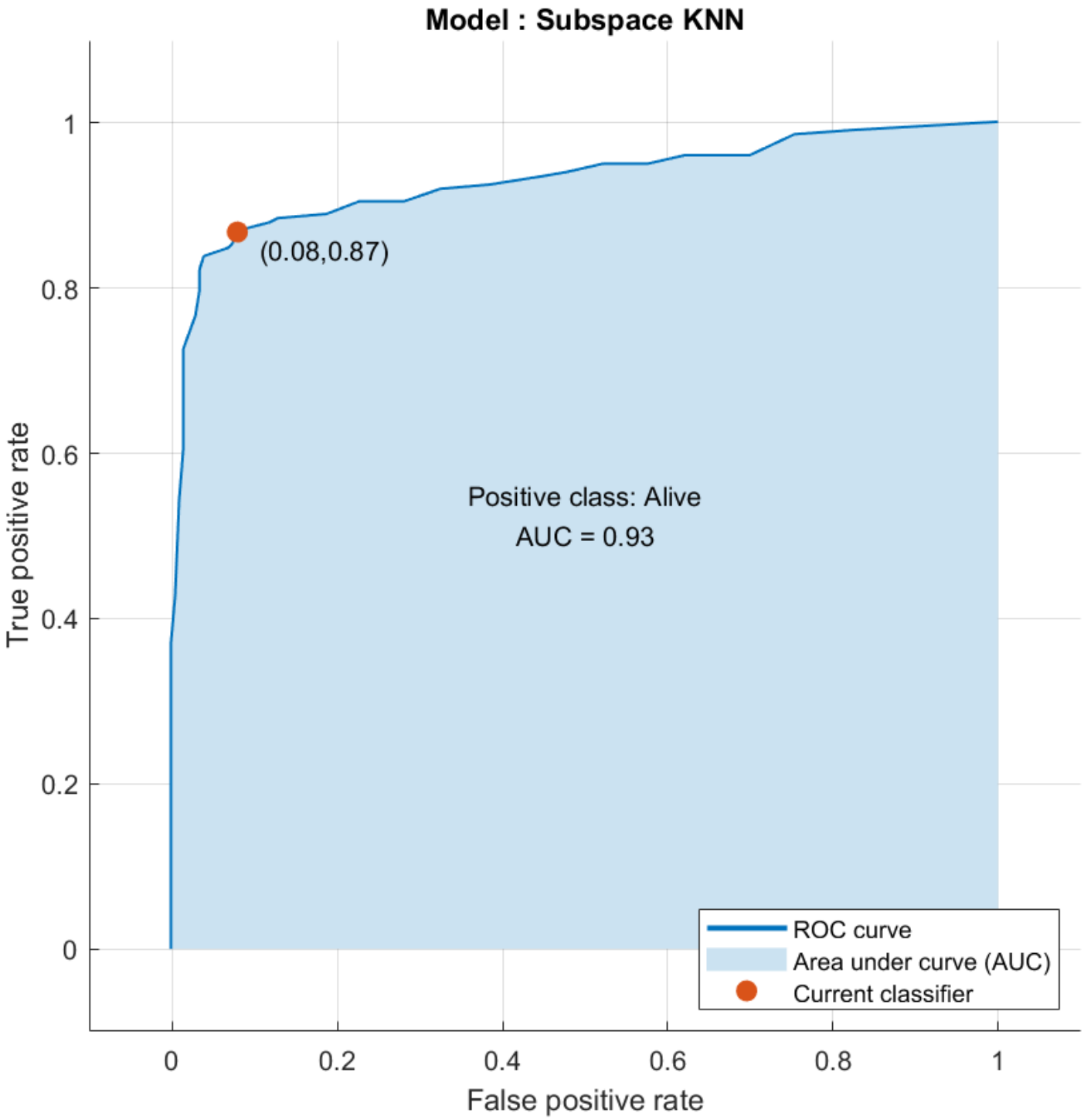

Figure 2

The area under the receiver operator characteristic curve for the ensemble-based Subspace KNN model.