Healthcare Analytics 3 (2023) 100182

Contents lists available at ScienceDirect

Healthcare Analytics

journal homepage: www.elsevier.com/locate/health

A machine learning approach for risk factors analysis and survival prediction of Heart Failure patients

Md. Mamun Ali [a], Vian S. Al-Doori [b], Nubogh Mirzah [c], Asifa Afsari Hemu [d], Imran Mahmud [a], Sami Azam [e], Kusay Faisal Al-tabatabaie [f], Kawsar Ahmed [g,h,*], Francis M. Bui [g], Mohammad Ali Moni [i,**]

[a] Department of Software Engineering (SWE), Daffodil International University (DIU), Sukrabad, Dhaka 1207, Bangladesh
[b] Medical Instrumentation Techniques Engineering, Al-Rafidain University College, Baghdad 10064, Iraq
[c] Computer Technologies Engineering, Al-Turath University College, Baghdad, Iraq
[d] Department of Computer Science and Engineering (CSE), Daffodil International University (DIU), Sukrabad, Dhaka 1207, Bangladesh
[e] College of Engineering, IT and Environment, Charles Darwin University, Casuarina, NT 0909, Australia
[f] Middle Technical University, Electrical Engineering Technical College, Baghdad, Iraq
[g] Department of Electrical and Computer Engineering, University of Saskatchewan, 57 Campus Drive, Saskatoon, SK S7N 5A9, Canada
[h] Group of Biophotomatiχ, Department of Information and Communication Technology, Mawlana Bhashani Science and Technology University, Santosh, Tangail 1902, Bangladesh
[i] Artificial Intelligence & Digital Health, School of Health and Rehabilitation Sciences, Faculty of Health and Behavioural Sciences, The University of Queensland, St Lucia, QLD 4072, Australia

ARTICLE INFO

Keywords:
Machine learning
Heart failure
Risk factor analysis
Feature importance
Survival prediction

ABSTRACT

In this study, we propose machine learning (ML) for risk factors analysis and survival prediction of Heart Failure (HF) patients using a survival dataset. Five supervised ML methods are applied to the dataset: Decision Tree (DT), Decision Tree Regressor (DTR), Random Forest (RF), XGBoost, and Gradient Boosting (GB) algorithms. We compare the applied algorithms' performances based on accuracy, precision, recall, F-measure, and log loss value and show RF provides the highest accuracy of 97.78%. The analysis of the risk factors shows the most predictive features based on coefficients and feature importance. The top six risk factors for HF patients are serum creatinine (SC), age, ejection fraction (EF), platelets, creatinine phosphokinase (CPK), and SS (SS). Further analysis of these factors shows significant clustering of the features. The survival analysis finds that the increment of SC, age, and SS and the decrement of EF are the most significant risk factors for HF patients. Our results suggest that HF survival prediction is possible with higher accuracy using the proposed model. Our ML models are useful in clinical settings for screening patients with HF probability.

1. Introduction

Cardiovascular diseases (CVDs) including coronary heart disease (heart attacks), strokes, and HF [1] form a major burden of diseases globally. According to the World Health Organization (WHO), CVDs including HF are responsible for 31% of total death worldwide [2]. HF occurs when the heart is incapable of delivering enough blood throughout the body. This may be affected by high blood pressure, diabetes, coronary heart diseases, and other heart problems or disorders [3].

There are two categories of severe HF: HF with retained EF (HFPEF) and HF with reduced EF (HFREF) [4]. In a clinical environment, the difference between HFPEF and HFREF is significant. HFREF is most common in male patients and is caused by cardiomyocyte loss. HFPEF, on the other hand, is frequently detected in older female patients who have (a group of) non-cardiac comorbidities such as hypertension, T2DM, stroke, anaemia, pulmonary illness, liver problems, sleep apnea, gout, and cancer [5]. Though the prognosis tends to be similar between the two HF subsets, there are significant variations in cause-specific survival, which may be critical in risk stratification and disease prevention [6]. The best way to differentiate between HFREF and HFPEF is to use echocardiography data. While echocardiography should be conducted at some stage for all HF patients, this test has not always

* Corresponding author at: Department of Electrical and Computer Engineering, University of Saskatchewan, 57 Campus Drive, Saskatoon, SK S7N 5A9, Canada.
** Corresponding author.
E-mail addresses: mamun35-274@diu.edu.bd (M.M. Ali), vian.kasim@ruc.edu.iq (V.S. Al-Doori), nobojh.hussein@turath.edu.iq (N. Mirzah), asifa15-1716@diu.edu.bd (A.A. Hemu), imranmahmud@daffodilvarsity.edu.bd (I. Mahmud), sami.azam@cdu.edu.au (S. Azam), kusay.faisal@mtu.edu.iq (K.F. Al-tabatabaie), kawsar.ict@mbstu.ac.bd, k.ahmed.bd@ieee.org, k.ahmed@usask.ca (K. Ahmed), francis.bui@usask.ca (F.M. Bui), m.moni@uq.edu.au (M.A. Moni).

https://doi.org/10.1016/j.health.2023.100182
Received 19 February 2023; Received in revised form 27 March 2023; Accepted 18 April 2023

been performed, and medical decisions may need to be taken in the absence of echocardiographic results. More than one-third of HF patients in one US Medicare cohort did not have echocardiography performed in the clinic [7].

HF is a widespread problem. To predict HF survival, it is of crucial importance to analyze the risk factors. Though many studies have been performed on heart disease prediction, very limited work is conducted on the survival prediction of a HF patient. Pocock et al. in 2006 developed a prognostic model using multivariate cox regression employing baseline candidate variable to identify the causes related to mortality of chronic HF patients [8]. The model found 21 predictor variables for survival prediction. The proposed model is not technically up to date since more modern machine leaning algorithms performing better with less computational cost. In 2013, Pocock et al. conducted a clinical meta-analysis based on 39,372 patients to predict survival [9] and found thirteen mortality predictors, which has no specific explaination. Kleinbaum and Klein in 2012 proposed a model to predict mortality for HF patients based on Kaplan–Meier survival curves and the log-rank test [10]. They proposed a model to predict the survival of HF patients. But their proposed model is not well performed in terms of performance measurement metrics. In 2013, Austin et al. conducted a study on HF and proposed a model to predict HF, and death and survival for HF with preserved EF (HFPEF) and HFREF [4]. They found that tree-based methods performed better than conventional classifiers. However, their proposed model is not properly validated that their model is actually capable of predicting HF death events.

Ahmad et al. conducted a survival analysis and proposed a model to predict death or survival events using the COX Regression model in 2017 [11]. Regression model is not well performed in terms of classification. On the other hand, they did not validate that their model is well performed in terms of computational cost. Zahid et al. in 2019 introduced a gender-based model to predict mortality for HF using biostatic methods rather than a ML model [12]. Chicco and Jurman in 2020 proposed a model to predict the survival and death events [1]. They demonstrated that it is possible to predict the survival of an HF patient based on SCe and EF only. However, only two attributes are not capable of significantly predict the mortality of HF patience since other factors are also associated with HF. Wussler et al. in 2020 compared two models such as biomarkers (noninvasive and highly reproducible quantitative tools), and a statistical score-based model to predict the survival of HF patients [13] and found that biomarker identification is the best tool for the survival prediction. However, the biomarker tool cannot effectively be used by for non-clinicians. This study therefore aims to develop a ML based model, which can predict HF patient survival and analyze the risk factors responsible for HF patient mortality. Cho et al. in 2021 compared pre-existing methods and built a ML model to predict the cardiovascular risk prediction [14]. But they did not perform survival prediction o f those high-risk patients.

In this study, we analyze an HF survival event dataset. After preprocessing the HF dataset, we perform statistical analysis, exploratory data analysis (EDA) and ML analysis. Statistical analysis and EDA extract some important information related to death and survival events due to HF. A ML approach is applied to build a model to predict whether an HF patient will survive or not. Using this ML approach, the features are ranked based on the coefficient values for each classification algorithm. This results in the identification of the features, that foremost contribute to mortality due to HF.

This study is designed to build an efficient ML model to predict the survival of HF patients. Knowledge of the risk factors is also crucial. Therefore, this study aims to build a ML model and analyze the risk factors to predict the survival HF patients.

Our Contributions in this study is mentioned as follows:

a. Building a well performed ML model by tuning hyper parameter and processing the dataset, which has given a better accuracy compare to previously proposed models.
b. Validating the supervised ML algorithm's classification by unsupervised ML.
c. Finding the errors what the model was doing through unsupervised ML such as model based clustering and principal component analysis (PCA).
d. Finding the significant risk factors for HF patients.

2. Materials & methods

In this study, Python programming language in google colab was employed to conduct the study. The overall research methodology is depicted in Fig. 1.

2.1. Details of dataset

In this study, a HF survival event dataset is employed to analyze and build the model. The dataset is collected from Kaggle [15]. The dataset contains 299 instances, where in 96 cases the outcome was death and whereas the remaining ones were from patients who survived. The details of the dataset are described in Table 1.

2.2. Data preprocessing

Usually, raw data contains noisy and inconsistent instances, which is a barrier to a good prediction and ML analysis result [16]. Data preprocessing is an important task for ML analysis since it prepares a dataset to get a better analytical result. The HF dataset was therefore preprocessed. First missing values in the dataset were dealt with, according to data type. Then feature engineering is performed to transform the feature types where necessary as the raw data contains both string and numeric features. Some of the features were converted to numeric to nominal types in the feature engineering stages. Outliers and extreme values are removed from the dataset to increase the quality of the dataset. Finally, the dataset is balanced using SMOTE (Synthetic Minority Oversampling Techniques). Though dataset is small, it can be used to build ML model [17].

2.3. Statistical & exploratory data analysis

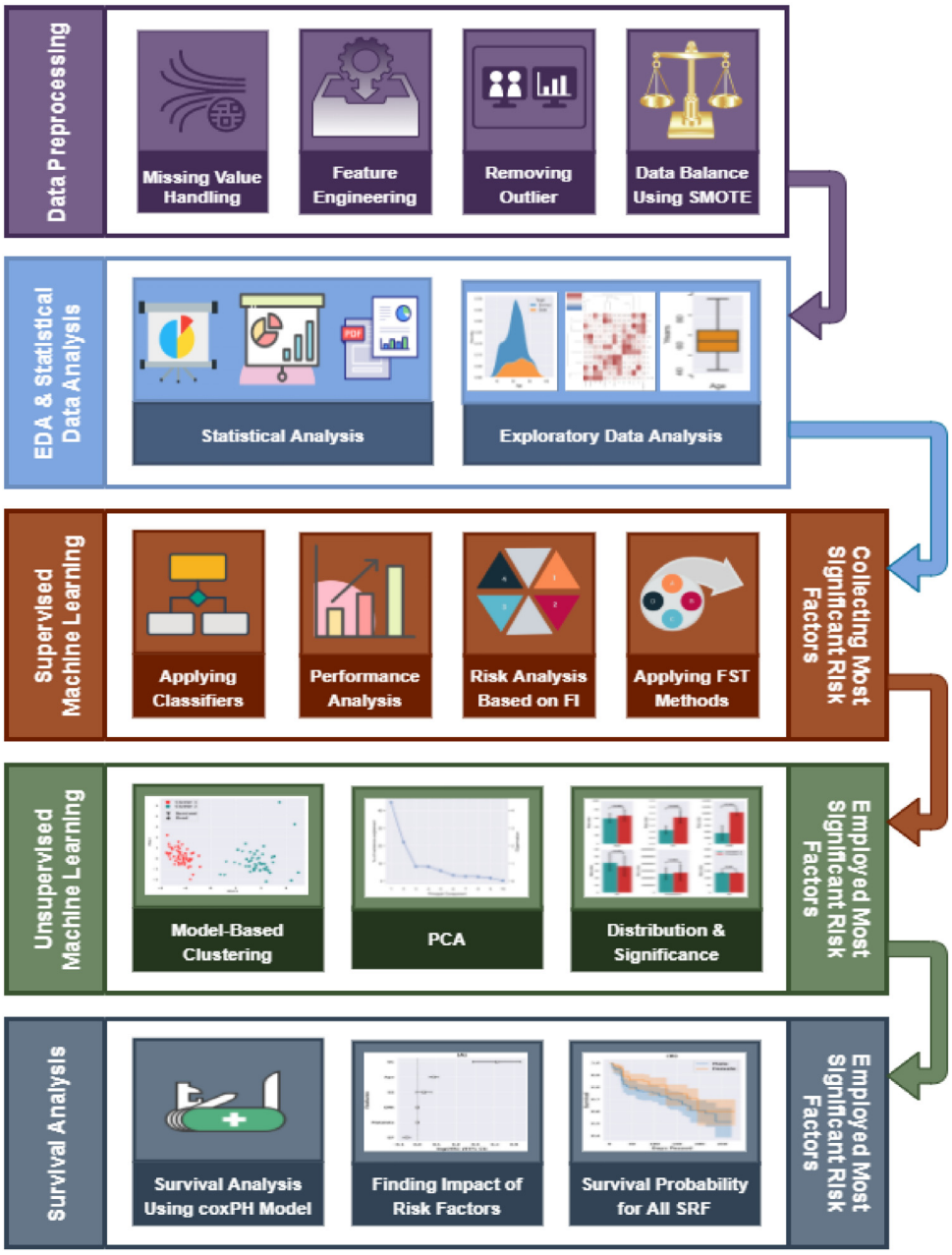
Generally, a dataset contains different types of important information, which cannot be found easily. By analyzing the dataset, this information can be extracted. In this study, we have employed some statistical and EDA on the dataset. in order to uncover hidden patterns and trends [18]. EDA is a method to analyze datasets and characterize their major properties, mainly using graphical approaches. Before the modeling process, EDA is used to examine the data [19,20].

2.4. Supervised ML analysis

After data preprocessing, five different supervised ML algorithms, DT, RF, XGB, GB, and DTR, are applied for further analysis. These algorithms are selected based on literature review. In this study, train test split method was employed to train and test the models. 70% of the data were employed to train the model and other 30% of dataset was employed to test the models. The details of these algorithms are as follows:

- **Decision Tree (DT)**

One of the earliest and most well-known supervised classification algorithms is the DT. Logically a decision is represented by a DT to categorize data objects into a data structure. A DT consists of nodes and edges, and typically these nodes contain several tiers, with the starting or the first node referred to as the root or parent node and others referred to as child or leaf nodes. Internal nodes (nodes with at least one leaf) display input variables or attribute checks. The supervised ML classifiers branch towards the required leaf node based on the evaluation results, and this approach of testing and branching



- FI: Feature importance and SRF: Significant risk factors

Fig. 1. Pipeline of our proposed ML models and data analysis framework.

Table 1
Data description.

Feature name	Feature type	Explanation of feature
Age	Float	Age of the participant
Anaemia	Boolean	Decrease of red blood cells or hemoglobin
CPK	Integer	The level of CPK enzyme in the blood.
Diabetes	Boolean	The patient has diabetes or not
EF	Integer	Percentage of blood leaving the heart at each contraction.
HBP	Boolean	The patient has high blood pressure or not
platelets	Float	Platelets in the blood
SC	Float	The level of creatinine in the blood
SS	Integer	The level of sodium in the blood
Sex	Boolean	Male or Female
Smoking	Boolean	Do smoke or not
Time	Integer	Follow-up period
DEATH_EVENT	Boolean	If the patient died during the follow-up period (Target attribute)

is repeated until the leaf node is reached [21]. The leaf or child nodes represent the result of a DT. DTs are simple to read and understand, and they are now a standard part of patient monitoring protocols [22]. When exploring the tree to classify a test observation, the output of all trials conducted at each node along the pipeline would offer sufficient information to predict the node's category. Gini criterion was chosen for this algorithm to conduct this study and max depth was 5.

• **Random Forest (RF)**

A RF is a type of classification algorithm, which consists of many DTs, analogous to how a forest has many trees [23]. Deep DTs can introduce an issue known as overfitting at the training stage with a training dataset, which results in a large change in classification outcomes for minor differences in the test sample. They are susceptible to their training results, making them prone to make mistakes in the test observations. The various DTs, which are part of RF, are trained with various sections of the training dataset. The sample's input values must be sent along with each DT of the forest to identify a new sample. Each DT then uses a particular part of the input values and returns a result as a classification output. The forest then selects the output with the highest number of "votes" (for the outputs of categorical segmentation) or the sum of all trees in the forest (for the outputs of numeric segmentation). Since the results of several DTs are considered by the RF, the variation caused by a single DT for a similar dataset will be reduced [24]. In this study, n_estimator was selected 65 and max depth was selected 8.

• **XGBoost**

Extreme GB(XGBoost) is another ML algorithm, which is designed based on the concept of DTs. It employs a GBsystem [25] and is a modular method of tree boosting commonly used in ML. XGBoost has the strong ability to provide solutions for real-world problems, with limited resources. XGBoost offers parallel tree boosting (also known as GBDT, GBM) to address a range of data science problems quickly and accurately [26]. For this study, the learning rate for this algorithm was set 0.59.

• **Gradient Boosting (GB)**

GB is a type of classifier which is designed following the DT method with a fixed size that incorporates several basic predictors [27]. It constructs the model in a similar way as other boosting methods, and it adds them using optimization with a self-assertive loss function. A principal goal of the GBM is to find a function *F(x),* which constrains its loss function *L (y, F(x)*. A weighted line of the base learners is an approved projected prototype. In this algorithm number of estimator was chosen 100. Other parameters were default in this study.

• **Decision Tree Regressor (DTR)**

In the context of tree systems, the DTR builds regression or classification prototypes. It splits a dataset into successively smaller sections while creating a related DT, creating a tree containing leaf nodes and decision nodes. A decision node (such as Outlook) has two or three components (such as Sunny, Overcast, and Rainy), each of which represents a score for the feature being assessed. A leaf node (such as Hours Played) represents a quantitative target decision. The root of a tree is the uppermost contributing decision node. A DTR can handle both category and continuous data. Squared_error was chosen for criterion and best was selected as splitter.

2.5. Performance evaluation criteria

This study has used Accuracy, Precision, Recall, F-Measure, and Log Loss to find the best performing classification algorithm. The risk factors are analyzed by calculating coefficient values. This identifies the important risk features. The equations to find the value of these metrics are mentioned as following [28–32]:

$$Ac = \frac{TP + TN}{TP + FP + TN + FN} \tag{1}$$

$$Precision = \frac{TP}{TP + FP} \tag{2}$$

$$Recall = \frac{TP}{TP + FN} \tag{3}$$

$$F1 = \frac{2 * Precision * Recall}{Precision + Recall} \tag{4}$$

$$L_{(log)}(y, p) = (y\, log\, (p) + (1 - y)\, log(1 - p)) \tag{5}$$

Here, *Ac* refers to accuracy. *TP, FP, FN,* and *TN* represent true positive, false positive, false negative, and true negative.

2.6. Risk factor analysis using feature selection methods

In this study, risk factors are analyzed using two methods. First one is based on feature importance score of an individual algorithm. Another one is based on feature selection methods. Four different feature selection methods were used in this study such as Correlation based Feature Subset Evaluation (CFSSE), Gain Ratio Attribute Evaluation (GRAE), Info Gain Attribute Evaluation (IGAE), and Relief Attribute Evaluation (RFAE). All of these methods provide a score for each feature and higher score refers to higher important and more significant risk factors for HF patients to predict death or survival.

2.7. Model based clustering

Clustering is a data analysis method that divides data into multiple homogeneous groups in order to comprehend or interpret the phenomena being investigated. Model-based clustering is a clustering technique for high dimensionality data [26]. In this study, the Gaussian mixture model is applied to perform model-based clustering. In this model, each element is considered a multivariate Gaussian distribution [27]. The element that generates a certain instance specifies the cluster to which the instance belongs [33].

2.8. Principal Component Analysis (PCA)

PCA, also known as dimensionality reduction, is a multivariate approach that investigates a data frame in which events are characterized by numerous inter-correlated numeric target variables [34]. Its purpose is to retrieve the key points from the data frame, characterize it as a set of new orthogonal parameters known as principal components, and depict the linked pattern of occurrences and parameters as dots on a diagram [35]. In this study, PCA is performed to represent the found clusters in two-dimensional pattern so that it can be found the data points can be correctly identified by the models. At first, the dataset was standardized to perform the PCA. Then the scaled dataset is used to perform PCA. The number of components was chosen 2 for PCA in this study.

2.9. Survival analysis & prediction

Survival analysis is a set of statistical processes for ML in which the outcome variable of concern is the amount of time before an event takes place [36]. A survival model analyzes time-to-event past data and generates estimates, known as survival curves, that show how the chance of the event happening increases with time [33,37–40]. In this study, we employ the CoxPHFitter model for survival analysis and survival time prediction. CoxPHFitter is a Cox proportional hazard model, which proposes that a patient's log-hazard is a linear function of their covariates and a population-level baseline hazard that varies with time [41,42].

3. Results & discussion

3.1. Statistical & EDA

Table 2 gives an overview of the descriptive statistical analysis result of 203 patients HF patients who survived and 96 HF patients who

Table 2
Descriptive statistical representation of information for patients who survived and who died (mcg/L: micrograms per liter. mL: microliter. mEq/L: milliequivalents per liter).

Categorical feature							
Features	Category	All patients, N = 299 (%)	Patient's condition				P Value
			Dead 96 (%)		Survived 203 (%)		
Sex							<0.001
	Male	179 (59.87)	59 (61.46)		120 (59.11)		
	Female	120 (40.13)	37 (38.54)		83 (40.89)		
Smoking							0.536
	Yes	89 (29.77)	31 (32.29)		58 (28.57)		
	No	210 (70.23)	65 (67.71)		145 (71.43)		
Anaemia							0.720
	Yes	113 (37.79)	42 (43.75)		71 (34.98)		
	No	186 (62.21)	54 (56.25)		132 (65.02)		
Diabetes							<0.001
	Yes	167 (55.85)	48 (50)		119 (58.62)		
	No	132 (44.15)	48 (50)		84 (41.38)		
HBP							0.001
	Yes	133 (44.48)	52 (54.17)		81 (39.90)		
	No	166 (55.52)	44 (45.83)		122 (60.10)		
Numeric feature							
Features	All patients		Patient's condition				P Value
			Dead		Survived		
	Mean (STD)	Median (IQR)	Mean (STD)	Median (IQR)	Mean (STD)	Median (IQR)	
Age (Years)	62.09 (11.74)	62 (54–70)	67.11 (12.64)	70 (54.75–77)	59.71 (10.53)	61(52.50 –67)	0
CPK (mcg/L)	607.01 (956.03)	244 (103-582)	871.73 (1400.51)	422.50 (122–855)	481.81 (613.89)	203 (96.50–582)	<0.001
EF (%)	39.37 (12.37)	38 (30–45)	36.78 (14.47)	35 (25–45)	40.60 (11.08)	38 (35–50)	<0.001
Platelets (kilo-platelets/mL)	267 018.25 (93 832.28)	263 358.03 (221 000–298 000)	268 642.23 (104 494.70)	263 358.03 (210 000–321 000)	266 250.26 (88 609.97)	263 358.03 (224 000–293 000	<0.001
SC (mg/dL)	1.52 (1.40)	1.1 (0.90–1.70)	2.20 (2.17)	1.35 (1.0-2.10)	1.20 (0.60)	1.10 (0.90–1.30)	<0.001
SS (mEq/L)	136.70 (5.76)	137 (134–140)	136.51 (5.42)	136 (134–140)	136.78 (5.92)	138 (134–140)	0
Time (Days)	127.17 (75.81)	120 (73–196)	77.27 (67.30)	41 (26–126)	150.77 (67.88)	83 (46–212)	<0.001

died. The total number of patients was 299. The table lists the number of dead and survived patients for each group of nominal attributes. Mean, Standard Deviation (STD), Median and interquartile ranges are also given in the table. P values for each feature are also given. According to Table 2, it is found that smoking is mostly responsible for death of HF patients. Besides, survival rate is high among anaemia negative patients. High blood is also one of the most risk factors. The mortality rate is high among the patients who have high blood pressure. Higher age, CPK, and SC are also risky for the HF patients.

Fig. 2 depicts a cluster map to visualize the correlation and clusters of features based on the relationship of the features. Clustering is the process of grouping features based on the correlation between the features. Correlation and clustering are based on p-values of all the features with respect to each other. In this figure, a pair of features are considered statistically significant when the p-value is less than or equal to 0.05 (p-value ≤ 0.05). According to Fig. 2, SC, Age, HBP, time of follow up, and CPK are the mostly statistically significant for death event of a HF patient. In addition to that some other attributes are also significant those are depicted in Fig. 2.

Fig. 3 is a boxplot, where the data are represented before and after removing the outliers. The figure represents the numeric attributes, such as minimum range, maximum range and the first, second (known as medium), and third quartile. Instances outside the minimum and maximum range are considered outliers and are removed from the dataset. The removal of outliers is essential to increase the data quality. Outliers affect the performances of the ML model's performance. Fig. 3 shows that the applied dataset is now ready for applying ML model.

Fig. 4 depicts a Kernel Density Estimate (KDE) plot for all numeric attributes for both survived and perished patients due to HF. This is an approach to illustrate the distribution of instances in a dataset. Fig. 4 shows the data using a continuous probability density curve in two dimensions. The figure indicates that HF is very risky for people aged 50 years or more. The KDE plot for each of the numeric attributes gives a clear overview of the ranges of values for attributes which are risk factors for HF patients. According to Fig. 4, the range of CPK for survival patient is below 1500. So, below 1500 CPK is safe for the HF patients. Below 30 EF is the risky for the EF patients. Platelets between 180 000 and 360 000 is the safe for the HF patients. Besides, 0.5–2.0 SC is safe and above 130 SS is safe for the HF patients according to Fig. 4.

3.2. Result of supervised ML

A model was designed for the ML analysis as depicted in Fig. 1. Both supervised ML and unsupervised ML were performed after pre-processing the data. In the data preprocessing phase, the dataset is processed in such a way that it can work with different ML algorithms. Then two ML approaches were applied to the dataset. This section describes the supervised ML results. We applied different classification

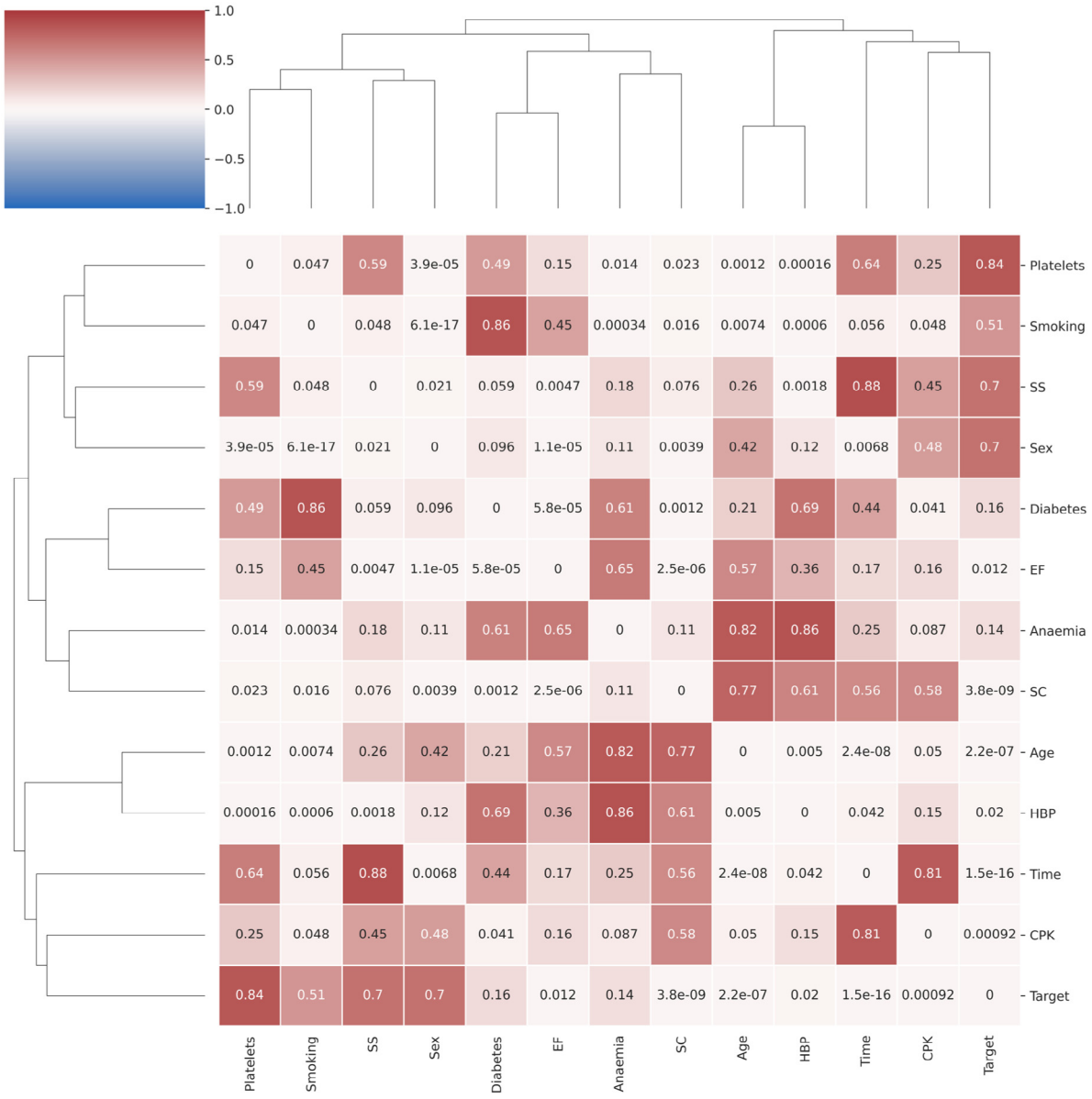


Fig. 2. Illustration of correlation and clustering among all the features based on P-value.

Table 3
Hyper tuned parameters for all the applied models.

Algorithms	Hyper tuned parameters
DTR	Criterion = "squared_error" and splitter = "best"
XGB	Learning_rate = 0.59.
DT	Criterion = "gini" and max_depth = 5.
GB	n_estimator = 100
RF	n_estimator = 65 and max_depth = 8.

methods to find the best performing classification algorithms based on evaluation metrics such as accuracy, precision, recall, F-Measure, and log loss value. Five different ML classification algorithms were applied: RF, XGB, DT, DTR, and GB, implemented using python programming language (Python 3). The hyper tuned parameters for these algorithms have been represented in Table 3. The importance of the individual features for each classifier is determined, based on the feature importance value which are found by the applied individual ML algorithms. The coefficient value was calculated to identify the significant features which are most responsible for HF death events.

In this study, we found that all algorithms had satisfactory classification results to classify the HF dataset. Details of the supervised ML results are described below.

3.2.1. Performance measurement of ML approaches

Table 4 gives an overview of the performance of all the applied classification algorithms. As can be seen in the table, DTR achieved the lowest accuracy (85.24%), while XGB produced the lowest result for precision (94%), recall (94%), F-score (94%) and log loss (1.919). The best performing classification algorithm is RF, with a 97.78% accuracy, 97% precision, 97% recall, 97% F-score and a log loss of 0.767.

Fig. 5 represents the area under the ROC (AUROC) Curve and area under the Precision–Recall Curve (AUPRC) respectively. Both of the curves show that RF is the best performing classification algorithm, covering 90.6% area under PRC and 99% area under the ROC curve. Based on these curves DTR is the worst performing classification algorithm to predict the survival of HF patients. The higher value of area of ROC, and PRC refer to better performance of an algorithm.

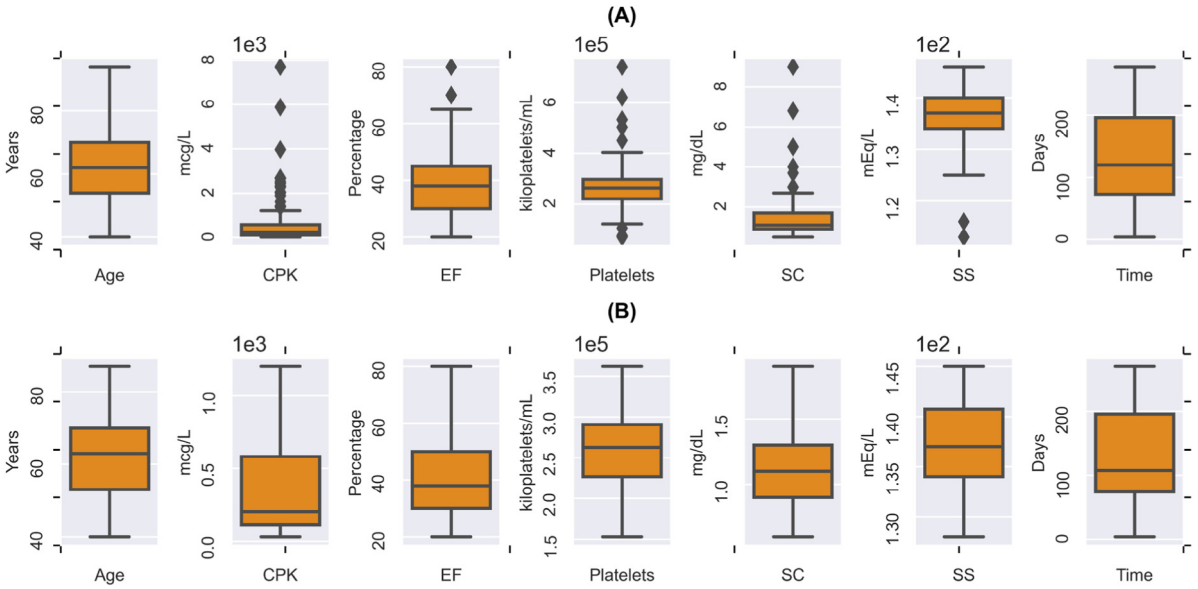


Fig. 3. Numeric data distributions using box plot before and after removing outliers. (A) Before removing outliers and (B) After removing outliers.

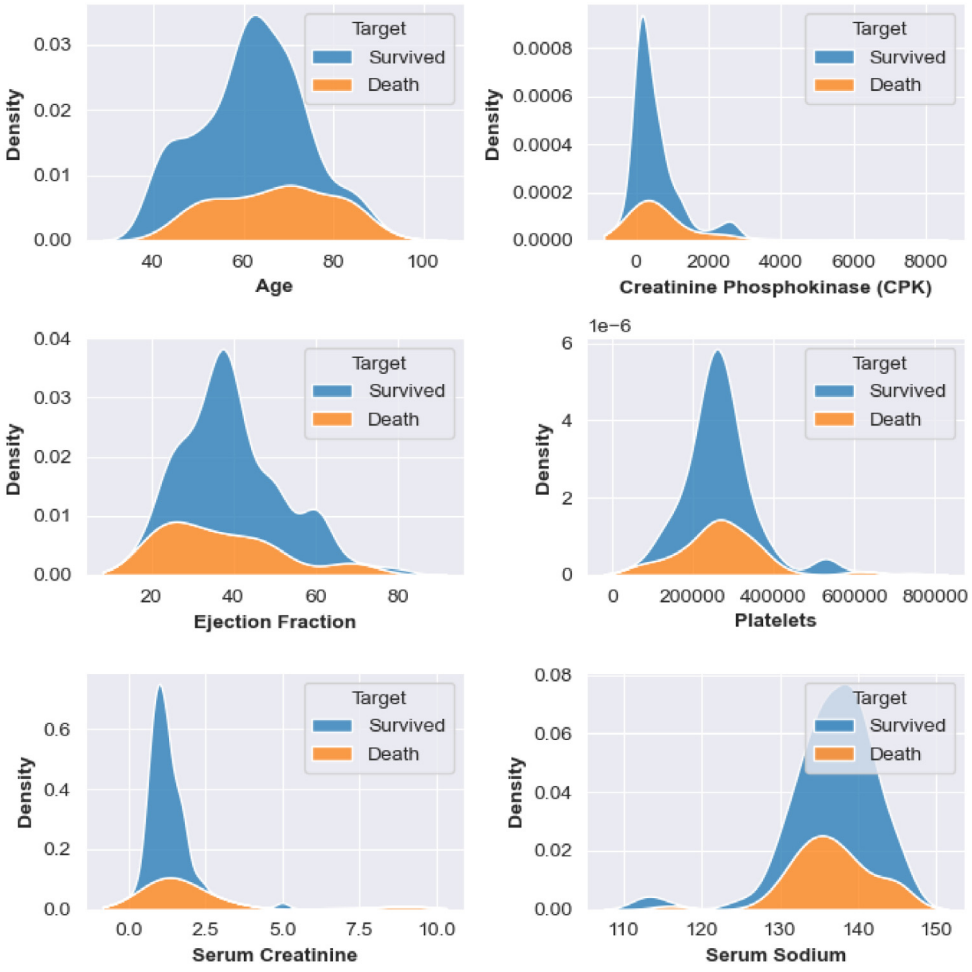


Fig. 4. Representation of continuous probability density curve for all the numeric attributes for both survived and death patients due to HF.

3.2.2. Risk factors analysis for HF death event

Table 5 represents the feature importance values of all the features for each classification algorithm (DTR, XGB, DT, GB, and RF) applied to the HF dataset. The coefficient value describes the contribution to predicting HF survival prediction. This value is the most important criterion to determine the importance of each feature.

Fig. 6 depicts the feature importance of each algorithm applied to the HF dataset. It indicates how much an attribute contributes to

Table 4
Performance comparisons of different ML approaches.

Algorithm	Accuracy (%)	Precision (%)	Recall (%)	F-Measure (%)	Log loss
DTR	85.24	96	96	96	1.151
XGB	94.44	94	94	94	1.919
DT	96.67	96	96	96	1.151
GB	96.67	96	96	96	1.151
RF	97.78	97	97	97	0.767

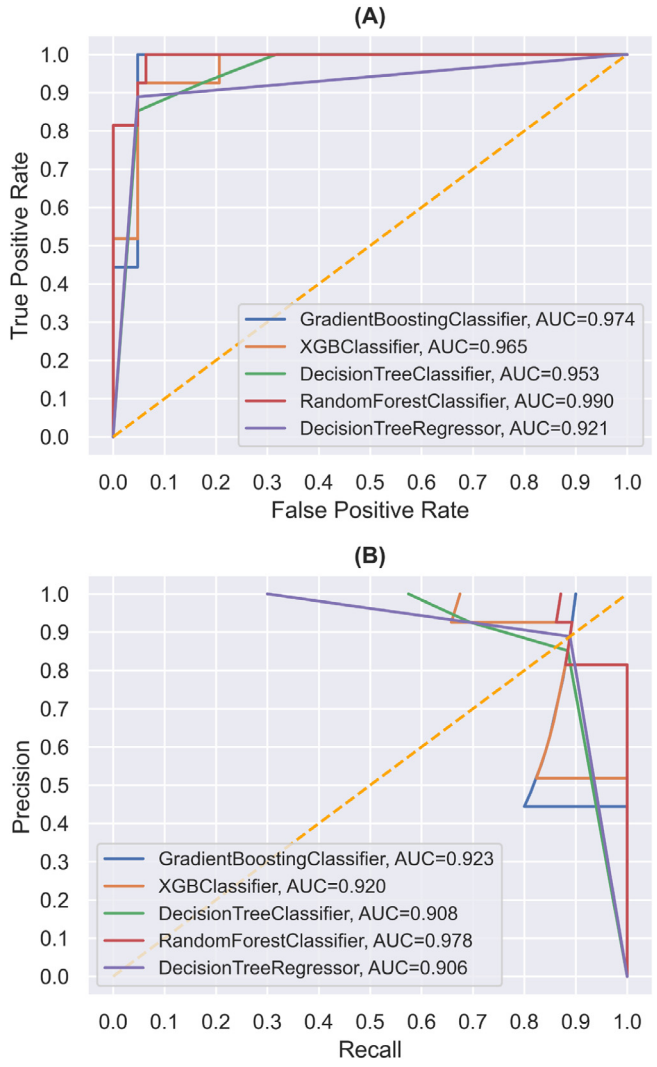


Fig. 5. Area under Receiving Operating Characteristics (ROC) Curve. & Area under Precision–Recall (PRC) Curve (A) ROC Curve, (B) PRC Curve.

Table 5
Coefficient values for each feature for each ML approaches.

Features name	DTR	XGB	DT	GB	RF
SC	0.308270	0.206512	0.251917	0.263667	0.206589
Age	0.136021	0.114136	0.218072	0.184684	0.190757
EF	0.146988	0.159115	0.131726	0.161314	0.143404
Platelets	0.068017	0.049214	0.164056	0.115776	0.110675
CPK	0.218642	0.077942	0.149070	0.115561	0.110963
Smoking	0.047535	0.044533	0	0.021235	0.038716
Sex	0.032697	0.009594	0.030074	0.004935	0.011984
Diabetes	0	0.041988	0	0.007518	0.026773
SS	0.041829	0.058180	0.055085	0.088232	0.116862
Anaemia	0	0.176489	0	0.014350	0.028064
HBP	0	0.062297	0	0.022727	0.015212

Table 6
Top six significant features related to HF death events.

Algorithm	Top six features					
	1st	2nd	3rd	4th	5th	6th
DTR	SC	CPK	EF	Age	Platelets	Smoking
XGB	SC	Age	EF	Platelets	CPK	HBP
DT	SC	Age	Platelets	CPK	EF	SS
GB	SC	Age	EF	Platelets	CPK	SS
RF	SC	Age	EF	Platelets	CPK	SS

Table 7
Result of FST Approach.

Feature Name	CFSSE	GRAE	IGAE	RFAE
Age	0.084	0.070282	0.309342	0.2214
Anaemia	0.0845	0.005337	0.005105	0.0839
CPK	0.0871	0.119346	0.701707	0.2836
Diabetes	0.0811	0.004772	0.004725	0.1134
EF	0.1295	0.085268	0.282563	0.3482
HBP	0.134	0.013036	0.012921	0.0522
Platelets	0.0988	0.122306	0.712986	0.2692
SC	0.0838	0.076617	0.315989	0.2806
SS	0.0936	0.049986	0.200659	0.3003
Sex	0.038	0.001176	0.000361	0.1452
Smoking	0.0223	0.000371	0.001033	0.0789

predicting HF survival and death events. The figure also depicts the risk factors for HF patients according to all the applied classification algorithms.

Table 6 represents the six most significant features, which are responsible for HF survival and death event. These are the most important risk factors for HF patients. SCe, age, and the EF are the most important attributes for HF patients. Platelets, CPK, and SS are also risk factors for HF patients.

Table 7 represents the feature selection technique results. The table demonstrates the attribute weight related to predicting the survival of an HF patient based on four feature selection techniques, CFSSE, GRAE, IGAE, and RFAE. These weighted feature values indicate the impact on the survival of an HF patient. The most important feature is the largest risk factor.

3.3. Result of unsupervised ML

During supervised ML analysis, we identified the six most significant risk factors based on feature importance to predict the survival of an HF patient. These features are employed in the clustering the HF patients into two groups, which is a part of unsupervised ML. Unsupervised ML has been used to verify that the proposed supervised ML model is capable of identifying the death and survival event of HF patients.

Model-based clustering is performed on the most significant features of the HF dataset to create two clusters: survived and dead. Fig. 7(A) illustrates the mean value of each feature of two clusters along with the standard deviation. In addition to that the correlation between the two clusters for the same feature is depicted in Fig. 7(A). The significance is based on p values. The figure depicts that the differences between the two clusters are statistically significant for all these features except for SS. Overall, Fig. 7(A) gives an idea about the distribution of data for two clusters and relation between them,

After applying model-based clustering, PCA is performed on the clustered results to reduce the dimensionality. The result is represented in Fig. 7(B) which shows that the two target groups, survived and dead, are clustered well. A few survived data, marked by the red circle, which are supposed to be in cluster 1 belong to cluster 2. Their characteristics are similar to the dead group. According to the characteristics, these patients are not supposed to survive but they did. It means some external factors may be associated with these patients along with HF, those responsible for survival. Further research should be conducted for these red circled instances. The result indicates that the proposed model

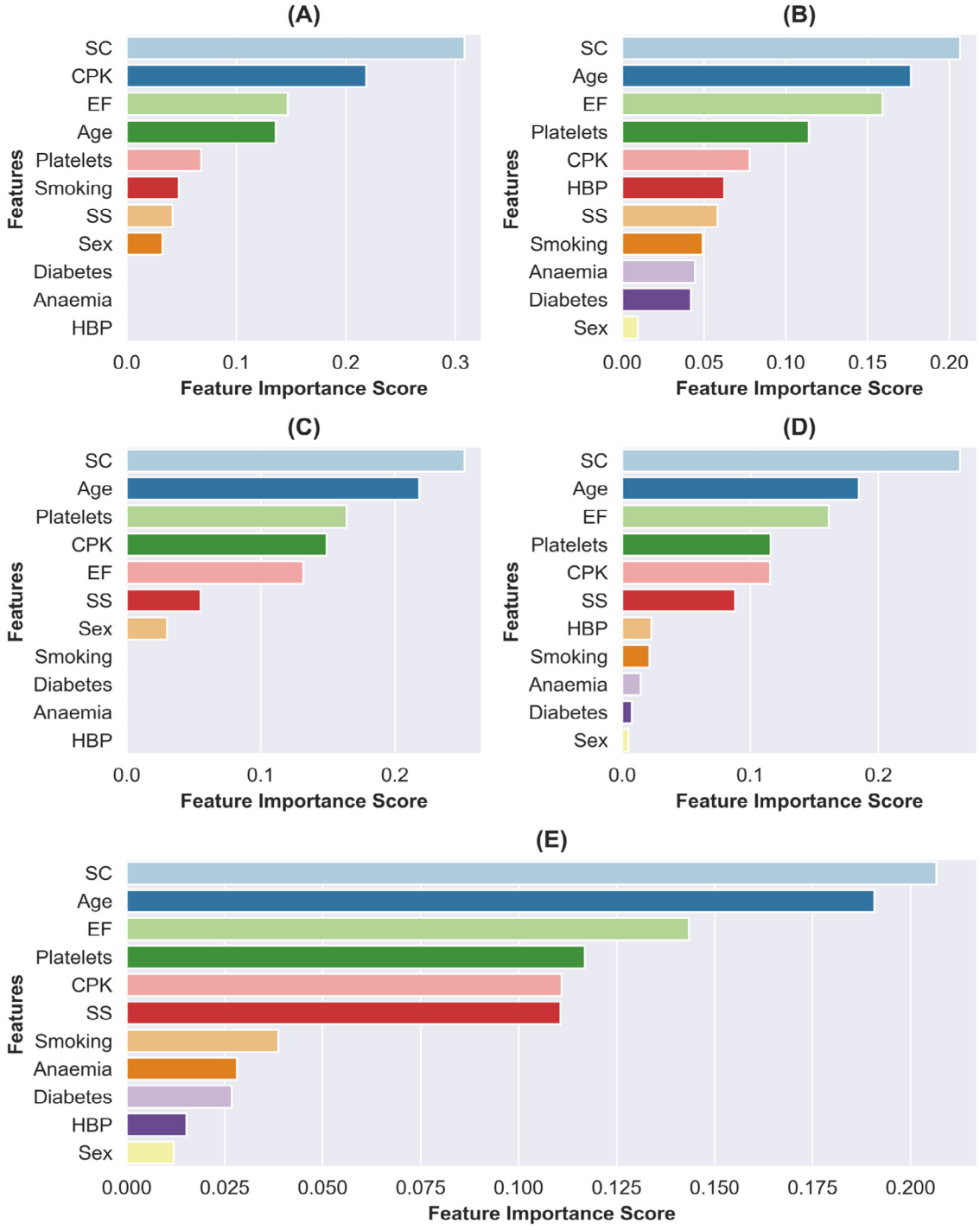


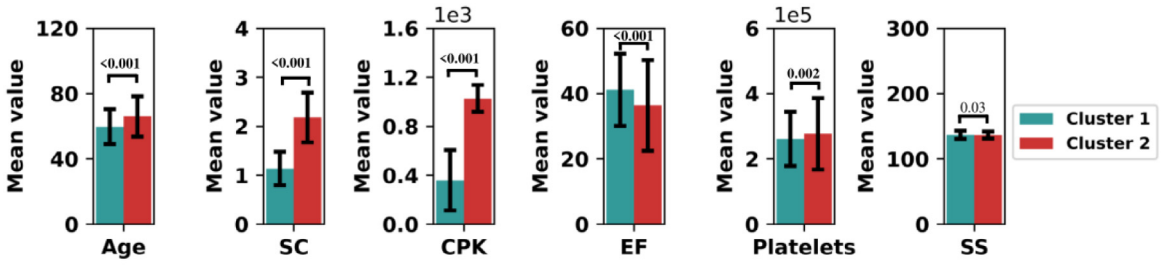
Fig. 6. Feature importance for HF death events (A) DTR Classifier, (b) XGBoost classifier, (C) DT classifier, (D) GB classifier, (E) RF classifier.

has been able to differentiate two groups successfully. Overall result of the unsupervised result depicted in Fig. 7(B) validates the result of supervised ML results performances and found that the proposed model is highly capable for survival and death prediction of HF patients.

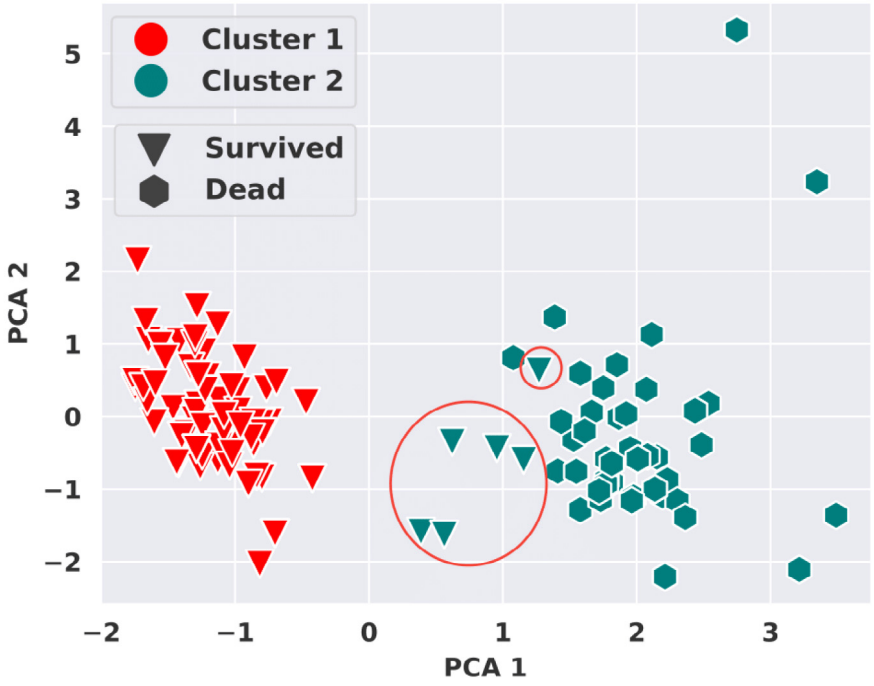
3.4. Result of survival analysis

Table 8 describes the survival analysis results and the impact of the most significant features for HF patients. According to the survival analysis results, age, SC, CPK, and EF are statistically most significant and correlated with a patient's survival after HF. Besides, higher age, SC, and SS are a risk factor, whereas a low EF is a major risk factor for an HF patient. So, it can be said that age, SC, CPK, EF, and SS are the most significant risk factors for HF patients.

Fig. 8(A–G) illustrates the survival probability of male and female HF patients, and of the 40% of patients with the lowest and the 40% with the highest SC, age, SS, CPK, platelets, EF, whereas Fig. 8(H) visualizes the impact of the most significant risk factors on HF patients. Fig. 8(A) shows that the probability of a female patient surviving HF is higher than the probability of a male patient surviving HF. The 40% of HF patients who have the lowest SC, age, or the number of platelets have a higher survival probability compared to the 40% with the highest SC, age, or number of platelets. However, HF patients who belong to the upper 40% group for SS, and EF have a better chance of



(A)



(B)

Fig. 7. (A). Distribution of data and significance between the same features of two clusters based on p value (B). Scatter plot for representing survived and death group based on clustering.

Table 8
The impact of the most significant risk factors based on survival analysis.

Features	HR	Z Score	P Value	−log2(p)
Age	1.09	7.25	<0.005	40.07
SC	1.51	6.53	<0.005	33.86
CPK	1.00	3.58	<0.005	11.51
EF	0.95	−5.41	<0.005	23.91
Platelets	1.00	−0.73	0.46	1.11
SS	1.03	1.60	0.11	3.18

survival than others. The probability of survival changes with follow up times in terms of CPK. In terms of risk factors, Fig. 8(H) indicates that SC is the most important risk factor, for HF patients. The second most important risk factor is SS, while age and EF are the third and fourth important risk factors related to survival for HF patients. This agrees with the results of Chicco and Jurman in 2020 who also found that SC, age, SS and EF are the most important risk factors for HF patients.

4. Conclusion

This study aimed to build a ML model to predict HF survival and identify the most important risk factors. It is found that RF achieved the maximum accuracy (97.78%) along with 0.97 precision, recall, F-Measure and 0.767 log loss. Based on coefficient values for each algorithm, the most important risk factors were identified. The most important attributes for HF patients are SCe, age, EF, platelets and SS. Overall, it can be said that the model is capable of predicting survival events for an HF patient and analyzing the risk factors. Our ML model might be useful in the clinical settings for screening HF patients by the clinicians and experts. The limitations of the study are a small dataset. The dataset is not big enough, which will be solved in our future study. In future, we want to collect a larger dataset and employ the most updated technology to build such models.

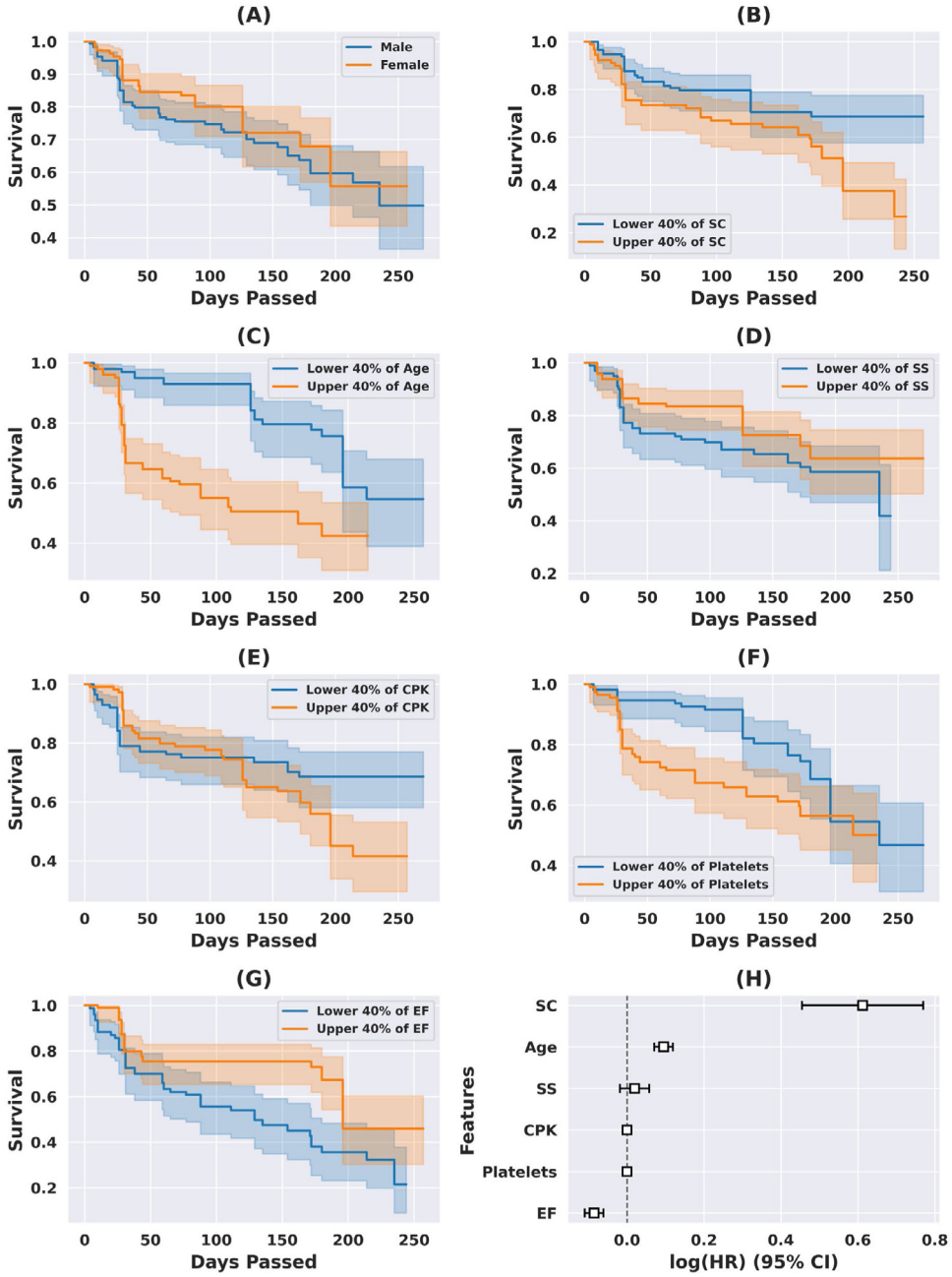


Fig. 8. Visual representation of survival analysis. (A) Survival probability comparison between males and females who are HF patients. (B) Survival probability comparison between two groups of the patients who the lower 40% and upper 40% of SC. (C) Survival probability comparison between two groups of the patients who are the lower 40% and upper 40% of age. (D) Survival probability comparison between two groups of the patients who the lower 40% and upper 40% of SS. (E) Survival probability comparison between two groups of patients who the lower 40% and upper 40% of CPK. (F) Survival probability comparison between two groups of the patients who the lower 40% and upper 40% of platelets. (G) Survival probability comparison between two group of the patients who the lower 40% and upper 40% of EF. (H) The impact of risk factors on survival of an HF patient.

CRediT authorship contribution statement

Md. Mamun Ali: Analyzed the data, Wrote the manuscript. **Vian S. Al-Doori:** Helped perform the experimental analysis with constructive discussions. **Nubogh Mirzah:** Helped perform the experimental analysis with constructive discussions. **Asifa Afsari Hemu:** Analyzed the data, Wrote the manuscript. **Imran Mahmud:** Helped perform the experimental analysis with constructive discussions, Reviewed the work. **Sami Azam:** Reviewed the work. **Kusay Faisal Al-tabatabaie:** Helped perform the experimental analysis with constructive discussions. **Kawsar Ahmed:** Provided the idea, Designed the experiments, Analyzed the data, Wrote the manuscript, Reviewed the work. **Francis M. Bui:** Provided the idea, Designed the experiments, Helped perform the experimental analysis with constructive discussions, Reviewed the work. **Mohammad Ali Moni:** Provided the idea, Designed the experiments, Reviewed the work.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

This work was supported in part by funding from the Natural Sciences and Engineering Research Council of Canada (NSERC).

Data availability

The data that has been used is confidential.

References

[1] D. Chicco, G. Jurman, ML can predict survival of patients with heart failure from SCe and EF alone, BMC Med. Inf. Decis. Mak. 20 (1) (2020) 16.
[2] World Heart Day, World Health Organization, 2021, Available at: https://www.who.int/cardiovascular_diseases/world-heart-day/en/ [Accessed on 18-04-2021].
[3] NHLBI, What is heart failure?, 2022, Available at: https://www.nhlbi.nih.gov/health-topics/heart-failure [Accessed on 18-04-2022].
[4] P.C. Austin, J.V. Tu, J.E. Ho, D. Levy, D.S. Lee, Using methods from the data-mining and machine-learning literature for disease classification and prediction: a case study examining classification of heart failure subtypes, J. Clin. Epidemiol. 66 (4) (2013) 398–407.
[5] S.A. Hunt, W.T. Abraham, M.H. Chin, A.M. Feldman, G.S. Francis, T.G. Ganiats, M. Jessup, M.A. Konstam, D.M. Mancini, K. Michl, J.A. Oates, 2009 Focused update incorporated into the ACC/AHA 2005 guidelines for the diagnosis and management of heart failure in adults: a report of the American college of cardiology foundation/American heart association task force on practice guidelines developed in collaboration with the international society for heart and lung transplantation, J. Am. Coll. Cardiol. 53 (15) (2009) e1–e90.
[6] D.S. Lee, P. Gona, R.S. Vasan, M.G. Larson, E.J. Benjamin, T.J. Wang, J.V. Tu, D. Levy, Relation of disease etiology and risk factors to heart failure with preserved or reduced ejection fraction: insights from the national heart, lung, and blood institute's framingham heart study, Circulation 119 (24) (2009) 3070.
[7] F.A. Masoudi, E.P. Havranek, G. Smith, R.H. Fish, J.F. Steiner, D.L. Ordin, H.M. Krumholz, Gender, age, and hear failure with preserved left ventricular systolic function, J. Am. Coll. Cardiol. 41 (2) (2003) 217–223.
[8] S.J. Pocock, D. Wang, M.A. Pfeffer, S. Yusuf, J.J. McMurray, K.B. Swedberg, J. Ostergren, E.L. Michelson, K.S. Pieper, C.B. Granger, Predictors of mortality and morbidity in patients with chronic heart failure, Eur. Heart J. 27 (1) (2006) 65–75.
[9] S.J. Pocock, C.A. Ariti, J.J. McMurray, A. Maggioni, L. Køber, I.B. Squire, K. Swedberg, J. Dobson, K.K. Poppe, G.A. Whalley, R.N. Doughty, Predicting survival in heart failure: a risk score based on 39 372 patients from 30 studies, Eur. Heart J. 34 (19) (2013) 1404–1413.
[10] D.G. Kleinbaum, M. Klein, Kaplan–meier survival curves and the log-rank test, in: Survival Analysis, Springer, New York, NY, 2012, pp. 55–96.
[11] T. Ahmad, A. Munir, S.H. Bhatti, M. Aftab, M.A. Raza, Survival analysis of heart failure patients: A case study, PLoS One 12 (7) (2017) e0181001.
[12] F.M. Zahid, S. Ramzan, S. Faisal, I. Hussain, Gender based survival prediction models for heart failure patients: a case study in Pakistan, PLoS One 14 (2) (2019) e0210602.
[13] D. Wussler, E. Michou, M. Belkin, N. Kozhuharov, M. Diebold, M.D. Gualandro, T. Breidthardt, C. Mueller, Mortality prediction in acute heart failure: scores or biomarkers? Swiss Med. Wkly. (2020).
[14] S.Y. Cho, S.H. Kim, S.H. Kang, K.J. Lee, D. Choi, S. Kang, S.J. Park, T. Kim, C.H. Yoon, T.J. Youn, I.H. Chae, Pre-existing and ML-based models for cardiovascular risk prediction, Sci. Rep. 11 (1) (2021) 8886.
[15] LARXEL, Heart failure prediction, 2020, Available at: https://www.kaggle.com/andrewmvd/heart-failure-clinical-data [Accessed on 18-04-2021].
[16] P. Ghosh, S. Azam, M. Jonkman, A. Karim, F.M. Shamrat, E. Ignatious, S. Shultana, A.R. Beeravolu, F. De Boer, Efficient prediction of cardiovascular disease using ML algorithms with relief and lasso feature selection techniques, IEEE Access 9 (2021) 19304–19326, http://dx.doi.org/10.1109/access.2021.3053759.
[17] A.C.D. Silveira, Á. Sobrinho, L.D.D. Silva, E.D.B. Costa, M.E. Pinheiro, A. Perkusich, Exploring early prediction of chronic kidney disease using ML algorithms for small and imbalanced datasets, Appl. Sci. 12 (7) (2022) 3673.
[18] W.J. Dixon, F.J. Massey Jr., Introduction to statistical analysis, 1951.
[19] J.W. Tukey, EDA, Vol. 2, 1977, pp. 131–160.
[20] C.H. Yu, EDA in the context of data mining and resampling, Int. J. Psychol. Res. 3 (1) (2010) 9–22.
[21] J.R. Quinlan, Induction of DTs, Mach. Learn. (1986).
[22] J.A. Cruz, D.S. Wishart, Applications of ML in cancer prediction and prognosis, Cancer Inform. 2 (2006) 117693510600200030.
[23] M.M. Raihan, E. Ahmed, A. Karim, S. Azam, M. Raihan, L. Akter, M.M. Hassan, Chronic renal disease prediction using clinical data and different ML techniques, in: 2021 2nd International Informatics and Software Engineering Conference, IISEC, 2021, http://dx.doi.org/10.1109/iisec54230.2021.9672365.
[24] S. Uddin, A. Khan, M.E. Hossain, M.A. Moni, Comparing different supervised ML algorithms for disease prediction, BMC Med. Inf. Decis. Mak. 19 (1) (2019) 1–16.
[25] M.M. Ahamad, S. Aktar, M. Rashed-Al-Mahfuz, S. Uddin, P. Liò, H. Xu, M.A. Summers, J.M. Quinn, M.A. Moni, A ML model to identify early stage symptoms of SARS-cov-2 infected patients, Expert Syst. Appl. 160 (2020) 113661.
[26] Read the Docs, XGBoost documentation, 2022, Available at: https://xgboost.readthedocs.io/en/latest/ [Accessed on 21-04-2021].
[27] C. Bouveyron, C. Brunet-Saumard, Model-based clustering of high-dimensional data: A review, Comput. Statist. Data Anal. 71 (2014) 52–78.
[28] M.M. Ali, B.K. Paul, K. Ahmed, F.M. Bui, J.M. Quinn, M.A. Moni, Heart disease prediction using supervised ML algorithms: Performance analysis and comparison, Comput. Biol. Med. 136 (2021) 104672.
[29] M.M. Ali, K. Ahmed, F.M. Bui, B.K. Paul, S.M. Ibrahim, J.M. Quinn, M.A. Moni, ML-based statistical analysis for early stage detection of cervical cancer, Comput. Biol. Med. 139 (2021) 104985.
[30] A.A. Hemu, R.B. Mim, M.M. Ali, M. Nayer, K. Ahmed, F.M. Bui, Identification of significant risk factors and impact for ASD prediction among children using ML approach, in: 2022 Second International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies, ICAECT, IEEE, 2022, pp. 1–6.
[31] M.S. Hossain, M.H. Shovo, M.M. Ali, M. Nayer, K. Ahmed, F.M. Bui, December. ML_Sps: Stroke prediction system employing ML approach, in: Artificial Intelligence and Data Science: First International Conference, ICAIDS 2021, Hyderabad, India, December (2021) 17–18, Revised Selected Papers, Springer Nature Switzerland, Cham, 2022, pp. 215–226.
[32] K. Ahmed, T. Jesmin, M.Z. Rahman, Early prevention and detection of skin cancer risk using data mining, Int. J. Comput. Appl. 62 (2013) 4.
[33] A. Banerjee, H. Shan, Model-based clustering, in: C. Sammut, G.I. Webb (Eds.), Encyclopedia of ML, Springer, Boston, MA, 2011, http://dx.doi.org/10.1007/978-0-387-30164-8_554.
[34] G. Gan, C. Ma, J. Wu, Data clustering: theory, algorithms, and applications, Soc. Ind. Appl. Math. (2020).
[35] S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemometr. Intell. Lab. Syst. 2 (1–3) (1987) 37–52.
[36] H. Wang, G. Ni, J. Chen, J. Qu, Research on rolling bearing state health monitoring and life prediction based on PCA and internet of things with multi-sensor, Measurement 157 (2020) 107657.
[37] M.J. Hossain, U.N. Chowdhury, M.B. Islam, S. Uddin, M.B. Ahmed, J.M. Quinn, M.A. Moni, ML and network-based models to identify genetic risk factors to the progression and survival of colorectal cancer, Comput. Biol. Med. (2021) 104539.
[38] M.A. Hossain, S.M.S. Islam, J.M. Quinn, F. Huq, M.A. Moni, ML and bioinformatics models to identify gene expression patterns of ovarian cancer associated with disease progression and mortality, J. Biomed. Inform. 100 (2019) 103313.
[39] F.M. Shamrat, S. Azam, A. Karim, R. Islam, Z. Tasnim, P. Ghosh, F. De Boer, Lungnet22: A fine-tuned model for multiclass classification and prediction of lung disease using X-ray images, J. Pers. Med. 12 (2022) 680, http://dx.doi.org/10.3390/jpm12050680.
[40] M.R. Rahman, T. Islam, M.A. Al-Mamun, T. Zaman, M.R. Karim, M.A. Moni, The influence of depression on ovarian cancer: Discovering molecular pathways that identify novel biomarkers and therapeutic targets, Inform. Med. Unlocked 16 (2019) 100207.
[41] W. Kim, J.J. Park, H.Y. Lee, K.H. Kim, B.S. Yoo, S.M. Kang, S.H. Baek, E.S. Jeon, J.J. Kim, M.C. Cho, S.C. Chae, Predicting survival in heart failure: a risk score based on machine-learning and change point algorithm, Clin. Res. Cardiol. 110 (8) (2021) 1321–1333.
[42] H.K. Rana, M.R. Akhtar, M.B. Islam, M.B. Ahmed, P. Lió, F. Huq, J.M. Quinn, M.A. Moni, ML and bioinformatics models to identify pathways that mediate influences of welding fumes on cancer progression, Sci. Rep. 10 (1) (2020) 1–15.