Sigma Journal of Engineering and Natural Sciences
Web page info: https://sigma.yildiz.edu.tr
DOI: 10.14744/sigma.2023.00095

Research Article

Machine learning classification models for the patients who have heart failure

Şevval Tuğçe BADİK[1], Mutlu AKAR[1*]

[1]*Department of Mathematics, College of Arts & Sciences, Yildiz Technical University, Istanbul, 34210, Türkiye*

ARTICLE INFO

Article history
Received: 13 February 2023
Revised: 04 April 2023
Accepted: 31 July 2023

Keywords:
Heart Failure; Machine Learning; Classification Algorithms; Principal Component Analysis; CatBoost

ABSTRACT

Heart failure is a cardiovascular disease with significant morbidity and mortality, affecting a growing number of people worldwide [1]. The aim of this paper is to predict the probability of survival of patients by looking at their various characteristics, diseases, and lifestyles in the most successful way by using various machine learning methods. The 299 patients in the data set we use, had left ventricular systolic dysfunction in 2015 and are classified as New York Heart Association (NYHA) class III and IV. The probability of survival of patients is estimated by applying various machine learning methods on the data set. In this study, there are two versions. In the first version of the study, Principal Component Analysis (PCA) is used to reduce the size of the data set. The performance of the machine learning algorithms is then evaluated using a variety of metrics. In the second version, the data set is only subjected to machine learning techniques, and performance is then assessed. Accuracy, Matthews correlation coefficient (MCC), sensitivity, specifity, F_1 score, receiver operating characteristic-area under the curve (ROC-AUC), and precision-recall area under the curve (PR-AUC) values are calculated to measure success. Comparing the two versions reveals that all machine learning algorithms in general have performed better in the second version without PCA. In the second version, the CatBoost algorithm gave the most successful result. Patients with heart failure can have their mortality status predicted using machine learning techniques. The goal of this paper is to look at a variety of characteristics in order to assess the patient's mortality status. The condition of the patient can be improved by selecting the proper treatment based on the mortality situation.

Cite this article as: Badik ŞT, Akar M. Machine learning classification models for the patients who have heart failure. Sigma J Eng Nat Sci 2024;42(1):235–244.

INTRODUCTION

Heart failure (HF) is a cardiovascular disease that causes substantial morbidity and mortality in a growing number of people around the World [1]. Heart failure affects about 6.5 million people in the United States, more than 14 million people in Europe, and 26 million people worldwide, and the number continues to rise [1]. Heart failure affects more men than women, and its incidence rises rapidly as people get older [2].

***Corresponding author.**
*E-mail address: makar@yildiz.edu.tr

This paper was recommended for publication in revised form by Regional Editor Ahmet Selim Dalkilic

Heart failure is a cardiac structural or functional disorder that causes the heart to fail to provide enough oxygen to meet the metabolic needs of the tissues despite normal filling pressures or only at the expense of increased filling pressures. Heart failure is a clinical syndrome caused by structural or functional impairment in the heart, in which patients have typical symptoms such as shortness of breath, ankle swelling and weakness and signs such as elevated jugular venous pressure, pulmonary crackles, and displaced apex beat [3].

Heart failure (HF) is generally classified based on the left ventricular ejection fraction (LVEF), with HF with LVEF greater than 50% being referred to as HF with preserved ejection fraction (HFpEF) and HF with LVEF less than 50% being referred to as HF with reduced ejection fraction (HFrEF). In recent years, the HFrEF has been split into two parts. The LVEF value is classified as a mid-range ejection fraction in the 40%-49% range and a reduced ejection fraction when the LVEF value is lower than 40% [4]. Left ventricular systolic dysfunction is defined as an LVEF less than 40%.

In this work, we use various machine learning strategies to produce the best possible results according to various performance metrics. In the related work section, the studies conducted with this data set are mentioned. In the Data Set section, the variables in the data set were mentioned and examined. The methods applied and the results obtained are mentioned in the Methods section. In the Conclusion section, it is mentioned what purpose this study can be used for.

RELATED WORK

The data set used in this paper was made available in an paper published in 2017. In that paper [5], the survival status of 299 patients who have heart failure in Pakistan in 2015 was analyzed by Ahmad, Munir, Bhatti, Aftab, and Raza using the Cox Regression and Kaplan Meier plot. Following that, a gender-based survival analysis study was conducted by Zahid, Ramzan, Faisal, and Hussain [6] on this dataset using statistical techniques such as Cox Regression in 2019. Finally, Chicco and Jurman [7] applied machine learning algorithms to this data set to determine the patients' survival status in 2020. Chicco and Jurman [7] stated that the survival analysis of patients can be performed by looking at the creatinine and ejection fraction features in the data set using machine learning algorithms. By adding the time feature to these two features, logistic regression is the method that gave the most successful result among various machine learning algorithms [7].

DATA SET

Our data set consists of 299 patients who went to the Faisalabad-Pakistan Cardiology Institute and Allied hospital between April-December (2015). The dataset includes 105 female patients and 194 male patients. In the data set, the age of the patients, whether they have a smoking habit, various analysis results and whether they have diabetes, blood pressure, and anaemia are given. The paper [5] did not mention whether patients had another significant disease. The ages of the patients are between 40 and 95 and the average age is 60.83. Creatinine phosphokinase (CPK), platelets, ejection fraction, creatinine, sodium values were examined in the blood of patients. These variables are given as continuous variables in the dataset. Smoking, diabetes, anemia, blood pressure (BP) and gender are categorical variables in the dataset. In addition to these variables, the patients' death or survival times are given in days, a variable known as time. All of the patients have left ventricular systolic dysfunction and are classified as NYHA class III and IV. Time is 4-285 days with an average of 130 days [5]. The data set contains no missing values. Using these variables, we aim to classify patients' chances of death and survival as accurately as possible. In the data set, death=1 denotes a deceased patient, while death=0 denotes a survived patient.

The enzyme CPK catalyzes the reaction of creatine and adenosine triphosphate (ATP) into phosphocreatine and adenosine diphosphate (ADP). CPK levels of 20 to 200 IU/L are considered natural. CPK levels may be affected by a variety of factors, involving rhabdomyolysis, heart disease, kidney disease, and even some drugs [8].

Platelets are specialized disk-shaped cells in the bloodstream that help to build blood clots and are involved in heart attacks, strokes, and peripheral vascular disease [9]. In patients with HFrEF, there is a connection between platelets and increased mortality. Platelets may be a way to predict poor results in patients with HFrEF [10].

Creatinine refers to serum creatinine, which is a variable included in the data set. Higher creatinine levels and a greater rise in serum creatinine have been linked to a longer stay in the hospital, higher long term mortality, and higher rehospitalization rates [11].

Ejection fraction is the percentage of blood volume ejected in each cardiac cycle. Ejection fraction shows left ventricular systolic performance [12].

Sodium is a mineral that is used in diets all over the world. Sodium is essential for maintaining appropriate blood volume and blood pressure [13].

The total number of patients who died was 96, while the number of patients who survived was 203, as shown in Table 1. Therefore, we may assume that the data set is unbalanced in this situation. The means of continuous variables by dead, survived, and total patients are shown in Table 2.

Table 1. Distribution of patients by categorical variables

Categorical Variables	Number of Dead Patients	Number of Survived Patients
Smoking (1: true)	30	66
Smoking (0: false)	66	137
Diabetes (1: true)	40	85
Diabetes (0: false)	56	118
Anaemia (1: true)	46	83
Anaemia (0: false)	50	120
BP (1: true)	39	66
BP (0: false)	57	137
Gender (1: men)	62	132
Gender (0: women)	34	71
All number of the patients	96	203

Table 2. Means of continuous variables by patients

Continuous Variables	Mean of Dead Patients	Mean of Survived Patients	Mean of All Patients
Age	65.21	58.76	60.83
Ejection Fraction	33.47	40.27	38.08
Sodium	135.38	137.22	136.63
Creatinine	1.83	1.18	1.39
Pletelets	256381.04	266657.49	263358.03
CPK	670.20	540.05	581.84
Time	70.89	158.34	130.26

MATERIALS AND METHODS

It is possible for the model to obtain more successful results by determining the various parameter values in each classification method with the GridSearch CV [14] method. In addition, more successful results can be obtained with various changes on the kernels in the support vector machine method. For example, Akar and Sirakov [15] applied classification methods on the skin lesions dataset by increasing the size of the data using Clifford algebra. Apart from increasing the size by using the kernel, Akar, Sirakov, and Mete [16] obtained a new kernel using Clifford algebra and obtained more successful results with these kernels.

Principal Component Analysis (PCA)

Principal Component Analysis (PCA) can be used to simplify practically any data matrix. A model of the behavior of a physical or chemical system can be created using PCA in conjunction with a well chosen set of objects and variables [17].

Logistic Regression

Many of the fundamental presumptions of ordinary least squares-based linear regression models, such as the linearity of the relationship between the dependent and independent variables, the normality of the error distribution, and the measurement level of the independent variables, are not true in the case of logistic regression. Because it uses a non-linear log transformation of the linear regression, logistic regression can deal with relationships between dependent and independent variables that are not linear [18].

Naïve Bayes

A given example that is described by its feature vector is given the most likely class using a Bayesian classifier. By assuming that features are independent of class, that is, where is a feature vector and is a class, learning such classifiers can be considerably sped up. In spite of this irrational presumption, the resulting classifier, naïve Bayes, is very effective in practice, frequently outperforming much more sophisticated methods. Numerous real-world applications, such as text classification, medical diagnosis, and system performance monitoring, have demonstrated the efficacy of naïve Bayes [19].

K-Nearest Neighbors (KNN)

The k-Nearest Neighbor (KNN) approach applies the classification of the closest of a group of previously classified points to an unclassified sample point [20].

Support Vector Machine (SVM)

Support Vector Machine (SVM) maximizes the geometric margin while minimizing the empirical classification error. For a specific kernel, selecting the right kernel function and parameter values is crucial for the amount of data available. Text categorization, handwritten digit recognition, tone recognition, image classification and object detection, micro-array gene expression data analysis, and data classification are just a few of the real-world issues that SVM have been used to solve [21].

Artificial Neural Network (ANN)

An Artificial Neural Network (ANN) is a simulation of the human brain. The ANN is made up of processing units called neurons. An artificial neuron tries to copy the structure and behavior of a natural neuron. A neuron is made up of inputs and outputs. The output can be thought of as a synapse on the axon in the human brain, the input as dendrites in the human brain. Neurons have the function of determining the activation of neurons. Data classification, pattern recognition and application of data that are not clear can be carried out using ANN [22].

Decision Tree

By grouping the original inputs with major subgroups, the decision tree simplifies complex relationships among input variables and target variables. A method of classifying the population into branches such as segments that are an inverted tree with root nodes, internal nodes and leaf nodes is used. The decision tree algorithm is non-parametric, which allows it to effectively handle large, complex datasets and does not require a complicated parametric structure. In particular when used with a small data set the main disadvantage is that the Decision Tree algorithm is prone to overfitting and underfitting [23].

Random Forest

Multiple decision trees are generated by forest; randomization occurs in two different ways: random data sampling for bootstrap samples, as is done in bagging, and random feature selection for creating individual base decision trees. The main factors that determine a random forest classifier's generalization error are the power of each decision tree classifier and the correlation among base trees. An efficient method for estimating missing data, random forest has the ability to process thousands of input variables without deleting any, provides estimates of significant variables, generates an internal, unbiased generalization error estimate as the forest grows, and maintains accuracy even when a large amount of the data is missing. Random forest classification has been applied in certain areas such as handwriting recognition, detection of hidden web search interfaces, land map classification, multilabel classification [24].

Gradient Boosting

A gradient descent based formulation for boosting methods has been developed to establish an association with the statistical framework. It was referred to as a gradient boosting machine because of this formulation of the boost methods and corresponding models. In gradient boosting machines, the learning process is applied consecutively to new models for more accurate estimation of response variables. The basic idea of the gradient boosting algorithm is to build new baselearners that are maximally correlated with the negative gradient of the loss function associated with the entire ensemble [25].

XGBoost

A popular and highly efficient machine learning technique is tree boosting. XGBoost is a scalable end- to-end tree boosting system. The scalability of XGBoost in all scenarios is the main factor that contributes to its success. In distributed or memory-constrained environments, the system expands to billions of samples and performs more than ten times quicker than currently used popular solutions on a single machine [26].

CatBoost

Yandex created CatBoost, an improved version of the gradient enhanced decision trees (GDBT) algorithm. In a variety of classification and regression tasks, it excelled. CatBoost is better at handling categorical features than more sophisticated gradient boosting algorithms like XGBoost and lightBGM. It is advised to use ordered boosting rather than the traditional GDBT gradient estimation approach to handle the gradient bias and prediction drift issues in CatBoost. CatBoost has reduced the need for super parameter tuning [27].

The data set consists of a total of 12 independent variables along with the follow up time variable and a dependent variable, i.e. the death variable. This study aims to create a model that will make the most successful classification according to these variables.

Firstly, the "event" expression is replaced with "death" in the data set. If death = 1, the patient is dead, if death = 0, the patient is survived. After that, standard scaler is used to scale the data set by eliminating the categorical variables. The categorical variables are is added later back into the data set.

After scaling the continuous variables in the data set, they display a distribution in a certain range as shown in Figure 1. The aim of using standard scaler to scale variables is to see the importance of the variables more clearly by putting the variables in the data set into a certain range and making the model more successful.

Pearson correlation coefficient (PCC) [28] is used to examine the relationship between the dependent variable "death" and the independent variables after scaling the data set.

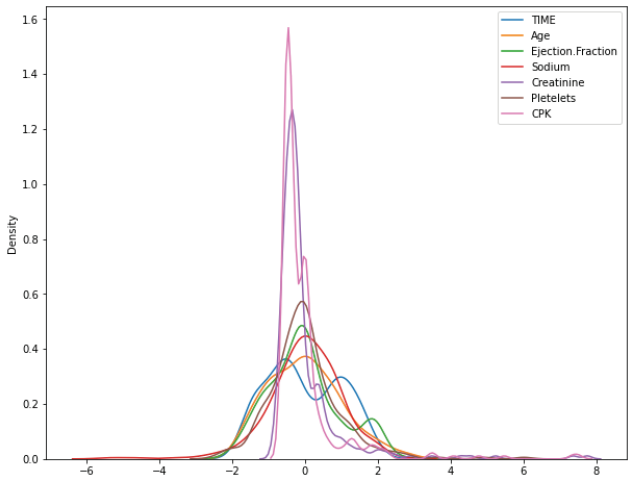

Figure 1. Distribution of continuous variables after standard scaler.

Table 3. Pearson correlation coefficient of variables according to "death"

Variable	Pearson Correlation Coefficient
Time	-0.5270
Age	0.2537
Ejection Fraction	-0.2690
Sodium	-0.1952
Creatinine	0.2942
Pletelets	-0.0491
CPK	0.0627
Gender	-0.0043
Smoking	-0.0126
Diabetes	-0.0019
BP	0.0793
Anaemia	0.0662

According to Table 3, the most significant variables for the variable death are age, time, creatinine, and ejection fraction, as shown by the PCC.

The methods are applied without removing any variables from the data set. First of all, 33% of the data set is separated as test set and 67% of the data set is separated as train set. In other words, 200 patients are allocated for train set and 99 patients for test set. The random state is used to prevent different results every time we run the model. After these stages, two different methods are applied to the data set.

In the first version, 12 independent variables is reduced to 6 components by applying PCA [17] to the model. After that, logistic regression [18], naïve Bayes [19], KNN [20], SVM [21], ANN [22], decision tree [23], random forest [24], gradient boosting [25], XGBoost [26] and CatBoost [27] classifier algorithms are applied to the model. After applying PCA to the model, the importance level of the 6 components as a percentage according to the random forest feature selection is as in Figure 2.

In the second version, the classifier algorithms such as logistic regression, naïve Bayes, KNN, SVM, ANN, decision tree, random forest, gradient boosting, XGBoost, and CatBoost are used. Figure 3 shows the importance level of all variables in percentages using the random forest feature selection.

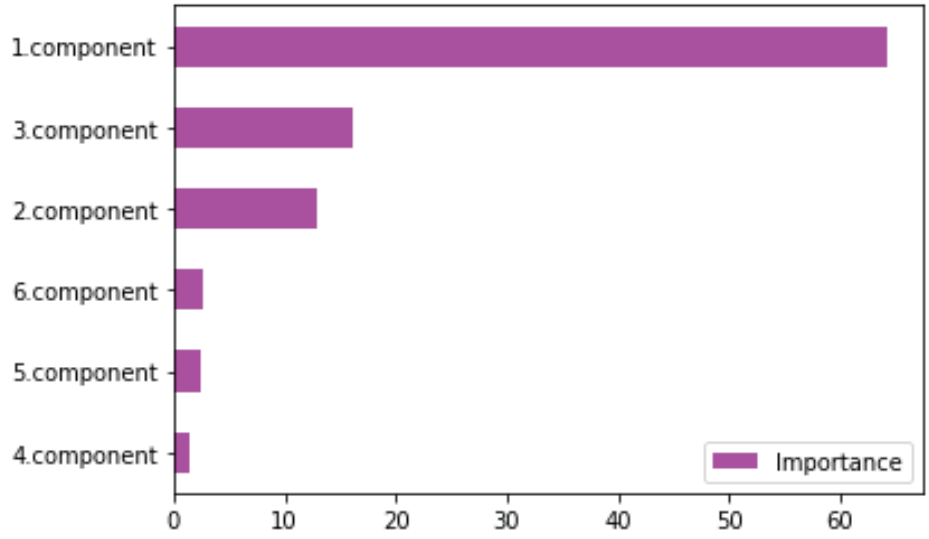

Figure 2. Importance levels of components.

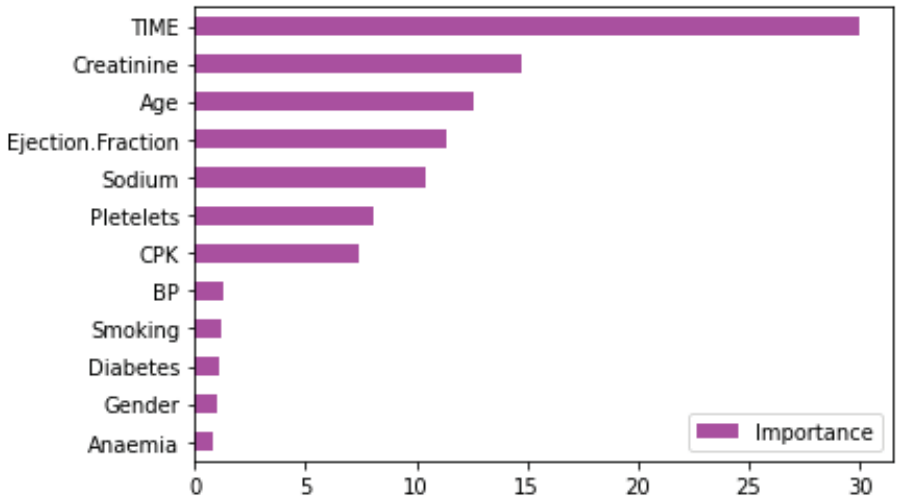

Figure 3. Importance levels of all variables.

RESULTS AND DISCUSSION

Hyper-parameter optimization has been applied to the classification algorithms in two versions. The most suitable parameters for the model are chosen using the GridSearch CV, and then the performance of the model is measured. Accuracy, MCC [29], sensitivity, specifity, F_1 score, ROC-AUC, and PR-AUC values are calculated to measure success.

Logistic regression is applied to the model with solver 'liblinear'. Linear and rbf are selected as kernels in SVM and compared two different results. GridSearch CV is used to find the best parameters for all models except naïve Bayes and logistic regression.

Table 4 shows the performance measurements of the methods after applying PCA to the model. After applying GridSearch CV optimization, n_neighbors= 3 for KNN, C=0.1 for SVM Linear, C=50 and gamma=0.01 for SVM RBF, activation='relu', alpha=0.01, hidden_layer_sizes=(30,30) and solver='adam' for ANN, max_depth=6, min_samples_split=5 for Decision Tree, max_depth=2, max_features=5, n_estimators=1000, min_samples_split=5 for random forest, learning_rate=0.01, n_estimators=500, max_depth=5, min_samples_split=5 for Gradient Boosting, learning_rate=0.01, n_estimators=100, max_depth=3, min_samples_split=2, subsample=0.6 for XGBoost, learning_rate=0.1, iterations=200, depth=5 were selected for CatBoost.

SVM Linear is the most successful method in accuracy, MCC, ROC-AUC, PR-AUC metrics. ANN gave the most succesful results in the sensitivity, F_1 score. Logistic regression, KNN and SVM Linear in the specifity metric gave the same result, as shown in Table 4. MCC values for all models are shown in Figure 4. SVM Linear has the highest MCC value.

In the second version, the methods are applied on 12 independent variables without applying PCA to the model. In this version, no hyper-parameter optimization

Table 4. Methods and performance measurement metrics after applying PCA to the model

Method	Accuracy	MCC	Sensitivity	Specifity	F_1 score	ROC-AUC	PR-AUC
Logistic Regression	0.7677	0.4739	0.5714	0.8750	0.6349	0.7232	0.5597
Naïve Bayes	0.6465	0.1733	0.3429	0.8125	0.4068	0.5777	0.4038
KNN	0.7374	0.3969	0.4857	0.8750	0.5667	0.6804	0.5121
SVM Linear	0.7777	0.4990	0.6000	0.8750	0.6563	0.7375	0.5759
SVM RBF	0.7576	0.4525	0.5714	0.8594	0.6250	0.7154	0.5456
ANN	0.7677	0.4827	0.6286	0.8437	0.6567	0.7361	0.5634
Decision Tree	0.7273	0.3924	0.5714	0.8125	0.5970	0.6920	0.5087
Random Forest	0.7374	0.4011	0.5142	0.8594	0.5806	0.6868	0.5146
Gradient Boosting	0.6970	0.3208	0.5143	0.7969	0.5455	0.6556	0.4703
XGBoost	0.7576	0.4525	0.5714	0.8594	0.6250	0.7154	0.5456
CatBoost	0.7172	0.3735	0.5714	0.7969	0.5882	0.6842	0.4978

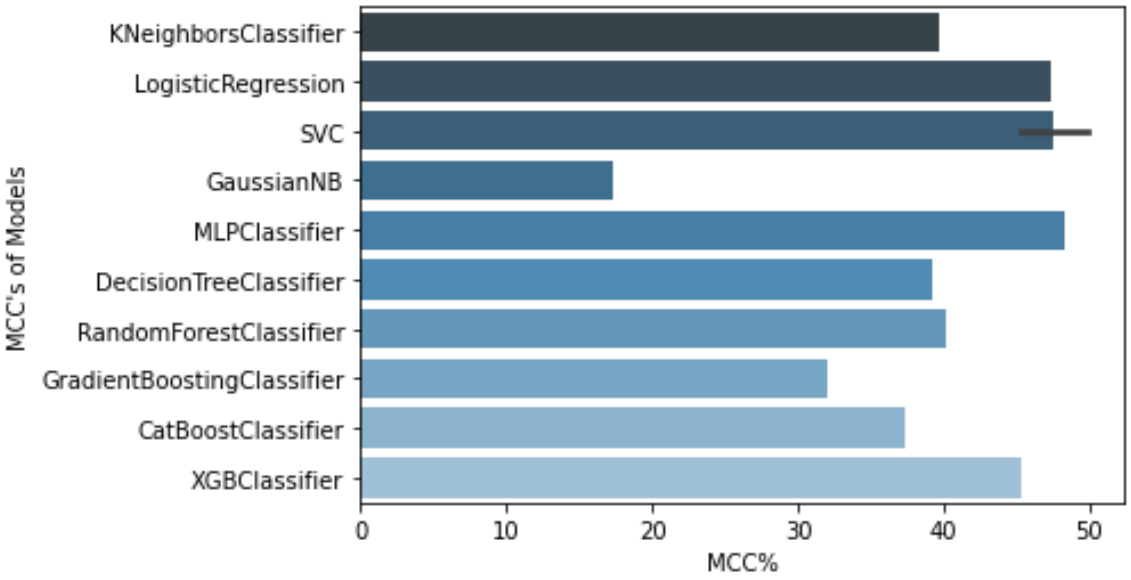

Figure 4. MCC's of the model.

has been made for XGBoost. The lack of hyper-parameter optimization is due to the fact that XGBoost achieves better results without using hyper-parameter tuning. After applying GridSearch CV optimization, n_neighbors=7 for KNN, C=6.0 for SVM Linear, C=50 and gamma=0.01 for SVM RBF, activation='relu', alpha=0.005, hidden_layer_sizes=(50,) and solver='adam' for ANN, max_depth=4, min_samples_split=16 for decision tree, max_depth=10, max_features=2, n_estimators=1000, min_samples_split=10 for random forest, learning_rate=0.01, n_estimators=500, max_depth=3, min_samples_split=10 for gradient boosting, learning_rate=0.01, learning_rate=0.1, iterations=200, depth=5 were chosen for CatBoost. Table 5 shows the results of the performance measurement for the second version, and the CatBoost classification method yield the most successful results in accuracy, MCC, F_1 score, and PR-AUC metrics. On the other hand, the most successful sensitivity result is Gradient Boosting, the most successful specifity result is KNN and the most successful ROC-AUC result is XGBoost. Figure 5 shows the MCC values of the model. Accordingly, it is seen that the highest MCC value belongs to CatBoost, followed by XGBoost and Gradient Boosting classification methods.

The version with PCA applied to the model produces noticeably lower successful results in all performance assessment measures when we compare the two versions.

In the second version, tree-based machine learning methods such as random forest, Gradient Boosting, XGBoost, and CatBoost produce better results than other methods. Among these methods, CatBoost is the most

Table 5. Methods and performance measurement metrics without PCA

Method	Accuracy	MCC	Sensitivity	Specifity	F_1 score	ROC-AUC	PR-AUC
Logistic Regression	0.7778	0.5139	0.6857	0.8281	0.6857	0.7569	0.5813
Naïve Bayes	0.7172	0.3483	0.4571	0.8594	0.5333	0.6583	0.4845
KNN	0.7778	0.4949	0.4857	0.9375	0.6071	0.7116	0.5750
SVM Linear	0.7879	0.5330	0.6857	0.8438	0.6957	0.7647	0.5951
SVM RBF	0.8081	0.5775	0.7143	0.8594	0.7246	0.7868	0.6262
ANN	0.8081	0.5830	0.7429	0.8438	0.7324	0.7933	0.6274
Decision Tree	0.8182	0.6403	0.8857	0.7813	0.7750	0.8335	0.6506
Random Forest	0.8586	0.6906	0.8000	0.8906	0.8000	0.8453	0.7107
Gradient Boosting	0.8586	0.7160	0.9143	0.8281	0.8205	0.8712	0.7107
XGBoost	0.8687	0.7259	0.8857	0.8593	0.8267	0.8725	0.7268
CatBoost	0.8788	0.7348	0.8286	0.9063	0.8286	0.8674	0.7471

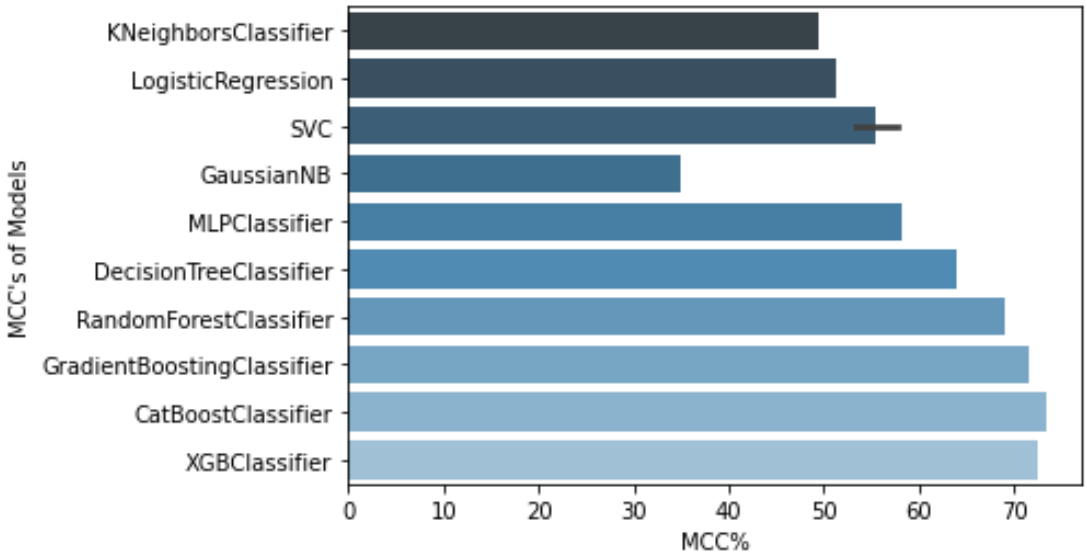

Figure 5. MCC's of the model without PCA.

Table 6. Comparison of the most successful results of the two articles

Method	Accuracy	MCC	Sensitivity	Specifity	F_1 score	ROC-AUC	PR-AUC
Logistic Regression	0.8380	0.6160	0.7850	0.8600	0.7190	0.8220	0.6170
CatBoost	0.8788	0.7348	0.8286	0.9063	0.8286	0.8674	0.7471

successful method for most of the performance measurement metrics.

Table 6 shows a comparison of the most effective machine learning approach for this data set used in Chicco and Jurman's study [7] with the method used in this study.

Logistic Regression produced the best results in Chicco and Jurman's study. They used logistic regression to analyze ejection fraction, serum creatinine, and time variables, and they averaged the results after running it 100 times. They scaled the data set and split the 80% train set into the 20% test set when using this method.

In this study the data set is standardized, and 67% of the training set is split into 33% of the test set. We divide the data set in this manner because the data set is very small, and we want to avoid an unbalanced distribution of the data set as a result. In this work, machine learning algorithms were used on all variables, as opposed to Chicco and Jurman's study.

A comparison of the two results reveals that the CatBoost method produces better results for each performance metric. For instance, while the accuracy in Chicco and Jurman's study was 0.8380, it is now 0.8788 in this study, and the MCC value was 0.6160 in Chicco and Jurman's study, but it is now 0.7348 in this paper.

Thus, in contrast to the Chicco and Jurman's study, performance measurement metrics clearly show that our process of separating the data set, selecting variables, and using the CatBoost machine learning algorithm increases success.

CONCLUSION

Heart failure (HF) is a serious public health problem that affects about 26 million individuals throughout the world. When compared to people with other chronic conditions and the general population, people with heart failure have a much worse quality of life [30]. Despite advances in HF treatment, the rate of morbidity and mortality is still very high. Mortality rates following HF hospitalization are reported to be 10% after 30 days and 22% after a year [31]. HF patients and the death rate from HF are increasing day by day.

In recent years, it has become important to recognize diseases with high mortality rates around the world, such as infectious diseases and cancer [32]. Machine learning can predict the mortality status of patients with HF. This study aims to help determine the patient's mortality status by looking at various variables. The patient's condition can be changed by choosing the appropriate treatment according to the mortality situation.

RESEARCH DATA

The dataset used in this study can be found at the following site.

https://plos.figshare.com/articles/dataset/Survival_analysis_of_heart_failure_patients_A_case_study/5227684/1

ACKNOWLEDGEMENTS

This work was supported by the Research Fund of the Yildiz Technical University, Project Number: FYL-2021-4416. The authors would like to thank for this support.

AUTHORSHIP CONTRIBUTIONS

Authors equally contributed to this work.

DATA AVAILABILITY STATEMENT

The authors confirm that the data that supports the findings of this study are available within the article. Raw data that support the finding of this study are available from the corresponding author, upon reasonable request.

CONFLICT OF INTEREST

The author declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

ETHICS

There are no ethical issues with the publication of this manuscript.

REFERENCES

[1] Morris JH, Chen L. Exercise training and heart failure: A review of the literature. Card Fail Rev 2019;5:57–61. [CrossRef]

[2] Bui AL, Horwich TB, Fonarow GC. Epidemiology and risk profile of heart failure. Nat Rev Cardiol. 2011;8:30–41. [CrossRef]

[3] McMurray JJ, Adamopoulos S, Anker SD, Auricchio A, Böhm M, Dickstein K, et al. ESC Guidelines for the diagnosis and treatment of acute and chronic heart failure 2012: The Task Force for the Diagnosis and Treatment of Acute and Chronic Heart Failure 2012 of the European Society of Cardiology. Eur Heart J 2012;33:1787–1847.

[4] Gong FF, Jelinek MV, Castro JM, Coller JM, McGrady M, Boffa U, et al. Risk factors for incident heart failure with preserved or reduced ejection fraction, and valvular heart failure, in a community-based cohort. Open Heart 2018;5:e000782. [CrossRef]

[5] Ahmad T, Munir A, Bhatti SH, Aftab M, Raza MA. Survival analysis of heart failure patients: a case study. PLoS One 2017;12:e0181001. [CrossRef]

[6] Zahid FM, Ramzan S, Faisal S, Hussain I. Gender-based survival prediction models for heart failure patients: a case study in Pakistan. PLoS One 2019;14:e0210602. [CrossRef]

[7] 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 2020;20:16. [CrossRef]

[8] Aujla RS, Patel R. Creatine Phosphokinase. In: StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2021.

[9] Gregg D, Goldschmidt-Clermont PJ. Cardiology patient page. Platelets and cardiovascular disease. Circulation 2003;108:e88–e90. [CrossRef]

[10] Mojadidi MK, Galeas JN, Goodman-Meza D, Eshtehardi P, Msaouel P, Kelesidis I, et al. Thrombocytopaenia as a prognostic indicator in heart failure with reduced ejection fraction. Heart Lung Circ 2016;25:568–575. [CrossRef]

[11] Metra M, Cotter G, Gheorghiade M, Dei Cas L, Voors AA. The role of the kidney in heart failure. Eur Heart J 2012;33:2135–2142. [CrossRef]

[12] Bamira D, Picard MH. Imaging: Echocardiology-Assessment of Cardiac Structure and Function. In: Encyclopedia of Cardiovascular Research and Medicine. 2018. p. 35–54. [CrossRef]

[13] Patel Y, Joseph J. Sodium intake and heart failure. Int J Mol Sci 2020;21:9474. [CrossRef]

[14] Bergstra J, Bengio Y. Random search for hyper-parameter optimization. J Mach Learn Res 2012;13:281–305.

[15] Akar M, Sirakov NM. Support vector machine skin lesion classification in Clifford algebra subspaces. Appl Math 2019;64:581–598. [CrossRef]

[16] Akar M, Sirakov NM, Mete M. Clifford algebra multivectors and kernels for melanoma classification. Math Methods Appl Sci 2022;45:4056–4068. [CrossRef]

[17] Wold S, Esbensen K, Geladi P. Principal component analysis. Chemom Intell Lab Syst 1987;2:37–52. [CrossRef]

[18] Park HA. An introduction to logistic regression: from basic concepts to interpretation with particular attention to nursing domain. J Korean Acad Nurs 2013;43:154–164. [CrossRef]

[19] Rish I. An empirical study of the Naive Bayes classifier. In: IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence. 2001. p. 41–46.

[20] Cover T, Hart P. Nearest neighbor pattern classification. IEEE Trans Inform Theory 1967;13:21–27. [CrossRef]

[21] Durgesh S, Bhambhu L. Data classification using support vector machine. J Theor Appl Inf Technol 2010;12:1–7.

[22] Shiruru K. An introduction to artificial neural networks. Int J Adv Res Innov Ideas Educ 2016;1:27–30.

[23] Song YY, Lu Y. Decision tree methods: applications for classification and prediction. Shanghai Arch Psychiatry 2015;27:130–135.

[24] Kulkarni V, Sinha P. Random forest classifiers: A survey and future research directions. Int J Adv Comput 2013;36:1144–1153.

[25] Natekin A, Knoll A. Gradient boosting machines, a tutorial. Front Neurorobot 2013;7:21. [CrossRef]

[26] 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 Aug 13-17; San Francisco, CA. New York: ACM; 2016. p. 785–794. [CrossRef]

[27] Wang Y, Guo R, Huang L, Yang S, Hu X, He K. m6AGE: A predictor for N6-methyladenosine sites identification utilizing sequence characteristics and graph embedding-based geometrical information. Front Genet 2021;12:670852. [CrossRef]

[28] Benesty J, Chen J, Huang Y, Cohen I. Pearson correlation coefficient. In: Noise Reduction in Speech Processing. Springer Topics in Signal Processing Vol 2. Heidelberg: Springer; 2009. p. 1–4. [CrossRef]

[29] Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta 1975;405:442–451. [CrossRef]

[30] Heo S, Lennie TA, Okoli C, Moser DK. Quality of life in patients with heart failure: ask the patients. Heart Lung 2009;38:100–108. [CrossRef]

[31] Çavuşoğlu Y, Zoghi M, Eren M, Bozçalı E, Kozdağ G, Şentürk T, et al. Post-discharge heart failure monitoring program in Turkey: Hit-PoinT. Anatol J Cardiol 2017;17:107–112. [CrossRef]

[32] Maayah B, Abu Arqub O, Alnabulsi S, Alsulami H. Numerical solutions and geometric attractors of a fractional model of the cancer-immune based on the Atangana-Baleanu-Caputo derivative and the reproducing kernel scheme. Chin J Phys 2022;80:463–483. [CrossRef]