Predicting heart failure from patient medical history using machine learning models based on the MIMIC-IV database

Winston Koh[1], Keyi Kang[1], Jialin Gao[1]
Client Partner: Jiancheng Ye, PhD[2]
Faculty Advisor: Yushu Shi, PhD[1]

[1] *Department of Population Health Sciences, Weill Cornell Medicine, New York, NY, USA*
[2] *Department of Emergency Medicine, Weill Cornell Medicine, New York, NY, USA*

Introduction

Heart failure is a critical public health challenge. It affects about 2% of the adult population worldwide, and it is a leading cause for hospitalization in the US[1]. The prevalence and incidence of heart failure are expected to increase due to factors such as an aging population and improvements in the treatment of acute cardiovascular diseases. This increase is expected to lead to large increases in expenditures related to heart failure treatment[2]. To enable clinicians to quickly treat heart failure patients and improve patient outcomes, we propose using prediction models to identify patients at high risk of experiencing heart failure.

Guidelines for diagnosis of heart failure recommend a variety of tests, including physical examination, electrocardiograms, and laboratory tests for biomarkers[3]. Prediction models may be useful for distinguishing low and high risk patients, thus relieving patients and healthcare providers of unnecessary visits and testing. Using statistical and machine learning techniques applied to electronic health record (EHR) data, the models could integrate various predictors, including demographics, medical history, vital signs, laboratory results, and treatment records. Prediction models can also be used to measure the importance of these risk factors, allowing

researchers to determine which factors contribute most significantly to heart failure. By leveraging predictive models as screening tools, healthcare providers can effectively manage patient risk, reduce the burden on healthcare systems, and ultimately enhance patient outcomes and quality of life.

Previous studies have employed various models to predict outcomes related to heart failure based on clinical variables. Several studies have used statistical models to predict incident heart failure. Kalogeropoulos et al.[4] created and validated a model which used nine routinely available clinical variables to predict incident heart failure in elderly individuals. Another outcome of interest is survival after heart failure. Taslimitehrani et al.[5] applied Contrast Pattern Aided Logistic Regression to create a prognostic model that predicted survival at several time points after heart failure diagnosis. Averbuch et al.[6] further explored the use of machine learning in various other applications related to heart failure, including early diagnosis, classification, and risk stratification.

Our study aims to to predict heart failure diagnoses by applying statistical and machine learning models to predictor variables found in EHR data. Other studies have also applied machine learning methods to predict heart failure diagnoses. Kwon et al.[7] developed a deep neural network to predict heart failure diagnoses using demographic and echocardiography features. Masetic et al.[8] used random forests to predict heart failure from electrocardiogram time series data. Alonso-Betanzos et al.[9] used support vector machines to classify the type of heart failure diagnosis using ventricular volume data. While machine learning models are capable of effectively modelling clinical outcomes, most of them prioritize predictive accuracy and lack emphasis on clinical interpretability. Enhancing model interpretability, especially

to uncover the most significant risk factors for heart failure, remains an important research direction in this field.

We propose that Shapley additive explanation values (SHAP) are a useful tool for understanding clinical prediction models that use complex machine learning methods despite their black-box nature. SHAP is a model-agnostic method for interpreting machine learning models by calculating the contribution of each feature to an individual prediction compared to the average prediction for a dataset[10]. Local explanations can then be combined to allow for global understanding of a model's behavior[11], which will provide insights that enhance trust in the prediction outputs. In this manner, we aim to develop models that bridge the gap between predictive accuracy and clinical usability.

Material and methods

Data Source

MIMIC-IV 2.2 is a freely available database of de-identified EHR data from patients admitted to Beth Israel Deaconess Medical Center, an academic medical center in Boston, MA, USA, from 2008 to 2019[12–14]. The outcome of interest was whether patients had any type of heart failure diagnosis at their last visit, which was identified by billed diagnosis codes (ICD 9: 428; ICD 10: I50). Predictor variables include patient demographics (genotypic sex, age, insurance, language, marital status, race), morbidity history (hypertension, atrial fibrillation, diabetes type 1, diabetes type 2, chronic obstructive pulmonary disease, asthma, liver disease, chronic kidney disease, malignant neoplasms, depression, osteoarthritis, anemia), medication history (enalapril, lisinopril, ramipril, carvedilol, metoprolol succinate, bisoprolol, furosemide, bumetanide, spironolactone, warfarin, apixaban, rivaroxaban), and the most recent vital signs and lab measurements (systolic and diastolic blood pressure, blood oxygen saturation, NT-

proBNP, creatinine, blood urea nitrogen, sodium, potassium, aspartate transaminase, alanine transaminase, troponin T, complete blood count).

Methods

Variables with excessive amounts of missing values were excluded, including white blood cell type counts, blood oxygen, NT-proBNP, and troponin T. Race/ethnicity categories were combined into US Census categories, and some census categories were combined due to small sizes[15]. Implausible values for vital signs and lab test results were removed from the dataset. Variables were compared between the groups with and without heart failure using a two-sample t-test for continuous variables and Fisher's exact test for categorical variables. Missing values were imputed by conditional mean imputation, in which regression models are used to impute a single value for each missing value[16].

Data was randomly split into three sets for training (72%), validation (8%), and testing (20%). We used several different classification algorithms to create models that predict whether patients experienced heart failure. These algorithms include logistic regression, linear discriminant analysis (LDA), naive Bayes, decision trees, gradient boosted trees, random forests, support vector machines, and neural networks. Oversampling was used to balance the distribution of the outcome variable in order to estimate a decision function that does not favor the majority class. Models were evaluated on the validation set based on their area under the ROC curve (AUC), Brier score, sensitivity, and specificity.

Feature importance was evaluated in each model using SHAP values. Features with larger magnitude SHAP values contribute more to a model's prediction. For each model, we calculated the SHAP values and used them to rank features based on their mean absolute value value across many predictions. The ten most important features across all models were selected and

used to fit a logistic regression model and a gradient boosted tree model. The performances of the models were evaluated on the test set.

Data was extracted from the MIMIC-IV database using DuckDB and cleaned using Python 3.12 and Polars[17,18]. Univariate statistical tests to compare variables between groups were done using the compareGroups package in R 4.4.1[19,20]. Imputation was done using miceforest[21]. Model fitting and evaluation was done using scikit-learn, XGBoost, and PyTorch[22–24]. SHAP values were calculated using the SHAP package[11].

Results

Study Cohort Characteristics

The MIMIC-IV 2.2 database contains information on a total of 180640 patients. 20354 patients experienced heart failure at their most recent visit, and 160286 patients were not diagnosed with heart failure. Significant differences were found in most variables used for modelling ($p < 0.05$). Univariate analysis (Table 1) shows that patients with heart failure were generally men (53% vs 46%), older (75 vs 55), on Medicare insurance (63% vs 28%), had a race that was White/European (73% vs 67%), and more likely to be widowed and less likely to be single (24% vs 8%, 21% vs 39%). For patient medical history, heart failure patients were more likely to be prescribed the various medications and diagnosed with the various morbidities. Heart failure patients also had significant differences in many different lab results, such as creatinine (1.64 vs 0.97 mg/dL), blood urea nitrogen (33 vs 16.8 mg/dL), aspartate transaminase (98.9 vs 54.1 IU/L), and alanine transaminase (58.7 vs 41 IU/L). Patients without heart failure had more missing values for medication history and lab test results (Table 2).

Table 1: Patient characteristics. Continuous variables show mean (SD). Categorical variables show N (%). Morbidity and medication variables: N(%) for variable = Yes.

	No Heart Failure; N=160286	**Heart Failure; N=20354**	**p-value**
Gender			<0.001
F	86104 (53.7%)	9571 (47.0%)	
M	74182 (46.3%)	10783 (53.0%)	
Insurance			<0.001
Medicaid	13519 (8.43%)	653 (3.21%)	
Medicare	44978 (28.1%)	12740 (62.6%)	
Other	101789 (63.5%)	6961 (34.2%)	
Language			<0.001
English	145409 (90.7%)	18082 (88.8%)	
Unknown	14877 (9.28%)	2272 (11.2%)	
Marital status			<0.001
Divorced	10035 (6.26%)	1529 (7.51%)	
Married	68872 (43.0%)	8682 (42.7%)	
Single	62885 (39.2%)	4309 (21.2%)	
Widowed	12650 (7.89%)	4817 (23.7%)	
Unknown	5844 (3.65%)	1017 (5.00%)	
Race			<0.001
Asian	7167 (4.47%)	414 (2.03%)	
Black	21259 (13.3%)	2265 (11.1%)	
Hispanic/Latino/South American	9381 (5.85%)	640 (3.14%)	
Native American/Pacific Islander	526 (0.33%)	73 (0.36%)	
Other	7158 (4.47%)	557 (2.74%)	
White/European	106966 (66.7%)	14943 (73.4%)	
Unknown	7829 (4.88%)	1462 (7.18%)	
Age	54.6 (20.0)	75.4 (13.1)	<0.001
Hypertension	63348 (39.5%)	10297 (50.6%)	<0.001
Atrial fibrillation	15784 (9.85%)	10890 (53.5%)	<0.001
Diabetes type 1	1950 (1.22%)	443 (2.18%)	<0.001
Diabetes type 2	25052 (15.6%)	8662 (42.6%)	<0.001
Chronic obstructive pulmonary disease	8412 (5.25%)	4814 (23.7%)	<0.001
Asthma	15827 (9.87%)	2735 (13.4%)	<0.001
Liver disease	9603 (5.99%)	2070 (10.2%)	<0.001

	No Heart Failure; N=160286	Heart Failure; N=20354	p-value
Chronic kidney disease	13115 (8.18%)	9263 (45.5%)	<0.001
Cancer	28552 (17.8%)	4846 (23.8%)	<0.001
Depression	29712 (18.5%)	4727 (23.2%)	<0.001
Osteoarthritis	10799 (6.74%)	3026 (14.9%)	<0.001
Anemia	40832 (25.5%)	11792 (57.9%)	<0.001
Enalapril	1433 (1.04%)	494 (2.44%)	<0.001
Lisinopril	23754 (17.2%)	8687 (42.9%)	<0.001
Ramipril	280 (0.20%)	106 (0.52%)	<0.001
Carvedilol	3094 (2.24%)	3760 (18.6%)	<0.001
Metoprolol succinate	13218 (9.57%)	8437 (41.7%)	<0.001
Bisoprolol	97 (0.07%)	46 (0.23%)	<0.001
Furosemide	25546 (18.5%)	16740 (82.7%)	<0.001
Bumetanide	271 (0.20%)	800 (3.95%)	<0.001
Spironolactone	3702 (2.68%)	3215 (15.9%)	<0.001
Warfarin	11177 (8.09%)	7096 (35.1%)	<0.001
Apixaban	2809 (2.03%)	1949 (9.63%)	<0.001
Rivaroxaban	1891 (1.37%)	881 (4.35%)	<0.001
Systolic BP	126 (17.9)	127 (20.8)	0.027
Diastolic BP	74.1 (11.4)	68.7 (12.7)	<0.001
Creatinine (mg/dL)	0.97 (0.81)	1.64 (1.43)	<0.001
Blood urea nitrogen (mg/dL)	16.8 (12.0)	33.0 (24.0)	<0.001
Sodium (mEq/L)	139 (3.46)	139 (4.54)	<0.001
Potassium (mEq/L)	4.14 (0.48)	4.27 (0.58)	<0.001
Aspartate transaminase (IU/L)	54.1 (413)	98.9 (711)	<0.001
Alanine transaminase (IU/L)	41.0 (178)	58.7 (288)	<0.001
HGB (g/dL)	12.2 (2.12)	10.7 (2.13)	<0.001
HCT (%)	36.8 (5.98)	33.2 (6.27)	<0.001
MCV (fL)	90.3 (6.46)	92.2 (7.31)	<0.001
MCH (pg)	29.8 (2.48)	29.6 (2.72)	<0.001
MCHC (g/dL)	33.1 (1.50)	32.1 (1.59)	<0.001
RDW (%)	14.1 (1.98)	15.8 (2.55)	<0.001
PLT (10^3/μL)	247 (99.9)	230 (110)	<0.001
WBC (10^3/μL)	8.51 (5.78)	9.43 (7.55)	<0.001
RBC (10^6/μL)	4.10 (0.72)	3.62 (0.75)	<0.001
Neutrophil (%)	68.0 (13.8)	73.7 (13.1)	<0.001
Lymphocyte (%)	22.4 (12.0)	15.9 (10.6)	<0.001
Monocyte (%)	6.63 (3.16)	6.96 (3.74)	<0.001

	No Heart Failure; N=160286	Heart Failure; N=20354	p-value
Eosinophil (%)	1.91 (2.07)	1.91 (2.43)	0.71
Basophil (%)	0.51 (0.39)	0.42 (0.38)	<0.001

Table 2: N (%) of missing values for variables before imputation. Some variables did not have any missing values: gender, insurance, language, marital status, race, age, and all morbidity variables.

	No Heart Failure; N=160286	Heart Failure; N=20354
Enalapril	22,175 (14%)	116 (0.6%)
Lisinopril	22,175 (14%)	116 (0.6%)
Ramipril	22,175 (14%)	116 (0.6%)
Carvedilol	22,175 (14%)	116 (0.6%)
Metoprolol succinate	22,175 (14%)	116 (0.6%)
Bisoprolol	22,175 (14%)	116 (0.6%)
Furosemide	22,175 (14%)	116 (0.6%)
Bumetanide	22,175 (14%)	116 (0.6%)
Spironolactone	22,175 (14%)	116 (0.6%)
Warfarin	22,175 (14%)	116 (0.6%)
Apixaban	22,175 (14%)	116 (0.6%)
Rivaroxaban	22,175 (14%)	116 (0.6%)
Systolic BP	71,078 (44%)	8,167 (40%)
Diastolic BP	71,211 (44%)	8,299 (41%)
Creatinine	6,424 (4.0%)	27 (0.1%)
Blood urea nitrogen	7,241 (4.5%)	26 (0.1%)
Sodium	7,901 (4.9%)	30 (0.1%)
Potassium	7,841 (4.9%)	24 (0.1%)
Aspartate transaminase	51,651 (32%)	2,640 (13%)
Alanine transaminase	50,679 (32%)	2,658 (13%)
HGB	4,200 (2.6%)	42 (0.2%)
HCT	4,046 (2.5%)	29 (0.1%)
MCV	4,219 (2.6%)	45 (0.2%)
MCH	4,220 (2.6%)	45 (0.2%)
MCHC	4,219 (2.6%)	45 (0.2%)
RDW	4,225 (2.6%)	45 (0.2%)
PLT	4,134 (2.6%)	34 (0.2%)
WBC	4,216 (2.6%)	44 (0.2%)

	No Heart Failure; N=160286	Heart Failure; N=20354
RBC	4,219 (2.6%)	45 (0.2%)
Neutrophil %	17,223 (11%)	1,051 (5.2%)
Lymphocyte %	17,220 (11%)	1,051 (5.2%)
Monocyte %	17,223 (11%)	1,051 (5.2%)
Eosinophil %	17,223 (11%)	1,051 (5.2%)
Basophil %	17,223 (11%)	1,051 (5.2%)

Model Development

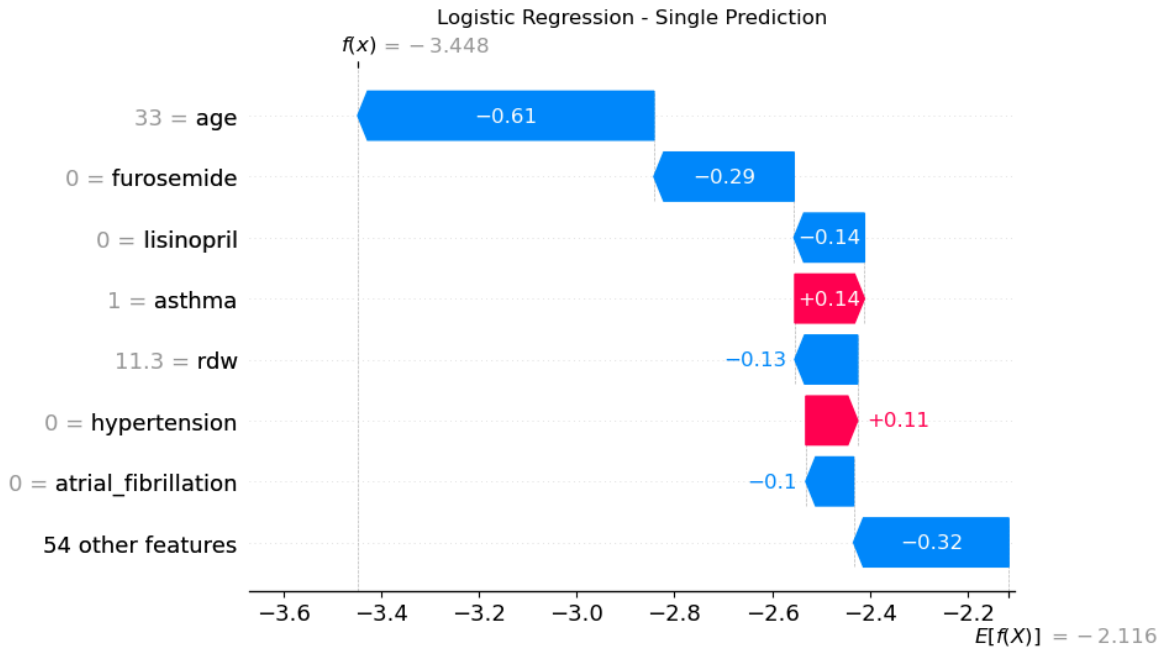
Eight models were fitted using different classification algorithms and evaluated on the validation set (Table 3). Most models achieved a similar level of performance based on their ROC-AUC, Brier score, sensitivity, and specificity, with the exception of linear discriminant analysis and naive Bayes.

Table 3: Average model performance on the validation set with 95% confidence intervals.

Model	AUC	Brier Score	Sensitivity	Specificity
Decision Tree	0.896	0.108	0.846	0.850
	[0.886, 0.906]	[0.104, 0.112]	[0.829, 0.863]	[0.844, 0.856]
Gradient Boosted Tree	0.935	0.104	0.874	0.853
	[0.930, 0.940]	[0.1, 0.108]	[0.858, 0.890]	[0.847, 0.859]
LDA	0.929	0.067	0.612	0.951
	[0.923, 0.935]	[0.064, 0.071]	[0.588, 0.635]	[0.947, 0.955]
Logistic Regression	0.931	0.107	0.871	0.852
	[0.926, 0.936]	[0.104, 0.111]	[0.855, 0.887]	[0.845, 0.858]
Naive Bayes	0.895	0.145	0.742	0.862
	[0.888, 0.902]	[0.139, 0.15]	[0.721, 0.763]	[0.856, 0.868]
Neural Network	0.939	0.105	0.888	0.837
	[0.934, 0.944]	[0.102, 0.109]	[0.873 , 0.903]	[0.831 , 0.844]
Random Forest	0.933	0.1	0.868	0.852
	[0.928, 0.938]	[0.097, 0.103]	[0.851, 0.884]	[0.846, 0.858]
Support Vector Machine	0.935	0.102	0.874	0.858
	[0.930, 0.941]	[0.099, 0.106]	[0.858, 0.890]	[0.852, 0.864]

SHAP values for six models with an acceptable level of performance (decision tree, gradient boosted trees, logistic regression, neural network, random forest, support vector machine) were calculated to understand how each feature contributes to the model's prediction (Figure 1). The ten most important features based on SHAP values were identified as age; blood urea nitrogen; red blood cell distribution width; history of atrial fibrillation, hypertension, and chronic kidney disease; and medication history for furosemide, lisinopril, and carvedilol. These features were used to fit two smaller models (logistic regression and gradient boosted trees). We then observed that hypertension unexpectedly reduces both models' predictions, so we fitted two additional models without this predictor (Figure 2). The final models performed similarly to the larger models and to each other (Table 5). Coefficients and odds ratios for the final logistic regression model are shown in Table 6.

(a)



(b)

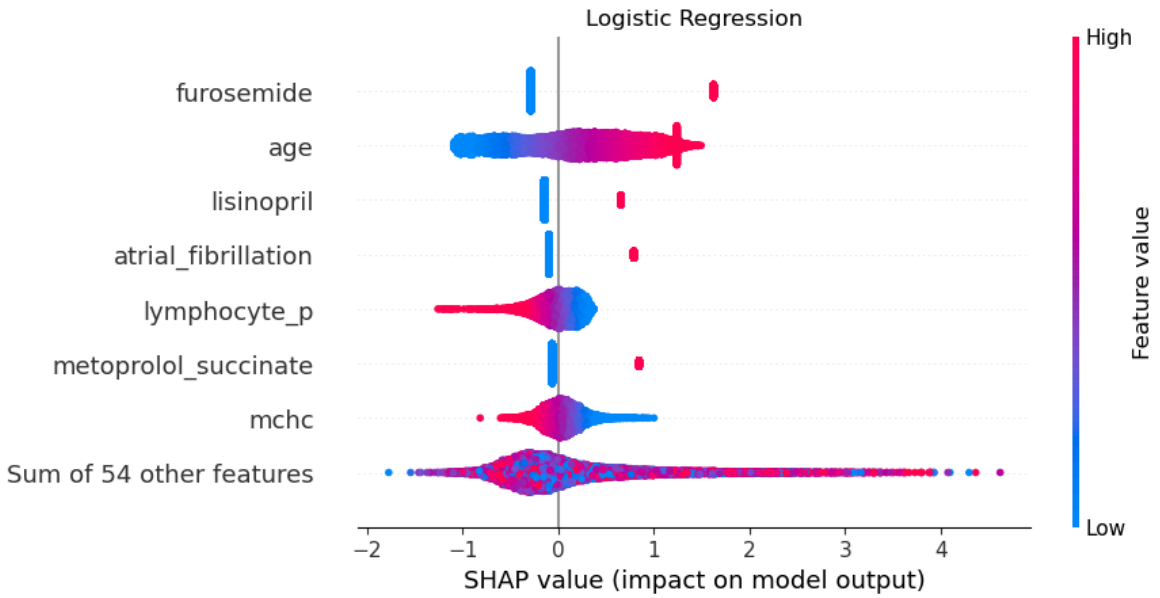


Figure 1: SHAP values for a logistic regression model. **(a)** Shapley values show the contribution of each feature to for a single prediction compared to the average prediction (predictions for this model are log odds). The sum of of the Shapley values is equal to the difference between the predicted value and the average value over many observations. **(b)** Shapley values for many predictions are aggregated to interpret a model's global behavior. The most important features are those with the highest mean absolute value.

Table 4: Top 10 features ranked by their their mean absolute SHAP values for six models (decision tree, gradient boosted tree, logistic regression, neural network, random forest, support vector machine).

Feature	DT	GBT	LR	NN	RF	SVM	Mean
Furosemide	1.0	2.0	1.0	1.0	1.0	1.0	1.167
Age	2.0	1.0	2.0	2.0	2.0	2.0	1.833
Atrial_fibrillation	3.0	4.0	4.0	4.0	3.0	4.0	3.667
Lisinopril	4.0	3.0	3.0	3.0	4.0	5.0	3.667
Metoprolol_succinate	7.0	7.0	6.0	8.0	7.0	6.0	6.833
Hypertension	6.0	6.0	10.0	11.0	13.0	3.0	8.167
RDW	8.0	5.0	19.0	6.0	6.0	11.0	9.167
Blood urea nitrogen	10.0	8.0	15.0	5.0	5.0	15.0	9.667
CKD	5.0	12.0	13.0	17.0	8.0	12.0	11.167
Carvedilol	9.0	10.0	24.0	16.0	12.0	9.0	13.333

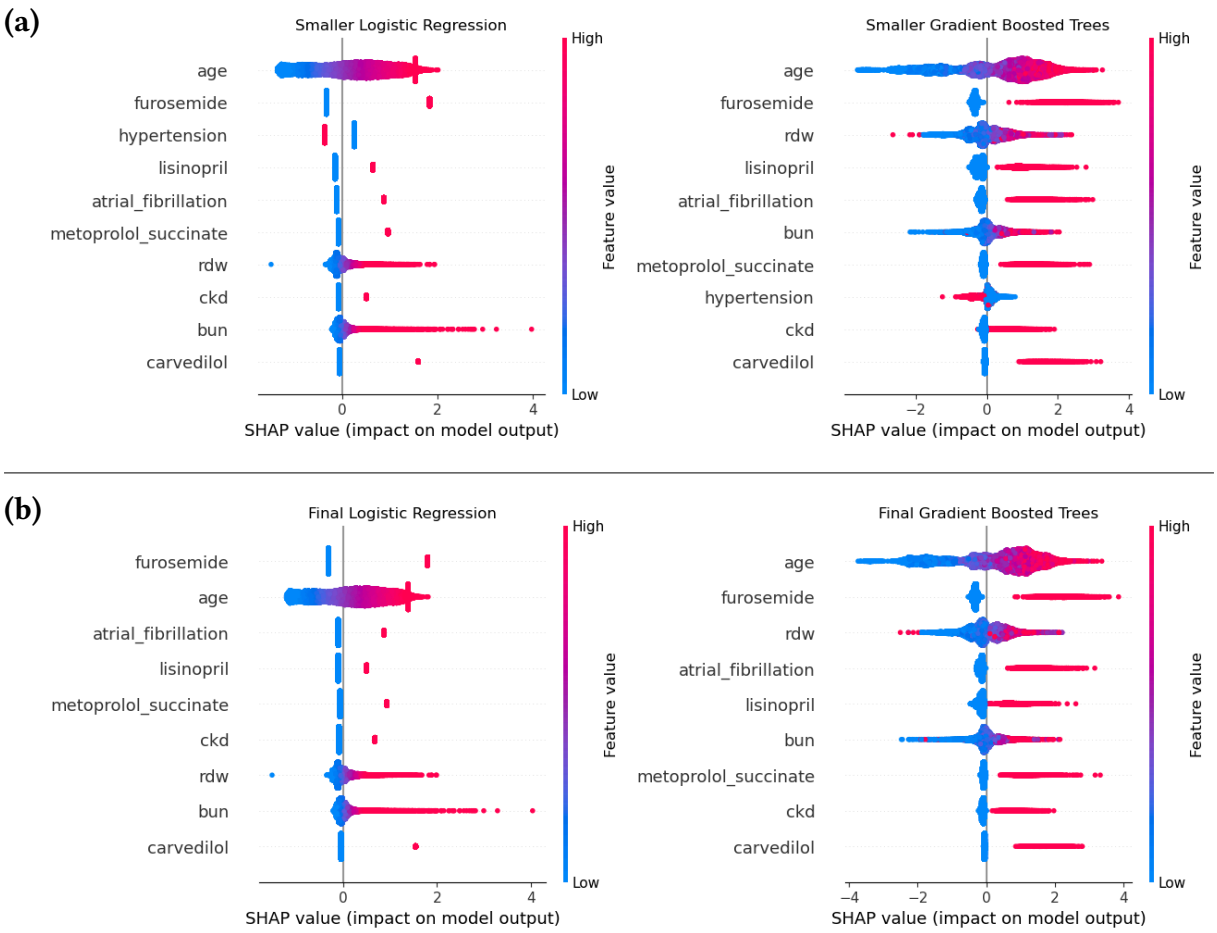


Figure 2: Summary plots for the SHAP values of the smaller models. **(a)** SHAP values for the ten most important features. A history of hypertension appears to decrease the value of the predictions in both models. This would result in a lower odds or probability of heart failure, contrary to our knowledge of hypertension being a risk factor. **(b)** SHAP values for models fitted on the same set of features without hypertension as a predictor.

Table 5: Average model performance with 95% confidence intervals on the test set for the final two models, fitted using the 10 most important features.

	AUC	**Brier Score**	**Sensitivity**	**Specificity**
Gradient Boosted Tree	0.934	0.116	0.891	0.829
	[0.931, 0.937]	[0.113, 0.118]	[0.882, 0.901]	[0.824, 0.833]
Logistic Regression	0.932	0.106	0.867	0.851
	[0.929 , 0.936]	[0.104, 0.108]	[0.857, 0.878]	[0.847, 0.855]
Gradient Boosted Tree	0.931	0.120	0.896	0.822
(No Hypertension)	[0.927, 0.934]	[0.118, 0.122]	[0.887, 0.906]	[0.817, 0.826]
Logistic Regression	0.931	0.108	0.869	0.847
(No Hypertension)	[0.927, 0.934]	[0.106, 0.111]	[0.859, 0.879]	[0.843, 0.851]

Table 6: Coefficients and odds ratios for predictors in the logistic regression model without hypertension, with 95% confidence intervals calculated by bootstrapping.

Predictor	Coefficient	Odds Ratio	Odds Ratio CI Lower	Odds Ratio CI Upper
Furosemide	2.113	8.276	7.458	9.228
Age	0.036	1.036	1.031	1.038
Atrial fibrillation	0.974	2.65	2.477	3.002
Lisinopril	0.605	1.832	1.61	1.991
Metoprolol succinate	0.997	2.711	2.52	3.002
RDW	0.108	1.114	1.063	1.126
Blood urea nitrogen	0.015	1.015	1.013	1.019
Chronic kidney disease	0.765	2.149	1.641	2.317
Carvedilol	1.592	4.911	4.48	9.293

Discussion

We used SHAP values to identify important features related to heart failure diagnoses, used these features to create two models, and achieved fairly good performance in internal validation. Since our models have a high negative predictive value, they could be utilized to filter out patients with low risk of heart failure and reduce unnecessary visits and testing. However, the models have a low positive predictive value due to the relatively low occurrence of heart failure, so a high proportion of positive results would be false positives if applied to the general population. The ten predictors selected for the final models are related to a patient's medical history as well as results from routine lab tests, all of which can be easily obtained from an EHR database. This differs from similar studies which often used electrocardiogram data to achieve a similar or higher levels of predictive performance. Kwon et al.[7] used a neural network on ECG data and achieved a 0.889 AUC score; Masetic et al.[8] achieved nearly 1.00 AUC score using random forests on time series data, albeit without external validation; and Yu et al.[25] used support vector machines to achieve a 0.98 accuracy.

When we used the ten most important features we found that hypertension counterintuitively reduced the outputs of both models, corresponding to an odds ratio for hypertension that is less than 1 in the logistic regression model. This would indicate that an individual with a history of hypertension has a lower odds of heart failure compared to a similar individual without hypertension, despite the fact that hypertension is more prevalent among heart failure patients than patients without heart failure. One possible explanation for this result is that some coefficients in our model may be biased because the model has not adjusted for a confounder variable or has inappropriately adjusted for a collider variable[26,27]. This does not affect the predictive capabilities of the models, as seen in the similar performance of models with and without hypertension (Table 5). If statistical inference for these coefficients is desired, further research and domain knowledge are necessary to specify a model that accounts for the causal relationships of the variables in order to get unbiased estimates.

We have demonstrated how SHAP can be used for interpretation of various types of prediction models. Different models tended to agree on which features contributed the most to their predictions as well as the direction of the contribution despite the differences in the models' assumptions and the algorithms used to fit them, though the features with lower ranks diverge greatly. We also showed how SHAP can identify the presence of possible confounders or colliders. Additional analysis of the SHAP values can be done to further understand the contributions of each feature such as studying feature interactions[28].

The findings of this study have other limitations. Our gradient boosted tree model is less interpretable compared to logistic regression while achieving a similar level of performance, so decision makers may not find it preferable to the logistic regression model. The methods

used to develop our models also have limitations. Model development and validation was done entirely by using random splitting data from a single medical center. External validation using data from a different location would provide stronger support for our final models[29]. The presence of missing values is another issue. We used single imputation to fill in the missing data, but this has issues. Single imputation treats the imputed values the same as observed values, and for statistical inference it can result in underestimation of standard errors and p-values that are too small[30]. Multiple imputation is preferred over single imputation for handling missing data, so it could be used as a more robust method for estimating the coefficients of a logistic regression model as well as the associated confidence intervals and p-values. However, further investigation is needed to determine how to best apply multiple imputation to create nonparametric prediction models such as decision trees or neural networks.

In conclusion, we have developed a logistic regression and gradient boosted tree model to predict a diagnosis of heart failure using age, blood urea nitrogen, red blood cell distribution width; history of atrial fibrillation, hypertension, and chronic kidney disease, and medication history for furosemide, lisinopril, and carvedilol. Our models can effectively predict a heart failure diagnosis and can potentially be used for early diagnosis of heart failure patients.

Bibliography

1 Bui A L, Horwich T B, Fonarow G C. Epidemiology and risk profile of heart failure. *Nature Reviews Cardiology* 2011; **8**: 30–41.

2 Metra M, Teerlink J R. Heart failure. *The Lancet* 2017; **390**: 1981–95.

3 Yancy C W, Jessup M, Bozkurt B, *et al.* 2013 ACCF/AHA Guideline for the Management of Heart Failure. *Journal of the American College of Cardiology* 2013; **62**: e147–239.

4 Kalogeropoulos A, Psaty B M, Vasan R S, *et al.* Validation of the Health ABC Heart Failure Model for Incident Heart Failure Risk Prediction: The Cardiovascular Health Study. *Circulation: Heart Failure* 2010; **3**: 495–502.

5 Taslimitehrani V, Dong G, Pereira N L, Panahiazar M, Pathak J. Developing EHR-driven heart failure risk prediction models using CPXR(Log) with the probabilistic loss function. *Journal of Biomedical Informatics* 2016; **60**: 260–9.

6 Averbuch T, Sullivan K, Sauer A, *et al.* Applications of artificial intelligence and machine learning in heart failure. *European Heart Journal - Digital Health* 2022; **3**: 311–22.

7 Kwon J-m, Kim K-H, Jeon K-H, *et al.* Development and Validation of Deep-Learning Algorithm for Electrocardiography-Based Heart Failure Identification. *Korean Circulation Journal* 2019; **49**: 629.

8 Masetic Z, Subasi A. Congestive heart failure detection using random forest classifier. *Computer Methods and Programs in Biomedicine* 2016; **130**: 54–64.

9 Alonso-Betanzos A, Bolón-Canedo V, Heyndrickx G R, Kerkhof P L. Exploring Guidelines for Classification of Major Heart Failure Subtypes by Using Machine Learning. *Clinical Medicine Insights: Cardiology* 2015; : CMC.S18746.

10 Lundberg S M, Lee S-I. A Unified Approach to Interpreting Model Predictions. Advances in Neural Information Processing Systems 30. 2017; : 4765–74.

11 Lundberg S M, Erion G, Chen H, *et al.* From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence* 2020; **2**: 56–67.

12 Johnson A E W, Bulgarelli L, Shen L, *et al.* MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data* 2023; **10**: 1.

13 Johnson A E W, Bulgarelli L, Pollard T J, Celi L A, Mark R G. MIMIC-IV (version 2.2). *PhysioNet* 2023. https://doi.org/10.13026/6mm1-ek67

14 Goldberger A L, Amaral L A N, Glass L, *et al.* PhysioBank, PhysioToolkit, and PhysioNet: Components of a New Research Resource for Complex Physiologic Signals. *Circulation* 2000; **101**. DOI:10.1161/01.CIR.101.23.e215

15 About the Topic of Race. 2022; published online March. https://www.census.gov/topics/population/race/about.html

16 Austin P C, White I R, Lee D S, Van Buuren S. Missing Data in Clinical Research: A Tutorial on Multiple Imputation. *Canadian Journal of Cardiology* 2021; **37**: 1322–31.

17 Raasveldt M, Muehleisen H. DuckDB. https://github.com/duckdb/duckdb

18 Vink R, Gooijer S de, Beedie A, *et al.* pola-rs/polars: Python Polars. 2024; published online July. DOI:10.5281/ZENODO.13872436

19 Subirana I, Sanz H, Vila J. Building Bivariate Tables: The compareGroups Package for R. *Journal of Statistical Software* 2014; **57**: 1–16.

20 R Core Team. R: A Language and Environment for Statistical Computing. 2024. https://www.r-project.org/

21 miceforest: Fast, Memory Efficient Imputation with LightGBM. https://github.com/AnotherSamWilson/miceforest

22 Pedregosa F, Varoquaux G, Gramfort A, *et al.* Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 2011; **12**: 2825–30.

23 Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016: 785–94.

24 Ansel J, Yang E, He H, *et al.* PyTorch 2: Faster Machine Learning Through Dynamic Python Bytecode Transformation and Graph Compilation. In: 29th ACM International Conference on Architectural Support for Programming Languages and Operating Systems, Volume 2 (ASPLOS '24). ACM, 2024. DOI:10.1145/3620665.3640366

25 Yu S-N, Lee M-Y. Conditional mutual information-based feature selection for congestive heart failure recognition using heart rate variability. *Computer Methods and Programs in Biomedicine* 2012; **108**: 299–309.

26 VanderWeele T J. Principles of confounder selection. *European Journal of Epidemiology* 2019; **34**: 211–9.

27 Pearce N, Richiardi L. Commentary: Three worlds collide: Berkson's bias, selection bias and collider bias. *International Journal of Epidemiology* 2014; **43**: 521–4.

28 Molnar C. Interpretable machine learning: a guide for making black box models explainable, Second edition. Munich, Germany: Christoph Molnar, 2022

29 Moons K G, Altman D G, Reitsma J B, *et al.* Transparent Reporting of a multivariable prediction model for Individual Prognosis Or Diagnosis (TRIPOD): Explanation and Elaboration. *Annals of Internal Medicine* 2015; **162**: W1–73.

30 Donders A R T, Van Der Heijden G J, Stijnen T, Moons K G. Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology* 2006; **59**: 1087–91.