CHAPTER-I

Introduction
CHAPTER-I

INTRODUCTION

1.1 Introduction

Data is growing at a phenomenal rate today and the users expect more sophisticated information from this data. There is need for new techniques and tools that can automatically generate useful information and knowledge from large volumes of data.

Data mining refers to extracting or mining interesting knowledge from large amounts of data. It provides a means of extracting previously unknown, predictive information from the base of accessible data in data warehouses. Data mining is the task of discovering interesting patterns from large amounts of data where the data can be stored in databases, data warehouses, or other information repositories. It is also popularly referred to as Knowledge Discovery in Databases (KDD). Since the inception data mining has achieved marvelous success so various problems emerged during the tenure have been solved by data mining techniques. Data mining technology is an application oriented technology and is having extensive applications in various fields. It also evaluates, integrates and reasons to guide the solution of practical problems and find the relationships between events. Furthermore predictions of further activities can be easily made by using the prevailing data.

1.2 Origin of Data Mining

Since ancient times, our ancestors have been searching for useful information from data by hand. However, with the rapidly increasing, in volume of data in modern times, more automatic and effective mining approaches are required. Early methods such as Bayes' theorem in the 1700s and regression analysis in the 1800s
were some of the first techniques used to identify patterns in data. After the 1900s, with the proliferation, ubiquity, and continuously developing power of computer technology, data collection and data storage were remarkably enlarged. As Datasets have grown in size and complexity, direct hands-on data analysis has increasingly been augmented with indirect, automatic data processing. This has been aided by other discoveries in computer science, such as neural networks, clustering genetic algorithms. Data mining involves an integration of techniques from multiple disciplines such as database technology, statistics, machine learning, high-performance computing, pattern recognition, neural networks, data visualization, information retrieval, etc. Figure 1.1 shows Data Mining Technology.

Figure 1.1: Data Mining Technology
Bioinformatics is an emerging field that has its roots in molecular biology, mathematics and computer science. It deals with generation, management and analysis of the biological data, which is obtained from various experiments and techniques, often resulting in large data. The analysis of such enormous biological data requires use of sophisticated algorithms, which can process the data and help in visualizing the data and extract information from it. This led to the evolution of Bioinformatics, an interdisciplinary field involving both biologist and computer scientists.

Data mining is field which is being applied in all domains now a day.

- Signal processing
- Social media analytics
- Medical science
- Government domain
- Finance analysis (Stock prediction, customer behavior prediction etc.)

1.3 Data Mining Architecture

Figure 1.2 shows the architecture of a typical data mining system may have the following major components database, data warehouse, or other information repository; their server which is responsible for fetching the relevant data based on the user’s data mining request; knowledge base which is used to guide the search, or evaluate the interestingness of resulting patterns; data mining engine which consists of a set of functional modules for tasks; pattern evaluation module which interacts with the data mining modules so as to focus the search towards interesting patterns; and graphical user interface which communicates between users and the data mining system, allowing the user interaction with system. The major components of any data
mining system are data source, data warehouse server, data mining engine, pattern evaluation module, graphical user interface and knowledge base.

a) Data Sources:

Database, data warehouse, World Wide Web (www), text files and other documents are the actual sources of data. Organizations usually store data in databases or data warehouses. Data warehouses may contain one or more databases, text files, spreadsheets or other kinds of information repositories. Sometimes, data may reside even in plain text files or spreadsheets. World Wide Web or the Internet is another big source of data.

Different Processes

The data needs to be cleaned, integrated and selected before passing it to the database or data warehouse server. As the data is from different sources and in different formats, it cannot be used directly for the data mining process because the data might not be complete and reliable. So, first data needs to be cleaned and integrated. Again, more data than required will be collected from different data sources and only the data of interest needs to be selected and passed to the server. These processes are not as simple as we think. A number of techniques may be performed on the data as part of cleaning, integration and selection.

b) Database or Data Warehouse Server:

The database or data warehouse server contains the actual data that is ready to be processed. Hence, the server is responsible for retrieving the relevant data based on the data mining request of the user.
c) Data Mining Engine:

The data mining engine is the core component of any data mining system. It consists of a number of modules for performing data mining tasks including association, classification, characterization, clustering, prediction, time-series analysis etc.

d) Pattern Evaluation Modules:

The pattern evaluation module is mainly responsible for the measure of interestingness of the pattern by using a threshold value. It interacts with the data mining engine to focus the search towards interesting patterns.

e) Graphical User Interface:

The graphical user interface module communicates between the user and the data mining system. This module helps the user use the system easily and efficiently without knowing the real complexity behind the process. When the user specifies a query or a task, this module interacts with the data mining system and displays the result in an easily understandable manner.

f) Knowledge Base:

The knowledge base is helpful in the whole data mining process. It might be useful for guiding the search or evaluating the interestingness of the result patterns. The knowledge base might even contain user beliefs and data from user experiences that can be useful in the process of data mining. The data mining engine might get inputs from the knowledge base to make the result more accurate and reliable. The pattern evaluation module interacts with the knowledge base on a regular basis to get inputs and also to update it.
1.4 Data Mining Techniques

Data mining tools use sophisticated, automated algorithms to discover hidden patterns, correlations, and relationships among organizational data. These tools are used to predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. Figure 1.3 shows various data mining techniques.
1.4.1 Descriptive Modeling

The descriptive function deals with the general properties of data in the database.

- Mining of Clusters
- Mining of Associations

Mining of Clusters:

Cluster refers to a group of similar kind of objects. Cluster analysis refers to forming group of objects that are very similar to each other but are highly different from the objects in other clusters.
Mining of Association:

Associations are used in retail sales to identify patterns that are frequently purchased together. This process refers to the process of uncovering the relationship among data and determining association rules.

1.4.2 Predictive Modeling

The term Predictive Data Mining is usually applied to identify data mining projects with the goal to identify a statistical or neural network model or set of models that can be used to predict some response of interest. Each model is made up of a number of predictors, which are variables that are likely to influence future results.

a) Classification and Prediction process:

There are two forms of data analysis that can be used for extracting models describing important classes or to predict future data trends. These two forms are as follows:

- Classification
- Prediction

b) Classification:

Classification is the problem of identifying which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known. The individual observations are analyzed into a set of quantifiable properties, known as various explanatory variables, features, etc. These properties may variously be categorical (e.g. “A”, “B”, “AB” or “O”, for blood type), ordinal (e.g. “large”,...
“medium” or “small”), integer-valued (e.g. the number of occurrences of a particular word in an email) or real-valued (e.g. a measurement of blood pressure).

c) Binary and Multiclass Classification

Classification can be thought of as two separate problems - binary classification and multiclass classification. In binary classification, a better understood task, only two classes are involved, whereas in multiclass classification involves, assigning an object to one of several classes. Since many classification methods have been developed specifically for binary classification, multiclass classification often requires the combined use of multiple binary classifiers.

Classification models predict categorical class labels; and prediction models predict continuous valued functions.

d) Prediction:

Prediction task predicts the possible values of missing or future data. Prediction involves developing a model based on the available data and this model is used in predicting future values of a new data set of interest. For example, a model can predict the income of an employee based on education, experience and other demographic factors like place of stay, gender etc. Also prediction analysis is used in different areas including medical diagnosis, fraud detection etc.

1.5 Data Classification Process

The Data Classification process includes two steps:

- Building the Classifier or Model
- Using Classifier for Classification
Building the Classifier or Model:

- This step is the learning step or the learning phase.
- In this step the classification algorithms build the classifier.
- The classifier is built from the training set made up of database tuples and their associated class labels.
- Each tuple that constitutes the training set is referred to as a category or class. These tuples can also be referred to as sample, object or data points.

In this step, the classifier is used for classification. Here the test data is used to estimate the accuracy of classification rules. The classification rules can be applied to the new data tuples if the accuracy is considered acceptable.

Classification and Prediction Issues:

The major issue is preparing the data for Classification and Prediction.

Preparing the data involves the following activities:

- Data Cleaning - Data cleaning involves removing the noise and treatment of missing values. The noise is removed by applying smoothing techniques and the problem of missing values is solved by replacing a missing value with most commonly occurring value for that attribute.
- Relevance Analysis - Database may also have the irrelevant attributes. Correlation analysis is used to know whether any two given attributes are related.
- Data Transformation and reduction - The data can be transformed by any of the following methods.
- Normalization-The data is transformed using normalization. Normalization involves scaling all values for given attribute in order to make them fall
within a small specified range. Normalization is used when in the learning step, the neural networks or the methods involving measurements are used.

- Generalization - The data can also be transformed by generalizing it to the higher concept. For this purpose we can use the concept hierarchies.

1.5.1 Comparison of Classification and Prediction Methods

Here are the criteria for comparing the methods of Classification and Prediction.

- Accuracy - Accuracy of classifier refers to the ability of classifier. It predicts the class label correctly and the accuracy of the predictor refers to how well a given predictor can guess the value of predicted attribute for a new data.

- Speed - This refers to the computational cost in generating and using the classifier or predictor.

- Robustness - It refers to the ability of classifier or predictor to make correct predictions from given noisy data.

- Scalability - Scalability refers to the ability to construct the classifier or predictor efficiently; given large amount of data.

- Interpretability - It refers to what extent the classifier or predictor understands.

There are many algorithms which are used for classification in data mining shown above. Following are some classification algorithms.
1.6 Classification Algorithms

1.6.1 Neural Networks

Artificial neural networks are a system loosely modeled on the human brain. Neural networks are an attempt to simulate within specialized hardware or software, the multiple layers of simple processing elements called neurons. Each neuron is linked to certain of its neighbors with varying coefficients of connectivity that represent the strengths of these connections. Learning is accomplished by adjusting these strengths to cause the overall network to output appropriate results. Artificial neurons are much simpler than the biological neuron. Neural networks are used for feature extraction, association and optimization, function fitting and modeling. Neural networks can be divided into two major categories:

- Supervised Neural Networks
- Unsupervised Neural Networks

Supervised Neural Networks consist of input patterns associated with known output patterns, while unsupervised Neural Networks have structures found in the input patterns.

1.6.2 Decision Tree

A decision tree is a model that is both predictive and descriptive. Decision trees are commonly used for classification but can also be used for regression analysis. Decision trees are advantageous tools for making corporate or financial decisions where a lot of complex information has to be considered. Decision trees provide a functional framework in which alternate decisions and the ramifications of making those decisions can be laid down and evaluated. Decision trees also help in forming a balanced.
1.6.3 Bayesian Network

A Bayesian network is a representation of a joint probability distribution of a set of random variables with a possible mutual causal relationship. The network consists of nodes representing the random variables, edges between pairs of nodes representing the causal relationship of these nodes, and a conditional probability distribution in each of the nodes. The main objective of the method is to model the posterior conditional probability distribution of outcome (often causal) variable(s) after observing new evidence. Bayesian networks may be constructed either manually with knowledge of the underlying domain, or automatically from a large dataset by appropriate software.

A Bayesian network, Bayes network, belief network, Bayes model or probabilistic directed acyclic graphical model is a probabilistic graphical model (a type of statistical model) that represents a set of random variables and their conditional dependencies via a Directed Acyclic Graph (DAG).

1.6.4 Genetic Algorithms

Genetic algorithms refer to simulated evolutionary systems that dictate how populations should be formed, evaluated and modified. One of a variety of algorithms known as optimization techniques genetic algorithms are in their infancy and more experience with them is required before a mine-related use can be proposed.

1.6.5 Artificial Neural Network

Neural Networks are analytic techniques modeled after the (hypothesized) processes of learning in the cognitive system and the neurological functions of the brain and capable of predicting new observations (on specific variables) from other observations (on the same or other variables) after executing a process of so called
learning from existing data. Neural Networks is one of the Data Mining techniques. The first step is to design a specific network architecture (that includes a specific number of “layers” each consisting of a certain number of “neurons”). Network is then subjected to the process of “training”. In that phase, neurons apply an iterative process to the number of inputs to adjust the weights of the network in order to optimally predict the sample data on which the “training” is performed. After the phase of learning from an existing data set, the new network is ready and it can then be used to generate predictions. The resulting “network” developed in the process of “learning” represents a pattern detected in the data.

1.6.6 Nearest Neighbor Classifier

The k-Nearest Neighbor’s algorithm (k-NN) is a method for classifying objects based on closest training examples in the feature space. K-NN is a type of instance-based learning, or lazy learning. It can also be used for regression. The k-nearest neighbor algorithm is amongst the simplest of all machine-learning algorithms. The space is partitioned into regions by locations and labels of the training samples. A point in the space is assigned to the class c if it is the most frequent class label among the k nearest training samples. Usually Euclidean distance is used as the distance metric; however this will only work with numerical values. In cases such as text classification another metric, such as the overlap metric (or Hamming distance) can be used.

1.6.7 Naives Bayesian Classifier

Bayesian classification is based on Bayes’ Theorem. Bayesian classifiers are the statistical classifiers. Bayesian classifiers can predict class membership probabilities such as the probability that a given tuple belongs to a particular class.
Baye's Theorem:

Bayes' Theorem is named after Thomas Bayes. There are two types of probabilities.

- Posterior Probability \( P(H/X) \)
- Prior Probability \( P(H) \)

where \( X \) is data tuple and \( H \) is some hypothesis.

According to Bayes' Theorem,

\[
P(H/X) = \frac{P(X/H)P(H)}{P(X)}\]

1.6.8 Apriori principle

If an item set is frequent, then all of its subsets must also be frequent, or if an item set is infrequent then all its supersets must also be infrequent. Apriori principle holds due to the following property of the support measure. Apriori algorithm is a classical algorithm in data mining. It is used for mining frequent item sets and relevant association rules. It is devised to operate on a database containing a lot of transactions, for instance, items brought by customers in a store.

1.7 Machine Learning

Machine learning, computational learning theory, and similar terms are often used in the context of Data Mining, to denote the application of generic model-fitting or classification algorithms for predictive data mining. Unlike traditional statistical data analysis, which is usually concerned with the estimation of population parameters by statistical inference, the emphasis in data mining (and machine learning) is usually on the accuracy of prediction (predicted classification), regardless of whether or not the “models” or techniques that are used to generate the prediction is interpretable or open to simple explanation. Good examples of this type of
technique often applied to predictive data mining are neural networks or meta-learning techniques such as boosting, etc. These methods usually involve the fitting of very complex “generic” models that are not related to any reasoning or theoretical understanding of underlying causal processes; instead, these techniques can be shown to generate accurate predictions or classification in cross validation samples.

1.7.1 Feature Selection

One of the preliminary stage in predictive data mining, when the data set includes more variables than could be included (or would be efficient to include) in the actual model building phase (or even in initial exploratory operations), is to select predictors from a large list of candidates. For example, when data are collected via automated (computerized) methods, it is not uncommon that measurements are recorded for thousands or hundreds of thousands (or more) of predictors. The standard analytic methods for predictive data mining, such as neural network analyses, classification and regression trees, generalized linear models, or general linear models become impractical when the number of predictors exceed more than a few hundred variables.

Feature selection selects a subset of predictors from a large list of candidate predictors without assuming that the relationships between the predictors and the dependent or outcome variables of interest are linear, or even monotone. Therefore, this is used as a pre-processor for predictive data mining, to select manageable sets of predictors that are likely related to the dependent (outcome) variables of interest, for further analyses with any of the other methods for regression and classification.
1.7.2 Optimization Techniques

Optimization is used to optimize the input as effective and functional as possible. The technology optimization is a set of methods and techniques for the design and use of technical systems as fully as possible within the parameters. Even though the clustering algorithms are simple and effective, they are sensitive to initialization and is easily trapped in local optima. There are many optimization algorithms to overcome the drawbacks of clustering techniques.

1.8 Medical Data Mining Framework

Nowadays most of the research area are working on data mining techniques in medical data.

Knowledge discovery and data mining have found numerous applications in business and scientific domain. Valuable knowledge can be discovered from application of data mining techniques in healthcare system. In this work briefly examine the potential use of classification based data mining techniques such as rule based, decision tree, machine learning algorithms like Support Vector Machines, Principle Component Analysis etc., Rough Set Theory and Fuzzy logic. In particular consider a case study using classification techniques on a medical data set of Heart disease, cancer patients.
Figure 1.4: Medical Data Mining Framework

Figure 1.4 shows different Techniques for Medical data mining. In terms of prediction and decision making, Data mining techniques have substantial expansion in medical industry with respect to various diseases like heart disease, cancer.

Clinical databases have accumulated large quantities of information about patients and their medical diagnosis reports which describe their condition. Relationships and patterns within this data could provide new medical knowledge. Many methodologies have been developed and applied to discover this hidden knowledge. In this research, the techniques of data mining were used to search for relationships and multi dimensions in a large medical database.

It analyzed few of the above data mining techniques on the medical diseases and related complications data. This work depends on applying different mining techniques on medical reports of different patients with different kinds of medical complications. Researchers from different areas are continuously trying to apply these data mining techniques to medical data. Health care industry produces enormous quantity of data that clutches complex information relating to patients and their
medical conditions. Data mining is gaining popularity in different research areas due to its infinite applications and methodologies to mine the information in correct manner. Data mining techniques have the capabilities to discover hidden patterns or relationships among the objects in the medical data. In last decade, there has been increase in usage of data mining techniques on medical data for determining useful trends or patterns that are used in analysis and decision making. Data mining has an infinite potential to utilize healthcare data more efficiently and effectively to predict different kind of disease. This work investigates various Data Mining techniques such as classification, clustering, association and also highlights related work to analyze and predict human disease.

1.8.1 Research Frame Work

![Figure 1.5: Research Frame Work Flow](image)

In this system, pre-processing is carried out on the medical dataset. As a next step, an efficient clustering or classification model has been built and the model has been optimized using evolutionary algorithms. Figure 1.5 shows Research frame work flow.
The overall process of the proposed framework is divided into three steps:

1) Efficient Feature Selection.

2) Classification using Hybrid kernel based ISVM.

3) Optimization Technique based on EGWO.

Here, for experimentation, the dataset given in the UCI machine learning repository heart, cancer dataset will be subjected to analyze the performance of the proposed Methodology.

1.9 Applications of Data Mining

Data Mining is a young discipline with wide and diverse applications; there is still a non-trivial gap between general principles of data mining and domain specific, effective data mining tools for particular applications. Various fields of applications are listed below:

- Data mining applications in sales/ marketing
- Data mining applications in banking / finance
- Data mining applications in Health Care and Insurance
- Data Mining for the Retail Industry
- Data mining for the Telecommunications industry
- Data Mining Application in Higher Education
- Data mining for instruction Detection
- Data mining for Biological Data Analysis
- Data mining for Certain Scientific Applications
• **Data mining applications in sales/marketing**

  Data mining is the process of extracting unknown patterns from database which help in planning, organizing, managing and launching new market in a cost effective way. Data mining plays an important role in Market Basket. It gives information relevant to item sets that are purchased together, their sequence and when they were bought. This information helps business encouragement and to make it most profitable.

• **Data mining applications in banking/finance:**

  There are numerous fields in which data mining can be used like in financial and banking sector for credit analysis, fraudulent transactions, customer segmentation and profitability, optimizing stocks portfolios, predicting payment default, ranking investments, marketing, high risk loan applicants, cash management and forecasting operations and most profitable credit card customers and cross selling.

• **Data mining applications in Health Care and Insurance:**

  Insurance industry growth is completely depends on the ability of transforming data into information regarding customers, competitors and its market. The insurance industries have implemented the Data Mining successfully and have achieved tremendous competitive advantages. The data mining applications in insurance industry can be used in the form that, data mining is applied in claims analysis such as identifying the medical procedures which are claimed together. Data mining enables to forecasts the potential customers who will buy new schemes. This data mining also proactive insurance companies to detect risky customer’s behavior patterns. Data mining also helps in detecting fraudulent behavior.
• **Data Mining for the Retail Industry:**

Retail industry assemble huge amount of data related to sales and customer history of shopping. Retail data mining helps in analyzing client behavior, client patterns of shopping and trends which increases the quality of client service, enhance things consumption ratios, design more effective goods transportations and distribution policies achieve better customer retention and satisfaction and to minimize the cost of business.

• **Data mining for the Telecommunications industry:**

Telecommunication industries generally generate and store large amount of high quality data, having a very huge customer base, and operate in rapidly changing and highly competitive environment. Telecommunication companies use data mining to enhance their marketing efforts to detect fraud and to betterment of their telecommunication networks.

• **Data Mining Application in Higher Education:**

Data mining can be effectively used to address students and alumni challenges. Data mining facilitate organizations to use their current reporting capabilities to uncover and understand hidden patterns in huge databases. These patterns are then built into data mining models and used to predict individual behavior accurately. As a result of their insight, institutions are able to allocate resources and staff efficiently. This data mining can provide an entity the information necessary to take action before a student drops out, or to efficiently allocate resource with an accurate estimate of how many students will take a particular course.
• **Data mining for instruction Detection:**

Instructions are the set of actions that threatens the availability and integrity of a network resource. Network instruction detection has been considered to be one of the most promising method for defending complex and dynamic intrusion behaviors. Intrusion detection techniques using data mining have attracted more and more interests in recent years. Data mining techniques used for intrusion detection are frequent modalities for mining, classification, clustering and mining data streams etc.

• **Data mining for Biological Data Analysis:**

Data mining has also made significant contribution to biological data analysis like genomics, proteomics, functional Genomics and biomedical research. It helps in analysis by semantic integration of heterogeneous, distributed genomic and proteomic databases; association and path analysis, visualization tools in genetic data analysis and more.

• **Data mining for Certain Scientific Applications:**

It also helps in the analysis of large amounts of data from domains such as geosciences, astronomy, etc. Other scientific applications such as climate and ecosystem modeling, chemical engineering, and fluid dynamics which constantly generate large amounts of data are also domains that benefit quite a lot from data mining.

1.10 Challenges in Medical Data Mining

The various issues in medical data mining are heterogeneity of medical data such as volume and complexity of the medical data, the physician’s interpretation of patient’s data. Some of the ethical issues are data ownership, lawsuits dealing with the details of the patients and privacy, security of the human data. The medical data
are heterogeneous. It may be collected from various sources such as images, interviews from physician and patients. This heterogeneity requires high capacity data storage devices and new tools to analyze such data. It is obviously very difficult for an unaided human to process those medical records. The physician’s interpretation of images, signals, or any other clinical data, is written in unstructured free-text English, that is very difficult to standardize and thus difficult to mine. Even specialists from the same discipline cannot agree on unambiguous terms to be used in describing a patient’s condition. It has been suggested that computer translation may hold part of the solution for processing the physician’s interpretation. Another unique feature is privacy and security concerns. The ethical issues are obtaining the consent of patients for what their medical records will be used. The patients are so concerned that their medical information could be widely used and dispensed without their knowledge. These concerns have led to various efforts by international organizations to protect the confidentiality of those medical records. The following section will describe some of the classifiers used for intelligent decision making system.

1.11 Medical Data Classification

Development of computational methods for the diagnosis of heart disease has been the focus of many researchers. These methods are used to build knowledge based decision support systems which use the knowledge acquired from medical experts for effective decision making. However, the existing computer based decision support systems lack in terms of analysis of the past and prediction of the future. Hence, it is necessary to propose new classification techniques which can perform learning, analysis of the past and prediction of the future. In this research work, new classification techniques have been proposed and implemented for the effective analysis of multivariate heart datasets which are temporal in nature. The proposed
system has been tested with two benchmark datasets namely heart, cancer disease experimented by Das et al., (2009). The main advantage of the proposed classification algorithms are increasing in classification and prediction accuracy for providing effective medical decisions on heart, cancer diseases.

1.1.1 Input Dataset

Therefore, this proposed research work uses the benchmark dataset compiled for the Data Mining Repository of the University of California, Irvine (UCI). Finally, the system is validated using Datasets from Cleveland, Hungary and Switzerland. The main advantage of using this dataset is that the proposed research work is capable of providing significant data that are easily shared with other researchers and developers. The feedbacks obtained from other researchers allow improving the result of this proposed work.

a). Heart Dataset:

The main reason for selecting UCI Repository dataset is that currently, it is the mostly used comprehensive dataset that is shared by many researchers. The Cleveland heart disease database is the data set that has been used by ML researchers to this date. All published experiments related to using a subset of 14 of the 76 attributes present in the processed Cleveland heart disease database. Specifically, ML researchers use only the Cleveland database till today. The existence of heart disease in the patient is indicated in the “goal” field by means of an integer that can take any value from 0 (no presence) to 4. Distinguishing disease existence (values 1-4) from nonexistence (value 0) has been the focus of the experiments conducted in the Cleveland database (Blake and Mertz, 2004). A GA-based algorithm thus has the potential to evolve the feature subset, and may be developed as a GA-based pruning
method for hybrid fuzzy neural network classifiers. The description of the dataset is
given in Table-1.1 (http://archive.ics.uci.edu/ml/datasets/Heart+Disease).

- **Cleveland data**

  This data base encloses 76 characteristics; on the other hand all allocated tests
refer to exploiting a subset of 14 of them. Particularly, ML researchers employ only
the Cleveland database still today. The “goal” field refers to the presence of heart
disease in the patient. It is integer valued from 0 (no presence) to 4. Experiments with
the Cleveland database have contemplated on simply attempting to differentiate
presence (values 1, 2, 3, 4) from absence (value 0). The names and social security
numbers of the patients were lately eliminated from the database, substituted with
dummy values. Six of the examples have been rejected as they had missing values.
Class distributions are 54% heart disease absent, 46% heart disease present.

- **Hungarian data**

  Owing to a vast percentage of missing values three of the characteristics have
been rejected however the format of the data is precisely the similar as that of the
Cleveland data. Thirty-four examples of the database were rejected on account of
missing values and 261 examples were present. Class distributions are 62.5% heart
disease not present and 37.5% heart disease present.

- **Switzerland data**

  More number of missing values is in Switzerland data. It encloses 123 data
instances and 14 features. Class distributions are 6.5% heart disease not present and
93.5% heart disease present.
### Table-1.1:
Heart Disease Dataset Description

<table>
<thead>
<tr>
<th>Attribute No.</th>
<th>Attribute</th>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Age</td>
<td>Age in years</td>
<td>numeric</td>
</tr>
<tr>
<td>4</td>
<td>Sex</td>
<td>Sex</td>
<td>1 = Male; 0 = Female</td>
</tr>
</tbody>
</table>
| 9             | Cp        | Chest pain type | Value-1: typical angina  
Value-2: atypical angina  
Value-3: non-anginal pain  
Value-4: asymptomatic |
| 10            | Trestbps  | resting blood pressure | (in mm hg on admission to the hospital) |
| 12            | Chol      | serum cholesterol in | mg/dl |
| 16            | Fbs       | fasting blood sugar > 120 mg/dl, 1 = true; 0 = false |
| 19            | Restecg   | Resting electrocardiographic results | Value-0: normal  
Value-1: having st-t wave abnormality (t wave inversions and/or st elevation or depression of > 0.05 mv)  
Value-2: showing probable or definite left ventricular hypertrophy by estes' criteria |
| 32            | Thalach   | maximum heart rate achieved |
| 38            | Exang     | exercise induced angina | (1 = yes; 0 = no) |
| 40            | Oldpeak   | st depression induced by exercise relative to rest |
| 41            | Slope     | The slope of the peak exercise st segment | Value-1: upsloping  
Value-2: flat  
Value-3: downsloping |
| 44            | Ca        | number of major colored by fluorooscopy | vessels (0-3) |
| 51            | Thal      | 3 = normal; 6 = fixed defect; 7 = reversible defect |
| 58            | Num       | Diagnosis of heart disease (angiographic disease status) | Value-0: < 50% diameter narrowing  
Value-1: > 50% |
b). Cancer Dataset:

- **Leukemia Cancer Dataset**

  **Description:** The leukemia data set contains expression levels of 7129 genes taken over 72 samples. Labels indicate which of two variants of leukemia is present in the sample (AML, 25 samples or ALL, 47 samples). This dataset is of the same type as the colon cancer dataset and can therefore be used for the same kind of experiments. In fact, most of the papers that use the colon cancer data also use the leukemia data (http://mlearn.ics.uci.edu/databases/).

  **Size:**
  
  - 72 samples, split into 38 training and 34 test samples
  - 7129 genes
  - 3.8 MB

- **Breast and Lung Dataset**

  Breast and Lung cancer dataset adenocarcinoma tissues were collected from patients and from some of these patients, paired normal colon tissue also was obtained. Gene expression in 40 tumor and 22 normal colon tissue samples was analyzed. The data set contains the expression of the 2000 genes with highest minimal intensity across the 62 tissues. Each gene intensity has been derived from the about 20 feature pairs that correspond to the gene on the chip by using a filtering process. The data is otherwise unprocessed, i.e. no normalization has been performed yet. The training set consists of 40 colon tissues of which 14 are normal and 26 tumor samples. The test set consists of 22 tissues of which 8 are normal and 14 tumor samples. The number of gene expression levels is 2000. The goal here is to classify the tissues as being cancerous or noncancerous.
1.12 Motivation

The data mining community inherits the data mining techniques development in statistics and machine learning, and applies them to various real world problems. As data mining becomes more popular, recently classification techniques are increasingly applied to provide decision support in medical field, business, financial analysis, telecommunication, automatic recognition of letters or digits based on handwriting samples, intrusion detection in computer networks and biomedicine. To better solve application specific problems, there has been a trend toward the development of more application specific data mining systems.

1.13 Aim and Objectives

This thesis concentrates on the data mining techniques effectively than other Statistical models to medical data. The broad objectives of the research work are:

- To find the minimum number of most relevant attributes from a given medical dataset.
- To examined an improved HSVM for effectively predicting the disease of patients. To exploit the maximum potential of SVM
- An enhanced GWO strategy was established to search for the optimal feature subset for HISVM classification

1.14 Organization of the Thesis

The thesis is organized as follows, it consists of seven chapters.

Chapter-I focuses the basic concepts with historical developments of data mining and a detailed description of the Data mining techniques exploring multivariate medical data.
Chapter-II literature review reviews related work in medical data mining, Feature selection, classification and optimization.

Chapter-III Data mining using the efficient feature selection technique based on modified fuzzy C-means clustering with rough set theory and traditional SVM classification. This chapter reviews the FRS, FCM-RS algorithm, also discussed and produce the selected features of specified medical datasets.

Chapter-IV Data Mining using the Hybrid kernel improved SVM classifier applied to investigates the given medical dataset performance and presents combination kernel function of support vector machine. Through the experiment results and comparison results show that its performance is better than that of other SVMs constructed by ordinary kernel function. Four approaches have been highlighted in SVM based classification including HSVM, Naïve bays Classification Model, Nearest Neighbor Classification, in this PSO as Optimizer within Other Learning Algorithms. The performance is compared with classifiers including SVM, Decision Tree, and k-Nearest Neighbor. Also, some GWO variants for data classification are discussed in this chapter.

Chapter-V Data Mining using Enhanced Grey wolf Optimization Algorithm Presents the Simplified Grey Wolf Optimization (SGWO) algorithm that is based on Traditional PSO for data mining. This is followed by the principle of the GWO. It provides some introduction to population-based optimization algorithms with more emphasis on the PSO.

Chapter-VI investigates Results and Discussion. It finds computational results of proposed methodology and comparative analysis of existing model in this various
techniques have been compared with the help of implementation provided by Mat lab machine learning tool. The efficiency of algorithms has been compared on the basis of the following measures like, classification Accuracy, Specificity, Sensitivity, FNR and FPR.

Chapter-VII Discuss Conclusion and Future Work, it summarizes the main contributions and offers suggestions for further research on the topics of the thesis.

Several relevant references regarding the present study have been documented under a separate title Bibliography.