Heart Failure Identification and Classification using Unstructured Dataset of Cardiac Patients

Author
Muhammad Saqlain

Registration Number
NUST201464582MCEME35414F

Supervisor
Dr. Nazar Abbas Saqib

DEPARTMENT OF COMPUTER ENGINEERING
COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING
NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY
ISLAMABAD

2016

Heart Failure Identification and Classification using Unstructured Dataset of Cardiac Patients

Author
Muhammad Saqlain

Registration Number

NUST201464582MCEME35414F

A thesis submitted in partial fulfilment of the requirements for the degree of

Master of Science in Computer Software Engineering

Thesis Supervisor:

Dr. Nazar Abbas Saqib

Thesis Supervisor's Signature: ____________________________________

DEPARTMENT OF COMPUTER ENGINEERING
COLLEGE OF ELECTRICAL & MECHANICAL ENGINEERING
NATIONAL UNIVERSITY OF SCIENCES AND TECHNOLOGY
ISLAMABAD

Declaration

I certify that this research work titled “*Heart Failure Identification and Classification using Unstructured Dataset of Cardiac Patients”* is my own work. I hereby declare that I have written this thesis work totally on the basis of my individual efforts under the kind guidance of my supervisor Dr. Nazar Abbas Saqib. All resourced used in this thesis work have been cited and contents of this work have plagiarism free. No part of the work existing in this thesis has been submitted in support of any application for any other degree of qualification to this or any other institute or university of learning.

Student Signature
Muhammad Saqlain

Language Correctness Certificate

This thesis work has been read by an English expert and is free of typing, semantic, syntax, grammatical, and spelling mistakes. Thesis is also according to the format given by the university.

Signature of Student

Muhammad Saqlain

Registration Number

NUST201464582MCEME35414F

Signature of Supervisor

Dr. Nazar Abbas Saqib

Copyright Statement

Acknowledgments

All praises to Allah, the most loving and kind and the creator of this universe, for giving me courage and leading me through. This dissertation would not be possible without the help of ALLAH Almighty and after Him so many people in so many ways.

There are a number of people who generously supported me in conducting this research. Without their help in solving the smaller and bigger obstacles that emerged along the way, this work would not have been possible. I honor all of them for helping me out in my research work First of all, I would like to express my gratitude and appreciation to my supervisor Dr. Nazar Abbas Saqib, Prof. at Computer Engineering Department E&ME, for his guidance, supervision and enduring patience. I gratefully acknowledge my committee members Dr. Wasi Haider Butt and Dr. Arsalan Shoukat for their support and precious time.

I am also blessed and grateful to my parents their never ending support, love and encouragement. I honour my friends and colleagues Mazhar Hameed and Rao Muzammil Liaqat with very deep gratitude for being a support and helping companions throughout this work.

ABSTRACT

Heart Failure (HF) has become a major health problem throughout the world. Its occurrences increase with the age of patients, and it has become the main cause of the high mortality rate in elders. A successful progression and identification of HF can be very useful to lessen the social and individual problem from this syndrome. A lot of raw clinical data are available in health care institutions in the form of patients' medical reports, electronic test results and medication history. There is a need to convert this data set in electronic form and to get hidden information and patterns. It would help the medical practitioners to make earlier intelligent decisions about the risks of HF. Data mining techniques have great potential to extract these hidden information and patterns from such data set. This research study contains a real data set of cardiac disease patients from a renowned cardiology hospital in Pakistan to develop an HF identification and classification model using this data set. This data set has divided into three groups according to the patient's age, namely young, adult, and old and then further classified each age group into four classes according to present physical situation of the patients, namely normal, low risk, high risk, and critical. Latest data mining algorithms have applied to each separate class of every age group to identify and classify the HF patients. The results of this study show that Decision Tree (DT) gives the highest accuracy result of 90% and outperform all other state-of-the-art algorithms. Our proposed model correctly identifies various stages of cardiac patients for each age group and it can be very beneficial for the early detection and prediction of HF risk factors. This study also provides a summary of modern strategies for treatment of HF patients for each physical class that have appeared in the past few years.

Keywords: Classification Model, Data Mining, Decision Tree, Heart Failure, Knowledge Discovery in Data Set

LIST OF FIGURES

LIST OF TABLES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

Abbreviation	Illustration
ACEI	Angiotensin Converting Enzyme Inhibitors
ACS	Acute Coronary Syndrome
AFIC	Armed Forces Institute of Cardiology
ANFIS	Adaptive Niuro-Fuzzy Interference System
ANN	Artificial Neural Network
ARB	Angiotensin Receptor Blocker
AUC	Area under the Curve
BMI	Body Mass Index
Bpm	Beats per Minute
CAD	Coronary Artery Disease
CDSS	Clinical Decision Support System
CHIAD	Chi-squared Automatic Interaction Detection
COPD	Chronic Obstructive Pulmonary Disease
CPXR	Contrast Pattern Added Regression
CPXR (log)	Contrast Pattern Added Logistic Regression
CRT	Cardiac Resynchronization Therapy
DBP-BI	Diastolic Blood Pressure Before Infection
DBP-MA	Diastolic Blood Pressure Maximum Achieved
DNA	Deoxyribonucleic Acid

DT	Decision Tree
ECG	Electrocardiogram
ECHO	Echocardiography
EF	Ejection Fraction
EHR	Electronic Health Record
EP	Electrophysiology
FN	False Negative
FP	False Positive
GWTG	Get With the Guidelines
HF	Heart Failure
HIV	Human Immunodeficiency Virus
HR-BI	Heart Rate Before Infection
HR-MA	Heart Rate Maximum Achieved
IEHAPS	Intelligent and Effective Heart Attack Prediction System
IHDPS	Intelligent Heart Disease Prediction System
IVF	In Vitro Fertilization
Kg/m2	Kilogram per meter square
LR	Logistic Regression
LVEF	Left Ventricular Ejection Fraction
MAFIA	Maximum Frequent Itemset Algorithm
MDMP	Medical Decision Model with Pruning
mmHg	Millimeter of Mercury
NB	Naïve Bayes

NLP	Natural Language Processing
RCT	Randomized Clinical Trial
RF	Random Forest
SBP-BI	Systolic Blood Pressure Before Infection
SBP-MA	Systolic Blood Pressure Maximum Achieved
SVM	Support Vector Machine
TBI	Traumatic Brain Injury
TN	True Negative
TP	True Positive
WAC	Weight Association Classifier

Table of Contents

CHAPTER 1

1. INTRODUCTION

This chapter provides the complete introduction of the research study. The basic determination is to present the research motivation, main concepts, problem definition and research contribution.

Heart failure (HF) is a medical syndrome that is increasingly affects the cardiac patients all over the world and it has become the most costly ailment as well. It frequently leads to hazardous cardiac incidents such as heart attacks and stroke [1]. It has also become the prominent death source throughout the world, both men and women and the major cause of the economic problem for health care organizations [2]. In earlier times, mostly research studies of cardiac diseases had only focused on the elderly, but it has also an excessive effect on younger generation [3, 4]. That is the reason, latest research studies about cardiac disease have commonly defined adult and young heart failure affected patients [5, 6].

1.1. Heart Failure Significance

Heart failure (HF) has become a foremost reason for cardiovascular morbidity and mortality [7], and its occurrence is increasing day by day [8]. In 2012, the total cost for diagnosis and care of HF patients was $30.7 billion and expected to $69.8 billion till 2030 [9]. In common population, the chance of getting HF for a healthy person at 40 years of age is 1 in 5 [10]. It has become the key public health care precedence to control high HF patient's mortality rate [11]. It is the major goal for healthcare organizations to identify the cost-effective techniques to minimize the occurrence of hospitalization.

Cardiac diseases are the main killer in emerging countries like Pakistan and the situation is alarming because the number has frequently risen. 30-40% of all human deaths in Pakistan are due to cardiac diseases [12]. To overcome this problem, we should put attention to the community about this syndrome, its signs and symptoms, its risk factors, and its protective measures.

1.2. Types of Heart Failure

There are three basic types of HF syndrome as discus below.

1.2.1. Left-Sided HF

The left ventricle of heart is bigger than others so it delivers mostly power for pumping. In this type of HF, heart become unable to contract or relax normally, so it becomes difficult to supply the proper amount of blood to the whole body.

1.2.2. Right-Sided HF

Right-sided HF generally occurs due to left-sided HF. When the left side is not properly working it will increase pressure onto the right side for transferring blood back to the lungs that will ultimately badly affect the right ventricle. When the right ventricle fail to pump blood it backs in the veins and causes for swelling in ankles and legs.

1.2.3. Congestive HF

Congestive HF occurs when blood flow through the heart slows, fluid gathers in the lungs and restricts with breathing that may happen shortness of breath, particularly when someone is lying down.

1.3. Causes of HF

It is a common concept that HF happens mostly in aged people, but modern research proves it wrong because many cases of HF have found for young as well. As HF is a very heterogeneous disease so there are many causes that can lead to it. If someone has been diagnosed with any one of the following conditions, then it is difficult to prevent the onset of HF.

1.3.1. Coronary Artery Disease (CAD)

If cholesterol deposits create in arteries of the heart, then less quantity of blood will reach the muscle of the heart. This situation is known as coronary artery disease. It may cause of angina (chest pain) or even heart attack.

1.3.2. Hypertension

Hypertension can increase the risk of HF occurrence by two to three times. It puts high pressure in the blood vessels, so heart pump harder for circulating the blood in the body than normal situation. It can also make heart’s chambers weaker and larger.

1.3.3. Past Heart Attack

When the required amount of blood of the heart muscle could not reach due to blockage of the arteries then heart attack occurs. It damages or weaken the muscle tissues of heart that could not contract properly and heart lose its ability of blood pumping.

1.3.4. Diabetes

Diabetes also increases the risk of HF occurrence because it is the cause for developing atherosclerosis and hypertension. Both of these have been closely linked to HF.

1.3.5. Sleep Apnea

Sleep apnea is a sleep syndrome in which throat's tissues collapse and resist the airway. It also increases the risk of diabetes, stroke, heart failure, and high blood pressure.

1.3.6. Obesity

It has been seen that the heart pump much harder for obese person than for non-obese. Obesity can also cause of hypertension and sleep apnea, which finally develop HF.

1.4. Major Signs & Symptoms of HF

If someone has one or more following symptoms, then he should report himself to cardiologist even he has not been detected with any cardiac problem [13, 14].

1.4.1. Shortness of Breath

Breathlessness at rest, during activity, or during sleep may cause of HF. Cardiac patients often face difficulty in breathing while they are lying flat and they complain of restless and waking up tired.

1.4.2. Edema (Build-up of fluid)

If someone feels swollen in the legs, feet, ankles, or even suddenly gain weight, then he/she is facing edema which is the major cause of HF.

1.4.3. Chronic Coughing (Wheezing)

HF patients may face wheezing, that can produces pink or white blood tinged mucus. Basically this fluid creates in the lungs.

1.4.4. High Heart Rate

Cardiac patients normally face variable heart rate and even sometime feel their heart is throbbing or racing. All these situations are alarming for developing HF.

1.4.5. Tiredness (Fatigue)

If a person is feeling tired all the time and facing difficulty to do daily activities, such as walking, shopping or climbing stairs, then he must visit some heart specialist earlier because he is suffering fatigue that is one of the common causes for onset of HF.

1.5. Data Mining Methodology for Cardiac Patient's Data

Unstructured raw data collected from the patients report in the form of patient history, electronics test results, previous treatment records, and complex medical reports can be very beneficial for clinical organizations if they get successfully the important buried information from it [15], and this information will be used to develop prognostic models for medical researchers and practitioners to making intelligent decisions, cost saving, and control diseases. There is a need to advance approaches for discovering knowledge of medical care system's databases that is accessible as clinical and administrative level. Data Mining is one of the most significant approach for Knowledge Discovery in Databases (KDD) from healthcare data sets and it can apply for mining hidden patterns and extracting useful information from it [16].

Data Mining has its applications in numerous fields such as engineering, marketing, management, customer relationship, disease prediction, Web mining, and crime analysis. In disease prediction the medical researchers can identify, predict, and diagnose patients of various diseases with the help of data mining approaches. Data mining approaches used in the research study [17], to diagnose numerous diseases like cancer, hepatitis, and diabetes. We applied supervised learning data mining strategy in this study, which use a training data set to train model parameters. The importance of data mining methodologies for extraction of hidden pattern and buried information increases with data set size.

1.6. Treatment Procedure of HF in AFIC

Armed Forces Institute of Cardiology (AFIC) is a main tertiary cardiac treatment hospital in Rawalpindi, Pakistan that operates various patients of cardiology from all over the region. Patients come to hospital to pursue identify, diagnosis, and treatment of various cardiac problems. Patients' reports are kept in an unstructured format and even converted into a hard form in the department.

Cardiac patients are moved through different processes of test phases to identify their current condition. After the result of each test report further treatment decision has taken by the cardiologists. Some of important tests are given below.

1.6.1. Physical Examination

During physical examination doctor asked the patients about their medical symptoms and history. Healthcare associate checks the blood pressure, weight, and listen lungs and heart with a stethoscope for the patients.

1.6.2. Blood Tests

A sample of blood has taken from a patient's arm and forward to the lab. The results of blood tests have used to analyze of different important elements, such as albumin, sodium and potassium, and creatinine.

1.6.3. Chest X-Rays

X-ray is taken to check the size of heart or lung congestion from the front, back or even sides. It is normally taken in the radiology lab and is a painless test.

1.6.4. Electrocardiogram (ECG)

In this test small round plastic electrodes are adjusted on the patient's chest. Wires connect the electrodes to the ECG machine which records heart beats frequency and electrical conditions.

1.6.5. Exercise Stress Test

This test has taken in different situation to check physical fitness of patients. They are asked to walk on the treadmill and the speed is increased step by step. At the result breathing, heart rate, heart rhythm, how much patient tired are monitored.

1.6.6. Echocardiography (echo)

This is an ultrasound test that helps to understand the motion and structure of the heart using sound waves. A specialist moves the ultrasound device on the chest of lying patient that create clear images of the valves and chamber of the heart.

1.7. Research Motivation

There are a lot of existing databases for health care organizations in structured and structured forms such as images, radiology reports, medication profiles, signals, patient history, pathology report, and treatment records. This type of data can be very heterogeneous, complex, uncertain, and noisy [18]. Therefore, Data Mining offers some methodologies that can extract useful hidden patterns and information from such databases. These valuable patterns and information can be very helpful for clinical researchers and practitioners in making intelligent decisions and early prediction of diseases.

The heart failure risk calculation is very critical to find avoidance opportunities. The elementary steps of cardiac disease risk calculation are: track and identify the progression of cardiac disease risk cases. Today, the main concern of all major health care organizations is the high mortality rate of HF patients [11]. Due to lack of an effective means for prediction of HF, we have examined a very little development for controlling HF progression. The individual and social burden can be reduced by early identification and prediction of HF, changing lifestyle and starting defensive therapies.

1.8. Problem Definition

HF is a very complex and heterogeneous disease which is challenging to identify due to the variety of unusual symptoms. Some of the common HF risk elements are: hypertension, very low Left Ventricular Ejection Fraction (LVEF), diabetes, hyperlipidemia, anemia, family history, and smoking history. Raw data are accessible in the form of patient's medical history, complex reports, and electronic clinical test results. These clinical reports are available in structured and unstructured form. Structured data can be easily used for predictive model, but there is a need to mine valuable buried information from unstructured data because this data is very noisy, complex, and multi-dimensional.

Another major problem for HF identification is the age of cardiac patients because it does not equally affect the each age group. In earlier times, mostly research studies of cardiac diseases had only focused on the elderly, but it has also an excessive effect on young generation. That is the reason, latest research studies about cardiac disease have commonly defined adult and young heart failure affected patients. All cardiac patients appearing in hospital have different physical condition, so they cannot be equally treated. For accurate prediction and identification each age group with different physical condition should treated in a separate manner.

1.9. Research Contribution

This study proposes a data mining classification model for the identification and progression of HF in cardiac patients. It consists of a real data set of 500 cardiac patients which were admitted in the Armed Force Institute of Cardiology (AFIC), Pakistan in the last few years. We manually extract 32 important features from structured and unstructured data that were available in the form of the patient's history, test reports, medication profiles, and radiology reports. Then we implement all necessary preparation and pre-processing step to make a final database into MS Excel. Our major concerning attributes are LVEF (Left Ventricular Ejection Fraction), BMI (Body Mass Index), SBP (Systolic Blood Pressure), DBP (Diastolic Blood Pressure), and Comorbidities (Hypertension, Ischemic Heart Disease, Atrial Fibrillation, Diabetes, Chronic Kidney Disease) as these are highly impacting factors for HF.

The main contribution of this research is a classification model that classifies each cardiac patient into three age groups, namely young (<45 years), adult (45-64 years), and old (>64 years). Each age group has further divided into four classes according to the patient's current physical condition and with the help of cardiologists assign the values of major concerning attributes for each class of every group. We define classes as normal, low risk, high risk, and critical from Class 1 to Class 4. Now we have total twelve models, four in each age group and every patient of the data set belongs to a specific model. Finally, we implement state-of-the-art data mining, classification algorithms and found results for each model individually. We use accuracy as performance measurement and results of this study show that the Decision Tree algorithm outperforms all the other state-of-the-art algorithms with 90% of accuracy. On behalf of numerous cardiac research, we also provide a treatment plan for cardiac patients belonging to different models of our proposed study. This treatment plan can be very useful for patients and clinical researchers to overcome the HF syndrome.

1.10. Thesis Outline

Section 2 explains the literature review providing the works of various other researchers in data mining and biomedical. Section 3 presents the methodology followed in our research and section 4 contains the details proposed framework of this study. Section 5 contains results, discussion and comparison of decision tree with other state-of-the-art data mining classification models and a proposed treatment plan also presented in the same section. The conclusion is presented in in the last section 6.

CHAPTER 2

2. LITERATURE REVIEW

Heart failure has become more common and more costly disease these days. Cardiac diseases are the leading cause of death in USA and these are impacting the whole world economy very badly [20, 21]. Medical practitioners and researcher are focusing on various new strategies to control the mortality rate and minimizing the hospitalization due to heart failure. We found different research studies that represent the use of many data mining, classification algorithms for early prediction of various diseases. The most important classification algorithms using data mining and biomedical field are: Artificial Neural Network (ANN), Naïve Bayes (NB), Support Vector Machine (SVM), Decision Trees (DT), Logistic Regression, and Random Forest (RF).

2.1. Data Mining Strategies for Diseases Prediction: Previous Work

Data mining predictive modeling of various diseases has become a very broad area of research by using electronic health record (EHR). Researchers extract useful information and hidden patterns from the raw data of patients available in structured and unstructured forms and use them to create prediction modeling. These models can be very helpful for clinical researchers to overcome different diseases. Authors of [10], take dataset from the National Health and Nutrition Examination Survey (NHANES) and propose a classification approach using support vector machine (SVM) algorithm to identify persons having diabetes and prediabetes condition in US population cross-sectional illustrative sample. They found an area under the curve for each class 84% and 73%, respectively.

Traumatic brain injury (TBI) is a vital community health problem and a leading disability and death cause all over the world. Taslimitehrani et al. work on this syndrome and develop a new Logistic Regression (LR) model name as Contrast Pattern Aided logistic regression model called CPXR (Log) to classify TBI patients and got AUC as 93% [23]. They also explain that their CPXR (Log) model of brain injury at admission time data, including local models and patterns, and provided new predictor variable's odds ratios based on CPXR (Log), together with those whose odds ratio vary from the SLogR model created odds ratios. The results of this model outperform all the state-of-the-art regression models.

Authors of [24], propose a methodology of the decision tree (DT) and suggestion analysis for predicting the hypertension and for making a good policy of information management. This study observed the features of the knowledge detection and data mining classifiers to establish how they able to use for prediction of health consequences and deliver policy evidence for hypertension managing by using the database of the Korea Medical Insurance Corporation. Precisely, this study authorized the prediction power of data mining classification algorithms by relating the performance of two decision tree algorithms, Chi-squared Automatic

Interaction Detection (CHIAD) and C5.0, and logistic regression using the testing set of 4588 recipients and the training set of 13,689 recipients. Conflicting to the earlier study, the CHIAD algorithm achieved better results than the LR in the prediction of hypertension, and C5.0 had the lowest prediction results. Furthermore, the CHIAD association rule and algorithm also presented the segment-specific information for the target group and risk factors that can be used in a policy examination for management of hypertension.

Authors of paper [25], represent the support vector machine (SVM) algorithm as a high performance for resolving many Bioinformatics classification problems. They have presented a de novo SVM classification model miPred to report exactly the challenges for improvement of the classification accuracy of current (quasi) de novo methods. Their inclusive analysis described that it yielded significantly or comparable improved predictive performances than current classifiers for distinctive non-conserved practical pre-miRs from genomic quasi hairpins and non-pre-miRs with higher discriminative accuracy.

Data mining clustering and classification techniques have proposed for predicting health cost and to control the increasing rate of health care organization expenses [26]. They developed an algorithm based on latest data-mining techniques provide computable predictions of health costs and present, a powerful and significant tool for the estimation of health care costs. They also claim that R2, which has conventionally been used to prediction accuracy of the report, has some restrictions, and the use of more expressive error procedures, particularly considered for the application at hand, and may give improved vision into the expectation accuracy. Regardless of the relative plenty of medical information involved in their data sets, they initiate that for all but the main cost patients, major cost material was the most correct forecaster of true costs. This procedure can be used for predictions of cost for groups and individuals and as a basis for patient interference in health care supervision.

Authors of [33], suggested an improved Decision Tree algorithm for effective modeling of medical data that also include of pre-pruning as well as post-pruning. They have presented a Medical Decision Model with Pruning (MDMP) to examine the clinical protection data for controlling the making of intelligent decision. They first examine the features of clinical data by the FP-growth to remove the attribute sets of candidate, and then calculate the information achievement value of the attribute to regulate whether it is rejected or not. Then the presented MDMP accepts the trimming process in decision tree creation. Their experimentations with real clinical data have established that the presented decision tree model outperform all the other models in terms of the classification accuracy.

The authors of the research [34] develop a model to accept various approaches of machine learning to understand hidden patterns in a huge data set. Determining whether to patients operate with medically restricted prostate cancer regularly involves the urologist for classification of patients into predictable groups such as ‘recur’ or ‘remission’. In their study, they show that classification models for prostate cancer reappearance that may possibly provision the urologist’s conclusion, making can be encouraged from data using ordinary machine learning methods, providing that follow-up and repressing has been suitably measured. Finally, they present a weighting plan that permits to contain data records of non-recurrent people with small follow-up times in the modeling dataset. They used three algorithms for propose their model that is Cox regression, NB, and DT for predicting reappearance of cancer. They also provide results comparison of these algorithms.

2.2. Heart Failure Prediction Systems

Although heart failure (HF) mostly affects the elders, but the most costly and aggressive therapeutic mediations are measured in young patients [27, 28]. HF is a very heterogeneous disease and there can be many causes of it for young generation, such as coronary artery disease (CAD), myocarditis, cardiomyopathies, and alcoholic related myocardial abrasions [29]. Despite improvements in our HF understanding, its identification remains dependent on inaccurate measures that may cause of misclassification and corresponding diagnostic labels [30]. Actually, the deficiencies in existing HF classification have been suggested as a possible description for why we have realized such little improvement in developing new managements for this syndrome [31, 32].

Data mining classification approaches have been used for prediction and identification of heart diseases. Authors of research [35], provide a comparison of performance for accuracy of various algorithms such as: SVM, ANN, RIPPER, and DT. The result of this study shows that SVM outperform all the other algorithm with highest accuracy of 84% and RIPPER was a second best algorithm with an accuracy of 81%, whereas the DT shows lowest accuracy. An isolated heart disease identifying model has developed for care service and avoidance of heart failure by research study [36]. They developed a remote cardiac monitoring system for protective care, which is made of measuring tools, mobile gateways, a monitoring server and PoC tools. Mobile gateways assignment physiological indications collected from numerous measuring tools to the monitoring server and PoC tools allow physicians to observe a physiological signs of patients remotely. For the meantime, a CDSS in the checking server examines the signals constantly by taking advantage of prediction algorithms and arrhythmia detection. The CDSS also assesses medical rules based on the examination result and may attentive patients or doctors consequently. To ease defensive care, they proposed an algorithm for prediction of PAF whether onset would happen after 20 minutes. They attained the 88% of prediction accuracy by fully discovering an ECG segment detached from a PAF onset.

A research study [37], in which authors found a real data set of cardiac patients from VA Medical Center, California. They applied three algorithms LR, ANN, and DT and using lift charts and the error rate to define the performance of each algorithm. They also provide a comparison of performance to declare the best algorithm for prediction of HF. This study shows the ANN is the best algorithm with maximum accuracy and DT represent the lowest accuracy. Authors of [38], developed a prediction model for demonstration of HF patient's and define their survival risk using some common data mining classification algorithms such as LR, DT, SVM, and RF. This study shows that LR is the best algorithm with highest accuracy.

A research study [39] propose a different regression algorithm called Contrast Pattern Added Regression (CPXR) model. Their strategy was to extract valuable hidden patterns from a high dimensional and heterogeneous data set and define an individual model for each pattern. They proved that their model outperformed all the other latest data mining regression models. They also upgraded their model and introduced another regression model called Contrast Pattern Added Logistic Regression [40]. They worked on electronic health record data to propose a prediction model that predict survival risk of 1, 2, and 5-years of HF patient by using the probable loss function. Their study resulted an accuracy and AUC of 91% and 94%, respectively.

The authors of the research study [41], develop a classification model called Clinical Decision Support System (CDSS) that identify and diagnosis causes of cardiac diseases. The system hires four diverse machine learning classification algorithms (Support Vector Machine, Naïve Bayes, Decision Tree and Artificial Neural Network), combines the prediction outcomes of each classification algorithm in a collective machine, and produces additional information for diagnosis of disease. The method was accomplished with four diverse disease data sets containing about 242 cases with cardiac syndrome and the data sets were amplified by taking for classifier collective training. The trial outcome with cross-validation displays that the suggested model predicts the level of syndromes with a comparatively high accuracy of more than 94%. In specific, connection information between the main 10 proteins for cardiac disease analysis was initiated by the model as an applicable set of probable biomarkers, which currently need further medical verification.

An Intelligent Heart Disease Predication System (IHDPS) has proposed by a study [42], in which authors extract hidden patterns from a large cardiac patient's data set and relation among these patterns with heart failure. This system is created by the combination of three data mining classification strategies: ANN, NB, and DT. Using the left chart and classification metric they trained various models for prediction of HF and found a prediction model with highest accuracy. This study concluded that NB algorithm provides a more accurate prediction model by following ANN and DT, respectively.

A research study [43], in which authors proposed a prediction model of HF in cardiac patients call Intelligent and Effective Heart Attack Prediction System (IEHAPS). They extract hidden information and patterns from a complex date set of cardiac patients that mainly affect for HF. They used different prediction approaches as frequent pattern mining, clustering, and ANN. They applied a filtering technique called Maximum Frequent Itemset Algorithm (MAFIA) to filter out most important patterns and trained ANN algorithms with these patterns to predict HF patients effectively.

In research study [44], the authors have developed an Effective and Intelligent Heart Disease Prediction System in which they use CAR procedures produced by Weighted Association Classifier (WAC) to predict cardiac disease. This system is user-friendly, Web-based, and reliable. The result of this study shows that WAC is very accurate for extraction of important hidden patterns from data set. In a cardiac study [19] authors differentiate the factors of heart failure for men, women and adults. They explain that in the public setting of heart failure, men differ from women, and younger patients from elderly ones, in ways that may affect the results. Additionally, results and features of old patients differ in a progressed manner, not only dichotomously crossways the full choice of ages. Regardless of these alterations, females and the old treated with HF carvedilol in the public experience medical results like to those of males and their younger complements. These results are like to those of females and old patients' attractive carvedilol in RCTs. Their explanations of simpler HF signs and symptoms, smaller use of vital medications, and additional HF hospitalizations in females and the old propose that more aggressive and accurate treatment, with the use of suggested β-blockers, may advance their medical results. Only correctly considered RCTs, which their results maintenance, can decisively define the efficiency and protection of β-blockers in females and old with heart failure.

In a research study [55], the authors proposed a program for quality improvement using data composed as part of the GWTG-HF. They established that though age was related with diffident decrements in supreme guidelines-based heart failure treatments, use of the main indication created treatments (β-blockers and ACEI/ARB) still continued high between older HF patients, even adults of more than 85 years old. These extracted patterns vary from previous studies, screening critical drops in evidence based treatment in relative to age among HF and ACS patients and thus propose that medical researcher may have become comparatively more yielding with guidelines-based satisfying approvals for their older patients, mainly in the model of a guidelines-assessment platform.

Authors of the study [56] explain the in-hospital results and characteristic non-adherence HF patients. They found that between GWTG-HF hospitals, non-adherence patients as a feature for heart failure hospitalization inclined to be young age and extra social demographically loss. Regardless of proof of larger volume burden and lesser EF, this populace had improved in-hospital results. This lesser risk-adjusted humanity and LOS recommends that it may be calmer to become constant patients with non-adherent by reinstituting fluid and/or sodium constraint and resuming suitable clinical treatment. Non-adherence patients were similarly or more likely to accept the Joint Commission heart failure core measures at ejection. Though, charges of extra guideline-based maintenance were suboptimal, which may add to unfortunate long-term results. Given the great resource consumption in non-adherence patients, battered exertions to measure predictors of heart failure readmission and to recover rates of obedience with all indications based care for this exposed population have the probable to decrease health differences, lower hospital costs, and improve quality of care.

Authors of paper [57] have conducted worldwide assessments of cTAKES for two huge scale studies of phenotype-extraction: (1) Pharmacogenomics breast cancer classification treatment study inside the PGRN and (2) determining cardiac risk factors for a case-control level of the outer arterial syndrome using the EMR inside the eMERGE. Contract outcomes are less than 90 when associated with a gold standard of expert-abstracted, which they define in distinctly submitted. They develop a cTAKES model and extensions to the detection of syndrome development from text free neuroradiology records. They are developing a universal system assessment of cTAKES production beside a manually inattentive patient cohort gold standard identification for 25 medical research studies that will be defined in another research paper. They stretched cTAKES to contribute in the first i2b2 NLP test for the mission of recognizing the smoking status of document-level patient's and have stretched it additional to patient summarization. Some of the boundaries are a finer grained inevitability discovery and chronological determination. An autonomous assessment of cTAKES established their entrance in the second i2b2 NLP test importance cTAKES convenience to data from other organizations.

In a research paper [58] authors proposed a model which effectively extracted EF impressions and related EF standards from text free echocardiogram results. This model has a precision value of 95% and a sensitivity of 88.9% at the perception level, and precision value of 100% and a sensitivity of 98.41% at the text level. Text format may disturb model performance for text level classification as of variable applicable segment position and a greater number of segments, but this will require to be additional discovered in future research. Though, meaningful the EF helps earners define decision of treatment. Using an automatic model to detention this data is more competent than manual evaluation and progresses contact to data. Heart failure specialists may recognize significant extra use cases for this model, and upcoming research may study application approaches for this valuable tool. Classification automation of EF <40% is a vital first step in providing that serious information for quality development in the context of expressive use of electronic health records for patients with HF.

Authors of research paper [59] proposed methodology to recognize the heart failure syndrome in elderly patients having stable chronic obstructive pulmonary disease. The response rate of their study was 34% and may appear diffident but was only somewhat lesser than in studies based on population evaluating heart failure in aged people. As they engaged aged patients with constant COPD, they would suppose lower reaction rates as they asked numerous patients with pretty high levels of infirmity. Though, certainly, they calculated only a collection of the accessible patients, collection partiality appears doubtful because known and relevant cardiac risk factors of HF and co-morbidities were only somewhat lesser in contributors than in non-responders. Significantly, the medical applicability of their outcomes is high as they involved those patients who were capable to experience the pertinent diagnostic inquiries that is, patients in whom dealing is likely to be started in daily practice. In deduction, numerous easily gotten medical parameters and an uncommon additional analytical inquiries particularly, natriuretic electrocardiography and peptide and may advance the discovery of connected primary care heart failure patients with COPD. The use of these constraints should growth assurance about the judgment of heart failure and will support GPs to choose about the need for extra treatment or echocardiography in patients with COPD. This study also adds four easily measurable medical items (history of ischemic disease, high body mass index, laterally displaced apex beat and raised heart rate) offer independent analytical information about the absence or presence of attendant heart failure in the separate primary care patient having COPD.

Similar author of this research has already proposed classification and identification of HF in cardiac patients. In [45], we diagnose and identify the HF patients by using structured and unstructured data and describe the mortality rate of one year of more before the actual heart failure occurs. We also provide performance analysis of state-of-the-art data mining classification algorithms and found NB as high performance algorithms with an accuracy of 87%. Another research study [46], in which we present a framework, by using a data set of cardiac patients from AFIC, Pakistan and manually abstract the valuable hidden patterns with the help of cardiologists. After implementing all data mining pre-processing steps we applied a classification algorithm Support Vector Machine to classify our data set into 4 classes according to the physical condition of patients. Our planned classification model presented the accuracy of 82% and beat all state-of-the-art algorithms. This model with its outstanding result is very supportive for cardiologists and medical practitioners to recognize the reasons of heart failure and they can make a primary prediction and intelligent conclusions to control the situations of patients on behalf of these outcomes.

2.3. Disease Prediction Systems Using Decision Tree Algorithm

As decision the tree developed, it turned out to have lots of important features, both in the traditional areas of engineering and science and in a range of applied fields, including data mining and business intelligence. Recently, the decision tree algorithm has been implemented in various areas like medicine, disease prediction and security projects. It can be used for projects that are concentrated on either prediction or insight. Decision trees tend to meet very fast on such models, even on data sets having very many columns.

An important research article [49], which gives knowledge on gene manifestation levels of a number of genes in a cell in only experimentation. Microarray data of DNA is a powerful implement in the cancer diagnosis. Many struggles have been performed to use gene appearance profiles to progress tumor classification precision. This study provides a comparison of prediction accuracy of classes for two diverse classifiers, Genetically Evolved Decision Trees and Genetic Programming, which was carried out by the best 10 and 20 genes graded by the mutual information and t-statistic. Genetic Programming demonstrated out to be the superior classifier for this data set based on area under the curve (AUC) and overall accuracy using mutual information based on selection of features. The authors concluded that organized Genetic Programming with common information based selection of features is the most effective substitute to the present colon cancer prediction methods.

Authors of the study [50] perform prediction model on breast cancer patients and applied seven algorithms (Naïve Bayes, Artificial Neural Network, Logistic Regression, Bayes Net, and Decision Tree) in which decision tree was the best performer algorithm with 85.6% of accuracy. Another research study [51] in which authors proposed a prediction model for In Vitro Fertilization (IVF) using decision tree learning and integrating genetic algorithm. For this purpose they predict very accurate and timely IVF treatment that is important for both physicians and patients. It is very difficult for medical practitioners to diagnose patterns and automatically decide how to improve accuracy for every infertile couple. This study presents a combined intelligence technique that integrating decision learning methods and genetic algorithm for mining patterns and information from an IVF clinical database.

A research paper [52] proposed a prediction model by using a hybrid adaptive neuro-fuzzy inference system (ANFIS) and decision tree for predicting the inhibitory action of anti-human immunodeficiency virus (HIV). The decision tree algorithm has used to extract important prediction variables for this model and provide these variables as ANFIS input. The results of this model were matched with other models which represents that performs better than all of them.

CHAPTER 3

3. METHODOLOGY AND DATA PROCESSING

3.1. Methodology Used

In this research study of Heart Failure patterns and features extraction, we have followed a new defined CPXR (Contrast Pattern Added Regression) methodology. The CPXR is a regression technique that can be used to develop knowledge discovery and data mining models. It is very a successful methodology and got higher accuracy of 91% than any other data mining technique because it is based on real world and practical experience [40]. They used real data set of heart failure (HF) patients from electronic health records (EHRs) of Mayo Clinic. Their approach was to divide whole data set into 5 categories, namely demographics, lab results, vital, medication, and co-morbidities. In this way they defined the impact of each attribute of cardiac patient's data set for heart failure. As our data set is also real and it's mostly features matches with this study that's why we decide to follow CPXR methodology and we got very accurate results.

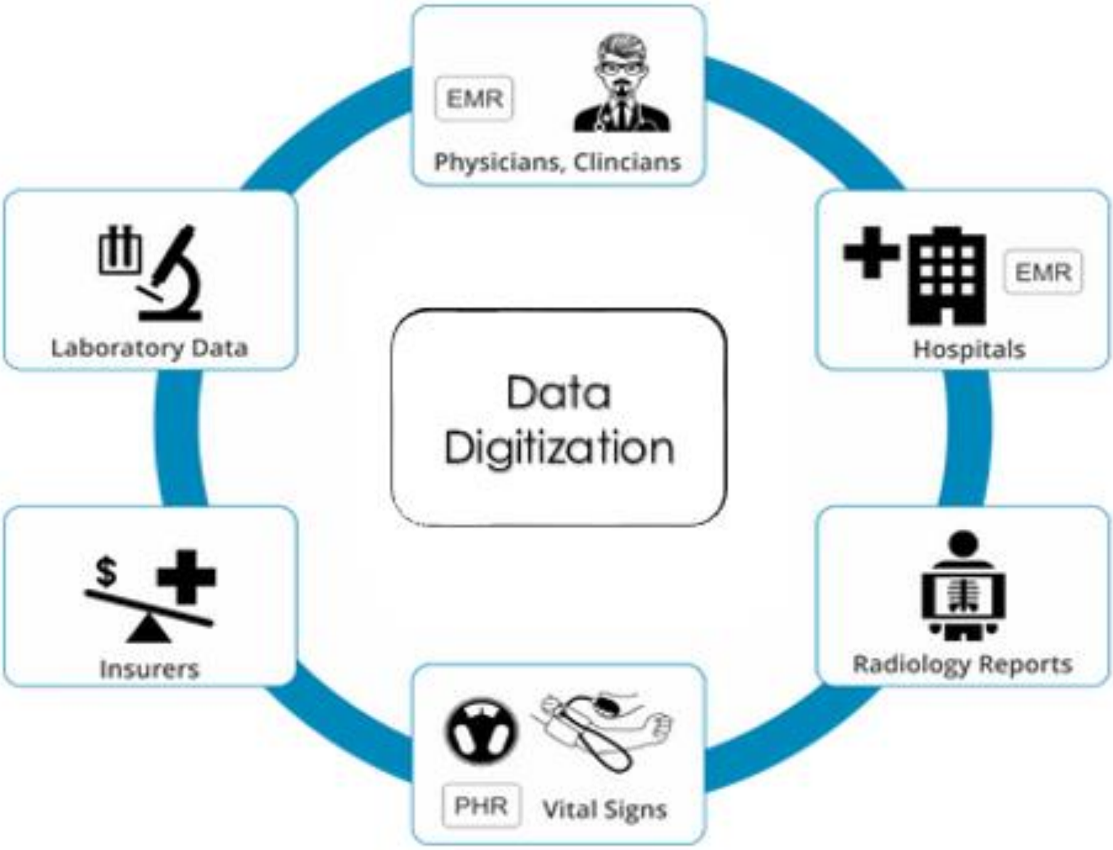


Figure 1: Digitization of dataset

3.2. Data Digitization and Pre-processing

Raw data set of 500 cardiac patients has taken from their profiles, medical reports, test results and medical history of AFIC, Pakistan. This data set is available in unstructured form which is very heterogeneous, noisy and complex. We digitize this raw data from hard form to soft form by making MS Excel files as shown in Figure 1. With the close cooperation of medical practitioners and cardiac specialists, we extract important 30 attributes of this unstructured data that have a major impact on heart failure. We build a good understanding with the help of cardiologist's bout the selected attributes from medical reports of cardiac patients and makes assured that these attributes are sufficient to get useful results for our classification model. This method provides us a deeper understanding and knowledge about cardiac diseases and helps us to recognize the problem domain of this syndrome.

To create a useful database, these extracted attributes are stored in MS Excel. Various machine learning algorithms are applied to make unstructured data set into a structured data set. Such as, mapping table has used to change textual form data into numeric form, because many classifying algorithms accept only numeric and binary data. Finally, we classify these attributes into five categories as given below and explain in Table 1.

- **Demographics**

  It contains patient's age, gender, family history and smoking history.

- **Vital**

  Vital contains results of various initial tests such as Body Mass Index (BMI), Left Ventricular Ejection Fraction (LVEF), Systolic Blood Pressure before infection and maximum achieved, Diastolic Blood Pressure before infection and maximum achieved, Heart Rate beats before infection and maximum achieved.

- **Lab results**

  It includes various lab results such as Cholesterol, Hemoglobin, Lymphocytes and Sodium.

- **Medications**

  It contains medication records from the patient's medical history such as Beta-Blocker, ACE inhibitor, Diuretic user and STATIN user.

- **Co-morbidities**

  As heart failure is a very heterogeneous syndrome and many other diseases are the main cause for HF occurrence so we select such disease from the patient's medical history. These include Hypertension, Diabetes, Coronary Artery Disease, Chronic Kidney

Disease, Atrial fibrillation, Anemia, Rheumatoid Arthritis, Cataract, Hyperlipidemia and pulmonary disease.

TABLE 1. FIVE CATEGORIES OF DATA SET

Category	Attribute	Values Variation
Demographics	Age (years)	39-89
	Gender (Male)	69%
	Family History	Present, not present
	Smoking History	Current, past, ever, never, unknown
Vital	BMI ((kg/m2))	15-31
	LVEF	18-71
	SBP-BI (mmHg)	100-200
	DBP-BI (mmHg)	70-100
	SBP-MA (mmHg)	100-180
	DBP-MA (mmHg)	70-90
	HR-BI (bpm)	52-129
	HR-MA (bpm)	53-130
Lab results	Hemoglobin (g/dL)	10.5-13
	Sodium (mEq/L)	123-132-2
	Cholesterol (mg/dL)	109-179
	Lymphocytes(*10(9)/L)	0.62-2.02
Medications	Beta-Blocker	49.0%
	ACE inhibitor	53.2%
	Diuretic	48.6%
	STATIN	49.2%
Co-morbidities	Hypertension	70.0%
	Diabetes	22.8%
	Coronary Artery Disease	37.7%
	Chronic Kidney Disease	25.8%
	Atrial fibrillation	24.6%
	Anemia	30.4%
	Rheumatoid Arthritis	23.0%
	Cataract	17.2%
	Hyperlipidemia	29.4%
	Pulmonary disease	19.6%

In pre-processing phase various sub-processes have done such as data reduction, data cleaning, and data transformation. Out of 30 attributes 17 contains binary values: including gender, family history, smoking history, all co-morbidities and medication attributes. And 13 attributes contain numerical values: including age, LVEF, BMI and lab results. Missing, inconsistent and identical values are handled by replacing and removing with correct values for the cleaning purpose of the data set. For standardization of data set another cleaning operator called Normalize has applied. In the reduction process irrelevant and redundant values have removed and final attributes are selected using Feature Subset Selection operator of machine learning. This data set is then uploaded to Rapid Miner modelling tool which one again clean the data. It uses Replace Missing Value operator to exchange the missing values with the average values of that specific attribute.

3.3. Tools used for Modelling

In this research study we used two main modelling tolls that are given below.

3.3.1. Rapid Miner 5

Rapid miner is the extensive and mostly used data mining open source solution throughout the world. It contains computer scientists, mathematicians and statisticians in the one hand, who is involved in the methodology of machine learning, data mining and statistical methods. It makes easier and possible to apply new approaches and analysis methods. It can also be used as a tool because it offers a wide variety of techniques from simple statistical calculations such as clustering, parameter optimization, regression, correlation, classification and dimension reduction [47].

3.3.2. Microsoft Excel 2010

MS Excel is used to create the initial database of this study. To represent the performance result for various data mining classification models different chart were also built with MS Excel. It is very simple to modify, filter and sort data in this tool.

CHAPTER 4

4. PROPOSED FRAMEWORK: HF CLASSIFICATION MODEL

This chapter introduces proposed classification framework of this research for identification of heart failure in cardiac patients. The main purpose of this research study is to describe a valid data mining biomedical approach based on data preparation and pre-processing from unstructured cardiac patient reports to extract useful patterns and information that will further used to classify heart failure patients according to their current physical condition and it will be helpful for medical practitioners to make intelligent decisions. This framework starts from patients report in the form of unstructured data and passing through different steps ends at useful knowledge discovery for medical practitioners as shown in Figure 2. Data preparation and pre-processing steps are already explained in chapter 3 and the remaining steps are discussed below.

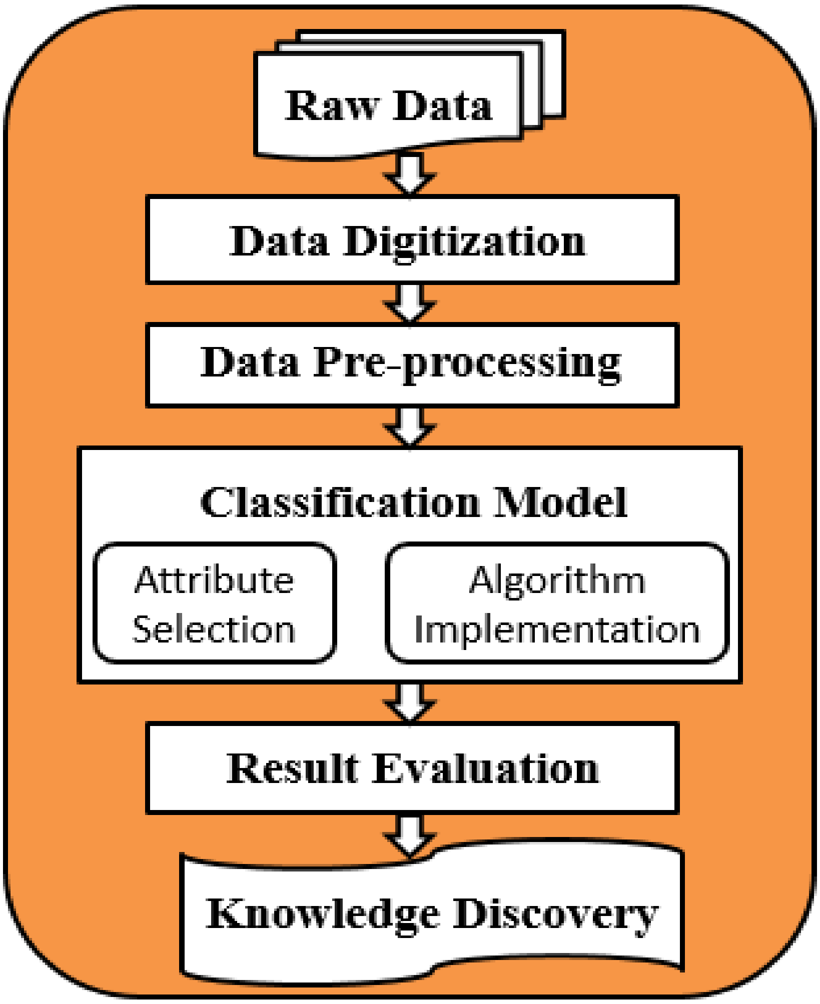


Figure 2. Step by Step Flow of Proposed Methodology

4.1. Classification Model

When we have applied all preparation and pre-processing phases to prepare a valuable structured database in MS Excel, then we upload this data set in Rapid Miner repository. Normally, heart failure cases are found in older, but in these days this syndrome is becoming very common for young generation as well. That's why we divide our database into three major groups according to the patient's age, i.e. young group (<45 years), adult group (45-64 years), and old group (>64 years). This grouping makes it easy to find the influence of heart failure for different age group separately. The significances of this grouping methodology are very important as we can compare and generate the outcomes for each age group that will also improve the accuracy of this study.

All cardiac patients having various physical condition at the time of admission in hospital. That is why we classify all age group patients into four classes according to their current physical condition. These classes describe below.

4.1.1. Class 1: Normal

These people belong to normal class as they have not found any sign and symptom of heart failure till present time. Many of them appear here for annual checkups and found clear results for each test report.

4.1.2. Class 2: Low Risk – People having the risk of HF

The people of this class also have not found sign and symptom of heart failure, but they frequently faced heart abnormality and irregularity. Such as they may have sometimes abnormal heart beat or they can face high blood pressure in different situations.

4.1.3. Class 3: High Risk – People have symptoms of HF

These are the regular heart failure patients and they normally faced some of the signs and symptoms of HF such as hypertension, high heart rate, shortness of breath and swelling in legs. They can also be affected by many other co-morbidities such as diabetes, coronary artery disease, kidney disease and high blood pressure.

4.1.4. Class 4: Critical – People reached to critical stage of HF

These people belong to critical stage of heart failure and having very rare chances of survival. They have faced many symptoms of heart failure and at the same time they have various co-morbidities as well. Many of them have also faced stroke in previous history.

By this classification our study becomes more useful and valuable because now it will focus all cardiac patients with a specific physical condition having special age group. At the end, we have total twelve models as given in Table 2. With the close collaboration of cardiologists we select five main concerning attributes such as BMI, LVEF, SBP, DBP, and Co-morbidities for each model because they are high effecting factors for heart failure. The Filter

Example operation has applied in Rapid Miner tool to select and assign special values for each main concerning attribute as explained in Table 2.

TABLE 2. CHARACTERISTICS OF HF PATIENTS ACCORDING TO PHYSICAL CONDITION

Age Group	**Attributes**	**Class 1 (Normal)**	**Class 2 (Risk)**	**Class 3 (High Risk)**	**Class 4 (Critical)**
Young (<45 Y)	BMI	18-25	25-28	28-30	>30
	LVEF	>55	45-55	40-45	<40
	S-BP	90-120	120-130	130-140	>140
	D-BP	60-80	80-90	80-90	>90
	Co-morbidities		Hypertension	Diabetes, Hypertension	Hypertension, Coronary Artery Disease
Adult (45-64 Y)	BMI	20-26	26-28	28-29	>29
	LVEF	>50	45-50	40-45	<40
	S-BP	100-130	130-140	140-150	>150
	D-BP	60-80	80-90	>90	>90
	Co-morbidities		Hypertension	Diabetes, Hypertension, Coronary Artery Disease	Hypertension, Coronary Artery Disease, Atrial Fibrillation
Old (>64)	BMI	22-24	24-26	26-28	>28
	LVEF	>45	40-45	35-40	<40
	S-BP	110-140	140-150	150-160	>160
	D-BP	70-80	80-90	>90	>90
	Co-morbidities		Hypertension, Diabetes	Hypertension, Ischemic Heart Disease	Hypertension, Ischemic Heart Disease, Chronic Kidney Disease

4.2. Implementation of Data Mining Classification Algorithms

When we have implemented all grouping and classifying phases, now we have been applying various commonly used data mining bioinformatics classification algorithms such as: Artificial Neural Network (ANN), Naïve Bayes (NB), Support Vector Machine (SVM), Decision Trees (DT), Logistic Regression, and Random Forest (RF) on each model. We preferred Rapid Miner studio to define the final results of this research study as it being used for prediction of various diseases throughout the world and it provide built in platform for all state-of-the-art classification algorithms. We uploaded the data set from a database repository by using “Retrieve” operator. Then “Set Role” operator has applied to label the predictive and result-ing attributes. In the next step we applied “Filter Example” operator to filter and to give the values of main concerning attributes. We follow the study [48] for splitting the training and testing data into 70% and 30% respectively using “Split Data” operator. Finally, we applied the data mining classifiers to find performance results for each Algorithm. We used four performance measures such as: accuracy, area under the curve (AUC), precision and recall to get overall results of this study. We represent the whole classification process in Figure 3. As decision tree (DT) is the leading result algorithm for this study so in this figure only DT algorithm has represented.

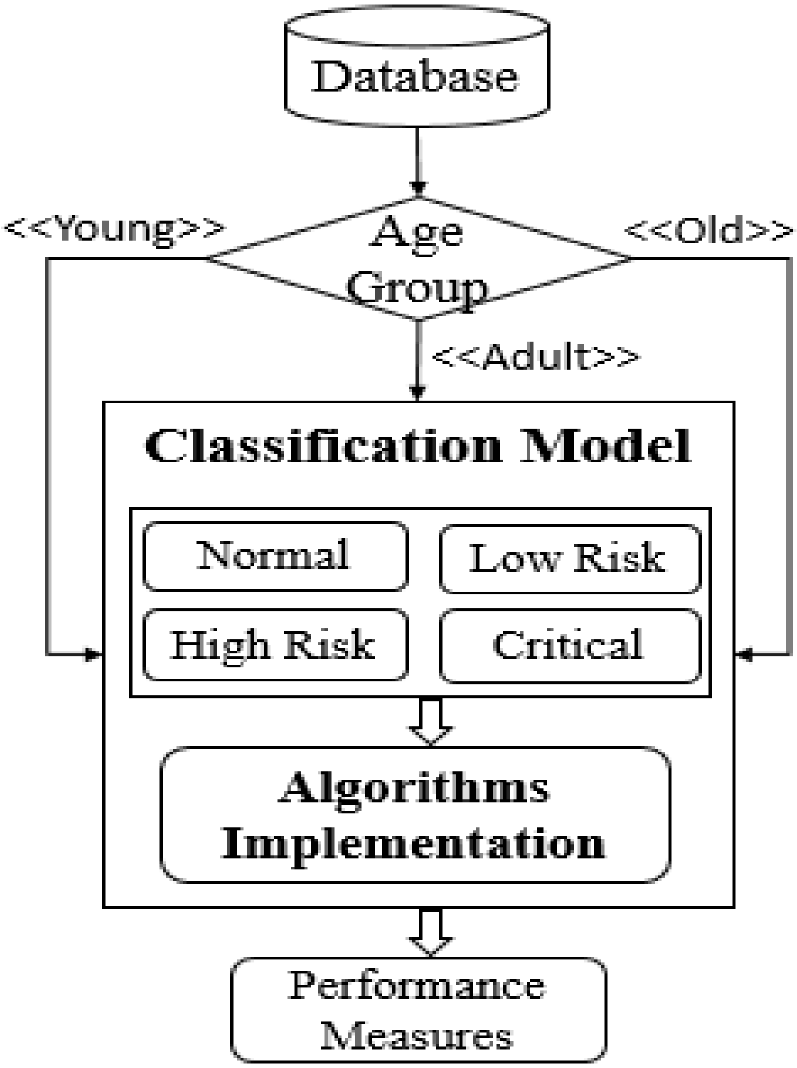


Figure 3. Classification Using Data Mining Algorithms for Each Age Group

This classification model firstly classifies the cardiac patients into three groups, young, adult and old according their ages. As heart failure affects differently for different age groups that is why we distribute data set into age groups so that each patient should be focused individually and correctly. When a specific patient has moved to its special age group, next he/she will be classified according to his/her current physical condition. For this purpose patients have to pass out from different initial test (BP, BMI and HR) and lab test (LVEF, X-Rays, Echo, ECG and blood tests) phases. On behalf of these test results patients are moved to one of the classes out of Normal, Risk, High Risk and Critical which define their current physical condition. Now we can apply any state-of-the-art classification algorithms to find the final result evaluations. Figure 4 and 5 shows the overall process implementation of the proposed classification model using Rapid Miner Studio. As Decision Tree is the best performer algorithm so we represent in this figure.

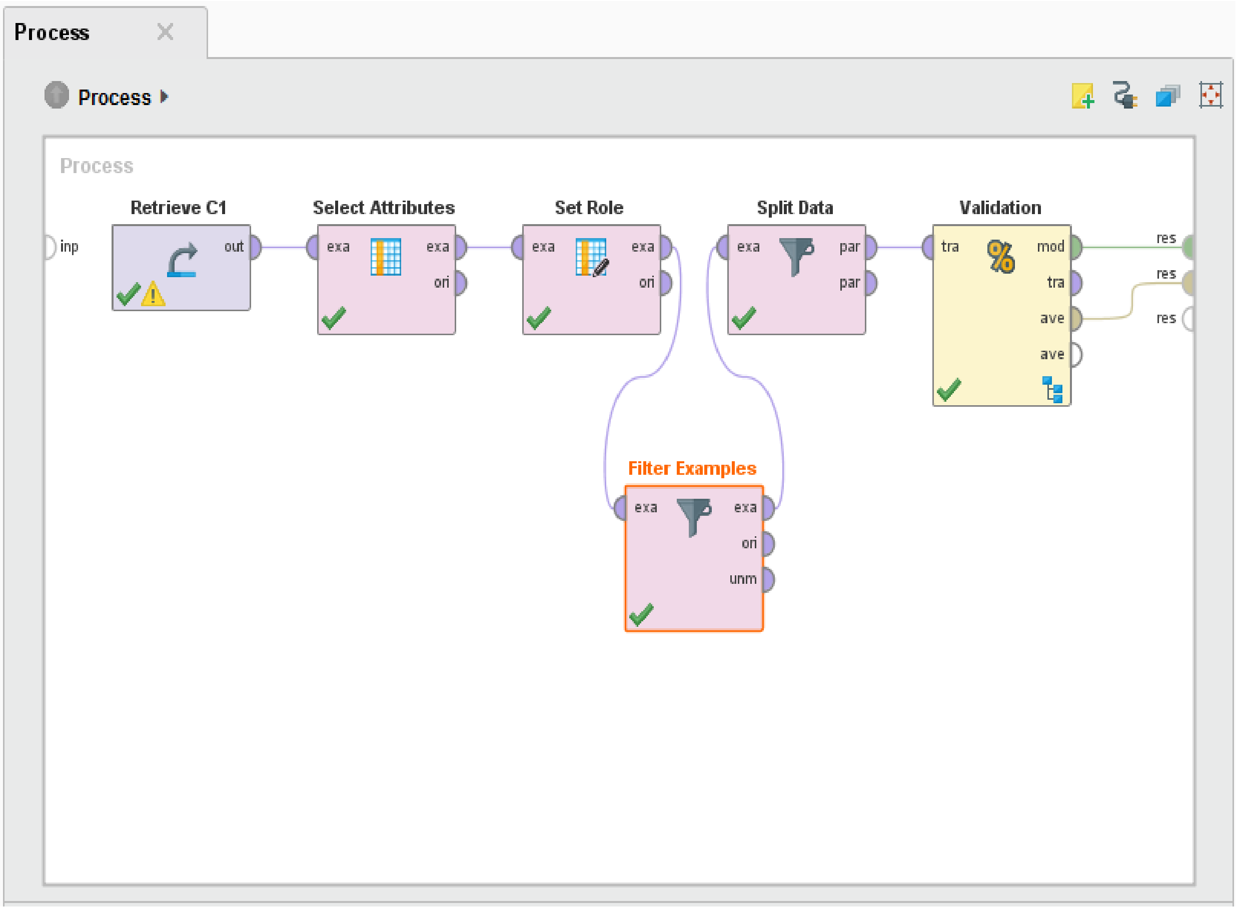


Figure 4. Rapid Miner Process Representation

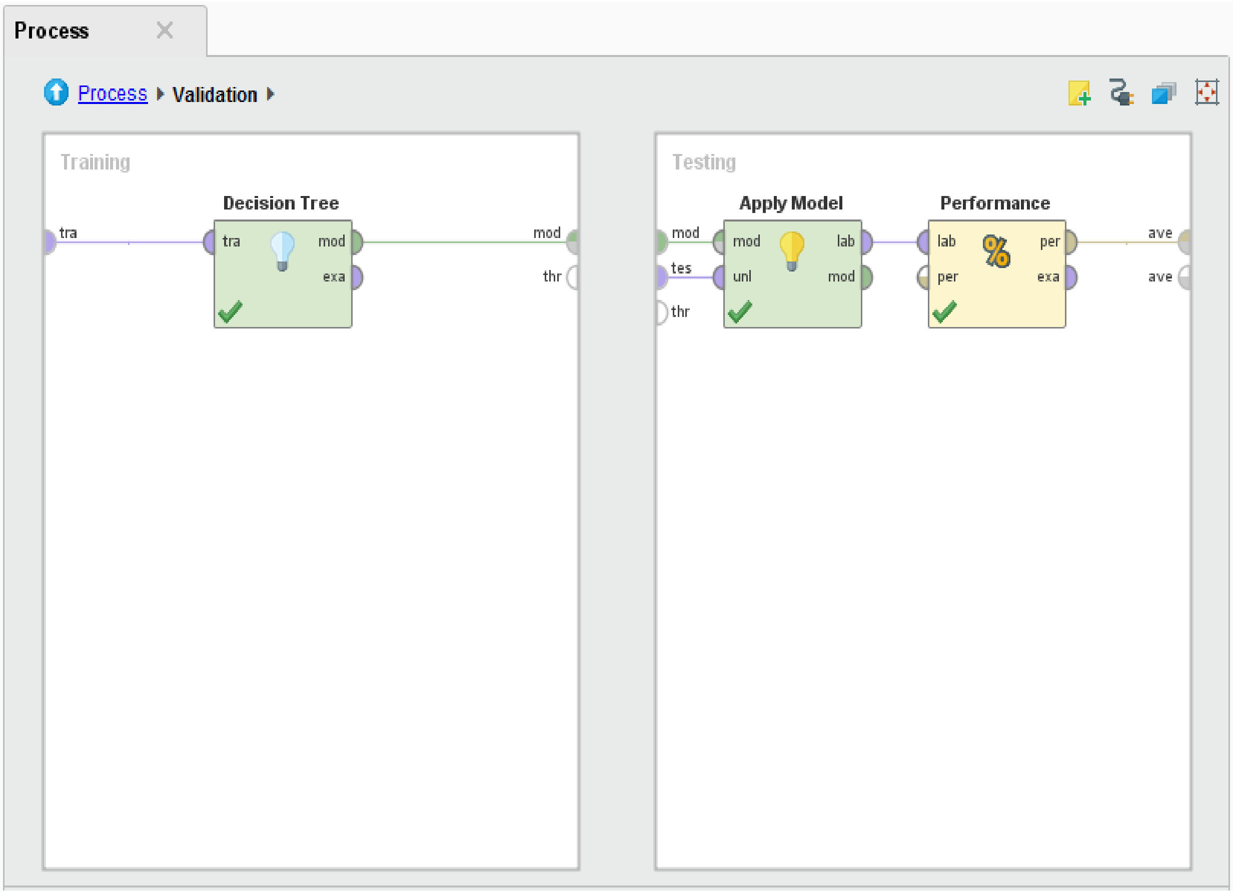
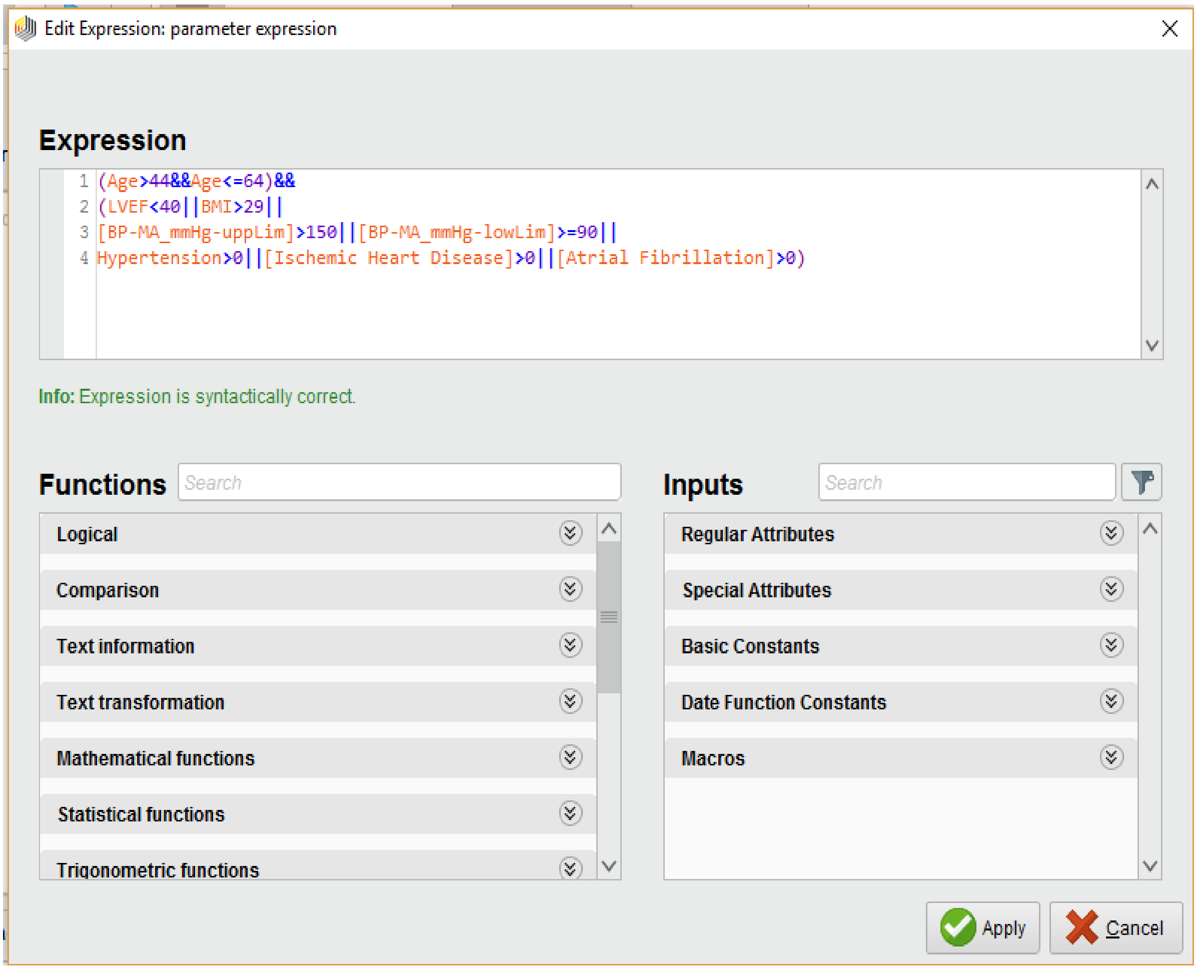


Figure 5. Decision Tree Classification Algorithm Representation

4.3. Filter Example Operator

In Figure 4, the "Filter Example" operator is very important as it has the main role for classification of our data set. We have used this operator for grouping and classifying phases according to Table 1. This operator provides a parameter expression box in which we can set all logical, mathematical and statistical functions according to our choice for each prediction variable. Main concerning attribute have selected by using input filter and then applied specific value using different functions. For example, Figure 6 shows a parameter expression box for patients who belong to an adult age group and they are facing critical stage of heart failure. We define the values of all main concerning attributes according to proposed classification model and applied various logical and mathematical functions. If there would be any syntax or logical error it will show in the red color statement and we cannot proceed it next. When all the functions are error free it represent in the green color statement "Expression is syntactically correct", and now we can apply it using the Apply button.



```
(Age>44&&Age<=64)&&
(LVEF<40||BMI>29||
[BP-MA_mmHg-uppLim]>150||[BP-MA_mmHg-lowLim]>=90||
Hypertension>0||[Ischemic Heart Disease]>0||[Atrial Fibrillation]>0)
```

Info: Expression is syntactically correct.

Figure 6. Filter Operation for Adult Age Group Having Critical HF Condition

4.4. Graphical representation of Decision Tree

The decision tree is a class of data mining methodologies that have roots in traditional statistical fields such as cognitive science and linear regression. It provides a platform in which we can develop chance events, possible outcomes and decision alternatives schematically. The graphical methodology is mostly useful in outcome dependencies and comprehending sequential decisions. DT contains various nodes that create a rooted tree which means it is a directed tree having a node with no incoming edged called “root node”. All the remaining nodes have at least one incoming edge and a node having outgoing edges is called a test or internal node. All nodes with no outgoing edge are called leaves, decision or terminal nodes.

To explain the graphical representation of the decision tree algorithm we have taken a graph of the adult age group with the risky physical condition as shown in Figure 7. Decision tree intelligently takes all those predictive variables that have a major impact on this class of cardiac patients. As left ventricular ejection fraction (LVEF) has the most impacting variable so DT makes it the root node and adjust its value according to according to Table 2. Other variables such as age, systolic blood pressure, and diastolic blood pressure are succeeded to make their position as internal nodes. Remaining variables heart rate, smoking history and family history become leave nodes. This graph presents the importance of each prediction variable from top to down as well. As LEEF is on top of the graph and behave as root node so we can say that major factor of heart failure for this class of cardiac patients. Other important factors for heart failure are systolic blood pressure, heart beat, diastolic blood pressure, family history, age and smoking history respectively. In this way we can find the sequence and the importance of each prediction variable for every model of our study individually.

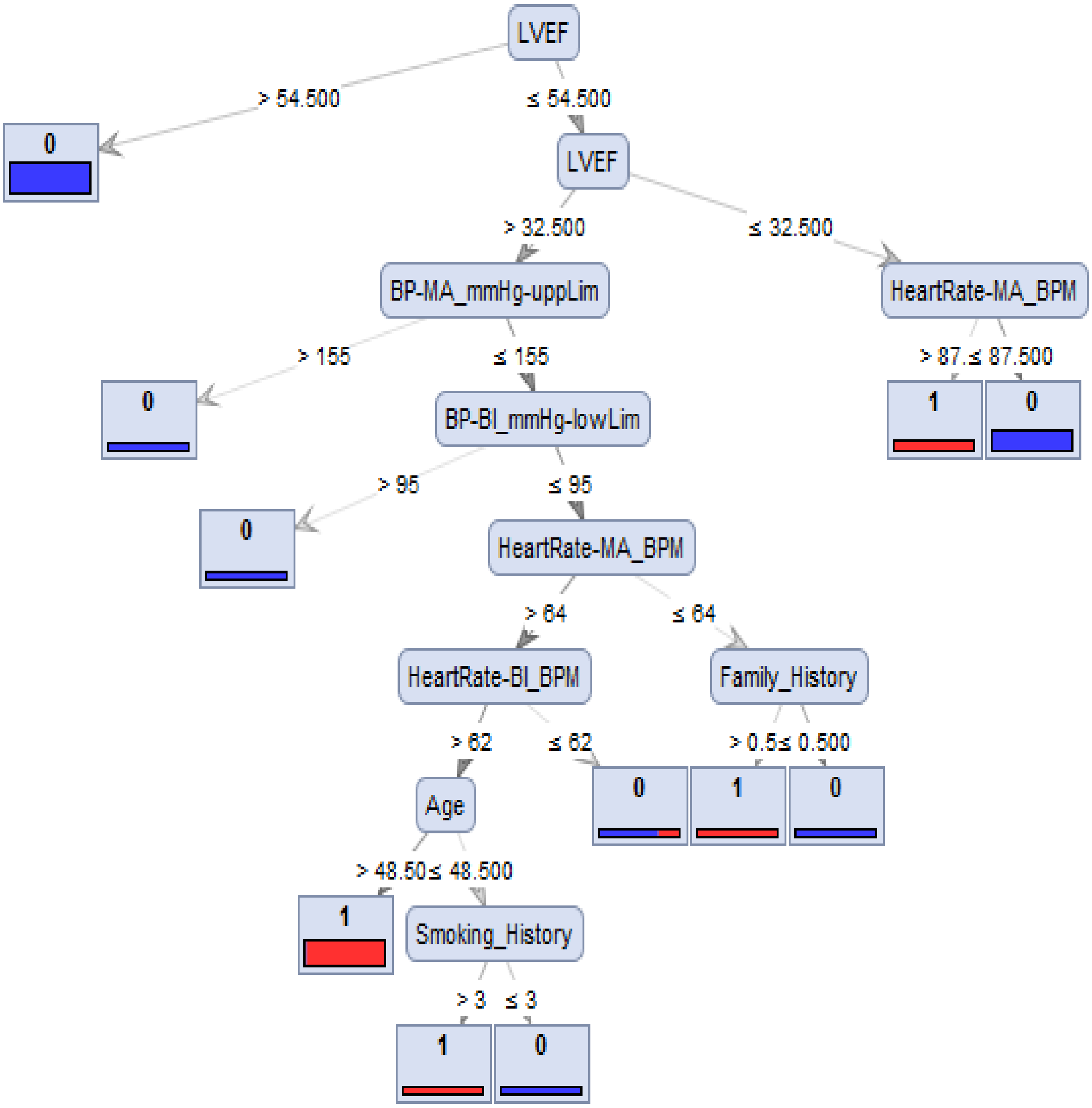


Figure 7. Graphical Representation of DT for Adult Age Group Having Risk Condition

4.5. Advantages of Using Decision Tree Algorithm

Decision tree offers some advantages over other algorithms of classification and analyzing alternatives. Here we discus some advantages and important features of the decision tree algorithm that make it more valuable than other classification algorithms.

- Decision tree freely integrates numerous levels of measurement, including quantitative and qualitative measures.
- It can freely adjust too many turns and twists in the data set such as nested effects, unbalanced effects, interactions and nonlinearities, and offsetting effects, that may commonly defeat other multi-way and one-way numeric and statistical approaches.
- We can express complex alternatives very clearly and quickly using decision trees. The modification of DT is very easy as new dataset becomes available.
- It produces results that interconnect very well in visual and symbolic terms. Decision trees are easy to understand, easy to produce and easy to use as well. One beneficial feature is the capability to integrate multiple predictor variables in a simple step-by-step manner. The capability to incrementally constructed very high complex sets of rule is both difficult and simple.
- We can represent possible outcomes, chance events and decision alternatives schematically. Its visual approach is most helpful in realizing sequential outcome and decision dependencies.
- It can be used in conjunction with various other tools of project management.
- It can handle both numerical and nominal attributes.
- It is easy to follow and self-explanatory. The decision tree is considered as very comprehensible as it can be easily converted into a set of rules.
- Decision tree can handle a dataset with missing values as well.
- Decision trees are highly robust and nonparametric and produce comparable effects irrespective of the measurement’s level of the fields that are used to build branches of the decision tree.
- Decision tree has no expectations about the class structure and space distribution, so it is considered as nonparametric technique.
- Any discrete value classifier can be represented by a decision tree.
- It is capable for management a dataset with a lot of errors.

CHAPTER 5

5. RESULT ANALYSIS AND PROPOSED TREATMENT PLAN

5.1. Performance Measure Criterions

Lot of efforts have been made to get the valuable results of proposed classification study by applying different attributes values used in modeling to start the model. Rapid Miner tool automatically calculate the results by using the values of true positive (TP), true negative (TN), false positive (FP) and false negative (FN) assessments. One by one, we upload the data set for each class into Rapid Miner tool and generate twelve models. Performance operator has been used to analyze the performance of each model. We used four performance measure criterions as given below.

5.1.1. Accuracy

Accuracy measure of a classification model can be defined as: the proportion of occurrences whose class the classifier can appropriately predict. It is the degree of confidence of measurements of a specific quantity to the true value of that specific quantity.

$$\text{Accracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} - - - - - - - - \text{(I)}$$

5.1.2. Precision

For any classification system of information retrieval and pattern recognition, precision is the fraction of recovered occurrences that are exactly relevant. It is also called positive or true predictive value.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} - - - - - - - - \text{(II)}$$

5.1.3. Recall

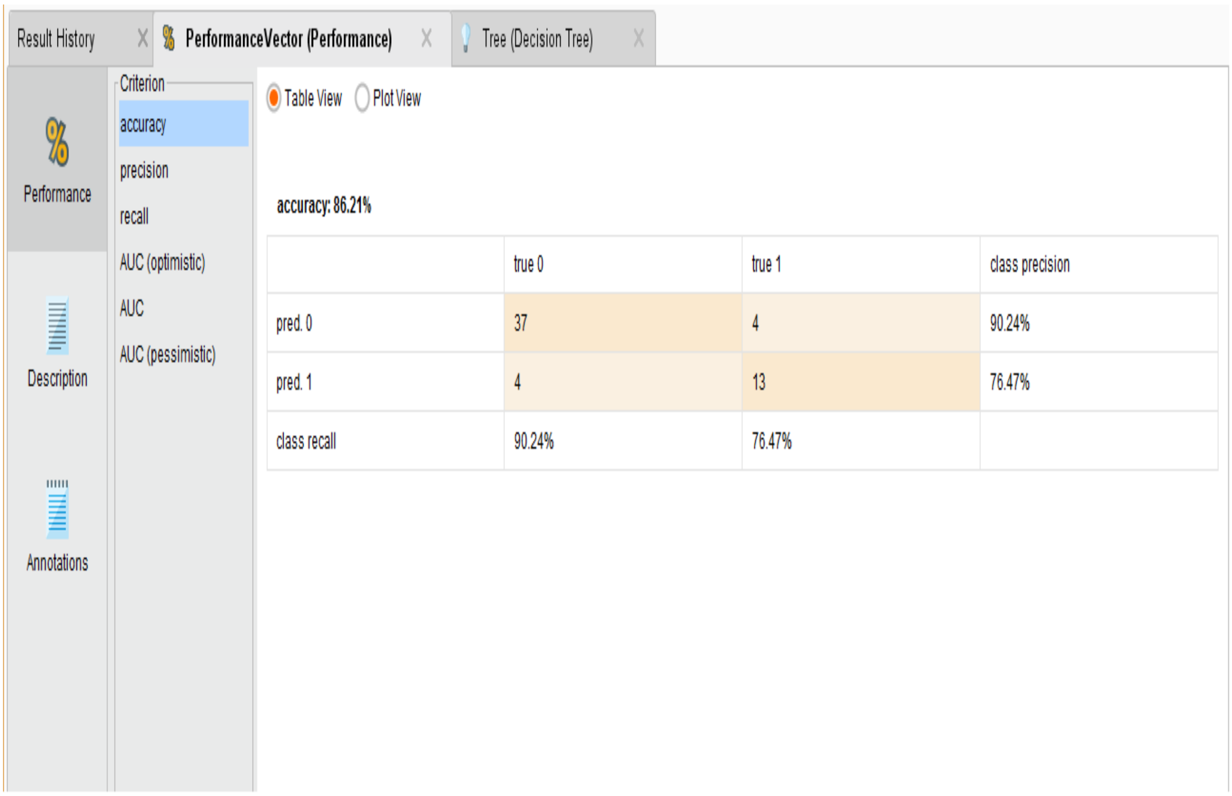
Recall is based on measure of relevance and understanding and is defined as: the fraction of significant occurrences that are retrieved. It is also called sensitivity.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} - - - - - - - - \text{(III)}$$

5.1.4. Area under the curve (AUC)

The AUC is an estimation of the probability that a classification model will rank a randomly selected positive occurrences, higher than a randomly selected negative occurrences.

The Figure 8 presents the diagrammatic view of Rapid Miner to show the decision tree algorithm’s accuracy measure of adult patient having risky condition of heart failure using TP, TN, FP and FN assessments. Here we present the table view in the same way plot view can also be seen by clicking plot view option. All the other performance measure criterions are given in the left of the table, we can view them on the same page.



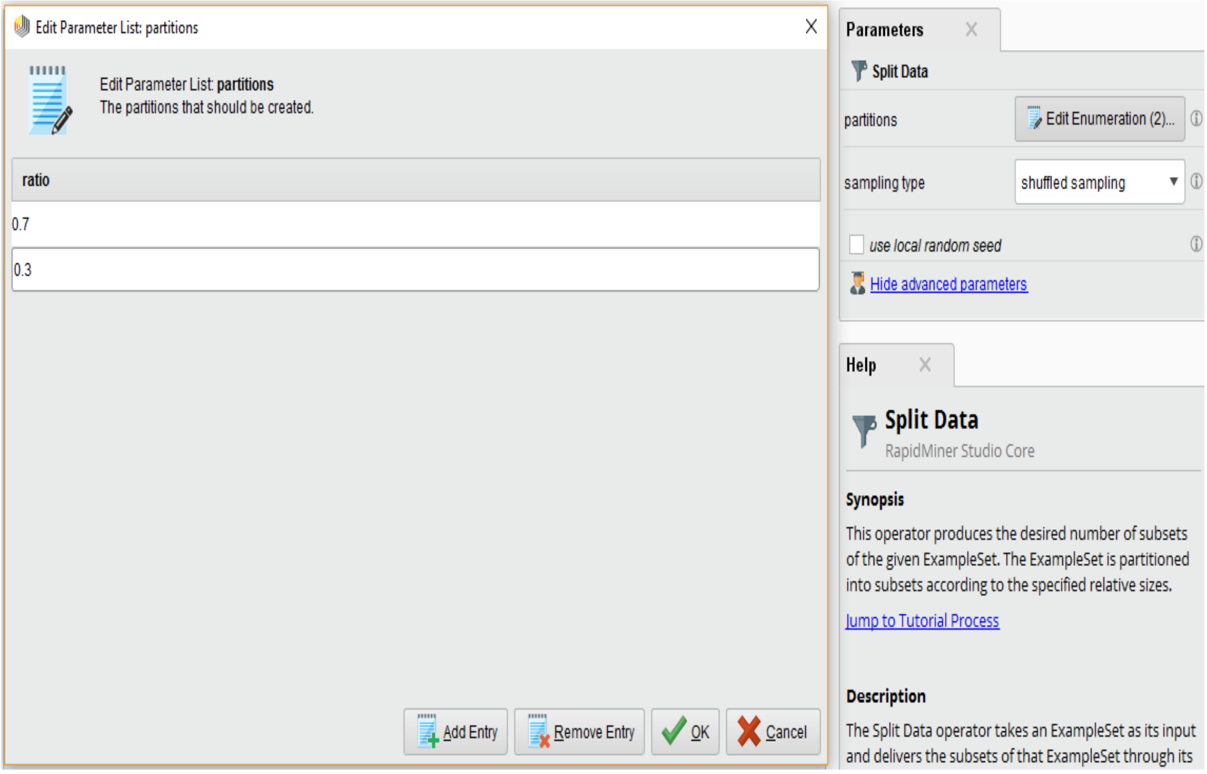
accuracy: 86.21%

	true 0	true 1	class precision
pred. 0	37	4	90.24%
pred. 1	4	13	76.47%
class recall	90.24%	76.47%	

Figure 8. Accuracy Measure of DT Algorithm for Adult Age Group Having Risk Condition

5.2. Split Validation Process

We use split validation operator as it perform simple validation by splitting data set into training and testing set and applies the model. Generally it performs split validation on unseen data set to estimate the various performance measure criterions of a learning operator. It can easily found that how precisely a model will practically perform which was trained by a specific learning operator. We split data set into 70% and 30% for training and testing set respectively as it is the default value of Rapid Miner and it is an ideal ratio to get highest accuracy for a model as shown in Figure 8. I used shuffled sampling because it builds subsets of data set randomly. Examples are also chosen randomly for creating these subsets. The whole process of split validation is done in only one iteration, as compare to another validation operator called X-validation iterates various times using numerous subsets for training and testing purposes as represented in Figure 9.



Edit Parameter List: **partitions**
The partitions that should be created.

ratio
0.7
0.3

Split Data

partitions: Edit Enumeration (2)...

sampling type: shuffled sampling

□ *use local random seed*

Hide advanced parameters

Split Data

RapidMiner Studio Core

Synopsis

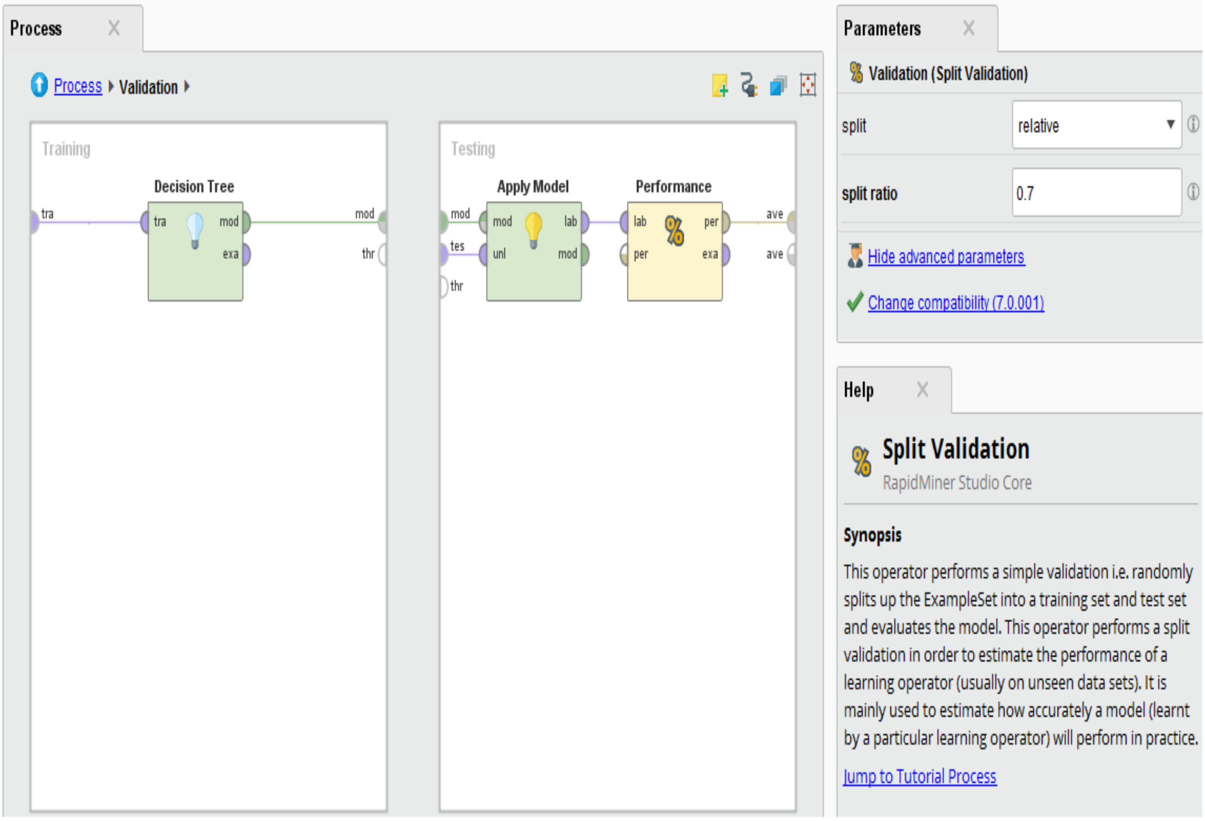
This operator produces the desired number of subsets of the given ExampleSet. The ExampleSet is partitioned into subsets according to the specified relative sizes.

Jump to Tutorial Process

Description

The Split Data operator takes an ExampleSet as its input and delivers the subsets of that ExampleSet through its

Figure 9. Splitting dataset into training and testing groups



Validation (Split Validation)

split: relative

split ratio: 0.7

Hide advanced parameters

✓ Change compatibility (7.0.001)

Split Validation

RapidMiner Studio Core

Synopsis

This operator performs a simple validation i.e. randomly splits up the ExampleSet into a training set and test set and evaluates the model. This operator performs a split validation in order to estimate the performance of a learning operator (usually on unseen data sets). It is mainly used to estimate how accurately a model (learnt by a particular learning operator) will perform in practice.

Jump to Tutorial Process

Figure 10. Single iteration Split Validation Process

5.3. Decision Tree Algorithm Performance Analysis

We one-by-one uploaded the data set of each model into Rapid Miner and applied the decision tree algorithm to found the final results of all the models. Table 3 shows the results of all performance measure criterions for each model individually. Young age group out performed all the other age groups with 94% of average accuracy. For more analysis of results, with the help of line chart we present the accuracy of the decision tree algorithm for each age group individually, as shown in Figure 9. This line chart shows that that decision tree accomplished very accurately for patients of young age when they reached at Class 3 (high risk of HF) and Class 4 (critical stage of HF). But on the same time its performance for Class 1(Normal) and Class 2 (risk of HF) is not so accurate. There can be numerous reasons for this issuer, such as it may be due to problem of splitting data into testing and training data sets or due to lack of data set for these classes. On the other hand decision tree shows the higher accuracy for Class 1 patients of adult age than all the remaining classes. Decision tree algorithm performs accurately for the patients of older age group where it shows more than 90% of accuracy of for Class 4 patients as they have reached to critical stage of heart failure and there are very rare chance of their survival.

TABLE 3. RESULTS OF DT ALGORITHM FOR ALL PERFORMANCE MEASURE CRITERION

Age Group	Performance Measure (%)	Class 1 (Normal)	Class 2 (Risk)	Class 3 (High Risk)	Class 4 (Critical)	Average Result (%)
Young	**Accuracy**	**83.3**	**91.7**	**100**	**100**	**93.8**
	Precision	100	100	0	100	75.0
	Recall	77.9	85.7	0	100	66.0
	AUC	50.0	50.0	50.0	0	37.5
Adult	**Accuracy**	**98.5**	**86.2**	**86.2**	**89.2**	**90.0**
	Precision	96.7	76.5	62.5	93.1	82.2
	Recall	100	76.5	45.5	94.7	79.2
	AUC	98.7	83.1	77.9	50.0	77.4
Old	**Accuracy**	**88.9**	**80.0**	**85.0**	**90.9**	**86.2**
	Precision	100	80.0	0	100	70.0
	Recall	71.4	80.0	0	89.5	60.2
	AUC	50.0	88.0	70.8	96,5	76.3

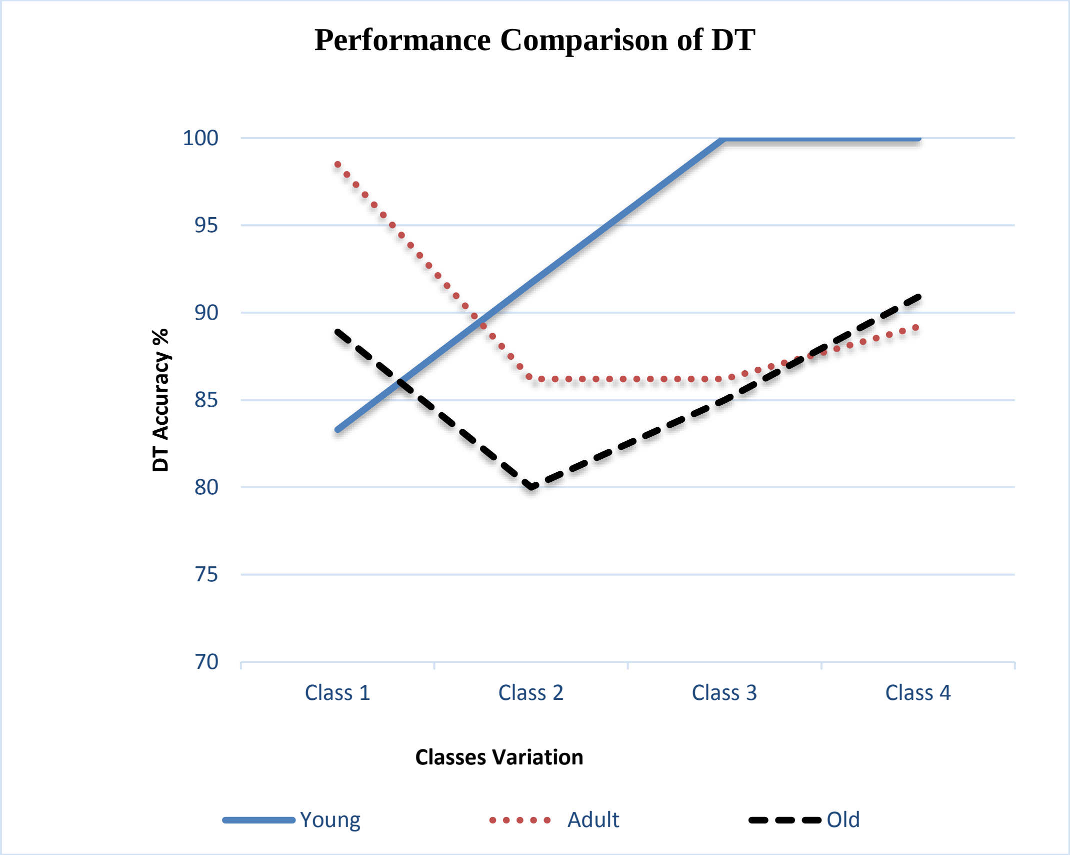


Figure 9. Accuracy Curves of Decision Tree Classification Model for each age groups.

5.4. Result Comparison for State-of-the-art Classification Algorithms

We have applied six frequently used data mining classification algorithms that are: Decision Trees (DT), Support Vector Machine (SVM), Logistic Regression (LR), Random Forest (RF), Naïve Bayes (NB) and Artificial Neural Network (ANN), and found the performance measure criterions for each classifier of every model defined in this study as presented in Table 2. Results of each classifier algorithm vary from model to model. We analyze the accuracy of each classifier for every models as given in Table 4.

Average results of classifiers show their overall accuracy for a specific age group patients. For example the random forest is the leading result classifier with average accuracy of 95.8% for the young age group patients, that's why we can prefer it whenever our concerning patient belongs to young age group. Same is the case for the patients of adult age group where decision tree outperform all the other algorithm with an average accuracy of 90% and the patients of old age groups where logistic regression is the leading result algorithm with 87.4% of average accuracy.

We have calculated the average accuracy for each classification algorithm of every age group to found the overall performance every model of our study. Table 5 presents that decision tree algorithm has outperformed all the other state-of-the-art classification algorithms with an average accuracy of 90%. That’s why we have selected it as the best performance algorithm for our research study. Decision tree algorithm is also a very good option when the prediction data set is in unstructured form because it indirectly perform feature selection and pattern extraction processes. As heart failure is a very heterogeneous syndrome and there can be numerous factors of its happening, so there are several nonlinear relationships between its parameters. Decision tree is a good performer algorithm even for nonlinear relationships parameters that’s why it gave the highest accuracy in our study [53].

TABLE 4. STATE-OF-THE-ART CLASSIFICATION ALGORITHM’S ACCURACY MEASURE FOR EACH MODEL

Age Group	**Classification Algorithm (%)**	**Class 1**	**Class 2**	**Class 3**	**Class 4**	**Average Accuracy (%)**
Young	**DT**	**83.3**	**91.7**	**100**	**100**	**93.8**
	SVM	88.2	81.2	93.8	100	90.8
	LR	58.8	75.0	93.8	0	56.9
	RF	100	83.3	100	100	95.8
	NB	100	66.7	100	100	91.7
	ANN	75.0	91.7	100	100	91.7
Adult	**DT**	**98.5**	**86.2**	**86.2**	**89.2**	**90.0**
	SVM	99.0	69.9	81.5	89.4	84.9
	LR	96.9	69.9	77.2	91.5	83.9
	RF	85.1	74.1	83.1	87.7	82.5
	NB	85.1	77.6	76.9	90.8	82.6
	ANN	86.6	62.1	75.4	87.7	78.0
Old	**DT**	**88.9**	**80.0**	**85.0**	**90.9**	**86.2**
	SVM	96.0	57.1	89.7	93.6	84.1
	LR	92.0	67.9	93.1	96.8	87.4
	RF	61.1	70.0	90.0	87.7	77.2
	NB	77.8	55.0	80.0	86.4	74.8
	ANN	83.3	65.0	90.0	87.7	81.5

TABLE 5. AVERAGE RESULT COMPARESEN OF ALL CLASSIFIERS FOR EACH AGE GROUP

Classifier	Young	Adult	Old	Average Result
DT	**93.8**	**90.0**	**86.2**	**90.0**
SVM	90.8	84.9	87.4	86.6
LR	56.9	83.9	87.4	76.1
RF	95.8	82.5	77.2	85.2
NB	91.7	82.6	74.8	83.0
ANN	91.7	78.0	81.5	83.7

5.5. Knowledge Discovery

As heart failure is a heterogeneous syndrome and there can be multiple factors and co-morbidities for its occurrence, so the data set of heart failure is also very complex and noisy. To overcome this problem we divide predictor variable of HF patient's data set into five categories including vital, demographics, medication, lab results and co-morbidities. Generally, when the number of predictor variables increase the performance of classification algorithms decreases, but the only decision tree algorithm performs positively in this situation. That's why, decision tree gave the highest accuracy than all the other state-of-the-art classification algorithms. Heart failure syndrome doesn't affect equally to all age groups cardiac patients, so this study classify them into three age groups namely young, adult and old.

Further, at the time of admission to hospital all cardiac patients have different physical condition and they are treated in different manner according to their condition. Our classification model also divide each age group cardiac patients into four classes according to their physical condition. With the help of medical practitioners and cardiologists we define the main concerning attributes for each model. Special values have been assigned to these attributes for each model to get the final results. State-of-the-art classification algorithms have been applied with different performance measures including; accuracy, precision, recall and area under the curve, but decision tree outperforms all the other algorithms with 90% of accuracy. Here are some important features and outcomes of this research study:

- *To overcome the complexity problem of the data set, it has divide into five categories namely; vital, demographics, medication, lab results and co-morbidities.*
- *Data set of this study contains 69% male cardiac patients, which mean men have more chances to get heart failure syndrome than women.*

- *Cardiac patients have divide into three age group namely; young, adult and old.*
- *Data set contains more than 70% patients with age≥50 that conclude adult and old age group patients have more chances to get heart failure.*
- *Each age group has further classified into four classes according to current physical condition of cardiac patients containing normal, risk, high risk and critical condition.*
- *70% of cardiac patients admitted to hospital having complain of hypertension, which means this syndrome has the main cause of heart failure.*
- *Cardiac patients having LVEF<50 have more chances to get this heart failure.*

5.6. Proposed Treatment Plane

On behalf of numerous cardiac studies and results of our proposed classification model, we recommend a valuable treatment plan for heart failure patients having any physical condition, explain in Figure 10. This treatment plan is very useful for physicians and medical practitioners as well as for cardiac patients. We distribute this treatment plan according to physical condition defined from Class 1 to Class 4 in this study. After following our classification model and this treatment plan, a physician and cardiologist can recommend more accurate medicine and treatments for each class, and cardiac patients can follow these medicine and treatments more confidently. An accurate heart failure classification model of cardiac patients is also very useful for clinical researchers and practitioners for developing medical trials for critical stage and higher risk heart failure patients.

This classification model presents that people of Class 1 are in normal stage but they are worry about their health due to their family history and some relevant co-morbidities that may cause of heart failure. They may are suffering by diabetes or hypertension and comes to hospital for early checkup of heart failure. Even they have no sign and symptoms of heart failure but physicians should treat their risk factors such diabetes, coronary artery disease and hypertension. They should be recommended regular checkup after every 6 months. They should be force to improve their healthy life style by regular exercise, losing body weight and quitting smoking. Authors of research study [54] claimed that in a common population of a society, nearly 56% of total population belongs to Class 1 and Class 2 of heart failure, so we should focus more both of these classes.

Patients of Class 2 are common in our society as they didn’t pay any attention for the risk of heart failure due to ignorance of this syndrome. As these patients are already suffering by diabetes, hyperlipidemia, hypertension and coronary artery disease so they have a risk of heart failure as well. Heart failure is a very heterogeneous syndrome and there are various co-morbidities and factors that may cause of it. Therefore people with such conditions should never forget the risk of heart failure and they should concern to relevant physicians and cardiologists before they got any dangerous heart failure symptom. These patients should also focus to improve their lifestyle and make sure to get regular relevant tests for heart failure after every three months. If they are facing less ejection fraction than they can use of angiotensin converting enzyme-inhibitor (ACE-I), angiotensin receptor blocker (ARB), and beta blocker. And if they have history of heart failure than they can consider electrophysiology (EP) consultation.

Class 3 contains the high risk heart failure patients that may have different signs and symptoms of HF and they are facing structural irregularities in their cardiac system. Due to ignorance of these signs and symptoms of heart failure patients could not find this syndrome on time which may lead them to any serious heart injury or stoke. Patients of this class generally face very low left ventricle ejection fraction (LVEF) which is the main cause of heart failure, so these patients should make sure of look after their LVEF values with the help of cardiologists. They should also improve their health by controlling other co-morbidities. As these patients are on the high risk of heart failure so they should be treated on emergency bases so that they couldn't reach the critical stage of heart failure. These patients should avoid the use of salt and fatty food. By concerning the cardiologists some selected patients can also be treated by cardiac resynchronization therapy (CRT) and in case of fluid overload diuretics should use to control the blood pressure.

Patients of Class 4 are in the most dangerous zone of heart failure as they have reached to the critical stage. Mostly these patients are found in adult and old age groups due to carelessness and ignorance. As these patients reached sequentially to this stage with the passage of time so there are very rare chances of their survival. Their treatment is also very expensive and time consuming for medical practitioners. They should consider use of some advance therapies such as mechanical circulatory support (MCS) and heart transplant. They should advise to use of palliative medicine.

In this way, first we present a classification model to identify the heart failure patients from an unstructured data set of cardiac patients. Then we classify these patients into three age groups and further each age group into four classes according to physical condition of these patients. Our proposed treatment plan is very helpful for clinical practitioners and cardiologists as well as for heart failure patients of any physical condition for effectively treatment of cardiac disease. This plan can also be useful to reduce the huge financial burden on healthcare organizations due to heart failure syndrome. We should publically promote this plan so that people could get awareness of this disease. If the people of initial classes follow this treatment plan, they can easily avoid to move on critical stage of heart failure and many human lives could be save.

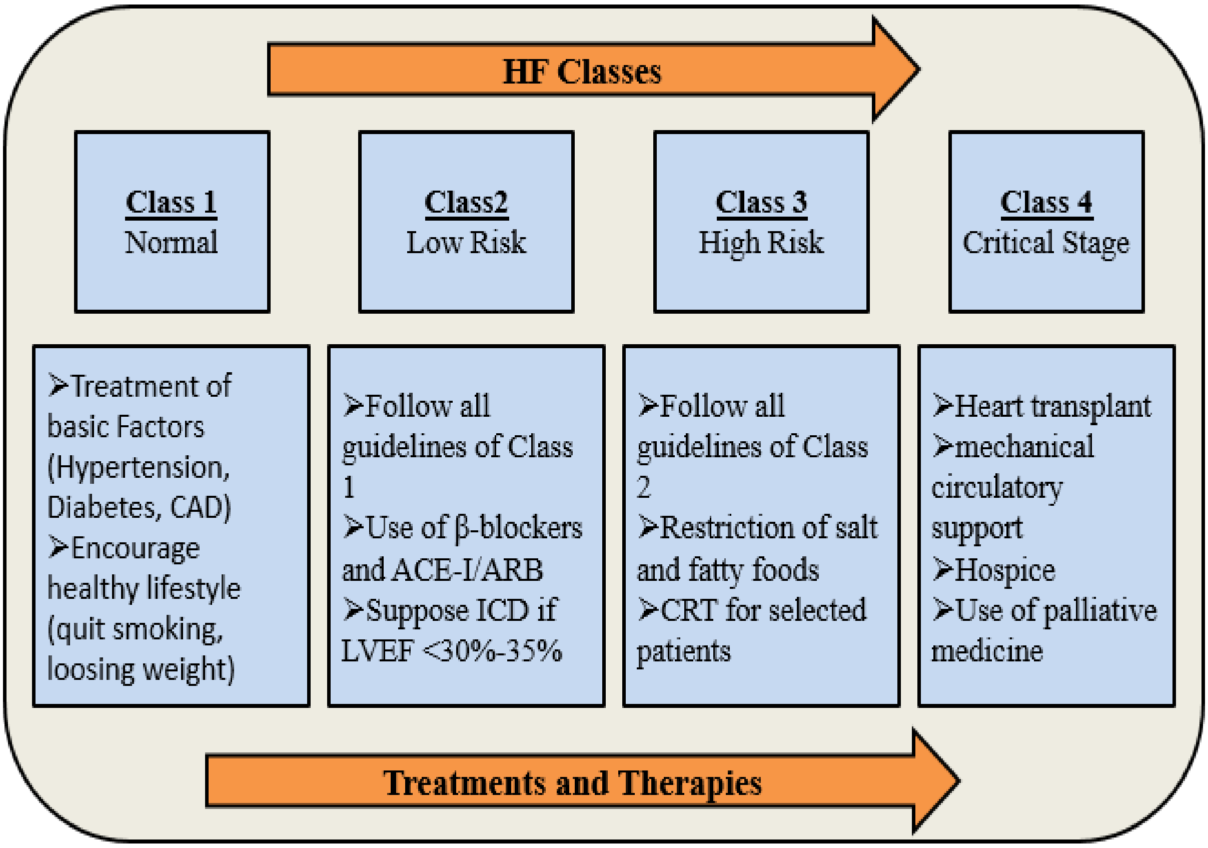

HF Classes
Class 1
Normal
Class2
Low Risk
Class 3
High Risk
Class 4
Critical Stage
➢Treatment of basic Factors (Hypertension, Diabetes, CAD)
➢Encourage healthy lifestyle (quit smoking, loosing weight)
➢Follow all guidelines of Class 1
➢Use of β-blockers and ACE-I/ARB
➢Suppose ICD if LVEF <30%-35%
➢Follow all guidelines of Class 2
➢Restriction of salt and fatty foods
➢CRT for selected patients
➢Heart transplant
➢mechanical circulatory support
➢Hospice
➢Use of palliative medicine
Treatments and Therapies

Figure 10. Proposed Treatment Plan for each class of Heart Failure

CHAPTER 6

6. CONCLUSION

In this research study we have successfully identified the heart failure patients from a real data set of 500 cardiac syndrome patients. Our proposed classification model successfully distribute cardiac patients into three age groups (young, adult and old) and each age group into four classes (normal, low risk, high risk and critical) according to physical condition of these patients. Main concerning attributes of the data set (BMI, LVEF, S-BP, D-BP and Co-morbidities) have filtered out and assigned specific values for a special patients. We used four performance measure criterions (accuracy, precision, recall and area under the curve) for six state-of-the-art classification algorithms (decision tree, support vector machine, logistic regression, random forest, naïve Bayes and artificial neural network) to find the overall results of this study. We have finally twelve models and each model describe a group of patients having special age group and belong to the specific physical class. Results of each performance measure criterions for all classification algorithms have found individually for each model. In our study Decision Tree has approved itself as a best performing algorithm among all other state-of-the-art algorithms with highest overall accuracy of 90%.

Average results describe the overall performance of all algorithms for patients of a specific age group. Results present that young age group is the top performer with average accuracy of 91%. These valuable results can very useful for successfully identification and early prediction of heart failure syndrome by clinical practitioners and cardiologists and it will help to make intelligent decisions before the actual heart failure occurs. Using this research study medical researchers can accurately diagnose heart failure syndrome for each age group patients with any physical condition and it can also helpful for progression of this disease. We also recommend a valuable treatment plan for cardiac patients having different classes according to proposed model. This flowchart is very useful for medical practitioners and cardiologists as well as for cardiac patients, because by implementing this plan they can treat cardiac syndrome more confidently and accurately. This study will help to cardiac patients of each physical class to make sure that they are not moving toward the next more dangerous class of this disease.

References

[1] Jonnagaddala J., Liaw S.T., Ray P., Kumar M., Dai H.J. and Hsu C.Y., “Identification and Progression of Heart Disease Risk Factors in Diabetic Patients from Longitudinal Electronic Health Records,” Bio Medical Research International, 2015.

[2] D. Yach, C. Hawkes, C.L. Gould and K.J. Hofman, “The global burden of chronic diseases: overcoming impediments to prevention and control,” The Journal of the American Medical Association, vol. 291, pp. 2616–2622, 2004.

[3] Cleland J.G., Swedberg K. and Follath F., “The Euro Heart Failure survey program, a survey on the quality of care among patients with heart failure in Europe. Part 1: patient characteristics and diagnosis,” Euro Heart Journal, vol. 63, 2003.

[4] Roger V.L., Go A.S. and Lloyd-Jones D.M., “Heart disease and stroke statistics–2012,” A report from the American Heart Association, vol. 220, 2012.

[5] Mogensen U.M., Ersboll M. and Andersen M., “Clinical characteristics and major comorbidities in heart failure patients more than 85 years of age compared with younger age groups,” Euro Journal Heart Fail, vol. 23, 2011.

[6] Muntwyler J., Cohen-Solal A., Freemantle N., Eastaugh J., Cleland J.G. and Follath F., “Relation of sex, age and concomitant diseases to drug prescription for heart failure in primary care in Europe,” Euro Journal Heart Fail, vol. 8, 2004.

[7] McCullough P.A., Philbin E.F. and Spertus J.A., “Confirmation of a heart failure epidemic: findings from the Resource Utilization among Congestive Heart Failure (REACH) study,” 2002.

[8] McMurray J.J., Adamopoulos S. and Anker S.D., “ESC guidelines for the diagnosis and treatment of acute and chronic heart failure 2012: the Task Force for the Diagnosis and Treatment of Acute and Chronic Heart Failure 2012 of the European Society of Cardiology,” 2012.

[9] S.M. Dunlay, N.L. Pereira and S.S. Kushwaha, “Contemporary strategies in the diagnosis and management of heart failure,” Mayo Clinic Process, vol. 89, 2014.

[10] W. Ouwerkerk, A.V. Adriaan and A.H. Zwinderman, “Factors influencing the predictive power of models for predicting mortality and/or heart failure hospitalization in patients with heart failure,” JACC: Heart Fail, vol. 2, pp. 429-436, April 15, 2014.

[11] Y. Gerber, S.A. Weston, M.M. Redfiled, A.M. Chamberlain, S.M. Manemann, J.M. Killian and V.L. Roger, “A contemporary appraisal of the heart failure epidemic in Olmsted County, Minnesota, 2000 to 2010,” JAMA Internal Med, April 20, 2015.

[12] Cardiovascular-diseases, http://www.thenews.com.pk/Todays-News-6-134656-Cardiovascular-diseases-claim-200000-lives-annually-in-Pakistan

[13] K. Aziz, S. Aziz, “Evaluation and Comparison of Coronary Heart Disease Risk Factor Profiles of Children in a Country with Developing Economy”.

[14] K. Rajeswari, V. Vaithiyanathan and P. Amirtharaj, “Prediction of Risk Score for Heart Disease in India Using Machine Intelligence,” International Conference on Information and Network Technology, vol.4, pp. 18-22, 2011.

[15] Tan G. and Cbye H., “Data mining applications in healthcare,” Journal of Healthcare Information Management, vol. 19, 2004.

[16] Giudici, P, “Applied Data Mining: Statistical Methods for Business and Industry,” New York: John Wiley, 2003.

[17] Rafael S. Parpinelli, “An Ant Colony Based System for Data Mining: Applications to Medical Data,” GECCO, pp. 791-797, 2001.

[18] Cios K.J. and Moore G.W., “Uniqueness of medical data mining” Intell Med, vol. 26, pp. 1-24. 2002.

[19] Franciosa J.A., Nelson J.J. and Lukas M.A., “Heart failure in community practice: relationship to age and sex in a beta-blocker registry,” Congest Heart Fail, vol. 12, pp. 17-23, 2006.

[20] Trogdon J.G., Finkelstein E.A., Nwaise I.A., Tangka F.K. and Orenstein D., “The economic burden of chronic cardiovascular disease for major insurers,” Health Promot Pract., vol. 8, 2007.

[21] Cohen J.W. and Krauss N.A., “Spending and service use among people with the fifteen most costly medical conditions,” 1997 Health Aff (Millwood), vol. 138, 2003.

[22] Y. Wei, T. Liu, R. Valdez, M. Gwinn and M.J. Khoury, “Application of support vector machine modeling for prediction of common diseases: the case of diabetes and prediabetes,” BMC Med. Inform. Decis. Mak., vol. 16, 2010.

[23] V. Taslimitehrani and G. Dong, “A new CPXR based logistic regression method and clinical prognostic modeling results using the method on traumatic brain injury,” in: Proc. of IEEE International Conference on Bioinformatics and Bioengineering, pp. 283–290, November 2014.

[24] Y. M. Chae, S. H. Ho, K. W. Cho, “Data mining approach to policy analysis in a health insurance domain,” International journal of medical informatics,” vol. 62, pp.103-111, 2001.

[25] Ng K.L. and Mishra S.K., “De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures,” Bioinformatics, vol. 23, 2007.

[26] D. Bertsimas, M. V. Bjarnadottir and M. Kane, “Algorithmic prediction of health-care costs,” Operations Research, vol. 56, pp. 1382-1392, 2008.

[27] Moreno S.G., Novielli N. and Cooper N.J., “Cost-effectiveness of the implantable Heart Mate II left ventricular assist device for patients awaiting heart transplantation,” Journal of Heart Lung Transplant, vol. 8, 2012.

[28] Mulloy D.P., Bhamidipati C.M., Stone M.L., Ailawadi G., Kron I.L. and Kern J.A., “Orthotropic heart transplant versus left ventricular assist device: a national comparison of cost and survival,” Journal of Thorac. Cardiovasc. Surg., vol. 73, 2013.

[29] Bibbins-Domingo K., Pletcher M.J., Lin F., Vittinghoff E., Gardin J.M., Arynchyn A., Lewis C.E., Williams O.D. and Hulley S.B., “Racial differences in incident heart failure among young adults,” N. Engl. J. Med., vol. 360, 2009.

[30] Loscalzo J., “Personalized cardiovascular medicine and drug development: time for a new paradigm,” Circulation, vol. 45, 2012.

[31] Butler J., Fonarow G.C. and Gheorghiade M., “Strategies and opportunities for drug development in heart failure,” J. A. M. A., vol. 4, 2013.

[32] Vaduganathan M., Greene S.J., Ambrosy A.P., Gheorghiade M. and Butler J., “Disconnect between phase II and phase III trials of drugs for heart failure,” Nat Rev Cardiol., vol. 10, 2013.

[33] G. Duan, D. Ding, Y. Tian, X. You, “An Improved Medical Decision Model Based on Decision Tree Algorithms,” International Conferences on Big Data and Cloud Computing, IEEE, 2016.

[34] B. Zupan, J. Demsar, M.W. Kattan, J.R. Beck and I. Bratko, “Machine learning for survival analysis: a case study on recurrence of prostate cancer,” Artif. Intell. Med., vol. 20, pp. 59-75, 2000.

[35] Milan Kumari and Sunila Godara, “Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction,” IJCST, vol. 2, 2009.

[36] Kwon K., Hwang H., Kang H., Woo K. G. and Shim, K., “A remote cardiac monitoring system for preventive care,” In ICCE, Proc. IEEE, pp. 197-200, 2013.

[37] Kurt I., Ture M. and Kurum A. T., “Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease,” Expert Systems with Applications, vol. 34, pp. 366-374, 2008.

[38] M. Panahiazar, V. Taslimitehrani, N. Pereira and J. Pathak, “Using EHRs and machine learning for heart failure survival analysis,” Studies in health technology and informatics, Med. Info., vol. 216, pp. 40-44, 2015.

[39] G. Dong and V. Taslimitehrani, “Pattern-aided regression modeling and prediction model analysis,” IEEE Transactions on Knowledge Data Engineering, vol. 27, pp. 2452–2465, September 2015.

[40] V. Taslimitehrani and G. Dong, “Developing EHR-driven heart failure risk prediction models using CPXR (Log) with the probabilistic loss function,” Journal of Biomedical Informatics, vol. 60, pp. 260-269, 2016.

[41] Hong J., Kim S. and Zhang B., “AptaCDSS-E: A classifier ensemble-based clinical decision support system for cardiovascular disease level prediction,” Expert Syst. Appl., vol. 34, 2008.

[42] Awang R. and Palaniappan S., “Intelligent heart disease predication system using data mining technique,” IJCSNS International Journal of Computer Science and Network Security, vol. 8, 2008.

[43] Patil S. and Kumaraswamy Y., “Intelligent and effective heart attack prediction system using data mining and artificial neural network,” European Journal of Science Research, vol. 31, 2009.

[44] Jyoti Soni, Uzma Ansari and Dipesh Sharma, “Intelligent and Effective Heart Disease Prediction System using Weighted Associative Classifiers,” International Journal on Computer Science and Engineering (IJCSE), vol. 3, pp. 2385-2392, 2011.

[45] Saqlain M., Wahid H., Nazar A.S. and Muazzam A.K, “Identification of Heart Failure by Using Unstructured Data of Cardiac Patients,” 45th International Conference on Parallel Processing Workshops, IEEE, 2016.

[46] Saqlain M., Athar A., Nazar A.S. and Muazzam A.K., “Developing a Classification Model for an Effective Treatment of Heart Failure,” International Journal of Computer Science and Information Security (IJCSIS), vol. 14, 2016.

[47] Sebastian Land, Simon Fischer, “Rapid Miner 5- Rapid Miner in academic use,” Rapid-I, www.rapid-i.com, 2012.

[48] http://information-gain.blogspot.com/2012/07/why-split-data-in-ratio-7030.html

[49] Ashwinikumar Kulkarni, B.S.C. Naveen Kumar, Vadlamani Ravi, Upadhyayula and Suryanarayana Murthym, “Colon cancer prediction with genetics profiles using evolutionary techniques,” Expert Systems with Applications, vol. 38, pp. 2752-2757, 2010.

[50] Arihito Endo, Taeko Shibata and Hiroshi Tanaka, “Comparison of Seven Algorithms to Predict Breast Cancer Survival,” Biomedical Soft Computing and Human Sciences, vol.13, pp. 11-16, 2008.

[51] Ruey-Shiang Guh, Tsung-Chieh Jackson Wu and Shao-Ping Weng, "Integrating Genetic Algorithm and Decision Tree learning for Assistance in Predicting in Vitro Fertilization outcomes," Expert Systems with Applications, Article in Press, Corrected Proof, 2010.

[52] Ruey-Shiang Guh, Tsung-Chieh Jackson Wu and Shao-Ping Weng, "A hybrid decision trees – adaptive neuro fuzzy inference system in prediction of anti HIV molecules," Expert Systems with Applications, Article in Press, Corrected Proof, 2010.

[53] http://www.simafore.com/blog/bid/62333/4-key-advantages-of-using-decision-trees-for-predictive-analytics

[54] Ammar K.A., Jacobsen S.J. and Mahoney D.W., "Prevalence and prognostic significance of heart failure stages: application of the American College of Cardiology/American Heart Association heart failure staging criteria in the community," vol. 115, pp. 1563-1570, 2007.

[55] Forman D.E., Cannon C.P., Hernandez A.F., Liang L., Yancy C. and Fonarow G.C., "Influence of age on the management of heart failure: findings from Get With the Guidelines-Heart Failure (GWTG-HF)," Am Heart J, vol. 157, 2009.

[56] Ambardekar A.V., Fonarow G.C., Hernandez A.F., Pan W., Yancy C.W. and Krantz M.J., "Characteristics and in-hospital outcomes for non-adherent patients with heart failure: findings from Get With The Guidelines Heart Failure (GWTG-HF)," Am Heart J, vol. 158, pp. 44-52, 2009.

[57] G.K. Savova, J.J. Masanz and P.V. Ogren, "Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications," J. Am. Med. Inform. Assoc., vol. 17. pp. 507-513 2010.

[58] J.H. Garvin, S.L. DuVall and B.R. South, "Automated extraction of ejection fraction for quality measurement using regular expressions in Unstructured Information Management Architecture (UIMA) for heart failure," J. Am. Med. Inform. Assoc., http://dx.doi.org/10.1136/amiajnl-2011-000535, 2012.

[59] F.H. Rutten, K.G.M. Moons, M.J.M. Cramer, D.E. Grobbee, N.P.A. Zuithoff, J.W.J. Lammers and A.W. Hoes, "Recognizing heart failure in elderly patients with stable chronic obstructive pulmonary disease in primary care: cross sectional diagnostic study," BMJ, vol. 331, 2005.