# An Intelligent Cardiac Ailment Prediction Using Efficient ROCK Algorithm, K-Means and C4.5 Algorithm

Yusuf Perwej, Firoj Parwej, and Nikhat Akhtar

***Abstract*—The data mining techniques have the ability to discover hidden patterns or correlation among the objects in the medical data. There are many areas that adapt data mining techniques, namely marketing, stock, health care sector and so on. In the health care industry produces gigantic quantities of data that clutches complex information relating to the sick person and their medical conditions. The data mining has an infinite potential to make use of healthcare data more effectually and efficiently to predict various kinds of disease. The present-time healthcare industry heart ailment is a term that assigns to an enormous number of health care circumstances related to heart. These medical circumstances relate to the unexpected health circumstance that straight control the cardiac. In this paper we are using a ROCK algorithm because it uses Jaccard coefficient on the contrary using the distance measures to find the similarity between the data or documents to classify the clusters and the contrivance for classifying the clusters based on the similarity measure shall be used over a given set of data. Afterward, C4.5 algorithm is used as the training algorithm to show the rank of a cardiac ailment with the decision tree. The C4.5 can be referred as the statistic classifier as well as this algorithm uses avail radio for feature selection and to build the decision tree. The C4.5 algorithm is widely used because of its expeditious classification and high exactitude. Lastly, the cardiac ailment database is clustered using the K-means clustering, which will alienate the data convenient to cardiac sickness from the database.**

***Index Terms*—Clustering; C4.5 Algorithm; Data Mining; K-Means; ROCK Algorithm.**

## I. Introduction

Today scenario, the fast development of networking, data storage, and the data collection capacity, big data are now process of classification via enormous data sets to recognize trends and patterns and set up relationships [1]. The data expeditiously expanding in all science [1] and engineering domains, including physical, healthcare sector, biological and biomedical sciences [2]. The Data mining is the automated mining and knowledge discovery in databases are repeatedly treated as [3] alternative, data mining is in fact part of the knowledge discovery process [4]. Data mining is the process of pattern finding and extraction where an enormous amount of data is involved. Again, both the healthcare industry and data mining have emerged some of credible early detection systems and other different health care [5] respective systems for the diagnosis and clinical data. The healthcare sector can be regarded as a place with wealthy data as they originate enormous amounts of data, including electronic medical records, administrative reports and other standard discovery [6]. In this paper, we are analyzing the cardiac ailment prediction using different classification algorithms [7]. In the healthcare sector data mining techniques like clustering, classification algorithms such as ROCK, decision tree, C4.5 algorithms are executed to analyze the different kinds of Cardiac ailment issue. The C4.5Algorithm and Clustering Algorithm like K-Means are the data mining techniques applied in the cardiac ailment prediction [6].

## II. The Circumstance that Rise Hazard for Cardiac Ailment

There are several risk factors associated with the coronary Cardiac ailment and stroke. The various health conditions, your lifestyle, and your age and family history can rise your hazard for Cardiac ailment. The situation or habits that make a person more likely to develop a [8] ailment are hazard factors. They can also encourage the probability of a current ailment will get worse. Further the few risk factors, such as family history, cannot alter, while other risk factors, like high blood pressure, can be altered with medication.

### A. The Uncontrollable Risk Factors

This section explains overview of the several uncontrollable risk factors associated with the coronary Cardiac ailment and stroke.

a) *Age:* The Cardiac belonging disease usually occurs in men above the age of and in 40women after menopause, and most likely people who die of heart attacks are above the age of 60.
b) *Race:* Ethnicity can have an influence on your hazard level as well. Genetic distinction and environmental factors play a role in your risk for developing Cardiac disease. American Indians, African Americans, and Mexican Americans are more likely to have a Cardiac ailment than Caucasians.
c) *Gender:* Men have got more risk of Cardiac attack than compare to women, and men generally suffer from Cardiac attacks at the earlier age of life
d) *Family History:* For the human being who is having a close relative who had Cardiac attack may be at risk of Cardiac ailment.

### B. The Controllable Risk Factors

This section explains overview of the several controllable risk factors associated with the coronary Cardiac ailment

Published on December 29, 2018.

Y. Perwej is with the Department of Information Technology, Al Baha University, Al Baha, Kingdom of Saudi Arabia (KSA) (e-mail: yusufperwej@gmail.com).

F. Parwej is with the Department of Computer Science & Engineering, Singhania University, Pacheri Bari, Jhunjhunu, Rajasthan, India (e-mail: dr.firojparwej@hotmail.com).

N. Akhtar is with the Department of Computer Science & Engineering, Babu Banarasi Das University, Lucknow, India (e-mail: dr.nikhatakhtar@gmail.com).

DOI: http://dx.doi.org/10.24018/ejers.2018.3.12.989

and stroke.

a) *Tobacco* Use: The chemicals in tobacco smoke increase the development of blood clots and promote the cause Cardiac attacks by building-up of plaque in artery walls.
b) *Stress:* Stress is detrimental to your health in several ways. Firstly, enhanced periods of stress cause your body to release the stress hormones cortisol and adrenaline.
c) *Weight:* Supposing body pound rise, the risk of Cardiac disease also rises. This is particularly factual for people who carry extra body fat around the waist. To detract the risk of Cardiac illness promiscuous dietary factors that can be used.
d) *Bad Fats:* A lot of food we eat every day contains saturated and trans fats. They lift up the blood cholesterol level, which may potentially clog up the arteries, putting one at greater risk of Cardiac ailment.
e) *High Cholesterol:* The exorbitant cholesterol in the blood building up in the walls of the arteries can cause a process called atherosclerosis, a form of Cardiac illness.
f) *Diabetes:* Diabetes can cause Cardiac disease by increasing the risk of high blood pressure and high cholesterol in the blood. It stimulates injury to the artery walls and formation of blood clots.
g) *Blood Pressure:* Blood pressure is the force of the blood opposed to the inner walls of the blood vessels, generated when the heart pumps blood. When a person has hypertension, the arteries are under enhanced pressure and the Cardiac has to pump harder, which may lead to injury of the artery walls, atherosclerosis, and coronary Cardiac illness.

## III. Related Works

In this section elaborates the respective work done in the medical domain. During the literature study, we are exploring that the amount of research done in this area is quite enormous. There are various approaches and various techniques used in the Cardiac disease prediction. The various data mining techniques have been used by them for diagnosis & achieved various probabilities for various techniques. Jyoti Sonia, et.al. [9] in year 2011 produced three classifiers Decision Tree, Naïve Bayes and Classification via clustering to diagnose the presence of Cardiac disease in patients. Classification via clustering: Clustering is the process of grouping similar elements. This method may be used as a preprocessing step prior to feeding the data to the classifying model. To increase the prediction of classifiers, genetic quest was inclusive. Shadab et al. [10] proposed a Naive Bayes method in the year 2012 using 15 attributes in the dataset for the Cardiac diagnosis in the Cardiac prediction system. The system extracts hidden knowledge from a historical Cardiac disease database. This is the most emphatic model to predict patients with Cardiac disease. This model could answer intricate queries, each with its own strength with respect to ease of model interpretation, access to extensive knowledge and precision.

Ishtake et al. [11] presented a prediction system for Cardiac diagnosis using decision trees, Neural Network and Naive Bayes method using 15 attributes in the year 2013. A prototype Cardiac disease prediction system is developed using three data mining classification modeling method. The system draws out hidden information from a historical Cardiac disease database. All three models could answer complicated queries, each with its own strength with respect to ease of model interpretation, access to extensive knowledge and precision. Nishara Banu et al. [12] offered to C4.5 algorithm, MAFIA and K-means clustering. The frequent patterns can be classified using C4.5 algorithm as the training algorithm using the concept of knowledge entropy. The outcome showed that the designed prediction system is competent of predicting the Cardiac attack successfully and in the year 2014.

She is using 13 attributes in the dataset achieving 89 percent precision. Hlaudi Daniel Masethe et al. [13] produced a model for prediction of Cardiac disease using J48, Bayes Net, and Naïve Bayes, Straightforward CART and REPTREE algorithms using patient data set from a doctor. Assessment of the confusion matrix showed that J48, REPTREE and straightforward CART show a prediction model of 89 cases with a hazard factor positive for Cardiac illness. J.Vijayashree et al. [14] presented a widespread about Cardiac related cardiovascular disease and to brief about current decision support systems for the prediction and diagnosis of Cardiac illness supported by data mining and hybrid intelligent method. Arabasadi et al. [15] offered a highly actual hybrid method for the diagnosis of coronary artery disease which had increased the performance of the neural network by approximately 10% by increase its initial weights using genetic algorithm. The authors had declared that they have achieved accuracy, sensitivity and specificity rates of 93.85%, 97% and 92% respectively, on Z-Alizadeh Sani dataset. Yusuf Perwej, Mohammed Y. Alzahrani, et al. [6] proposed a C5.0 algorithm, CHARM algorithms and K-means clustering in the year 2018. 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. 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.

## IV. Restriction of Traditional Clustering with Categorical Data

We investigated and experiments demonstrated that the distance measures cannot lead to high-quality clusters when clustering categorical data. In addition, majority of clustering algorithms merge most similarity points in a single cluster at every step and this localized method is prone to errors [16]. The response to the stated issue is ROCK, which takes a more global method for clustering that is, if two similar points having similarity vicinity, then only the two points can be combined to the same cluster [17]. The Robust Clustering Using Links (ROCK) hierarchical clustering algorithm along with the Jaccard coefficient is being used to decide the group of various subject areas and to receive the similarity among the

DOI: http://dx.doi.org/10.24018/ejers.2018.3.12.989

different data. The primary consideration behind this process is to cluster the similarity data with each other. The ROCK algorithm is best favorable for clustering [18] categorical data because it does not use distance measures as an alternative it uses coefficient to discover the similarity between the two data.

## V. The Rock Technicalities

In this section, we are talking about few technicalities used in the ROCK clustering algorithm that is based on the notion of links and neighbors.

### A. *Categorical Data*

The categorical variables represent the types of data which may be echeloned into groups. The categorical attributes are also referred to as titular attributes, which are merely used as names, such as the brands of motor vehicle and names of bank branches. At the instance of categorical variables are field of race, gender, age classification, study, educational level and nationality [16]. The latter two variables may also contemplate in a numerical manner by using appropriate values for age and highest grade completed, it is frequently more auxiliary to categorize such variables into a relatively small number of groups and this data typically is of fixed dimension.

### B. *Goodness Measure*

The performing clustering the purpose of using good measure is to maximize the criterion function and to recognize the finest pair of clusters to be merged at every stage of ROCK. In goodness measure order to take the plunge whether or not to merge clusters MC*x* and MC*y*, the goodness measure, its mean the proportionality measure between two clusters Goodness *(*MC*x,* MC*y)* for merging clusters MC*x* and MC*y* (1) is defined as

$$\text{Goodness}(MCx, MCy) = \frac{link(MCx, MCy)}{(Sx+Sy)^{1+2f(\theta)} - Sx^{1+2f(\theta)} - Sy^{1+2f(\theta)}} \tag{1}$$

where link [MC*x,* MC*y*] is the number of cross-links between clusters MC*x* and MC*y*, (2) at the instance of

$$link(MCx, MCy) = \sum_{FPi \in MCx, FPj \in MCy} link(FPi, FPj) \tag{2}$$

The pair of clusters for which the above goodness measure is maximum is the finest pair of clusters to be merged at any given stage. This naive procedure may work well for well- echeloned clusters, however in this case of clusters with points that are vicinal, a huge cluster may ingest other clusters and thus, points from various clusters may be merged into a single cluster. In view of the fact, that a huge cluster typically would have a huge number of cross links with other clusters.

### C. *Links*

The number of links between two items is explained as the number of usual vicinal they have. The algorithms are a hierarchical structure agglomerative algorithm using the number of links as the parity measures rather than a measure based on gap. The ROCK algorithm uses the Jaccard coefficient to measure parity [17]. The ROCK clustering algorithm utilizes the knowledge about links between points when making opinion on the points to be merged into a single cluster, and it is very powerful. The Jaccard coefficient JC endows a suitable measure of the gap between points, clusters are infrequently not well separated and so a new measure of parity between points was proposed that meditate the neighborhood of a point. Supposing siml($A_i$, $A_j$) is the similarity between points, $A_i$ and $A_j$, (3) and $0 \le \theta \le 1$ is a parameter, then

$$link(Ai, Aj) = \left| \{JC : siml(Ai, B) \ge \theta\} \cap \{Q : siml(Aj, B) \ge \theta\} \right. \tag{3}$$

In other words, link (xi, xj) is the number of shared neighbors of xi and xj. The opinion is that two points will be close only if they share a relatively huge number of neighbors. Namely the policy is introduced to manage the issue of boundary points, which are nearby to each other, but convenient to various clusters.

### D. *Criterion Function*

For a clustering technique, arise an essential question is the following is it feasible to characterize the finest clusters. The purpose is to maximize the criterion function to get the finest quality clusters. Through maximizing we mean maximizing the sum of links of intra cluster point pairs while keeps down the sum of links of inter cluster point pairs. A criterion function is described (4) here to be

$$E_l = \sum_{z=1}^{nc} sc_z \times \sum_{FPi, FPj \in MCx} \frac{link(FPi, FPj)}{sc_z^{1+2f(\theta)}} \tag{4}$$

where nc is the number of clusters, $sc_z$ is the size of cluster *z*, MC*x,* notify the cluster *x* and *f (θ)* is a function of the parameter *θ*. To determine the function *f (θ)* is an arduous issue (similarity threshold). This criterion function will make sure that points with a huge number of links between them are allocated to the same cluster, it does not inhibit a clustering in which all points are allocated to a single cluster.

### E. *Neighbors*

If parity between two points exceeds certain parity threshold (θ), they are neighbors i.e., if siml(C, D)≥θ then only two points C, D are neighbors (5), where similarity is a parity function and θ is a user-specified threshold. The given threshold θ between 0 and 1, a pair of points C, D are described to be neighbors if the under mentioned hold.

$$siml(C, D) \ge \theta \tag{5}$$

## VI. The Rock Algorithm

The algorithm ROCK (Robust Clustering uses links) is produced by Sudipto Guha, Rastogi and Kyuseok Shim is a hierarchical clustering algorithm that employs the links to merge clusters. ROCK performs agglomerative hierarchical clustering and discover the idea of links for data with categorical attributes. This algorithm uses the link-based parity measure to measure the parity between two data points and between two clusters. ROCK employs the knowledge about links between data points when making appropriate decisions on the data points to be merged into a single cluster. The following Fig. 1 shows the general concept of the ROCK algorithm.

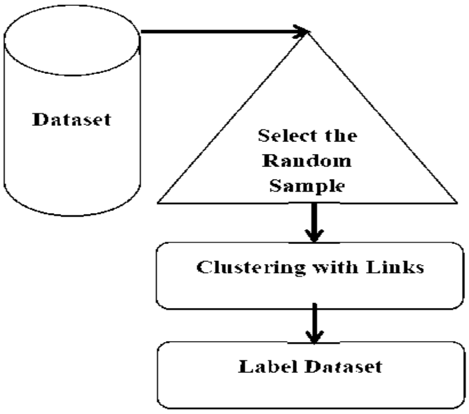
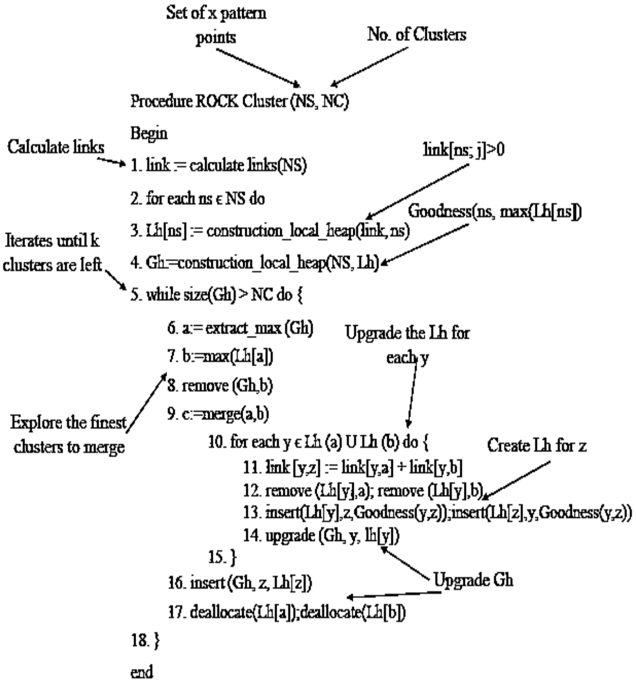


Fig. 1. The General Concept ROCK Algorithm Steps

ROCK algorithm is the finest favorable algorithm for clustering categorical data because it may use Jaccard or Cosine similarity coefficients to explore the similarity between the two data points and additional it uses the concept of links to determine the neighbors. It is arduous to handle and manage the enormous chunks of data, hereupon clustering can help groups them in order. We have noticed that the task of discovery or explore some document out of the huge amount of data is unwieldy. In addition, the reaction time of exploring the document is very high due to the high scale discovery among the data. This technique is to cluster the data in order to segregate the data into some conglomeration with alike features and hence to lack the query reaction time by exploring the clusters procured alternately whole database or data warehouse. This algorithm is echeloned into three general parts: firstly, procured a random sample of the data. Secondly execute clustering of the data using the link agglomerative method in other words goodness measure is used to determine which pair of points is merged at each step and thirdly using these clusters the still existing data on disk are assigned to them.

Now, we are present pseudo-code for the ROCK algorithm. It receives as input the set NS of n sampled points to be clustered in other words that are drawn arbitrarily from the real data set, and the number of required clusters NC. The series of steps begins by computing the number of links between pairs of points in stage 1. At the beginning, each point is a separate cluster. For each cluster ns, we build a local heap Lh [ns] and keep the heap during the implementation of the algorithm. Lh [ns] holds every cluster y like that link [ns, y] is non-zero. The clusters y in Lh [ns] is sequenced in the lessen order of the goodness measure with regard to ns Goodness (x, y). Besides to the local heaps Lh [ns] for each cluster nc, the algorithm also keeps an extra global heap Gh that holds all the clusters. Additionally, the clusters in Gh are sequenced in the diminishing sequenced of their best goodness measures. Consequently, Goodness (y, max(Lh[y])) is used to sequence the different clusters y in Gh, where max(Lh[y]), the max element in Lh[y], is the best cluster to merge with cluster y. At each stage, the max cluster y in Gh and the max cluster in Lh[y] are the best pair of clusters to be merged.

The while-loop in stage 5 iterates until only NC clusters stay in the global heap Gh. Therewith, it also inhibits clustering if the number of links between every pair of the still existing clusters becomes zero. At each stage of the while-loop, the max cluster a is extracted from Gh by extract max and Lh(a) is used to determine the finest cluster b for it. So far as clusters a and b will be merged, entries for a and b are no longer requisite and can be remove from Gh. Clusters a and b are then merged in stage 9 to create a cluster c containing (yay+yby) points.



```
Procedure ROCK Cluster (NS, NC)
Begin
1. link := calculate links(NS)
2. for each ns ϵ NS do
3. Lh[ns] := construction_local_heap(link, ns)
4. Gh:=construction_local_heap(NS, Lh)
5. while size(Gh) > NC do {
    6. a:= extract_max (Gh)
    7. b:=max(Lh[a])
    8. remove (Gh,b)
    9. c:=merge(a,b)
        10. for each y ϵ Lh (a) U Lh (b) do {
            11. link [y,z] := link[y,a] + link[y,b]
            12. remove (Lh[y],a); remove (Lh[y],b)
            13. insert(Lh[y],z,Goodness(y,z));insert(Lh[z],y,Goodness(y,z))
            14. upgrade (Gh, y, lh[y])
        15. }
    16. insert (Gh, z, Lh[z])
    17. deallocate(Lh[a]);deallocate(Lh[b])
18. }
end
```

Fig. 2. The ROCK Algorithm Iteration

There are two tasks that require to be abolished once clusters a and b are merged firstly for every cluster that hold a or b in its local heap, the elements a and b requirement to be replaced with the latest merged cluster c and the local heap necessity to be updated, and secondly a new local heap for c necessity to be created. These jobs are carried out in the for-loop of stage 10 to 15 shown in Fig. 2. The number of links between clusters y and c is straightforwardly the sum of the number of links between y and a, and y and b. This is used to calculate Goodness(y,z), the latest goodness measure for the pair of clusters y and z, and the two clusters are pushed into each other's local heap. Pay attention that Lh[z] can only hold clusters that were earlier either in Lh(a) or Lh(b) since these are the only clusters that have non-zero links with cluster z. As well as, pay attention that, as an outcome of merging clusters a and b, it is feasible that the cluster a or b was earlier the finest to be merged with y and at the moment z becomes the finest one for being merged. Moreover, it is also feasible that neither a nor b was the finest cluster to merge with y, in spite of that z is a finest cluster to merge with y. In such circumstances, meanwhile the max cluster in the local heap for y transformation, the algorithm necessity to transfer y in Gh to reect information relating to the new best cluster for y namely stage 14. The process also exigency to make certain that Gh holds the finest cluster to be merged for the latest cluster z.

## VII. The C4.5 Classification Algorithm

The decision trees are powerful and famous tools for classification and prediction [19]. Classification a method most appropriate in the area of medical diagnosis for automatic search of valid, novel, unknown, useful and conceivable knowledge from very huge databases be made up of data acquired from many applications. The C4.5 is a spaciousness of ID3, developed by Quinlan in 1993. This algorithm provides a decision tree for the given crop pest training data by recursively break apart that data [20]. The C4.5 algorithm is a continued development of the earlier algorithm, namely the ID3 algorithm. Consequently, the genuine ID3 and C4.5 algorithms have the same fundamental concept. It transforms the trained trees into sets of if-then precept and its manage discrete and continuous attributes. The decision trees generated by C4.5 can be used for classification since it is often referred to as a statistical classifier. The C4.5 different from its forerunner, that is capable to manage attributes with discrete or continuous type and capability to manage empty attribute (lost value). Also, can do pruning of branches and choice is done using a calculation attribute gain ratio. Now, at this place we are discussing the three principles of the work done by the C4.5 algorithm according to [1] and [20].

a) Firstly, the perform decision tree building. The main intention of this decision tree building algorithm is to create a model of a set of training data that will be utilized to predict the class of a latest data.

b) Secondly, the decision tree pruning, because the outcome of decision tree building can be bulky and not convenient to read, the C4.5 algorithm can make easier the decision tree with pruning based on the value of the level of assuredness. The pruning also purposes to detract the prediction inaccuracy rate on the latest data.

c) Thirdly, construction the regulation for the decision tree that has been building. The regulation is in if-then form that obtain knowledge from the decision tree by tracing from the root node to the leaf node.

### *A. Steps to Generate C4.5 Decision Tree Algorithm*

Input: Data training pattern, list of attributes and attribute_selection_ procedure

Ouput: Decision Tree

```
1. Create a node DTN,
2. If the pattern has the identical class, DTC
3. Then comeback DTN as a leaf node with class DTC
     label
4. If list of attributes is vacant
5. Then comeback DTN as a leaf node with class label
     that is the  most class in the samples.
6. Select test-attribute, that has the most Gain Ratio using
     attribute_selection_ procedure
7. Give node DTN with test-attribute label
8. For each ai pada test-attribute
9. Concatenate branch in node DTN to test-attribute = ai
10. Make division for pattern si from the pattern where
     test-attribute = ai;
11. If si is vacant
12. Then Connect leaf node with the most class in pattern
13. Else connect node that generates by
     Generate_decision_tree(si,attributelist, test-attribute)
14. end for
15. Comeback DTN;
```

Based on above algorithm decision tree model can be illustrated as follows. Suppose that there is one set of training data pattern P that have the attributes ($X_1$, $X_2$, $X_3$, ...) and classes consisting of ($C_1$, $C_2$, $C_3$, ...). The C4.5 algorithm will execute as follows. If P is not vacant and all the pattern has the identical class of $C_Y$, then the decision tree for P is a leaf node with label $C_Y$. If the attribute is vacant, then the decision tree contains a leaf node with label Cz where Cz is the highest class in the training pattern P. If P consists of a pattern that has a dissimilar class of the division P into $P_1$, $P_2$, $P_3$, ... $P_n$. The training pattern P division by isolating values of attribute XC, which at the time became the parent node. Assume that XC consists of 3 types of values that are $v_1$, $v_2$, $v_3$, then T will be division into three subsets, such as that the value of XC = $v_1$, $v_2$ = XC, and XC = $v_3$. This procedure sustains recursively with the base case of move 1 and move 2. Attribute that will serve as the parent node or attribute that will division the data is done by compute the gain. The gain is used to choose the attributes to be tested based on knowledge theory idea of entropy.

## VIII. The K-Means Clustering Algorithm

The K-mean is the most renowned partitioning technique of clustering. It was firstly proposed by MacQueen in 1967. K-mean is a numerical, unsupervised, iterative, non-deterministic technique of clustering [21]. In this object are classified as related to one of K-groups. The outcome of partitioning technique is a set of K clusters, each object of data set related to one cluster. In each cluster there may be a centroid or a cluster representative. K-means is a data mining algorithm which carries out clustering of the data pattern. As described earlier, [22] clustering means the split of a dataset into a number of groups, namely identical items fall or related to identical groups. In order to cluster the database, K-means algorithm uses an iterative procedure. The input in this case is the number of required clusters and the preparatory means and also produces finishing means as output. If in the algorithm need is to present K clusters then there will be K initial means and final means also [23]. Subsequently, cessation of this clustering algorithm, each object of dataset becomes a member of one cluster. The cluster is determined by discovery throughout the means for the intention to search the cluster having nearest mean [24] to the object. The cluster with the lowest distanced mean is cluster to which investigate objects be suited to. In this matter K-means algorithm, it attempts to group the data items in the dataset into desired number of clusters. To execute this task well it makes some iteration until some commingle criteria meet [25]. Subsequently, each iteration, lately calculated means are updated such that they become

DOI: http://dx.doi.org/10.24018/ejers.2018.3.12.989

near to the final means as well as at final, the algorithm intermingles and then inhibits performing iterations.

### *A. Steps of K-Means Clustering Algorithm*

The K-Means clustering algorithm is an opinion, in which there is required to classify the given data set into K clusters, the value of K (no. of clusters) is defined by the user which is determinate. In this first the centroid of [21] each cluster is chosen for clustering and then pursuant to the selected centroid, the data points having least distance from the given cluster, is allocated for that distinctive cluster. In the k-means using Euclidean Distance for calculating the distance of data point from the particular centroid. This algorithm be made up of four steps.

a) Initialization: In this first step data set, number of clusters and the centroid that we defined for each cluster.

b) Classification: In this second step distance is calculated for each data point from the centroid and the data point having least distance from the centriod of a cluster is allocated to that particular cluster.

c) Centroid Recalculation: In this third step cluster generated in advance, the centriod is again often calculated means recalculation of the centriod.

d) Convergence Situation: In this fourth step lateness when reaching a given or defined number of iterations as well as lateness when there is no exchange of data points between the clusters. Lateness when a threshold value is attained.

Supposing all of the above situation are not acquiescent, then move to step 2 and the whole process repeat once more, until the given situation are not acquiescent.

### *B. The Forecast of Cardiac Ailment using k-Means Clustering*

The calculate the centroid or mean of all objects in every cluster. So far as encore steps 2, 3 and 4 while the same points are entrusted to each cluster in consecutive rounds. The alteration doesn't make a material abnormality in the definition of the clusters. The clustering is accomplished on preprocessed data set using the K-means algorithm with the K values so as to educt relevant data to the cardiac ailment. The k-means is relatively an efficient technique and straightforward algorithm that has been appropriate for many healthcare problem domains. The clusters don't overlap characters and are also non-hierarchical in disposition. The K-means intense, substantial and easy to understand and its gives best outcome when the data set is distinct or well separated from each other data set, then it gives the best outcome.

## IX. The Proposed System Architecture

In this section, we are talking about proposed system architecture. The proposed architecture is shown in Fig. 3. In recent times, modern medicine generates an enormous amount of knowledge stored in the medical database. It is essential to extract useful information and providing scientific decision-making for the diagnosis and treatment of cardiac disease from the database increasingly becomes essential. Data mining in medicine can deal with this cardiac ailment issue. It can also ameliorate the management, quality of hospital information and encourage the growth of telemedicine and community medicine.

The main purpose of this system is to build cardiac ailment prediction system using a historical cardiac database that gives the diagnosis of cardiac disease. To build this system, medical terms, namely blood pressure, gender, cholesterol, chest pain, sugar etc 14 input attributes are used. The architecture uses an application of cardiac ailment medical data mining based on computation intelligence such as ROCK algorithm, C4.5 classification algorithm, and K-means clustering have been introduced. This Algorithm enables to take the cardiac ailment dataset and classify whether a person is having a cardiac ailment or not. The above algorithm contain cardiac ailment medical dataset, performs clustering relevant data and ROCK algorithm for categorical attributes item-sets. Afterwards, choose categorical attribute item data sets and classify the pattern using C4.5 algorithm is used to display accuracy and effective cardiac ailment level.

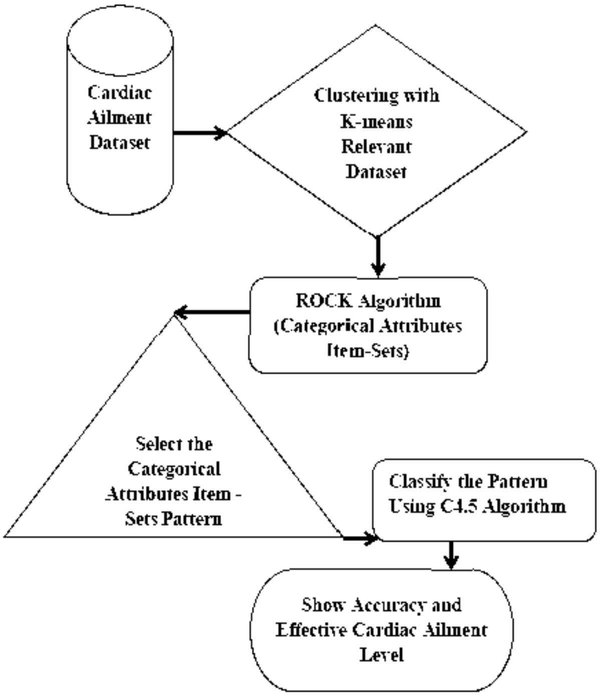

Cardiac Ailment Dataset
Clustering with K-means Relevant Dataset
ROCK Algorithm (Categorical Attributes Item-Sets)
Select the Categorical Attributes Item - Sets Pattern
Classify the Pattern Using C4.5 Algorithm
Show Accuracy and Effective Cardiac Ailment Level

Fig. 3. The Proposed Cardiac Ailment Prediction System Architecture Steps

## X. The Data Source and Outcome

In this section, we are used dataset with input attributes is obtained from Cleveland heart disease database [26]. With the assist of the recordset, the cardiac attack prediction with considerable patterns are extracted. The attribute diagnosis with value 1 is indicated as cardiac disease prediction and value 0 is indicated as no cardiac disease prediction for patients. Therein key attributes are patient Id and other attributes is used as input. The experimental outcome in recognizing essential patterns for predicting the cardiac ailment. The cardiac ailment database is preprocessed successfully by removing corresponding records and endow missing values as shown in below.

### A. Predictable Attribute

1. Diagnosis (value 0: <50% diameter narrowing (no cardiac disease); value 1: >50% diameter narrowing (has cardiac disease))

### B. Key Attribute

1. Patient Id – Patient's identification no

### C. Input Attribute

1. Sex (value 1: Male; value 0: Female)
2. Age in Year
3. Oldpeak – ST depression induced by exercise
4. Restecg – resting electrographic results (value 0: normal; value 1: having ST-T wave abnormality; value 2: showing probable or definite left ventricular hypertrophy)
5. Fasting Blood Sugar (value 1: >120 mg/dl; value 0: <120 mg/dl)
6. Slope – the slope of the peak exercise ST segment (value 1: unsloping; value 2: flat; value 3: downsloping)
7. Exang - exercise induced angina (value 1: yes; value 0: no)
8. Serum Cholesterol (mg/dl)
9. Trest Blood Pressure (mm Hg on admission to the hospital)
10. Thal (value 3: normal; value 6: fixed defect; value 7: reversible defect)
11. CA – number of major vessels colored by fluoroscopy (value 0-3)
12. Thalach – maximum heart rate achieved
13. Chest Pain Type (value 1:typical type 1 angina, value 2: typical type angina, value 3:non-angina pain; value 4: asymptomatic)
14. famhist: family history of coronary artery disease (1 = yes; 0 = no)

The elegant cardiac ailment data set, consequence from preprocessing, is then conjunct by K-means algorithm with the K value of 2 [27]. The conjunct contains the data associated with the cardiac ailment as shown in above and the further contains the left over number. Then the regular forms are mined efficiently from the collection applicable cardiac ailment, using the ROCK algorithm. The system consortiums of cardiac disease parameters for general and risk level belonging to their values and levels are listed above in that, ID lesser than of hexadecimal (000001) of weight contains the usual level of prediction cardiac ailment and higher ID other than hexadecimal (000001) comprise the higher risk levels cardiac disease and mention the prescription IDs. Next section displays the parameters of the cardiac ailment prediction with equivalent prescription hexadecimal key ID and their levels.

### D. The C4.5 Algorithm Decision Tree Structure

If Number of Years (Age)= < 40 and Overweight= True
and Liquor= Nevermore
Then
Cardiac Ailment Level is Least
(Or)
If Number of Years (Age)= >40 and Blood
pressure=High and Tobacco Smoking = Current
Time
Then
Cardiac Ailment Level is Utmost

The next section shows the instance of training data to forecast the cardiac ailment level and then Fig. 4 shows the efficient cardiac disease level with a tree using the C4.5 by information obtain.

TABLE I: THE ELEGANT CARDIAC AILMENT DATA SET

| Number Key ID | Hexadecimal Reference ID | Cardiac Ailment Attributes Description |
|---|---|---|
| 1 | 000001 | Gender |
| 2 | 000002 | Number of Years (Age) |
| 3 | 000005 | Painloc: Chest Pain Suffering Location |
| 4 | 00000A | CP: Chest Pain Type |
| 5 | 00000C | Relrest |
| 6 | 00001F | Chol: Serum Cholesterol in mg/dl |
| 7 | 00002C | Trestbps: Resting Blood Pressure |
| 8 | 00002D | Tobacco Smoking |
| 9 | 000030 | Cigarettes Per Day |
| 10 | 00003A | Number of Years as a Tobacco Smoker |
| 11 | 00003B | dm (1 = History of Diabetes; 0 = No Such History) |
| 12 | 00003C | fbs: (Fasting Blood Sugar > 120 mg/dl) |
| 13 | 00004A | Famhist: Family History Coronary Artery Disease |
| 14 | 00004B | Thalach: Maximum Cardiac Rate Achieved |
| 15 | 00004C | Sedentary Lifestyle and Inactivity |
| 16 | 00004F | Exang: Exercise Induced Angina |
| 17 | 000059 | Ca: Number of Major Vessels (0-3) Colored by Fluoroscopy |
| 18 | 00005A | Num: treatment of Cardiac Ailment |

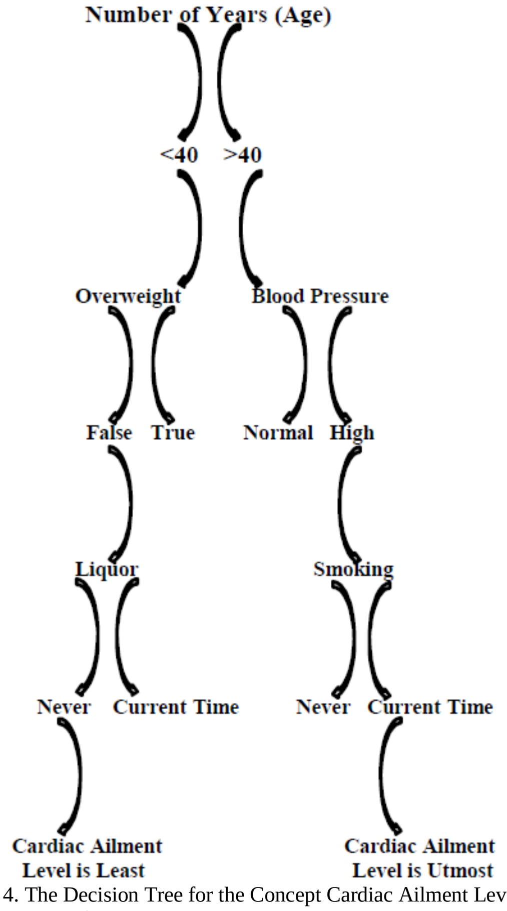


Fig. 4. The Decision Tree for the Concept Cardiac Ailment Level by Information Obtain Through C4.5 Algorithm

DOI: http://dx.doi.org/10.24018/ejers.2018.3.12.989

TABLE II: THE TRAINING DATA TO FORECAST THE CARDIAC AILMENT LEVEL

| **Cardiac Ailment** | **Cardiac Ailment** | **Cardiac Ailment** |
|---|---|---|
| **Hexadecimal Risk Level** | **Weights** | **Parameter Description** |
| 000001 | Ages<40 | Gender (Male and Female) |
| 000008 | Ages<40 | Gender (Male and Female) |
| 000009 | Overweight | Acknowledge |
| 000001 | | Not Acknowledge |
| 000001 | Alcohol | Nevermore |
| 000007 | | Current Time |
| 000003 | | Past |
| 000001 | Tobacco Smoking | Nevermore |
| 000006 | | Current Time |
| 000003 | | Past |
| 000009 | Exalted Saturated Meal | Acknowledge |
| 000001 | | Not Acknowledge |
| 000001 | | Exercise Regularly |
| 000006 | | Not Ever |
| 000008 | Exalted Salt Dies | Acknowledge |
| 000001 | | Not Acknowledge |
| 000007 | Sedentary Life | Acknowledge |
| 000001 | | Not Acknowledge |
| 000008 | Noxious Cholesterol | High |
| 000001 | | Usual |
| 000005 | Blood Sugar | High(>120&<400) |
| 000001 | | Usual(>90&<120) |
| 000004 | | Low (<90) |
| 000009 | Cardiac | Low (< 60bpm) |
| 000001 | | Usual(60 to 100) |
| 00000A | | High (>100bpm) |
| 000001 | Blood Pressure | Usual (130/89) |
| 000008 | | Low (< 119/79) |
| 00000B | | High (>200/160) |

The experimental outcome of our proposed approach as presented in Table III. The aim is to have high accuracy, as well as high precision and recall metrics. These can be without difficulty converted to accurate- definite (AD) and erroneous - definite (ED) metrics. The accurate - definite (AD) is the total percentage of members classified as class Z allied to class Z and erroneous - definite (ED) is the total percentage of members of class Z but does not allied to class Z. Thereafter, accurate - indefinite (AI) is the total percentage of members which do not allied to class Z are classified not a part of class Z. The erroneous - indefinite (EI) is the total percentage of members of class Z incorrectly classified as not allied to class Z. It can also be given as hundred percentages erroneous definite.

**Precision = AD / (AD +ED)**

**Recall = AD / (AD +EI)**

TABLE III. THE DIFFERENTIATION BETWEEN ROCK AND K-MEANS BASED ROCK WITH C4.5

| **Technique Used** | **Precision** | **Recall** | **Percentage of Accuracy (%)** |
|---|---|---|---|
| K-Means Based Rock | 0.79 | 0.71 | 76% |
| K-Means Based Rock with ID3 | 0.82 | 0.87 | 87% |
| K-Means Based Rock with C4.5 Classification Algorithm | 0.84 | 0.95 | 95% |

## XI. CONCLUSION

The data mining is the designated to as data or knowledge discovery, is the process of analyzing data and metamorphose it into intuition that informs business decisions. The Data mining software entitles to organizations to analyze data from various sources in order to detect patterns. Data mining technique in the healthcare sector is being used mainly for predicting different ailment as well as in assisting in diagnosis for the medical practitioner in making their clinical decision. The diagnosis is extensively being used in predicting diseases and they are extensively used in medical diagnosing. In this paper, we are proposing a cardiac Ailment prediction system using ROCK, C4.5 classification algorithm and k-means clustering. The ROCK is a clustering algorithm designed to deal with categorical data and execution well with actual categorical data, and creditably with time-series data. In this research work using ROCK, ID3, and C4.5 compare with each other. Among all classification technique the best appropriate classifier for the present study consisting of healthcare sector data is C4.5 decision tree classification algorithm. The accuracy of K-means based with ROCK 76%, K-Mean based ROCK with ID3 87%, and K-Mean based on ROCK with C4.5 classification algorithm 95% respectively. The ROCK algorithm and K-mean based on ID3 are studied and compared to K-mean based with C4.5 classification algorithm to achieve an efficient consequence in cardiac ailment diagnosis and to formulate a medication plan.

## REFERENCES

[1] Jiawei Han, Micheline Kamber, 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, USA, 2012.

[2] Yusuf Perwej, “An Experiential Study of the Big Data,” for published in the International Transaction of Electrical and Computer Engineers System (ITECES), USA, ISSN (Print): 2373-1273 ISSN (Online): 2373-1281, Vol. 4, No. 1, page 14-25, March 2017 DOI:10.12691/iteces-4-1-3.

[3] D. Luo, C. Ding, H. Huang, "Parallelization with Multiplicative Algorithms for Big Data Mining", Proc. IEEE 12th Int'l Conf Data Mining, pp. 489-498, 2012.

[4] Marco Viceconti, 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.

[5] Tan, P., Steinbach, M. and Kumar, V. Introduction to Data Mining, Addison-Wesley, Boston, 2006.

[6] Yusuf Perwej, Mohammed Y. Alzahrani, F. A. Mazarbhuiya, Md. Husamuddin, “The State of the Art Cardiac Illness Prediction Using Novel Data Mining Technique” International Journal of Engineering Sciences & Research Technology (IJESRT), ISSN: 2277-9655, Vol. 7, Issue 2, Page no. 725-739, February -2018. DOI: 10.5281/zenodo.1184068

[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] Fryar CD, Chen T, Li X. Prevalence of Uncontrolled Risk Factors for Cardiovascular Disease: United States, NCHS Data Brief, No. 103. Hyattsville, MD: National Center for Health Statistics, Centers for Disease Control and Prevention, US Dept of Health and Human Services; 2012.

[9] Jyoti Soni et.al. Predictive Data Mining for Medical Diagnosis: An Overview of Heart Disease Prediction; International Journal of Computer Applications (0975 – 8887) Volume 17– No.8, March 2011.

[10] Shadab Adam Pattekari and Asma Parveen, prediction system for heart disease using naïve bayes, International Journal of Advanced

Computer and Mathematical Sciences, ISSN 2230-9624. Vol 3, Issue 3, pp 290-294, 2012.

[11] Ms. Ishtake S.H, Prof. Sanap S.A., “Intelligent Heart Disease Prediction System Using Data Mining Techniques”, International J. of Healthcare & Biomedical Research, Volume: 1, Issue: 3, Pages 94-101, April 2013

[12] M.A.Nishara Banu and B.Gomathy,” Disease Forecasting System Using Data Mining Methods”, ISBN: 978-1-4799-3966-4, pp: 130-133, 2014

[13] Hlaudi Daniel Masethe, Mosima Anna Masethe-prediction of Heart Disease using Classification Algorithms; Proceedings of the World Congress on Engineering and Computer Science 2014.

[14] J.Vijayashree and N.Ch.SrimanNarayanaIyengar,” Heart Disease Prediction System Using Data Mining and Hybrid Intelligent Techniques: A Review”, International Journal of Bio-Science and Bio-Technology , Vol.8, No.4, pp. 139-148, 2016

[15] Zeinab Arabasadi, Roohallah Alizadehsani, Mohamad Roshanzamir, Hossein Moosaei, Ali Asghar Yarifard, Computer aided decision making for heart disease detection using hybrid neural network-Genetic algorithm, Computer Methods and Programs in Biomedicine, Volume 141, Pages 19-26, April 2017

[16] M. Al-Razgan, C. Domeniconi, and D. Barbara, “Random Subspace Ensembles for Clustering Categorical Data,” Supervised and Unsupervised Ensemble Methods and Their Applications, pp. 31-48, Springer, 2008.

[17] S. Guha, R. Rastogi, and K. Shim,” ROCK: A Robust Clustering Algorithm for Categorical Attributes”, 15th International Conference on Data Engineering, pp. 512-521, 2000

[18] Qiongbing Zhang, Lixin Ding, Shanshan Zhang, “A Genetic Evolutionary ROCK Algorithm” International Conference on Computer Application and System Modeling (ICCASM), 2010

[19] B Mobasher, R Cooley, S Jaideep et al., Comments on decision tree, New York:IEEE Press, 1999.

[20] Quinlan, J. R. C4.5: Programs for Machine Learning. Morgan Kaufmann Publishers, 1993.

[21] J. Macqueen, "Some methods for classification and analysis of multivariate observations", 5th Berkeley Symp. Math. Statist. Prob, pp. 281-297, 1967.

[22] M. Erisoglu, N. Calis, and S. Sakallioglu, "A New Algorithm for Initial Cluster Centers in K-means Agorithm, " Pattern Recognition Letters, vol. 32, no. 14, pp. 1701-1705, 2011.

[23] 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

[24] Y Jiang, C H. Y. Zhang, "Clustering Algorithm for Data-Mining[J]", Journal of Electronic and Information, vol. 27, no. 4, pp. 655-622, 2005.

[25] J Dong, M. Qi. K-means Optimization Algorithm for Solving Clustering Problem. Knowledge Discovery and Data Mining, pp52-55, 2009.

[26] David W. Aha & Dennis Kibler. "Instance-based prediction of heart-disease presence with the Cleveland database."

[27] B. Kovalerchuk, E. Vityaev, and J. Ruiz, "Consistent and complete data and "expert" mining in medicine'. In: Cios, K. (ed.) Medical Data Mining and Knowledge Discovery, Springer, Heidelberg, pp. 238-280, 2001.