A SECURED PREDICTION SYSTEM FOR HUMAN DISEASES USING A GENETIC ALGORITHM APPROACH TO DATA MINING

Miss. Snehal Ganesh Shinde [1], Dr. L.M.R.J Lobo [2]

[1] P.G Student, Department of Computer Science and Engineering, WIT, Solapur University, Solapur, India

[2] Associate Professor, Department of Computer Science and Engineering, WIT, Solapur University, Solapur, India

Abstract : Many changes are happening in life styles of people in growing countries like India in recent days. Such changes in environment, diet, pollution and stress have led to the scenario that human beings are affected by microorganisms causing fatal diseases. In India, human diseases have become a major reason of deaths. A number of people have worked in this area to detect a particular disease, but it may happen that a person may be suffering from more than one disease at a time. Our attempt therefore is to detect the diseases a patient is suffering from with the use of detail symptoms given by a patient regarding his health status. This detection will enable him to have an appropriate treatment in time. In the health sector Data Mining techniques can be used to play a major role to detect a disease. The aim of the system is to predict different types of arrhythmia by using Association Rule Mining name Apriori and generating optimized association rules using Genetic Algorithm. The predicted output is the best association rules which will be used for the optimal prediction of diseases. This system can be used in medical sector. It is observed that many people lose their lives due to untimely detection of diseases. Such a system would detect a disease in time and required treatment can be given to save the patient. Thus it has a social cause. The results achieved by the experimentation conducted by us indicates that rules with good accuracy (confidence) are achieved. The result of Apriori association rule algorithm achieved by the implemented system is compared with WEKA.
.

***IndexTerms* – Diseases, Association Rule Mining, Genetic Algorithm, Rules.**

I. INTRODUCTION

Today's health-care services have come a long way to provide medical care to the patients and protect them from different diseases. Human diseases are the main reason of death throughout the world, and the larger number of deaths arises in low and middle income countries like India. Medical practitioners continuously generate large amount of data in the field of biomedical. This data can be used for the early detection of the human diseases, which can support to reduce the number of diseases. Now a day’s many changes are happening in life styles of peoples in growing countries like India, human diseases have become a major reason of deaths. [5, 12].

Association Rule Mining is the most powerful technique in data mining to generate rules. Generation of rule involves two phases. The frequent itemsets are found in the first phase and second phase generates the rule. There are number of techniques for generating association rules. Apriori is most important algorithm for generating association rules. We have thus used Apriori Algorithm to generate strong and valid association rules. These rules are then optimized using Genetic Algorithm to get new rules. Using the best association rules, we predict different types of arrhythmia.

1.1 Introduction to Arrhythmia (Heart Disease and Abnormal Heart Rhythm)

Cardiac Arrhythmia occurs when electrical impulses in the heart works improperly. Heart rate about 50 to 100 beats per minute are normal. Abnormal heart rates and Arrhythmias don't necessarily occur together. Arrhythmias may occur with a normal heart rate, or with heart rates that are slow (called bradyarrhythmias -- less than 50 beats per minute). Arrhythmias may also occur with rapid heart rates (called tachyarrhythmia -- faster than 100 beats per minute). An arrhythmia can not cause any symptoms and be silent. During a physical exam, a doctor can detect an irregular heartbeat by taking your pulse or through an electrocardiogram (ECG).

The Types of Arrhythmias:

- Premature ventricular contractions (PVCs): These are amongst the most common arrhythmias and occur in people with and without heart disease. This is the skipped heartbeat which we all occasionally experience. In some people, it may be related to stress, too much caffeine or nicotine, or too much exercise. But sometimes, even heart disease or electrolyte imbalance can cause PVCs. People having a lot of PVCs, and/or symptoms associated with them, should be evaluated by a heart doctor. However, in most people, PVCs rarely need treatment and are usually harmless.
- Atrial fibrillation: Atrial fibrillation is a very common irregular heart rhythm which causes the atria, the upper chambers of the heart, to contract abnormally.
- Atrial flutter: This is an arrhythmia which is caused by one or more rapid circuits in the atrium. Compared to atrial fibrillation, atrial flutter is more organized and regular. This arrhythmia mostly occurs in people with heart disease and in the first week after heart surgery. It often converts to atrial fibrillation.
- Heart block: It is a delay or complete block of the electrical impulse as it travels from the sinus node to the ventricles. The level of the block or delay can occur in the AV node or HIS-Purkinje system. The heart may beat irregularly and, even, more slowly. Heart block is treated with a pacemaker, if serious.
- Sinus node dysfunction: It is a slow heart rhythm due to an abnormal SA (sinus) node. Significant SA node dysfunction that causes symptoms is treated with a pacemaker.
- Sinus tachycardy
- Sinus bradycardy

- Left bundle branch block
- Right bundle branch block
- Left ventricule hypertrophy
- Ischemic changes (Coronary Artery Disease)
- Old Anterior Myocardial Infarction
- Old Inferior Myocardial Infarction
- Supraventricular Premature Contraction

1.2 Data Mining:

Data Mining is a data analysis methodology used to identify hidden patterns from large amount of data. It has been successfully used in different areas for knowledge discovery. Data Mining or knowledge discovery has come out into view, as one of the most progressive areas in Communication Engineering, Information Technology and Biomedical Science. Nowadays different types of data mining methods have been used and developed. [6, 7]. Data mining is an important area of research and is preferably used in Healthcare domain which is an active interdisciplinary area of research. [11]

1.3 Algorithms used:

1.3.1 Association Rule Mining:

Association rules are if / then statements that help unwrap relationships between apparently unrelated data in a relational database or other information repository. An association rule comprises of two parts, an antecedent (if) and a consequent (then). An antecedent (if) is an item found in the data. A consequent (then) is found in combination with the antecedent.

Association rules are made through breaking down information for regular if/then examples and utilizing the parameters support and confidence for distinguishing the most imperative connections.

It is proposed to distinguish robust rules in databases utilizing two distinct proportions of intriguing quality. The first one is support which generates frequent item set from the provided database and the other one is confidence which is focuses on rule generation.

1.3.1.1 Apriori Algorithm:

Apriori algorithm is used to mine the frequent patterns in database. Support and Confidence are the normal methods used to measure the quality of association rule

Support –Support decides how frequently a given formulated rules is pertinent to a given informational index.

Confidence- Confidence decides how as often as possible things in Y show up in exchanges that contain X.

Support, $s(X\text{->}Y) = \dfrac{\sigma(X \cup Y)}{N}$

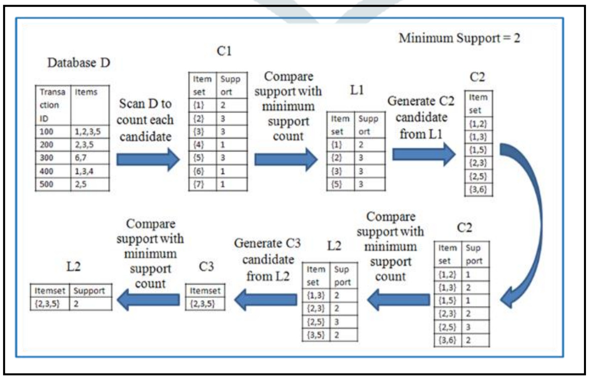
Confidence, $c(X\text{->}Y) = \dfrac{\sigma(X \cup Y)}{\sigma(X)}$

Where, X and Y disjoint item set. [10]

Terms related to this algorithm are as follows:

- Frequent Item Sets: The set of item which has minimum support and it is denoted by $L_{i\ \text{for}}$ i^{th} itemset.
- Apriori Property: Any subset of frequent itemset much be frequent.
- Join Operation: To find L_k, by joining L_{k-1} with itself a set of candidate k-itemsets is generated.
- Join Step: By joining Lk-1 with itself Candidate item C_k is generated.
- Prune Step: Any (k-1)-itemset cannot be a subset of a frequent k-itemset, if it is not frequent.

Figure 1 shows the generation of itemsets & frequent itemsets where the minimum support count is 2


Minimum Support = 2
Database D
Scan D to count each candidate
C1
Compare support with minimum support count
L1
Generate C2 candidate from L1
C2
C2
Compare support with minimum support count
L2
Generate C3 candidate from L2
C3
Compare support with minimum support count
L2

Transaction ID	Items
100	1,2,3,5
200	2,3,5
300	6,7
400	1,3,4
500	2,5

Item set	Support
{1}	2
{2}	3
{3}	3
{4}	1
{5}	3
{6}	1
{7}	1

Item set	Support
{1}	2
{2}	3
{3}	3
{5}	3

Item set
{1,2}
{1,3}
{1,5}
{2,3}
{2,5}
{3,6}

Item set	Support
{1,2}	1
{1,3}	2
{1,5}	1
{2,3}	2
{2,5}	3
{3,6}	2

Item set	Support
{1,3}	2
{2,3}	2
{2,5}	3
{3,5}	2

Itemset
{2,3,5}

Itemset	Support
{2,3,5}	2

Figure: 1. Generation of Itemsets & Frequent Itemsets

We use the following rule to generate the association rule from frequent itemset:

- Find all nonempty subset of L, for each frequent itemset L.
- write the association rule S → (L-S) For each nonempty subset of L, if support count of L/support count of S >= Minimum Confidence

Calculation of the best rule from the itemset L= {2, 3, 5} are as follows:

Consider the minimum support is 2 & minimum confidence is 0.7 (70%). All nonempty subset of {2,3,5} are: {2,3},{2,5},{3,5},{2},{3},{5}.

Rule 1: {2, 3} → {5}
Confidence = Support Count of ({2, 3, 5})/ Support Count of ({2, 3}) = 2/2 = 1 (100%)
Rule 2: {2, 5} → {3}
Confidence = Support Count of ({2, 3, 5})/ Support Count of ({2, 5}) = 2/3 = 0.67 (67%)
Rule 3: {3, 5} → {2}
Confidence = Support Count of ({2, 3, 5})/ Support Count of ({3, 5}) = 2/2 = 1 (100%)
Rule 4: {2} → {3, 5}
Confidence = Support Count of ({2, 3, 5})/ Support Count of ({2}) = 2/3 = 0.67 (67%)
Rule 5: {3} → {2, 5}
Confidence = Support Count of ({2, 3, 5})/ Support Count of ({3}) = 2/3 = 0.67 (67%)
Rule 6: {5} → {2, 3}
Confidence = Support Count of ({2, 3, 5})/ Support Count of ({5}) = 2/3 = 0.67 (67%)
Hence Rule 1 & Rule 3 are accepted rules as the confidence of these rules is greater than 70%

1.3.2 Genetic Algorithm

Genetic Algorithm (GA) depends on the standards of Genetics and Natural Selection, which has been connected in machine learning and streamlining issues and is a pursuit based enhancement method. [3]. It is usually used to discover ideal or close ideal answers for non-tractable issues which generally would take a lifetime to generate solutions. It is as often as possible used to tackle improvement issues.

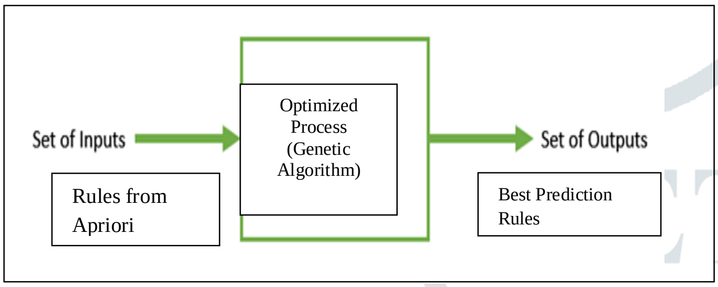


Figure: 2. Flow Chart of Genetic Algorithm

GA depends on similarity with the hereditary structure and conduct of chromosomes inside the number of inhabitants in people (chromosomes) having the establishment that people in a populace go after assets and mates. Those people which are best in every 'opposition' will create more posterity than those people that perform ineffectively. Qualities from `good' people proliferate all through the populace so two great guardians will in some cases create posterity which are superior to either parent. In this manner each progressive age will turn out to be more suited to their condition. The main ingredients of GA are Chromosomes, Selection, Recombination and Mutation.

1.3.2.1 Selection: An extent of the current populace is chosen to breed another age amid each progressive age. Fitness based process is utilized to choose singular arrangements where fitter arrangements (as estimated by a wellness work) are regularly more prone to be chosen. In this stage elitism could be utilized – the best n people are straightforwardly exchanged to the people to come. The elitism guarantees, that the estimation of the improvement work can't get most exceedingly bad (once the extremis is achieved it would be kept).

1.3.2.1Crossover: The most widely recognized compose is single point hybrid. In single point hybrid, we pick a locus time when you swap the rest of the alleles from one parent to the next. The kids take one segment of the chromosome from every individual parent. Chromosome is broken dependent on the arbitrarily chose hybrid point. This specific technique is called single point hybrid on the grounds that here just a single hybrid point exists. Some of the time just a single youngster is made, however by and large both posterity are made and put into the new populace. Hybrid does not generally happen. Once in a while, in light of a set likelihood, no hybrid happens and the guardians are straightforwardly duplicated to the new populace.

1.3.2.2 Mutation: We have another populace loaded with people (Chromosomes) where some are straightforwardly duplicated, and others are created by hybridisation. With the end goal to guarantee that the people are not all the very same, we permit a little possibility of change. We experience every one of the alleles of the considerable number of people, and if that allele is chosen for transformation, we either transform it by a little sum or supplant it with esteem. Change is genuinely basic. Notwithstanding, Mutation is crucial to guaranteeing hereditary assorted variety inside the populace. Genetic Algorithm is a randomized calculation that could be kept running for quite a while to acquire an ideal arrangement. [3]

II. RELATED WORK

Rahman et al. [1] Designed and executed a specialist framework dependent on continuous patient criticism that expects to give a reasonable choice, which would help in customizing the administration and additionally recognizing and distinguishing of any shunt glitches without the need to contact or visit the doctor's facility, for hydrocephalus administration and shunt conclusion. The win-prolog programming environment was used for developing the patient feedback expert system. The patient's details and symptoms are inputs, and the result of patient feedback analysis system is either that the patient needs to contact the physician or the problem is handled by modifying the opening times of the valve or assure him that the cause of the symptoms is not due to shunt complication. The system has the ability to identify the shunt state, i.e. problem existing or not and if yes identify such problem.

Kaysi et.al. [2] Developed an investigation that planned to foresee which patients become more cheerful and insight because of tDCS treatment by examining electroencephalography (EEG) of MDD patients that was gathered toward the beginning of tDCS treatment. This was accomplished through ordering power otherworldly thickness (PSD) of resting-state EEG utilizing bolster vector machine (SVM), straight segregate examination (LDA) and outrageous learning machine (ELM). Members were named as enhanced or not enhanced dependent on the adjustment in state of mind and subjective scores. The obtained classification results of all channel pair combinations were used to identify the most relevant brain regions and channels for this classification task. This represented an encouraging sign that EEG-based classification may help to tailor the selection of patients for treatment with tDCS brain stimulation.

Syam et.al [3] used Medical images for retrieval and the feature extraction along with colour, shape and texture feature extraction to extract the query image from the database medical images. At the point when a question picture was given, the highlights were separated and after that the Genetic Algorithm-based closeness measure was performed between the inquiry picture highlights and the database picture highlights. The Squared Euclidean Distance (SED) registered the comparability measure in deciding the Genetic Algorithm fitness. Consequently, from the Genetic Algorithm-based similitude measure, the database pictures that were pertinent to the given question picture were recovered. The CBIR procedure was assessed by questioning distinctive restorative pictures and the recovery effectiveness was assessed in the recovery results.

Dragulescu and Albu [4] developed an expert system to make some predictions regarding the hepatitis infection. This system implementation used three algorithms, Bayes's theorem, Aitken's formula and Logistic model
The system presented three important parts:
-The First part was a logical inference which was used to decide what type of hepatitis virus present for a new patient. The possibilities were B, B+D and C.
-The second part of the system was used to see the type and the grade using methods from statistical inference.
-The third part of the system - is made for the patients infected with hepatitis C virus and it predicted the biological parameters evolution during the treatment using artificial neural networks.

Singh et.al. [5] Developed a framework based on associative classification techniques on heart dataset for early diagnosis of heart based diseases. The attributes considered related to cause of heart diseases were - gender, age, chest pain type, blood pressure, blood sugar. The used Data mining algorithms namely Apriori, FP-Growth, Naive bayes, ZeroR, OneR, J48 and k-nearest neighbour. On basis of best results, using hybrid technique for classification associative they achieved a prediction accuracy of 99.19%

Nahar et.al. [6] Developed a research which assessed the performance of six well-known classification algorithms: Naive Bayes, SMO, IBK, AdaBoostM1, J48 and PART, using a number of performance matrices. The research explored medical (domain knowledge) knowledge based feature selection (MFS) on a real-life dataset and noted performance improvements compared to computer-automated feature selection for majority of the considered techniques and across majority of the performance measures. The findings could assist the design of a heart disease CAD and might also guide exploring the complicated symptoms, risk and prevention factors of the different disease for particular group of population.
Saraee et.al. [7] Proposed an approach for using data mining in classifying mortality rate related to accidents in children under 15. These data were gathered from the patient files that are recorded in the medical record section of the AL Zahra Hospital in Isfahan. The data mining methods used are decision tree and Bayes' theorem. DM techniques were applied to the data to bring about very interesting and valuable results.

Ranganatha et.al. [8] Developed a task to store restorative data of patients who sought hospitalization for coronary illness and calculations were kept running on that data and result were given as client reasonable words and chart. Data mining algorithms ID3 and Naïve Bayesian were used. The user had to login to enter the patient information. After login, patient information page is displayed where the user fills patient history form and it is given as an input to the algorithms. The algorithms are executed to give the result in the form of decision tree in case of ID3 and probability in case of Naïve Bayesian. The attributes of heart disease dataset included the Name, Age, Gender, Chest Pain Type, Rest ECG, CA (Coronary Angioplasty), Exang, Slope, FBS (Fasting Blood Sugar).

Palaniappan and Awang [9] built up an Intelligent Heart Disease Prediction System utilizing information mining strategies. Information mining procedures are Decision Trees, Naïve Bayes and Neural Network. It could answer complex questions for diagnosing coronary illness. To assemble the mining models the CRISP-DM technique were utilized. Business understanding, information understanding, information arrangement, displaying, assessment, sending are six noteworthy stages. To construct and access the models DMX question dialect and capacities are utilized. Lift Chart and Classification Matrix techniques were utilized to assess the adequacy of the models.

Song et.al. [10] Developed a project which establishes the intelligent diagnosis model of lung cancer based on the testing data from clinical diagnoses and by means of interactive regulation tapping. They respectively analysed 12 targets of 100 patients through multi-subjects combined means of radioimmunology, enzyme linked immunosorbent assay and chemistry so as to discover the interactive regulation between cancer and its likely causes and guide the diagnosis and prevention of lung cancer with the regulation model. The result showed that this method is superior to the conventional statistics of quantitative medicine.

Pandey [11] built up an investigation that intended to give a survey of information mining in the domain of human services. They talked about that information mining can be gainful in therapeutic space. Focal points and inconveniences of every now and again utilized information mining systems in the area of human services and therapeutic information have been looked at. Distinctive information mining methods, their points of interest and disadvantages are investigated. For compelling usage of these systems in human services space, there was a need to improve and secure wellbeing information sharing among different gatherings. Uniqueness of information mining as for restorative information is likewise tended to. The requirements and troubles identified with protection affectability and vast volume of medicinal information assume imperative job in determination of the specific information mining procedure. Moral and legitimate parts of medicinal information were additionally essential viewpoints. In view of its appropriateness to all individuals restorative information can have an extraordinary status.

Xao [12] built up a forecast model of coming year's FBG, in view of four years' verifiable restorative examination information, utilizing customary information mining procedures with a novel calculation to appraise the FBG change likelihood and a proposed highlight determination calculation, that joins the component significance scores of gathering learning and Sequential Backward Selection (SBS) calculation to choose an ideal element subset. Exploratory information was gathered from a medicinal examination database containing 108,386 clients, in which 7,136 individuals have four years' records. By contrasting the exploratory outcomes and arbitrary woodland and SVM, the element determination technique can adequately enhance the model execution.

III. METHODOLOGY

3.1 Steps for implementation of proposed system

Figure 3 shows an execution of proposed system. In this system the patients' disease details are collected and stored in a dataset. This is given to an association rule mining algorithm name Apriori. The algorithm returns rules that represent the dataset. However these rules are too many and not optimized. Genetic algorithm is then applied to generate the best rules that predict the disease suffered by the patient.

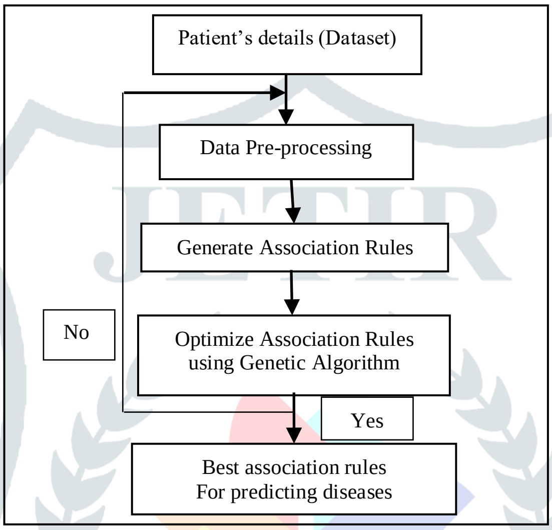

Patient's details (Dataset)
Data Pre-processing
Generate Association Rules
No
Optimize Association Rules using Genetic Algorithm
Yes
Best association rules For predicting diseases

Figure: 3.An Execution of System

IV. RESULT AND DISCUSSION

4.1 Dataset

Cardiac Arrhythmia Data Set was collected from the UCI Machine Learning Repository from which 452 instances and 27 attributes are taken as dataset. This dataset contains attribute keys, attribute values and classes. Here, one instance represents one patient's details. The attribute information contains patient's personal details and ECG recordings which are used to determine the type of arrhythmia. Attribute information is shown in Table 1.

Table: 1. Description of Attributes

Attributes	Description
0 Age	Age in years.
1 Sex	Sex (0 = male; 1 = female)
2 Height	Height in centimetres.
3 Weight	Weight in kilograms.
4 QRS duration	Average of QRS duration in msec
5 P-R interval	Average duration between onset of P and Q waves in msec.
6 Q-T interval	Average duration between onset of Q and offset of T waves in msec.
7 T interval	Average duration of T wave in msec.
8 P interval	Average duration of P wave in msec.
Vector angles in degrees on front plane of: 9 QRS 10 T	

11 P 12 QRST 13 J 14 Heart rate: Number of heart beats per minute
Of channel DI: Average width, in msec., of: 15 Q wave 16 R wave 17 S wave 18 R' wave, small peak just after R 19 S' wave
20 Number of intrinsic deflections, 21 Existence of ragged R wave 22 Existence of diphasic derivation of R wave 23 Existence of ragged P wave 24 Existence of diphasic derivation of P wave 25 Existence of ragged T wave 26 Existence of diphasic derivation of T wave

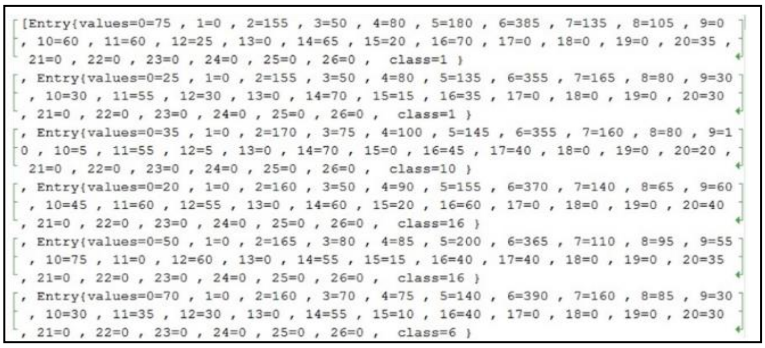
The dataset distinguishes between the presence and absence of cardiac arrhythmia and classifies it in one of the 16 classes. Normal ECG is represented by Class 01, different classes of arrhythmia by classes 02 to 15 and rest of unclassified ones by class 16. Different classes of arrhythmia are shown in Table 2.

Table: 2. Classes Of Arrhythmia

Class code	**Class**	**Number of instances**
01	Normal	245
02	Ischemic changes (Coronary Artery Disease)	44
03	Old Anterior Myocardial Infarction	15
04	Old Inferior Myocardial Infarction	15
05	Sinus tachycardy	13
06	Sinus bradycardy	25
07	Ventricular Premature Contraction (PVC)	3
08	Supraventricular Premature Contraction	2
09	Left bundle branch block	9
10	Right bundle branch block	50
11	1. degree AtrioVentricular block	0
12	2. degree AV block	0
13	3. degree AV block	0
14	Left ventricule hypertrophy	4
15	Atrial Fibrillation or Flutter	5
16	Others	22

4.2 Data Pre-Processing:

The aim of this step is to produce meaningful dataset from UCI dataset as input. In this step, The UCI dataset is converted into correct format for implementation of associative techniques. We create a key – value dataset, which contains combination of attribute keys, attribute values and classes. The missing attribute values in dataset are replaced by zero and we have quantized the attribute values i.e. large set of values are reduced to a discrete set, to make the dataset meaningful. The dataset is as shown in the figure 4.



```
[Entry{values=0=75 , 1=0 , 2=155 , 3=50 , 4=80 , 5=180 , 6=385 , 7=135 , 8=105 , 9=0
, 10=60 , 11=60 , 12=25 , 13=0 , 14=65 , 15=20 , 16=70 , 17=0 , 18=0 , 19=0 , 20=35 ,
 21=0 , 22=0 , 23=0 , 24=0 , 25=0 , 26=0 ,  class=1 }
, Entry{values=0=25 , 1=0 , 2=155 , 3=50 , 4=80 , 5=135 , 6=355 , 7=165 , 8=80 , 9=30
 , 10=30 , 11=55 , 12=30 , 13=0 , 14=70 , 15=15 , 16=35 , 17=0 , 18=0 , 19=0 , 20=30
, 21=0 , 22=0 , 23=0 , 24=0 , 25=0 , 26=0 ,  class=1 }
, Entry{values=0=35 , 1=0 , 2=170 , 3=75 , 4=100 , 5=145 , 6=355 , 7=160 , 8=80 , 9=1
0 , 10=5 , 11=55 , 12=5 , 13=0 , 14=70 , 15=0 , 16=45 , 17=40 , 18=0 , 19=0 , 20=20 ,
 21=0 , 22=0 , 23=0 , 24=0 , 25=0 , 26=0 ,  class=10 }
, Entry{values=0=20 , 1=0 , 2=160 , 3=50 , 4=90 , 5=155 , 6=370 , 7=140 , 8=65 , 9=60
 , 10=45 , 11=60 , 12=55 , 13=0 , 14=60 , 15=20 , 16=60 , 17=0 , 18=0 , 19=0 , 20=40
, 21=0 , 22=0 , 23=0 , 24=0 , 25=0 , 26=0 ,  class=16 }
, Entry{values=0=50 , 1=0 , 2=165 , 3=80 , 4=85 , 5=200 , 6=365 , 7=110 , 8=95 , 9=55
 , 10=75 , 11=0 , 12=60 , 13=0 , 14=55 , 15=15 , 16=40 , 17=40 , 18=0 , 19=0 , 20=35
, 21=0 , 22=0 , 23=0 , 24=0 , 25=0 , 26=0 ,  class=16 }
, Entry{values=0=70 , 1=0 , 2=160 , 3=70 , 4=75 , 5=140 , 6=390 , 7=160 , 8=85 , 9=30
 , 10=30 , 11=35 , 12=30 , 13=0 , 14=55 , 15=10 , 16=40 , 17=0 , 18=0 , 19=0 , 20=30
, 21=0 , 22=0 , 23=0 , 24=0 , 25=0 , 26=0 ,  class=6 }
```

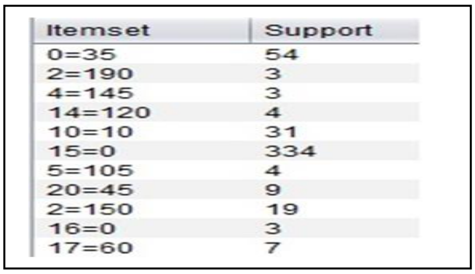
Figure: 4. Snapshot Of Dataset

Here 0 = 75 is one item in dataset. (In 0=75, 0 is a key of attribute of age and 75 a value of key i.e. attribute value).

4.3 Algorithms:

4.3.1 Apriori Algorithm:

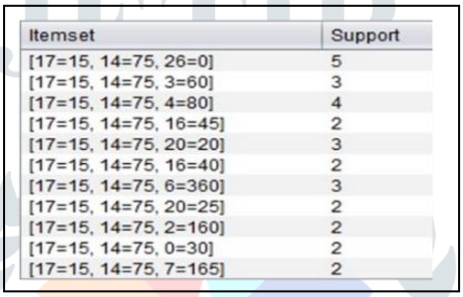
4.3.1.1 Candidate Generation: In this step, we generate candidates that finally give frequent itemsets. For candidate generation first we collect all items from the dataset. Then we find the support count (number of times each item has occurred in the dataset) for each item by using support formula. These are called as candidates. Only those candidates are collected which satisfies the pre-specified minimum support threshold (support > or = minsupport). This resulting set is called as List1. Then, List 1 is combined with itself to find new candidates. Again we find support for each of these candidates and only those candidates are collected which satisfies the minimum support threshold. This resulting set is called as List 2. And this process is continued further until no more itemsets are found. Figure 5 shows the candidate itemset.



Itemset	Support
0=35	54
2=190	3
4=145	3
14=120	4
10=10	31
15=0	334
5=105	4
20=45	9
2=150	19
16=0	3
17=60	7

Figure 5: Snapshot of Candidate Itemset

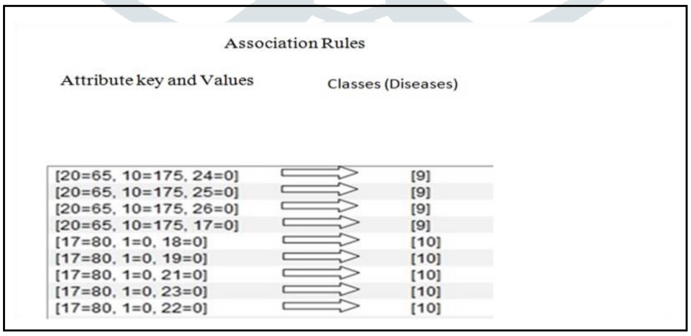
4.3.1.2 Frequent Itemset: The itemsets generated as the result of the above candidate generation step are called as frequent itemsets. The support of a frequent itemset satisfies the minimum support threshold. Each subset of frequent itemsets is also frequent, if not, that frequent itemset is rejected. Frequent itemsets are used for generating association rules. Frequent itemset is as shown in Figure 6.



Itemset	Support
[17=15, 14=75, 26=0]	5
[17=15, 14=75, 3=60]	3
[17=15, 14=75, 4=80]	4
[17=15, 14=75, 16=45]	2
[17=15, 14=75, 20=20]	3
[17=15, 14=75, 16=40]	2
[17=15, 14=75, 6=360]	3
[17=15, 14=75, 20=25]	2
[17=15, 14=75, 2=160]	2
[17=15, 14=75, 0=30]	2
[17=15, 14=75, 7=165]	2

Figure: 6. Snapshot of Frequent Itemset

4.3.1.3 Generating Rules: After generating frequent itemset, we generate strong association rules from the frequent itemset. First, we check classes for each frequent itemset from the dataset. An association rule involves an 'if' and a 'then' statement. Here, frequent itemsets are 'if' statement and classes (diseases) are 'then' statement. We calculate the confidence value of association rules which shows the number of times the if/then statements have found to be true. The association rules having the confidence value greater than minimum confidence are considered as final rules. All these rules satisfy both minimum confidence and minimum support threshold. Figure 7 shows the generation of rules.



Association Rules

Attribute key and Values		Classes (Diseases)
[20=65, 10=175, 24=0]	⇨	[9]
[20=65, 10=175, 25=0]	⇨	[9]
[20=65, 10=175, 26=0]	⇨	[9]
[20=65, 10=175, 17=0]	⇨	[9]
[17=80, 1=0, 18=0]	⇨	[10]
[17=80, 1=0, 19=0]	⇨	[10]
[17=80, 1=0, 21=0]	⇨	[10]
[17=80, 1=0, 23=0]	⇨	[10]
[17=80, 1=0, 22=0]	⇨	[10]

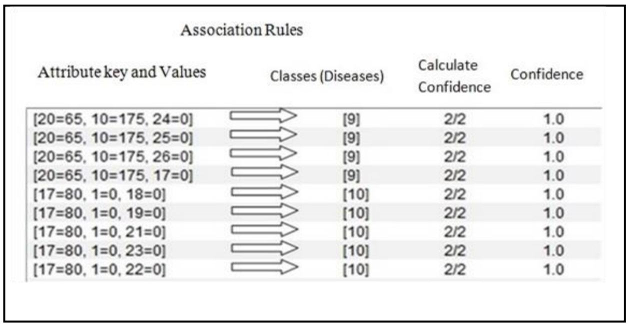
Figure: 7. Snapshot of Generating Rules

Output of Association Rule:

The result of the system using Apriori Association Rule Algorithm with minimum support 2, minimum confidence 0.7 and 3 iterations is obtained. Figure 8 shows the output of association rules.

Here, each attribute keys and attribute values represent patient's details and the corresponding classes represent the type of arrhythmia that the patient may have.

To find best rules from it, we optimize these rules using genetic algorithm.



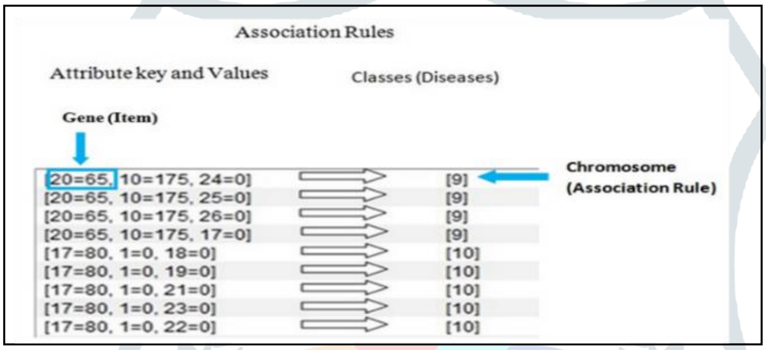
Association Rules

Attribute key and Values		Classes (Diseases)	Calculate Confidence	Confidence
[20=65, 10=175, 24=0]	⇨	[9]	2/2	1.0
[20=65, 10=175, 25=0]	⇨	[9]	2/2	1.0
[20=65, 10=175, 26=0]	⇨	[9]	2/2	1.0
[20=65, 10=175, 17=0]	⇨	[9]	2/2	1.0
[17=80, 1=0, 18=0]	⇨	[10]	2/2	1.0
[17=80, 1=0, 19=0]	⇨	[10]	2/2	1.0
[17=80, 1=0, 21=0]	⇨	[10]	2/2	1.0
[17=80, 1=0, 23=0]	⇨	[10]	2/2	1.0
[17=80, 1=0, 22=0]	⇨	[10]	2/2	1.0

Figure: 8. Snapshot of Association Rules Output

4.3.2 Genetic Algorithm:

Genes - In this system, genes are items in the association rules.
Chromosomes - In this system, association rules are chromosomes.
4.3.2.1 Population: In this system, Population is the set of chromosomes i.e. set of Association Rules. Figure 9 shows the population.



Association Rules

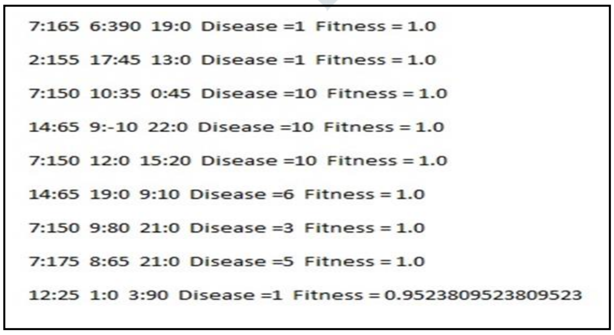
Attribute key and Values		Classes (Diseases)
[20=65, 10=175, 24=0]	⇨	[9]
[20=65, 10=175, 25=0]	⇨	[9]
[20=65, 10=175, 26=0]	⇨	[9]
[20=65, 10=175, 17=0]	⇨	[9]
[17=80, 1=0, 18=0]	⇨	[10]
[17=80, 1=0, 19=0]	⇨	[10]
[17=80, 1=0, 21=0]	⇨	[10]
[17=80, 1=0, 23=0]	⇨	[10]
[17=80, 1=0, 22=0]	⇨	[10]

Figure: 9. Snapshot of Population

4.3.2.2 Fitness function: In this step, we find the Fitness value of each rule (chromosomes) in the population. This fitness values shows the fittest value of rules (chromosomes). The confidence of item is considered to be fitness value on the following condition,
Fitness value = Association Rule Confidence, if Confidence > 0.8
4.3.2.3 Selection and Crossover: In selection step, we select the rules from the population according to their fitness value. After selection, we use single point crossover method for crossover. Here, we choose a random crossover point from where the items (genes) on either side of the rule (parent) are replaced. Crossover rate is 40%. As a result, it forms a best rule (offspring).
4.3.2.4 Mutation: In this step, to ensure that the best rules (offspring) are not exactly same to each other, we allow a small chance of mutation. Here, we change the item (gene) by a small amount. We have set the mutation rate to 5%.

Output of Genetic Algorithm:

Output of Genetic Algorithm using maximum generation 100 is generation of best rules from Association Rules. Figure 10 shows the output of genetic algorithm as best Association Rules.



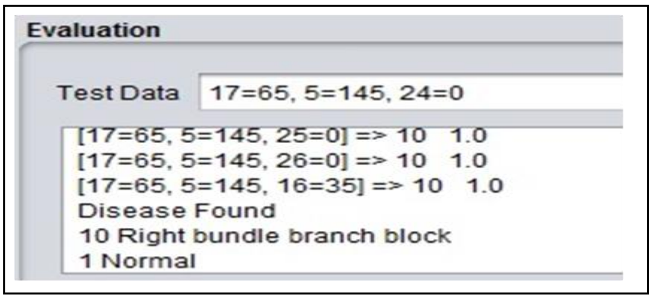
```
7:165 6:390 19:0 Disease =1 Fitness = 1.0

2:155 17:45 13:0 Disease =1 Fitness = 1.0

7:150 10:35 0:45 Disease =10 Fitness = 1.0

14:65 9:-10 22:0 Disease =10 Fitness = 1.0

7:150 12:0 15:20 Disease =10 Fitness = 1.0

14:65 19:0 9:10 Disease =6 Fitness = 1.0

7:150 9:80 21:0 Disease =3 Fitness = 1.0

7:175 8:65 21:0 Disease =5 Fitness = 1.0

12:25 1:0 3:90 Disease =1 Fitness = 0.9523809523809523
```

Figure: 10. Snapshot of Output of Genetic Algorithm-Best
Association Rules

Here, each attribute keys and attribute values represent patient's details and the corresponding classes (disease) represent the type of arrhythmia that the patient may have.

4.4 *Evaluation (Predict disease):*

In Evaluation, we predict if the person is normal or suffering from one of the types of arrhythmia. In this, we provide a person's details. These details are then matched with the generated association rules. The matched rules and the prediction of diseased or normal is displayed. Figure 11 shows the evaluation of given person details.



Evaluation

Test Data 17=65, 5=145, 24=0

```
[17=65, 5=145, 25=0] => 10   1.0
[17=65, 5=145, 26=0] => 10   1.0
[17=65, 5=145, 16=35] => 10   1.0
Disease Found
10 Right bundle branch block
1 Normal
```

Figure: 11. Snapshot of Evaluation (Predict Disease)

V. CONCLUSION AND FUTURE SCOPE

In this system Association Rule Mining Algorithm is used for generating the rules from the disease dataset. These rules are then optimized using Genetic Algorithms to get the best rules. The predicted output is the best association rules which will be used for the optimal prediction of diseases. When we compared the result of Apriori association rule algorithm with that of open source data mining tool WEKA, we get Apriori association rules only for symptoms to symptoms mapping with normal person symptoms, while our system produces Apriori association rules for symptoms to disease mapping. This system can be used in the medical sector. Such a proposed system would detect a disease in time and required treatment can be given to save the patient. Thus it has a social cause.

REFERENCES

[1] AbdelRahmanAlkharabsheh,LinaMomani,NayelAl-Zu'bi,WaleedAl-Nuaimy" An expert system for hydrocephalus patient feedback" in Annual International IEEE Conference of the IEEE Engineering in Medicine and Biology,2010

[2] Alaa M. Al-Kaysi, Ahmed Al-Ani, Colleen K. Loo, Michael Breakspear,Tjeerd W. Boonstra "Predicting brain stimulation treatment outcomes of depressed patientsthrough the classification of EEG oscillations" in proceedings of 38th Annual International IEEE Conference of the Engineering in Medicine and Biology Society (EMBC),2016

[3] B. Syam, J. Sharon Rose Victor, Y. SrinivasaRao"Efficient similarity measure via Genetic algorithm for content basedmedical image retrieval with extensive features" in proceedings of IEEE International Mutli-Conference on Automation, Computing, Communication, Control and Compressed Sensing (iMac4s),2013

[4] DoinaDragulescu, Adriana Albu "Expert System for Medical Predictions" in proceedings of 4th International IEEE Symposium on Applied Computational Intelligence and Informatics,2007

[5] Jagdeep Singh, AmitKamra and Harbhag Singh "Prediction of heart diseases using associative classification" in proceedings of 5th International IEEE Conference on Wireless Networks and Embedded Systems (WECON), 2016

[6] JesminNahar, Tasadduq Imam, Kevin S. Tickle, Debora Garcia-Alonso "Medical Knowledge based Data Mining forCardiac Stress Test Diagnostics"in proceeding of 2nd Asia-Pacific World Congress on Computer Science and Engineering (APWC on CSE),2015

[7] MohammadHosseinSaraee, ZahraEhghaghi,HodaMeamarzadeh,BahareZibanezhad"Applying data mining in medical data with focus on mortality related to accident in children" in proceedings of International IEEE Multitopic Conference,2008

[8] S.Ranganatha, H.R.Pooja Raj, C.Anusha, S.K. Vinay "Medical data mining and analysis for heart disease dataset using classification technique" in proceedings of National IEEE Conference on Challenges in Research & Technology in the Coming Decades (CRT 2013), 2013

[9] SellappanPalaniappan,RafiahAwang"Intelligent heart disease prediction system using data mining techniques" in proceedings of International IEEE/ACS Conference on Computer Systems and Applications,2008

[10] Shaoyun Song, Yu Ma"The Research and Application of Technology in the Diagnosis of Lung Cancer Warning Association Rule Mining" in proceedings of 8th International Conference on Information Technology in Medicine and Education (ITME),2016

[11] Subhash Chandra Pandey "Data mining techniques for medical data: A review" in proceedings of International IEEE Conference on Signal Processing, Communication, Power and Embedded System (SCOPES),2016

[12] WenxiangXao, Fengjing Shao, Jun Ji, Rencheng Sun, ChunxiaoXing"Fasting Blood GlucoseChangePredictionModelBasedon MedicalExamination Data and Data Mining Technique" in proceedings ofInternational IEEE Conference on Smart City/SocialCom/SustainCom (SmartCity),2015