Journal of Electronics, Electromedical Engineering, and Medical Informatics
Multidisciplinary : Rapid Review : Open Access Journal **ISSN: 2656-8632**

RESEARCH ARTICLE

OPEN ACCESS

Manuscript received June 21, 2022; revised July 23, 2022; accepted July 25, 2022; date of publication July 29, 2022
Digital Object Identifier (DOI): https://doi.org/10.35882/jeeemi.v4i3.225

A Comparative Study for Time-to-Event Analysis and Survival Prediction for Heart Failure Condition using Machine Learning Techniques

Saurav Mishra

Corresponding author: Saurav Mishra (e-mail: Saurav.Mishra@live.com).

I would want to thank and express my gratefulness to the researchers at the Institute of Cardiology and Allied Hospital, Faisalabad for creating and publishing the heart failure dataset publicly accessible for research activities.

ABSTRACT Heart Failure, an ailment in which the heart isn't functioning as effectively as it should, causing in an insufficient cardiac output. The effectual functioning of the human body is dependent on how well the heart is able to pump oxygenated, and nutrient rich blood to the tissues and cells. Heart failure falls into the category of cardiovascular diseases - the disorders of the heart and blood vessels. One of the leading causes of global deaths resulting in an estimated 17.9 million deaths globally every year. The condition of heart failure results out of structural changes to the cardiac muscles majorly in the left ventricle. The weakened muscles cause the ventricle to lose its ability to contract completely. Since the left ventricle generates the required pressure for blood circulation, any kind of a failure condition results in the reduction of cardiac power output. This study aims to conduct a thorough survival analysis and survival prediction on the data of 299 patients classified into the class III/IV of heart failure and diagnosed with left ventricular systolic dysfunction. Survival analysis involves the study of the effect of a mediation assessed by measuring the number of subjects survived after that mediation over a period of time. The time starting from a distinct point to the occurrence of a certain event, for example death is known as survival time and the corresponding analysis is known as survival analysis. The analysis was performed using the methods of Kaplan-Meier (KM) estimates and Cox Potential Hazard regression. KM plots showed the survival estimates as a function of each clinical feature and how each feature at various levels affect survival over the period of time. Cox regression modelled the hazard of death event around the clinical features used for the study. As a result of the analysis, ejection fraction, serum creatinine, time and age were identified as highly significant and major risk factors in the advanced stages of heart failure. Age and rise in level of serum creatinine have a deleterious effect on the survival chances. Ejection Fraction has a beneficial effect on survival and with a unit increase in the in the EF level the probability of death event decreases by ~5.2%. Higher rate of mortality is observed during the initial days post diagnosis and the hazard gradually decreases if patients have lived for a certain number of days. Hypertension and anemic condition also seem to be high risk factors. Machine learning classification models for survival prediction were built using the most significant variables found from survival analysis. SVM, decision tree, random forest, XGBoost, and LightGBM algorithm were implemented, and all the models seem to perform well enough. However, the availability of more data will make the models more stable and robust. Smart solutions, like this can reduce the risk of heart failure condition by providing accurate prognosis, survival projections, and risk predictions. Technology and data can combine together to address any disparities in treatment, design better care plan, and improve patient health outcomes. Smart health AI solutions would enhance healthcare policies, enable physicians to look beyond the conventional practices, and increase the patient satisfaction levels not only in case of heart failure conditions but healthcare in general.

INDEX TERMS Machine Learning, Cardiovascular Disease, Heart Failure, Survival Classification

I. INTRODUCTION

Heart Failure, a disorder that has no cure. A risk condition in which the heart isn't pumping blood as efficiently as it should be resulting in an inadequate cardiac output. The effective functioning of the body is dependent on how well the heart is able to pump blood and provide the most essential fuel i.e., oxygenated, and nutrient rich blood to the cells.

Heart failure falls into the category of cardiovascular diseases (CVDs) which are disorders of the heart and blood vessels and one of the leading causes of global deaths. CVDs led to an estimated 32% (17.9 million) of all global deaths annually. CVDs and related heart failure deaths majorly happen due to heart attacks, strokes, and almost 1/3rd of the deaths occur prematurely under the age of 70 [1]. CVDs are associated with certain behavioral factors of individuals such as unhealthy diet, lack of physical activity, consumption of tobacco and alcohol whose overall harmful effects may show up as high blood pressure, higher levels of blood sugar, obesity, etc. CVDs and heart failure conditions are usually long-term and tend to get worse with time if not properly cared and controlled. The symptom of a worsening heart includes breathlessness even in the resting condition, feeling of tiredness, laziness, swelling in the limbs, faster heart rate, dizziness, loss in balance and coordination, and feeling unconscious. Heart failure can be categorized into the *acute* kind which develops suddenly or the *chronic* kind where the heart's function deteriorates gradually over time. In most of the cases, heart failure is seen as a result of underlying medical conditions which affects the functioning of the heart.

Before diving into the details regarding the anatomical causes and physiology of heart failure and CVDs, first lets briefly discuss about the normal functioning of the heart. A normal healthy heart is anatomically a muscular pump responsible for circulation the oxygenated blood throughout the body via the circulatory system. The heart contains a total of 4 chambers with 2 on each side – the right and the left. The upper chambers are known as atrium and the lower chambers are known as ventricles. The chambers of the right side deal with deoxygenated blood while the left chambers deal with the oxygenated blood. The right atrium takes in the deoxygenated blood from the body and pumps it to the right ventricle. Once the right ventricle is full it sends the blood through the pulmonary artery to the lungs where blood gets oxygenated. Oxygen rich blood flows into the left atrium via the pulmonary veins and finally reaching the left ventricle. Once the left ventricle is full, oxygenated blood is circulated throughout the body via the aortic valve and the aorta. The cardiac cycle of the heart is completely based on the rhythmic contraction and relaxation of the heart muscles.

In conditions of heart failure, there are structural changes to the heart mainly in the left ventricle. As Andrew JS talks about in the study - The pathophysiology of chronic heart failure [2], the failing ventricle loses its coordination and contractions resulting in an inefficient functioning. Since the left ventricle generates both pressure and flow of the blood, any kind of a failure condition results in a reduced cardiac power output. The weakened cardiac muscles cause the ventricle to lose the ability to contract and may stretch to a point beyond which the heart can't pump the required amount of blood needed by the body. Based on the amount of blood pumped out with each beat heart failure is categorized into three types. Type 1 is heart failure due to reduced ejection fraction - HFrEF, also known as heart failure due to left ventricular systolic dysfunction and is characterized by an EF < 40%. Type 2 is heart failure with mid-range ejection fraction – HfmrEF with an EF in the range of 40-49%. Type 3 is heart failure with preserved ejection fraction - HFpEF or the diastolic heart failure characterized by an EF ≥ 50% [3]. The major risk factors for developing the failure condition includes ischemic heart disease, hypertension, smoking, obesity, elevated cholesterol levels, and diabetes. Bui et. al., [4] mention that heart failure can have various traits based on the age, sex, race, and ethnicity. Based on the functional capacity, disease progression, and severity of failure, the New York Heart Association (NYHA) has categorized failure into four classes (Class I, II, III, and IV). As per Bredy et. al., [5] the various classes are based on capacity to perform various physical activities based on the metabolic equivalent of task, *a measure of how much energy is expended compared to remaining at rest, relative to the mass of the person* and could differ among individuals. There are a numerous health conditions that could lead to a heart failure like atherosclerosis – a condition where arteries supplying blood to the heart are blocked due to build-up of fatty substances (plaque), high blood pressure – which puts additional stress on the heart, cardiomyopathy – a condition in which the heart muscles & walls of the 4 chambers in the heart gets stretched, arrhythmia – problems related to the rhythm of the heartbeats, congenital heart problems, hyperthyroidism, or pulmonary hypertension [6].

The objective of this study is to build a robust survival analysis system and survival classification system for heart failure that help in timely and accurate prediction of the survival function. The study also builds and compares models for survival classification for patients with heart failure. Survival analysis involves the techniques and procedures for evaluation of data to determine the time until an event occurs. The event of interest in the context of this study is the death event and the corresponding time to event is the time from when the heart failure is diagnosed to the occurrence of the death event. This study follows the techniques like Kaplan-Meier analysis and Cox Potential Hazard regression models for estimating the time to the death event for heart failure. Survival classification involves predicting the survival of patients by analyzing the available clinical data. These models aid the physicians and act as a decision support system for making better assessments and thereby increasing the patient outcome levels. The models and experiments designed in this study have the potential to be packaged to create a smart AI solution that does survival prediction, provides self-explainable survival analysis for time to event analysis for heart failure patients, and also establishes the most significant risk factors for a patient diagnosed with the heart failure condition. Thus, it is reasonable to conduct the proposed experiments.

The rest of the paper is organized into the following sections - literature review to discuss about the methods applied in other studies related to survival analysis and classification, material, and methods to discuss about the methods applied in the scope of this study, results, and discussion to analyze the results acquired from the experiments conducted, and the conclusion which summarizes the study and mentions the scope for future work.

II. LITERATURE REVIEW

Heart failure is one kind of condition within the human body which deteriorates the cardiac functioning gradually over time ultimately leading to the death event in severe conditions. Once the failure is diagnosed many different analysis and studies can be performed to study the effect of the failure. Various analytical studies have been accepted in the past to assess the survival analysis of individuals developing some kind of a heart failure. The survival analysis experiments are constructed using diverse set of algorithms like the Kaplan-Meier estimates, Cox Proportional Hazard regression methods, and various other bio-statistical analysis using methods like Mann–Whitney U test, Pearson correlation coefficient , and chi square test to evaluate the feature ranking for the data. Many studies have also applied machine learning algorithms like the - Support Vector Machines, Logistic Regression, Multi-Layer Perceptron, Artificial Neural Networks, Random Forest, Decision Tree, Ensemble Learning approaches, and boosting algorithms for heart failure survival classification and risk prediction. Many studies have also been conducted to predict hospital readmissions. In this section we discuss on the similar studies that have happened in the recent past.

Survival analysis is the major focus area in the study the heart failure conditions and analyzing the time to the event of interest (death event) is one of the major attention points. Ahmad et. al., [7] utilize the capabilities of Cox Regression and Kaplan Meier estimates to study the over-all trends of survival for patients detected with class III/IV of heart failure. Bayesian approach is another approach to assess the probability of the event of interest occurring. Bayesian methods are more useful in a clinical setting for relevant data analysis and considered to be much better in comparison to other methods. Ashine et. al., [8] used the Bayesian method with Deviance Information Criteria as the model selection scheme. The model with least deviance value is the preferred one. The overall survival time is impacted by the severity of conditions like chronic kidney disorder, diabetes mellitus, high blood pressure, anemia, smoking, and advanced stages of heart failure. The study also points out that patients who receive appropriate treatment and maintain an active lifestyle have higher probability of survival. Zheng et. al., [9] performed a time-to-event analysis by evaluating risk stratification model based on the LightGBM model and plotting the Kaplan-Meier estimates and hazard ratio using the Cox PH model to predict the 1, 2, and 3 year all-cause mortality of patients with chronic heart failure. Chicco et. al., [10] conduct a feature ranking analysis using the established biostatistical analysis and compare the results with the corresponding machine learning feature importance given by the models. The authors also design several survival prediction models based on the number of data features considered for model building. The analysis says that it may be possible to predict the survival of heart failure patients considering only ejection fraction and serum creatinine levels. Jia et. al., [11] design a risk identification model to identify the major risk aspects that would help to compute the 10 year risk of developing cardiovascular disease. The risk model based on Cox PH regression, involved a novel risk factor – heart rate which is classified as a significant factor for estimating the risk. Cheraghi et. al., [12] design a prognostic framework to determine factors affecting the 6 month survival and follow-up using the Cox PH and Kepler-Meier analysis. With a 45.8% mortality rate, the results for a 6-month prognosis lacked precision. Awan et. al., [13] study the 30 day hospital re-admission for patients aged over 65 diagnosed with the heart failure condition. The authors implement a multi-layer perceptron based model and compare the results with other ML algorithms. The MLP model was tuned to set the weights of minority class to three times that of the majority class to deal with the class imbalance for better generalization and turned out to be the best (AUC – 0.628) in predicting the 30 day readmission or death. Desai et. al., [14] perform a prognostic study to evaluate the 1 year follow-up period by comparing ML algorithms with the traditional logistic regression. ML methods provide only a marginal improvement over logistic regression in predicting the heart failure outcomes. However, the inclusion of clinical features extracted from electronic medical records enhances the model performance considerably with gradient boosting methods achieving a p-value <0.001. Venkatesh et. al., [15] in a multi ethnic evaluation of atherosclerosis use 735 variables to predict outcomes over a period of 12 year follow-up. The authors use random survival forest for top_most predictors for each cardiac outcome. The structural changes of the left ventricle, presence of cardiac troponin-T, higher creatinine level, and age stood out as the top 3 risk factors for incident heart failure. Sanchez [16] designs and compares several heart failure survival prediction models built using random forests, decision trees, extra trees classifier, and boosting algorithms. Sanchez performs a detailed study on the explainability of the models to build transparency and self-explainable systems using techniques like the partial dependency plots, Shapley values, feature importance. The extra tree classifier turned out as the best performing in terms of both classification and explainability of decision. Tabassian et. al., [17] design a framework built with statistical modelling and machine learning techniques for the diagnosis of heart failure with preserved ejection fraction by analyzing the spatiotemporal patterns of echocardiograph curves. The ML model based on the distance weighted k-nearest neighbors achieved an area under the curve of 0.89. Vosough et. al., [18] compare the performance of 6 difference algorithms – SVM, least-square SVM, random forest, bagging classifiers, ada-boost, and naïve bayes to predict hospital readmission for heart failure patients with random forest showing the most optimal performance with an accuracy of 0.90. Mortazavi et. al., [19] study the usefulness of several aggregated models for analyzing 472 clinical features to predict a possible of four readmission outcomes - 30-day all-cause, 180-day all-cause, 30-day due to heart failure, and 180-day due to heart failure. The machine learning techniques show better performance and significant improvements (15 – 25 %) in comparison to the traditional logistic regression. Golas et. al., [20] conduct a retrospective study of the EMR data and aim to predict the 30-day hospital readmission risk by implementing a deep unified networks and take advantage of the mesh-like architecture to avoid overfitting. The system achieved an accuracy of 76.4%. Kwon et. al., [21] implement deep learning and machine learning systems to predict the mortality of patients diagnosed with heart failure. The deep model composed of 3 hidden layers each with 33 nodes, batch normalization, and dropouts. The ML algorithms utilized were logistic regression, bayesian network, support vector machine, and random forests. Deep learning system performed the best achieving an AUCs of 0.88, 0.782, and 0.813 for in-hospital, 12-month, and 36-month mortality respectively. Adler et. al., [22] implement boosted decision trees based out of the adaboost algorithm to improve the risk prediction for heart failure condition achieving an AUC of 0.88. Kumar et. al., [23] propose a secured IoT based framework fitted with machine learning (random forest & xgboost) and statistical survival models (Kaplan-Meier & Cox PH regression) to recognize the various risk factors and cardio vascular diseases related to heart failure. The major risk factors affecting survival include age, ejection fraction, serum creatinine, creatinine phosphokinase, and platelet count.

Most of the study cited above more or less point to similar health risk factors that trigger and effect the heart failure condition. Renal dysfunction (kidney disorder), diabetes mellitus, high blood pressure, anemia, and smoking are cited as the major risk factors that trigger a heart failure condition. Parameters like the Ejection Fraction, Sodium Creatinine levels, age of the patient, and the current stage of heart failure are the most important influencing parameters for deciding the outcome of heart failure at the advanced stages. Various results obtained from these studies suggest that machine learning methods could be well equipped to provide meaningful, accurate, and explainable risk prediction techniques which could not only save time but also increase the patient outcomes and satisfaction levels.

III. MATERIALS AND METHODS

A. DATASET

1) DESCRIPTION

The dataset used for this study is taken from the Faisalabad Institute of Cardiology [7]. The dataset contains cardiovascular medical records taken from 299 patients. The patient cohort comprised of 105 women and 194 men between 40 and 95 years in age. All patients in the cohort were diagnosed with the systolic dysfunction of the left ventricle and had previous history of heart failures. As a result of their previous history every patient was classified into either class III or class IV of New York Heart Association (NYHA) classification for various stages of heart failure as defined by Bredy et al., [5]. A detailed description of the dataset is explained in the next section.

2) FEATURE DESCRIPTION

The dataset comprises of 13 data features namely – Age, Anemia, High Blood Pressure, Creatinine phosphokinase, Diabetes, Ejection Fraction, Sex, Platelets, Serum Creatinine, Serum Sodium, Smoking, Time, and Death Event. Out of these 13 features, 5 are binary features - anemia, high blood pressure, diabetes, sex, and smoking. TABLE 1 gives an overview of the data features. Let’s discuss each feature and its medical significance in detail.

Age: Considered as one of the major risk factors for cardiovascular disease, increasing age causes deterioration in the cardiac structure and functioning which makes it vulnerable for heart failure. As per Li et. al., [24] ~1% of individuals over 50 years of age are affected by heart failure which in known to double with every passing decade of life. Strait et. al., [25] point out with increasing age, structurally the heart muscles become think and stiff which enforce an extra burden of the functional responsibilities of the heart. Functionally, since the heart is unable to function normally causes a number of deficits, lowering of cardiac reverse (*the difference between the rate at which the heart pumps blood at a particular time and its maximum capacity for pumping blood*). Aging also decreases the ability of the heart to undergo an effective repair which also declines with age.

Anaemia: *A condition in which the number of red blood cells or the haemoglobin concentration within them is lower than normal* and is often considered as a comorbidity with heart failure. Haemoglobin acts as the oxygen carrier and if one shows a decrease in red blood cell count causing a decrease haemoglobin level, there will be a decreased capacity of the blood to carry oxygen to the tissues and organs. Anemic condition is common in cases of heart failure and the common triggers of anemia involve age, gender, nutritional iron deficiency, deficits in folate, vitamin B12, & vitamin A, chronic kidney disease, and cytokine production as noted in the studies by Shah et. al., [26] and Taylor [27].

High Blood Pressure (Hypertension): Termed as a *silent killer*, hypertension is a condition which occurs when the force with which the blood pushes the walls of the blood vessels in always on the higher side. To understand this a bit more, we first try to understand the term blood pressure. When heart beats, blood is pumped out of the heart with

TABLE 1
Dataset Features

Feature	Explanation	Measurement	Range
Age	Age of the patient	Years	[40 -- 95]
Anaemia	Decrease of red blood cells or haemoglobin level in the blood.	Binary	0, 1
High Blood Pressure	Indicative of whether the patient has hypertension.	Binary	0, 1
Creatinine Phosphokinase	Level of the CPK enzyme in the blood	mcg/L	[23 -- 7861]
Diabetes	If the patient is diabetic	Binary	0, 1
Ejection Fraction	Percentage of blood leaving the heart at each contraction	Percentage	[14 -- 80]
Sex	Gender of the patient.	Binary	0, 1
Platelets	Platelets in the blood	kiloplatelets/mL	[25.01 -- 850.00]
Serum Creatinine	Level of creatinine in the blood	mg/dL	[0.50 -- 9.40]
Serum Sodium	Level of sodium in the blood	mEq/L	[114 -- 148]
Smoking	Indicates if the patient has smoking habit.	Binary	0, 1
Time	Follow-up period for the next doctor visit.	Days	[4 -- 285]
Death Event	If the patient died during the follow-up period.	Binary	0, 1

some amount of force or pressure to circulate the blood throughout the body via the circulatory system. This pressure is made of two force components – (1) Systolic pressure – the force with which blood is pumped out of the heart into the circulatory system, and (2) Diastolic pressure – the force generated when the heart rests between heart beats. High blood pressure triggers harm by increasing the workload of the heart and blood vessels and forcing them work harder and inadequately. Higher pressure increases the friction and harms the subtle tissues inside the arteries making them narrower as mentioned in a study by the American Heart Association [28].

Creatinine Phosphokinase (CPK): An enzyme (protein that helps to elicit chemical changes in the body) found in the heart, brain, and skeletal muscles. Any kind of damage to the muscle tissue causes the enzyme to leak into the blood stream.

Consequently, high levels of CPK typically indicate a sort of elevated stress to the heart or other muscles. The normal range of CPK in males is between 39 – 308 U/L and 26 – 192 U/L in females. Identifying the specific type of CPK helps determine what kind of a tissue could be damaged. In the condition of a heart failure the levels of CPK2 (CK-MD) are at elevated levels and could point to a myocardial muscle damage, electrical injury, or heart attack [29], [30].

Diabetes: A chronic disease that occurs when the blood glucose levels are too high. The pancreas is no longer able to prepare the required amount of insulin. Insulin acts as a bridge to let the glucose from food flow from the blood stream into the body cells to generate energy. Patients with history of diabetes are at a higher risk of developing heart failure. High levels of blood sugar can potentially damage the blood vessels and nerves that control the heart. Kenny et. al., [31] and Rosano et. al., [32] states that patients with diabetes mellitus are at a two time more risk of developing heart failure due to the abnormal cardiac handling of blood glucose.

Ejection fraction (EF): The measurement how much blood is pumped out of the left ventricle with each contraction. The ideal range for EF may lie somewhere between 50 to 70 percent. An EF of 60 indicates that 60% of the total blood volume in the left ventricle is pushed out with each heartbeat. An EF of less than 40% is indicative of heart failure or cardiomyopathy usually categorized as *"systolic"* heart failure. Heart failure triggered due to EF can be categorized into – (1) Heart failure with reduced ejection fraction (HFrEF): EF <= 40%, (2) Heart failure with preserved EF (HFpEF): EF >= 50%, and (3) Heart failure with mid-range EF (HFmrEF): EF between 41 – 49 (both inclusive) percent range [33], [34].

Sex: Although the overall lifetime risk of developing a heart failure for both men and women stands similar, there are striking differences in both the genders about the nature of heart failure. Men are known to have a higher incidence rate for heart failure than women. As reviewed by Stromberg et. al., [35] women are known to survive longer after the outset of heart failure. As discussed in the paper by Lam et. al., [36] men are more prone towards developing the HFrEF type, whereas women predominate in HFpEF type of heart failure. The higher risk of HFrEF in men is attributed to the greater disposition to macrovascular coronary artery disease and myocardial infarction both of which are distinguished precursor for HFrEF. Macrovascular disease refers to the deposition of fat, plaque, calcium, and blood clots in the larger blood vessels. These in turn increase the risk of developing type 2 diabetes which men are almost twice more likely to develop. And as already discussed patients with history of diabetes are at a higher risk of developing heart failure. These are some of the known gender differences in the condition of heart failure.

Platelets: Also known as thrombocytes, are cells that circulate within the blood and play a major role in blood clotting mechanism by binding together when some kind of a damage or injury to a blood vessel is recognized. The normal range for platelet counts in the body ranges between 150,000 to 450,000 per µL of blood. The condition of having greater than 450,000 platelets is known as thrombocytosis whereas a count less than 150,000 is known as thrombocytopenia. Chung et. al., [37] in their study mention the association of heart failure with increased risk of venous thromboembolism (blood clots in the veins), stroke, and sudden death. However, the contribution of platelets to the thromboembolic complications of heart failure is still uncertain. Mojadidi et. al., [38] discuss the effect of thrombocytopenic condition on mortality in the case of HFrEF condition and the utilization of total platelet count as a predictive indicator for heart failure.

Serum Creatinine: Creatinine, a chemical waste formed as a by-product of normal muscle functioning present in the blood stream which is filtered in the kidney and eliminated through urine. The normal creatinine levels in men and women are 0.6 to 1.2 milligrams/deciliters (mg/dL) and 0.5 to 1.1 mg/dL respectively. Men usually have higher creatinine levels compared to women since men, on an average, have more muscle mass. Shlipak et. al., [39] warn about the rising level of creatinine in the blood stream and the importance of renal functioning as a prognosis for heart failure. Any degradation in the functioning of the renal system contributes directly towards poorer outcomes and increased heart failure risks. Metra et. al., [40] point out that renal dysfunction and heart failure are closely related and associated with a high mortality rate. The interaction between cardiac and renal dysfunction is considered critical for prognosis and progression of heart failure conditions. Hence, a complete evaluation of the renal system is very much important to extract the current hemodynamic status of blood circulation and an accurate prognostic assessment for heart failures.

Serum Sodium: An essential electrolyte, sodium helps in maintaining the balance of water level in and around the cells. Maintaining proper sodium levels is important for proper functioning of muscles, nerves, and maintain stable blood pressure levels. The normal sodium level is between 135-145 milliequivalents per liter. The condition of sodium level less than 135mEq/L is known as hyponatremia. Abebe el. al., [41] discuss the impact of sodium levels in prognosis of a heart failure condition and mention that hyponatremia is one of the vital factors in the prognosis of heart failure condition. Adrogué [42] discusses the impact of low sodium levels at the early and late stages of heart failure, the ways of preventing hypotonic hyponatremia (excess of free water), and managing hyponatremia in heart failure patients.

Smoking: People with habit of smoking have a major risk of developing ischemic heart disease. Ischemic disease happens due to building up of plaque within the coronary artery. As mentioned in the publication article by the National Heart, Lung, And Blood Institute [43], [44] plaque can choke the arteries by forming blood clots, thereby limiting the flow of blood to the heart muscles. In the event of heart not receiving enough blood, depletes it from getting the adequate amount of oxygen and nutrients for appropriate functioning. This condition is called ischemia. Insufficient blood supply to the heart muscles puts the person at risk for a heart attack. The chemicals that go in with smoke triggers the buildup of plaque in the arteries, damaging the blood vessels, and altering the way they work by disturbing the normal heart rhythm. Kamimura et. al., [45] find that cigarette smoking involves a critical hazard factor on the structure and functioning of left ventricle and heart failure hospitalization. Ahmed et. al., [46] study the effect of quitting smoking on the risk of developing a future heart failure and find that cessation for >15 years brings down the risk of heart failure and death to the same level as that of a non-smoker.

Time: The time period where a heart failure patient transits from the in-patient setting to the out-patient setting is considered critical in managing heart failure condition. Mueller et. al., [47] discuss the importance of a well-designed and structured follow-up program for refining the treatment outcomes. The authors highlight the importance of patient education during the follow-up visits for self-monitoring the signs & symptoms of any kind of decline in the heart health. Similarly, Agostinho et. al., [48] mention the how a well-structured and protocol based follow-up program can lead to reduction in hospital re-admission and mortality rates. McAlister et. al., [49] point out a follow-up within 14 days post discharge due to the heart failure condition is associated with improved treatment outcomes.

Death Event: Heart disease or cardiovascular disorders are one of the leading causes of death globally. A trigger for heart failure could be plaque building up within the blood vessels which reduces or blocks the flow of blood, dysfunction of the renal system causing high levels of creatinine, low sodium levels, fluctuating ejection fraction, or other cardiac abnormalities. The severity of the above mentioned factors could determine the criticality of the condition and death can happen due to acute myocardial infarction, progressive heart failure, sudden death, or other cardiovascular irregularities. The event of death may vary depending upon the gender, race, and ethnicity.

3) DATA SPLIT

The dataset contains a total of 299 records. So, to have as much data available for training the train to test split ratio is kept at 85:15 respectively. The training data gets 254 records while the test data gets a 45 record share. In the 254 records in train set, there are 171 cases that survived the heart failure condition while 83 cases succumbed to the condition. The 171:83 ratio looks highly imbalanced, and the model could easily drift towards learning/predicting the survival cases resulting in a bias. The imbalance in the train set is corrected by the application of Synthetic Minority Oversampling Technique (SMOTE) data augmentation technique that synthesizes duplicate data records from the existing examples for the minority class. Once SMOTE is applied, we have a balanced set with 171 records each for both the classes. The model would follow a balanced learning approach and not get biased towards survival cases.

B. SYSTEM DESIGN AND ARCHITECTURE

1) PROGRAMMING RESOURCES

All of the programming task in the study was done on a cloud based system with CUDA enabled Nvidia Tesla K80 GPU, 4 core CPUs, 20 GB RAM. The programs are written in Python 3.8.8 using the web-based Anaconda Jupyter environment. Dataset creation, loading, and manipulation were done using the Pandas library. NumPy was utilized for any scientific computing with the data and numerical analysis. Matplotlib and Seaborn libraries were also utilized for basic analysis of categorical variables. Plotly and Plotly Express libraries were extensively utilized to conduct an in-depth and interactive Exploratory data analysis. Since Data augmentation and class imbalance was dealt with using the Imbalanced-learn package. All the data preprocessing APIs, classification metrics APIs, and machine learning algorithms were consumed from the scikit-learn package. XGBoost and LightGBM package were utilized to create boosting models. Lifelines python library was utilized for survival analysis experiments.

2) MACHINE LEARNING ALGORITHMS

The following machine learning algorithms were applied, and their performance analyzed on the dataset for a comparative study for the best classifier to predict the event of death or survival based on the various health parameters being considered.

Support Vector Classifier (SVC): To start with, a basic version of SVC with Radial Basis Function (RBF) kernel was implemented.

Decision Tree: A supervised learning methodology based on tree like structure which learns/predicts using simple decision rules reasoned from the various data features.

Random Forest: A supervised learning methodology built on top of decision trees. The individual trees are built on different samples and makes an evaluation based on the majority vote for classification.

Extreme Gradient Boosting (XGBoost): Coming from the class of gradient boosting methods based on decision trees, XGBoost provides an efficient form of boosting algorithm with faster implementation, training, and delivering high performance. Each new tree focusses on the errors made by the previous tree to create a strong predictor model.

Light Gradient Boosted Machine (LGBM): Another algorithm based out of the gradient boosting mechanism and built on top of decision trees, LGBM works by carrying out leaf-wise vertical growth of the system resulting in a better reduction of loss function and pushing towards a higher accuracy.

All the above learning modes are tuned to extract the best learning parameters and the best estimator using the RandomizedSearchCV approach where not all hyperparameter values are considered. Instead, a set number of hyperparameter setting is experimented from the specified probability distributions. A 5-fold cross validation configuration is used to fit the RandomizedSearchCV method.

3) FEATURE RANKING and MODEL EXPLAINABILITY

As part of the machine learning survival classification experiments, each of the model trial was extended to verify the importance assigned to all the data features by the algorithm. This would voice out the most important features in arriving at the classification results and in turn contribute to creating better self-explainable models. The feature ranking APIs from the respective machine learning algorithm were utilized to obtain the importance each algorithm assigns to respective features. In addition, SHAP (SHapley Additive exPlanations) library was utilized to explain the output of models.

4) SURVIVAL ANALYSIS

In clinical trials and survival analysis, the effect of any mediation is assessed by measuring the number of subjects survived after that mediation over a period of time. The time starting from a distinct point to the occurrence of a certain event, for example death is known as survival time and the corresponding analysis is known as survival analysis. Time-to-event in the clinical setting is a variable for each subject with a definite beginning and an end defined somewhere along the time line of the complete duration of the study. The time begins as the subject is enrolled into the study or at the start of a particular treatment and ends when the event of interest happens, or the subject becomes censored from the study. A series of experiments were conducted to evaluate the survival probability using the Kaplan-Meier (KM) estimates and Cox regression models.

KM estimate is one of the most efficient approaches available to measure the fraction of subjects living for a definite amount of time after a treatment. The KM estimate survival curve is defined as the probability of surviving in a given length of time while considering time in many small intervals. There are three assumptions used in the KM analysis.

1. At any point in time, the censored patients have the same survival forecasts as those who continue to remain as part of the study and are followed up.
2. The survival likelihoods are the same for all the subjects irrespective of when they are enrolled to be a part of the study.
3. The event happens at the time specified.

The survival probability at any certain time interval is calculated as the number of subjects surviving at the particular time divided by the number of patients at risk as determined at the start of the study and is calculated as per the formula defined as per equation (**1**):

$$S_t = \frac{(Number\ of\ subjects\ living\ at\ the\ start\ - \\ Number\ of\ subjects\ died)}{Number\ of\ subjects\ at\ the\ start} \qquad (1)$$

Subjects who die or drop out of the study are not counted as *at risk* and are considered as censored, ignored in the denominator [50] [51].

Cox Proportional Hazards Model introduced by D.Cox [52] allows to compute survival regressions on the time duration using the censored subjects and the features available as part of the dataset based on small intervals containing at most one event of interest. The dependent variable is a hazard function at any given time and the model follows the concept of gradient descent for optimization. Hazard is basically the reverse of survival, or the likelihood of failure (death event). Cox model works in a way that the logarithm hazard function for a particular subject act as a linear function of the relevant static features while the hazard function at the population-level changes over time. Mathematically, the hazard model is defined as per equation (**2**) by Davidson-Pilon [53].

$$h(t|x) = b_0(t) \exp\left(\sum_{i=1}^{n} b_i(x_i - \bar{x}_i)\right) \qquad (2)$$

where,

- h(t|x) represents the hazard function.
- $b_0(t)$ represents the baseline hazard.
- the summation term represents the log-partial hazard.
- the full exp(...) term represents the partial hazard.

Cox regression makes following basic assumptions -

1. All individuals have the same hazard function, but a unique scaling factor in front.
2. The explanatory variables act multiplicatively on the hazard function.
3. The relationship between the log hazard and each covariate is linear.

IV. RESULTS AND DISCUSSIONS

The entire study was conducted in different phases comprising of Exploratory Data Analysis, Survival Analysis, Survival Prediction, and Feature Ranking & Importance. IN this section we discuss the important finding in each of the phase and the inferences drawn from the results.

A. EXPLORATORY DATA ANALYSIS

All the 13 features from the dataset were evaluated in a detailed study as part of the exploratory analysis. Let's discuss each feature and the hidden information that were extracted from this analysis.

CATEGORICAL FEATURES

Out of the 13 features, 'anemia', 'high_blood_pressure', 'diabetes', 'sex', 'smoking', and 'DEATH_EVENT' are binary and fall into the categorical bucket. A detailed analysis reveals the following information –

a. ~43% of the population under study have anemic symptoms while ~57% are non-anemic.
b. ~35% of the population have hypertension or high blood pressure while ~65% have normal blood pressure.
c. ~42% of the population are diabetic while ~58% are non-diabetic.
d. ~65% of the population are male while ~35% are female.
e. ~32% of the population have smoking habits while ~68% are non-smokers.
f. 32.11% of the cases succumbed while 67.89% cases survived the condition.

FIGURE 1 shows the bar chart representation for distribution of the categorical variables.

1) ANAEMIA

Among the population under consideration, 43.1% are anemic while 56.9% do not show any such symptoms. In the anemic population, 27.8% have survived the heart failure condition while 15.4% people succumbed to the condition. In the non-anemic cohort, 40.1% of the population survived the failure while 16.7% have succumbed to the condition. The pie chart in FIGURE 2 validates the numbers just discussed.

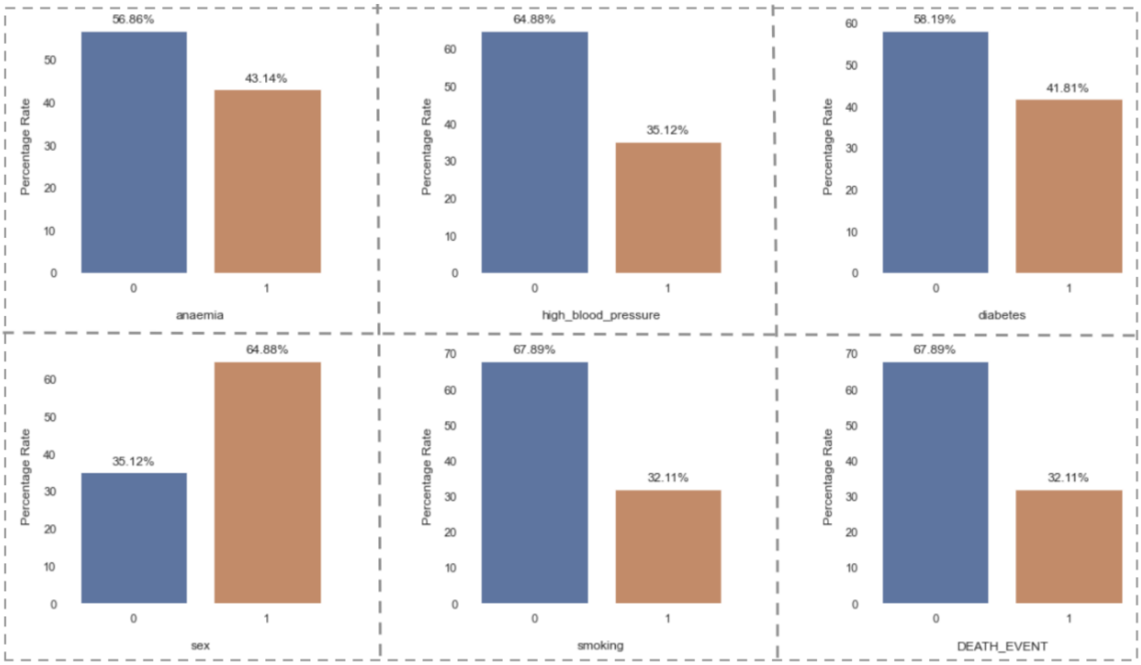


FIGURE 1. Categorical features

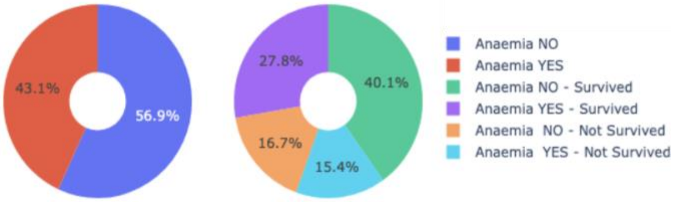


FIGURE 2. Effect of anemia on survival

2) HYPERTENSION

Out of the total population, around 35.1% of the patients suffer from high blood pressure or hypertension. Among these 22.1% survived the heart failure condition while 13% succumbed to the condition. From the 64.9% that do not have hypertension, 45.8% survived the event of a heart failure while 19.1% succumbed to the condition. FIGURE 4 shows the chart depicting these observations.

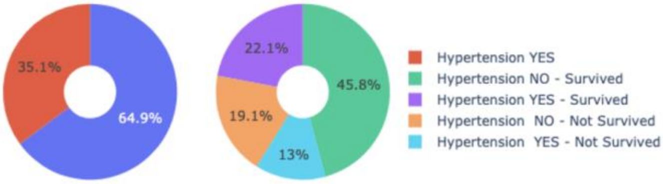


FIGURE 4. Hypertension and survival

3) DIABETES

Out of the total population, ~42% of the people have diabetes while ~58% do not have diabetes. In the diabetic cohort, 28.4% of the population have survived the heart failure condition while 13.4% people have succumbed. In the non-diabetic category, 39.5% of the population survived the event of a heart failure while 18.7% people succumbed to the condition. FIGURE 3 displays the chart showing these observations.

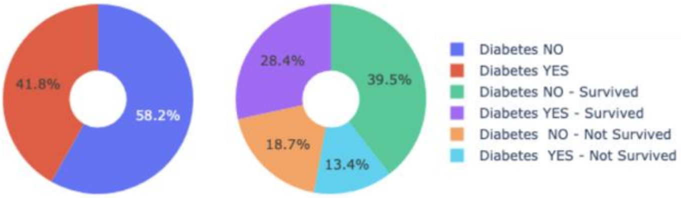


FIGURE 3. Effect of diabetes on survival

4) SMOKING

Out of the total population, ~32% of the people have smoking habits while the remaining 68% non-smokers. Among the smoking group, 22.1% have survived the heart failure condition while 10% people have succumbed to the condition. In the non-smoking cohort, 45.8% of the population have survived the heart failure condition while 22.1% have succumbed to the condition. FIGURE 6 demonstrates the numbers discussed.

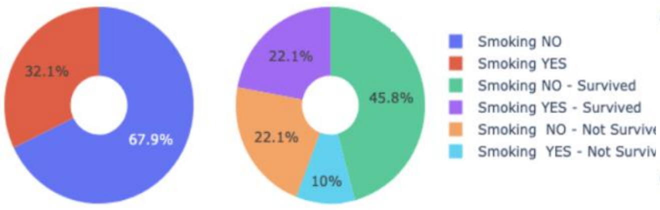


FIGURE 6. **Effect of smoking on survival**

5) SEX

In the entire cohort, 44.1% of the male population survived while 20.7% succumbed to the condition. Similarly, 23.7% of the female population survived and 11.4% succumbed to the heart failure. FIGURE 7 displays the chart showing these observations.

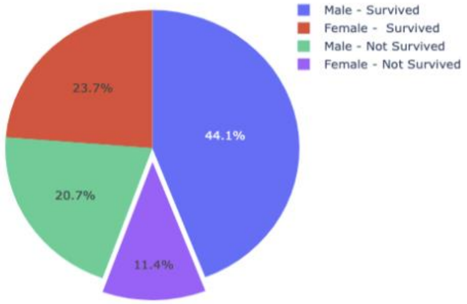


FIGURE 7. **Gender and survival**

QUANTITATIVE FEATURES

1) AGE

The age for the population under consideration ranges between [40 – 95] with prominent spikes in the population density at certain age intervals around [44-46], [50-52], [60-62] (highest density), [64-66], [70-72]. The complete distribution can be seen in FIGURE 9.

The analysis on the effect of age on survival rate tells the following -

1. The survival rate is more within the age group 50 to 70.

2. The probability of not surviving the heart failure condition is prevalent across all age groups with maximum around the 60's age group. The chance of survival decreases drastically beyond the age of 80. FIGURE 5 shows the effect of age on survival.

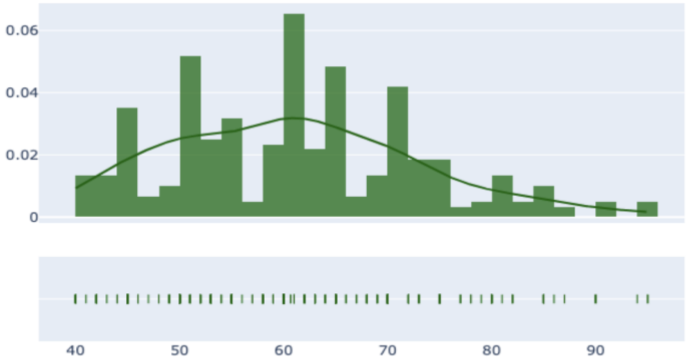


FIGURE 9. **Distribution of age**

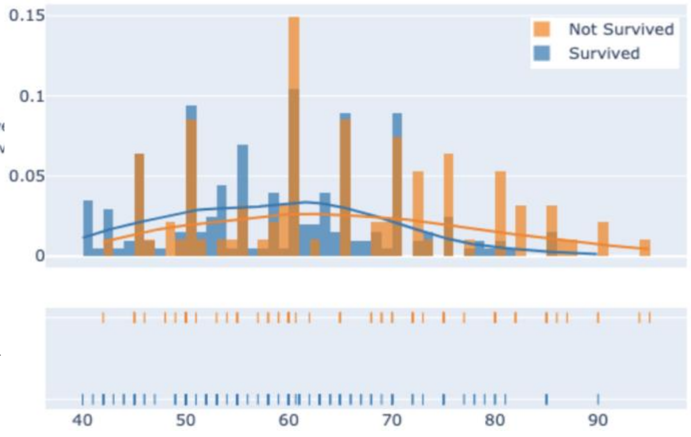


FIGURE 5. **Effect of age on survival**

2) CREATININE PHOSPHOKINASE

The CPK levels for patients who did not survive the heart failure are on the higher side with a few patients show abnormally high CPK levels. FIGURE 8 shows the distribution plot for the CPK levels across the population.

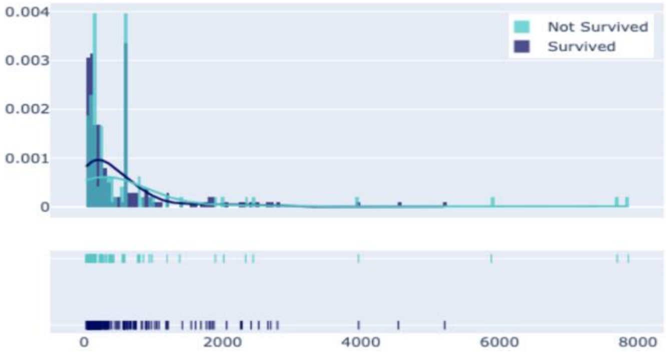


FIGURE 8. **Distribution of cpk**

3) EJECTION FRACTION

Majority of the person who succumb due to heart failure show a lower than normal ejection fraction values. Only 25 - 45 (%) of the blood was being pumped out of the heart. FIGURE 10 shows the distribution plot for the levels of ejection fraction across the population.

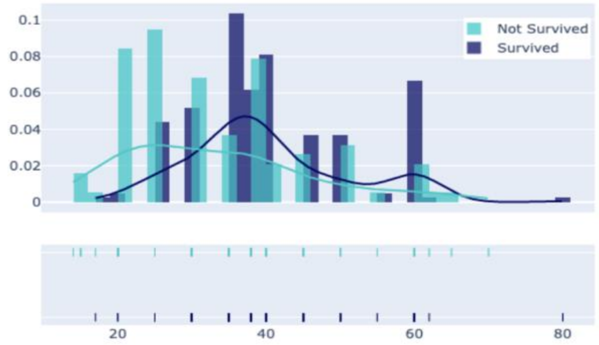


FIGURE 10. **Distribution of ejection fraction**

4) PLATELETS

Majority of the person who succumbed to the heart failure condition have platelets count within the normal range. There are a few cases of death where the platelet count hovers around the lower boundary limit. A few cases report abnormally high count of platelets in both the survival and death cases. FIGURE 12 shows the notched box plot for platelets distribution across the population.

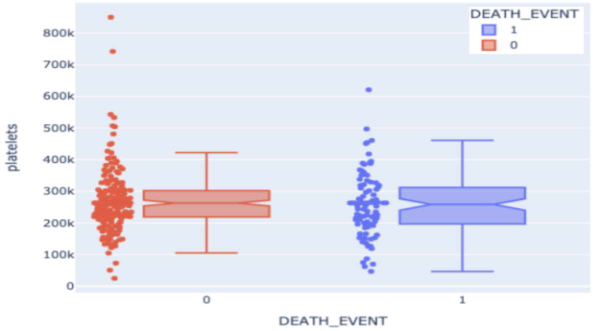


FIGURE 12. Distribution of platelets

5) SERUM CREATININE

Among the population under this study, there are 96 cases who have succumbed to the heart failure condition. Out of these 96 subjects, 87 have reported serum levels hovering around the upper limit or even higher level. The distribution of the blood serum levels against the probability density is shown in FIGURE 13.

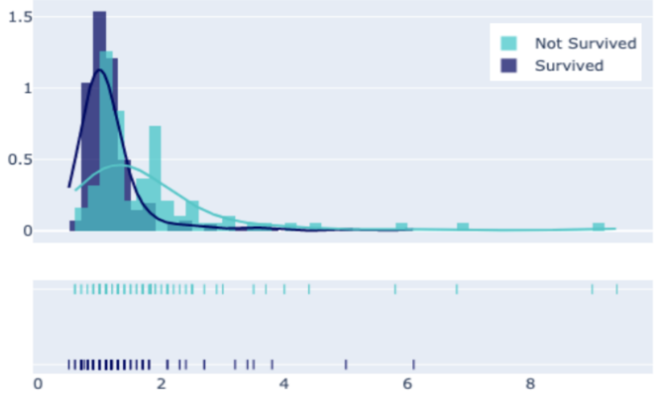


FIGURE 13. Distribution of serum creatinine

6) SERUM SODIUM

Out of these 96 subjects that have succumbed to the heart failure condition, 59 (46.5%) cases have reported lower than normal sodium levels. Out of which 26.8% cases survived the heart failure condition while the remaining 19.7% succumbed. Of the cases that have sodium levels in the acceptable range, 41.1% cases survived the heart failure condition while a small number of 12.4% cases succumbed to the condition. FIGURE 11 shows the distribution for serum sodium against the probability density across the population who survived vs the ones who did not survive.

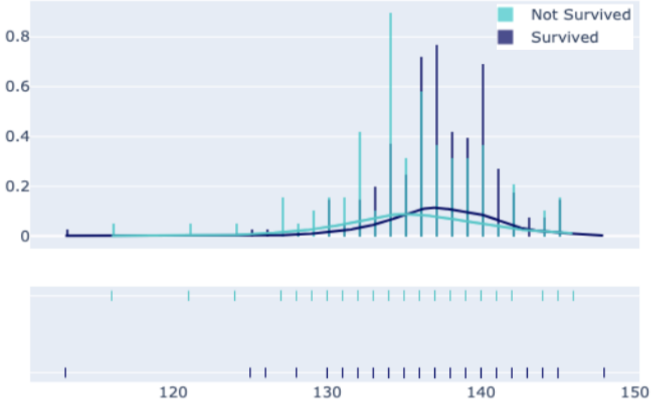


FIGURE 11. Distribution of serum sodium

7) DEATH EVENT

Out of the 299 subjects in the cohort, 96 (32.11%) have succumbed to heart failure while 203 (67.89%) cases survived. The value 1 implies the subject survived whereas 0 indicated the subject succumbed to the heart failure condition. FIGURE 14 shows the distribution of the death event.

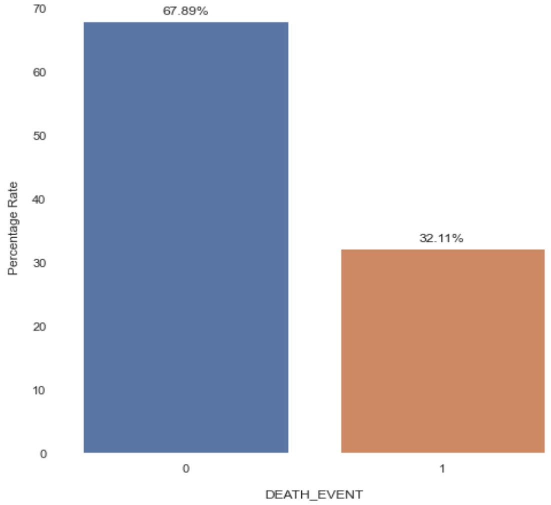


FIGURE 14. Distribution of death event

3. SURVIVAL ANALYSIS

A series of experiments were conducted to analyze the survival probability using the Kaplan-Meier (KM) estimates and Cox regression models.

1) KAPLAN MEIER ESTIMATE

KM estimate curves were plotted for each of the categorical and continuous features available. We discuss each of the category in detail in the sections below.

CATEGORICAL FEATURES

FIGURE 15 shows the plots for each of the categorical feature. A complete understanding of each plot is discussed next.

Anaemia: Anemia is common in heart failure condition in which the number of red blood cells or the haemoglobin concentration is lower than normal thereby reducing the capacity of the blood to supply oxygen to the cells and tissues. The KM estimate plot for anemia shown a decreased probability for survival if the person is known to be anemic.

Diabetes: Diabetes is known to trigger heart failure conditions and patients with diabetes are at an increased risk of developing heart failure symptoms at initial stages. However, the KM estimate curve for diabetes and survival probability shows almost comparable trend for both diabetic and non-diabetic patients making diabetes a non-significant factor. This could be due to the fact that the population considered as part of this study had a history of Left Ventricular Systolic Dysfunction and had previous heart failures that put them in classes III or IV of New York Heart Association classification of the stages of heart failure.

High Blood Pressure: High BP increases the heart's workload with narrow and less flexible arteries making it difficult for the blood to travel efficiently throughout the body. The KM estimate curve also shows a similar trend where patients with high blood pressure (hypertension) are at an increased risk of survival due to heart failure with significant lower survival probability.

Sex: The occurrence of heart failure is lower in women compared to men for all ages. However, due to rise in incidence with age the overall number of men and women living with heart failure are nearly comparable. The KM estimate curve also shows a similar trend and since the population under consideration are at an advanced stage of heart failure, the survival probability curve shows similar trend for both men and women.

Smoking: Smoking is a major risk factor for developing initial stages of ischemic heart disease, a condition in which plaque builds up inside the coronary arteries. People who smoke are at an increased risk of developing heart failure condition. However, the KM estimate curve for smoking shows almost comparable trend for both smokers and non-smokers making smoking a non-significant factor. This could be due to the fact that the population considered as part of this study had Left Ventricular Systolic Dysfunction and had previous heart failures that put them in classes III or IV of New York Heart Association (NYHA) classification of the stages of heart failure.

Thus, from the KM estimates for categorical features, Anemia and High Blood Pressure are known to have a major impact on the survival probability of patients who are at an advanced stage of heart failure condition.

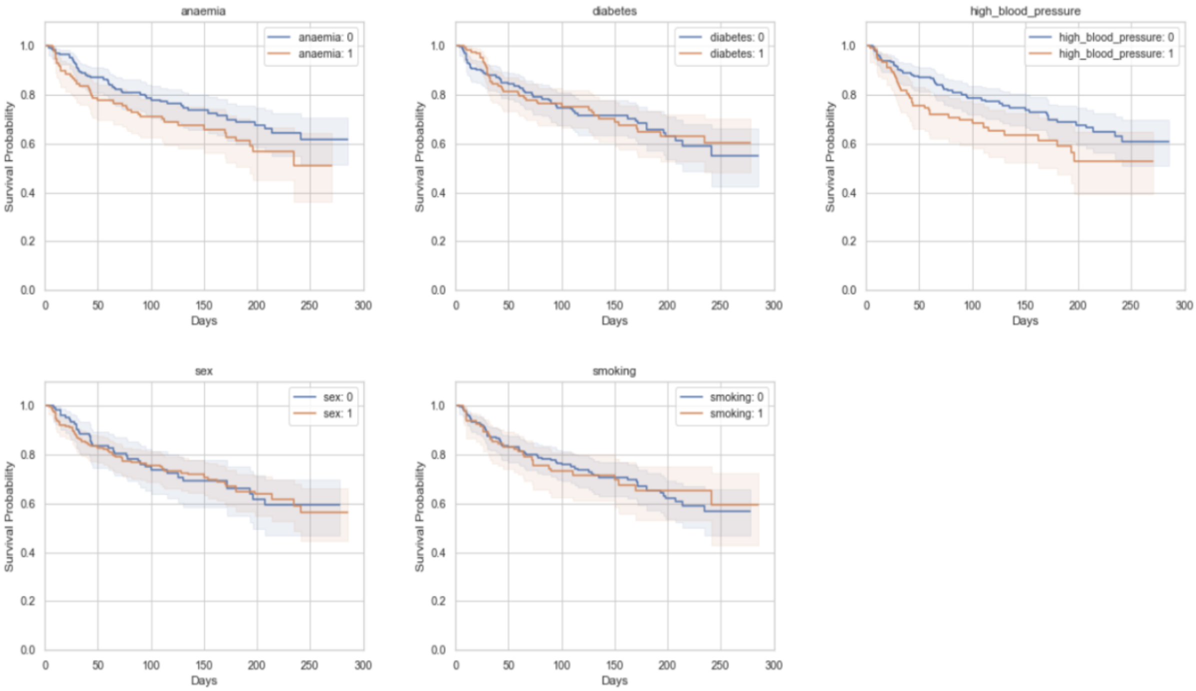


FIGURE 15. KM estimates: categorical features

CONTINUOUS FEATURES

FIGURE 16 and FIGURE 17 shows the plots for each of the numerical feature. A detailed interpretation of the respective plot is discussed below.

Age: The cohort was split into 3 categories of age that ranges from 40 to 95. The age buckets were [39 – 60], [60 – 80], and [80 – 100]. The KM estimate curve for survival probability shows the least survival probability for people greater than 80 years.

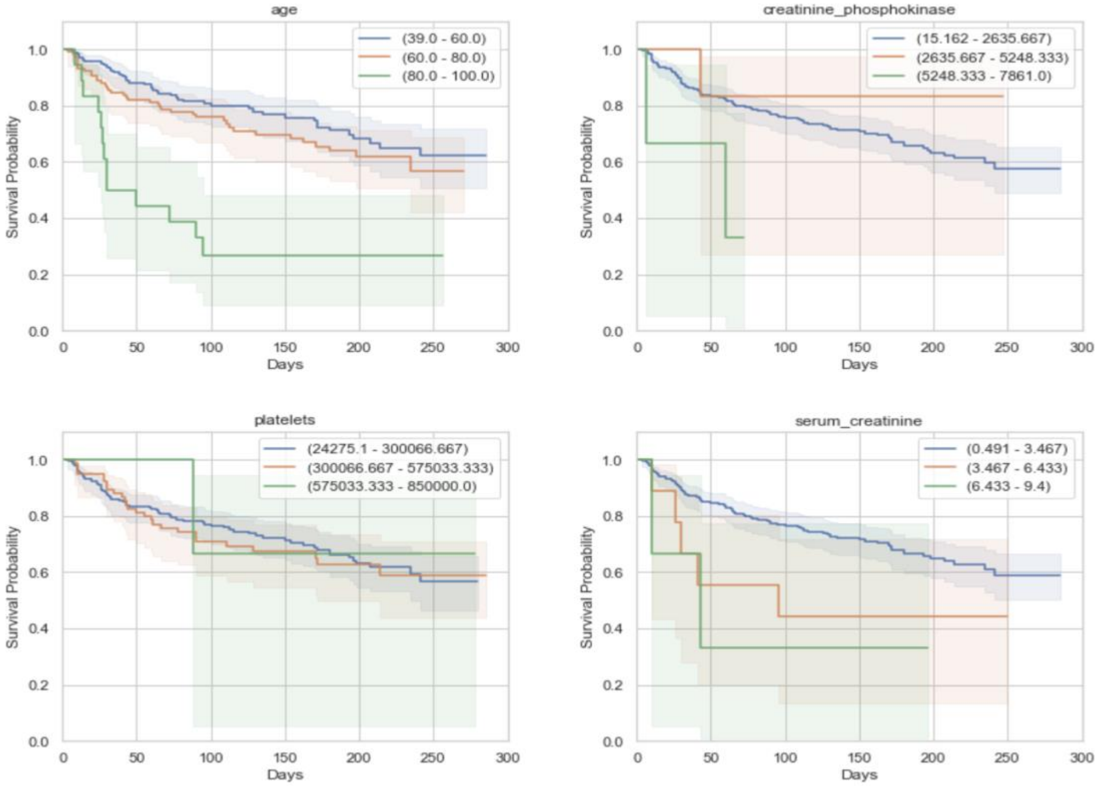


FIGURE 16. KM estimates for continuous features I

Creatinine Phosphokinase (CPK): The CPK values for the population is spread out in the range [23 -- 7861]. This range was split into 3 buckets of [15.162 - 2635.667], [2635.667 - 5248.333], and [5248.333 - 7861.0]. The survival probability stands lowest for the higher bucket of [5248.333 - 7861.0]. A majority of the population fall under the 1st bucket who show a decent chance of surviving. This makes CPK levels somewhat insignificant.

Platelets: The normal platelet count ranges from 150,000 to 450,000 platelets per microliter of blood. The platelets count for the population ranges from [25100 - 850000] which was divided into 3 buckets for KM estimate curves. All the 3 groups of platelets range show a comparable and decent survival probability. This is indicative that platelets be an insignificant feature for computing the survival estimate and prediction.

Serum Creatinine: Heart failure condition usually report an increase in serum creatinine levels in the scale of ≥ 0.3 mg/dL. The range of serum creatinine levels for the population ranges from [0.5 - 9.4] which was divided into 3 buckets - [0.49 - 3.46], [3.46 - 6.43], and [6.43 - 9.4] for KM estimate curves. The curve for the 3rd bucket showed the least survival probability followed by the 2nd and 1st buckets respectively. This makes it evident that increased serum creatinine levels pose a good amount of during heart failure conditions.

Ejection Fraction (EF): The EF range for the population under consideration in [14 -- 80]. To study the survival probability the EF range was converted into buckets of [0 - 30], [30 - 45], and [45, 100]. KM estimate curves show that population falling under the bucket [0 - 30] show the least chance of surviving a heart failure condition. The remaining 2 buckets show a comparable and decent chance of survival.

Serum Sodium: The normal blood sodium level is between 135 - 145 milliequivalents per liter. The range of serum sodium levels for the population ranges from [113 - 148] which was divided into 3 buckets - [113 - 124], [124 - 136], and [136 - 148] for KM estimate curves. From the KM plots, it is observed that survival probability is least for population falling under the 1st bucket with extremely low sodium levels (However, it looks like there are not many people falling into this bucket.). Bucket 2 and bucket 3 population showed better probability with population falling under bucket 3 having the normal sodium levels and the best chances of survival a heart failure condition.

Time: The time column denotes the follow-up time for a patient. Follow-up time can be interpreted as a time to event where the event in this case would be DEATH or the patient becoming censored. So, time can be interpreted as a target column predicting the time to death in case of survival analysis for a patient with heart failure condition.

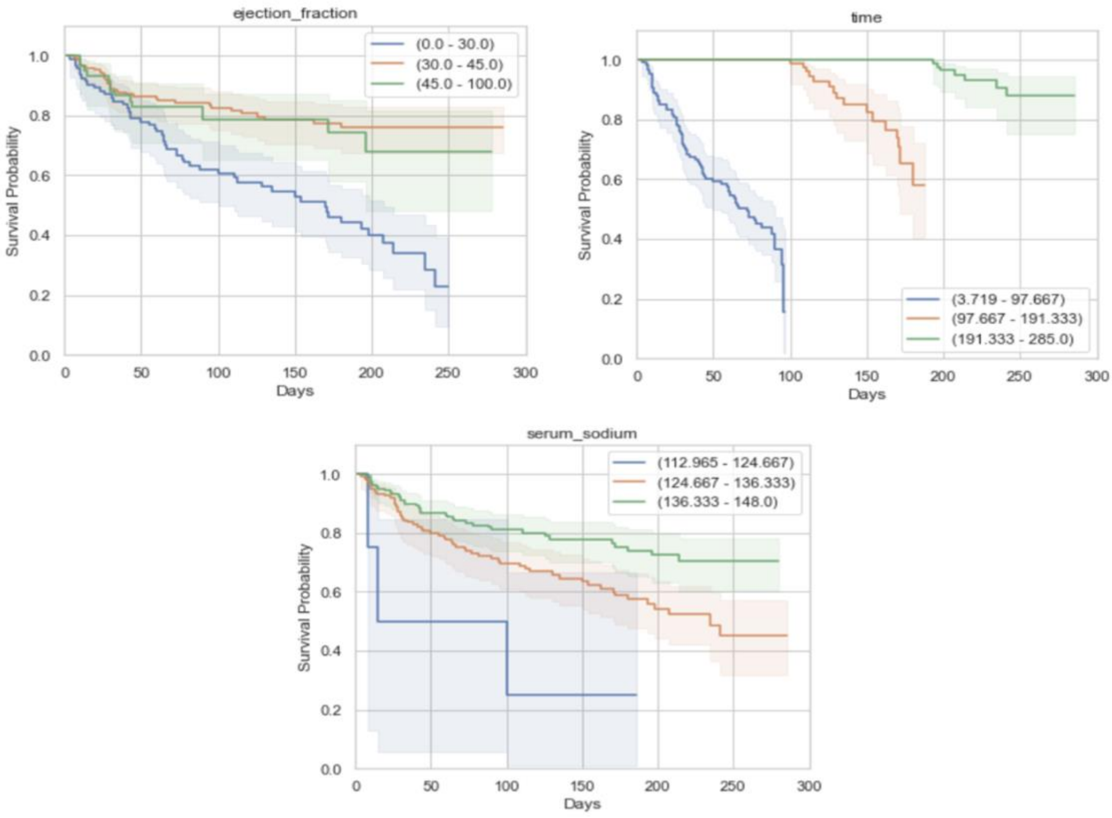


FIGURE 17. KM estimates for continuous features II

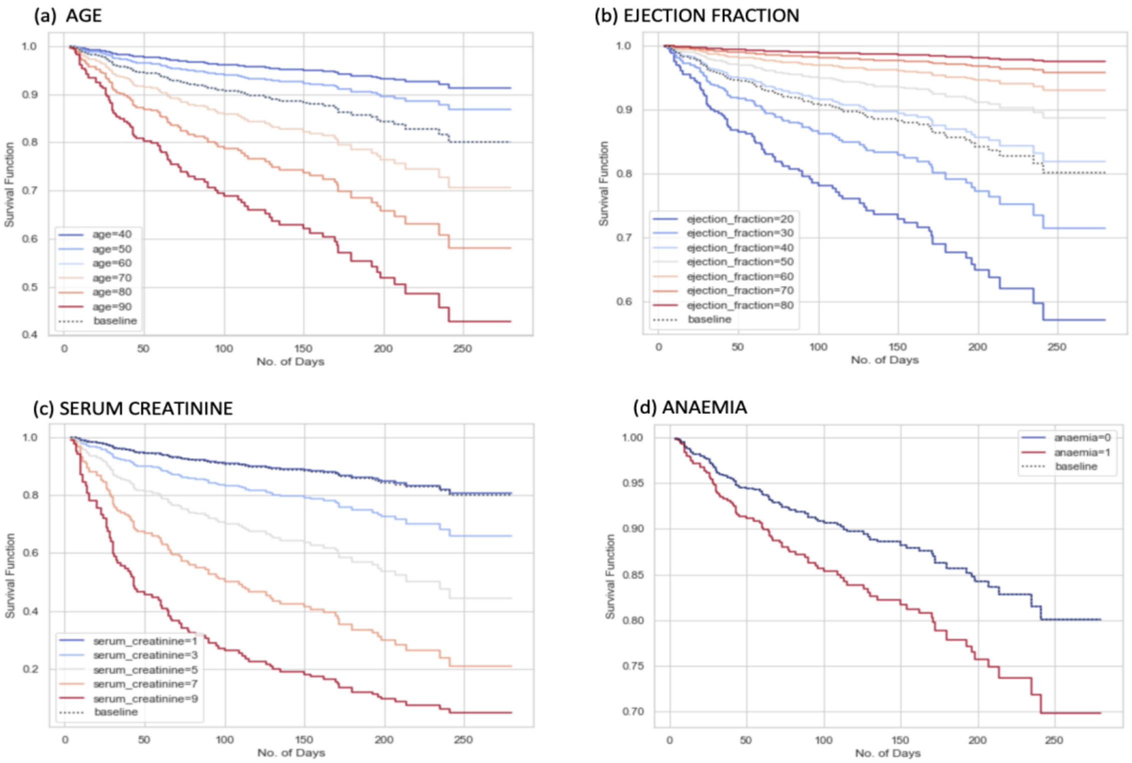
Follow-up time is highly correlated to the DEATH event because the death of a patient would directly impact the time value. As we also observe in the KM estimate plots for time, the curve with least time bucket shows a very poor chances of survival. However, this could be due to the fact that the person died and thus the follow-up time is less. For the remaining 2 buckets with higher values of time, the person has either survived the heart failure condition OR might have left the study mid-way (censored) and have not returned for a follow-up.

Thus, from the KM estimates for continuous features, Ejection Fraction, Serum Creatinine, and Time are known to have a major impact on the survival probability of patients who are at an advanced stages of heart failure condition.

2) COX PROPORTIONAL HAZARDS MODEL

The CoxPHFitter model is fitted with fitted with 294 total observations with 200 right-censored observations. Features like Age, Serum Creatinine, and Ejection Fraction are highly significant with p value less than 0.0005 making them highly correlated to the death event with 99.9995% or higher confidence level. The coefficient and the respective exponentials (hazard ratio) of the features resembles the probability increase of experiencing the death event due to a unit increase in the feature value. The hazard of death due to increasing age in heart failure conditions rises by ~3.9% with every year passing by. The probability of death event increases by 28.1% for each unit increase in the serum creatinine. Increasing age and rise in level of serum creatinine have a deleterious effect on the survival chances. However, the probability of death event decreases by ~5.2% for each unit increase in the EF level indicative of a beneficial effect. Anaemia turned out to be significant with greater than 96% confidence level and patients with anemic history stand at 61% higher risk of hazard compared to non-anemic patients. FIGURE 18 show the partial effects of the most significant features on the survival function at various levels. Smoking, Sex, Platelets, and Diabetes, all have comparatively higher p-values, have large standard errors, and correspondingly wide confidence intervals indicating that they cannot be considered significant.

The survival function for individuals in the test cohort in shown in FIGURE 19. This prediction assumes that the individual has just entered the study (i.e., it is not conditioned on how long the subject has already lived for). A few patients show moderate survival expectancy while one patient shows a very poor chances of survival. Two of the patients show very good projections of survival with high probabilities.


(a) AGE

(b) EJECTION FRACTION

(c) SERUM CREATININE

(d) ANAEMIA

FIGURE 18. Partial effects of features on survival function

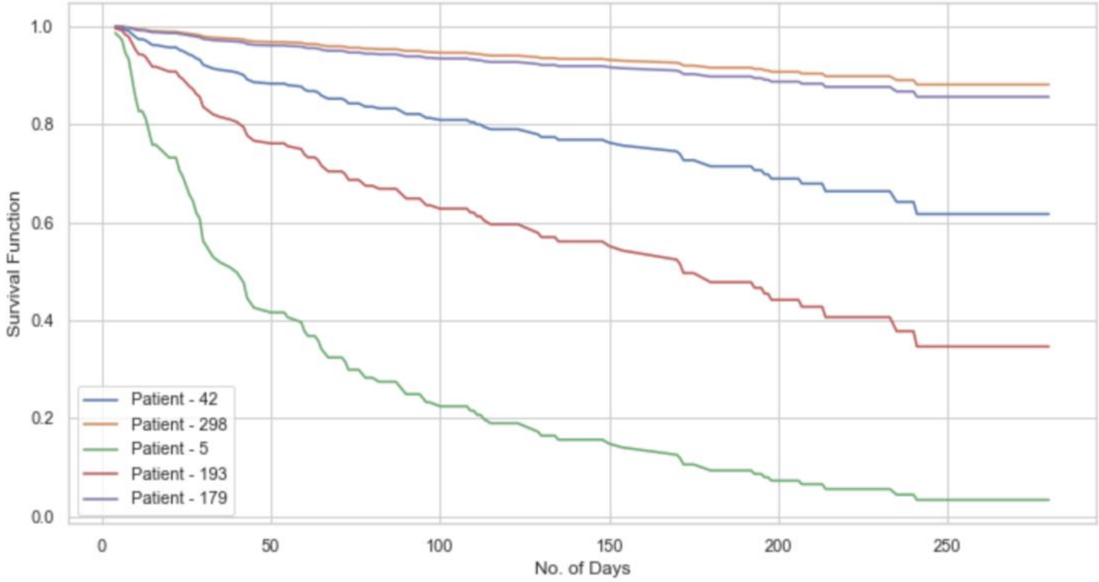


FIGURE 19. Survival function for individuals

C. SURVIVAL PREDICTION

Machine learning analysis was performed for classification of the death event due to the heart failure condition. The main features considered for the classification task were chosen as the most important features that are highly correlated to the death event identified from the exploratory data analysis and the survival analysis experiments. The highest correlated features were identified to be Age, Ejection Fraction, Serum Creatinine, and time. Time, although is highly correlated to death was not selected as a valid feature since it would lead to target leakage. Target (data) leakage happens when the training data contains any kind of information about what the model is trying to predict, and which would not be available during the test time. This is primarily a time-to-event problem where both time and DEATH_EVENT act as the target variables. DEATH_EVENT translates to whether the patient died (1) or whether the patient was censored (0) from the study. Time captures the time at which the patient either died or got censored compared to the start of the study. So, time is directly impacted by the occurrence of the event of interest and that is why we see a very high correlation among the two. We may have the time information during the training phase but in a real time scenario when making on ground predictions time as a feature will not be available. A real time use case scenario for such a situation would be – a patient goes to the physician complaining of a probable cardiac complication. The physician as part of the examination will want to run through the model to assess the severity of any complication and starts ordering tests to collect the relevant data. When it comes to record the time field the physician would not be sure about what to input since for this patient, we would not know the start of the event. Thus, time cannot be considered as a valid feature for the classification model.

Several experiments were conducted based on SVC, Decision trees, Random Forest, XGBoost, and LGBM models with a subset of data having Age, Ejection Fraction, and Serum Creatinine as features. The models were first trained in the default mode and then were put on the randomized cross validation to extract the best set of hyperparameters and model.

The results of each of the experiment is summarized in TABLE 2. Though the models seem to be performing decently the amount of data available for training and validation do not seem sufficient. More data can boost the model performance and help the model become more reliable for better on the ground performance. A detailed analysis of the implementation and cross validation can be found in the linked code repository.

TABLE 2
ML classification results

Model	Confusion Matrix			F1 Score	Accuracy	Precision	Recall	MCC	AUC ROC	AUC PR
SVM	0 1	6 2 0	3 19 1	88.37	83.33	86.36	90.48	59.22	78.6	74.2
Decision Tree	0 1	5 5 0	4 16 1	78.05	70.00	80.00	76.19	30.86	65.9	86.4
Random Forest	0 1	7 4 0	2 17 1	85.00	80.00	89.47	80.95	55.85	79.4	83.8
XGBoost	0 1	7 4 0	2 17 1	85.00	80.00	89.47	80.95	55.85	79.4	83.8
Light GBM	0 1	6 3 0	3 16 1	85.71	80.00	85.71	85.71	52.38	76.2	80.14

D. FEATURE IMPORTANCE

An additional feature ranking analysis was performed to check the importance of the variables based on each learning algorithm and as well as the measure of mutual information each feature shares with the DEATH_EVENT. Mutual information amongst two random variables is a value typically greater than zero, which measures how much a given variable can explain the other target variable. It is equal to zero iff the two variables are independent, and larger values point towards a greater degree of dependency. The most important features turned out to be serum creatinine, ejection fraction, and age. These variables contain a large amount of information regarding the death event and tend to increase the classification performance metrics. Features like sex, platelets, diabetes, anemia are not considered relevant and do not seem to contain any relevant information about the death event. This could be mainly due to the fact that the population under study fall into either class III or IV of NYHA classification for heart failure and the insignificant features are mostly known to trigger the heart failure conditions at the initial stages. FIGURE 20 shows the mutual information plot for each feature with respect to the target death event.

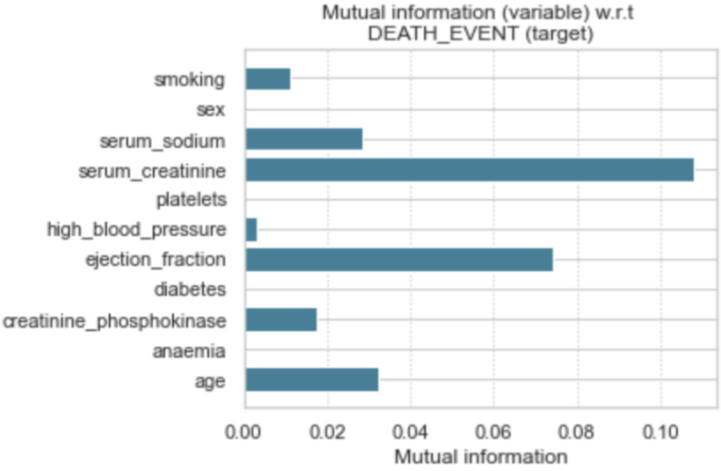


FIGURE 20. Mutual information

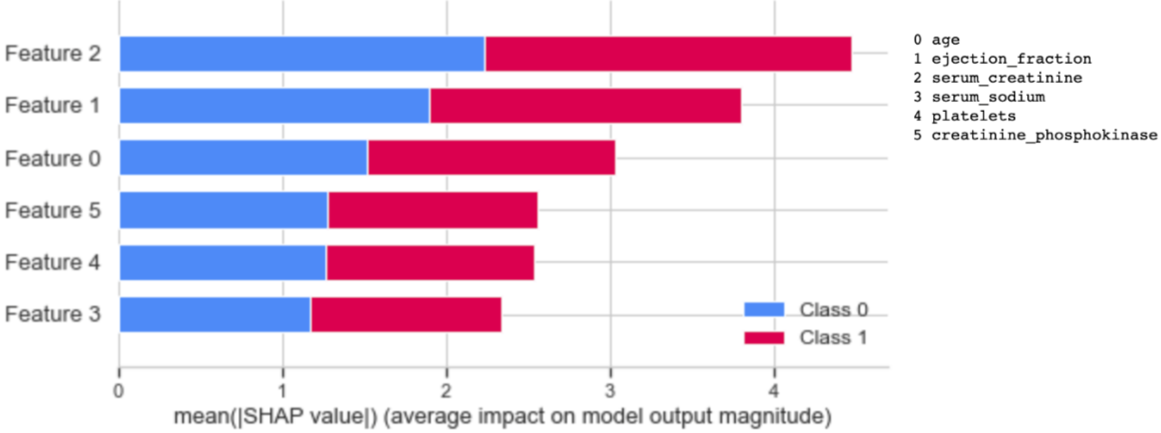


FIGURE 21. AVG. impact of features on the output

The feature importance from the ML experiments also suggests a similar ranking scheme of the features. A majority of the models assign the highest scores to serum creatinine, ejection fraction, and age. FIGURE 21 shows the average impact of the selected features on the Light GBM model output. Feature scores will help the physician interpret the results of the model and in turn contribute towards developing a better AI eco system in the healthcare domain. Visuals and feature importance like these will help answer questions like - why a particular prediction was made? Or what did the model see to arrive at the particular decision? It will narrow the gap and make acceptance of AI into the healthcare domain simpler and true-to-life.

E. RESULT COMPARISON

Several studies as already discussed have performed analysis for classification and survival analysis of heart failure cases. Still, it is not feasible and hard to compare the results of all these studies since every study is conducted based on different clinical parameters extracted from various sources. The data used in other studies are also extracted from different geographies and demographics. However, in general the findings of this study are consistent with the results of related studies that used the same data source. For example, Chicco et. al., [10] also state the top most parameters being ejection fraction and serum creatinine. Ahmad et al., [7] show a similar trend in the survival analysis segment. Sanchez [16] explainable models also show similar feature importance results with serum creatinine, ejection fraction, and age as the most important risk factors. Similar results were seen in the study conducted by Kumar et. al., [23].

V. CONCLUSION

A number of experiments were performed as part of the study. An extensive exploratory data analysis followed by a detailed survival analysis to predict time-to-event for the patients diagnosed with a heart failure condition. Survival analysis involved studying the Kaplan-Meier estimates and Cox regression models. KM estimates showed a detailed trend on the effect of each data feature has on the survival probability. Each of the feature was studied independently and it turned out that serum creatinine, ejection fraction, time, and age are the most important factors affecting the chances of survival. Cox regression also flagged out Age, Serum Creatinine, and Ejection Fraction as highly significant and strongly correlated to the death event with 99.9995% or higher confidence level. Increasing age and rising level of serum creatinine have a deleterious effect whereas a unit increase in the EF level has a beneficial effect on the survival chances. Various machine learning models were experimented for survival prediction given the health parameters. For survival prediction time was omitted from the features list for models to train on since that would introduce the problem of target leakage. ML model like SVM, decision trees, random forest, XGBoost, and Light GBM were trained on the data. The models performed decently with SVM, LightGBM, and XGBoost emerging as the top performers with F1 scores of 88.37, 85.71, and 85 respectively. SVM (5) and LGBM (6) resulted in the least number of false predictions. The analysis for feature explanation tells what patterns in the data the model is looking at to arrive at a decision. It speaks about the most important features that are involved in the assessment. Serum Creatinine, Ejection Fraction, and age hold the highest amount of mutual information with respect to the death event. Feature importance and explanations make the model transparent, explainable, and reliable which are the most essential considerations for integrating AI into the healthcare network.

All the survival projections and predictions accomplished as part of this study come with a few limitations. The population cohort has a limited number (299) of subjects being examined. The results can be enhanced by enrolling more subjects into the study. It would not only help the models to train on more diverse data but also enhance the stability and accuracy in terms of model performance and predictions in a real time healthcare setting. In the current study, the models are validated with a test set extracted out of the same population. This does not present the real estimates and capabilities of model validations. If the models were to be deployed in a different geographical region, the performance of the model would be unknown. Had there been similar data from external population belonging to different topographies available, the models could be trained and validated against those data making them more robust in their task. External validation is a very important consideration before deploying AI models in the healthcare domain. These are a few limitations to the current study.

As part of the future work, the study can be extended to support data with similar feature set from various geographies around the world to make the model robust and stable. The data used in the study has subjects who are in the Stage III & IV of NYHA heart failure classification. Datasets can also be extended in future to include various other health parameters that may have an impact during the initial stages (I & II) of heart failure as per NYHA.

To conclude, heart failure is always linked with high mortality rate, high morbidity, and the condition can lead to several chronic cardiovascular illness. Medical field generates a huge amount of data in terms of EHR/EMR records, clinical notes, medical history, pathological test results, radiological image data which when put to efficient utilization can uncover many hidden patterns that can be used to design better AI solutions to aid in the cardiovascular problem diagnosis and propose efficient treatment plans. Such smart solutions, as already discussed can reduce the risk of heart failure condition by providing accurate prognosis, hospital readmission risks, and survival projections. Technology and data can combine together to address the disparities in treatment, design better care plan, and improve health outcomes. A tactical design, development, and deployment of smart health AI solutions would advance healthcare strategies, empower physicians to look beyond the conventional practices, and eventually assist to increase the patient satisfaction levels not only in case of heart failure conditions but healthcare in general. Furthermore, high level patient education and awareness programs that communicate the benefits of a heart healthy diet and physically active lifestyle could prove to be effective and may help to reduce the mortality rate due to heart failure conditions.

Code Repository: The python codes for the experiments in this study can be found at - https://github.com/sauravmishra1710/Heart-Failure-Condition-And-Survival-Analysis.

REFERENCES

[1] N. D. WHO Team, “Cardiovascular diseases.” https://www.who.int/health-topics/cardiovascular-diseases#tab=tab_1 (accessed May 25, 2022).

[2] A. J. Coats, “The pathophysiology of chronic heart failure,” 2000.

[3] J. F. Nauta, X. Jin, Y. M. Hummel, and A. A. Voors, “Markers of left ventricular systolic dysfunction when left ventricular ejection fraction is normal,” *Eur. J. Heart Fail.*, vol. 20, no. 12, pp. 1636–1638, Dec. 2018, doi: 10.1002/EJHF.1326.

[4] A. L. Bui, T. B. Horwich, and G. C. Fonarow, “Epidemiology and risk profile of heart failure,” *Nat. Rev. Cardiol.*, vol. 8, no. 1, Jan. 2011, doi: 10.1038/NRCARDIO.2010.165.

[5] C. Bredy *et al.*, “New York Heart Association (NYHA) classification in adults with congenital heart disease: relation to objective measures of exercise and outcome,” *Eur. Hear. J. - Qual. Care Clin. Outcomes*, vol. 4, no. 1, pp. 51–58, Jan. 2018, doi: 10.1093/EHJQCCO/QCX031.

[6] National Health Service, “Heart failure - NHS.” https://www.nhs.uk/conditions/heart-failure/ (accessed May 31, 2022).

[7] T. Ahmad, A. Munir, S. H. Bhatti, M. Aftab, and M. A. Raza, “Survival analysis of heart failure patients: A case study,” *PLoS One*, vol. 12, no. 7, p. e0181001, Jul. 2017, Accessed: Feb. 07, 2022. [Online]. Available:

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

[8] T. Ashine, G. Muleta, and K. Tadesse, "Assessing survival time of heart failure patients: using Bayesian approach," *J. Big Data*, vol. 8, no. 1, pp. 1–18, Dec. 2021, doi: 10.1186/S40537-021-00537-4/TABLES/5.

[9] C. Zheng *et al.*, "Time-to-event prediction analysis of patients with chronic heart failure comorbid with atrial fibrillation: a LightGBM model," *BMC Cardiovasc. Disord.*, vol. 21, no. 1, pp. 1–12, Dec. 2021, doi: 10.1186/S12872-021-02188-Y/TABLES/5.

[10] D. Chicco and G. Jurman, "Machine learning can predict survival of patients with heart failure from serum creatinine and ejection fraction alone," *BMC Med. Inform. Decis. Mak.*, vol. 20, no. 1, pp. 1–16, Feb. 2020, doi: 10.1186/S12911-020-1023-5/TABLES/11.

[11] X. Jia, M. M. Baig, F. Mirza, and H. GholamHosseini, "A Cox-Based Risk Prediction Model for Early Detection of Cardiovascular Disease: Identification of Key Risk Factors for the Development of a 10-Year CVD Risk Prediction," *Adv. Prev. Med.*, vol. 2019, pp. 1–11, Apr. 2019, doi: 10.1155/2019/8392348.

[12] M. Cheraghi, M. Sadeghi, N. Sarrafzadegan, A. Pourmoghadas, and M. A. Ramezani, "Prognostic Factors for Survival at 6-Month Follow-up of Hospitalized Patients with Decompensated Congestive Heart Failure," *ARYA Atheroscler.*, vol. 6, no. 3, p. 112, 2010, Accessed: Jun. 09, 2022. [Online]. Available: /pmc/articles/PMC3347826/.

[13] S. E. Awan, M. Bennamoun, F. Sohel, F. M. Sanfilippo, and G. Dwivedi, "Machine learning-based prediction of heart failure readmission or death: implications of choosing the right model and the right metrics," vol. 6, no. 2, pp. 428–435, Apr. 2019, Accessed: Jun. 08, 2022. [Online]. Available: https://onlinelibrary.wiley.com/doi/full/10.1002/ehf2.12419.

[14] R. J. Desai, S. V. Wang, M. Vaduganathan, T. Evers, and S. Schneeweiss, "Comparison of Machine Learning Methods With Traditional Models for Use of Administrative Claims With Electronic Medical Records to Predict Heart Failure Outcomes," *JAMA Netw. Open*, vol. 3, no. 1, pp. e1918962–e1918962, Jan. 2020, doi: 10.1001/JAMANETWORKOPEN.2019.18962.

[15] B. Ambale-Venkatesh *et al.*, "Cardiovascular Event Prediction by Machine Learning: The Multi-Ethnic Study of Atherosclerosis," *Circ. Res.*, vol. 121, no. 9, pp. 1092–1101, Oct. 2017, doi: 10.1161/CIRCRESAHA.117.311312.

[16] P. A. Moreno-Sanchez, "Improvement of a Prediction Model for Heart Failure Survival through Explainable Artificial Intelligence," Aug. 2021, doi: 10.48550/arxiv.2108.10717.

[17] M. Tabassian *et al.*, "Diagnosis of Heart Failure With Preserved Ejection Fraction: Machine Learning of Spatiotemporal Variations in Left Ventricular Deformation," *J. Am. Soc. Echocardiogr.*, vol. 31, no. 12, pp. 1272-1284.e9, Dec. 2018, doi: 10.1016/J.ECHO.2018.07.013.

[18] R. Najafi-Vosough, J. Faradmal, S. K. Hosseini, A. Moghimbeigi, and H. Mahjub, "Predicting Hospital Readmission in Heart Failure Patients in Iran: A Comparison of Various Machine Learning Methods," *Healthc. Inform. Res.*, vol. 27, no. 4, p. 307, Oct. 2021, doi: 10.4258/HIR.2021.27.4.307.

[19] B. J. Mortazavi *et al.*, "Analysis of Machine Learning Techniques for Heart Failure Readmissions," *Circ. Cardiovasc. Qual. Outcomes*, vol. 9, no. 6, pp. 629–640, Nov. 2016, doi: 10.1161/CIRCOUTCOMES.116.003039.

[20] S. B. Golas *et al.*, "A machine learning model to predict the risk of 30-day readmissions in patients with heart failure: A retrospective analysis of electronic medical records data," *BMC Med. Inform. Decis. Mak.*, vol. 18, no. 1, Jun. 2018, doi: 10.1186/s12911-018-0620-z.

[21] J. M. Kwon *et al.*, "Artificial intelligence algorithm for predicting mortality of patients with acute heart failure," *PLoS One*, vol. 14, no. 7, p. e0219302, Jul. 2019, doi: 10.1371/JOURNAL.PONE.0219302.

[22] E. D. Adler *et al.*, "Improving risk prediction in heart failure using machine learning," *Eur. J. Heart Fail.*, vol. 22, no. 1, pp. 139–147, Jan. 2020, doi: 10.1002/EJHF.1628.

[23] D. Kumar *et al.*, "Cardiac Diagnostic Feature and Demographic Identification (CDF-DI): An IoT Enabled Healthcare Framework Using Machine Learning," *Sensors 2021, Vol. 21, Page 6584*, vol. 21, no. 19, p. 6584, Oct. 2021, doi: 10.3390/S21196584.

[24] H. Li, M. H. Hastings, J. Rhee, L. E. Trager, J. D. Roh, and A. Rosenzweig, "Targeting Age-Related Pathways in Heart Failure," *Circ. Res.*, pp. 533–551, Feb. 2020, doi: 10.1161/CIRCRESAHA.119.315889.

[25] J. B. Strait and E. G. Lakatta, "Aging-associated cardiovascular changes and their relationship to heart failure," *Heart Fail. Clin.*, vol. 8, no. 1, p. 143, Jan. 2012, doi: 10.1016/J.HFC.2011.08.011.

[26] R. Shah and A. K. Agarwal, "Anemia associated with chronic heart failure: current concepts," *Clin. Interv. Aging*, vol. 8, p. 111, Feb. 2013, doi: 10.2147/CIA.S27105.

[27] A. Taylor, "Anaemia." https://www.who.int/health-topics/anaemia#tab=tab_1 (accessed Mar. 29, 2022).

[28] A. H. Association, "How High Blood Pressure Can Lead to Heart Failure | American Heart Association," *American Heart Association*, 2016. https://www.heart.org/en/health-topics/high-blood-pressure/health-threats-from-high-blood-pressure/how-high-blood-pressure-can-lead-to-heart-failure (accessed Mar. 30, 2022).

[29] R. S. Aujla and R. Patel, "Creatine Phosphokinase," *StatPearls*, Apr. 2022, Accessed: Apr. 02, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK546624/.

[30] "CPK Test - About, Normal Range, Preparation, Test Results & More." https://www.portea.com/labs/diagnostic-tests/creatine-phosphokinase-cpk-ck-mb-bb-mm-test-83/ (accessed Apr. 02, 2022).

[31] H. C. Kenny and E. D. Abel, "Heart Failure in Type 2 Diabetes Mellitus," *Circ. Res.*, vol. 124, no. 1, pp. 121–141, Jan. 2019, doi: 10.1161/CIRCRESAHA.118.311371.

[32] G. M. Rosano, C. Vitale, and P. Seferovic, "Heart Failure in Patients with Diabetes Mellitus," *Card. Fail. Rev.*, vol. 3, no. 1, p. 52, 2017, doi: 10.15420/CFR.2016:20:2.

[33] S. Hajouli and D. Ludhwani, "Heart Failure And Ejection Fraction," *StatPearls*, Aug. 2021, Accessed: Apr. 02, 2022. [Online]. Available: https://www.ncbi.nlm.nih.gov/books/NBK553115/.

[34] "Ejection Fraction Heart Failure Measurement | American Heart Association." https://www.heart.org/en/health-topics/heart-failure/diagnosing-heart-failure/ejection-fraction-heart-failure-measurement (accessed Apr. 02, 2022).

[35] A. Strömberg and J. Mårtensson, "Gender differences in patients with heart failure," *Eur. J. Cardiovasc. Nurs.*, vol. 2, no. 1, pp. 7–18, Apr. 2003, doi: 10.1016/S1474-5151(03)00002-1.

[36] C. S. P. Lam *et al.*, "Sex differences in heart failure," *Eur. Heart J.*, vol. 40, no. 47, pp. 3859-3868c, Dec. 2019, doi: 10.1093/EURHEARTJ/EHZ835.

[37] I. Chung and G. Y. H. Lip, "Platelets and heart failure," *Eur. Heart J.*, vol. 27, no. 22, pp. 2623–2631, Nov. 2006, doi: 10.1093/EURHEARTJ/EHL305.

[38] M. K. Mojadidi *et al.*, "Thrombocytopaenia as a Prognostic Indicator in Heart Failure with Reduced Ejection Fraction," *Heart. Lung Circ.*, vol. 25, no. 6, pp. 568–575, Jun. 2016, doi: 10.1016/J.HLC.2015.11.010.

[39] M. G. Shlipak, G. C. Chertow, and B. M. Massie, "Beware the rising creatinine level," *J. Card. Fail.*, vol. 9, no. 1, pp. 26–28, Feb. 2003, doi: 10.1054/JCAF.2003.10.

[40] M. Metra, G. Cotter, M. Gheorghiade, L. Dei Cas, and A. A. Voors, "The role of the kidney in heart failure," *Eur. Heart J.*, vol. 33, no. 17, pp. 2135–2142, Sep. 2012, doi: 10.1093/EURHEARTJ/EHS205.

[41] T. B. Abebe *et al.*, "The prognosis of heart failure patients: Does sodium level play a significant role?," *PLoS One*, vol. 13, no. 11, Nov. 2018, doi: 10.1371/JOURNAL.PONE.0207242.

[42] H. J. Adrogué, "Hyponatremia in Heart Failure," *Methodist Debakey Cardiovasc. J.*, vol. 13, no. 1, p. 40, Jan. 2017, doi: 10.14797/MDCJ-13-1-40.

[43] NHLBI, "Smoking and Your Heart - How Smoking Affects the Heart and Blood Vessels | NHLBI, NIH," *National Heart, Lung, and Blood Institute*, 2022. https://www.nhlbi.nih.gov/health/heart/smoking (accessed Apr. 01, 2022).

[44] NHLBI, "Smoking and Your Heart - Smoking Risks | NHLBI, NIH," *National Heart, Lung, and Blood Institute*, 2022. https://www.nhlbi.nih.gov/health/heart/smoking/risks (accessed Apr. 01, 2022).

[45] D. Kamimura *et al.*, "Cigarette smoking and incident heart failure:

Insights from the jackson heart study," *Circulation*, vol. 137, no. 24, pp. 2572–2582, Jun. 2018, doi: 10.1161/CIRCULATIONAHA.117.031912.

[46] A. A. Ahmed *et al.*, "Risk of heart failure and death after prolonged smoking cessation: Role of amount and duration of prior smoking," *Circ. Hear. Fail.*, vol. 8, no. 4, pp. 694–701, May 2015, doi: 10.1161/CIRCHEARTFAILURE.114.001885.

[47] C. Mueller *et al.*, "Roadmap for the treatment of heart failure patients after hospital discharge: an interdisciplinary consensus paper," *Swiss Med. Wkly. 2020 5*, vol. 150, no. 5, p. w20159, Feb. 2020, doi: 10.4414/SMW.2020.20159.

[48] J. R. Agostinho *et al.*, "Protocol-based follow-up program for heart failure patients: Impact on prognosis and quality of life," *Rev. Port. Cardiol.*, vol. 38, no. 11, pp. 755–764, Nov. 2019, doi: 10.1016/J.REPC.2019.03.006.

[49] F. A. Mcalister, E. Youngson, P. Kaul, and J. A. Ezekowitz, "Early follow-up after a heart failure exacerbation," *Circ. Hear. Fail.*, vol. 9, no. 9, Sep. 2016, doi: 10.1161/CIRCHEARTFAILURE.116.003194.

[50] E. L. Kaplan and P. Meier, "Nonparametric Estimation from Incomplete Observations," *J. Am. Stat. Assoc.*, vol. 53, no. 282, pp. 457–481, Jun. 1958, doi: 10.1080/01621459.1958.10501452.

[51] M. K. Goel, P. Khanna, and J. Kishore, "Understanding survival analysis: Kaplan-Meier estimate," *Int. J. Ayurveda Res.*, vol. 1, no. 4, p. 274, 2010, doi: 10.4103/0974-7788.76794.

[52] D. R. Cox, "Regression Models and Life-Tables," *J. R. Stat. Soc. Ser. B*, vol. 34, no. 2, pp. 187–220, Apr. 1972, [Online]. Available: http://www.jstor.org/stable/2985181.

[53] C. Davidson-Pilon *et al.*, "lifelines: survival analysis in Python," *J. Open Source Softw.*, vol. 4, no. 40, p. 1317, May 2019, doi: 10.5281/ZENODO.4816284.