Article

Cardiac Diagnostic Feature and Demographic Identification Models : A Futuristic Approach for Smart Healthcare using Machine Learning

Deepak Kumar [1], Chaman Verma [2,*], Sanjay Dahiya [3], Pradeep Kumar Singh [4,*], Maria Simona Raboaca [5,6,7,8*]

[1] Department of Computer Science and Applications, Guru Kashi University, 151302, Talwandi Sabo, Punjab, India; Dr.d.k.mehta81@gmail.com
[2] Department of Media and Educational Informatics, Faculty of Informatics, Eötvös Loránd University, 1053, Budapest, Hungary; chaman@inf.elte.hu
[3] Department of Computer Science and Engineering, Ch. Devi Lal State Institute of Engineering Technology, 125077, Sirsa, Haryana, India; sanjaydahiyakkr@gmail.com
[4] Department of Computer Science and Engineering, ABES Engineering College, 201009, Ghaziabad, Uttar Pradesh, India; pradeep_84cs@yahoo.com
[5] ICSI Energy, National Research and Development Institute for Cryogenic and Isotopic Technologies, 240050, Ramnicu Valcea, Romania
[6] Faculty of Electrical Engineering and Computer Science, "Stefan cel Mare" University of Suceava, 720229, Suceava, Romania
[7] Technical University of Cluj-Napoca, 400114 Cluj-Napoca, Romania
[8] Doctoral School Polytechnic University of Bucharest, Romania;simona.raboaca@icsi.ro
* Correspondence :Dr.d.k.mehta81@gmail.com (D.K.), chaman@inf.elte.hu (C.V.), pradeep_84cs@yahoo.com (P.K.S.), simona.raboaca@icsi.ro (M.S.R.)

Abstract: Around the world, every year, about 17 million people death cause happen due to CardioVascular Diseases (CVD). As per clinical records, primarily sufferers exhibit myocardial infarctions and Heart Failures (HF). Creatinine is a Musculo - skeletal waste product. The kidneys filter creatinine from the blood and excrete it through the urine in a healthy body. High creatinine levels can suggest renal problems. Elevated Serum Creatinine (SC) has been well established in the HF. Patients' electronic medical records can be used to quantify symptoms and other related clinical laboratory test values, which would then be utilized to direct biostatistics exploration to uncover patterns and associations that doctors would otherwise miss. The latest American Heart Association guidelines for 1500 mg/d sodium tend to be sufficiently relevant for patients with stage A and B with HF. In this article, we used a dataset of the year 2015 of heart patients records of 299 patients. The present paper used the data analytic and statistical tools to verify the significant differences between alive and dead patients' SC and Serum Sodium (SS). It also demonstrates the impact of significant features on abnormal SC and SS on the Survival-Status levels. The Age-Group feature, which is derived from age attribute and, Ejection Fraction (EF), anemia, platelets, Creatinine Phosphokinase (CPK), Blood-Pressure (BP), gender, diabetes, and smoking-status were utilized to determine the potential contributing features to mortality with Cox regression model. The Kaplan Meier plot was used to investigate the overall pattern of survival concerning age-group. During pre-processing of the dataset, Age and SS were removed due to multicollinear features during performing machine learning algorithms experiments. This paper also predicted patients' survival, age group, and gender using supervised machine learning classifiers. Detection of significant features would help in making informed decisions to balance the lifestyle of heart patients. The author revealed that the patient's follow-up months, as well as SC, EF, CPK, and platelets, are sufficient key features to predict heart patient survival using Random Forest (RF) stratified 10-fold CV method with accuracy (96%) with 5% Standard Deviation (SD) from medical records dataset. We identified the age-group and gender of the patient, and the RF model outperformed others with the best accuracy 96% and 94% in both cases having 11% SD. Also, prominent features such as CPK, SC, follow-up month, platelets, and ejection were found to be significant factors in predicting the patient's age-group. Smoking habits, CPK, platelets, follow-up month, and SC of each patient were discovered to be significant predictors of patient gender. The hypothetical study proved that SC and SS making substantial differences in the survival of patients ($p < 0.05$) and failed to reject that anemia, diabetes, and BP making a significant impact on the creatinine and sodium of each patient ($p > 0.05$). With $\chi^2(1) = 8.565$, the Kaplan Meier plot revealed that mortality was high in the extremely elder age-group. The finding has possible effects on clinical practice and becomes a new medical support system when predicting whether a patient can survive a heart attack or not. The doctor should primarily concentrate on follow-up month, SC and EF, CPK, and platelet count since the aim is to understand whether a patient survives after HF.

Keywords: Serum Creatinine; Serum Sodium; Ejection Fraction; Creatinine Phosphokinase; Multicollinearity; Matthew Correlation coefficient.

1. Introduction

HF is a condition in which the muscles in the heart wall deteriorate and swell, limiting the heart's ability to pump blood. The heart's ventricles can become rigid and stop filling correctly between beats. With time, the heart fails to meet the good demand for blood in the body, and the individual begins to have trouble breathing. Coronary heart disease, diabetes, high BP, and other illnesses such as HIV, alcoholism or drug abuse, thyroid disease, an excess of vitamin E in the body, radiation or chemotherapy, among others, are the most common causes of HF [1,2]. Occasionally, the symptoms of CVD differ by gender. For example, a male patient is more likely to experience chest pain. In contrast, a female patient is more likely to experience other symptoms associated with chest pain, such as nausea, excessive exhaustion, and shortness of breath [3]. Researchers have been experimenting with a range of strategies to predict Heart Disease (HD), their age-groups and the gender of HD patients. Still, prediction is challenging at any stage due to a variety of variables, including but not limited to difficulty, execution time, and approach precision [4]. Early detection can also help avoid HD, which can lead to death. Angiography is the most precise and effective tool for predicting cardiac artery disease [5], but it is costly, making it out of reach for low-income families. When extracting valuable information from large amounts of data, data mining plays a critical role. It's used in nearly every field of life, including medicine, engineering, industry, and education. Data mining is a technique for examining data to retrieve important decision-making information that has been concealed in the past repository. Several machine learning algorithms have been used to understand the complexity and non-linear interaction between various variables by reducing prediction and natural results. We need to use machine learning algorithms to assist medical healthcare practitioners in analyzing data and make accurate and precise diagnostic decisions due to the ever-increasing amount of medical data. Different classification algorithms are used in medical data mining to estimate CVD in patients and death predictions due to heart attacks [6]. In older people, age plays a vital role in the degradation of cardiovascular function, which leads to an increased risk of CVD [7,8]. The American health association recorded CVD in men and women in the United States were 40 to 59 years old and also found 75% and 86% in 60 to 79 and over 80 years old in men respectively [9].

In this study, we examine a dataset of medical records from patients with HF that were published in July 2017 and 2020 by Ahmad T. et.al [14] and Chico D. [31] . From the medical records of 299 Pakistani patients with HF, Ahmad T. used conventional biostatistics time-dependent models (Cox regression) to predict mortality and identify key features. Following that, Chico D. used the same dataset to predict Survival-Status various state-of-the-art machine learning algorithms. We hope to close this gap by employing a variety of data mining approaches to first estimate patient's gender,age-group and Survival-Status with statistical methods to determine the impact of SC and SS on various health-related problems like diabetes, anemia, High BP level and also confirm the effect on the Survival-Status levels. Eight machines learning models Table5 were used: Decision Tree (DT), Logistic Regression (LR), and RF, Gaussian Naïve Bayes (GNB), Gradient boosting Machines (GBM), Support Vector Machine (SVM),k-Nearest Neighbour (k-NN), and Extreme Gradient boosting (XGB). The HD study literature motivates us to do a study on the following points on the target dataset:

- Motivated to develop a decision-making method that accurately predicts the age-group (Adult and Very Old) and Gender (Woman/Man) and Survival-Status of cardiac patients.
- 10-fold cross-validation technique can be used for extracting out best-performing predictors.
- Significant features from the dataset that influence the machine learning algorithm's output can be identified to examine the critical risk factors.

2. Literature Review and Objectives, Hypotheses Development

2.1. Effect of Abnormal SC and SS on HF Patient

As we know that healthcare professional can assess the quantity of sodium in our blood with a sodium blood test (also known as a serum sodium test). SS is routinely measured in order to examine electrolyte, acid-base, and water balance, as well as renal function. If the patient is not in renal failure or has severe hyperglycemia, sodium accounts for approximately 95% of the osmotically active chemicals in the extracellular compartment.The ideal SS reference range is 135-147 mmol/L.

Since the widely accepted idea that increased sodium consumption contributes to increased fluid retention during cardiac failure, a sodium-free diet as proposed for the general public is expected to boost the results of HF sufferers. There is more evidence present which linking to sodium intake with BP [10], the occurrence of hypertension, [11], CVD, [12] other HF risk factors, even the latest American Heart Association guidelines for 1500 mg/d sodium tend to be sufficiently relevant for patients with stage A and B HF. Yash Patel et.al [13] have found in their review study that due to reduced renal perfusion, sodium, and water reabsorption from renal tubules, HF is characterized by sympathetic system activation and renin-angiotensin-aldosterone system (RAAS) activation. The activation of the antidiuretic and anti-natriuretic systems has been linked to a sodium-restricted diet in HF patients. Latest Cochrane reviews from 185 clinical trials of low vs. high-sodium diets found a statistically significant rise relative to high intake groupings of sodium in the range of renin, aldosterone, noradrenaline, adrenaline, cholesterol, and triglycerides. Tanvir Ahmad et al. [14] studied and focused on heart failure patient's survival and used cox regression model. They found 32% mortality rate due to CVD and found EF, age, creatinine (creatinine > 1.5 renal dysfunction), sodium, anemia, and BP as significant mortality rates. Smoking and diabetes were not found necessary toward HF death. Sagar B. Dugani et al. [15] examined a coronary HD dataset w.r.t gender to see whether risk profiles differed by age at the start and diabetes, lipoprotein insulin resistance had the highest relative risk of more than 50 clinical and biomarker risk factors. In the women, diabetes had the highest aHR of any clinical factor. Mohammed W Akhtar et. al [16] finding was on renal insufficiency (abnormal SC) in HF patients and found patients had a five-fold increased risk of death after being discharged from the hospital. Several studies have generally identified and supported the close connection between hypertension and dietary sodium intake. Reducing dietary sodium not only reduces BP and hypertension but also leads to decreased cardiovascular risk. Regardless of sex or ethnic group, both hypertensive and normotensive people experience a significant drop in BP after a sustained slight reduction in salt intake, with more significant decreases in systolic BP for more considerable reductions in dietary salt. Water accumulation, increased systemic peripheral resistance, changes in endothelial function, structural changes and function of broad elastic arteries, changes in sympathetic activity, and autonomic neural regulation of the cardiovascular system are all linked to a high sodium intake and an increase in BP [17]. Also, in its pathophysiological context and clinical implications, the subject of salt-sensitivity, which refers to individual susceptibility concerning BP variations following changes in dietary salt intake [18]. Blood sugar levels in people with diabetes can rise to dangerously high levels, causing health problems such as kidney disease [19].

The study by Abede Tamrat et al. [20] included 370 patients. Anemia was shown to be expected in 41.90 % of the study cohorts, with most of the participants being female (64.59 %). Between anemic and non-anemic patients, there was a substantial difference in hemoglobin, creatinine, and salt levels. Angiotensin-converting enzyme inhibitors were used less commonly by anemic patients with HF.

- **Objective 1**: To explore an impact of SC and SS on Survival-Status level of the patient.
  1. H_{01}: *No significant difference between Alive and Dead towards SC and SS.*
- **Objective 2**: To explore an impact of SC and SS on anemia, diabetes, and High BP levels of the patients.
  1. H_{02a}: *No significant difference between non-anemic and anemic levels towards SC and SS.*
  2. H_{02b}: *No significant difference between non-diabetic and diabetic levels towards SC and SS.*
  3. H_{02c}: *No significant difference between Normal BP and High BP towards SC and SS.*

2.2. Smoking Habits exists among Gender specific HF Patients

Ambuj Roy et. al [21] researched atherosclerotic CVD and found it caused by frequent smoking habits. They found a relation between various types of tobacco use and cardiovascular manifestation was high, and magnitude was significant. In younger people who smoke more cigarettes a day, between women and men, and in some ethnic groups, such as the South Asians, the risk seems to be higher. Huxley RR et. al [22] examined the sexual disparity of smokers in the risk of coronary HD and underlined uncertainty in gender-specific smoking habits towards HD.

- **Objective 3**: To explore the association of the gender and smoking habit of the patient.
  1. H_{03}: *No significant association between gender with smoking habits.*

2.3. HF Patients among Particular Age-Group

Majidur Rahman et.al [23] research mainly focused on detecting a specific age-group based on cancer diagnosis and another related factor. They found the highest accuracy (59.09%) with Artificial Neural Network (ANN) among proposed classifiers (LR, SVM, ANN). The 1st degree relative, GERD, alcohol abuse, T. Cell, Industrial hazard, estrogen exposure, and papillomavirus were significant features using recursive feature elimination technique for identifying cancer patients' age-group. Adam S. Vaughan et. al [24] examined recent developments in the age-group mortality of the cardiac disease. The authors used the data from 3098 US countries on HD mortality from 2010 to 2015 using the Baysian statistic model in four age-group (35-44, 45-54, 55-64, and 65-74 years old). They concluded that mortality of HD decreased in all age-groups except between 55-64 years of age. The observed rise in cardiovascular mortality at the local and age-groups represents a problem in people, societies and the country in their entirety following over 40 years of declines at the national and county levels. Sagar B. Dugani et. al [15] studies on coronary HD dataset and researched on whether risk profiles vary by age at the beginning. Authors used 28024 women dataset and found diabetes and lipoprotein insulin resistance had a relative risk of over 50 clinical and biomarker risk factors. Diabetes had the highest aHR of any clinical factor in the women, ranging from 10.71 (95 percent CI, 5.57-20.60) at CHD onset in those younger than 55 years to 3.47 (95 percent CI, 2.47-4.87) at Cardio HD onset in those 75 years or older. Metabolic syndrome (aHR, 6.09; 95 percent CI, 3.60-10.29), hypertension (aHR, 4.58; 95 percent CI, 2.76-7.60), obesity (aHR, 4.33; 95 percent CI, 2.31-8.11), and smoking (aHR, 4.33; 95 percent CI, 2.31-8.11) were all found to be risk factors for CHD onset in people under 55.

- **Objective 4**: To explore the impact of age-group on the survival-Status levels(censored/Dead) of the patient.

  1. H_{04}: *No significant association between Age-Group levels and Survival-Status levels.*

- **Objective 5**: To predict the Age-group of HF patient based on significant features

2.4. HF among Particular Gender

Steve Horvath et. al [25] researched seven different racial/ethnic groups by using their blood, saliva, and brain samples tested. Researchers examined blood's intrinsic epigenetic aging (regardless of the number of blood cells) and extrinsic epigenetic aging rates w.r.t blood cell counts and by tracking the immune system age). Their findings were sex, race/ethnicity, and Cardio HD (CHD) risk factors to a lesser degree with epigenetic aging rates, but not to incident CHD outcomes. Hispanics, older African-Americans, and women have lower mortality rates than predicted. Jennifer L. Rodgers et. al [26] research findings were on gender and aging-related to HF risks. They found that CVD was more common in the elderly and those above the age of 65. In adults, age was an independent risk factor for CVD, but other factors such as frailty, obesity, and diabetes compound these risks. They also found gender-related findings that older females were confirmed to have a higher risk of CVD than age-matched men. The chances of CVD rise with age in both men and women, which corresponds to a general decrease in sex hormones, especially estrogen and testosterone. Benjamin et. al [27] found that several variations in CVD risk factors were associated with sex in aged adults. They found that although age was a risk factor for CVD in both men and women, it was clear that older women were more susceptible to some HD-related complications. Villa et. al [28] examination on HD patients that women were generally safe from CVD before menopause, but their risk increases dramatically after menopause. In both men and women, the reduction in sex hormones had been shown to play an essential role in developing CVD with the onset of advanced age. Edward Korot et. al [29] used the Auto ML model with images of the retinal fundus and predicted the gender from the UK Biobank dataset with accuracy (88.8%).

- **Objective 6**: To predict the gender of HF patient based on significant features

2.5. HF Patients' Survival-Status Identification

Jian Ping Li et. al [30] designed an HD identification method using supervised ML classification algorithms Table 5 and the accuracy of the system was drawn 92.37% with SVM. Cleveland dataset was used with LOSSCV hyperparameters tuning, and FCMIM was identified as the best feature selector. In another study, Davide Chicco et. al [31] predicted patient's survival binary classification problem with many contemporary classification methods 5 and extracted out two most relevant features (EF and SC) with Matthew Correlation Coefficient(MCC) (61.6%) and accuracy (83.8%) on HF dataset. ABID ISHAQ et. al [32] also designed and developed patient's survival classification but with Synthetic Minority Over sampling (SMOTE) technique due to the existence of imbalance in the target variable. Authors found improved accuracy (92.62%) with ETC classifier and RF feature selector algorithm used on same HF dataset.

- **Objective 7**: To predict the HF patient's Survival-Status levels based on the significant features

3. Problem Statement

For a better understanding, the limitations and significance of the proposed HD diagnosis approaches have been summarised in Table 1. Many of these existing approaches used various techniques to detect HD and other related demographics early. However, all of these methods have a low prediction accuracy, precision, and recall rate for predictions. According to Table 1, the HD and other mentioned identification method's prediction performance metrics needs to be improved for more efficient and reliable early detection for better care and recovery. The study also found women are more vulnerable to CVD and SC found the most significant feature in every mentioned classification. Therefore, it is imperative to study SC, SS, and gender-related features and found the significant impact of smoking, BP, and diabetes. The Cox regression method for survival analysis may be employed to examine the effect of various existing demographic dataset features on time-specific occurrence. New methods for accurate identification of CVD and other features from given dataset are needed to address these issues. Prediction accuracy without using SMOTE and further related performance enhancement is a significant challenge and gap of the research.

Table 1. Previous research Versus Extant research.

Ref.	Tech.	DV	FS Algo.	CV-HP	Multcol.
[30]	LR, K-NN, ANN, SVM (RBF), SVM(Linear), NB, DT, LOSOCV, Feature Selection	HD (Present/Absent)	Relief, MRMR, LASSO, LLBFS, FCMIM	LOSO	X
[16]	T-Test, Fisher Exact Test	SC, Renal Insufficiency	×	×	×
[14]	COX Regression Model	EF levels (EF 45) with Survival (Time to event)	Statistical Analysis	X	X
[31]	RF, DT, GBM, LR, ANN, NB, SVM (RBF), SVM (Linear), KNN, MCC	Survival Check (Survived / Dead)	RF	Grid Search	×
[34]	SMOTE, DT, Ada-boost, LR, SGD, RF, GBM, ETC, NB, SVM	Survival Check (Survived / Dead)	RF	×	×
[32]	Stratified cox proportional Hazard regression model	4 Age-Group examined with Survival Analysis (time-to-event)	Statistical Analysis	×	×
[29]	Code free deep learning model	Gender	×	×	×
Present	Mann-Whitney *U*-Test, χ^2 test, Cox Regression, DT, LR, GBM, GNB, RF, SVM (RBF), KNN, XGB, VIF, MCC	Survival-Status (Alive/Dead), Age-Group, Gender, SC and SS	RF, χ^2, XGB	Grid Search	√

Source: Own elaboration.

4. Research Organization

The remainder of the paper is structured as follows: Section 5 elaborates the state-of-art- research design and methodology. Section 6 explains the machine learning Models with Hyperparameter Tuning. Section 7 focuses on the basics of applied machine learning algorithms. Section 8 discusses the feature ranking and selection mechanism of machine learning algorithms. Section 9 debates the on the various performance evaluation metrics. Section 10 reflects the results of seven experiments. Section 11 discusses the findings of exhibited in Section 10 with the existing literature. Section 12 concludes the real crux of the extant research with the future proceeding.

5. Materials and Methods

5.1. Design and Contribution

A hybrid-python-based application has been designed and created for automating the inferential, differential, and predictive analysis of the patterns of heart disease data. The present model called the Cardiac Diagnostic Feature, and Demographic Identification (CDF-DI) can be used as a cardiac diagnostic aid. Figure 1 depicting the pictorial view of the CDF-DI.

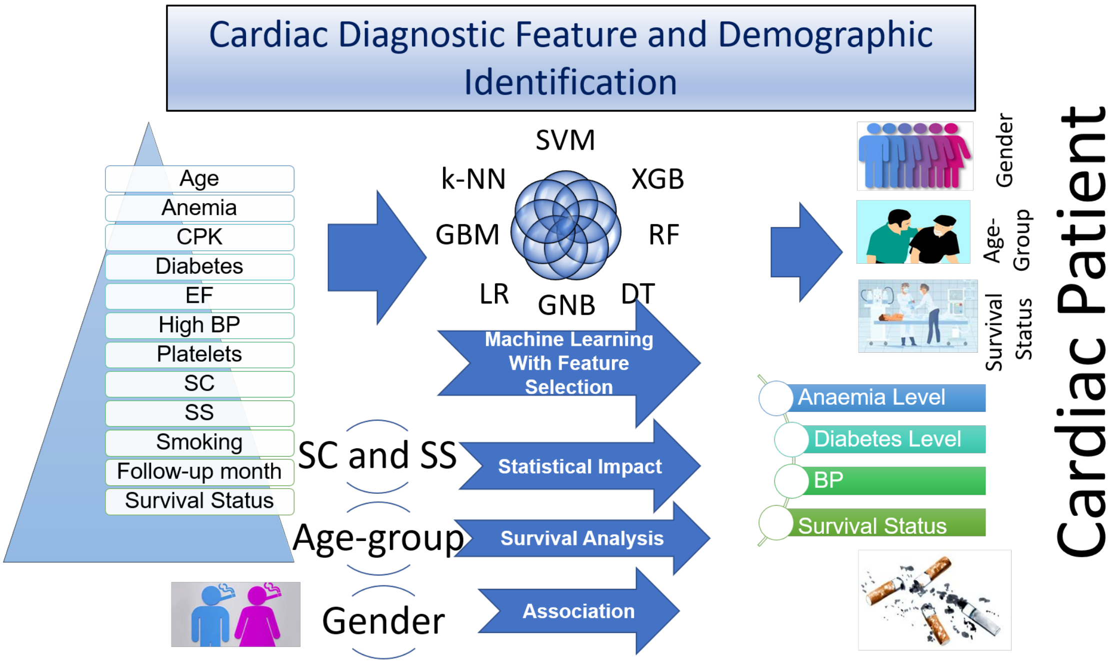


Figure 1. Cardiac Diagnostic Feature and Demographic Identification (CDF-DI) Models.

The model can predict patient's Survival-Status level (Alive/Dead), their gender (Woman/Man), and also their age group (Adult/Very Old). For this, the model is using eight (8) new machine learning algorithms (k-NN, SVM, XGB, LR, DT, RF, GBM, GNB) for classification purposes. These predictive algorithms are implemented, and their results are compared in terms of several performance metrics. The model can be used for confirming the impact of survival status, anemia, diabetics, and BP levels towards SC and SS of the patient. The proposed CDF-DI can assist doctors in efficiently diagnose patients, consequently enhancing the clinical decision-making processes for heart disease. Early treatment can therefore be done to reduce deaths resulting from the detection of late cardiac disease. Our research contributions can be summarised as follows.

– It is enhancing the diagnostic accuracy of Survival-Status, age-group, and gender of heart disease patients. The present paper proposed a model - CDF-DI, in which it integrates the

VIF technique for eliminating multi-collinearity among attributes. Further, it is adopting a 10-fold CV method for finding a robust predictive model among eight machine learning model classifiers based on MCC, F1-Score, and accuracy ranking performance metrics.

- Analysis the performance of machine learning models and comparison with state-of-the-art models. The presented CDF-DI is compared to the findings of previous studies and evaluated against another classification model. The present study also included a statistical analysis to confirm the proposed model's significance compared to other models.
- Verifying the impact of patients' complication levels towards SC and SS. In identifying the effects and association, the present article used patient's Survival-Status levels, BP levels, anemic level, and diabetic level towards SC and SS with Mann-Whitney U Test (non-parametric test). The gender and smoking level association was also verified with χ^2 (non-parametric test). Survival analysis is also done to ascertain the impact of age-group levels on Survival-Status levels with patients' follow-up months.

5.2. Dataset Description

The HF clinical record dataset used in this study was obtained from the UCI Machine Learning repository [45,46]. Every patient profile has 13 clinical characteristics, and the dataset includes medical reports of 299 patients who had heart problems that were collected during the follow-up months. There are 194 men and 105 women among the 299 records. The average follow-up period was 130 days, ranging from 4 to 285 days. A cardiac echo study or specialist notes is used to diagnose the disease. Renal dysfunction is indicated by an SC level higher than the average level(1.5). There are no missing values in the dataset.Table 2 shows a high-level description of the dataset, including narrative, ranges, and their units of measurement.

All six(6) categorical variables are binary types with 0 and 1, in Gender 0 representing the woman and 1 representing man. In smoking level 0 representing the patient has no habit of smoking, and 1 describes addiction to smoking. Diabetes level, anemic level, and High BP level-0 represent patients with no complications of diabetes, anemic and high BP, and values 1 illustrate suffering from diabetes, anemic, and high BP. In Survival-Status 0 level -representing patient is Alive, and 1 representing patient has died from cardiac arrest. In the dataset, patients' age lies between 40 to 95, with a mean value of 60.83. SC and SS values lie between 0.5 to 9.4 mg/dL and 114 to 148 mEq/L with mean values 1.39 and 136.63, respectively. CPK values lie between 23 to 861 Mcg/L with a mean value of 581.84. The EF ranges from 14 to 80, with a mean of 38.08. It is found that 203 alive patients have a mean CPK level of 540.05 ±52.91, and 96 dead patients mean the number of CPK levels was 670.20±134.37. Alive patient's CPK range was found in between 30 to 5209, and dead patients have range 23 to 7861 as displayed CPK enzyme level w.r.t Survival-Status levels (Alive/Dead) in Figure 2(a).

Figure 2(b) is displaying SC level w.r.t anemic levels. It is found that 170 non-anemic patients had mean SC was 1.35±0.06, and 129 anemic patients had a mean SC was 1.46±0.11. SC's minimum value was 0.50 and maximum was 6.80 w.r.t non-anemic level and anemic patients had SC value between 0.60 to 9.40. Figure 2(c) is showing SC level w.r.t diabetic level. It is found that non-diabetic patient means SC was 1.43±0.09 and diabetic patient mean SC was found 1.34±0.07. Non-diabetic patients SC was found in between 0.50 to 9.40, and diabetic patient SC range found 6.20. Also, 194 non-BP patients' mean SC level was 1.40±0.06, and 105 high-BP patients' mean SC was 1.39±0.13 found. Non-BP patient minimum SC was 0.60, and maximum 6.80 found, and high-BP patient SC range was found between 0.50 to 9.40 as depicted in Figure 2(d). In Figure 2(e), SS is displayed w.r.t Survival-Status also. It is found that alive patients had mean SS was 137.22±0.28, and dead patients had 135.38±0.51. Alive patients' SS range was found between 113 to 148, and dead patients' SS range was 116 to 146. As depicted in Figure 2(f), 170 non-anemic patients' mean SS was 136.46±0.34 found, and 136.84±0.39 mean SS value was found in 129 anemic patients. Minimum 113 and maximum 148 SS were found in non-anemic patients.

doi:10.20944/preprints202106.0201.v1

For patients who are suffering from anemia, minimum and maximum SS were found between 116 to 145.

SS w.r.t diabetic level whisker-plot in Figure 2(g) showing non-diabetic and diabetic patients overall summary. It is found that non-diabetic patient means SS was 139.96±0.29 and diabetic patients had SS was 136.16±0.46. High-BP patient means SS was 136.85±0.40 found and non-BP patient 136.51±0.33. It is a very minute difference between them, as depicted in Figure 2(h). High-BP patient SS range found between 124 to 148, and non-BP patient SS was in between 113 to 146.

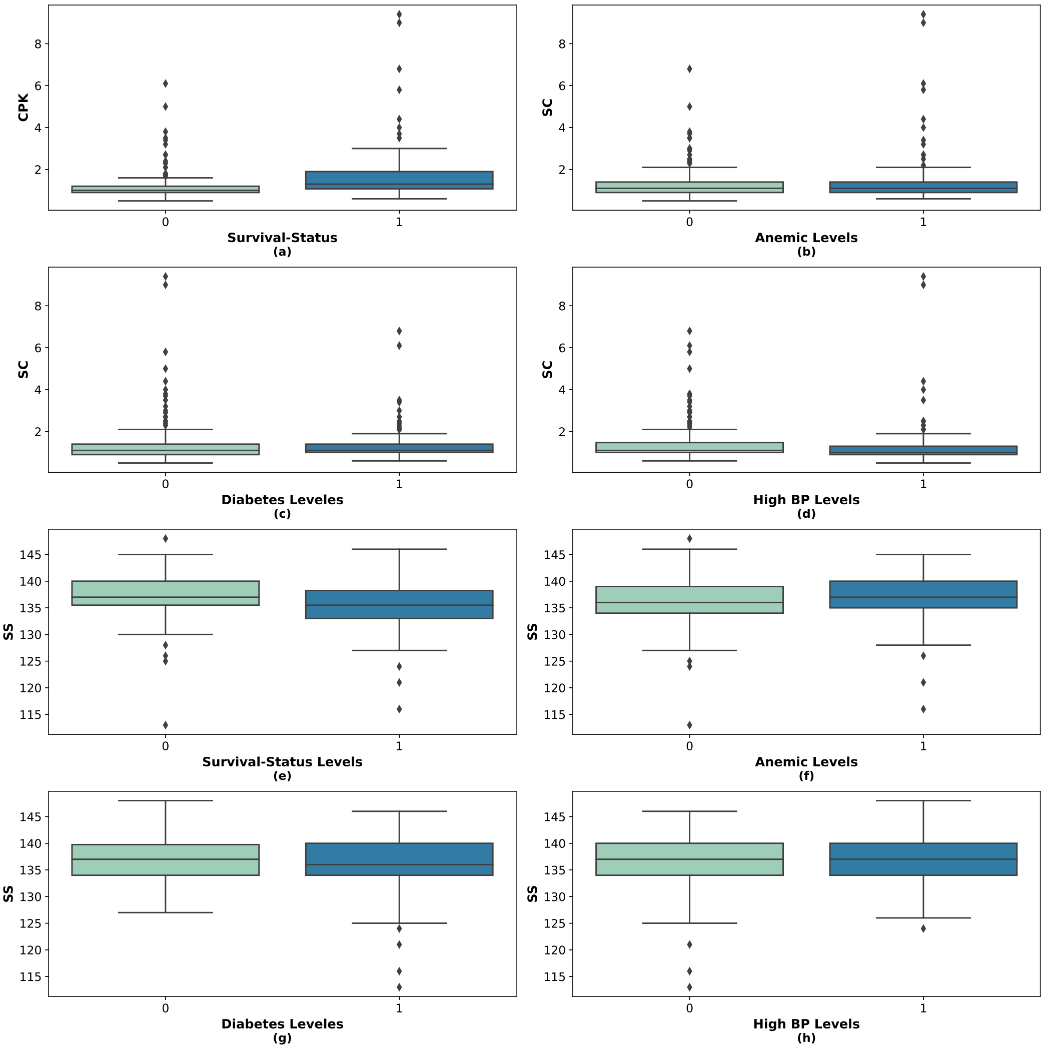


Figure 2. Dataset Box-Plot (**a**) CPK Vs Survival Status, and SC Vs. (**b**) Anemic Level, (**c**) Diabetic Level, (**d**) High BP Level, and SS Vs. (**e**) Survival Status Level, (**f**) Anemic Level, (**g**) Diabetic Level, (**h**) High BP Level.

Distribution plot displayed in Figure 3 of metric variables available in dataset. It is found in Figure 3(a) that EF had non-normal distribution ($p < 0.05$) with mean 38.08 and 11.83 SD was found with 36.74 lower bound and 39.43 upper bound at 95% confidence interval. The non-Normal distribution was also found in the platelets with a mean 263358.03 with 6.21 kurtosis and 1.46 skewness with 97804.24 SD in Figure 3(c). At 95% confidence interval, 252226.94 as lower bound and 274489.12 as upper bound found. A total of 299 samples of SC density distribution plot displayed in Figure 3(d). It is also non-normal distribution found with a mean of 1.40. Kurtosis and skewness were 25.83 and 4.46, respectively, found. SS distribution plot also displayed in Figure 3(e), and it is evident that the SS has non-normal distribution with 136.63 means. The lower bound and upper bound were 136.12 and 137.13, respectively, with 4.41 SD. Patients' Follow-up months distribution plot is displayed in Figure 3(f). From day 4 to day 285, the follow-up month mean was 130.26 found. The absence of normality was also there with 77.61 SD.

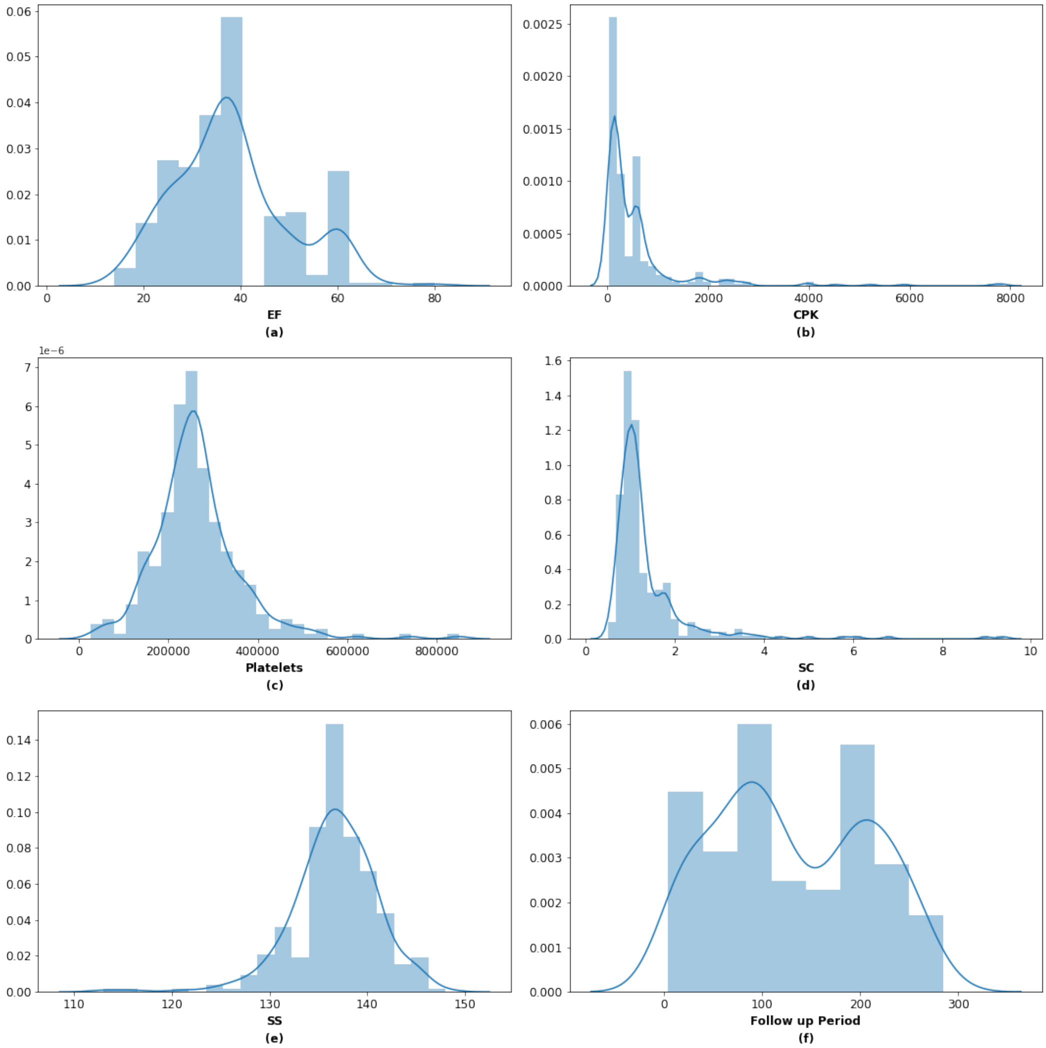


Figure 3. Dataset Distribution (**a**) EF, (**b**) CPK,(**c**) C-platelets,(**d**) SC, (**e**) SS, (**f**) Follow-up.

5.3. Features Correlation

The correlation between attributes can influence the machine learning model's performance. Data correlation can be used as a measurement tool to evaluate the relationship among features using Pearson's correlation. These values range from -1 to +1, indicating a negative or positive relationship between the attributes. Figure 4, a value close to zero indicates a low correlation between features for the dataset. The light-blue color implies that the correlation is close to zero, while the dark blue and dark orange colors mean that the correlation is close to +1 and -1, respectively. Diabetes and Sex are observed to have near zero, indicating very low or no correlation with the target attribute (Survival-Status). Accordingly, Survival-Status and SC attribute also have close to zero relationships with the Gender (Sex) target class. Thus, we could remove these features to improve the performance of our proposed model.

Table 2. Dataset Description.

Continuous variables				Categorical Variables		
Attribute Name	Description	Range	Measured In	Attribute Name	Description	Range
Platelets	Platelets in blood	25100-85000	kiloplatelets /mL	Gender	Woman/man	0-1
Age	Age of Patient	40-95	Years	Smoking	Yes/No	0-1
SS	135.39	114-148	mEq/L	Diabetes	40 (42%)	0-1
SC	Level of creatinine in the blood	.50-9.40	mg/dL	High BP	Yes/No	0-1
EF	Percentage of leaving the heart at each concentration	14-80	Percentage	Anaemia	Decrease in Red Blood Cell/Haemoglobin	0-1
CPK	Level of CPK enzymes in the blood	23-7861	Mcg/L	Survival-Status	Died / Alive	0-1
Time	Follow up Month	4-285	Days			

Source: Own elaboration.

doi:10.20944/preprints202106.0201.v1

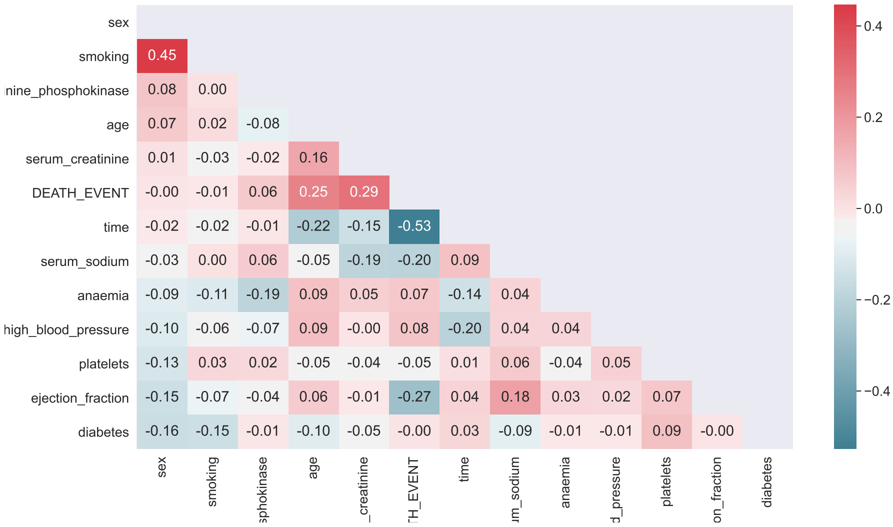


Figure 4. Feature Correlation.

5.4. Preprocessing

This derived feature is used as the target feature in this article. The Variation Inflation Factor (VIF) in Equation (1) is a measure of how multicollinear the dataset's features are concerning a given target feature. The VIF value of an attribute $a \in A$ is determined from a dataset D = (A, X) using a standard linear regression with an as the prediction target. Then, given R, the linear regression coefficient of determination, we have:

$$VIF(a) = \frac{1}{1 - R^2} \tag{1}$$

A variance inflation factor more significant than 10 implies high multicollinearity of the attribute with other dataset attributes. This arbitrary number 10 is used as a norm in several publications [48]. Furthermore, when a feature is excluded from the dataset, the VIF of the multicollinear features decreases. By calculating the VIF of each feature (considered a target) in the dataset and comparing it to a new VIF calculation with a feature removed, we can detect groups of attributes automatically. The paper also used backward elimination used for feature selection among the aforementioned dataset features. Tests for correlation and multicollinearity among features were tested using the VIF.

In this article, the present research article created a new derived attribute, “Age-group” from on age attribute from our target HD dataset. More patients in the very old (170) age-group level than "Adult" (129), the level found in derived Age-group attribute.

The model has significant multicollinearity based on VIF (>10) when gender and age group as dependent variables (Table 3).

Table 3. Feature's VIF of Gender and Age.

	Towards Gender		Towards Age	
Features	VIF Score	VIF score (backward elimination)	VIF Score	VIF score (backward elimination)
Age	30.12	Dropped	30.40	Dropped
Anaemia	1.90	1.79	1.91	1.79
CPK	1.45	1.40	1.46	1.42
Diabetes	1.78	1.75	1.79	1.75
EF	13.09	7.78	13.35	7.91
High BP	1.63	1.57	1.65	1.57
Platelets	8.49	7.02	8.64	7.02
SC	3.13	3.03	3.13	3.05
SS	58.37	Dropped	61.55	Dropped
Smoking	1.55	1.47	3.81	3.29
Time	5.65	4.02	1.89	1.89
Survival-Status	2.46	1.94	5.66	4.16
			2.47	1.97

Source: Own elaboration.

6. Models Hyperparameter Tuning

ML model's efficiency can be increased on the given dataset by tuning hyperparameters. The method of hyperparameter selection is one of the most critical characteristics of ML models. More time to adjust the hyperparameters for an effective result is generally required. Optimization of hyperparameters may be described as follows in Equation (2):

$$x^* = arg.m_x \varepsilon_X f(x) \tag{2}$$

Here f (x) denoting the objective score to minimize the validation set errors, x* is the minimum score hyperparameters collection, and x may assume any domain value of X. The present article using the objective score to maximizing the MCC score by minimizing the validation set errors using grid search with 10-fold cross-validation. The following listed parameters were utilized for ml models. The RF works by adopting various decision tree classifications for different dataset sub-samples and uses averages to increase predictive precision and track overfitting. The max sample parameter is managed if bootstrap=True (default), otherwise the entire dataset will be used to make a tree each [51]. For all proposed classifiers (Age-Group, Gender, Survival Prediction) Gini was used as a criterion and used to measure the quality of a split, max-feature set to 7 for the best split, min-sample-leaf set to 2 as the least sample number necessary for the leaf node. The dividing point is to be considered at any depth only when at least min samples leaf samples are left on each branch of the left and right. Min-sample-split set to 2 to split an internal node and number of trees in the forest set to 50 by setting n-estimators. In the decision tree (DT) classifier, a criterion is the same as the random forest set as Gini with max-features is set to log. The max depth and minimum sample split are set to 50. SVM library in machine learning is popular due to high-dimensional space and uses a subset of training samples in decision function. It has a dynamic approach to solve the problem due to kernels exist [52]. The SVM classifier is used in this paper for classification purposes. This C (10) parameter is used here for regularization parameter and must always be greater than zero and with radial basis kernel. Gamma value is used kernel coefficient set as 0.001 (Table 13). Gradient boosting classifier (GBM) creates a forward-looking additive model; it enables arbitrary differentiable loss functions to be optimized. The n classes regression trees in each step are fitted to the negative gradient of the binomial or multinomial default function. Binary classification is a special case in which only one tree is caused [53]. Learning rate (0.001) reduces each tree's contribution by learning rate. A compromise exists between the learning rate and n estimators. The number of nodes in the tree is limited by the maximum depth. This parameter should be tweaked for optimal performance; the interaction of the input variables determines the best value is set as 3 with a number of an estimator (nestimator) is 1000 (Table 13). SVM Linear classifier with stochastic gradient descent (SGD) learning to implement regularised linear models: the gradient of the loss is measured one sample at a time, and the model is modified with a decreasing intensity schedule along the way (aka learning rate) [54]. The alpha value of sgd classifier uses as multiple of regularization term set here as 0.1 and loss used for training time in an adjustment of several weight updates here set as log value. Penalty used for regularization term set to as none Table 4. Tree boosting is a common and successful machine learning technique. We explain XGBoost (XGB), a scalable end-to-end tree boosting method commonly used by data scientists to achieve state-of-the-art results on a variety of machine learning challenges [55]. To avoid overfitting, step size shrinkage was used in the update (eta aka learning rate) set as 0.1 and gamma set as 0. If the max delta step is zero, then there is no constraint to be followed but here set as 2 updates to follow more conservative. Maximum depth of tree set as 6 and increase this result model become more complex with minimum child weight set as 4 with 200 trees in the forest (nestimator). k-NN works on the principle of nearest neighbor methods on the training samples closest to the new point and uses them to predict the mark and the number decided by nneighbour set as 3 with mahanntan metric and weight set as uniform. Despite its name, logistic regression is not a regression but an algorithm of classification. It is used to estimate discrete values (0 or 1, yes/no. true/false) based on a given set of independent variables. It is also known as logit or MaxEnt [56,57]. Newton-cg set as a solver to handle multiclass problem and newton-cg handle only l2 penality with regularization parameter c set as 1.0.

Table 4. Model Hyperparameter Tuning.

Classifier	Model Tuning Parameters
RF	criterion= 'gini', max_features= 7, min_samples_leaf=2, min_samples_split= 2, n_estimators=50
DT	criterion='gini', max_depth=50, max_features= 'log2', min_samples_leaf= 1, min_samples_split= 50
SVM	C=10, gamma= 0.001, kernel= 'rbf'
GBM	learning_rate=0.001, max_depth= 3, n_estimators=1000, subsample= 0.5
SGD	alpha= 0.1, loss='log', penalty='none'
XGB	Gamma = 0, learning_rate = 0.1, max_delta_step = 2, max_depth = 6, min_child_weight = 4, n_estimators = 200, reg_alpha = 0, reg_lambda = 8
k-NN	metric='manhattan', n_neighbors= 3, weights= 'uniform'
LR	C=1.0, penalty= 'l2', solver= 'newton-cg'

Source: Own elaboration.

7. Machine Learning Experiment Design

The present paper experimented on machine learning binary classifiers for prediction on given dataset using: k-NN [44], GBM [36], DT[33], Linear SVM [38] Radial SVM [38], RF [34], XGB [43], LR [35], GNB [37]. All proposed machine learning models' definitions can be found in Table 5. The experiment trained and tested all classifiers with it 10-fold cross-validation using the grid search CV method. The paper trained each model with a different hyper-parameter on the training set, applied it to the validation set and then chose the model with the highest MCC as the final model to apply to the test set. In this experiment, we repeated experiments ten times for all classifiers and documented the highest result for MCC. The paper then arranges the result table according to ranking based on MCC first, F1-Score (second) and then finally, results are arranged based on accuracy. Results are displayed on the theme of different metrics, different ranks. The three rankings we applied to the report yielded the same results, revealing intriguing features resulting, when ranking based on MCC, F1 -score, or the accuracy, the top classifier changes.

Table 5. Machine Learning Models.

Model	Description	Reference
DT	DT is an algorithm of classification which works well on categorical and numerical forms of data. It is generally used to build tree-like structures. Medical data can be analysed easily with good accuracy.	[33]
RF	RF is a model of tree-based ensemble learning that produces exact prediction by combining several weak learners. This model uses the bagging technique for training a range of decision tree with different bootstrap samples	[34]
LR	LR typically predictive analysis based on the concept of probability. Binary categorical variable is predicted by one or more independent variable using sigmoid function	[35]
GBM	Many weak classifiers work together to build a powerful model for learning on the GBM. It usually time taking process due to creation of many independent tree. It has ability to deal with missing values	[36]
GNB	GNB is a naive bayes variant that works with gaussian distributions and is used for continuous data. The prior and posterior likelihood of the class in the data are involved in conjunction with a function that has constant values. All of the features are often assumed to obey a gaussian or regular distribution	[37]
SVM	SVM is a mathematical model-based supervised learning technique. It is used to solve problems including regression and classification problems. It classifies data by creating high-dimensional hyperplanes, also known as decision planes. Hyper planes are used to separate one form of data from another	[38]
XGB	The XGB is a popular ensemble learning algorithm that uses DT models in the background for computation. It is a highly effective scalable machine learning algorithm. It combines multiple weak-learner to build a strong classifier proved a better classifier.	[43]
k-NN	When compared to a collection of known data, the k-NN method allows us to identify unknown data by calculating the distance or similarity of an unknown datum. It assigns a class to the datum based on the number of neighbors with the same class who are the nearest to it. k controls or indicates the number of neighbors used in the decision.	[44]

Source: Own elaboration.

8. Feature Ranking and Selection

A subset of features reflects the characteristics of the original number of features and helps calculate the target feature. The feature importance help to reduce the computational cost by removing irrelevant features [58]. Fisher scores feature selection method uses a filter-based method by computing large distances between data points in different classes and a small distance between datapoints in the same classes approach [59]. Another popular feature ranking and selection method are Chi-square(χ^2) which works on χ^2 statistics w.r.t class labels. Higher the χ^2 value, the higher the related feature [60]. The χ^2 statistics and its related p-value is computed by cross-tabulation method [61]. RF offers two feature ranking techniques: mean accuracy reduction and Gini impurity reduction. As we know, RF generates various DT during training for working on the subsets of data and features. It observes all the outcomes of DT and selects its outcome based on the majority of votes. Mean accuracy reduction feature ranking technique tally the prediction accuracy on dropping of particular feature with the rest features' accuracy results. Rank the feature accordingly after observing the difference. It works on the principle that the bigger the precision decrease, the larger the feature importance is [62]. Another method of feature ranking also works on the same principle by using Gini as a metric instead of on the accuracy [63]. Regarding machine learning feature ranking, we used the RF, XGB inbuilt feature ranking algorithms ("Feature Ranking Results" Section) to extract the top most common features. RF and XGB are used to rank the feature as they proved to be a better classifier with the highest accuracy among all used classifiers in the present work.

9. Performance Evaluation Metrics

Scientific researches use a variety of performance matrices to assess prediction accuracy [64]. Still, no broad consensus is reached on a single elective measure yet despite the various best machine learning methods. In binary classification problems in machine learning, accuracy and F1-Score derived from Confusion matrics (CMs) have been (and continue to be) among the most widely used metrics for performance evaluation. These statistical measures with h imbalanced dataset could, however, dangerously demonstrate over optimal inflated results.

The proportion of actual negatives that were predicted as negatives is known as specificity (or true negative).This means that a part of true negatives will be predicted as positives, which could be considered as false positives. Equation described in Equation (3).

$$TNR = \frac{(TN)}{(TN + FP)} \tag{3}$$

Our model's recall is a metric for how well it can detect True Positives. As a result, recall tells us how many patients we accurately identified as having HD out of all those who have it. Mathematical Equation (4) describing this.

$$Recall = \frac{(TP)}{(TP + FN)} \tag{4}$$

Precision is the ratio of True Positives to all Positives in its most primitive sense as described in Equation (5).

$$Precision = \frac{(TP)}{(TP + FP)} \tag{5}$$

The ratio of the overall number of right predictions to the total number of predictions is known as accuracy as described in Equation (6).

$$Accuracy = \frac{(TP + TN)}{(TP + TN) + (FP + FN)} \tag{6}$$

The Harmonic Mean of Precision and Recall is the F1-Score as described in Equation (7).When we want to strike the right balance between precision and recall, we'll need the F1-Score.

$$F1 - Score = \frac{(2 \times TP)}{2 \times TP + FP + FN} \tag{7}$$

The popular Mathews Correlation Coefficient (*MCC*) in Equation (8) would be a more accurate statistical measure that only yields a high score if the prediction performed well in all four CMs groups (true positives, false negatives, true negatives, and false positives) despite imbalance dataset [49]. The *MCC* would produce a high score only if the binary predictor were able to correctly predict the majority of positive data instances and the majority of negative data instances while working with binary classification [50]. It has the worst value of -1 and the best value of +1.

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{8}$$

10. Experiments and Results

10.1. Experiment-1

This experiment was conducted to measure the SC and SS levels differences towards Survival-Status levels (0: Alive, 1: Dead) due to HF of 299 patients (H_{01}). The test was an experiment with a non-parametric Mann-Whitney U-test. Table 6 displays the U test statistics values differences between Alive and Dead patients towards SC and SS.

Table 6. Impact of SC and SS on Survival-Status of the Patient.

Parameter	Distribution normal	Homogeneity in variance	μ rank	U	Sig. 2-tailed *(p)*
SC	×	×	0: 128.10 1: 196.31	5298.00	0.00
SS	×	√	0: 162.40 1: 123.78	7226.50	0.00

Source: Own elaboration.

Surprisingly, across both populations, in case of SC and SS, the present paper found significant p-values for both (Alive and Dead) levels. This paper also found significant differences with unique μ rank in SC (μ rank : 128.10 & 196.31). The significant p value of SC suggested to consider significant differences (U = 5298, Z = -6.398, p < 0.05). In case of SS, significant differences could be observed in mean rank (μ rank : 162.40 & 123.78). Further, the SS significantly different towards Survival-Status levels (U = 7226.50, Z = -3.622, p < 0.05).

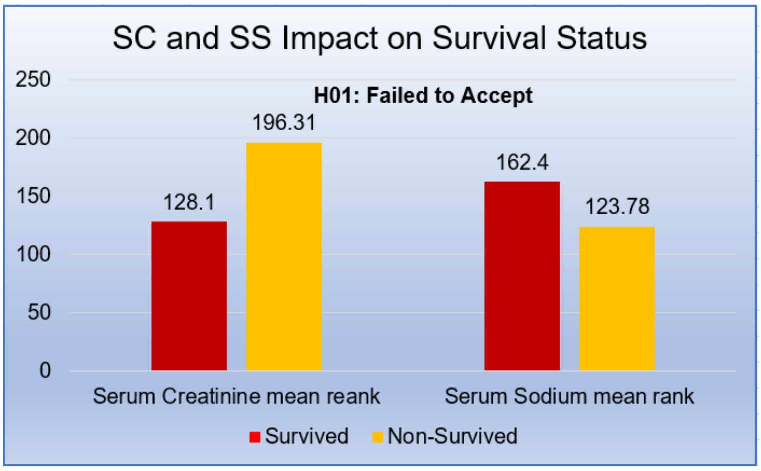


Figure 5. SC and SS Impact on Survival Status

Figure 5 exhibits the concrete differences for SC with mean rank (μ rank : 128.10 & 196.31). Hence, we found that Dead patient due to HF, their SC is greater. But, in the case of SS, the reverse scenario could be observed. Dead patient SS became low when a patient suffered from HF. The article suggests greater SS for a healthy life, and differences can be observed with their mean ranks (μ rank : 123.78 & 162.40). Therefore, this experiment results suggest that there is a significant impact of SC and SS on the Survival-Status levels, i.e., we reject the null hypothesis H_{01}.

10.2. Experiment-2

The experiment is conducted to test the three null hypotheses (H_{02}-H_{04}) to verify the differences in SC and SS w.r.t anemic levels, diabetic and high BP levels. All the hypotheses were tested with a Mann Whitney U test.

Table 7. Impact of SC and SS on Survival-Status levels of the Patient.

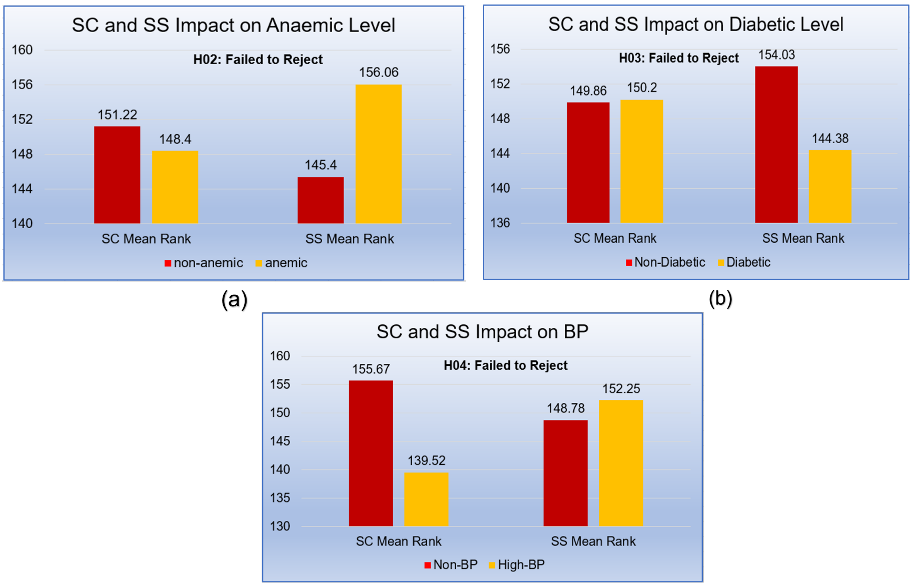
Variable & Assumptions			Anemic levels		Diabetic levels		BP levels		Sig. 2-tailed (p)
Parameter	Distribution normal	Homogeneity in variance	μ rank	U	μ rank	U	μ rank	U	
SC	×	√	0:151.22 1:148.4	10758	0:149.86 1:150.20	10850.0	0: 155.67 1: 139.52	9085	p >0.05
SS	×	√	0:145.40 1:156.06	10183.5	0:154.03 1:144.38	10173.5	0: 148.78 1: 152.25	9948.50	p >0.05

Source: Own elaboration.

Table 7 shows that the Mann Whitney U test statistics values to evaluate the hypotheses: H_{02a}, H_{02b} and H_{02c}. We found no significant p-values in anemic levels (Non-Anemic and Anemic), diabetic levels (Non-Diabetic, Diabetic), and BP level (Non-BP, BP) towards SC and SS. The SC w.r.t non-anemic (Median (Mdn) = 151.22) and anemic (Mdn = 148.40) patients are not significantly different, according to test results (U = 10758, p > 0.05). Same scenarios can be noted in the SS w.r.t non-anemic (Mdn = 145.40), and anemic (Mdn = 156.06) levels are not also significantly different with (U = 10183.50, p > 0.05). Therefore, finings suggests an insignificant p-value in the case of SC. In SS, we failed to reject "H_{02a}: No Significant difference between non-anemic and anemic levels towards SC and SS".

Further, the SC w.r.t non-diabetic (Mdn = 149.86) and diabetic (Mdn = 150.20) levels patients are not found statistically significantly different (U = 10850, p > 0.05). Same result patterns can be observed in the SS w.r.t non-diabetic (Mdn = 154.03) and diabetic (Mdn = 144.38) level patients are not also different with (U = 10173.50, $p > 0.05$) towards SS. therefore, we also failed to reject the null hypothesis (H_{02b}), i.e., there are no significant differences in diabetic and non-diabetic patients towards SC and SS.

Moreover, the SC in Non-BP (Mdn = 155.67) and BP (Mdn = 139.52) levels patients are not found fundamentally unique, as indicated by a U test statistical results (U = 9085, $p > 0.05$). Similar conduct can be noted in the SS w.r.t BP level in non-BP (Mdn = 148.78), and BP (Mdn = 152.25) patients are not likewise unique with (U = 9948.50, $p > 0.05$) towards SS. Therefore, we also failed to reject the null hypothesis (H_{02c}), i.e., and there are no significant differences found between Non-BP and BP patients towards SC and SS.



(a) (b)

(c)

Figure 6. SC and SS Impact on (**a**) Anemic level, (**b**) Diabetic Level, (**c**) BP.

As Figure 6(a) displaying, Non-Anemic patients had more excellent SC as compare to anemic patients ($\mu = 151.22 > \mu = 148.40$). But in the case of SC, anemic non-patient had less SS as reach anemic patients ($\mu = 145.4 < \mu = 156.06$). Finding suggests that there are no significant differences. From Figure 6(b), Non-diabetic patients had non-remarkable differences found in SC with diabetic patients with mean rank ($\mu = 149.89 < \mu = 150.20$). But in the case of SS, non-diabetic patients had a little-bit greater SS level as compared to diabetic patients ($\mu = 154.03 > \mu = 144.38$), but results were not statistically significant. As Figure 6(c) shows, Non-BP patients had more significant differences towards SC as compare to BP patients with mean rank ($\mu = 155.67 > \mu = 139.52$). But in the case of SS non-BP patients had minor differences toward SS level as compare to BP patients ($\mu = 148.78 < \mu = 152.25$) but experiment results unable to prove significance in results.

10.3. Experiment-3

This experiment was conducted to verify the association between gender and smoking levels of CHD patients. The experiment used a non-parametric χ^2 test to explore the association of gender level with smoking levels. We have two nominal categorical variables as the χ^2 test considers the non-metric variables and uses cross-tab as an input to explore the link.

Table 8. Observed value of Gender Vs. Smoking Status.

Gender	Non-Smoker	Smoker	Row Marginals (Row Sum)
Female	101	4	**105**
Male	102	92	**194**
Column Marginals (Columns Sum)	**203**	**96**	**299**

Source: Own elaboration.

Table 8 shows the cross-tab of observed values of gender and smoking levels, and Table 9 shows the expected values with χ^2 of each cell. Residuals (Observed-Expected) are also marked in this table. A positive residual cell χ^2 value means that the observed value is higher than the expected value. A negative cell residual χ^2 value (e.g., -29.7) means the observed cases are less than the expected number of cases.

Table 9. Expected value of Gender Vs. Smoking Status (χ^2 Values)

Gender	Non-Smoker	Smoker	*df*	χ^2	Sig. 2-tailed (*p*)
Female	71.30 (12.37)	33.7 (26.17)	1	59.45	0.00
Residual	29.7	-29.7			
Male	131.70 (6.70)	62.3 (14.16)			
Residual	-29.7	29.7			

Source: Own elaboration.

Table 9, it is observed that the most significant χ^2 value 26.17 is found in the second cell. It is because the observed female smoker patients were 04, whereas 33.7 was expected. Therefore, the 2nd cell consists of a much larger number of expected cases than observed. This means that the number of observed female smoker patients was significantly less than expected. The second-largest χ^2 value of 14.16 is located in the Male Smoking patient's cell. However, we find that the number of observed cases in this cell was significantly greater than expected (Observed = 92, Expected = 62.3). This indicates that a substantially higher number of male smokers is observed than what is actually expected. The third-largest cell χ^2 value of 12.37 is located in a non-smoker female's cell. The observed value of 101 and the predicted value of 71.30 for a non-smoker female were found. This means observed female non-smoker was significantly greater than expected. The last χ^2 value 6.70, in which expected non-male smokers (131.70) were considerably smaller than observed (102).

Further, It has been evident that the two groups were significantly associated ($p < 0.05$) with $\chi^2(1) = 59.45$. therefore, the findings suggest we reject the null hypothesis H_{03} that no significant association of gender with smoking level.

10.4. Experiment-4

In this experiment, the research article checks the age-group significance of association to the Survival-Status levels using follow-up months with the help of the survival analysis-cox regression model. Table 10 shows a descriptive summary that out of 299 patients, 129 are adults; 170 are from very old levels of age-group; 31 (24%) in adult and 65 (38.2) in very old age-group patient died due to CHD. 98 (76%) in adults and 105 (61.8%) in the very old age group found censored in the study to the end of follow-up months.

Table 10. Age-group Vs. Survival-Status Descriptive Summary

Case Processing Summary				
Age_Group	Total Patients	No. of Deaths	Censored	
			N	Percent
Adult	129	31	98	76.0%
Very Old	170	65	105	61.8%
Overall	299	96	203	67.9%

Source: Own elaboration.

When a patient is censored, it can be due to the follow-up study period is over and the patient has not experienced the event, or it can be due patient's follow-up is lost. Since the Cox regression is a semi-parametric model, model fitting did not estimate the intercept (baseline hazard). Diabetes, platelets, gender, and smoking were all shown to be non-significant factors in the Cox model in below Table 11.

Table 11. Significance Check towards Survival-Status

Variable	β coefficient	HR	Sig. 2-tailed (*p*)
Anemia	.497	1.644	.022
CPK	.000	1.0	.020
Diabetes	-.055	.946	.800
EF	-.046	.955	.000
high_BP	.490	1.632	.023
Platelets	.000	1.0	.972
Gender	-.134	.875	.590
smoking	.058	1.059	.818
Age-Group	-.596	.551	.008

Source: Own elaboration.

The negative age-group coefficient suggested that the risk of death from CVD in "Adults" have lower than "Very Old" in these data. The hazard ratio (HR) of Age-Group is exp (-0.60) = 0.55 here indicated that there are 55 deaths due to age group levels for every 100 deaths caused by CHD at each observation point. According to Table CPK, anemia, EF, high BP found significant towards Survival-Status levels ($p < 0.05$).

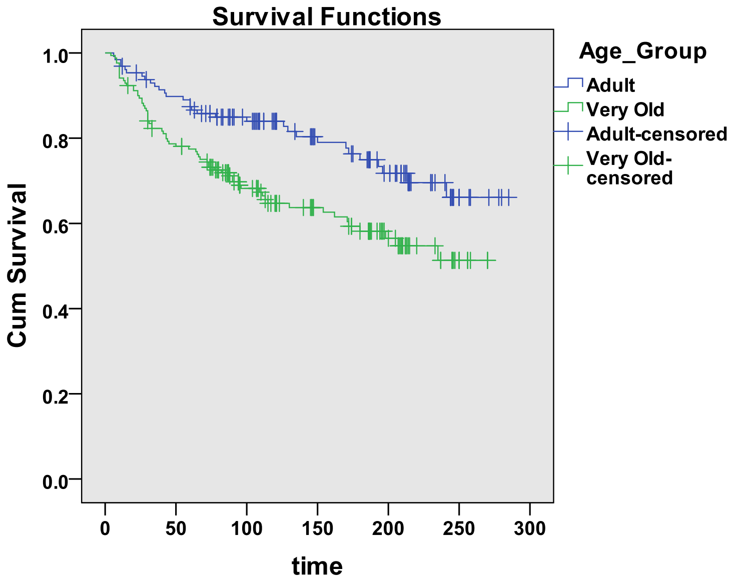


Figure 7. Kaplan Meier Curve of Follow-up months Vs. Age-Group Level.

Figure 7 visualizes the survival analysis of patient's follow-up months and age-group levels. Each level of Age-Group using the Kaplan Meier Survival Curve. It is evident that the survival rate for the Age-Group "Very Old" was lower than that of the "Adult.". It also indicates that difference between the two levels was statistically significant with log-rank p-value of .003 ($p < 0.05$) with χ^2 (8.565) at $df = 1$. Cross at curve indicating censored patients. Therefore experiments findings reject the null hypothesis H_{04} that no significant association between age-group with Survival-Status levels.

10.5. Experiment-5

This experiment predicted the gender (Woman, Man) of heart patients using various classifiers in Table 12. The present paper compared here eight state-of-the-art prediction algorithms (RF, GBM, DT, XGB, LR, SVM, k-NN, and GNB) that have a demonstrated track record for accuracy and efficiency in the research community. For all models, the paper used stratified 10-fold cross-validation and collected six performance metrics: MCC, F1-ratio, Accuracy, and Precision and recall, also called True Positive Rate (TPR), and also collected True Negative Rate (TNR), which is sometimes known as specificity. The finding revealed that the RF outperformed other models by achieving in terms of MCC (+0.87) with SD 0.25, with the highest accuracy (0.94) with SD 0.11, and F1-Score (0.95) with 0.09 SD. Furthermore, the results also demonstrated that the RF model achieved the highest TNR and TPR compared to other models, with a precision of 0.95 and recall of 0.95. GNB model in MCC, F1-Ratio and accuracy is the worst performer with +0.06, 0.71, and 0.59, respectively. In terms of recall, k-NN is the next following classifier after RF performing better with 0.80. GBM is following in terms of specificity after RF classifier with 0.68.

Table 12. Gender Classification Performance Metrics.

Classifier	MCC	F1-Score	Accuracy	Recall (TPR)	Precision	TNR
MCC Ranking:						
RF	**+0.87±0.25**	0.95±0.09	0.94±0.11	0.95±0.11	0.95±0.08	0.91±0.14
GBM	**+0.31±0.19**	0.71±0.06	0.65±0.08	0.64±0.08	0.79±0.09	0.68±0.17
LR	**+0.25±0.14**	0.75±0.04	0.67±0.05	0.78±0.08	0.73±0.05	0.45±0.14
Linear SVM	**+0.24±0.13**	0.75±0.04	0.66±0.05	0.78±0.07	0.72±0.05	0.45±0.13
DT	**+0.22±0.20**	0.72±0.06	0.64±0.07	0.72±0.13	0.73±0.08	0.50±0.24
XGB	**+0.21±0.18**	0.74±0.06	0.65±0.08	0.76±0.09	0.72±0.04	0.45±0.10
K-NN	**+0.12±0.11**	0.73±0.04	0.62±0.05	0.80±0.08	0.68±0.04	0.30±0.11
GNB	**+0.06±0.16**	0.71±0.05	0.59±0.06	0.76±0.09	0.66±0.04	0.29±0.12
F1-Score Ranking:						
RF	+0.87±0.25	**0.95±0.09**	0.94±0.11	0.95±0.11	0.95±0.08	0.91±0.14
LR	+0.25±0.14	**0.75±0.04**	0.67±0.05	0.78±0.08	0.73±0.05	0.45±0.14
SVM	+0.24±0.13	**0.75±0.04**	0.66±0.05	0.78±0.07	0.72±0.05	0.45±0.13
XGB	+0.21±0.18	**0.74±0.06**	0.65±0.08	0.76±0.09	0.72±0.04	0.45±0.10
k-NN	+0.12±0.11	**0.73±0.04**	0.62±0.05	0.80±0.08	0.68±0.04	0.30±0.11
DT	+0.22±0.20	**0.72±0.06**	0.64±0.07	0.72±0.13	0.73±0.08	0.50±0.24
GBM	+0.31±0.19	**0.71±0.06**	0.65±0.08	0.64±0.08	0.79±0.09	0.68±0.17
GNB	+0.06±0.16	**0.71±0.05**	0.59±0.06	0.76±0.09	0.66±0.04	0.29±0.12
Accuracy Ranking:						
RF	+0.87±0.25	0.95±0.09	**0.94±0.11**	0.95±0.11	0.95±0.08	0.91±0.14
LR	+0.25±0.14	0.75±0.04	**0.67±0.05**	0.78±0.08	0.73±0.05	0.45±0.14
SVM	+0.24±0.13	0.75±0.04	**0.66±0.05**	0.78±0.07	0.72±0.05	0.45±0.13
XGB	+0.21±0.18	0.74±0.06	**0.65±0.08**	0.76±0.09	0.72±0.04	0.45±0.10
GBM	+0.31±0.19	0.71±0.06	**0.65±0.08**	0.64±0.08	0.79±0.09	0.68±0.17
DT	+0.22±0.20	0.72±0.06	**0.64±0.07**	0.72±0.13	0.73±0.08	0.50±0.24
k-NN	+0.12±0.11	0.73±0.04	**0.62±0.05**	0.80±0.08	0.68±0.04	0.30±0.11
GNB	+0.06±0.16	0.71±0.05	**0.59±0.06**	0.76±0.09	0.66±0.04	0.29±0.12

Source: Own elaboration.

doi:10.20944/preprints202106.0201.v1

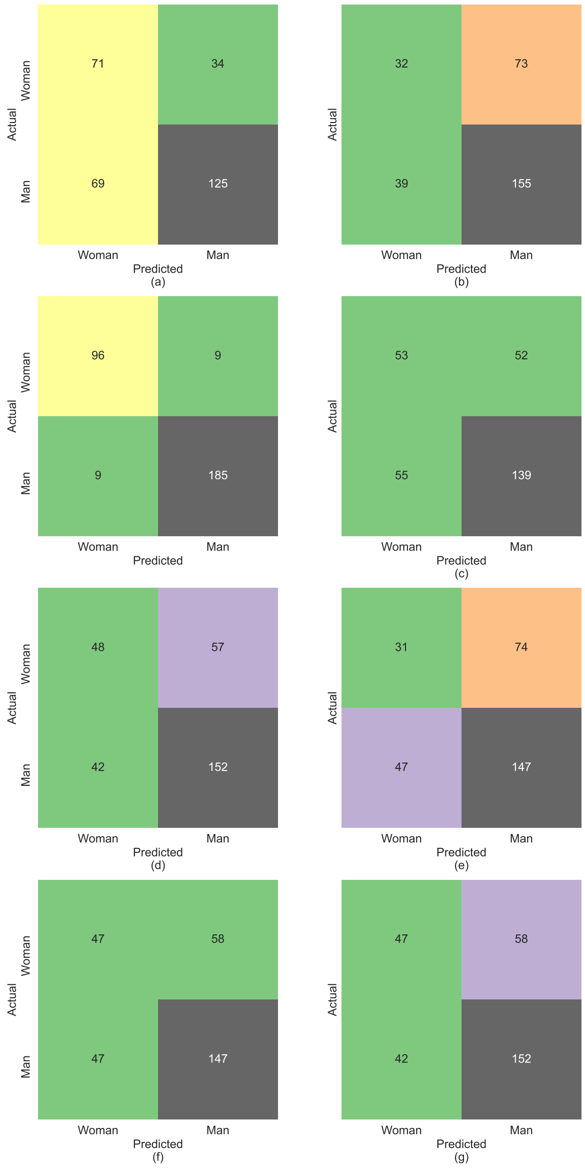


Figure 8. Gender classifier's CMs (**a**) GBM, (**b**) k-NN,(**c**) RF,(**d**) DT,(**e**) LR, (**f**), GNB,(**g**) XGB, (**h**) SVM.

Figure 8 depicts the gender classifiers' CMs for each proposed algorithm (GBM, k-NN, RF, DT, LR, GNB, XGB, SVM). The majority of the algorithms were able to obtain a classification accuracy of more than 59%. In Figure 8(c), the RF classifiers' CM TP representing men and TN rate representing women ratios are the highest among the offered techniques with 96 and 85, respectively. Only 9 women were misclassified as men, and 9 men misclassified as women in RF classifier compared to other classifier's CMs.The major misclassification occurred in GNB (Figure 8(f) and GBM (Figure 8(a) classifier. A majority count of 69 of men ratio misclassified as women in the GBM classifier, and a total of 74 majority of women misclassified as men in the GNB classifier.

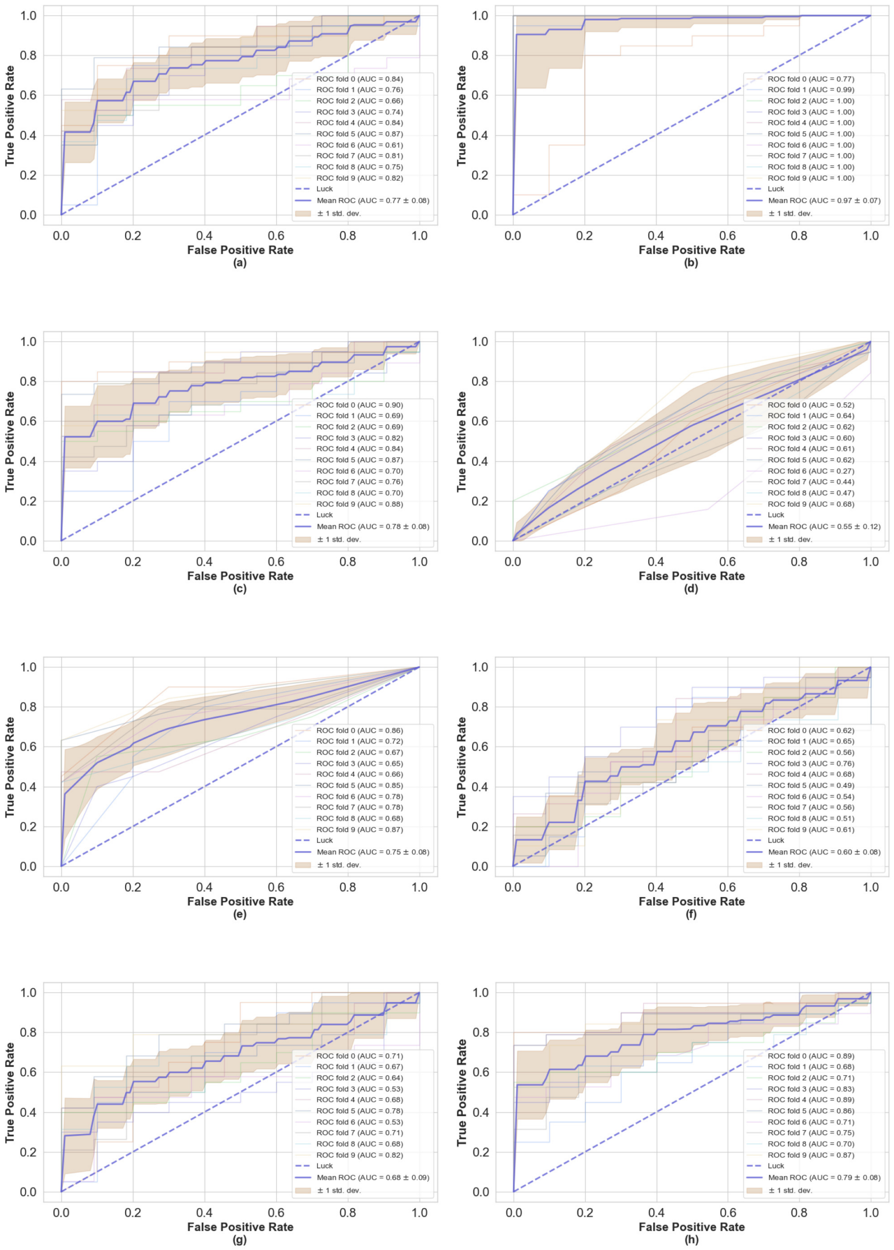


Figure 9. Gender classifiers' ROC's at dynamic thresholds (**a**) XGB, (**b**) RF,(**c**) LR,(**d**) k-NN,(**e**) DT, (**f**), GNB,(**g**) GBM, (**h**) SVM.

Figure 9 shows the TPR is plotted against the False Positive Rate (FPR) at various threshold values to construct ROCs, which is an essential tool for diagnostic test evaluation. The AUC (area under the ROC curve) is another method for determining a classifier's predictability. A classifier's superiority is measured by its AUC value, which is larger, the better. The best mean RF AUC (Figure 9(b) is 0.97 with SD of ±0.07 followed the AUC for SVM (Figure 9(h) is 0.79 with SD ±0.05. In a 10-fold cross-validation test, RF outperformed other classifiers with 96% accuracy, 97% sensitivity, and 95% specificity. The RF ROC appears to be more performing than the other proposed models based on their greater AUC, as shown in the ROC chart. The worst mean k-NN AUC (Figure 9(d)) is 0.55 with a SD of 0.12, followed by GNB AUC (Figure 9(f)) 0.60

with a SD of 0.05. According to experimental results, the proposed approach RF outperformed prior approaches addressed in the literature in terms of cross validiation accuracy.

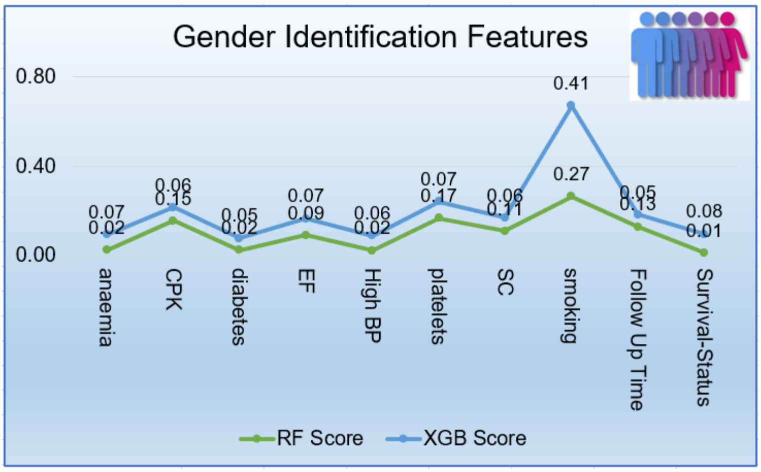


Figure 10. Gender identification Features.

Figure 10 visualizes the disparity between the two winner classifiers: RF and XGB, about their importance scores of features recognizing the patient gender. While predicting the gender of cardiac patients from a given dataset, the RF classifier gave the highest accuracy followed by the XGB classifier with accuracy, therefore extracting important features from them obvious. RF Classifier considered smoking (0.265), platelets(0.167), CPK(0.155), SC(0.109), EF(0.091) during follow-up months(0.129) as important features while predicting the gender as reflected in Figure 10. The XGB classifier also considered Smoking(0.406) as equal priority, but consider Survival-Status(0.084), EF(0.075), platelets(0.074), anemia(0.070) and High-BP(0.065) as important feature.

10.6. Experiment-6

This experiment was predicting the Age-Group (Adult, Very Old) of heart patients using various classifiers (Table 13). In fact, in terms of (MCC = +0.92) with SD 0.23, F1 score (0.96) with SD 0.11, and accuracy (0.96) with SD 0.11, the RF classifier is the best performing classifier. SVM classifier performs the worst of all listed algorithms in terms of (MCC=0.02) with SD 0.17, but k-NN classifier performs poorly in F1-Score ranking (0.62) with SD 0.08. In the accuracy, the ranking GNB classifier performing poorly with accuracy (0.55) with 0.05 SD. Indeed, as shown by our experiment, in both recalls (TP ratio = 0.97) with SD 0.07 and specificities (TN ratio = 0.95) with SD 0.16, with top MCC rating classification, RF was performing outstandingly. A significant number of patients' age groups could not be accurately predicted by k-NN(TPR=0.63) with SD 0.14 and adult group couldn't be predicted significantly (TNR=0.17) with SD 0.15 GNB classifier. Nevertheless, the researchers were again misled by higher accuracy values: a closer review of the findings showed that the SVM-SVC was badly negative (TNR = 0) in the case of GNB(TNR=0.17), with fewer patients correctly observed.

Table 13. Age-Group Classification Performance Metrics.

Classifier	MCC	F1-Score	Accuracy	Recall (TPR)	Precision	TNR
MCC Ranking:						
RF	**+0.92±0.23**	0.96±0.09	0.96±0.11	0.97±0.07	0.95±0.10	0.95±0.16
GBM	**+0.25±0.14**	0.73±0.05	0.64±0.05	0.86±0.10	0.64±0.04	0.35±0.13
DT	**+0.23±0.11**	0.68±0.07	0.62±0.05	0.69±0.14	0.66±0.06	0.53±0.17
XGB	**+0.23±0.15**	0.67±0.07	0.59±0.07	0.64±0.10	0.68±0.06	0.56±0.11
LR	**+0.16±0.13**	0.69±0.03	0.58±0.05	0.77±0.05	0.62±0.06	0.37±0.15
SVM	**0.12±0.14**	0.67±0.05	0.58±0.06	0.73±0.09	0.61±0.06	0.38±0.14
k-NN	**+0.09±0.11**	0.62±0.08	0.56±0.05	0.63±0.14	0.61±0.05	0.46±0.14
GNB	**0.02±0.17**	0.68±0.03	0.55±0.05	0.85±0.07	0.57±0.05	0.17±0.15
F1-Score Ranking						
RF	+0.92±0.23	**0.96±0.09**	0.96±0.11	0.97±0.07	0.95±0.10	0.95±0.16
GBM	+0.25±0.14	**0.73±0.05**	0.64±0.05	0.86±0.10	0.64±0.04	0.35±0.13
LR	+0.16±0.13	**0.69±0.03**	0.58±0.05	0.77±0.05	0.62±0.06	0.37±0.15
GNB	0.02±0.17	**0.68±0.03**	0.55±0.05	0.85±0.07	0.57±0.05	0.17±0.15
DT	+0.23±0.11	**0.68±0.07**	0.62±0.05	0.69±0.14	0.66±0.06	0.53±0.17
SVM	0.12±0.14	**0.67±0.05**	0.58±0.06	0.73±0.09	0.61±0.06	0.38±0.14
XGB	+0.23±0.15	**0.67±0.07**	0.59±0.07	0.64±0.10	0.68±0.06	0.56±0.11
K-NN	+0.09±0.11	**0.62±0.08**	0.56±0.05	0.63±0.14	0.61±0.05	0.46±0.14
Accuracy Ranking:						
RF	+0.92±0.23	0.96±0.09	**0.96±0.11**	0.97±0.07	0.95±0.10	0.95±0.16
GBM	+0.25±0.14	0.73±0.05	**0.64±0.05**	0.86±0.10	0.64±0.04	0.35±0.13
DT	+0.23±0.11	0.68±0.07	**0.62±0.05**	0.69±0.14	0.66±0.06	0.53±0.17
XGB	+0.23±0.15	0.67±0.07	**0.59±0.07**	0.64±0.10	0.68±0.06	0.56±0.11
LR	+0.16±0.13	0.69±0.03	**0.58±0.05**	0.77±0.05	0.62±0.06	0.37±0.15
SVM	0.12±0.14	0.67±0.05	**0.58±0.06**	0.73±0.09	0.61±0.06	0.38±0.14
k-NN	+0.09±0.11	0.62±0.08	**0.56±0.05**	0.63±0.14	0.61±0.05	0.46±0.14
GNB	0.02±0.17	0.68±0.03	**0.55±0.05**	0.85±0.07	0.57±0.05	0.17±0.15

Source: Own elaboration.

The Age-Group classifiers' CMs of all proposed algorithms are shown in Figure 11 GBM, k-NN, RF, DT, LR, GNB, XGB, SVC). The majority of the algorithms achieved classification accuracy greater than 55%. Among all specified proposed algorithms, the RF classifiers' CM Figure 11(c) TP (Very Old Age-Group Ratio) and TN (Adult Age-Group ratio) rate are greater with 165 and 122, respectively. Further, the XGB (Figure 11(g))Adult Age-Group ratio represented by TN is (0.56) and the GBM (Figure 11(a))-TPR, which is representing here very old age-group (0.86) is higher followed by the RF classifier. A substantial number of adult patients (107) were misclassified as Very-Old in the GNB(Figure 11(f)) classifier. In contrast, a considerable number of Very-Old patients (62) were misclassified as Adults in the k-NN classifier in Figure 11(b). As previously stated, the RF classifier has an extremely low rate of misclassification in Table 11, with just (07) Adult patients misclassified as Very-Old and only (05) Very-Old patients misclassified as Adult. TPR and True Negative Rate (TNR) of RF are 97% and 95%. Therefore, RF is outperforming among proposed classifiers.

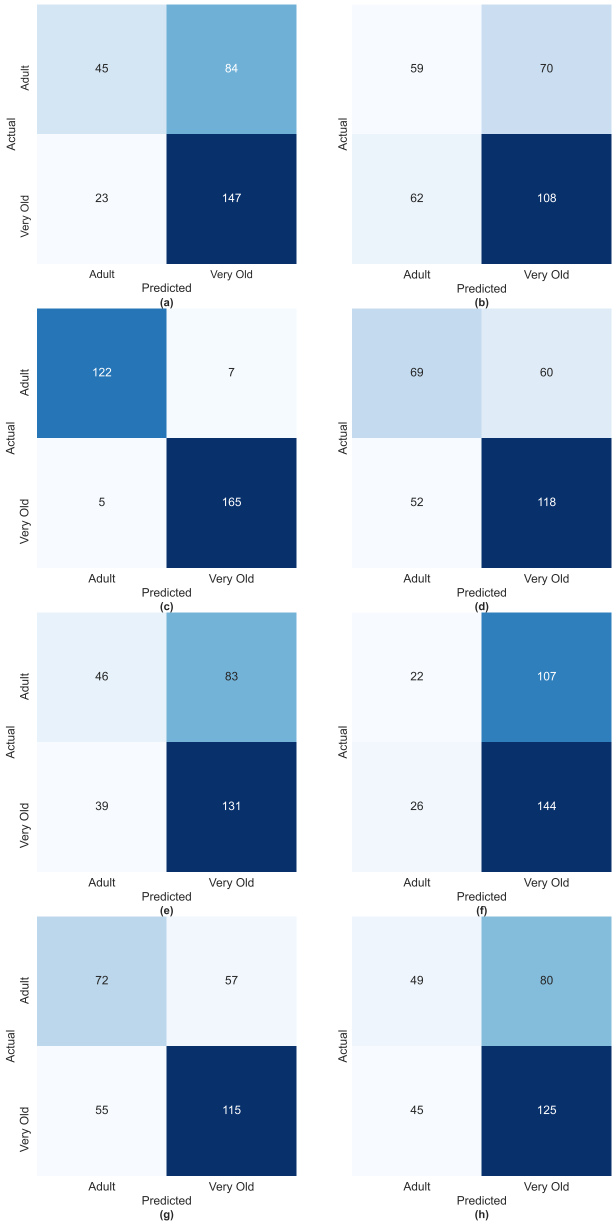


Figure 11. Age-group classifier's CMs (**a**) GBM, (**b**) k-NN,(**c**) RF, (**d**) DT, (**e**) LR, (**f**) GNB, (**g**) XGB, (**h**) SVM.

AUCROC score represents the capability of the model to distinguish among classes. From Figure 12(b), it can be clearly observed that RF (AUC =0.94) with 0.16 SD is best classifier followed by GNB (AUC = 0.67) and LR (AUC = 0.64) in Figure 12(f) and (g).

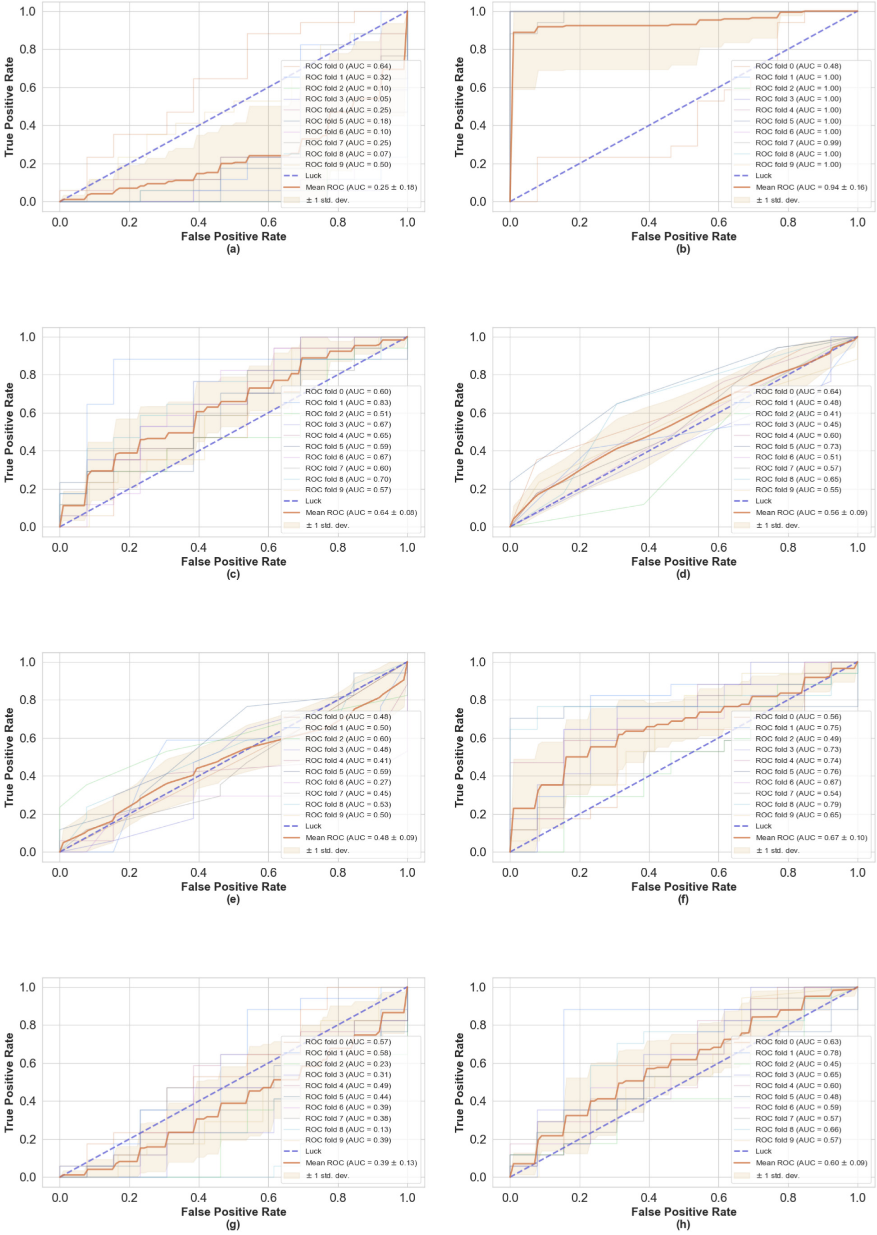


Figure 12. Age-Group classifier's ROC's at dynamic thresholds (**a**) XGB, (**b**) RF, (**c**) LR, (**d**) k-NN,(**e**) DT, (**f**) GNB, (**g**) GBM, (**h**) SVM.

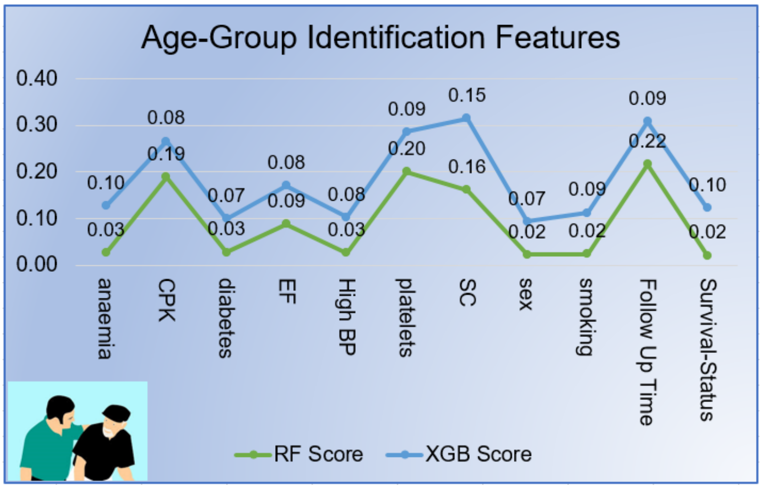


Figure 13. Patient Age-group Features.

Figure 13 displays the the most significant features that correctly identified the the age-group of patients. RF found highest coefficient of platelets(0.201), CPK(0.189), SC(0.161) and EF(0.088), diabetes(0.027) during Follow-up months(0.216) which are playing a significant role in predicting the Age-group as evident from Figure 13, it is also evident from visualization, the top significant features extracted by XGB classifier are SC (0.154), anemia(0.101), Survival-Status(0.101), smoking(0.088) and platelets (0.085) during follow-up months(0.093). Both classifier rank differently features to predict the Age-group.

10.7. Experiment-7

This experiment identified the survival (Survival-Status) of patients using various classifiers Table 14. In reality, RF classifier (MCC = +0.91) with ±0.11 SD is the best performing classifier in the MCC ranking, F1 score (0.94) with ±0.07 SD and accuracy (0.96) with ±0.06 SD rankings followed by DT in (MCC=+0.63) with ±0.11 SD as same to XGB (+0.63) but with high ±0.12 SD, F1-Score (0.75) with ±0.07 SD and accuracy (0.83) with ±0.05 SD. k-NN classifier performs worst among all specified algorithms by (MCC=+0.06) with ±0.16 SD, accuracy (0.61) with ±0.06 SD, but in F1-Score ranking, SVM-SVC is serving poor with 031 with ±0.11 SD. As previously mentioned, we tend to concentrate on the MCC rating for binary classifications like this because this rate only produces a high score if the classifier correctly predicted the most positive data instances and the majority of negative data instances. Indeed, the top MCC ranking classifier RF performed admirably on both recall (TP rate = 0.93) and specificity (TN rate = 0.98) with ±0.08 and ±0.07 SD, respectively. In conclusion, the F1 score and accuracy rankings hide a fundamental error in the top classifier: k-NN could not predict a large percentage of patients correctly. The MCC rating, on the other hand, takes this knowledge into account. However, accuracy values will deceive the researcher once more: a closer examination of the results reveals that the radial SVM performed poorly on the true positive (TP rate = 0.13), correctly observing fewer patients.

Table 14. Survival-Status Classification Performance Metrics.

Classifier	MCC	F1-Score	Accuracy	Recall (TPR)	Precision	TNR
MCC Ranking:						
RF	**+0.91±0.11**	0.94±0.07	0.96±0.05	0.93±0.08	0.95±0.08	0.97±0.05
DT	**+0.63±0.11**	0.75±0.07	0.83±0.05	0.80±0.10	0.71±0.09	0.85±0.06
XGB	**+0.63±0.12**	0.74±0.10	0.84±0.05	0.72±0.17	0.77±0.10	0.90±0.05
LR	**+0.59±0.10**	0.71±0.08	0.82±0.04	0.66±0.11	0.77±0.09	0.91±0.04
GBM	**+0.59±0.14**	0.72±0.10	0.83±0.05	0.69±0.15	0.75±0.12	0.89±0.06
GNB	**+0.53±0.17**	0.64±0.15	0.81±0.06	0.54±0.17	0.79±0.14	0.93±0.05
SVM	**+0.13±0.14**	0.21±0.11	0.68±0.03	0.13±0.08	0.52±0.24	0.94±0.03
k-NN	**+0.06±0.16**	0.33±0.12	0.61±0.06	0.30±0.12	0.37±0.14	0.77±0.07
F1-Score Ranking:						
RF	+0.91±0.11	**0.94±0.07**	0.96±0.05	0.93±0.08	0.95±0.08	0.97±0.05
DT	+0.63±0.11	**0.75±0.07**	0.83±0.05	0.80±0.10	0.71±0.09	0.85±0.06
XGB	+0.63±0.12	**0.74±0.10**	0.84±0.05	0.72±0.17	0.77±0.10	0.90±0.05
GBM	+0.59±0.14	**0.72±0.10**	0.83±0.05	0.69±0.15	0.75±0.12	0.89±0.06
LR	+0.59±0.10	**0.71±0.08**	0.82±0.04	0.66±0.11	0.77±0.09	0.91±0.04
GNB	+0.53±0.17	**0.64±0.15**	0.81±0.06	0.54±0.17	0.79±0.14	0.93±0.05
k-NN	+0.06±0.16	**0.33±0.12**	0.61±0.06	0.30±0.12	0.37±0.14	0.77±0.07
SVM	+0.13±0.14	**0.21±0.11**	0.68±0.03	0.13±0.08	0.52±0.24	0.94±0.03
Accuracy Ranking:						
RF	+0.91±0.11	0.94±0.07	**0.96±0.05**	0.93±0.08	0.95±0.08	0.97±0.05
XGB	+0.63±0.12	0.74±0.10	**0.84±0.05**	0.72±0.17	0.77±0.10	0.90±0.05
DT	+0.63±0.11	0.75±0.07	**0.83±0.05**	0.80±0.10	0.71±0.09	0.85±0.06
GBM	+0.59±0.14	0.72±0.10	**0.83±0.05**	0.69±0.15	0.75±0.12	0.89±0.06
LR	+0.59±0.10	0.71±0.08	**0.82±0.04**	0.66±0.11	0.77±0.09	0.91±0.04
GNB	+0.53±0.17	0.64±0.15	**0.81±0.06**	0.54±0.17	0.79±0.14	0.93±0.05
SVM	+0.13±0.14	0.21±0.11	**0.68±0.03**	0.13±0.08	0.52±0.24	0.94±0.03
k-NN	+0.06±0.16	0.33±0.12	**0.61±0.06**	0.30±0.12	0.37±0.14	0.77±0.07

Source: Own elaboration.

Figure 14 representing all CMs of Survival-Status classification algorithms (GBM, k-NN, RF, DT, LR, GNB, XGB, SVM). As observed from Table 14, the least accuracy (61%) was achieved by the k-NN algorithm. Among all specified proposed algorithms, the RF classifiers' CM displaying remarkable results. Here, TP represents the total number of right-predicted Dead patients, and TN represents the total number of right-predicted alive patients. Figure 14(c), The RF has achieved a greater number of TP and TN with 89 and 198, respectively. Further, the second-highest TN rate is achieved by the LR algorithm (Figure 14(e)), in which TN is 184, and also 2nd higher TP rate is achieved by the DT algorithm (Figure 14(d))-in which TPR can be noted down as 77 followed by the RF classifier TN and TP rate. A substantial number of alive patients (31) are misclassified as dead in the k-NN(Figure 14(b)) classifier, and a considerable number of patients killed (67) are misclassified as alive. As previously stated, the RF classifier has a shallow rate of misclassification in Table 14, the benchmark also followed in CM with only (07) dead patients misclassified as alive and only (05) alive patients misclassified as dead. As observed, TPR and TNR of RF are 93%, and 97% is higher. Therefore, RF can be considered as the best performing classifiers.

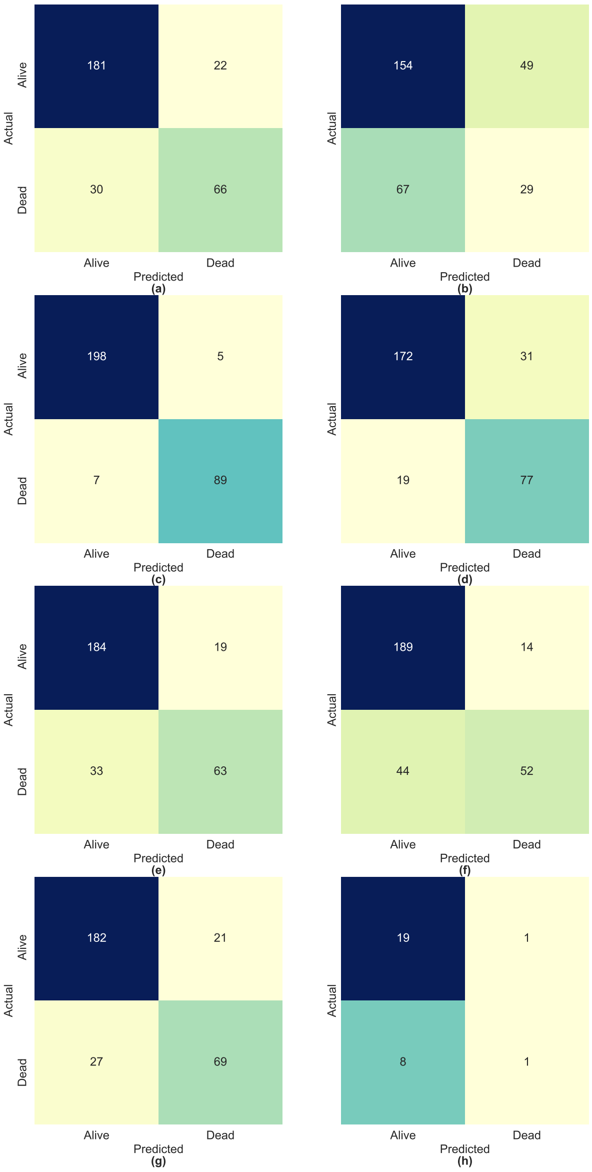


Figure 14. Survival Status classifier's CMs (**a**) GBM, (**b**) k-NN,(**c**) RF, (**d**) DT, (**e**) LR, (**f**) GNB, (**g**) XGB, (**h**) SVM.

In Figure 15, the ROC curve represents the TPR plot against the FPR at various threshold values. As we know, the classifier's superiority can also be measured by a larger AUC. The RF's best mean AUC (Figure 15(b) can be noted down as 0.97 with SD of ±0.08 followed the AUC for LR (Figure 9(c) is 0.96 with SD ±0.05. The worst mean k-NN AUC (Figure 15(d)) is 0.50 with a SD of 0.06, followed by next higher algorithm ,DT AUC (Figure 15(e)) 0.77 with a SD of 0.20.

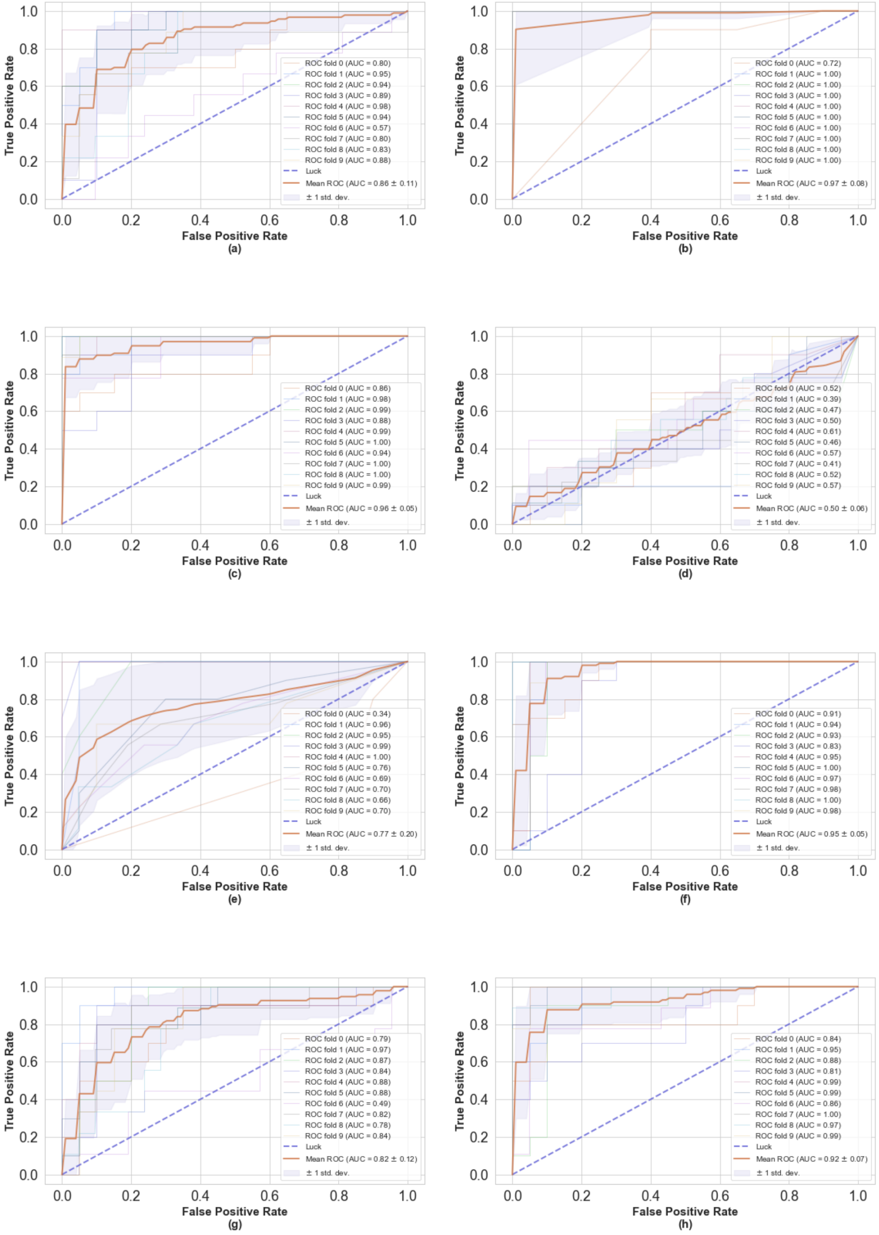


Figure 15. Survival Status classifier's ROC's at dynamic thresholds (**a**) XGB, (**b**) RF, (**c**) LR, (**d**) k-NN,(**e**) DT, (**f**) GNB, (**g**) GBM, (**h**) SVM.

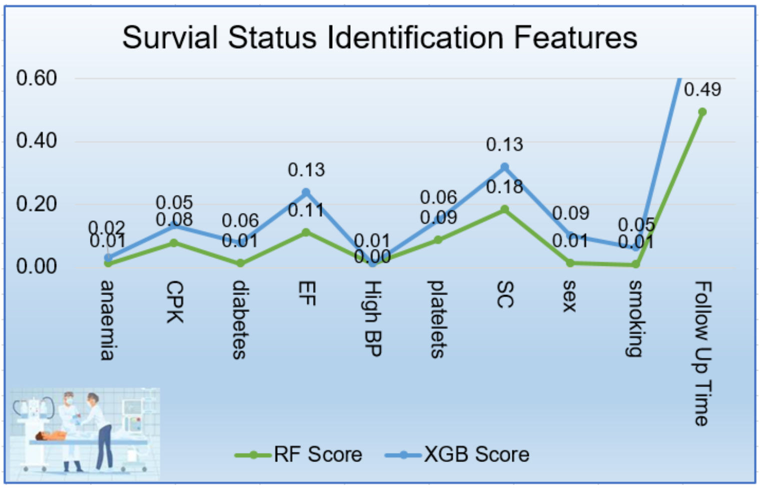


Figure 16. Patient Survival Features.

Figure 16 displays the most significant features that correctly identified the Survival-Status of the patients. RF found highest coefficient of SC(0.182), EF(0.110), platelets(0.086), CPK(0.077), gender(0.012) during Follow-up months(0.492) which are playing a significant role in predicting the Survival-Status, it is also evident from visualization, the top significant features extracted by XGB classifier are SC (0.134), EF(0.127), gender(0.087), diabetes(0.065) and platelets (0.064) during follow-up months (0.396).Both classifiers considering SC and EF as significant features for predicting the Survival-Status, but XGB treating gender at 3rd rank and RF treating at 5th rank.

11. Discussion

The significant p-value is critical for evaluating the hypothesis in statistical tests. To test the first two hypotheses, a significant Mann-Whitney U test is used in this study. It is playing a vital role due to the absence of normality in data. The present paper investigated the impact of SC and SS on Survival-Status levels, as well as it also verified the complications such as anemia, diabetes, and high BP impact of SC and SS; on the other hand, this also validated the association between gender and smoking habit among CVD patients, as well as the association between age-group and survival-status. In addition to this, an association χ^2 test is performed to investigate the relationship between gender and smoking habits. It also demonstrates the association between age-group and Survival-Status which is verified with a cox-regression model.

The first null hypothesis, "H_{01}: No significant difference in Alive and Dead towards SC and SS," is found to be rejected ($p < 0.05$). There are statistically significant differences found between the SC and SS w.r.t Survival-Status levels. The second null hypotheses "H_{02a}: No significant differences between non-anemic and anemic levels towards SC and SS", "H_{02b}: No significant differences between non-diabetic and diabetic levels towards SC and SS", and "H02c: No significant differences between non-BP and BP levels towards SC and SS" are all not found significant ($p < 0.05$). Furthermore, the gender of the patients and their smoking habits are found to be statistically significantly associated ($p < 0.05$). As a result, the third null hypothesis, "H_{03}: No significant relationship between gender and smoking level," is found rejected. During the study, it is found that actual Female Smoker patients (04) were significantly less than expected (33.7) as compared to non-smoker female patients. It is also observed that actual male smoker patients(92) were substantially more significant than expected(62.3) but not the same in male non-smoker patients. An actual male who is not habitual to smoking habits (102) was found significantly less than expected(131.70) during follow-up months. The present study also uses a cox regression model to explore the association between age-group and survival-status levels and demonstrate a statistically significant association between these two attributes. Therefore

"H_{04}: No significant association among age-groups and Survival-Status level" found to rejected ($p < 0.05$).

The paper's findings are self-evident: H01, the Survival-Status is linearly correlated with SC and SS (p < 0.05), supporting [12] [13] [14] [16]. In the case of (H_{02a}, H_{02b}, and H_{02c}) complications such as anemia, diabetes, and high blood pressure, the influence SC and SS were not found significant ($p > 0.05$), contradicting [10,17], and [19] but supporting [14]. The finding of the gender-smoking-habits association (H03) found significant ($p < 0.05$) support [21]. The age-group Survival-Status association (H_{04}) finding is significant ($p < 0.05$), rejected [24] but supported [15].

The reported CDF-DI was performed on a heart disease dataset and showed promising results compared to previous models in improving prediction accuracy. For comparison, we used eight state-of-the-art MLAs (GNB, LR, GBM, SVM, DT, XGB, k-NN, and RF) throughout the study that have an established track record for accuracy and efficiency in the research community. All models were subjected to 10-fold cross-validation, and six performance metrics were collected: accuracy, precision, TPR, F1-measure, MCC, TNR. RF classifier was found superior in all mentioned performance ranking (MCC score ranking, F1-Score ranking, and accuracy score ranking) in all machine learning prediction goals.

The proposed model outperformed with RF machine learning model in predicting patients' gender another model by obtaining accuracy up to 94%, and 95% in all rest performance metrics, i.e., precision, TPR, and F1-Score, respectively. The proposed CDF-DI model has the highest MCC values, up to 0.87, proving its superiority over other models. Furthermore, the proposed model had the lowest FPR and the highest TNR by up to 9% and 91%, respectively. The suggested model's low FPR and high TNR values demonstrate the CDF-DI model's capacity to reduce miss rates and improve prediction accuracy for both negative and positive subjects. Table 12 displays the comprehensive performance findings for predicting patients 'gender. GNB found the worst performer in MCC, F1, and Accuracy score ranking with 0.06, 71%, and 59% respectively.

Further, the predicting age group of patients has displayed encouraging results with the RF model. The proposed model was found best with RF in all key performance criteria, such as precision, TPR, F1-Score, and accuracy by up to 95%, 97%,96%, and 96%, respectively. The RF's most significant MCC values, up to 0.92, are found in the CDF-DI model, demonstrating its superiority over the rest proposed MLAs. Furthermore, the proposed model exhibited the lowest FPR and the greatest TNR by up to 5% and 95%, respectively. GNB model found the worst performing model in terms of MCC ranking and accuracy ranking with 0.02 and 55% scores, respectively. The k-NN model was found worst in F1-Score during the prediction of age-group of patients with 62% scores. The entire performance findings are shown in Table 15.

Table 15. Benchmark with previous study results.

CDF-DI(Extant Research)						Davide Chicco et. al [30]					
Model	Accuracy	F1-Score	MCC	Sen.	Spec.	Existing Accuracy	Existing F1-Score	Existing MCC	Sen.	Spec.	Extant Accuracy
RF	**0.96**	0.94	0.91	0.93	0.97	0.74	**0.55**	0.38	0.49	0.86	0.22 ↑
DT	**0.83**	0.75	0.63	0.80	0.97	0.74	**0.55**	0.38	0.53	0.83	0.09 ↑
GBM	**0.83**	0.72	0.59	0.69	0.89	0.74	**0.53**	0.37	0.48	0.86	0.09 ↑
LR	**0.82**	0.71	0.59	0.66	0.91	0.73	**0.47**	0.33	0.39	0.89	0.09 ↑
GNB	**0.81**	0.64	0.53	0.54	0.93	0.70	**0.36**	0.22	0.28	0.90	0.11 ↑
SVM	**0.68**	0.21	0.13	0.13	0.94	0.69	**0.18**	0.16	0.12	0.97	-0.01 ↓
k-NN	**0.61**	0.33	0.06	0.30	0.77	0.62	**0.15**	-0.02	0.12	0.87	-0.01 ↓

Source: Own elaboration.

Furthermore, the RF model has also shown promising results in predicting the Survival-Status of patients. All six significant performance measures, such as precision, TPR, F1-Score, accuracy, MCC, and TNR, are found to be better with RF by up to 95%, 94%, 96%, 0.91, and 97%, respectively, in the CDF-DI model. the k-NN model proved w.r.t. MCC and accuracy score ranking and in F1-Score ranking SVM model found weak model. Table 14 summarizes all of the performance findings.

This is incredibly encouraging for hospital settings: even if several laboratory test data and health conditions were absent from a patient's electronic health record, doctors could still predict patient survival by evaluating the EF, SC, and CPK values alone. The present research also yielded several intriguing outcomes that varied from the findings of the same dataset study [37]. Davide Chicco et al. identified EF, SC, age, CPK, and gender chosen as the top five features for predicting Survival-Status while Tanvir Ahmad et. al [16] also identified age, SC, High BP, EF, and anemia as top essential features. This study found SC, EF, platelets, CPK, and gender as important features which play an essential role in predicting Survival-Status with RF Classifier as depicted in Figure 16. We found EF at 2nd position and also found platelets as an essential feature which the previous study had not found. The present paper also improves the accuracy, F1-Score, and MCC by 0.22, 0.39, and 0.53 respectively in RF classifier and other models as depicted in Table 15.

The experiment results displayed that the supervised machine learning model performed a best role in predicting the age group and gender of heart failure patients very efficiently. Tree-based algorithms performed well on the imbalanced dataset using the 10-fold cross-validation method. As displayed in Figure 13, RF identified CPK, SC, follow-up month, platelets, and EF as significant features while predicting the age group (adult, very old) of patients. Also, the RF classifier identified smoking, CPK, platelets, follow-up month, and SC. These methods became beneficial in-patient care because doctors can predict a patient's age group based on only five significant characteristics. RF and XGB commonly extracted out SC and EF as crucial features in predicting the age-group target variable. Also, smoking, CPK, and platelets are found vital features in predicting the gender of patients commonly finding RF and XGB.

As it can be derived very quickly the top five input features which are playing vital role in predicting Survival-Status from Figure 16. The top 5 features selected by RF and XGB, feature selection techniques, are follow-up month, SC, EF, CPK, and anemia. The RF feature selector and the XGB feature selection have a lot in common. Follow-up month has the highest ranking (0.49), whereas anemia has a lower score (0.01) extracted by the RF feature selector. XGB also treats the follow-up month feature as the most important (0.40) and lowest rank to the anemia (0.02). As finding suggests that patients' SC, EF, platelets, and CPK needs special attention for their survival during their follow-up months.

12. Conclusion

The present research d confirmed that our traditional biostatic analysis finding signifies the importance of an appropriate level of sodium and normal creatinine level in the human body. The paper results revealed the significance of SC and SS towards the Survival-Status (Alive/Dead) of CVD patients. Also, it validated that health complications like anemia level, High BP level, and diabetic level have no significant effect on the SC and SS enzyme levels. This paper also found the significant association of smoking habits with specific patients towards gender. On the one hand, the authors found that the real count of non-smoker and smoker females is lower than the expected count. On the other hand, the actual male smoker and non-smoker female found significantly more significant than the expected count. The study also observed that the elderly patients (very old age-group) are more susceptible to HD mortality. Figure 7 confirmed that the patients who belong to the very old age-group are more mortality prone than adult patients. Further, the authors applied eight machine learning algorithms to identify the gender, age-group, and Survival-Status of the patients with improved accuracy as compared to the previous study.

The smoking, platelets, CPK, SC, and EF are found the most prominent predicting features. In addition to the earlier study's features EF, and SC [14,31], the authors recommended three more features: platelets, CPK, and gender to identify the Survival-Status of the patient.

The modest size of the dataset (299 patients) is a constraint of this study; a more extensive dataset would have allowed us to achieve more reliable results. Other information on the patients' physical characteristics (height, weight, BMI, hormones, etc.) and their work history might have helped detect additional risk factors for CVD.

Future work may include the principal component analysis [65], to transform the existing features to enhance the classification accuracy and apply the statistical algorithms to prove the comparative strength of the results [66]. Moreover, the real-time implementation of the extant research would support the medical support systems and helpful to the doctors for examining the cardiac patients. A novel diagnostic system can also be designed and developed for the IoT-enabled CDF-DI models.

Author Contributions: Conceptualization, D.K., C.V., Data curation, D.K.,C.V.,S.D., Methodology, D.K.,C.V., Formal analysis, D.K., Investigation, D.K., Resources, D.K., visualization, D.K., C.V., S.D., Validation, D.K., C.V., S.D., P.K.S., M.S.R., Writing–original draft preparation, D.K., writing–review and editing, D.K., C.V., S.D., P.K.S., M.S.R, Project Administration, C.V., Funding acquisition, C.V., M.S.R. All authors have read and agreed to the published version of the manuscript.

Funding: This research was not funded by any agency or institution.

Acknowledgments: The second author thank to Dr. Zoltán Illés (Eötvös Loránd University, Hungary) and Dr. Veronika Stoffová (Trnava University, Slovakia) to motivate him to contribute in this research.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CVD	CardioVascular Diseases
HF	Heart Failures
HD	Heart Disease
SC	Serum Creatinine
SS	Serum Sodium
EF	Ejection Fraction
CPK	Creatinine Phosphokinase
BP	Blood-Pressure
DT	Decision Tree
RF	Random Forest
SD	Standard Deviation
SVM	Support Vector Machine
KNN	k-Nearest Neighbors
XGB	eXtreme Gradient Boosting
GBM	Gradient Boosting Machines
LR	Logistic Regression
LNR	Linear Regression
AUC	Area Under Curve
VIF	Variation Inflation Factor
GNB	Gaussian Naive Bayes
Mdn	Median
TPR	True Positive Rate
TNR	True Negative Rate
CMs	Confusion matrics

References

1. WHO, Fact sheet on CVDs. Global Hearts. World Health Organization. 2016. Available online: https://www.who.int/health-topics/cardiovascular-diseases (accessed on 23 May 2021).
2. Fryar, CD.; Chen, TC.; Li, X. Prevalence of uncontrolled risk factors for cardiovascular disease: United States, 1999-2010. *NCHS Data Brief* **2012**, *103*, 1–8.
3. Medical Professionals. Cardiovascular Diseases. Available online: https://www.mayoclinic.org/medical-professionals/cardiovascular-diseases (accessed on 23 May 2021).
4. Allen, LA; Stevenson, LW; Grady, KL.; Goldstein, NE.; Matlock, DD.; Arnold, RM.; Cook NR.; Felker, GM.; Francis, GS.; Hauptman, PJ.;, Havranek, EP.; Krumholz, HM.; Mancini, D.; Riegel, B.; Spertus, JA. Decision making in advanced heart failure: a scientific statement from the American Heart Association. *American Heart Association, Council on Quality of Care and Outcomes Research, Council on Cardiovascular Nursing, Council on Clinical Cardiology, Council on Cardiovascular Radiology and Intervention, . . . Council on Cardiovascular Surgery and Anesthesia* **2012**, *125(15)*, 1928–1952.
5. Arabasadi, Z; Alizadehsani, R.; Roshanzamir, M.; Moosaei, H.; Yarifard, AA. Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm*Comput Methods Programs Biomed* , **2017**, *141*, 19–26.
6. Anbuselvan,P. Heart Disease Prediction using Machine Learning Techniques., *IJERT* , **Nov. 2020**, *9(11)*.
7. Curtis, AB.;Karki, R.; Hattoum, A.; Sharma. UC. Arrhythmias in Patients >=80 Years of Age: Pathophysiology, Management, and Outcomes, *J Am Coll Cardiol.*, **2018**, *71(18)*.
8. North, BJ.; Sinclair, DA. The intersection between aging and cardiovascular disease, *Circ Res.* , **2012**, *110(8)*,1097–1108.
9. Yazdanyar, A.; Newman, AB. The burden of cardiovascular disease in the elderly: morbidity, mortality, and costs *Clin Geriatr Med.* , **2009**, *25(4)*,563–577.
10. He, FJ.; Li J.; Macgregor, GA. Effect of longer-term modest salt reduction on blood pressure. *Cochrane Database Syst Rev.*, **2013**, *4*.
11. Effects of weight loss and sodium reduction intervention on blood pressure and hypertension incidence in overweight people with high-normal blood pressure. The Trials of Hypertension Prevention, phase II. The Trials of Hypertension Prevention Collaborative Research Group *Arch Intern Med.*, **1997**, *157(6)*,657–667.
12. Cook, NR.; Cutler, JA.; Obarzanek, E. et al. Long term effects of dietary sodium reduction on cardiovascular disease outcomes: observational follow-up of the trials of hypertension prevention (TOHP). *BMJ*, **2007**, *334(7599)*,885–888.
13. Patel, Y.; Joseph, J. Sodium Intake and Heart Failure.. *International Journal of Molecular Sciences*, **2020**, *21(24)*.
14. Ahmad, T.; Munir, A.; Bhatti, SH., Aftab M, Raza MA., Survival analysis of heart failure patients: A case study. *PLOS ONE*, **2017**, *12(07)*.
15. Dugani, SB; Moorthy, MV.; Li, C. et al. Association of Lipid, Inflammatory, and Metabolic Biomarkers With Age at Onset for Incident Coronary Heart Disease in Women. *JAMA Cardiol.*, **2021**, *06(04)*,443–447.
16. Akhter, MW.; Aronson, D.; Bitar, F. et al. Effect of elevated admission serum creatinine and its worsening on outcome in hospitalized patients with decompensated heart failure., *Am J Cardiol*, **2004**, *94(07)*,957–960.
17. Grillo, A.; Salvi, L.; Coruzzi, P., Salvi, P., Parati, G. Sodium Intake and Hypertension., *Nutrients.*, **2019**, *11(09)*.
18. Kurtz, TW; DiCarlo, SE.; Pravenec, M., Morris, RC Jr. The American Heart Association Scientific Statement on salt sensitivity of blood pressure: Prompting consideration of alternative conceptual frameworks for the pathogenesis of salt sensitivity? *J Hypertens.*, **2017**, *35(11), 2214–2225.*
19. High Creatinine Levels: Causes, Symptoms, and When to Seek Help., Available online: https://www.medicalnewstoday.com/articles/whentoworryaboutcreatininelevels#symptoms. (accessed on 23 May 2021).
20. Abebe, T.B., Gebreyohannes, E.A., Bhagavathula, A.S. et al. Anemia in severe heart failure patients: does it predict prognosis?. BMC Cardiovasc Disord 17, 248 (2017).

21. Roy, A.; Rawal,I; Jabbour, S; Prabhakaran.D. Tobacco and Cardiovascular Disease: A Summary of Evidence. In *Disease Control Priorities, Third Edition (Volume 5): Cardiovascular, Respiratory, and Related Disorders*; Prabhakaran. D., Anand S., Thomas A. Gaziano, Jean-Claude Mbanya, Yangfeng Wu, Rachel Nugent, Eds.; Publishing House: The World Bank, 2017; pp. 57–77.
22. Huxley, RR; Woodward, M. Cigarette smoking as a risk factor for coronary heart disease in women compared with men: a systematic review and meta-analysis of prospective cohort studies.*J Hypertens.*, **2017**, *35(11),2214–2225.*
23. Rahman, M.; Rashid, S. M. ; Ferdous Khan; M. Nayem; Biswas, A. ; Mahmud A., Symptom Wise Age Prediction of Cancer Patients using Classifier Comparison and Feature Selection. In 22nd International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, Country, 18-20 Dec. 2019; IEEE.
24. Vaughan, AS; Ritchey, MD.; Hannan, J; Kramer, MR.; Casper, M. Widespread recent increases in county-level heart disease mortality across age groups..*Ann Epidemiol.*, **2017**, *27(12),796–800.*
25. Horvath, S., Gurven, M., Levine, M.E. et al. An epigenetic clock analysis of race/ethnicity, sex, and coronary heart disease. *Genome Biol*, **2016**, *17(171),01–23.*
26. Rodgers, JL.; Jones, J.; Bolleddu, SI. et al. Cardiovascular Risks Associated with Gender and Aging. *J Cardiovasc Dev Dis.*, **2019**, *;6(2), 01—23.*
27. Benjamin, EJ; Muntner, P; Alonso, A. Heart Disease and Stroke Statistics-2019 Update: A Report From the American Heart Association.*Circulation*, **2019**, *;139(10), e56—e528.*
28. Villa, A.; Rizzi, N.; Vegeto, E. Estrogen accelerates the resolution of inflammation in macrophagic cells.*Sci Rep* , **2015**, *; 5, 15224.*
29. Korot, E.; Pontikos, N.;Liu, X. Predicting sex from retinal fundus photographs using automated deep learning. *Sci Rep*, **2021**, *11, 10286.*
30. Li, J.P.; Haq, A. U.; Din, S. U.; Khan, J.; Khan A.;Saboor,A. Heart Disease Identification Method Using Machine Learning Classification in E-Healthcare,*IEEE Access* , **2020**, *8, 107562-107582.*
31. 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* .
32. Ishaq, A.; Sadiq, S.; Umer, M.; Ullah, S.; Mirjalili, S.; Rupapara, V.; Nappi, M. Improving the Prediction of Heart Failure Patients' Survival Using SMOTE and Effective Data Mining Techniques,*IEEE Access* ,**2021**, *9, 39707-39716.*
33. Breiman, L.; Jerome H.; Friedman; Richard, A.; Olshen; Charles, J. Stone. Construction of Tree from a Learning Sample. In *Classification and Regression Trees, First Edition*; Publishing House: Taylor Francis Group, 1984; pp. 21–23.
34. Breiman, L. Random Forests,*Machine Learning*,**2001**, *45, 5–32* .
35. Boyd, CR; Tolson, MA; Copes, WS. Evaluating trauma care: the TRISS method. Trauma Score and the Injury Severity Score, *J. Trauma, Injury, Infection, Crit. Care*,**1987**, *27(4), 370-378.*
36. Friedman, J. Greedy Function Approximation: A Gradient Boosting Machine.*J. The Annals of Statistics*,**2001**, *29(5), 1189-1232.*
37. Pérez, A.; Larrañaga, P.;Inza, I. Supervised classification with conditional Gaussian networks: Increasing the structure complexity from naive Bayes.*J. International Journal of Approximate Reasoning* ,**2006**, *43(1), 1-25.*
38. Chapelle, O.; Schölkopf, B. Incorporating invariances in nonlinear Support Vector Machines, NIPS'01: Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic, MA United States, Jan. 2001; MIT Press: Cambridge, MA, USA, 2001.
39. Gardner, W.A. Learning characteristics of stochastic-gradient-descent algorithms: A general study, analysis, and critique.*Signal Processing* ,**1984**, *06(02), 113-133.*
40. Sharaff A., Gupta H. Extra-Tree Classifier with Metaheuristics Approach for Email Classification. In: Bhatia S., Tiwari S., Mishra K., Trivedi M. (eds) Advances in Computer Communication and Computational Sciences., Advances in Intelligent Systems and Computing, vol 924. Springer, Singapore, 2019.
41. Sklearn.Linearmodel.Perceptron — Scikit-Learn 0.24.1 Documentation.Available online: . https://scikitlearn.org/stable/modules/generated/sklearn. linearmodel.Perceptron.html. (accessed on 24 May 2021).

doi:10.20944/preprints202106.0201.v1

42. Perceptron - Wikipedia. Available online: https://en.wikipedia.org /wiki/Perceptron. (accessed on 24 May 2021).
43. Chen, M.; Liu, Q. ; Chen, S.; Liu, Y. ; Zhang, C.; Liu, R. XGBoost-Based Algorithm Interpretation and Application on Post-Fault Transient Stability Status Prediction of Power System.*IEEE Access*,**2019**, *07, 13149-13158*.
44. Chen, M.; Liu, Q. ; Chen, S.; Liu, Y. ; Zhang, C.; Liu, R. XGBoost-Based Algorithm Interpretation and Application on Post-Fault Transient Stability Status Prediction of Power System.*IEEE Access*,**2019**, *07, 13149-13158*.
45. Pawlovsky, A. P; An ensemble based on distances for a k-NN method for heart disease diagnosis. In: 2018 International Conference on Electronics, Information, and Communication (ICEIC), Honolulu, HI, USA, 2018.
46. Asuncion, A.; Newman, D. UCI machine learning repository, Tech. Rep., 2007. Available online: https://ergodicity.net/2013/07/ (accessed on 24 May 2021).
49. Padmanabhan, M.; Yuan, P.; Chada, G.; Nguyen, HV. Physician-Friendly Machine Learning: A Case Study with Cardiovascular Disease Risk Prediction. *J Clin Med.*,**2019**, *08(07), 1050*.
48. Makki,S. An Efficient Classification Model for Analyzing Skewed Data to Detect Frauds in the Financial Sector. Ph.D., Université de Lyon; Université libanaise, 2019. English. ⟨NNT : 2019LYSE1339⟩,Dec. 2019.
49. Chicco, D.; Jurman, G. The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*,**2020**, *21(6), 01-13* .
50. Jurman; Giuseppe; Riccadonna; Samantha; Furlanello; Cesare. A Comparison of MCC and CEN Error Measures in Multi-Class Prediction. *PLOS ONE*,**2012**, *07(08), 01-08* .
51. Sklearn.Ensemble.RandomForestClassifier-Scikit−Learn 0.24.1 Documentation. Available online: https:// scikit-learn.org/stable/modules/ generated/sklearn.ensemble.RandomForestClassifier.html. (accessed on 24 May 2021).
52. Support Vector Machines−Scikit−Learn 0.24.2 Documentation. Available online: https://scikit−learn.org/stable/modules/svm.html. (accessed on 24 May 2021).
53. Sklearn. Ensemble. Gradient Boosting Classifier−Scikit−Learn 0.24.2 Documentation. Available online: https://scikit−learn.org/ stable/modules/generated/sklearn. ensemble. GradientBoostingClassifier.html. (accessed on 24 May 2021).
54. Sklearn.Linearmodel.SGDClassifier−Scikit−Learn 0.24.2 Documentation. Available online: https://scikit-learn.org /stable /modules /generated /sklearn.linear model. SGDClassifier.html. (accessed on 24 May 2021).
55. Chen,T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In: KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining San Francisco California USA August, 2016.
56. Sklearn.Linearmodel.Logistic Regression-Scikit-Learn 0.24.2 Documentation. Available online: https://scikit-learn.org (accessed on 24 May 2021).
57. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; Vanderplas, J.; Passos, A.; Cournapeau, D.; Brucher, M.; Perrot, M. ; Duchesnay, E. Scikit-learn: Machine Learning in Python.*Journal of Machine Learning Research* ,**2012**, *07(08), 01-08* .
58. Ahmed; Hosameldin; Nandi; Asoke, K. Compressive Sampling and Feature Ranking Framework for Bearing Fault Classification With Vibration Signals*IEEE ACCESS*,**2018**, *06, 44731-44776* .
59. Gu, Q; Li, Z; Han, J. Generalized Fisher score for feature selection. In UAI'11: Proceedings of the Twenty-Seventh Conference on Uncertainty in Artificial Intelligence, July 2011, AUAI Press1911 N. Fort Meyer Drive, Suite 600ArlingtonVirginiaUnited States.
60. Yang,Y.; Pedersen, J. A comparative study on feature selection in text categorization, In ICML '97: Proceedings of the Fourteenth International Conference on Machine LearningJuly 1997, Morgan Kaufmann Publishers Inc.340 Pine Street, Sixth FloorSan, Francisco, CA, United States.
61. Mathworks.com. (2017). Cross-Tabulation—MATLAB Crosstab. Available online: https://uk.mathworks.com/help/stats/crosstab.html. (accessed on 24 May 2021).
62. Chicco; Davide; Rovelli; Cristina. Computational prediction of diagnosis and feature selection on mesothelioma patient health records, *PLOS ONE*, **2019**, *14(1), 01-28*.

63. Breiman, L; Cutler, A. Random forests Gini importance. Available online: https://www.stat.berkeley.edu / breiman/RandomForests/cchome.htmginiimp. (accessed on 24 May 2021).

64. Kumar, D.; Verma, C.; Singh, P.K.; Raboaca, M.; Felseghi, R.A.; Ghafoor, K.Z.; Computational Statistics and Machine Learning Techniques for Effective Decision Making on Student's Employment for Real-Time. *Mathematics* **2021**, *9*, *21*, 1–29.

65. Verma, C.; Zoltán, I.; Veronika, S.; Tanwar, S.; Kumar, N. Machine Learning-Based Student's Native Place Identification for Real-Time. *IEEE Access* **2020**, *8*, 130840–130854.

66. Verma, C.; Veronika, S.; Zoltán, I. Prediction of students' awareness level towards ICT and mobile technology in Indian and Hungarian University for the real-time: preliminary results. *Heliyon* **2019**, *5*, *6* 6321.