Innovative Data Mining Techniques for Predicting Cardiac Diseases: Current Trends and Future Directions

Dr. Ahmed El-Sayed[*1], Dr. Fatima Al-Hassan[2], Dr. Omar Abdulaziz[3] & Dr. Ranya Al-Sayed[4]
[1]Research Associate, Department of Computer Science, University of Sharjah, Sharjah, United Arab Emirates (UAE)
[2]Assistant Dean, College of Engineering, UAE University, Al Ain, United Arab Emirates (UAE)
[3]Professor, Department of Information Technology, Zayed University, Dubai, United Arab Emirates (UAE)
[4]Senior Lecturer, School of Computing, American University of Sharjah, Sharjah, United Arab Emirates (UAE)

ABSTRACT

Data Mining is an analytic process designed to find out data in search of harmonious patterns and methodical relationships between variables, and then to validate the extractive by applying the detected patterns to new subsets of data. The data mining is defined as the procedure of extracting information from enormous sets of data. In other words, we can say that data mining is mining knowledge from data. Afore, the scope of data mining has thoroughly been reviewed and surveyed by many researchers pertaining to the domain of healthcare industry which is an active interdisciplinary area of research. Actually, the task of knowledge extraction from the healthcare industry in medical data is a challenging effort and it is a very complex task. The present scenario in healthcare industry heart illness is a term that assigns to a huge number of health care circumstances related to heart. These medical situations relate to the unexpected health situation that straight control the cardiac. In healthcare industry data mining techniques like association rule mining, regression, classification, clustering is implemented to analyze the different kinds of cardiac based issue. Data mining techniques have the capabilities to explore hidden patterns or relationships among the objects in the medical data. In this paper we are using CHARM, an efficient algorithm for mining all frequent closed item set. The data classification is based on CHARM algorithms which result in accuracy, the data are estimated using entropy based cross validations and partition techniques and the results are compared. Subsequently, C5 algorithm is used as the training algorithm to show the rank of cardiac illness with the decision tree. The cardiac illness database is clustered using the K-means clustering algorithm, which will alienate the data appropriate to heart attackfrom the database.

Keywords: Data Mining, CHARM, Clustering, C5.0, K-Means, Medical Data, Classification

I. INTRODUCTION

Data mining is the process of exploringaactionable information from enormous sets of data. It is a uses mathematical analysis to derive patterns and trends that exist in the data [1]. It is a process of extracting earlier [2] unknown and process able information from enormous databases.The data mining is the process of using enormous data sets to gather important hidden knowledge [3]. It is disunite into seven steps like data integration, data selection, data cleaning, data transformation, data mining, pattern evaluation and knowledge presentation [4]. In present scenario, healthcare industry today produces enormousamounts of heteromorphic data about disease diagnosis, hospitals, resources,electronic patient records, etc.

Nowadays,the real world humans want to live a very luxurious life so they work like a machine in order to obtain lot of wealth. At a very young age, this type of way of living doesn't take a rest for themselves, which outcome in diabetics and blood pressure etc. It is a world known fact that heart is the most necessary part in the human body if that heart gets affected then it also affects the other parts of the human body [5]. The diagnosis is essential task and complicated that exigency to be executed accurately and proficiently. Thereupon, based on the doctor's experience& knowledge, the diagnosis are oftentimes made [6].Now,quality of service is a major challenge facing healthcare industry and its assurance diagnosing disease correctly & to provideadvantageous treatments to the sick person.In this paper, we are analyzing the cardiac illness prediction [5] [6] using different classification algorithms. At present, medical data are of the different types. It can be in the form of datasets, signals, images, wavelengths etc. The enormous amount of data is crucial to be [7] processed and scrutinized for knowledge extraction that empowers support for comprehensible the prevailing [8][9] situation in the healthcare industry. In the healthcare industry data mining techniques like association rule mining, Clustering,Classification Algorithms such as CHARM, Decision tree, C5.0 Algorithm are implemented to

analyze the different kinds of Cardiac illness problems. The C5.0 Algorithm and Clustering Algorithm like K-Means are the data mining techniques applied in the cardiac illness prediction.

II. KNOWLEDGE DISCOVERY PROCESS

The data mining is the salient part of the knowledge discovery process. In this, process may consist of the following steps the first step is chosen of data in which data is collected from different sources, the second step is pre-processing the chosen data, again third step is metamorphosis the data into an appropriate format [2] so that it can be processed further, the fourth step consist of data mining where appropriate data mining technique is applied on the transmute data for extracting valuable information and evaluation is the rearmost step. The term Knowledge Discovery in Databases or KDD for short, refers to the broad process [10] of discovery knowledge in data, and emphasizes the superior–level application of particular data mining methods. After the knowledge discovery two types of process first iterative and second interactive, that the process is iterative at each step, an interpretationthat moving back to foregoing steps may be needed.

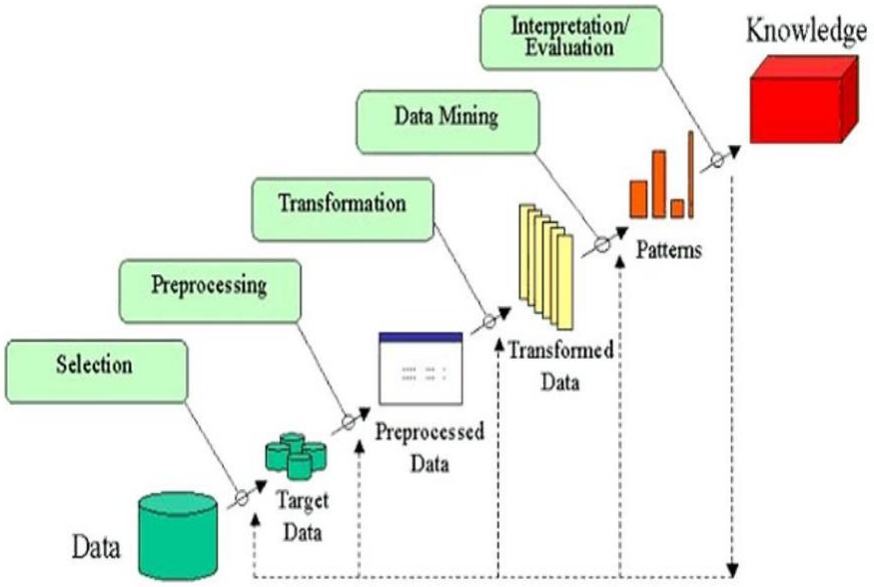


Figure 1. The Knowledge Discovery Process Steps

Knowledge Discovery in databases is the process of retrieving superior–level knowledge from small level data. In the preference step gather the miscellaneous data from varied sources for processing. In the real life medical data may be unaccomplished, noisy, inconsistent, complex, and inconsequent which needs a [4] selection process that pileup of data the essential data from which knowledge is to be extracted. The Pre-processing step performs fundamental operations of alienating the noisy data, try to discover the missing data or to develop a strategy for handling missing data, find out or remove outliers and extricate inconsistencies among the data. Afterwards, transformation step transforms the data into forms which is appropriate for [11] mining by performing tasks such as smoothing, normalization, generalization, aggregation, and discretization. The data deficiency task shrinks the data and appear for the same data in less volume, but produces the alike analytical outcomes. The unifying objective of the KDD process is to extract knowledge from data in the context of enormous data bases. It does this by utilizing data mining technique (algorithms) [2] to extract (identify) what is considered knowledge, pursuant to the specifications of measures and thresholds, using a database alongside any requisite preprocessing, sub sampling, and alteration of that database.

III. IMPORTANCE OF DATA MINING IN HEALTHCARE SECTOR

Currently, electronic health records are dynamically turning out to be more famous among healthcare establishments. To make better access to a significant amount of patient data, healthcare company is now in a position to maximize the performance and quality of their businesses with the assist of data mining.This technique's grip significant ability for the healthcare industry to enable health systems to respectively make use of data and analytics to cognize inefficiencies and best [4] practices that enhance care and bring down costs. The most basic definition of data mining is the analysis of enormous data sets to explore patterns and use those patterns to forecast or predict the probability of forthcoming events.In healthcare data,principally contains all the information relating to patients as well as the parties involved in healthcare industries. The storage of like type of data is increasing at a very swiftrate. Because of sustained increasing the size of electronic healthcare data a type of complexity is tobe living in it [12]. In other words, we can say that healthcare data

becomeintricate. By using the conventional methods it becomes very arduous in order to extract the significant information from it. Data mining is advantageous in such a circumstance where enormous collections of healthcare data are available. The data mining principally extracts the meaningful patterns which were earlier not known [6]. These patterns can then be unified into the knowledge and with the beneficence of this knowledge necessary decisions can become possible.

There are several techniques of data mining. Each and every medical informationassociated with patient as well as to healthcare organizations is subsidiary. It endows useful information in the scope of healthcare which may be then advantageous for management to take decisions namelydecision concerning health insurance policy, deciding on treatments, estimation of medical staff, disease prophecy etc. Data mining empowersyou to minimize costs [7] extensively by boosting efficiencies, [12] elongatedthe sick person's quality of life, and possibly even most outstandingly, assist in saving the lives of a lot more sick persons. The subsequent of thehealthcare sector is most likely based upon making use of data mining to decrease healthcare expenses, gauge credibility, cognize fraud insurance and healthcare claims, and in the end, risethe standard of sick persons service.

IV. DATA MINING TECHNIQUES

Data Mining is the process of extracting meaningful information and patterns from extensive data. Data Mining includes the collection, analysis, extraction, and statistics of data. It is also known as knowledge discovery process, knowledge mining from data or data and pattern analysis [2]. Data Mining is a logical process of discovering meaningful information to explore meaningful data. There are various data mining techniques have been developing and using in healthcare domain at the latest including classification, clustering, prediction, association, [4] sequential patterns, decision tree, visualization, as well as regression. Specifically the data mining techniques are a lot of conducive in predicting heart illness.

1. Association

The association is one of the renowned data mining techniques. The association rule learning areempower the finding of zestful relations (interdependencies) [2] between various variables in enormous databases. Again, association rule learning uncovers hidden patterns in the data that can be used to recognize [13] variables within the data and the co-occurrences of various variables that become visible with the maximum frequencies.

2. Classification

The classification is a transcendent data mining technique based on machine learning. In the data mining method, classification is the most normally used data mining technique which accommodate a set of pre classified pattern to create a model which can categorize the enormous [4] set of data. Classification analysis is anorderly process for acquiring vital

andpertinent information about data, and metadata data about data [14]. The classification analysis assiststo recognize the categories the data applicable. Classification analysis is nearly linked to cluster analysis as the classification can be used to cluster data.

3. Clustering

Clustering is one among the longstanding techniques used in data mining. Clustering analysis is the process of recognizingdata sets that are identical to each other, to perceive the dissimilarity as well as [2] symmetry within the data. The clustering technique illuminatesthe classes and puts objects in every class, till in the classification techniques, objects are assigned inpre determinedclasses [15]. Clustering is convenient to recognize various information because it correspond to with other examples so you can see where the equality and ranges agree. The clustering can work both ways and you can suppose that there is a cluster at a certain point and then use our identity criteria to see if you are right.

4. Prediction

Prediction in data mining is to recognize data points purely on the description of another respective data value. It is not inevitably respective to future events, but the used variables are unacquainted. The prediction is one of a data mining techniques that search the relationship between unconstrained variables [2] and the relationship between dependent and unconstrained variables [14]. It includesanalyzing trends, pattern matching, classification, and relation. By examiningpast phenomena or instances, you can make a prediction about future phenomena.

5. Sequential Patterns

The sequential patterns are one of data mining methodsthat try to search or recognize analogous patterns, steady events or trends in transaction data over a business term. More accurately, it consists of exploring enthralling subsequences in a set of sequences, where the interestingness [4] of a subsequence can be measured in terms of different criteria such as its phenomenon frequency, length, and benefit. Sequential pattern has countless real-life applications due to the reality that data is certainly encoded as sequences of symbols in many [16] region like as healthcare domain, e-learning, bioinformatics, texts, webpage click-stream analysis, and market basket analysis.

6. Decision Trees

A decision tree is a predictive model and the name itself suggestthat it looks like a tree. The decision tree is one of the most normally used data mining techniques because its model is simple to perceive for users [2]. In decision tree technique, the root of the decision tree is aneffortless question or condition that has multiple respond. Decision tree can be

believedas a partitioning of the original dataset where partitioning is done for a specific reason. Every data that come under a [17] portion has some equality in their information being predicted. Each answer, then direction to a set of questions or conditions that help us make up one's mind the data so that we can make the final conclusion based on it. This technique can be used for investigativeanalysis, data pre-processing and forecast work.

7. Visualization

Visualization is the most advantageous technique which is used to search for data patterns. This technique is used at the commencement of the data mining process. There are a lot of data mining technique which will produce advantageous patterns for nice data [18]. However,visualization is a technique which converts indigent data into nice data, letting different kinds of data mining technique to be used in exploringhidden patterns.

8. Regression

Regression endeavorsto define the dependency between variables. It suppose a [4] unilateral causal effect from one variable to the repercussion of another variable. Independent variables can be affected by each other, but it does not mean that this dependency is both ways as is the case with correlation analysis [19]. A regression explication can show that one variable is dependent on another but not conversely. This technique is used for time series modeling, forecasting, and discovery the causal effect relationship between the variables.

V. THE CHARM

The algorithm CHARM is produced by Mohammed j. Zaki and Chingjui Hsiao for mining all frequent closed itemset. The CHARM is aproficient algorithm for particularize the set of all frequent closed item-sets. CHARM is unprecedented in that it concurrently discoverboth the itemset space and transaction space, [20] dissimilar all foregoing association mining methods which only exploit the itemsetquest space. It also uses a very vital concept called diffsets [21] to detract the memory of intermediate computations.It means that it does not requisite enormous memory for storing the outcome of calculation expected for this algorithm. Moreover, CHARM avoids particularize all possible subsets of a closed itemset when particularize the closed frequent sets [22]. CHARM uses a highly proficient hybrid discovery method that bounce many levels of the IT-tree to swiftly cognize the frequent closed item-sets, alternately of having to numerate many possible subsets [23]. Now, we are present pseudo-code for the CHARM algorithm.

The algorithm beginsby initializing the prefix class [M], of nodes to be inquire into, the frequent single items and their tidsets. We suppose that the elements in [M] are instruction correspond to anappropriate total order f. The main computation is performed in the CHARM-Extend which regression the set of closed frequent itemsets B. CHARM-Extend is responsible for considering each combination of IT-pairs become visible in the prefix class [M]. For every IT-pair $Yi \times t(Yi)$ it amalgamateswith the IT-pairs $Yj \times t(Yj)$. Every Yi generates a new prefix class [Mi] which is initially blank. The two ITpairs are amalgamate to produce a new pair $Y \times Z$, where $Y = Yi \cup Yj$ and $Z = t(Yi) \cap t(Yj)$. After that experiment which of the four IT-pair properties can be applied by calling CHARM-property.Pay attention that this routine may alter the present class [M] by alienatingIT-pairs that are previously subsumed by the other pairs. It also pours out the newly generated IT-pairs in the new class [Mi]. Once upon all Yj have been processed, then recursively discover the new class [Mi] in a depth-first fashion. Then pour out the itemset Y, an extension of Yi, in the set of closed itemsets B, provided that X is not subsumed by anearlier found closed set. At this phase any closed itemset containing Yi has previously been generated as well as then sustain to process the next IT-pair in [M].

1. CHARM Pseudo-Code

CHARM (A, min supt):
[M] = {Yi × t(Yi) : Yi ∈ I ∧ σ(Yi) ≥ min supt}
CHARM-Extend ([M], B = ∅)
Return B //all unopened sets CHARM-Extend ([M], B):
For every Yi × t(Yi) in [M]
[Mi] = ∅ and Y = Yi
For every Yj × t(Yj) in [M], with Yj ≥f Yi
Y = Y ∪Yj and Z = t(Yi) ∩ t(Yj)
CHARM-Property([M], [Mi])
If ([Mi] = ∅) then CHARM-Extend ([Mi], B)
Alienate [Mi]
B = B ∪ Y //if Y is not subsumed CHARM-Property ([M], [Mi]):
If (σ(Y) ≥ minsupt) then
If t(Yi) = t(Yj) then //Property 1
Alienate Yj from [M]
Change all Yi with Y
Else if t(Yi) ⊂ t(Yj) then //Property 2
Change all Yi with Y
Else if t(Yi) ⊃ t(Yj) then //Property 3
Alienate Yj from [M]
Add Y × Z to [Mi] //use dictate f
Else if t(Yi) = t(Yj) then //Property 4
Add Y × Z to [Mi] //use dictate

VI. THE C5.0 CLASSIFICATION ALGORITHM

The decision trees are strong and famous tools for classification and prediction [24]. The C5.0 algorithm is a descendent of C4.5 machine learning algorithm. It is the classification algorithm which is appropriate for the enormous data set. It is superior than C4.5 on the efficiency, speed, and memory. The C5.0 is comfortably handled the multivaluedattribute and mislaid attribute from crop pest training data set. This algorithm is based on the decision trees. It is derived from anbefore system called ID3 where ID3 stands for induction of decision trees [25]. The decision trees are made out of list of feasible attributes and a set of training cases. These decision trees are used to categorize thereof sets of test cases. The moment used to createsthe rules using this algorithm are much lower and it create the rules which are even more actual. The C5.0 model works by disintegration the sample based on the field that provides the maximum information benefit [26]. The sample subset that is getting from the former disunite will be disunite afterward. The process will sustain so long as the sample subset cannot be disunite and is normally according to another field. Finally, inspect the lowest levelmsplit, those sample subsets that don't haveextraordinary role of the model will be refused.

Algorithm to Generate C5.0 Decision Tree

Input:

A. Data split,S, a set of training tuples and their related class labels.

B. Attribute_list, the set of patient attributes.

C. Attribute_choosing _method, a procedure to determine the fragmentation criterion dividing up the data tuples intopersonal classes. This standard consists of a dividing up _attribute and ,either a divided -point or divided subset.

Ouput:

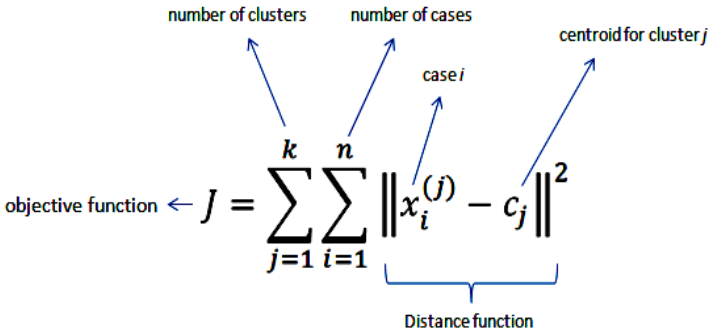
1. Create a node X
2. If tuples in S are all of the identical class, C, then
3. Return X as a leaf node labelled with the class C
4. If attribute_list is unfilled, then
5. Return X as a leaf node labelled with the most class in S
6. Enforce attribute_choosing_method(S, attribute_list) to discoverthe optimal divided _criterion
7. Label node X with divided _criterion

8. If divided_attribute is discrete-valued and multiwaydivided permit then
9. Attribute_listattribute_list- divided _attribute
10. For every outcome m of divided_criterionLet S_m be the set of data tuples in Stolerable outcome mif S_n is unfilled then attach a leaf labelled with most class in S to node Xelse, attach the node returned by generating C5.0 decision tree(S_m, attribute_list) to node X
11. Return X

The C5.0 algorithm gives the acknowledge on noise and mislaid data. If the difficulty of over fitting and error pruning is the solution to the C5.0 algorithm [27]. Scalability is enhanced by multi-threading, it signifiesC5.0 can take benefit of computers with multiple CPUs and cores. The C5.0 support all types of data like categorical, dates, continuous, times and timestamps. It can deal with missing values of crop pest data [28]. It specially supported enhance (It is a process of generation of various decision trees and they all are amalgamate to ameliorate the predictions) to refine the classifier precision.

VII. THE K-MEANS CLUSTERING ALGORITHM

The K-means is one of the straightforward unsupervised learning algorithms that solution to the well-known clustering issue. K-means clustering is a type [29] of unsupervised learning, which is used when you have unlabeled data. K-means clustering is a technique used for clustering analysis, in particular data mining and statistics [30]. It follows a straightforward procedure of classifying a given data set into a number of clusters, defined by the letter "k," which is fixed previously.



The clusters are then positioned at points and all observations or data points are related to the within reach cluster, computed, adjusted and then the process starts over using the new adaptation until a desired outcome is reached [31]. K-Means clustering purpose to split n objects into k clusters in which every object be suited to the cluster with the proximate mean. This method produces precisely k dissimilar clusters of greatest possible dissimilarity. The optimal number of clusters k leading to the greatest dissociation is not familiar as a priori and must be computed from the data. The purpose of K-Means clustering is to abbreviate total intra-cluster varianceand squared error function.

The prognostication of Cardiac Illness using k-Means clustering

In the K-means clustering the data into k category where k is predefined. Afterward, choose k points at random as cluster centers and allocate objects to their near cluster center in pursuance of the Euclidean distance function. Then calculate the centroid or mean of all objects in every cluster. So long as encore steps 2, 3 and 4 while the same points are assigned to each cluster in successive rounds. The transformation dose not make a material dissimilarity in the definition of the clusters [32]. The clustering is accomplishedon preprocessed data set using the K-means algorithm with the K values so as to extract pertinent data to the cardiac illness. K-Means is comparatively an efficient method and unpretentious algorithm that has been suitable for many healthcare problem domains. The K-means rapid, substantial and effortless to understand and its gives best outcome when the data set are isolated or well separated from each other [33].

VIII. THE CARDIAC ILLNESS

The heart is a crucial part of our body as well as our life is totally incumbent on the efficient working of the heart. The term cardiac illness is often used interchangeably with the term cardiovascular disease. The cardiovascular disease normally refers to conditions that involve narrowed or obstructed blood vessels that can lead to chest pain (angina), a heart attack, and stroke [34]. Othercardiac conditions, like as those that affect your heart's muscle, valves or rhythm, also are recognizance forms of heart disease. A heart attack occurs when the blood flow to a part of the heart is obstructed by a blood clot. If this clot cuts off the blood flow perfectly, the part of the heart muscle supplied by that artery begins to die [35]. Additional cardiovascular diseases include

angina (chest pain), stroke, high blood pressure, and rheumatic heart disease. Risk factors are conditions or habits that make a person more likely to emerge a cardiac illness. Thereupon, certain risk factors, such as age and family history of early cardiac illness, can't be altered. There are a number of factors which enlargement the risk of cardiac illness.

- Smoking
- High blood cholesterol
- Family history of heart disease
- Diabetes and prediabetes
- Being overweight or obese
- Poor diet
- High Blood Pressure
- Physical inactivity
- Having a history of preeclampsia during pregnancy
- Hypertension
- Age (55 or older for women)
- Sedentary lifestyle

Although, if you've had the condition, you should take extra care to try and control other cardiac illness risk factors.

IX. THE PROPOSED SYSTEM ARCHITECTURE

In this portion, we are discussing proposed system architecture. In recent times, contemporaneous medicine generates an enormous amount of information stored in the medical database. It is essential to extract utilitarian knowledge and providing scientific decision-making for the diagnosis and treatment of disease from the database increasingly becomes essential. Data mining in medicine can deal with this cardiac illness problem. It can also make better the management, quality of hospital information and promote the development of telemedicine and community medicine. In view of the fact that the medical information is multi-attribution, incompletion, specialty of redundancy, and very closely linked with time, medical data mining dissimilar from another one. The medical data mining involving pretreatment of medical data, fusion of various pattern and resource, fast and strong mining algorithms and credibility of mining outcome. This method and applications of cardiac illness, medical data mining based on computation intelligence, such as CHARM algorithm, C5.0 classification algorithm, and K-means clustering have been introduced.

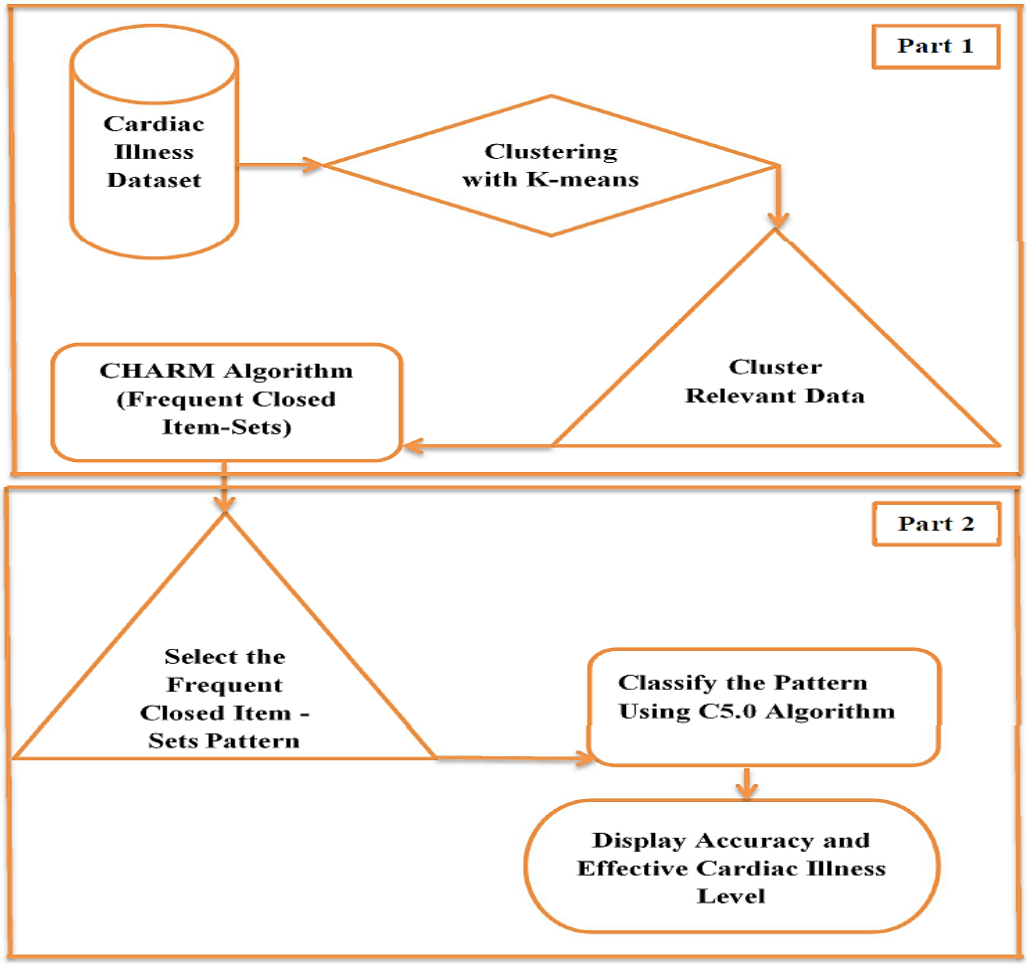


Figure 2. The Proposed System Architecture Steps

The algorithm takes the cardiac illness dataset and classify whether a person is having a cardiac illness or not. The above algorithm is divided into 2 parts shown in figure 2. The part 1 contains cardiac illness, medical dataset, performs clustering relevant data and CHARM algorithm for frequent closed item-sets. In part 2, select frequent closed item- data sets and classify the pattern using C5.0 algorithm is used to display accuracy and effective heart attack level.

X. THE EXPERIMENTAL OUTCOME

In this section, the experimental outcome in identifying necessary patterns for predicting the cardiac illness. The cardiac illness database is preprocessed successfully by deleting corresponding records and providing missing values as shown in table 1. The decorous cardiac illness data set, resulting from preprocessing, is then collected by K-means algorithm with the K value of 2. The collection contains the data associated with the cardiac illness as shown in table 2 and the further contains the left over binary data. Then the regular forms are mined efficiently from the collection appropriate cardiac illness, using the CHARM algorithm.

The model consortiums of cardiac illness parameters for general and risk level related to their values and levels are listed below in that, ID lesser than of (#1) of weight contains the normal level of prediction cardiac illness and higher ID other than #1 comprise the higher risk levels cardiac illness and mention the prescription IDs. Table 3 displays the parameters of the cardiac illness prediction with equivalent prescription binary key ID and their levels.

Ttable 1. The Cardiac Illness Dataset

Binary Key ID	Key Input Attributes Description
000001	Sick Person Id – Sick Person Identification Number
000010	Gender (Value 1: Male; Value 0: Female)
000011	Length of Life
000100	Chest Pain Type (Value 1:Typical Type 1 Angina, Value 2: Typical Type Angina, Value 3:Non-Angina Pain; Value 4: Asymptomatic)
000101	Fasting Blood Sugar (Value 1: >120 mg/dl; Value 0: <120 mg/dl)
000110	Restecg – Resting Electrographic Outcome (Value 0:Normal; Value 1: Having ST-T Wave Abnormality; Value 2: Showing Probable or Definite Left Ventricular Hypertrophy)
000111	Serum Cholesterol (mg/dl)
001000	Maximum Heart Rate Achieved : Value (0.0) 0.0 and <=80, Value (1.0) : >81 and <119
001001	Oldpeak – ST Depression Induced by Exercise
001010	Exang - Exercise Induced Angina (Value 1: Yes; Value 0: No)
001011	CA – Number of Major Vessels Colored by Fluoroscopy (Value 0-3)
001100	Slope – The Slope of the Peak Exercise ST Segment (Value 1:Unsloping; Value 2: Flat; Value 3: Downsloping)
001101	Thal (Value 3: Normal; Value 6: Fixed Defect; Value 7: Reversible Defect)
001110	Trest Blood Pressure (mm Hg on Admission to the Hospital)

The C5.0 Algorithm Decision Tree Structure
If Length of Life = < 35 and Overweight=Not Agree and Liquor=Not Once

Then
Cardiac Illness Level is Minimal
(Or)
If Length of Life = >35 and Blood pressure=High and
Smoking=Present Timet
Then Cardiac Illness Level is Maximal

Table 4 shows the example of training data to foresee the cardiac illness level and then figure 3 shows the efficient cardiac illness level with a tree using the C5.0 by information obtain. The experimental outcome of our approach as presented in table 4. The target is to have high accuracy, as well as high exactness and recall metrics. These can be easily converted to accurate- definite (AD) and inaccurate- definite (ID) metrics. The accurate - definite (AD) is the total percentage of members classified as class X relates to class X and inaccurate - definite (ID) is the total percentage of members of class X but does not relate to class X.

Table2. The Cluster Relevant Data Based Upon Cardiac Illness Dataset

Binary Key ID	Binary Reference ID	Attributes Description
000001	000001	Gender
000010	000010	Length of Life
000011	001001	Painloc: Chest Pain Location
000100	010010	CP: Chest Pain Type
000101	010000	Relrest
000110	011001	Chol: Serum Cholesterol in mg/dl
000111	010110	Trestbps: Resting Blood Pressure
001000	011100	Smoke
001001	011111	Cigarettes Per Day
001010	011110	Years (Number of Years as a Smoker)
001011	100101	dm (1 = History of Diabetes; 0 = No Such History)
001100	100010	fbs: (Fasting Blood Sugar > 120 mg/dl)
001101	101000	Famhist: Family History of Coronary Artery Disease
001110	101101	Thalach: Maximum Cardiac Rate Achieved
001111	110010	Sedentary Lifestyle/Inactivity
010000	101100	Exang: Exercise Induced Angina
010001	110100	Ca: Number of Major Vessels (0-3) Colored by Fluoroscopy
010010	110111	Num: Diagnosis of Cardiac Illness

$$\text{Exactness} = \text{AD} \,/\, \text{AD} + \text{ID}$$

$$\text{Recall} = \text{AD} \,/\, \text{AD} + \text{II}$$

Again, accurate - indefinite (AI) is the total percentage of members which do not relate to class X are classified not a part of class X . It can also be given as(100%-ID) and inaccurate - indefinite (II) is the total percentage of members of class X incorrectly classified as not related to class X.

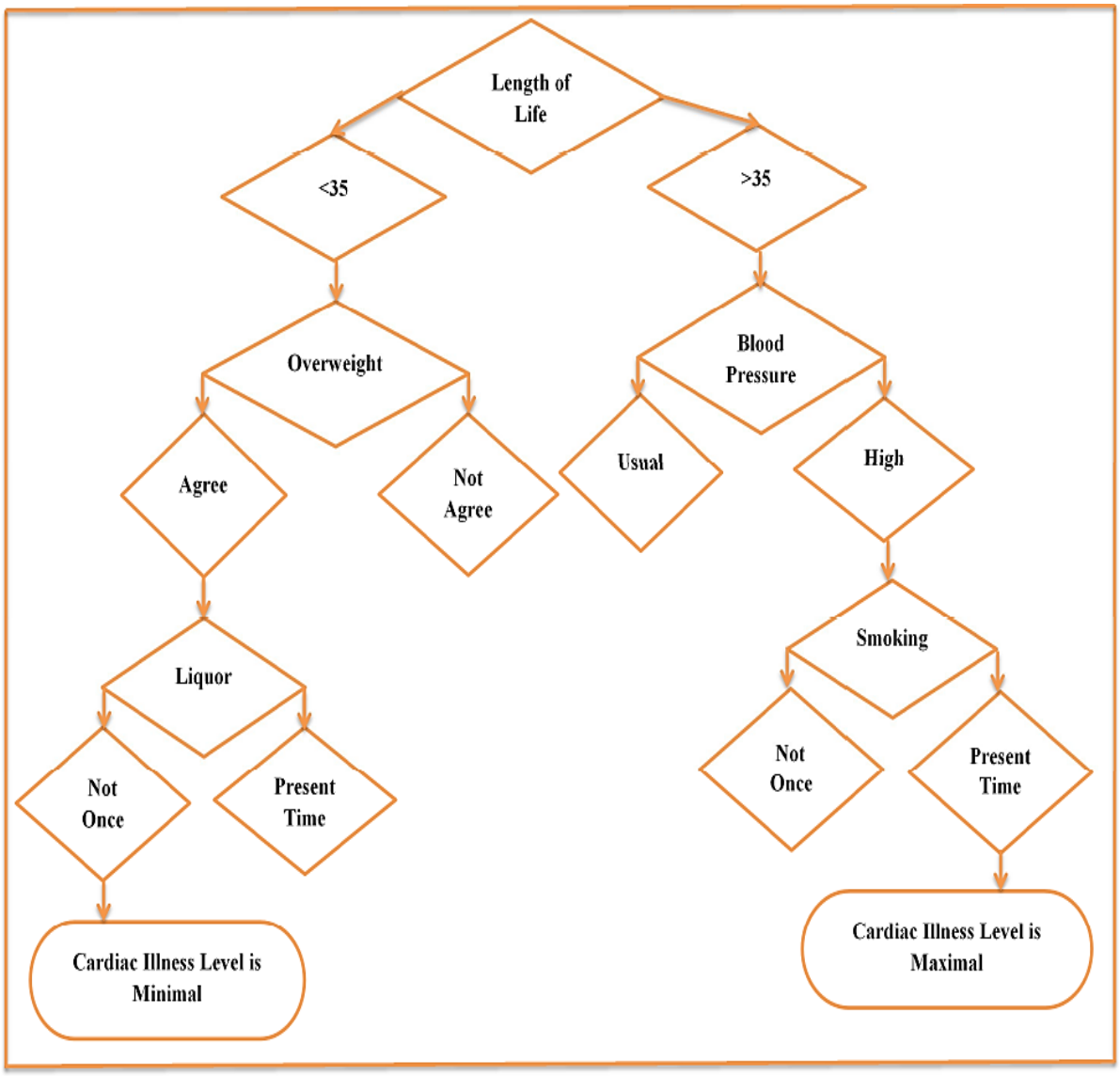


Figure 3. The Decision Tree for the Concept Cardiac Illness Level by Information Obtain Through C5.0 Algorithm

Table 3. The Cardiac Illness Parameters withCorresponding Prescripion Binary Key ID and Circumstance

Weights	Parameter Description	Cardiac Illness Binary Risk Level
Age<35 Age> 35	Gender (Male and Female)	000001 001000
Overweight	Agree Not Agree	001001 000001
Liquor	Not Once Present Time Past	000001 000111 000011
Smoking	Not Once Present Time Past	000001 000110 000011
Exalted Saturated Meal	Agree Not Agree	001001 000001
Working Out	Continually Not Ever	000001 000110
Exalted Salt Diet	Agree Not Agree	001000 000001
Sedentary Life	Agree Not Agree	000111 000001
Noxious Cholesterol	High Usual	001000 000001
Blood Sugar	High (>120&<400) Usual (>90&<120) Low (<90)	000101 000001 000100
Heart Rate	Low (< 60bpm) Usual (60 to 100) High (>100bpm)	001001 000001 001010
Blood Pressure	Usual (130/89) Low (< 119/79) High (>200/160)	000001 001000 001010

Table 4. The DifferentiationBetween CHARM and K-Means Based CHARM With C5.0

Using Technique	Exactness	Recall	Accuracy (%)
K-Means Based CHARM	0.80	0.72	78%
K-Means Based CHARM with ID3 and C4.5 Algorithm	0.83	0.93	93%
K-Means Based CHARM with C5.0 Classification Algorithm	0.86	0.96	96%

XI. CONCLUSION

The data mining is the process of discovery anomalies, patterns and relation within huge data sets to predict outcomes. The novel data mining techniques to sculpt precious information has been considered as an activist viewpoint to ameliorate the quality and accuracy of health care industry while reduces the health care cost and diagnosis time. In medical sciences prediction of cardiac illness is the most arduous task. Our real world, main causes of death are due to cardiac illness. The deaths due to cardiac illness in many countries occur due to sedentary life, mental stress, smoking, work overload, and many other issues and it is found as the main cause in adults is due to cardiac illness. In this paper, we are proposing a cardiac illness prediction system using CHARM, C5.0 classification algorithm and k-means clustering. The CHARM is an efficient algorithm for enumerating the set of all frequent closed items-sets. In this research work using CHARM, ID3, C4.5 and C5.0 compare with each other. Among all these classifiers C5.0 gives more accurate and efficient outcome. C5.0 is a classifier which classifies the data in less time compare to other classifier. For originate decision tree the memory usage is least and it also makes better the accuracy because error rate is low so accuracy in outcome set is towering. The accuracy of K-means based CHARM, K-Mean based CHARM with ID3 and C4.5 algorithm and K-Mean based on CHARM with C5.0 classification algorithm 78%, 93% and 96% respectively. The main objective of our paper using this technique presence of cardiac illness can be predicted accurately.

XII. ACKNOWLEDGMENTS

This research paper is made possible through the assistance and support from almost everyone. The author would like to thank the reviewers anonymous for their constructive comments. First and foremost, we would like to thank GOD for his unconditional guidance and wisdom as wedo my research. Second, we would like to thank my colleagues for his more support and encouragement for giving us this research. Finally, we are sincerely thank to my parents, family, who provide the advice and support.

XIII. REFERENCES

[1] Xindong Wu, Xingquan Zhu, Gong-Qing Wu, Wei Ding, "Data Mining with Big Data", IEEE transactions on Knowledge and Data Engineering, vol. 26, no. 1, January 2014.
[2] Jiawei Han, MichelineKamber, Jian Pei., “Data mining concepts and techniques “, 3rd ed, ISBN 978-0-12-381479-1, Morgan Kaufmann Publishers is an imprint of Elsevier. 225Wyman Street,Waltham, MA 02451, US
[3] MarcoViceconti, Peter Hunter, Rod Hose, "Big Data big knowledge: big data for personalised health care", IEEE Journal of Biomedical and Health Informatics, no. 99, February 2015.
[4] Tan, P., Steinbach, M. and Kumar, V. Introduction to Data Mining, Addison-Wesley, Boston, 2006.
[5] Chao-ton Su, Chieng-Hsing Yang et al., "Data mining for diagnosis of type III diabetes from 3d body surface anthropometrical scanning data" in Science Direct, Elsevier, 2006.
[6] Carlos Ordonez, Edward Omincenski and Levien de Braal "Mining Constraint Association Rules to Predict Heart Disease", Proceeding of 2001, IEEE International Conference of Data Mining, IEEE Computer Society, ISBN-0-7695-1119-8, 2001, pp: 433-440
[7] Boris Milovic*, Milan Milovic**, "Prediction and Decision Making in Health Care using Data Mining", International Journal of Public Health Science (IJPHS), vol. 1, no. 2, pp. 69-78, December 2012.

[8] Sufi Fahim, Ibrahim Khalil, "Diagnosis of cardiovascular abnormalities from compressed ECG: a data mining-based approach", Information Technology in Biomedicine IEEE Transactions on, vol. 15, no. 1, pp. 33-39, 2011.
[9] Fiscon Giulia, Emanuel Weitschek, Giovanni Felici, Paola Bertolazzi, Simona De Salvo, PlacidoBramanti, Maria Cristina De Cola, "Alzheimer's disease patients classification through EEG signals processing", Computational Intelligence and Data Mining (CIDM) 2014 IEEE Symposium on, pp. 105-112, 2014.
[10] Horeis, T.; Sick, B. Collaborative Knowledge Discovery & Data Mining: From Knowledge to Experience. Proceedings of IEEE Symposium on Computational Intelligence and Data Mining. pp. 421-428.
[11] U. Fayyad, G. Piatesky-Shapiro, P. Smyth, Advances in Knowledge Discovery and Data Mining, California:AAAI Press, 1996.
[12] M.-H. Kuo, T. Sahama, A. W. Kushniruk, E. M. Borycki, D. K. Grunwell, "Health big data analytics: current perspectives challenges and potential solutions", Int. J. Big Data Intell., vol. 1, no. 112, pp. 114-126, 2014.
[13] B.Ramasubbareddy,A.Govardhan,A.Ramamohanreddy,"Mining Positive and Negative Association Rules" in IEEE ICSE 2010, Hefeai, Chaina:, August 2010.
[14] Cheng Lin, Fan Yan," The Study on Classification and Prediction for Data Mining", Published in Measuring Technology and Mechatronics Automation (ICMTMA), Seventh International Conference on, Nanchang, China , 13-14 June 2015., DOI: 10.1109/ICMTMA.2015.318
[15] S. Guha, N. Mishra, R. Motwani, L.O' Callaghan, "Clustering data stream", In Proc.Of FOCS, pp. 359-366, 2000.
[16] H. Cheng, X. Yan, and J. Han.Incspan: Incremental mining of sequential patterns in large database. Proceedings of 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 527 - 532, 2004.
[17] B Mobasher, R Cooley, S Jaideep et al., Comments on decision tree, New York:IEEE Press, 1999.
[18] Tubao Ho, Trongdung Nguyen, "Visualization Support For User-Centered Model Selection In Knowledge Discovery and DataMining", International Journal On Artificial Intelligence Tools, vol. 110, pp. 691-713, 2001.
[19] Lazarevic A. Xu X, T. Fietz, Z. Obradovic, "Clustering Regression Ordering Steps for Knowledge Discovery in Spatial Databases", International Joint Conference on Neural Networks (IJCNN'99), July 10-16 (1999).
[20] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. InkeriVerkamo.Fast discovery of association rules. In U. Fayyad and et al, editors, Advances in Knowledge Discovery and Data Mining, pages 307–328. AAAI Press, Menlo Park, CA, 1996.
[21] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal.Discovering frequent closed itemsets for association rules.In 7th Intl. Conf. on Database Theory, January 1999.
[22] J. Zaki Mohammed, Ching-Jui Hsiao, "Efficient Algorithms for Mining Closed Itemsets and Their Lattice Structure", IEEE transactions on knowledge and data engineering, vol. 17, no. 4, APRIL 2005.
[23] M. Zaki, and C. Hsiao, CHARM: An efficient algorithm for closed itemset mining. In SDM'02, Arlington, VA, pp457-473, April 2002.
[24] Venkatadri.M, Lokanatha C. Reddy A Comparative Study On Decision Tree classification Algorithms In Data Mining International Journal Of Computer Applications Engineering, Technology And Sciences (Ij-Ca-Ets) Issn: 0974-3596 April '10-Sept '2010, Volume 2: Issue 2, Page: 24
[25] W. Peng et al., "An Implementation of ID3-Decision Tree Learning Algorithm" in University of New South Wales, Sydney, Australia:School of Computer Science & Engineering, 1990
[26] T. Bujlow, TahirRiaz, Jens MyrupPedersen,"Classification of HTTP traffic based on C5.0 Machine Learning Algorithm",Section for Networking and Security,Department of Electronic Systems Aalborg University
[27] Su-lin Pang, Ji-zhang Gong, "C5.0 Classification Algorithm and Application on Individual Credit Evaluation of Banks" in , Elsevier, 2009.
[28] M. Wang, K. Gao, L. Wang, X. Miu, "A Novel Hyperspectral Classification Method Based on C5.0 Decision Tree of Multiple Combined Classifiers", Fourth International Conference on Computational and Information Sciences, 2012.
[29] J. B. MacQueen,"Some Methods for classification and Analysis of Multivariate Observations", Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", Berkeley, University of California Press, vol 1, pp 281-297, 1967.
[30] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman, A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation", IEEE Trans. Pattern Anal. Mach. Intell., vol. 24, no. 7, pp. 881-892, Jul. 2002.
[31] Z Huang, "Extensions to the k-means algorithm for clustering large data sets with categorical values", Data Mining and Knowledge Discovery, vol. 2, pp. 283-304, 1998.

[32] K.A. Abdul Nazeer, M.P. Sebastian, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm", Proceeding of the World Congress on Engineering, vol. 1, 2009
[33] J Dong, M. Qi. K-means Optimization Algorithm for Solving Clustering Problem. Knowledge Discovery and Data Mining, pp52-55, 2009.
[34] Tu, M.C., D. Shin, and D. Shin, Effective Diagnosis of Heart Disease through Bagging Approach. Biomedical Engineering and Informatics, IEEE, 2009.
[35] Das, R., I. Turkoglu, and A. Sengur, Effective diagnosis of heart disease through neural networks ensembles. Expert Systems with Applications, Elsevier, 2009.36 (2009): p.76-75.