A STRATEGY TO CLINICALLY PRACTICE DATA MINING FOR PROGNOSTICATION AND ASSAY OF MALADIES.

Dissertation submitted in fulfilment of the requirements for the Degree of

MASTER OF TECHNOLOGY

IN

COMPUTER SCIENCE AND ENGINEERING

By

KHALID AMIN WANI

11607375

Supervisor

MS. MANJIT KAUR

School of Computer Science and Engineering

Lovely Professional University

Phagwara, Punjab (India)

November 2017

TOPIC APPROVAL PERFORMA

School of Computer Science and Engineering

Program: P172:: M.Tech. (Computer Science and Engineering) [Full Time]

COURSE CODE : CSE548 **REGULAR/BACKLOG :** Regular **GROUP NUMBER :** CSERGD0321

Supervisor Name : Manjit Kaur **UID :** 22361 **Designation :** Assistant Professor (Contract Basis)

Qualification :_________________________ **Research Experience :** _________________________

SR.NO.	NAME OF STUDENT	REGISTRATION NO	BATCH	SECTION	CONTACT NUMBER
1	Khalid Amin Wani	11607375	2016	K1637	9858445645

SPECIALIZATION AREA: Program Methodology and Design **Supervisor Signature:**____________________

PROPOSED TOPIC: A STRATEGY TO CLINICALLY PRACTICE DATA MINING FOR PROGNOSTICATION AND ASSAY OF MALADIES.

Qualitative Assessment of Proposed Topic by PAC		
Sr.No.	**Parameter**	**Rating (out of 10)**
1	Project Novelty: Potential of the project to create new knowledge	7.20
2	Project Feasibility: Project can be timely carried out in-house with low-cost and available resources in the University by the students.	7.20
3	Project Academic Inputs: Project topic is relevant and makes extensive use of academic inputs in UG program and serves as a culminating effort for core study area of the degree program.	7.40
4	Project Supervision: Project supervisor’s is technically competent to guide students, resolve any issues, and impart necessary skills.	6.80
5	Social Applicability: Project work intends to solve a practical problem.	7.00
6	Future Scope: Project has potential to become basis of future research work, publication or patent.	7.00

PAC Committee Members		
PAC Member 1 Name: Gaurav Pushkarna	UID: 11057	Recommended (Y/N): Yes
PAC Member 2 Name: Er.Dalwinder Singh	UID: 11265	Recommended (Y/N): Yes
PAC Member 3 Name: Harwant Singh Arri	UID: 12975	Recommended (Y/N): Yes
PAC Member 4 Name: Balraj Singh	UID: 13075	Recommended (Y/N): Yes

PAC Member 5 Name: Raj Karan Singh	UID: 14307	Recommended (Y/N): NA
PAC Member 6 Name: Harleen Kaur	UID: 14508	Recommended (Y/N): NA
PAC Member 7 Name: Sawal Tandon	UID: 14770	Recommended (Y/N): NA
PAC Member 8 Name: Tejinder Thind	UID: 15312	Recommended (Y/N): Yes
DAA Nominee Name: Kuldeep Kumar Kushwaha	UID: 17118	Recommended (Y/N): NA

Final Topic Approved by PAC: **A STRATEGY TO CLINICALLY PRACTICE DATA MINING FOR PROGNOSTICATION AND ASSAY OF MALADIES.**

Overall Remarks: Approved

PAC CHAIRPERSON Name: 11024: Amandeep Nagpal **Approval Date:** 04 Nov 201711/20/2017 7:04:57 PM

DECLARATION STATEMENT

I hereby declare that the research work reported in the dissertation proposal "A STRATEGY TO CLINICALLY PRACTICE DATA MINING FOR PROGNOSTICATION AND ASSAY OF MALADIES" in partial fulfilment of the requirement for the award of Degree for Master of Technology in Computer Science and Engineering at Lovely Professional University, Phagwara, Punjab is an authentic work carried out under supervision of my research supervisor Ms. Manjit Kaur. I have not submitted this work elsewhere for any degree or diploma.

I understand that the work presented herewith is in direct compliance with Lovely Professional University's Policy on plagiarism, intellectual property rights, and highest standards of moral and ethical conduct. Therefore, to the best of my knowledge, the content of this dissertation represents authentic and honest research effort conducted, in its entirety, by me. I am fully responsible for the contents of my dissertation work.

KHALID AMIN WANI

REG.NO: 11607375

SUPERVISOR'S CERTIFICATE

This is to certify that the work reported in the M.Tech Dissertation proposal "**A STRATEGY TO CLINICALLY PRACTICE DATA MINING FOR PROGNOSTICATION AND ASSAY OF MALADIES"**, submitted by **KHALID AMIN WANI** at **Lovely Professional University, Phagwara, India** is a bonafide record of his original work carried out under my supervision. This work has not been submitted elsewhere for any other degree.

Ms. Manjit Kaur

Date: 30-11-2017

ACKNOWLEDGEMENT

To start with, I would like to thank Almighty ALLAH for showering His unlimited blessings and for giving me the power to think and work. Never to forget praising ALLAH in the foremost, the Creator of treasures of knowledge and wisdom and prays for Prophet Mohammad (SAW).

I would like to express my sincere gratitude and appreciation to my mentor and it is my proud privilege and duty to acknowledge the kind of help and guidance received from my mentor Assistant Prof. Ms. Manjit Kaur. I am fortunate to work on this thesis under her supervision.

I would like to express an earnest gratitude to the faculty of Department of Computer Science & Engineering, Lovely Professional University for their helping hand time to time.

Last but not the least, I would like to thank my parents and friends, who have always been there for me with all their love and care that is worth millions**.**

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Data mining strategies have been generally used to mine educated data from therapeutic information bases. In data mining classification is a regulated discovery that can be utilized to configure models depicting critical information classes, where class quality is associated with the development of the classifier. Classification of coronary Heart Disease can be significant for the therapeutic professionals if it is mechanized with the true objective of brisk finding and correct outcome. Medicinal information bases are high volume in nature. On the off chance that the informational index contains repetitive and insignificant qualities, classification may create less exact outcome. In this work, a hybrid clustering with classification model comprising of improved clustering algorithm (Enhanced K-means) and support vector machine classification is proposed. To perform this, various data mining techniques for healthcare are studied and dataset is collected from UCI Dataset repository. This database consists of 76 attributes, but after all the published experiments, only a subset of 14 attributes are selected and used for the machine learning. To evaluate the results accuracy, precision, recall, root means squared error are used as performance parameters.

CHAPTER 1
INTRODUCTION

Heart Diseases remains the greatest cause of deaths for the last two decades. As of late computer technologies built up certain product's to help specialists in making choice of heart disease in the beginning time. Diagnosing the heart disease fundamentally relies upon clinical and obsessive information. Prediction of Heart can help medicinal specialists for anticipating heart disease current status in view of the clinical information of different patients. In biomedical field information mining assumes a basic part for expectation of ailments. For diagnosing, the data which has been given by the patients may incorporate comparative information and interrelated side effects particularly when the patient is experiencing more than one kind of ailments of the comparable class. The doctors are not sufficiently fit to analyse it effectively.

1.1 Introduction

Because of a wide accessibility of colossal amount of data and a need to change over this accessible gigantic measure of data to valuable information requires the utilization of data mining strategies. Data Mining and KDD have turned out to be famous as of late. The prominence of data mining and KDD shouldn't be a shock since the extent of the data accumulations that are accessible are extremely vast to be analyzed physically and even the strategies for programmed information examination in light of established measurements and machine adapting regularly confront issues when handling huge, dynamic information accumulations comprising of complex items.

The plenitude of data, combined with the requirement for capable data investigation apparatuses, has been depicted as an information rich however data poor circumstance. The quickly developing, huge measure of data, gathered and put away in vast and various data storehouses, has far surpassed our human capacity for cognizance without effective instruments. Thus, data gathered in substantial information storehouses progress toward becoming "information tombs"— information documents that are sometimes gone to. Thusly, essential choices are frequently made not just with respect to the data rich information put away in information vaults, yet in addition on a leader's instinct, basically in light of the fact that the chief does not have the instruments to extricate the profitable learning inserted in the immense measures of information. Moreover, consider master framework advances, which commonly depend on clients or space specialists to physically enter information into learning bases. Tragically, this system is inclined to predispositions and blunders, and is to a great degree tedious and expensive. Data mining apparatuses perform information examination and may reveal imperative information designs, contributing incredibly to business techniques, learning bases, and logical and restorative research. The enlarging hole amongst information and data requires an efficient advancement of information mining apparatuses that will transform information tombs into "brilliant chunks" of learning.

The method of extracting information from large set of databases and using it to make pivotal business decisions is termed as data mining [7]. It involves the procedure of examining data from various aspects and encapsulating it into valuable information. It extracts previously unknown and perspicuous information and allows users to examine data out of many dimensions, classifies it and summarizes the relationships that are identified.

Heart illness or cardiovascular infection is the greatest scourge tormenting the world. Similarly as with past scourge-bubonic torment, yellow fever, and smallpox; cardiovascular sickness not just strikes down a critical division of the populace all of a sudden however causes delayed enduring and handicap in a much bigger number. Luckily, look into concentrating on the reason, analysis, treatment, and anticipation of coronary illness is pushing forward quickly. The structure and capacity of the cardiovascular framework (both typical and unusual) and capacity to assess these parameters in the living patient, at times by methods for strategies that require entrance of the skin yet in addition, with expanding exactness, by non-intrusive techniques. At the same time, surprising advancement has been made in counteracting and treating cardiovascular maladies by restorative and surgical means.

1.1.1 Data Mining

Data Mining is defined as obtaining the valuable information from the large amount of data. It can be also defined as mining of knowledge from complex data. Some of the main applications of data mining are discussed below:

- **Predicting Trends and Behaviours -** The main application of data mining is prediction. This includes the process of automation of finding the prediction information in large set of databases. Earlier companies conduct objective/subjective surveys to collect information about their brands and products but now through opinion mining or we can say sentiment analysis companies can come to know about their quality of products and services without that questionnaire of papers. Targeted marketing is an example of prediction. Data mining makes use of data on past mailings to recognize the objectives to augment the arrival on venture for future viewpoints.

- **To discover previously unknown patterns -** Data mining tools can help in scanning through the databases to find out hidden patterns in one go. For an instance, discovering the retail sales data to scan for the unrelated products purchased together. This is a case of pattern discovery. Other case of pattern discovery involves the detection of fraudulent credit card transactions. The benefit of data mining is the automation of already built software and hardware platforms which can be implemented on new systems. The prevailing platforms are upgraded from previous version to new version and the new products are developed. At the point when connected on superior parallel processing systems data mining tools can dissect huge databases in couple of minutes. Faster processing of framework signifies that the users can experiment with more models to have knowledge of complex data. As a results of high speed analyzing huge quantities of data is useful to users. If there are larger databases, it can yield to improved predictions. Various stages of processing are:
  - Selection
  - Pre-processing
  - Transformation
  - Data mining
  - Interpretations/Evaluations.

1.1.2 Need of Data Mining

Information is most essential resources of the any organization or association yet discovering valuable data from the information is a complex undertaking. To discover valuable data from given informational collection we need to apply information mining procedures. Information mining is where we study and research methods for the mining information more successfully and effectively giving more sensible data which gives assistance in choice arranging, for instance, an organization performed mining on its yearly deal and discovered in which month it need to give offers to the customer which expanded general benefit. Today we have part of unstructured information and we need such effective methods which help giving productive investigation of this information. Information mining is a strategy which encourages us in mapping unstructured information to organized information with the assistance of a few systems, for example, Data Cleaning, Data Integration, and Data Transformation. These strategies give us data which helps in different fields of our everyday life for basic leadership in all the business and instructive field, lessens the un-valuable data and give us valid data which is important for the basic leadership in light of the fact that putting away the information is costly procedure.

1.1.3 Data Mining Applications

The data mining applications can be nonexclusive or area particular. The non-specific application is required to be a smart framework that by its own particular can takes certain choices like: selection of information, Determination of information mining procedures, introduction and translation of the outcome. Some non-specific information mining applications can't take its own particular these choices however manage clients for choice of information, choice of information digging strategy and for the understanding of the results. Following is the rundown of regions where data mining is broadly utilized:

- Financial Data Analysis
- Telecommunication Industry
- Biological Data Analysis
- Retail Industry
- Scientific Applications
- Intrusion Detection
- Mining of Clusters

1.1.4 How Data Mining works

The knowledge discovery in databases (KDD) is commonly defined with these stages:

1. Selection
2. Pre-processing
3. Transformation
4. Data mining
5. Interpretation/evaluation.

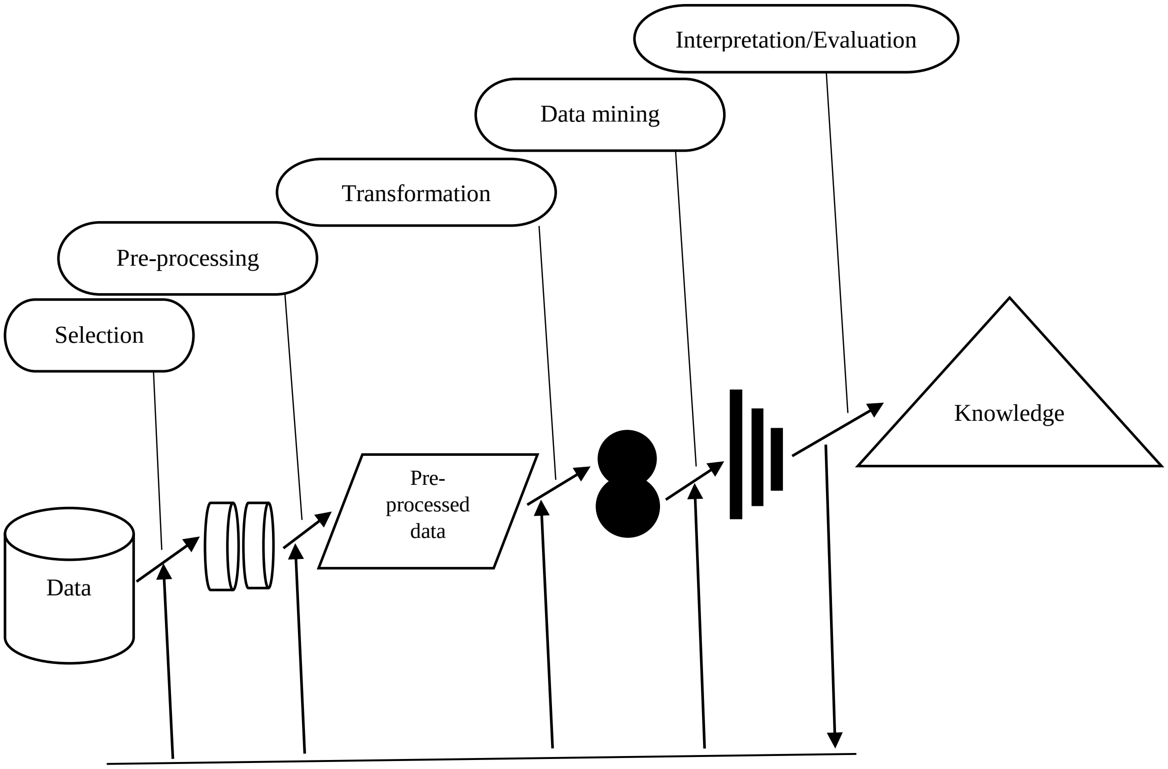

Figure1.1: Knowledge discovery in databases

1.2 Classification.

Classification is a mining utility which helps us to assign the items in a group to target category or classes. Main purpose of the classification is to predict exactly the target class for each case in the dataset. We can use various classifying model for identifying the loan applicants as low, medium, or high credit risks. Classification is the way toward finding a model (or capacity) that depicts and recognizes information classes or ideas, to be ready to utilize the model to foresee the class of articles whose class name is obscure. The inferred display depends on the examination of an arrangement of preparing information (i.e., information questions whose class mark is known) [2].

Some of the applications of classification are:

- Determining whether a particular credit card transaction is fraudulent
- Assessing whether a mortgage application is a good or bad credit risk.
- Diagnosing whether a particular disease is present or absent.

- Identifying whether or not certain financial or personal behaviour indicates a possible terrorist threat [5]

Some of the major tools used for constructing a classification model include Decision Trees, Artificial Neural Networks and Bayesian Classifier.

1.2.1 Decision Tree

Decision tree is defined as “a structure that can be used to divide up a large collection of records into successively smaller sets of records by applying a sequence of simple decision rules. With each successive division, the members of the resulting sets become more and more similar to one another”. A record enters the tree at the root hub. The root hub applies a test to figure out which youngster hub the record will experience next. There are diverse calculations for picking the underlying test, however the objective is dependably the same: to pick the test that best segregates among the objective classes. This procedure is rehashed until the point when the record touches base at a leaf hub. Every one of the records that end up at a given leaf of the tree are grouped a similar way. There is an extraordinary way from the root to each leaf. That way is a declaration of the control used to order the records. Distinctive leaves may make a similar characterization, albeit each leaf makes that arrangement for an alternate reason. For instance, in a tree that arranges leafy foods by shading, the leaves for apple, tomato, and cherry may all anticipate "red," though with fluctuating degrees of certainty since there are probably going to be cases of green apples, yellow tomatoes, and dark fruits too[4].

1.2.2 Artificial Neural Networks.

The motivation for neural systems was the acknowledgment that complex learning frameworks in human brains comprised of firmly interconnected arrangements of neurons. In spite of the fact that a specific neuron might be moderately basic in structure, thick systems of interconnected neurons could perform complex learning undertakings [5]. Artificial Neural Networks (ANNs) are one of a class of parameterized factual models that have pulled in significant consideration lately. The way that ANNs are exceedingly parameterized makes them extremely adaptable, with the goal that they can precisely display moderately little inconsistencies in capacities.

One of the upsides of utilizing neural systems is that they are very powerful concerning boisterous information. Since the system contains numerous hubs (manufactured neurons), with weights doled out to every association, the system can figure out how to function around these uninformative (or even wrong) cases in the dataset. Be that as it may, dissimilar to choice trees, which create natural decides that are reasonable to non-experts, neural systems are moderately obscure to human translation. Additionally, neural systems normally require longer preparing circumstances than choice trees, regularly reaching out into a few hours [1].

1.2.3 Bayesian Classifiers.

Bayesian classifiers are measurable classifiers. They can anticipate class participation probabilities, for example, the likelihood that a given tuple has a place with a specific class. Bayesian order depends on Bayes' hypothesis. Studies looking at order calculations have discovered a straightforward Bayesian classifier known as the innocent Bayesian classifier to be similar in execution with choice tree and chose neural system classifiers. Bayesian classifiers have likewise displayed high exactness and speed when connected to substantial databases. Innocent Bayesian classifiers expect that the impact of a trait esteem on a given class is autonomous of the estimations of alternate qualities. This supposition is called class contingent autonomy. It is made to disentangle the calculations included and, in this sense, is considered "credulous".

Different exact investigations of this classifier in contrast with choice tree and neural system classifiers have observed it to be tantamount in a few areas. In principle, Bayesian classifiers have the base mistake rate in contrast with every single other classifier. Nonetheless, practically speaking this isn't generally the case, attributable to mistakes in the suspicions made for its utilization, for example, class contingent autonomy, and the absence of accessible likelihood information [2].

1.3 Clustering

Unlike classification and prediction, which examine class-named data objects, clustering dissects data objects without counselling a known class name. By and large, the class names are absent in the preparation information essentially in light of the fact that they are not known regardless. Clustering can be utilized to create such names. The items are clustered or gathered in view of the guideline of amplifying the intra-class similitude and limiting the interclass comparability. That is, groups of items are shaped so questions inside a bunch have high similitude in contrast with each other, however are extremely not at all like protests in different groups. Each bunch that is framed can be seen as a class of articles, from which principles can be inferred.

Clustering task includes the following steps:

- Target promoting of a specialty item for a little capitalization business that does not have an extensive advertising spending plan.
- For bookkeeping inspecting purposes, to segmentize money related conduct into kind and suspicious classifications.
- As a measurement decrease device when the dataset has several qualities.
- For quality articulation grouping, where expansive amounts of qualities may show comparative conduct.

Clustering is regularly executed as a preparatory advance in a data mining process, with the subsequent groups being utilized as further contributions to an alternate system downstream, for example, neural systems. Because of the gigantic size of many present day databases, it is frequently useful to apply grouping investigation to begin with, to diminish the look space for the downstream calculations. Some of the major clustering algorithms include hierarchical clustering and K-means clustering. In hierarchical clustering, a tree like cluster structure (dendrogram) is created through recursive partitioning (divisive methods) or combining (agglomerative) of existing clusters. Agglomerative clustering techniques introduce every perception to be its very own little bunch. At that point, in succeeding advances, the two nearest groups are amassed into another joined bunch. Along these lines, the quantity of clusters in the dataset is decreased by one at each progression. In the end, all records are consolidated into a solitary immense bunch. Disruptive Clustering techniques start with every one of the records in one major group, with the most different records being divided from recursively, into a different group, until the point that each record speaks to its own particular bunch [5].

The K-means algorithm is a standout amongst the most ordinarily utilized dividing bunching calculations. The "K" in its name alludes to the way that the calculation searches for a settled number of groups which are characterized as far as nearness of information focuses to each other. The strategy is outlined utilizing two-dimensional charts. By and by the calculation is generally taking care of numerous more than two autonomous factors. This implies rather than directs relating toward two-component vectors (x1, x2), the focuses compare to n-component vectors (x1, x2 . . . xn). The system itself is unaltered [4].Descriptions of K-means and related algorithms gloss over the selection of K. Despite the fact that, K-means clustering algorithm is the least complex and most ordinarily utilized calculation it is extremely delicate to commotion and exception information focuses, on the grounds that few such information can considerably impact the mean esteem [3].

1.4 Heart Diseases

Before characterizing heart illness it is smarter to perceive what a heart is and works performed by our heart. The heart is a standout amongst the most essential organs in our body. Basically a pump, the heart is a muscle made up of four chambers isolated by valves and separated into two parts. Every half contains one chamber called a chamber and one called a ventricle. The atria (plural for chamber) gather blood, and the ventricles contract to drive blood out of the heart. The correct portion of the heart pumps oxygen-poor (blood that has a low measure of oxygen) to the lungs where platelets can acquire more oxygen. At that point, the recently oxygenated blood goes from the lungs into the left chamber and the left ventricle. The left ventricle pumps the recently oxygen-rich blood to the organs and tissues of the body. This oxygen furnishes your body with vitality and is basic to keep your body sound [6]. As indicated by the National Heart, Lung, and Blood organization (2008) Heart sickness is a general name for a wide assortment of infections, issue and conditions that influence the heart and now and then the veins also. Coronary illness is the main enemy of ladies and men in the United States, and more than a million Americans have myocardial areas of localized necrosis. Side effects of coronary illness shift contingent upon the particular sort of coronary illness. A great indication of coronary illness is chest torment. Be that as it may, with a few types of coronary illness, for example, atherosclerosis, there might be no indications in a few people until hazardous confusions create.

Despite the fact that coronary illness can happen in various structures, there is a typical arrangement of center hazard factors that impact whether somebody will eventually be in danger for coronary illness or not. The hazard factors incorporate age, sex, hypertension, diabetes, elevated cholesterol (hypercholesterolemia, hyperlipidemia), weight, and an inactive way of life.

CHAPTER 2
REVIEW OF LITERATURE

This chapter contains a review of research studies as well as relevant and general literature pertaining to the present research problem including various techniques which are being used either individually or in hybridized form.

Dr. Neeraj Bhargava et al.[8] in year 2017 undertook an experiment on application of mining algorithm (simple CART) in order to predict the heart attacks and to compare the best available method of prediction. The predictive accuracy determined by SIMPLE CART (79.90%) algorithm suggests that parameters used are reliable indicators to predict the presence of heart disease.

Anurag bhatt, et al.[9] Used data mining approach to predict and analyze cardiovascular disease. They provided the experimental analysis on the dataset provided by UCI machine learning repository using the WEKA tool. J48 and Naïve Bayes algorithm were used to perform this analysis. The datasets were analyzed using two different methods i.e. first only selected attributes were taken and then all attributes were taken together. Using J48 an accuracy of 82.3% with all attributes and 65.64% with selected attributes was achieved while as Naïve Bayes showed the highest accuracy of 98.64% using all attributes and 93.2% using selected attributes.

Jagdeep singh et al.[10] Proposed a hybrid model for prediction of heart disease in year 2016. In their study they proposed a technique that can generate classification association rules (CAR'S). Using this method they were able to predict which method gives the best prediction or accuracy in predicting heart disease. Their proposed work achieved an accuracy of 99.19% using ibk (nearest neighbor) with aprior associative algorithm.

Kalia Orphanou and Arianna Dagliati [11] used Naïve Bayes classifiers combined with temporal association rules for coronary heart disease diagnosis in year 2016. In their study they compared performance of two classifiers and reached to the conclusion that periodic classifiers are more accurate than baseline classifiers having accuracy of 71% and 68% respectively.

Parisa Naraei et al.[12] in year 2016 used two different algorithms, "Support Vector Machines" and "Multilayer Perceptron Neural Networks" utilized using WEKA tool to classify a heart disease dataset. The accuracy of classification results of Multilayer Perceptron and Support Vector Machines was found to be 84.48% and 80.52% respectively. In their study 8 attributes were chosen and result of their analysis showed the strength of SVMs in classification of medical data.

T.Santhanam and E. P. Ephzibah [13]proposed a system in year 2015 that will help physicians in earlier prognostication of heart disease. The proposed model is based on two computing techniques like genetic algorithms and fuzzy logic. Among all the other classification and prediction models their model provided an accuracy of 86%. With the help of use of genetic algorithms in their study they were able to reduce the number of attributes from thirteen to seven.

K. Prasanna Laxmi and Dr. C.R.K Reddy[14] in year 2015 proposed an efficient technique for heart disease prediction. They proposed a model based on stream associative classification heart disease prediction system. They used SACHDP on various datasets and it showed consistent results when compared with other associative classification techniques. The datasets used for their study were collected from UCI repository which included heart, breast, Pima, hepatitis and lymph disease datasets. As per the results SACHDP outperformed the other traditional associative classification techniques with an average accuracy of 94.94%.

Purushottam et al.[15] in year 2015 designed a system that was able to efficiently mine the rules to predict the risk level of patients based on the given parameters about their health. The experiment was performed on the Cleveland dataset which was collected from UCI repository using 14 attributes. WEKA tool was used for dataset analysis and knowledge extraction based on evolutionary learning (KEEL) was used to find out the classification decision rules. Total correctly classified instances were 86.66%.

M.A.Nishara Banu and B. Gomathy[16] used C4.5 algorithm, Maximal frequent item set algorithm (MAFIA) and K-means clustering in the year 2014 using 13 attributes in the dataset and achieved 89 percent accuracy. In their paper they used K-means to cluster relevant data in a database, MAFIA is then applied for finding maximal frequent patterns in heart disease database and C4.5 is used as a training algorithm to classify frequent patterns into different classes.

Ms.Ishtake and Prof. Sanap S.A.[17] Developed a prediction system for heart diagnosis using decision tree, Neural Network and Naive Bayes techniques using 15 attributes in the year 2013. The purpose of their study was to develop a heart disease prediction system using three data mining classification modelling techniques. All the three models were able to answer the complex queries each with its own strength with respect to the ease of model interpretation, access to detailed information and accuracy.

M. Akhil Jabbar et.al. in year 2013[18] proposed a method to predict heart disease using a technique called lazy associative classification. The proposed technique was used on 7 data sets from UCI repository and one real life dataset of heart patients from Andhra Pradesh. The average accuracy achieved by them was 81.66% using the same technique and in particular on heart diseases the accuracy achieved was 90% when compared with the other traditional techniques such as Naïve Bayes and J48.

Syed Umar Amin et al.[19] Presented a technique for prediction of heart disease using major risk factors. The technique involves two successful data mining tools viz. neural networks and genetic algorithms. The system was developed using Mat lab and predicted the risk of heart disease with an accuracy of 89%. According to results produced they were able to show that genetic algorithm and neural network approach gave better average prediction accuracy than traditional ANN.

Nidhi Bhatla and Kiran Jyoti[20] in year 2012 projected the study of different data mining techniques that can be employed in automated heart disease prediction systems. The analysis of their work proved that neural network with 15 attributes has shown the highest accuracy. On the other hand, Decision tree has also performed well with 99.62% accuracy by using 15 attributes.

Mia Shouman et al.[21] in year 2012 integrated decision tree and k-means clustering for diagnosis of heart disease. They also investigated the various number of centroid selection methods like inlier, outlier, range, random attribute values and random row methods. It was found out that among all these centroid selection methods inlier outperformed all other methods achieving an accuracy of 83.9%.

Chaitrali S. Dangare and Sulabha S. Apte [22] in year 2012 analyzed prediction system for heart disease using three data mining techniques namely Decision tree, Naïve Bayes and Neural Networks. In their paper they added two more attributes (obesity and smoking) to the previously defined 13 attributes and got the accuracy of 99.62%, 90.74% and 100% respectively. They concluded that out of these three prediction models Neural Networks predicts the heart disease with highest accuracy.

Yan Zhang et al. [23] in year 2012 made the use of support vector machine method which is based on statistical learning theory to diagnosis of coronary heart disease. Using a total number of 13 attributes the classification accuracy of three kernel functions was determined and they found out that Rbf kernel function has shown the highest classification accuracy of 88.6% followed by linear and then polynomial kernel function.

Beenish fida et.al.[24] in year 2011 proposed a classifier ensemble method for an efficient heart disease diagnosis. In this study classification is done through homogenous ensemble and final results obtained are optimized using genetic algorithms. They found out that the proposed method performed best on Cleveland dataset with an accuracy of 98.65%.

M. Anbarasi et.al[25] in year 2010 made use of three classifiers such as Decision tree, Classification via Clustering and Naïve Bayes for diagnosis of patients with heart disease. The classifiers were fed with reduced data set of 6 attributes. According to the results they found out that the Decision tree data mining technique outperforms the other two techniques with the accuracy of 99.2% after incorporating feature subset selection but with high model construction time (0.09s). Naïve Bayes performed consistent before and after reduction of attributes with same model construction time. Classification via Clustering performed poor as compared to two other methods.

Asha Rajkumar and Ms. G.Sophia Reena[26] in year 2010 compared the performance of three algorithms viz. Naïve Bayes, Decision tree, K Nearest Neighbor algorithms for heart disease diagnosis and found out that Naïve Bayes has best compact time for processing dataset (609ms) and gives an accuracy of 52.33% as compared to other algorithms.

Sellappan Palaniappan and Rafiah Awang[27] build an intelligent prediction system for heart disease diagnosis using three data mining models (Naïve Bayes, Decision tree and Neural Networks). Naive Bayes gives the highest probability (95%) with 432 supporting cases, followed closely by Decision Tree (94.93%) with 106 supporting cases and Neural Network (93.54%) with 298 supporting cases. In their study they found that intelligent heart disease prediction system was able to answer complex 'What If' queries which conventional decision support system cannot.

Table 2.1: Showing various techniques, year, accuracy and their associated authors used for the prediction of heart diseases.

S.no	Author	Year	Technique used	Accuracy	Attributes
1.	Dr. Neeraj Bhargava et al	2017	Simple CART	79.90	8
2	Anurag bhatt et al.	2017	J48 and Naïve Bayes algorithm	98.64% and 93.2%	14
3	Jagdeep Singh et al	2016	Association and classification algorithms	67.2%,97.31, 97.85%,99.19%,97.85%	13
4	Kalia Orphanou et al.	2016	Naïve Bayes and temporal association rules	71% and 68%	11
5	Parisa Naraei et al.	2016	SVM and Multilayer Perceptron Neural Networks	84.48% and 80.52%	14
6	T.Santhanam and E. P. Ephzibah	2015	Genetic algorithms and fuzzy logic	86%	7
7	K. Prasanna Laxmi and Dr. C.R.K Reddy	2015	Association and classification algorithms	94.94%	13
8	Purushottam et al.	2015	Decision tree	86.66%	14
9	M.A.Nishara Banu et al.	2014	C4.5 algorithm, MAFIA and K means clustering	96%	13
10	Ms.Ishtake , Prof. Sanap S.A.	2013	Naïve Bayes, Decision tree and Neural Networks	94.93%, 95%, 93.54%	15
11	M. Akhil Jabbar et.al.	2013	Lazy associative classification	90%	12
12	Syed Umar Amin et al.	2013	Neural networks and genetic algorithms	89%	12
13	Nidhi Bhatla et al.	2012	Decision tree, Neural Network	99.62%	13

S.no	Author	Year	Technique used	Accuracy	Attributes
14	Mia Shouman et al	2012	Decision tree and K-means clustering	83.9%	14
15	Chaitrali S. Dangare and Sulabha S. Apte	2012	Decision tree, Naïve Bayes and Neural Networks	99.62%, 90.74% and 100%	13
16	Yan Zhang et al.	2012	Support vector machines	88.6%	13
17	Beenish fida et.al	2010	SVM and Genetic Algorithm	98.65%	13
18	M. Anbarasi et.al	2010	Decision tree, Classification via Clustering and Naïve Bayes	99.2% for Decision tree	15
19	Asha Rajkumar and Ms. G.Sophia Reena	2010	Naïve Bayes, Decision tree, K Nearest Neighbor algorithms	52.33%	17
20	Sellappan Palaniappan et al.	2008	Naïve Bayes, Decision tree and Neural Networks	95%,94.93 %,and 93.54%	15

Table 2.2: Showing Total number of attributes with their description used for heart disease prediction.

S. no	**Author**	**Year**	**Total no. and description of attributes**	**Future work**
1.	Dr. Neeraj Bhargava et al	2017	8 (Age, chest pain, rest bp, blood sugar, rest electro, max heart rate, exercise angina, disease).	Using the CART algorithm on dataset we can generate rules that help us to predict correct cause of disease. Incorporating only the related rules can improve the accuracy.
2	Anurag bhatt et al.	2017	14(Patient ID, sex, fasting blood sugar, diagnosis, restecg, exang, slope, thal, rest blood pressure, pain loc, thalach, old peak, age in year, CA)	Comparative analysis and prediction can be performed on clinical outcomes using real patient data acquired from hospitals and medical research institutions.
3	Jagdeep Singh et al	2016	13 (Gender, fasting blood sugar, diagnosis, restecg, exang, slope, thal, rest bp, pain loc, thalach, old peak, age, CA)	The various parameters in this study can be further enhanced such as processing time, resources and memory used.
4	Kalia Orphanou et al.	2016	11(Medicines for blood pressure, Total cholesterol, HDL,LDL, Triglycerides ,Age ,Obesity ,Diet ,Exercise ,Diabetes ,Systolic blood pressure)	Incorporation of complex TARS as features of the classifier can improve its performance.
5	Parisa Naraei et al.	2016	14 (Patient ID, sex, fasting blood sugar, diagnosis, restecg, exang, slope, thal, rest blood pressure, pain loc, thalach, old peak, age in year, CA)	Support vector machines showing its strength in classification of medical data can be used for prediction of other diseases such as intracranial pressure in traumatic brain injured patients
6	T.Santhanam and E. P. Ephzibah	2015	13 (Gender, fasting blood sugar, diagnosis, restecg, exang, slope, thal, rest blood pressure, pain loc, thalach, old peak, age in year, CA)	The time and space complexities can be taken into consideration to improve the overall performance of the proposed system and also the work can be further improved for any data with uncertainty.
7	K. Prasanna Laxmi and Dr. C.R.K Reddy	2015	13 (Gender, fasting blood sugar, diagnosis, restecg, exang, slope, thal, rest blood pressure, pain loc, thalach, old peak, age in year, CA)	SACHDP can perform better if the number of generated rules are reduced to a certain extent.

S. no	**Author**	**Year**	**Total no. and description of attributes**	**Future work**
8	Purushottam et al.	2015	14 (Patient ID, sex, fasting blood sugar, diagnosis, restecg, exang, slope, thal, rest blood pressure, pain loc, thalach, old peak, age in year, CA)	The total number of correctly classified instances could be increased by incorporating more number of rules and making them complex as well.
9	M.A.Nishara Banu et al.	2014	13(Patient ID, sex, fasting blood sugar, diagnosis, restecg, exang, slope, thal, rest blood pressure, pain loc, thalach, old peak, age)	We can make use of different classification and clustering algorithms for efficient and effective prediction of diseases.
10	Ms.Ishtake , Prof. Sanap S.A.	2013	15(Patient ID, Diagnosis, sex, fasting blood sugar, diagnosis, restecg, exang, slope, thal, rest blood pressure, pain loc, thalach, old peak, age in year, CA)	The number of attribute list can be expandedto provide comprehensive diagnosis system. Furthermore the size of data set is relatively small thus a large data set can definitely produce better results.
11	M. Akhil Jabbar et.al.	2013	12(Age, gender, diabetic, BP systolic, BP dialic, height, weight, BMI, Hypertension, rural, urban, disease status)	Using the lazy associative classification technique complex Tar's can be incorporated to definitely improve the performance of the system. Also the number of attributes used can reduced.
12	Syed Umar Amin et al.	2013	12(Sex, age, blood cholesterol, blood pressure, hereditary, smoking, alcohol intake, physical activity, diabetes, diet, obesity, stress)	Using hybrid data mining techniques we could design more accurate clinical decision support system for diagnosis of diseases.
13	Nidhi Bhatla et al.	2012	13(Patient ID, sex, fasting blood sugar, diagnosis, restecg, exang, slope, thal, rest blood pressure, pain loc, thalach, old peak, age in year)	Use of same classifier in different data mining techniques can produce different results. The use of different data mining techniques can be used for efficient and effective prediction of heart disease.
14	Mia Shouman et al	2012	13(Patient ID, sex, fasting blood sugar, diagnosis, restecg, exang, slope, thal, rest blood pressure, pain loc, thalach, old peak, age)	Enhanced k-means clustering method can be used for dimensionality reduction and integrating it with decision tree can enhance its performance.

S. no	**Author**	**Year**	**Total no. and description of attributes**	**Future work**
15	Chaitrali S. Dangare and Sulabha S. Apte	2012	13(Patient ID, sex, fasting blood sugar, diagnosis, restecg, exang, slope, thal, rest blood pressure, pain loc, thalach, old peak, age in year).	The number of attributes used for the study can be reduced thus saving the time required to build the model.
16	Yan Zhang et al.	2012	13(Patient ID, sex, fasting blood sugar, diagnosis, restecg, exang, slope, thal, rest blood pressure, pain loc, thalach, old peak, age in year).	As the experimental results show that SVM can be effectively used to diagnose heart disease. Moreover accuracy of the svm can be enhanced by reducing the number of attributes used and also by extracting feature information only.
17	Beenish fida et.al	2010	13(Patient ID, sex, fasting blood sugar, diagnosis, restecg, exang, slope, thal, rest blood pressure, pain loc, thalach, old peak, age in year).	Genetic algorithms is a best technique for optimization and searching for quality solution thus using genetic algorithms with other classification algorithms and also reducing the number of attributes can improve the accuracy of the system.
18	M. Anbarasi et.al	2010	13(Patient ID, sex, fasting blood sugar, diagnosis, restecg, exang, slope, thal, rest blood pressure, pain loc, thalach, old peak, age in year).	The results can be improved by implementing fuzzy learning models to evaluate the intensity of diseases.
19	Asha Rajkumar and Ms. G.Sophia Reena	2010	17(Patient ID, sex, fasting blood sugar, diagnosis, restecg, exang, slope, thal, rest blood pressure, pain loc, thalach, old peak, age in year, CA, Thaldur, ekgmo, num)	The accuracy of the classifiers can be increased using the various hybrid techniques in case of data mining in the medical field.
20	Sellappan Palaniappan et al.	2008	15(Patient ID, sex, fasting blood sugar, diagnosis, restecg, exang, slope, ca, thal, rest blood pressure, serum cholesterol, thalach, oldpeak, age in year)	IHDPS can be further enhanced and expanded by incorporating other medical attributes than the mentioned ones and also continuous data can be used instead of categorical data.

CHAPTER 3
PROBLEM DEFINITION

Now-a-days individuals need to live extremely rich life so they buckle down keeping in mind the end goal to gain part of cash and live agreeable. Because of this, individuals neglect to take care of themselves which result in the adjustment in their way of life and food habits which prompted hypertension, cholesterol issue at exceptionally youthful age. They don't stress in the event that they are wiped out neither go for their own particular contemplation. Because of these activities, it prompted real issue called heart diseases. As in human body heart is most fundamental body part it might ruin the human well-being framework. In this manner, it is vital to analyze the heart maladies. Because of accessibility of tremendous measure of information, the data can't be recovered effectively, so information mining approaches are executed with a specific end goal to extricate proficient data for the survival of patient or to investigate significant reason for disease. In the existing technique, SVM and Neural networks are used to analyze the patterns from the dataset; and concluding SVM as a best classification technique.

A Hybrid clustering with classification model is proposed technique comprising of enhanced K-means (Clustering) technique with support vector machine (classification) technique which is used to mine the data to extract the useful patterns and to improve the accuracy of the classifier than the existing hybrid technique.

CHAPTER 4
SCOPE OF THE STUDY

The scope of this study is to use both clustering and classification algorithms i.e. clustered data is given to the classification algorithm for evaluating the mining patterns. For clustering most common algorithm K-means is used but it is enhanced in some terms such as dimensionality reduction. The enhanced K-means algorithm is applied for dimensionality reduction to remove outliers and noisy data. This optimized dataset is given as an input to Support Vector Machine classifier to find the useful patterns. According to the existing techniques surveyed, both K-means as well as SVM have been used either separately or with the other combination as well and it was observed that each technique will give better results but the issue that has been focused in this study is to enhance the k-means algorithm and then providing the same clustered data to classification algorithm which is Support vector machines, for better classification as well as improving the accuracy of the classifier.

CHAPTER 5

OBJECTIVES OF THE STUDY

1. To study and analyse various healthcare data mining techniques.
2. To propose a hybrid clustering with classification model comprising of enhanced k-means and Support vector machines.
3. To check the presence or absence of heart malady with the help of proposed clustering with classification model using only 14 selected attributes.
4. To enhance k-means algorithm and examine its performance as compared to simple k-means.
5. To evaluate and analyse the performance on the basis of accuracy, precision, recall, root means squared error and to compare the performance of the proposed technique with the existing hybrid techniques.

CHAPTER 6

PROPOSED RESEARCH METHODOLOGY

1. **Dataset collection:** Data has been collected from UCI dataset repository. This database consists of 76 attributes, but only a subset of 14 attributes are selected and used for this study.

Table 6.1: Showing the description of attributes to be used

Age	Sex	Chest pain	Resting blood pressure	Serum cholesterol	Fasting blood sugar	Resting electrocardiograph results
Max. heart rate achieved	Exercise induced angina	Old peak	Slope	Major vessels colored by fluoroscopy	Type of defect	Angiographic disease status

2. **Pre-processing:** The collected raw data is then pre-processed and formatted into a well-defined arff structure or comma delimited structure. If some missing values are there, it will handle all the missing values by either replacing those values or by removing.
3. **Proposed Technique:** The proposed technique comprises of clustering with classification i.e. clustered data is given to the classification for the evaluating the mining patterns. For the clustering of data, most commonly used algorithm is K-means but K-means algorithm have many limitations like:
   - K-means algorithm assumes that value of k (number of clusters) is known in advance which is not necessarily true in real-world applications.
   - The K-means algorithm is sensitive to initial centres selection.

   The enhanced K-means algorithm is applied for dimensionality reduction to remove outliers and noisy data. This optimized dataset is given as an input to Support Vector Machine classifier to find the useful patterns. According to the existing techniques surveyed, when K-means is combined SVM it will give better results but the issue is the problem of selecting the initial centroids in K-means; this problem is mainly focused in the proposed technique.

1. Data is partitioned into k equal parts. Then the arithmetic mean of each part is taken as the centroid point.
2. K-means is applied on the input dataset by finding the Euclidean distance of each data point from the centroid and clusters are defined. If the distance of centroid of the present nearest cluster is less than or equal to the previous distance, then the data point remains in that cluster and there is no need to find its distance from other cluster centroids.
3. Apply clustering on the dataset for dimensionality reduction and then classify that reduced dataset using Support Vector Machine classifier.

The initial centroids are randomly selected in case of simple K-means algorithm but it is not so in proposed algorithm. The proposed work is to select the initial centroids by partitioning the data into k equal parts and then the arithmetic mean of each part is taken as the centroid point. The efficiency and accuracy of enhanced K-means algorithm is more than simple K-means.

4. Evaluate the performance of the proposed technique on the basis of accuracy, precision, recall, Root mean squared error.

Consider an example where we have to make two clusters on the basis of two attributes which are height and weight. The values for the data sample are as

Height	Weight
185	72
170	56
168	60
179	68

Upon gathering the input data two clusters K1 and K2 are made. The first two observations are taken as initial centroids. Now calculate Euclidean distance from each of the clusters which is calculated using formula $\sqrt{}X_H - H_1) + (X_W - W_1)^2$.

Where X_H=observation value of variable height

H1= centroid value of cluster 1 for variable height

X_W= observation value of variable weight

W1= centroid value of cluster 1 for variable weight.

Upon calculating the Euclidean distance of each observation and the continuously updating the initial centroids by taking the mean of previous and new observation, following cluster assignments are obtained:

Euclidean distance from cluster 1	Euclidean distance from cluster 2	Assignment
185	72	1
170	56	2
168	60	2
179	68	1

As per the assignments in the above table it is clear that which observation is in which cluster. The observations are assigned accordingly to the nearest cluster to which the Euclidean distance is minimum and this is how clusters are made in k-means algorithm. After the clusters are made, the svm algorithm is used to classify the clusters. Likewise the above data can be used to classify the data to determine the short and tall categories. The two classes will be made with the class labels ‘short’ and ‘tall’. The svm classifier will be trained and after that it will classify the data on that basis. This approach can be used when the data set is often too large and impossible to classify using the traditional approaches.

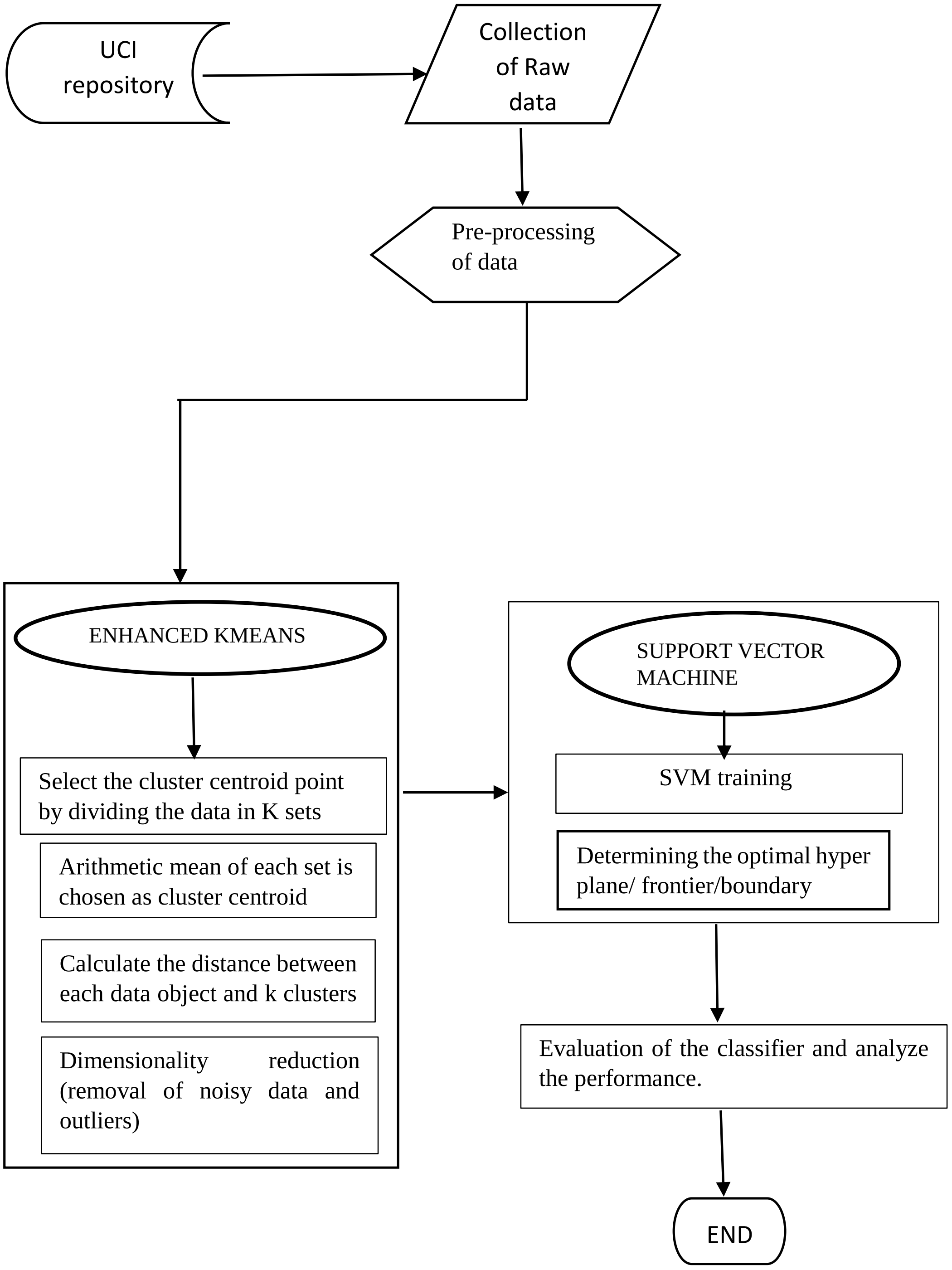
UCI repository
Collection of Raw data
Pre-processing of data
ENHANCED KMEANS
Select the cluster centroid point by dividing the data in K sets
Arithmetic mean of each set is chosen as cluster centroid
Calculate the distance between each data object and k clusters
Dimensionality reduction (removal of noisy data and outliers)
SUPPORT VECTOR MACHINE
SVM training
Determining the optimal hyper plane/ frontier/boundary
Evaluation of the classifier and analyze the performance.
END

Figure6.1: Flow chart of the proposed methodology

CHAPTER 7
EXPECTED OUTCOMES

Heart illness or cardiovascular infection is the best scourge harassing the world. Coronary illness portrays a scope of conditions that influence your heart. Ailments under the coronary illness umbrella incorporate vein ailments, for example, coronary corridor infection; heart musicality issues (arrhythmias); and heart abandons by birth (inherent heart absconds). Cardiovascular illness by and large alludes to conditions that include limited or blocked veins that can prompt a heart assault, chest torment (angina) or stroke. Other heart conditions, for example, those that influence your heart's muscle, valves or musicality, additionally are considered types of coronary illness. The Proposed system is developed to extract information about presence or absence of heart disease using various symptoms. Expected outcome of this system is to provide information about heart disease presence or absence by using enhanced k-means with support vector machines. The evaluation of the classifier is done on various parameters including Precision, Recall, f measure & accuracy. The proposed system is expected to provide better results than the previous systems which include simple k-means and support vector machines individually or using any other hybrid approach.

CHAPTER 8
SUMMARY AND CONCLUSIONS

In this work, a hybrid clustering with classification model comprising of enhanced k-means clustering algorithm and support vector machine classification technique is proposed. The purpose of this study is to study and analyze the various healthcare data mining techniques. Heart patient's raw data is collected from UCI Dataset repository to perform clustering and classification. For clustering hybrid k-means algorithm will be used and further for classification it is given to the support vector machines. The proposed work is to select the initial centroids by partitioning the data into k equal parts and then the arithmetic mean of each part is taken as the centroid point. The efficiency and accuracy of hybrid K-means algorithm is more than simple K-means so the proposed method is expected to perform better as compared to other methods or hybrid approaches that have been already used. Results are evaluated on the basis of accuracy, precision, recall, Root mean squared error.

CHAPTER 9
REFERENCES

[1] Hand, D., Mannila, H. and Smyth, P. (2001). Principles of Data Mining. The MIT Press, Massachusetts Institute of Technology, Massachusetts.

[2] Han, J. and Kamber, M. (2006). Data Mining: Concepts and Techniques. Second Edition, Morgan Kaufmann Publishers, San Francisco.

[3] Kantardzic, M. (2003). Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons, New Jersey.

[4] Berry J.A. Michael and Linoff S. Gordon (2004). Data Mining Techniques for Marketing, Sales, and Customer Relationship Management. Second Edition. Wiley Publishing, Inc., Indianapolis, Indiana.

[5] Larose T. Daniel (2005). Discovering Knowledge in Data: An Introduction to Data Mining. John Wiley & Sons, Inc., Hoboken, New Jersey.

[6] Benjamin, M. (2006). Introduction to Heart Disease. Available at http://www.mentalhelp.net/poc/view_doc.php?type=doc&id=4496 (Accessed 10 January 2011).

[7] Bora, Shital P. "Data mining and ware housing." Electronics Computer Technology (ICECT), 2011 3rd International Conference on, E-ISBN:978-1-4244-8679-3, Vol. 1. IEEE, 2011.

[8] N. Bhargava and S. Dayma, “An Approach for Classification using Simple CART Algorithm in Weka,” pp. 212–216, 2017.

[9] A. Bhatt, S. K. Dubey, and A. K. Bhatt, “Data Mining Approach to Predict and Analyze the Cardiovascular Disease,” 2017.

[10] J. Singh, A. Kamra, and H. Singh, “Classification,” 2016.

[11] K. Orphanou, A. Dagliati, L. Sacchi, A. Stassopoulou, E. Keravnou, and R. Bellazzi, “Combining Naive Bayes Classifiers with Temporal Association Rules for Coronary Heart Disease Diagnosis,” pp. 81–92, 2016.

[12] P. Naraei, V. Street, V. Street, and V. Street, “Application of Multilayer Perceptron Neural Networks and Support Vector Machines in Classification of Healthcare Data,” no. December, pp. 848–852, 2016.

[13] T. Santhanam and E. P. Ephzibah, “Heart Disease Prediction Using Hybrid Genetic Fuzzy Model,” vol. 8, no. May, pp. 797–803, 2015.

[14] K. P. Lakshmi, “Fast Rule-Based Heart Disease Prediction using Associative Classification Mining,” 2015.

[15] R. Sharma, “Efficient Heart Disease Prediction System using Decision Tree,” 2015.

[16] M. A. N. Banu and A. D. Preprocessing, “2014 International Conference on Intelligent Computing Applications Disease Forecasting System Using Data Mining Methods,” pp. 1–4, 2014.

[17] “‘ Intelligent Heart Disease Prediction System Using Data Mining Techniques ,’” no. April, pp. 94–101, 2013.

[18] M. A. Jabbar, “Heart Disease Prediction using Lazy Associative Classification,” pp. 40–46, 2013.

[19] S. U. Amin, K. Agarwal, and R. Beg, “Genetic Neural Network Based Data Mining in Prediction of Heart Disease Using Risk Factors,” no. Ict, pp. 1227–1231, 2013.

[20] K. Jyoti, “An Analysis of Heart Disease Prediction using Different Data Mining Techniques,” vol. 1, no. 8, pp. 1–4, 2012.

[21] M. Shouman, T. Turner, and R. Stocker, “Integrating Decision Tree and K-Means Clustering with Different Initial Centroid Selection Methods in the Diagnosis of Heart Disease Patients.”

[22] C. S. Dangare and M. E. Cse, “Improved Study of Heart Disease Prediction System using Data Mining Classification Techniques,” vol. 47, no. 10, pp. 44–48, 2012.

[23] Y. Zhang, F. Liu, D. Li, and X. Zhou, “Studies on application of Support Vector Machine in diagnose of coronary heart disease,” 2012.

[24] B. Fida, M. Nazir, N. Naveed, and S. Akram, “Heart Disease Classification Ensemble Optimization Using Genetic Algorithm,” pp. 19–24, 2011.

[25] “Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm Enhanced Prediction of Heart Disease with Feature Subset Selection using Genetic Algorithm,” no. October 2010, 2015.

[26] A. Rajkumar and G. S. Reena, “Algorithm,” vol. 10, no. 10, pp. 38–43, 2010.

[27] S. Palaniappan and R. Awang, “Intelligent Heart Disease Prediction System Using Data Mining Techniques,” pp. 108–115, 2008.