Journal of Current Science and Technology, May-August 2022 Vol .12 No .2, pp. 336-348

ISSN 2630-0656 (Online)

Cite this article: Panyamit, T., Sukvivatn, P., Chanma, P., Kim, Y., Premratanachai, P., & Pechprasarn, S. (2022, May). Identification of factors in the survival rate of heart failure patients using machine learning models and principal component analysis. *Journal of Current Science and Technology, 12*(2), 336-348. DOI:

Journal of Current Science and Technology

Journal homepage: https://jcst.rsu.ac.th

Identification of factors in the survival rate of heart failure patients using machine learning models and principal component analysis

Thanawan Panyamit[1], Phattharamon Sukvivatn[1], Paphada Chanma[1], Yejin Kim[1+], Premmika Premratanachai[1], and Suejit Pechprasarn[2]*

[1]Satriwitthaya School, Wat Bowon Niwet, Phra Nakhon, Bangkok 10200, Thailand
[2]College of Biomedical Engineering, Rangsit University, Patumthani 12000, Thailand
The authors have contributed equally.

* Corresponding author; E-mail: suejit.p@rsu.ac.th

Received 10 March 2022; Revised 3 May 2022; Accepted 5 May 2022
Published online 25 August 2022

Abstract

Heart failure (HF) and congestive heart failure (CHF) have recently been classified as a growing and widespread epidemic worldwide that significantly impacts morbidity and mortality, especially in the aged groups. This study used a publicly available clinical dataset on 299 HF patients with 12 variables potentially contributing to their mortality: age, anemia, creatinine phosphokinase, diabetes, ejection fraction, high blood pressure, platelets, serum creatinine and sodium levels, sex, smoking, and follow-up time. Several studies previously used this dataset to identify critical factors influencing patient mortality. Here, we curate the data to ensure it is unbiased, then apply principal component analysis and machine learning models to identify factors influencing crucial variables contributing to patient mortality. We investigate and compare the classification accuracy of different machine learning models, including the tree, linear discriminant, quadratic discriminant, logistic, naïve Bayes, support vector machine, nearest-neighbor ensemble, and kernel models. We found the ensemble bagged tree model to have the highest cross-validation classification accuracy of 96.4% and require only three variables: platelets, creatinine phosphokinase, and follow-up period.

Keywords**:** *heart failure classification; machine learning; mortality rate prediction; smart healthcare.*

1. Introduction

Heart failure (HF), and congestive heart failure (CHF) in particular, is a life-threatening disease that affects more than 64.3 million people worldwide (Groenewegen, Rutten, Mosterd, & Hoes, 2020). It is an alarming healthcare problem in numerous countries, including the United Kingdom (Taylor et al., 2019), the United States (Osenenko, Kuti, Deighton, Pimple, & Szabo, 2022), Thailand (Tankumpuan et al., 2019), Canada, Germany, India, and Northern China (Groenewegen et al., 2020). Moreover, it occurs in older people and healthy adolescents (Seferović et al., 2021). This ailment is caused by the inability of the heart muscle to efficiently pump blood and carry sufficient oxygen around the body, leading to reduced blood

.

flow and shortness of breath (van Riet et al., 2014). In addition, it may develop from coronary artery disease, high blood pressure, heart attack, high cholesterol, heart valve disease, or diabetes (Lehrke & Marx, 2017). This condition typically arises in older people since the main artery will become thicker, stiffer, and less flexible as we mature. In addition, its significant risk factors include unhealthy behaviors such as smoking, eating high-fat food, excessive alcohol intake, physical inactivity, and being overweight, as well as heredity, congenital disease, and sex (Prochaska et al., 2004).

Generally, the survival rate for patients diagnosed with heart failure is marginally low. About three-fourths of patients survive for one year, 45.5% for five years, 24.5% for ten years, and 12.7% for 15 years. Furthermore, women have worse short-term and long-term survival rates than men (Taylor et al., 2019). Sex, advanced age, history of diabetes, smoking, and chronic kidney disease are critical factors that can increase the mortality rate (Henkel, Redfield, Weston, Gerber, & Roger, 2008). Furthermore, the medical procedure for diagnosing HF includes an examination of the sounds of the heartbeat and lung congestion to diagnose any abnormal cardiac rhythm and performing physical tests such as an electrocardiogram (ECG), chest X-ray, magnetic resonance imaging (MRI), computerized tomography (CT) cardiac scan, or coronary angiogram (Fonseca, 2006).

The open-source dataset from the UC Irvine Machine Learning Repository has a database containing medical record information for 299 patients with HF, with 13 clinical features collected during their follow-up sessions. The patients consisted of 105 women and 194 men between 40 and 95 years old with left ventricular systolic dysfunction classified as class III and IV under New York Heart Association (NYHA) criteria and who have experienced HF. Chicco and Jurman (2020) reported the dataset and developed a machine learning (ML) model based on 13 clinical variables that can contribute to death: age, anemia, high blood pressure (HBP), creatinine phosphokinase (CPK), diabetes, ejection fraction (EF), platelets, sex, serum creatinine level, serum sodium level, smoking, follow-up period, and death event. In addition to providing the graphical forecast of survival probability and a nomogram, their study had the potential to alter clinical practice by providing clinicians with a new tool for predicting HF patient survival. Indeed, when determining whether a patient will survive HF, clinicians rely heavily on serum creatinine levels and EF.

Groenewegen et al. (2020) have shown HF to be a diverse condition, complicating case identification and patient categorization in epidemiological research. Left ventricular ejection fraction (LVEF) is a clinically valuable trait indicative of underlying pathophysiological processes and medication sensitivity. Currently, HF patients are classified as having HF with decreased (HFrEF; LVEF <40%), mid-range (HFmrEF; LVEF 40%–49%), or maintained EF (HFpEF; LVEF ≥50%). The thresholds are arbitrary and vary across standards, and LVEF classification has been criticized for simplifying a complicated illness.

Based on clinical trial data, Lehrke and Marx (2017) have shown that people with diabetes and HF may benefit most from treatment with sodium-glucose transport protein 2 (SGLT2) inhibitors that lower glucose levels. This effect might be related to glucose clearance via the kidney, with a net reduction in energy substrate availability due to SGLT2 inhibition and other plausible pathways. Lifestyle management also reduces energy substrate availability, which improves cardiac function in obese individuals with and without diabetes. Furthermore, there is evidence that metformin, which lowers energy substrate availability by lowering endogenous glucose synthesis, may benefit diabetic patients with HF. In contrast, no benefit in HF or potential adverse consequences has been observed with glucose-lowering techniques that enhance the availability of insulin either directly or indirectly.

2. Objectives

1. This research aims to use statistical models to identify crucial clinical variables that clinicians should pay particular attention to and evaluate their classification accuracy

for HF mortality to determine factors affecting prediction accuracy in HFpatients.
2. Use ML to predict the survival rates affected by each factor compared to those reported in the literature.
3. Apply principal component analysis (PCA) and ML models (Onan, 2020; 2022), comparing the HF classification and predictive accuracy of different statistical models (Onan, 2019; Onan & Korukoğlu, 2017): tree, linear discriminant, quadratic discriminant, logistic regression, naïve Bayes, support vector machine (SVM), and ensemble (Onan, Korukoğlu, & Bulut, 2016; Onan, Korukoğlu, & Bulut, 2017).

3. Materials and methods

3.1 Dataset and data curation

The HF dataset analyzed here was obtained from the open-source University of California Irvine (UCI) Machine Learning Repository (Chicco & Jurman, 2020; assessed January 14, 2022). It contains data for 299 patients (105 women and 194 men) between 40 and 95 years old who have a left ventricular systolic dysfunction classified as class III or IV based on the New York Heart Association (NYHA) criteria based on previous HF that include the following 13 features:

1. The death event was defined as the number of patients who died during the follow-up period, coded 0 if the patient is alive and 1 if they are deceased. Of the 299 patients in the dataset, 96 were dead, and 203 were alive. This dataset is biased toward patients who have survived. Since the main objective of this study is to identify factors affecting patient mortality rates, this parameter is the label used for supervised learning. The information for 107 random surviving patients was removed during the data curation process to create an unbiased dataset, leaving 192 entries, 96 coded 1 and 96 coded 0.
2. The age (in years) of these 192 patients was between 40 and 95, with an average of 61.7483 years and a standard deviation (SD) of 12.6811. The overall age distribution had a slight positive skewness.
3. Anemia was defined as decreasing red blood cells (hemoglobin), coded 0 if false and 1 if true. Of the 192 patients in the dataset, 78 had anemia, and 114 did not. The overall anemia distribution had a slight positive skewness.
4. CPK was defined as the blood level of the CPK enzyme (mcg/L). Across the 192 patients in the dataset, CPK was between 23 and 7861 mcg/L, with an average of 606.5938 mcg/L and an SD of 1030.6 mcg/L. The overall CPK distribution had a slight positive skewness.
5. Diabetes was defined as the patient's diabetes status, coded 0 if false and 1 if true. Of the 192 patients in the dataset, 83 had diabetes, and 109 did not. The overall diabetes distribution had a slight positive skewness.
6. EF was defined as the percentage of blood leaving the heart with each contraction. Of the 192 patients in the dataset, EF was between 14% and 70%, with an average of 36% and an SD of 11.2724%. The overall EF distribution had a slight negative skewness.
7. HBP was defined as the hypertension status of the patient, coded 0 if false and 1 if true. O the 192 patients in the dataset, 65 had HBP, and 127 did not. The overall HBP distribution had a slight positive skewness.
8. The platelet value was defined as the number of platelets in the blood (kilo-platelets/mL). Of the 192 patients in the dataset, the platelet value was between 25,100 and 742,000 kilo-platelets/mL, with an average of 263,700 kilo-platelets/mL and an SD of 102,820 kilo-platelets/mL. The overall platelet value distribution had a slight positive skewness.
9. Serum creatinine levels in the blood (mg/dL) were used. Of the 192 patients in the dataset, serum creatinine levels were between 0.5000 and 9.4000 mg/dL, with an average of 1.5141 mg/dL and an SD of 1.1735 mg/dL. The overall distribution of

.

serum creatinine levels had a slight positive skewness.

10. Serum sodium levels in the blood (mEq/L) were used. Of the 192 patients in the dataset, their serum sodium levels were between 113 and 148 mEq/L, with an average of 136.1198 mEq/L and an SD of 4.8139 mEq/L. The overall distribution of the serum sodium levels had a slight negative skewness.
11. Sex was coded as 0 if female and 1 if male. Of the 192 patients in the dataset, 69 were male, and 123 were female. The overall sex distribution had a slight negative skewness.
12. Smoking was defined as the patient's smoking status, coded 0 if false, and 1 if true. Of the 192 patients in the dataset, 59 were smokers, and 133 were nonsmokers. The overall smoker distribution had a slight positive skewness.
13. The follow-up time was recorded in days. Of the 192 patients in the dataset, follow-up time was between 4 and 285 days, with an average of 146.3750 days and an SD of 89.5326 days. The overall distribution of follow-up time had a slight positive skewness.

3.2 Training and applying MLmodels

The 13 variables were first treated as predictors and labels for supervised ML training and classification tasks. Death status was used as the label. The 12 predictor variables are summarized in Table 1.

Table 1 Classification model predictors and labels

Variables	Type	Variables	Type
Age	Predictor	HBP	Predictor
Anemia	Predictor	Platelets	Predictor
CPK	Predictor	Serum creatinine	Predictor
Diabetes	Predictor	Serum sodium	Predictor
EF	Predictor	Sex	Predictor
Smoking	Predictor	Death event	Label
Follow-up time	Predictor		

Table 2 The ML models evaluated

Model	Details	Model	Details
Tree	Fine tree	K-nearest neighbor (KNN)	Fine KNN
	Medium tree		Medium KNN
	Coarse tree		Coarse KNN
Linear Discriminant	Linear Discriminant		Cosine KNN
Quadratic Discriminant	Quadratic Discriminant		Cubic KNN
Logistic Regression	Logistic Regression		Weighted KNN
Naïve Bayes	Gaussian Naïve Bayes	Ensemble	Boosted Trees
	Kernel Naïve Bayes		Bagged Trees
SVM	Linear SVM		Subspace Discriminant
	Quadratic SVM		Subspace KNN
	Cubic SVM		RUSBoosted Trees
	Fine Gaussian SVM	Neural network (Onan, 2021)	Narrow Neural Network
	Medium Gaussian SVM		Medium Neural Network
	Coarse Gaussian SVM		Wide Neural Network
Kernel	SVM Kernel		Bilayered Neural Network
	Logistic Regression Kernel		Trilayered Neural Network

We used a built-in ML application in MATLAB 2021b to train the supervised ML models listed in Table 2. Classification accuracy was evaluated using five-fold cross-validation.

3.3 Principal components analysis

We used a PCA model with a statistical confidence of 95% to identify predictors contributing to the model classification accuracy. While several improved PCA methods have been reported (Biagetti, Crippa, Falaschetti, Orcioni, & Turchetti, 2016; Biagetti, Crippa, Falaschetti, Luzzi, & Turchetti, 2021), the conventional PCA method was used here.

4. Results

4.1 ML classification accuracy based on all 12 predictors

The ML models listed in Table 2 were first trained using all 12 predictors and the death event as the label (Table 1). The classification accuracy of each model is shown in Table 3. The top three ML models in terms of accuracy were the ensemble and *k*-nearest neighbor (KNN) bagged tree models with a classification accuracy of 94.8 % and the coarse tree model with 94.3%. The fine tree, medium tree, and logistic regression models ranked in the top third for classification accuracy, with a cross-validation accuracy of 93.2%. However, the boosted tree, logistic regression kernel, and SVM kernel models had notably low accuracy and were unlikely to be suitable for use as a proper model.

Table 3 Classification accuracy of ML models trained with all 12 predictors

Model	Details	Classification accuracy
Tree	Fine tree	93.2 %
	Medium tree	93.2 %
	Coarse tree	94.3 %
Linear discriminant	Linear discriminant	91.1 %
Quadratic discriminant	Quadratic discriminant	87.5 %
Logistic regression	Logistic regression	93.2 %
Naïve Bayes	Gaussian naïve Bayes	90.6 %
	Kernel naïve Bayes	91.7 %
SVM	Linear SVM	90.6 %
	Quadratic SVM	89.1 %
	Cubic SVM	89.1 %
	Fine gaussian SVM	66.7 %
	Medium gaussian SVM	91.1 %
	Coarse gaussian SVM	90.1 %
KNN	Fine KNN	81.8 %
	Medium KNN	82.8%
	Coarse KNN	79.7 %
	Cosine KNN	84.9 %
	Cubic KNN	81.2 %
	Weighted KNN	83.9 %
Ensemble	Boosted trees	49.5 %
(Onan, 2018)	Bagged trees	94.8 %
	Subspace discriminant	91.7 %
	Subspace KNN	56.2 %
	RUSBoosted trees	59.4 %
Neural network	Narrow neural network	90.6 %
	Medium neural network	92.7 %
	Wide neural network	91.7 %
	Bilayered neural network	89.1 %
	Trilayered neural network	91.7 %
Kernel	SVM kernel	38.0 %
	Logistic regression kernel	47.4 %

4.2 PCA analysis

4.2.1 PCA using one variable

ML models with the highest and lowest accuracy were explored with PCA using the one

.

predictor listed in Table 4. The PCA coefficients of the five ML models with the highest classification accuracy are shown in Table 5. The quadratic discriminant and Gaussian naïve Bayes models showed the highest accuracy, 52.6%, followed by the coarse KNN model with 51.0%. However, the three worst-performing ML models (ranked from best to worst) were the medium neural network model (40.1% accuracy), fine gaussian SVM model (38.5%), and wide neural network model (36.5%).

Table 4 Summary of the classification accuracy of ML models using one predictor identified via PCA with a 95% confidence level.

Variable	Model	Accuracy	Variable	Model	Accuracy
Platelets (100%)	Quadratic discriminant	52.6%	Platelets (100%)	Ensemble (subspace KNN)	47.9%
Platelets (100%)	Gaussian naïve Bayes	52.6%	Platelets (100%)	Ensemble (bagged trees)	47.4%
Platelets (100%)	Coarse KNN	51.0%	Platelets (100%)	Ensemble (boosted trees)	46.4%
Platelets (100%)	Bilayered neural network	50.5%	Platelets (100%)	Medium Gaussian SVM	45.8%
Platelets (100%)	Linear SVM	50.0%	Platelets (100%)	Narrow neural networks	45.8%
Platelets (100%)	Cosine KNN	50.0%	Platelets (100%)	Medium KNN	44.8%
Platelets (100%)	Coarse Gaussian SVM	49.5%	Platelets (100%)	Cubic KNN	44.8%
Platelets (100%)	Ensemble (RUSBoosted Trees)	49.5%	Platelets (100%)	Medium tree	44.3%
Platelets (100%)	SVM kernel	49.5%	Platelets (100%)	Weighted KNN	44.3%
Platelets (100%)	Logistic regression kernel	49.5%	Platelets (100%)	Coarse tree	43.8%
Platelets (100%)	Quadratic SVM	49.0%	Platelets (100%)	Trilayered neural network	42.7%
Platelets (100%)	Cubic SVM	49.0%	Platelets (100%)	Fine Ttee	42.2%
Platelets (100%)	Logistic regression	48.4%	Platelets (100%)	Medium neural network	40.1%
Platelets (100%)	Linear discriminant	48.4%	Platelets (100%)	Fine gaussian SVM	38.5%
Platelets (100%)	Kernel naïve Bayes	47.9%	Platelets (100%)	Wide neural network	36.5%
Platelets (100%)	Fine KNN	47.9%			

Table 5 Summary of the PCA coefficients of the five ML models with the highest classification accuracy using one predictor identified via PCA and a 95% confidence level.

One Variable	PCA coefficients				
Variable names	Quadratic discriminant (52.6%)	Gaussian naïve Bayes (52.6%)	Coarse KNN (51.0%)	Bilayered neural network (50.5%)	Linear SVM (50.0%)
Age	0.0000	0.0000	0.0000	0.0000	0.0000
Anemia	0.0000	0.0000	0.0000	0.0000	0.0000
CPK	0.0005	0.0005	0.0005	0.0005	0.0005
Diabetes	0.0000	0.0000	0.0000	0.0000	0.0000
EF	0.0000	0.0000	0.0000	0.0000	0.0000
HBP	0.0000	0.0000	0.0000	0.0000	0.0000
Platelets	1.0000	1.0000	1.0000	1.0000	1.0000
Serum creatinine	0.0000	0.0000	0.0000	0.0000	0.0000
Serum sodium	0.0000	0.0000	0.0000	0.0000	0.0000
Sex	0.0000	0.0000	0.0000	0.0000	0.0000

One Variable Variable names	PCA coefficients				
	Quadratic discriminant (52.6%)	Gaussian naïve Bayes (52.6%)	Coarse KNN (51.0%)	Bilayered neural network (50.5%)	Linear SVM (50.0%)
Smoking	0.0000	0.0000	0.0000	0.0000	0.0000
Follow-up time	0.0000	0.0000	0.0000	0.0000	0.0000

Table 6 Summary of the classification accuracy of ML models using two predictors identified via PCA with a 95% confidence level.

Variable	Model	Accuracy	Variable	Model	Accuracy
Platelets (100%), CPK (100%)	Wide neural network	57.3%	Platelets (100%), CPK (100%)	Quadratic discriminant	50.0%
Platelets (100%), CPK (100%)	Quadratic SVM	54.7%	Platelets (100%), CPK (100%)	Gaussian naïve Bayes	50.0%
Platelets (100%), CPK (100%)	Weighted KNN	54.7%	Platelets (100%), CPK (100%)	Fine KNN	50.0%
Platelets (100%), CPK (100%)	Bilayered neural network	54.7%	Platelets (100%), CPK (100%)	SVM kernel	50.0%
Platelets (100%), CPK (100%)	Medium neural network	54.2%	Platelets (100%), CPK (100%)	Coarse tree	49.5%
Platelets (100%), CPK (100%)	Linear discriminant	53.1%	Platelets (100%), CPK (100%)	Kernel naïve Bayes	49.5%
Platelets (100%), CPK (100%)	Logistic regression	53.1%	Platelets (100%), CPK (100%)	Linear SVM	49.5%
Platelets (100%), CPK (100%)	Medium Gaussian SVM	53.1%	Platelets (100%), CPK (100%)	Cubic SVM	49.5%
Platelets (100%), CPK (100%)	Coarse Gaussian SVM	52.6%	Platelets (100%), CPK (100%)	Ensemble (bagged trees)	49.5%
Platelets (100%), CPK (100%)	Ensemble (subspace discriminant)	52.6%	Platelets (100%), CPK (100%)	Ensemble (boosted trees)	48.4%
Platelets (100%), CPK (100%)	Ensemble (RUSBoosted Trees)	52.6%	Platelets (100%), CPK (100%)	Cosine KNN	47.4%
Platelets (100%), CPK (100%)	Fine tree	52.1%	Platelets (100%), CPK (100%)	Logistic regression kernel	47.4%
Platelets (100%), CPK (100%)	Trilayered neural network	51.0%	Platelets (100%), CPK (100%)	Narrow neural network	46.9%
Platelets (100%), CPK (100%)	Coarse KNN	50.5%	Platelets (100%), CPK (100%)	Narrow neural network	45.3%
Platelets (100%), CPK (100%)	Ensemble (subspace KNN)	50.5%	Platelets (100%), CPK (100%)	Medium KNN	43.2%
Platelets (100%), CPK (100%)	Medium tree	50.0%	Platelets (100%), CPK (100%)	Fine Gaussian SVM	42.2%

The PCA indicated that platelet number was the most significant parameter. However, this single variable was not sufficient to develop an accurate model.

4.2.2 PCA with two variables

The ML models developed and analyzed using two predictors in the PCA are shown in Table 6, and the PCA coefficients of the five ML models with the highest classification accuracy are shown in Table 7. The wide neural network model had the highest k-fold accuracy of 57.3%, followed by the quadratic SVM model, the weighted KNN model, and the bilayered neural network model, each with an accuracy of 54.7%. In contrast, the three worst-performing ML models (ranks from best to worst) were the narrow neural network model (46.9% accuracy), the cubic KNN model (45.3%), and the medium KNN model (43.2%).

.

Chicco and Jurman (2020) reported that serum creatinine and EF were the two most important variables for predicting HF patient survival. However, according to our PCA, the two most important varaibles were platelets and CPK. However, using only these two variables was insufficient to provide a good prediction accuracy.

Table 7 Summary of the PCA coefficients of the five ML models with the highest classification accuracy using two predictors identified via PCA with a 95% confidence level.

Two Variables	**PCA coefficients**									
Variable	**Wide neural network (57.3%)**		**Quadratic SVM (54.7%)**		**Weighted KNN (54.7%)**		**Bilayered neural network (54.7%)**		**Medium neural network (54.2%)**	
Age	0.0000	-0.0013	0.0000	-0.0013	0.0000	-0.0013	0.0000	-0.0013	0.0000	-0.0013
Anemia	0.0000	-0.0001	0.0000	-0.0001	s0.000 0	-0.0001	0.0000	-0.0001	0.0000	-0.0001
CPK	0.0005	1.0000	0.0005	1.0000	0.0005	1.0000	0.0005	1.0000	0.0005	1.0000
Diabetes	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
EF	0.0000	0.0001	0.0000	0.0001	0.0000	0.0001	0.0000	0.0001	0.0000	0.0001
HBP	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Platelets	1.0000	-0.0005	1.0000	-0.0005	1.0000	-0.0005	1.0000	-0.0005	1.0000	-0.0005
Serum creatinine	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Serum sodium	0.0000	0.0005	0.0000	0.0005	0.0000	0.0005	0.0000	0.0005	0.0000	0.0005
Sex	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Smoking	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
Follow-up time	0.0000	-0.0030	0.0000	-0.0030	0.0000	-0.0030	0.0000	-0.0030	0.0000	-0.0030

4.2.3 PCA with thee variables

The accuracy of ML models developed using three predictors in the PCA is shown in Table 8 (ranked from highest to lowest accuracy), and the PCA coefficients of the five ML models with the highest classification accuracy are shown in Table 9. The ensemble (bagged trees) model had the highest accuracy (96.4%), followed by the coarse tree model (95.3%), the fine tree model (94.8%), and the medium tree model (94.8%). However, the worst-performing ML models (ranks from best to worst) were the ensemble (boosted trees) and ensemble (RUSBoosted trees) models that had the same accuracy (49.5%), the logistic regression kernel model (47.4%), and the SVM Kernel model (45.3%).

Interestingly, we have achieved the same level of accuracy using three predictors as was possible with all 12 predictors. Therefore, a model based on three predictors is sufficient to identify the critical factors contributing to HF patient mortality.

Table 8 Summary of classification accuracies for ML models trained using three predictors identified via PCA with a 95% confidence level.

Variable	**Model**	**Accuracy**	**Variable**	**Model**	**Accuracy**
Platelets (100%), CPK (100%), follow-up time (99.87%)	SVM kernel	45.3%	Platelets (100%), CPK (100%), follow-up time (99.87%)	Coarse KNN	90.6%
Platelets (100%), CPK (100%), follow-up time (99.87%)	Logistic regression kernel	47.4%	Platelets (100%), CPK (100%), follow-up time (99.87%)	Medium neural Network	90.6%

Variable	Model	Accuracy	Variable	Model	Accuracy
Platelets (100%), CPK (100%), follow-up time (99.87%)	Ensemble (boosted trees)	49.5%	Platelets (100%), CPK (100%), follow-up time (99.87%)	Kernel naïve Bayes	91.1%
Platelets (100%), CPK (100%), follow-up time (99.87%)	Ensemble (RUSBoosted trees)	49.5%	Platelets (100%), CPK (100%), follow-up time (99.87%)	Weighted KNN	91.1%
Platelets (100%), CPK (100%), follow-up time (99.87%)	Ensemble (subspace KNN)	58.9%	Platelets (100%), CPK (100%), follow-up time (99.87%)	Linear discriminant	91.7%
Platelets (100%), CPK (100%), follow-up time (99.87%)	Narrow neural network	86.5%	Platelets (100%), CPK (100%), follow-up time (99.87%)	Gaussian naïve Bayes	91.7%
Platelets (100%), CPK (100%), follow-up time (99.87%)	Fine Gaussian SVM	88.0%	Platelets (100%), CPK (100%), follow-up time (99.87%)	Quadratic SVM	91.7%
Platelets (100%), CPK (100%), follow-up time (99.87%)	Fine KNN	88.5%	Platelets (100%), CPK (100%), follow-up time (99.87%)	Coarse Gaussian SVM	91.7%
Platelets (100%), CPK (100%), follow-up time (99.87%)	Cosine KNN	89.1%	Platelets (100%), CPK (100%), follow-up time (99.87%)	Ensemble (subspace discriminant)	91.7%
Platelets (100%), CPK (100%), follow-up time (99.87%)	Trilayered neural network	89.6%	Platelets (100%), CPK (100%), follow-up time (99.87%)	Wide neural network	91.7%
Platelets (100%), CPK (100%), follow-up time (99.87%)	Cubic KNN	90.1%	Platelets (100%), CPK (100%), follow-up time (99.87%)	Linear SVM	92.7%
Platelets (100%), CPK (100%), follow-up time (99.87%)	Bilayered neural network	90.1%	Platelets (100%), CPK (100%), follow-up time (99.87%)	Logistic regression	93.2%
Platelets (100%), CPK (100%), follow-up time (99.87%)	Quadratic discriminant	90.6%	Platelets (100%), CPK (100%), follow-up time (99.87%)	Fine tree	94.8%
Platelets (100%), CPK (100%), follow-up time (99.87%)	Cubic SVM	90.6%	Platelets (100%), CPK (100%), follow-up time (99.87%)	Medium tree	94.8%
Platelets (100%), CPK (100%), follow-up time (99.87%)	Medium Gaussian SVM	90.6%	Platelets (100%), CPK (100%), follow-up time (99.87%)	Coarse tree	95.3%
Platelets (100%), CPK (100%), follow-up time (99.87%)	Medium KNN	90.6%	Platelets (100%), CPK (100%), follow-up time (99.87%)	Ensemble (bagged trees)	96.4%

.

Table 9 Summary of the PCA coefficients of the five ML models with the highest classification accuracy using three predictors identified via PCA with a 95% confidence level.

Three variable	Variable											
PCA Coefficients	**Age**	**Anemia**	**CPK**	**Diabetes**	**EF**	**HBP**	**Platelets**	**Serum creatinine**	**Serum sodium**	**Sex**	**Smoking**	**Time**
Ensemble (96.5%)	0.0000	0.0000	0.0005	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	-0.0013	-0.0001	1.0000	0.0000	0.0001	-0.0000	-0.0005	0.0000	0.0005	-0.0000	0.0000	0.0030
	-0.0470	-0.0010	0.0029	0.0002	0.0188	-0.0013	0.0000	-0.0029	0.0089	-0.0001	-0.0002	0.9987
Coarse tree (95.3%)	0.0000	0.0000	0.0005	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	-0.0013	-0.0001	1.0000	0.0000	0.0001	-0.0000	-0.0005	0.0000	0.0005	-0.0000	0.0000	0.0030
	-0.0470	-0.0010	0.0029	0.0002	0.0188	-0.0013	0.0000	-0.0029	0.0089	-0.0001	-0.0002	0.9987
Fine tree (94.8%)	0.0000	0.0000	0.0005	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	-0.0013	-0.0001	1.0000	0.0000	0.0001	-0.0000	-0.0005	0.0000	0.0005	-0.0000	0.0000	0.0030
	-0.0470	-0.0010	0.0029	0.0002	0.0188	-0.0013	0.0000	-0.0029	0.0089	-0.0001	-0.0002	0.9987
Medium tree (94.8%)	0.0000	0.0000	0.0005	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	-0.0013	-0.0001	1.0000	0.0000	0.0001	-0.0000	-0.0005	0.0000	0.0005	-0.0000	0.0000	0.0030
	-0.0470	-0.0010	0.0029	0.0002	0.0188	-0.0013	0.0000	-0.0029	0.0089	-0.0001	-0.0002	0.9987
Logistic regression (93.2%)	0.0000	0.0000	0.0005	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	-0.0013	-0.0001	1.0000	0.0000	0.0001	-0.0000	-0.0005	0.0000	0.0005	-0.0000	0.0000	0.0030
	-0.0470	-0.0010	0.0029	0.0002	0.0188	-0.0013	0.0000	-0.0029	0.0089	-0.0001	-0.0002	0.9987
Linear SVM (92.7%)	0.0000	0.0000	0.0005	0.0000	0.0000	0.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000
	-0.0013	-0.0001	1.0000	0.0000	0.0001	-0.0000	-0.0005	0.0000	0.0005	-0.0000	0.0000	0.0030
	-0.0470	-0.0010	0.0029	0.0002	0.0188	-0.0013	0.0000	-0.0029	0.0089	-0.0001	-0.0002	0.9987

4.2.4 PCA with more than three variables

We found that three predictor variables were sufficient to build an accurate ML prediction model for HF patient survival. We computed the PCA responses for 4 to 12 predictor variables. For each number, the ML model with the highest accuracy is reported in Table 10. The accuracy increased gradually from one to two predictor variables, then markedly increased with three predictor variables (Figure 1). However, accuracy remained stable after three predictor variables. Therefore, three predictor variables are sufficient to develop an ML model to predict the mortality rate of HF patients.

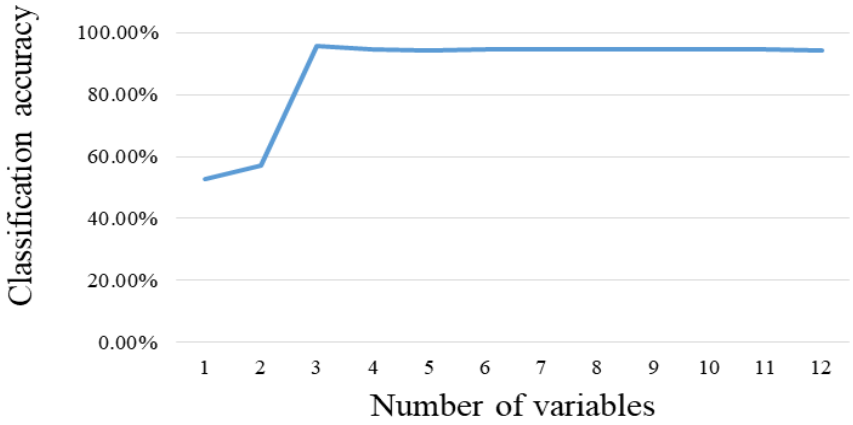

Figure 1 The relationship between the number of predictor variables and classification accuracy

Table 10. Summary of the classification accuracy of ML models trained using 1 to 12 predictors selected via PCA with a 95% confidence level.

Number of variables	Variables	Model	Accuracy
1	Platelets (100%)	Quadratic discriminant, Gaussian naïve Bayes	52.6%
2	Platelets (100%), CPK (100%)	Wide neural network	57.3%
3	Platelets (100%), CPK (100%), follow-up time (99.87%)	Ensemble (bagged trees)	96.4%
4	Platelets (100%), CPK (100%), follow-up time (99.87%), age (84.56%)	Coarse tree	94.8%
5	Platelets (100%), CPK (100%), follow-up time (99.87%), age (84.56%), EF (84.40%)	Fine tree, medium tree, coarse tree, bagged trees	94.3%
6	Platelets (100%), CPK (100%), follow-up time (99.87%), age (84.56%), EF (84.40%), serum sodium (99.51%)	Ensemble (bagged trees)	94.8%
7	Platelets (100%), CPK (100%), follow-up time (99.87%), age (84.56%), EF (84.40%), serum sodium (99.51%), serum creatinine (99.74%)	Ensemble (bagged trees)	94.8%
8	Platelets (100%), CPK (100%), follow-up time (99.87%), age (84.56%), EF (84.40%), serum sodium (99.51%), serum creatinine (99.74%), sex (66.64%)	Ensemble (bagged trees)	94.8%
9	Platelets (100%), CPK (100%), follow-up time (99.87%), age (84.56%), EF (84.40%), serum sodium (99.51%), serum creatinine (99.74%), sex (66.64%), anemia (81.54%)	Ensemble (bagged trees)	94.8%
10	Platelets (100%), CPK (100%), follow-up time (99.87%), age (84.56%), EF (84.40%), serum sodium (99.51%), serum creatinine (99.74%), sex (66.64%), anemia (81.54%), diabetes (75.45%)	Ensemble (bagged trees)	94.8%
11	Platelets (100%), CPK (100%), follow-up time (99.87%), age (84.56%), EF (84.40%), serum sodium (99.51%), serum creatinine (99.74%), sex (66.64%), anemia (81.54%), diabetes (75.45%), HBP (69.18%)	Ensemble (bagged trees)	94.8%
12	Platelets (100%), CPK (100%), follow-up time (99.87%), age (84.56%), EF (84.40%), serum sodium (99.51%), serum creatinine (99.74%), sex (66.64%), anemia (81.54%), diabetes (75.45%), HBP (69.18%), smoking (-67.33%)	Ensemble (bagged trees)	94.3%

.

5. Discussion

Based on an open-source dataset consisting of 13 variables, we have identified three predictor variables that contribute most to mortality rates in HF patients using the PCA. These variables are platelets, CPK, and follow-up time. We found no significant improvement in classification accuracy when more than three predictor variables were used in ML model training. Our findings will enable clinicians to pay particular attention to these three critical predictors when assessing HF patient survival. Moreover, the PCA-based framework we have used here is not limited to the HF dataset; it is equally applicable to other datasets that contain potentially predictive variables. Our approach will facilitate the elimination of uninformative variables, reducing the time and cost required to collect data from patients. The future application of more sophisticated algorithms and models can further enhance classification accuracy (Gianfelici, Turchetti, & Crippa, 2009).

6. Conclusion

We have identified critical factors contributing to mortality in HF patients using a dataset from the UCI Machine Learning Repository containing 13 mortality-associated variables and assessed the accuracy of ML models in predicting HF patient survival. We first curated the dataset, removing data for 107 randomly selected patients to minimize potential survivor bias and provide an unbiased training environment. Then, we developed and applied PCA and supervised ML models to predict HF patient mortality in MATLAB 2021, comparing them by their cross-validation classification accuracy. PCA models were developed using 1 to 12 variables, excluding the label. We found the critical predictive variables to be platelets, CPK, and follow-up time, which clinicians should pay special attention to when assessing HF patient survival.

7. Acknowledgements

We acknowledge the HF data obtained from the open-source UCI Machine Learning Repository; this study would not have been possible without access to it. We also acknowledge the Research Institute and the College of Biomedical Engineering, Rangsit University, and Satriwitthaya School for supporting and funding the research.

8. References

Biagetti, G., Crippa, P., Falaschetti, L., Orcioni, S., & Turchetti, C. (2016). Multivariate direction scoring for dimensionality reduction in classification problems. In *International Conference on Intelligent Decision Technologies* (pp. 413-423). Springer, Cham.

Biagetti, G., Crippa, P., Falaschetti, L., Luzzi, S., & Turchetti, C. (2021). Classification of Alzheimer's Disease from EEG Signal Using Robust-PCA Feature Extraction. *Procedia Computer Science*, *192*, 3114-3122.

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), 1-16.

Fonseca, C. (2006). Diagnosis of heart failure in primary care. *Heart failure reviews*, *11*(2), 95-107.

Groenewegen, A., Rutten, F. H., Mosterd, A., & Hoes, A. W. (2020). Epidemiology of heart failure.*European journal of heart failure*, *22*(8), 1342-1356.

Gianfelici, F., Turchetti, C., & Crippa, P. (2009). A non-probabilistic recognizer of stochastic signals based on KLT. *Signal Processing*, *89*(4), 422-437.

Henkel, D. M., Redfield, M. M., Weston, S. A., Gerber, Y., & Roger, V. L. (2008). Death in heart failure: a community perspective. *Circulation: Heart Failure*, *1*(2), 91-97.

Ishaq, A., Sadiq, S., Umer, M., Ullah, S., Mirjalili, S., Rupapara, V., & Nappi, M. (2021). Improving the prediction of heart failure patients' survival using SMOTE and

effective data mining techniques. *IEEE access*, *9*, 39707-39716.

Lehrke, M., & Marx, N. (2017). Diabetes mellitus and heart failure. *The American journal of cardiology*, *120*(1), S37-S47.

Onan, A., Korukoğlu, S., & Bulut, H. (2016). Ensemble of keyword extraction methods and classifiers in text classification. *Expert Systems with Applications*, *57*, 232-247.

Onan, A., & Korukoğlu, S. (2017). A feature selection model based on genetic rank aggregation for text sentiment classification. *Journal of Information Science*, *43*(1), 25-38.

Onan, A., Korukoğlu, S., & Bulut, H. (2017). A hybrid ensemble pruning approach based on consensus clustering and multi-objective evolutionary algorithm for sentiment classification. *Information Processing & Management*, *53*(4), 814-833.

Onan, A. (2018). An ensemble scheme based on language function analysis and feature engineering for text genre classification. *Journal of Information Science*, *44*(1), 28-47.

Onan, A. (2019). Two-stage topic extraction model for bibliometric data analysis based on word embeddings and clustering. *IEEE Access*, *7*, 145614-145633.

Onan, A. (2020). Mining opinions from instructor evaluation reviews: a deep learning approach. *Computer Applications in Engineering Education*, *28*(1), 117-138.

Onan, A. (2021). Sentiment analysis on product reviews based on weighted word embeddings and deep neural networks. *Concurrency and Computation: Practice and Experience*, *33*(23), e5909.

Onan, A. (2022). Bidirectional convolutional recurrent neural network architecture with group-wise enhancement mechanism for text sentiment classification. *Journal of King Saud University-Computer and Information Sciences*, *34*(5), 2098-2117.

Osenenko, K. M., Kuti, E., Deighton, A. M., Pimple, P., & Szabo, S. M. (2022). Burden of hospitalization for heart failure in the United States: a systematic literature review. Journal of Managed Care & Specialty Pharmacy, *28*(2), 157-167.

Prochaska, J. O., Velicer, W. F., Rossi, J. S., Redding, C. A., Greene, G. W., Rossi, S. R., ... & Plummer, B. A. (2004). Multiple risk expert systems interventions: impact of simultaneous stage-matched expert system interventions for smoking, high-fat diet, and sun exposure in a population of parents. *Health Psychology*, *23*(5), 503.

Seferović, P. M., Vardas, P., Jankowska, E. A., Maggioni, A. P., Timmis, A., Milinković, I., ... & Voronkov, L. (2021). The heart failure association atlas: heart failure epidemiology and management statistics 2019. *European Journal of Heart Failure*, *23*(6), 906-914.

Tankumpuan, T., Asano, R., Koirala, B., Dennison-Himmelfarb, C., Sindhu, S., & Davidson, P. M. (2019). Heart failure and social determinants of health in Thailand: An integrative review. *Heliyon*, *5*(5), e01658.

Taylor, C. J., Ordóñez-Mena, J. M., Roalfe, A. K., Lay-Flurrie, S., Jones, N. R., Marshall, T., & Hobbs, F. R. (2019). Trends in survival after a diagnosis of heart failure in the United Kingdom 2000-2017: population based cohort study. *bmj*, *364*.

Van Riet, E. E., Hoes, A. W., Limburg, A., Landman, M. A., van der Hoeven, H., & Rutten, F. H. (2014). Prevalence of unrecognized heart failure in older persons with shortness of breath on exertion. *European journal of heart failure*, *16*(7), 772-777