CHAPTER 2
LITERATURE REVIEW

2.1 INTRODUCTION

In this chapter, the related work for this research work and existing techniques for patients’ record extraction from distributed databases and the clustering of related records, frequent pattern identification and fuzzy classification for decision making in disease prediction using intelligent agents are discussed.

2.2 DISTRIBUTED DECISION SUPPORT SYSTEM

Pal & Sharma (2013) suggested an intelligent decision support system for establishment of new organizations on any geographical area using a graphical information system (GIS). It proposes an intelligent system to support making decisions for the establishment of new resources in any geographical area. Using GIS, the whole information can be retrieved from any geographical area. The GIS and Artificial Intelligence techniques are used here with a spatial database.

El-Sappagh & El-Masri (2014) suggested distributing clinical decision support system architecture. This technical architecture takes advantage of Electronic Health Record (EHR), data mining techniques, clinical records, proficient knowledge bases, feasible technologies and principles to offer decision-making assistance for healthcare professionals. The architecture will work extremely well in distributed EHR environments in
which each hospital possesses its local EHR, and it gratify the rapport, interaction and quantifiable objectives of an EHR. The model will also have a set of disseminating knowledge bases. Each knowledge base will be specialized in a specific disease domain, and the model achieves cooperation, integration and interoperability between these knowledge bases.

Amato et al (2013) reviewed a decentralized partially observable Markov decision process (Dec-POMDP). It is a rich framework to formulate sequential decision making and control problems for a distributed group of agents collaborating to achieve a common goal under uncertainty. Centralization may not be suitable for communication cost or for unreliability. In contrast, solutions to Dec-POMDPs yield decentralized control policies that the agents execute to collaboratively optimize the common objective. However, while many more specialized multi-agent models have been widely studied, the more general problem of quantification is Dec-POMDP solution methods with an adding the number of agents is still an open research question.

Rabelo (2012) designed a semi-automated distributed decision support system virtual enterprises presents a distributed, collaborative decision support framework for problem resolution in partnerships of autonomous and heterogeneous enterprises. The developed prototype for conceptual model's verification and validations is composed of three important elements: a collaborative discussion tool via argumentation with moderation, flexible decision protocol and tools for the previous impact of the decision by performance evaluation methods.

Agrebi et al (2014) proposed a decision support system for auditing distribution, logistics information systems are based on a multi-criteria hierarchical ranking and is built around a decision support system called SADAUDIT that allows the audit of complex systems. The work illustrates
the proposed methodology with the real case of three companies based in Tunisia and belonging to the same group.

2.3 AGENT-BASED DISTRIBUTED DATA MINING TECHNIQUES FOR DECISION SUPPORT

In the recent era business systems are moving towards globalization and their organizations’ data are stored in their local sites. There has been growing importance for the discovering useful trends or patterns of the large set of a database located in different sites to make efficient integrated decision making to overcome the uncertainties. Agent-based intelligent decision support system supports to achieve this feature.

An AGENT is an entity whose state is viewed as consisting of mental components such as beliefs, capabilities, choices and commitments (Shoham 1991). These mental states are precisely defined and have rough correspondence with their real world counterparts. Hence, what makes software or hardware component an Agent is a fact that one has chosen to analyze and control it in these mental states. Now the question is not what is an agent, but “What all entities can be viewed as having these mental states.” Although everything and anything can be described as having mental states, it is not always advantageous to do so.

Liu et al (2011) proposed a framework for an automated data mining system based on intelligent agents. This framework provides the base knowledge about agent system and for the developing intelligent agent-based automated data mining system for ordinary users who are lacking the knowledge of data mining.

Sycara & Zeng (1996) investigating techniques for developing distributed and adaptive collections of information agents that coordinate to
retrieve filter and fuse information relevant to the user, task, and the situation. In the work system of agents, information gathering is seamlessly integrated with decision support. In this paper, the work is focused on the distributed system architecture, agent collaboration interactions, and a reusable set of software components for structuring agents. The system has three types of agents: Interface agents interact with the user receiving user specifications and delivering results. The model is acquired and utilized for user preferences to guide system coordination in support of the user’s tasks. Task agents help users perform tasks by formulating the problem-solving plans and carrying out these plans through querying and exchanging information with other software agents. Information agents provide intelligent access to a heterogeneous collection of information sources. These system frameworks are implemented and are developing collaborating agents in diverse, complex real-world tasks, such as organizational decision making, and financial portfolio management.

“In Agent-Based Distributed Resource Allocation in Continuous Dynamic Systems”, Voos (2009) addressed the problem of distributed resource allocation and advantages of agent-based research and its applications. Here algorithms or methodologies have been developed that especially take into account of the decentralized system structure of multi-agent systems and their ability to communicate and coordinate. Well known methods of multi-agent based resource allocation comprise blackboard structures or auction-like algorithms. One method which will be investigated here is based on economic markets since resource allocation is also a basic problem in human societies. An abstract mathematical model of an idealized economic system consists of a certain number of commodities, agents and a price system. A commodity can either be a service or a good; any quantity of a commodity is a positive real number. Because of the limitation of the commodities, each is associated with a price while all prices together form the
price system. The agents can either be consumers or producers and the task of an agent is to make the decision on a quantity of his input or output for each commodity. Each producer chooses his supply based on his production factors and has the objective of profit maximization.

Haynes et al (2009) designed, intelligent agents capable of explaining themselves, which includes ontological, mechanistic, and operational explanations. These designs can help to develop better agents which support creating more usable and more affordable intelligent agents by encapsulating prior knowledge about how to generate explanations in concise representations. This system makes use of a knowledge base and algorithms to carry out their responsibilities. A knowledge base includes domain and problem-solving knowledge specific to the agent’s tasks. This design captures reviews and communicates the accumulated knowledge of intelligent system developers and researchers for explaining the structure and behavior of agents. Intelligent agents that can explain themselves expose the knowledge feed into their structure and behaviors and make them more efficient resources for their users and learners, and for agent developers who want to re-use and build on a proper work.

Sharma & Shadabi (2014) discovered research in multi-agents based data mining for intelligent decision support systems. The goal of this research is to mine a large set of data from distributed environment. The programming language used the Java programming language to leverage JADE and WEKA. The database technology used to store the dataset is MySQL, which is a well-known open source database. The data sets are divided into equal partitions and agent generation is based on partitions. Each Data Mining Agent operates on its own partition, or section and generates its own set of rules using decision tree classifiers. This research
work adopts Data mining and Multi-agent system (DMMAS) and the second experiment runs batch mining with all 5 derived “Adults Census Income” datasets. This study found that the processing speed is improved as the result of the multi-agent mining approach, although there can be a corresponding marginal loss of accuracy. In continuation that, the experiments have demonstrated the advantages of using DMMAS for classification tasks. The primary advantage of DMMAS is its ability to perform data mining tasks with base algorithms that are not able to load large data in memory. The improvement in efficiency of DMMAS becomes increasingly apparent as the dataset grows larger. In addition, the loss of accuracy that occurs with DMMAS compared to batch mining also becomes smaller as more data is made available.

Raj & George (2013) presented a distributed association rule mining algorithms and an agent-based distributed data mining architecture and an inside view of the existing agent-based association rule mining frameworks and point out the issues in the existing framework. The framework called mobile agent based distributed association rule mining (MAD-ARM) is presented, which attempts to reduce the communication overhead and ensures the mobile agent security.

Weng & Tran (2007) proposed an intelligent agent-based framework that allows traders to do business in remote locations by means of their mobile devices. The framework, which adheres to standardization efforts in the multi-agent field such as FIPA, paves a possible way towards a near future when mobile buying (and selling) agents can smoothly and securely travel among different agent-based marketplaces to carry out tasks on their users’ behalf. This proposed framework, aimed at providing new capabilities for advanced mobile e-business solutions, employs an approach that integrates the advantages of mobile agents and intelligent agents. In addition,
the intelligent and mobile agent technology is also a promising solution to the problems of low speed, high latency and limited computing ability that current wireless network is facing.

Moemeng et al (2008) explores an agent-based data mining integration for (1) formulation of a formal framework for multi-agent systems that allows extensibility, reusability, integrity of system components varied upon particular task; (2) providing a platform for data mining researchers which enhances research processes; and (3) providing a platform for agent researchers to extend functionalities of the system incorporated with data mining capability. The system is divided into three parts: control zone, optimization zone, and repository. Each part contains a collection of system components while agents work across these zones. The system flow starts from the control zone which determines the direction of the execution flow. Control Zone is where the user requests for experiment service and uploads his algorithm to the system, configures parameters and submits into an execution queue. Optimization Zone contains several instances of optimization units in, which may execute concurrently. A repository is a collection of system components, i.e. datasets, algorithms, and optimization methods; furthermore, it can also maintain links to on-line resources, e.g. remote databases. The repository concerns a high degree of concurrency control and security as some components are restricted to some particular users. A group of agents communicating in the defined set of channels. The system interface is a web-based implemented with PHP and AJAX to enhance usability and the underlying database is MySQL. JADE is used as the core agent container. Due to JADE’s FIPA compliance and extensibility, F-TRADE can be integrated further with other FIPA compliant applications.

(MAS). The work identifies various threads of work that exhibits, research and development activity in agent-based systems. Initially, the agent is defined and directed by the field of Artificial Intelligence. In this field, the symbolic representation of the agent’s environment is represented by first-order predicate logic. Then the agents are defined by using object oriented methodologies. By considering this scenario, agents have more different and advantages over objects. It differs with respect to the notion of flexible autonomous behavior. In agent systems, each agent considered having their own thread of control, whereas single thread of control in objects. And this review finds the definition of multi-agent system and the issues in future direction.

Decker et al (2006) proposed Intelligent adaptive information agents specify that the adaptation in open, multi-agent information gathering systems are important for several reasons. These reasons include the inability to accurately predict future problem-solving workloads, future changes in existing information requests, future failures and additions of agents and data supply resources, and other future task environment characteristic changes that require system reorganization. The system has developed a Multi-Agent distributed system infrastructure, REusable Task Structure-based Intelligent Network Agents (RESTINA) that handles adaptation in an open Internet environment. Adaptation occurs both at the individual agent level as well as at the overall agent organization level. The Retsina system has three types of agents. Interface agents interact with the user receiving user specifications and delivering results. They acquire, model, and utilize user preferences to guide system coordination in support of the user's tasks.

Ma et al (2009) proposed a formal method to establish trust theories for communication protocols in agent-based systems using temporal belief logic. The authentication protocol is defined and used for trust
management between agents. It used to verify and authorize agents acting on behalf of users to protect restricted information. After authentication, the two agents should be entitled to believe that they are communicating with each other and not with intruders. The research and commercial community have primarily focused on designing intelligent agent with specific domain knowledge with embedding rules and a trained data. Using the trained data and the intelligent agent, the information retrieval will provide flexible, modular and delegated solution.

Borowski et al (2011) explores integrating a reputation-based trust mechanism with an agent-based backup protection system to help protect against complex network failures. The goal of this research is to have the ability to find and react to malicious behavior while remaining compatible with traditional network security elements such as firewalls and data encryption. To achieve this goal, a reputation-based trust system has been integrated with an intelligent agent-based backup protection system, enabling agents to analyze other agents’ behavior. The success of this approach is determined by a performance comparison with the original agent-based backup protection scheme as well as with traditional protection mechanisms when malicious behavior interferes with communications between the communication-based agents.

Laftah Al-Yaseen et al (2015) proposed three agents, namely, coordinator, analysis, and communication agent. The basic concept underpinning the utilized MAS is dividing the largely captured network dataset into a number of subsets and distributing these to a number of agents depending on the data network size and core CPU availability. The proposed hybrid modified K-means with C4.5 classification in MAS is developed in JADE platform. The results show that compared to the current methods, the
intrusion detection system in a multiagent system (MAS-IDS) reduces the IDS processing time by up to 70%, while improving the detection accuracy.

Yang & Chang (2010) proposed a network management system by ontology supported multi-agent techniques. It is mainly designed for reducing the recovery time of network troubleshooting in the network management. This system integrated with the intelligent agent framework to effectively enhance and strengthen the network performance monitoring, and accordingly produces related quantification values of dynamic information. This system obtained the information from the coordination and cooperation of multi-agent system. The ontological free software is used for storing the operating information on network management in the backend database. The adaptability of the system is obtained from the use of open source code which make the user can precisely realize and control irregular phenomena in the network.

Yaskawa & Sakata (2003) present an application of intelligent agent technologies to different real world problems such as semiconductor manufacturing process, train operations and electric motor assembly lines. It breaks the conventional simulators which have limited to examination. In this agent construction, an additional platform Parallel Inference Machine (PIM) has been used that describes and executes multiple agents in the independent and parallel manner.

Don Gilbert (1998) explored the value of agents by examining some real agent-enhanced applications. This paper explains the application of intelligent agent like customer helpdesk, web browser intelligence, shopping assistance, etc. It lists the basic characteristics of the intelligent agent. Finally
the agent architecture of IBM named Agent Building Environment (ABE) is introduced to the basic functions.

2.3.1 Intelligent Agent Framework

Lo Piccolo et al (2013) reviewed the mobile agent JADE platform. The work aims at showing JADE performance measurements. First, assessed the effect of two different factors, the operating system, and the execution environment, on the container start-up operations, i.e., different OSs and Java JREs on the container startup. The ANOVA model has been applied to study the two factors and the relative interactions. Such approach has shown that GNU/Linux 2.6.7 OS on the MC host and Windows XP OS on the SC host leads to the low start up times regardless of the JRE version. To the best of the knowledge, the work provides a quantitative evaluation of JADE scalability in terms of agent migration and interaction with FIPA agents. The results exhibit a nearly linear scalability of the JADE platform.

An intelligent agent is implemented to equalize the human behavior and beliefs. Ronald & Sterling (2005) designed the agent-based model of pedestrian behavior in a real world environment. It defines the unpredictable nature of human decision making. Human decision exhibits different behaviors depending on their knowledge on that environment and other personal characters. BDI agent-based architecture model is implemented in this simulation which is useful for high-level decision making. Here the advantage of the agent is found that the capable of doing several things concurrently without trouble.

Pantic et al (2005) presented a flexible technique for teaching introductory artificial intelligence using Java and implemented simple agent framework. It uses an intelligent agent and other Artificial Intelligence
techniques to mention filter and retrieve relevant information from the World Wide Web.

Lee & Liu (2004) presented an integrated framework for data retrieval and data extraction in the context of internet shopping. To computerize a series of product search and selection activities, the work focuses on implementing agent technology with mining. It is based on the intelligent agent which is implemented by JADE - Java Agent Development Environment which is a multi-agent platform. The framework mainly concentrates on automatic fuzzification and defuzzification scheme.

Dastani (2008) presented a BDI-based agent-oriented programming language, named 2APL - A Practical Agent Programming Language. BDI is one of the agent architecture characterized by ‘mental state’ with three components called belief, desire, and intention. Belief is information that an agent has about the environment, desire is an option available for the agent to take a decision and intention represents the state of affairs that the agent has chosen and committed resources to. This language provides an effective integration of declarative and imperative style agent programming by introducing declarative beliefs and goals with events and plans.

One of the major innovative technologies of the agent is the development of distributed software system. (Bellifemine et al 2008) presented a framework JADE – Java Agent Development environment that facilitates the development of interoperable intelligent Multi-Agent systems. This paper presents JADE and its technological components together with a discussion of the possible reasons for multi-agent systems.

Finin et al (1994) describes the design and experimentation with the Knowledge Query and Manipulation Language (KQML), a new language and protocol for exchanging information and knowledge. This work is part of a
larger effort, the ARPA Knowledge Sharing Effort which is aimed at developing techniques and methodology for building large-scale knowledge bases which are shareable and reusable. KQML is both a message format and a message-handling protocol to support run-time knowledge sharing among agents. KQML focuses on an extensible set of performatives, which defines the permissible “speech acts” agents may use and comprise a substrate on which to develop higher-level models of inter-agent interaction such as contract nets and negotiation. In addition, KQML provides a basic architecture.

Imtiaz et al (2005) presented a framework for the unification of information extraction and data mining. This framework is designed by an intelligent agent. The intelligent agent is designed to extract features with the cross-feedback approach and also provides unified undirected graphical handling. It creates a good model for both entity and text level abstraction.

One of the main applications of the agent is network security. (Dasgupta et al 2005) describes the security agent architecture called Common intrusion detection system (CIDS) which is used as an administrative tool for intrusion detection. It allows easy inclusion of new detection, decision, and action plug-in, independently. It can also act as a monitoring tool for data gathering and visualization, which help to evaluate the behavior of any monitored network.

Arasu & Palanisamy (2007) presented Agent programming for processing, Bio-signal Waveforms. The paper demonstrates the concept of developing a Multi-Agent platform for the processing of Electroencephalogram / Electrocardiogram / Electromyogram waveforms. It also demonstrates the concept of developing agents using JADE – Java Agent Development framework. The agents are trained intelligent system that is
capable of setting up the platform for processing the Bio-signals. The agents themselves communicate with each other in the decision-making process.

The technical goal of this work is to develop a multi-agent platform for the processing of bio-signals aiming at assisting medical practitioners in developing standard examination procedures. After completing the steps of analysis and design, the features of JADE platform have been explored for the deployment of the Multi-Agent system. With reference to the design phase, the following agents have been developed using Java language: Generic Agent, Database Agent, Electroencephalogram Agent, Electrocardiogram Agent and Electromyogram Agent. The required behaviors and actions were implemented as per the design guidelines and FIPA recommendations. The required table has been created in Database which is accessible through Java Database Connectivity (JDBC). Based on the type of signal, the Generic Agent invokes the specific agent available in the agent platform which in turn processes the signal available in the data file. The result is being sent to the Generic Agent and it stores / updates the result in the Database using the Java database connectivity.

2.4 DISEASE PREDICTION TECHNIQUES

Clustering, Classification, and Pattern mining are an important data mining tasks for the prediction of disease from the related the symptom sets. Various algorithmic techniques are available to use for various medical applications.

2.4.1 Clustering Techniques

Jain & Maheswari (2012) survived the recent clustering techniques in data mining. Data clustering was a common technique for statistical analysis of the data. The data clustering was used in many fields such as
machine learning, pattern recognition, and bioinformatics. In the clustering process, the similar objects were classified into the different groups. The data were partitioned into the several subsets for grouping the medical dataset. Two types of an algorithm such as hierarchical were used in the data clustering process. The hierarchical algorithm measured the successive clusters based on the previously established clusters. The hierarchical algorithm was divided into divisive (top-down) or agglomerative (top-down) method. In the divisive method, the whole clusters were separated into successive small clusters. In the agglomerative method, all the smaller clusters were combined into larger clusters. In this survey, the conclusion is hierarchical algorithm was more efficient to attain the highest accuracy during the clustering of data.

Xiao & Hung (2007) has come out with a recent concept of queries and uncertain data which the academic communities have paid more attention. In the tasks such as clustering or nearest-neighbor queries, expected distance is often used as a distance measurement among uncertain data objects. Traditional database systems store uncertain objects using the expected (average) location in the data space. Distances can be calculated easily from the expected locations, but it poorly approximates the real expected distance values. Recent research work calculates the expected distance by calculating the weighted average of the pairwise distances among samples of two uncertain objects. However the pairwise distance calculation stake much longer time than the former method. The proposed method of this paper Approximation by Single Gaussian (ASG) is an efficient method to calculate the expected distance by a function of the means and variances of samples of uncertain objects. Theoretical and experimental studies show that ASG has both advantages of the latter method’s high accuracy and the former method’s fast execution time. The conclusion is ASG plays an important role in
reducing computational costs significantly in query processing and various data mining tasks such as clustering and outlier detection.

Haraty et al (2014) presented the concept that the huge amount of data generated by media sensors in health monitoring systems, medical diagnosis that produce media (audio, video, image, and text) content, and from health service providers is too complex, and voluminous to be processed and analyzed by traditional methods. Data mining approaches offer the methodology and technology to transform these heterogeneous data into meaningful information for decision making. This paper presents the studies of data mining applications in health care. Mainly, the study of k-means clustering algorithms on large datasets and present an enhancement to k-means clustering, which requires k or a lesser number of passes to a dataset. The proposed algorithm, which we call G-means, utilizes a greedy approach to producing the preliminary centroids and then takes k or lesser passes over the dataset to adjust these center points. In the experimental results, which were used in an increasing manner on the same dataset, shows that G-means outperforms k-means in terms of entropy and F-scores. The experiments also yield better results for G-means in terms of the coefficient of variance and the execution time.

Gu et al (2013) focused on the problem of MEDLINE clustering document. The ML and CL constraints were derived from two data sources, i.e. Medical Subject (MS) and Global Content (GC). The MS and GC constraints were automatically obtained from the database and it consists of large fraction pairs. The Must link (ML) and Cannot link (CL) constraints were generated independently based on the similarity measures. The automatic constraint generation was used in many other domains with multiple data sources. The noisy constraints were formed due to these
qualities of information sources. Due to the dissimilarity measures, approximately 10% of CL constraint was produced the incorrect clusters.

Ng & Jing (2009) introduced a new fuzzy K-modes clustering algorithm for categorical data. The dissimilarity of the clusters was measured for the categorical objects. The cluster optimization was analyzed based on the new fuzzy K-modes clustering algorithm. The frequency based fuzzy K-modes clustering technique was used to minimize the cost function of the clusters. The cluster centers were updated by calculating the distance between each object and each cluster centroid. The distances between two objects were computed by comparing the similarity between the clusters. The objective functions were minimized by using the new dissimilarity measures. Two data types, namely, numeric and categorical data types were used to map the data from the database. The numeric data type consists of real values. The categorical data type consists of finite and infinite values. The fuzzy K-modes clustering technique were used to cluster the data into the K-clusters by minimizing the objective functions. The fuzzy K-modes clustering method was not appropriate for the large medical database.

Bharathi & Vijayan (2014) developed a semi-supervised method for clustering the Medline document. Semi-supervised clustering technique was a combination of supervised and unsupervised learning. In the biomedical clustering method, the three different types of information were considered for grouping the data, such as the Global Content (GC) data, Local Content (LC) data and the Medical Subject Heading (MeSH) semantic information.

The LC information was obtained from the documents, whereas the GC data was retrieved from the entire MEDLINE collection. Two methods of constraints namely, Must Link (ML) constraint and Cannot Link (CL) constraint was used in the Semi-Supervised Spectral clustering (SSNcut) technique.
The ML and CL constraint information were retrieved from the medical subject heading semantics and global content. The similar data were provided by the ML constraint whereas CL constraint was provided the dissimilar data. Each clustering was evaluated by comparing predicted cluster with the true cluster. The ML constraint was more effective than CL constraint. To overcome this limitation, the K-means clustering technique was used for grouping the dissimilar data.

Soliman et al (2012) developed a hybrid fuzzy particle swarm and fuzzy K-modes clustering algorithm. The clustering algorithm was classified into two methods such as hard clustering algorithms and fuzzy clustering algorithms. In hard clustering algorithm, each pattern data was allocated to a single cluster during the clustering operation. In the fuzzy clustering method, the degrees of membership were assigned for each input pattern in the several databases. The overlapping of the clusters was reduced based on a Particle Swarm Optimization (PSO) method.

A hybrid Fuzzy Particle Swarm Optimization (FPSO) and Fuzzy K-modes (FK-modes) algorithm were used for clustering the categorical data. The concepts of FK-modes algorithm was used to handle the uncertainty phenomena. The FPSO technique was used to reach a global optimal solution for clustering optimization problems. The FPSO-FK-modes algorithm was implemented and evaluated by using the slandered benchmark data sets. The performance of the clustering data was also analyzed based on the FPSO-FK modes algorithm. This approach was suitable only for a limited number of the datasets.

Maulik et al (2010) introduced the integrating clustering and supervised learning techniques for categorical data analysis. The modified Differential Evolution (DE)-based Fuzzy C-Medoids (FCM) clustering was used for clustering the categorical data in the medical database. The FCM
algorithm was combined both the local as well as global data with adaptive weighting. The elements of the data were ordered based on the integrating clustering and supervised learning techniques. The data points were selected from different clusters based on their proximity to the respective medoids. The selected data points were used for training the Support Vector Machine (SVM) algorithm.

The clustering assignments of the remaining points were determined by using the trained classifier. The performance FCM algorithm was compared with the classical DE technique, the average linkage hierarchical clustering algorithm for four artificial and four real life categorical datasets. The limitation of the integrating clustering and supervised learning techniques were computationally too expensive for massive and high-dimensional datasets. Inability to make corrections once the splitting/merging decision were formed in the data mining process.

Bai et al (2013) introduced a novel fuzzy clustering algorithm for categorical data. In the fuzzy clustering method, the cluster information was integrated into the dataset. The within-cluster information was used for the dispersion process and the between-cluster information was used for the enhancement technique. Due to the integration process, the within-cluster and between-cluster were minimized simultaneously based on the fuzzy clustering method. The clusters were analyzed linearly depends on the number of data, object, attributes. The novel fuzzy algorithm was more effective and suitable for clustering categorical information sets. The fuzzy clustering method does not produce better accuracy for clustering the categorical database.

Cao et al (2012) proposed a K-modes clustering algorithm for measuring a dissimilarity criteria. The K-modes clustering technique were widely used for clustering the data based on the biological data and genetic taxonomy. The attribute values were calculated from the medical database.
The cluster features were extracted based on the attribute values. The distinct characteristics of the data were also analyzed for calculating a dissimilarity measure. The new dissimilarities were calculated between the nodes of the cluster and an object. The K-modes clustering algorithm was most effective in the case of a large set of the database. The convergence theorem and time complexity analysis were used to evaluate the clustering efficiency. The cluster scalability was analyzed based on the dissimilarity criteria. The K-modes clustering method was suitable only for the limited number of the datasets.

2.4.1.1 K-Means Clustering Techniques

Harale & Kulkarni (2011) introduced an iteratively improved K-means clustering algorithm. The data mining was an unsupervised learning technique for placing the data elements into related groups. The K-means clustering method was used to minimize the clustering error in the Euclidean space. The iterative K-means clustering algorithm was more efficient and produces better quality clusters. The cluster starting points was selected based on an optimization formulation method and a novel iterative method. The clustered method was applied to the Pima Indian medical diabetes dataset. The membership pattern of the each cluster was calculated based on the K-means clustering method. The K-means clustering method was used in the small data sets as well as the large datasets. The arbitrary shapes of the cluster were analyzed from the datasets. The high dimensional data were measured based on the K-means clustering method.

Prabhu & Anbazhagan (2011) suggested the K-Means clustering method for improving the performance of the high dimensional datasets. The dimensionality of the data was reduced through Principle Component Analysis (PCA) techniques for increasing the performance of the clustering process. The dimensionality reduction was the process of decreasing the
number of random variables. The features were selected from the random data variables. The PCA method was used to combine the optimal variable in a linear manner. The PCA technique was applied to the original data before clustering process and produces accurate results.

Chauhan et al (2010) formulated the data clustering method for discovering clusters in spatial cancer databases. The enormous amounts of hidden data were extracted from the medical datasets. The medical, as well as the spatial data, were analyzed by using the data mining techniques. The Spatial data mining includes the discovery of useful and interesting patterns from spatial databases by grouping the data into single clusters. The continuous and discrete spatial medical databases were analyzed for forming the efficient clusters. The arbitrary shapes of the clusters were formed while the data were continuous in nature.

Two data mining techniques, namely, K-Means clustering techniques and hierarchical agglomerative clustering were applied on the spatial data set to produce the efficient clusters. The K-means clustering algorithm was a classical clustering technique for grouping the large data sets into single clusters. The optimal clusters were analyzed based on the K-means clustering method. This algorithm was described as follows:

- The cluster center was selected arbitrarily for assigning the objects to the single clusters.

- The cluster means were updated for each cluster until there was no change in the database.

An agglomerative clustering method was one of the types of hierarchical clustering techniques. In agglomerative approach, the data were clustered based on bottom-up strategy. The atomic clusters were merged into
larger until all the data were merged into the distinct cluster. The drawbacks behind these clustering techniques were not able to produce the different shapes of the clusters. The missing values were not analyzed during the clustering process.

Bhatia & Khurana (2013) proposed the data clustering by using the K-Means and the modified algorithms. The clustering algorithm was useful to analyze the data from the different database. The data were grouped into the single clusters by using the K-Means clustering algorithm. The number of iterations, accuracy, Silhouette validity index and execution time were calculated to analyze the performance of the clustering process. The k-Means algorithm was one of the popular partitioning algorithms in the data mining techniques. The K-Means clustering techniques were used to classify the data into k clusters. The local minimum of the clusters was analyzed to estimate the dissimilarity measures. It consists of two different phases such as, in the first phase, the K centers were determined for each cluster randomly, and in the second phase, the distance between data points and the cluster centers were measured. Euclidean distance was used to determine the distance. The data point was assigned to its nearest cluster.

The original K-means algorithm was used to select the initial clusters. The input data space was partitioned into multiple segments based on the modified approach. The frequency of the each segment was calculated after the data space partitioning process. The initial centroid was measured to form the cluster with better accuracy. This method was not suitable for the complex dataset.

2.4.2 Classification Techniques

Qian & Zhou (2011) proposed a novel approach to tree kernel-based protein-protein interaction (PPI) extraction, where the tree
representation generated from a constituent syntactic parser is further refined using the shortest dependency path between two proteins derived from a dependency parser. Specifically, all the constituent tree nodes associated with the nodes on the shortest dependency path are kept intact, while other nodes are removed safely to make the constituent tree concise and precise for PPI extraction. Compared with previously used constituent tree setups, the proposed dependency-motivated constituent tree setup achieves the best results across five commonly used PPI corpora. Moreover, the tree kernel-based method outperforms other single kernel-based ones and performs comparably with some multiple kernel ones on the most commonly tested AIMed corpus.

Dash et al (2015) proposed the hybrid gravitational search and particle swarm based fuzzy Multi Layer Perceptron (MLP) method for medical data classification. The hybrid training algorithm for fuzzy MLP called MLP-GSPSO was used for classifying medical data set by merging two heuristic methods such as Particle Swarm Optimization (PSO) and gravitational Search (GS). The five datasets were used as the benchmark dataset for analyzing the performance of the fuzzy MLP-GSPSO model from the data warehouse. The classification procedure consists of two types: binary classification and multi-class classification. The data were divided between two classes based on the binary classification process. In the multi-class classification process, the data were separated between more than two classes. The MLP-GSPSO method has not provided the better classification accuracy in the medical database.

Çelik et al (2015) presented a study of evaluated the diagnostic accuracy of immune system algorithms with the aim of classifying the primary types of a headache that are not related to any organic etiology. They are divided into four types: a migraine, tension, cluster, and other primary
headaches. After we took this main objective into consideration, three different neurologists were required to fill in the medical records of 850 patients in our web-based expert system hosted on our project website. In the evaluation process, Artificial Immune Systems (AIS) were used as the classification algorithms. The AIS is classification algorithms that are inspired by the biological immune system mechanism that involves significant and distinct capabilities. These algorithms simulate the specialties of the immune system such as discrimination, learning, and the memorizing process in order to be used for classification, optimization, or pattern recognition. According to the results, the accuracy level of the classifier used in this study reached a success continuum ranging from 95% to 99%, except for the inconvenient one that yielded 71% accuracy.

Riano et al (2015) introduced an ontology to the care of chronically ill patients and implement two personalization processes and a decision support tool. The first personalization process adapts the contents of the ontology to the particularities observed in the healthcare record of a given concrete patient, automatically providing a personalized ontology containing only the clinical information that is relevant for health care professionals to manage that patient. The second personalization process uses the personalized ontology of a patient to automatically transform intervention plans describing health care general treatments into individual intervention plans. For co-morbid patients, this process concludes with the semi-automatic integration of several individual plans into a single personalized plan. Finally, the ontology is also used as the knowledge base of a decision support tool that helps health care professionals to detect anomalous circumstances, such as wrong diagnoses, unobserved co-morbidities, missing information, unobserved related diseases, or preventive actions.
Assareh & Moradi (2007) presented a new method for feature selection which considers all of the M/Z points as potential features, rather than just considering peaks. The proposed method is a weight factor in ‘T test’ feature selection algorithm to eliminate those features which are regionally correlated and thus likely to correspond to the same peptide. Intuitive linguistic rules are then built based on histogram analysis of the biomarkers intensities and the membership function are adjusted using a genetic algorithm. The main emphasis was to extract a limited number of simple rules, which can be understood by a human being as knowledge and furthermore, human expert can add his/her knowledge of the system by a number of rules. In the other word, the proposed method simplifies the “knowledge exchange” procedure between a human being and classification system. Moreover, from classifier perspective, and using only two simple linguistic rules, the proposed method achieved a satisfactory accuracy and outperformed two well-defined classification methods: LDA and K-NN.

Oswal et al (2014) surveyed the data mining methods to analyze the medical data information resources. The knowledge gathered from the data of patients collected in databases facilitate in the diagnosis process. The survey focused on (i) domain specific which is for making a diagnosis of particular diseases such as system involved with the only diagnosis of heart diseases. (ii) Second types of the system which concentrate on multiple disease diagnosis. The method divides the medical diagnosis system into two category such as a system which uses symptoms to identify diseases and not considering other variables such as family medical history, age, etc. and another kind of system which considers symptoms along with variables such as medical history, age etc. This survey describes various algorithmic models such as K-Means, K-NN, SVM, and C4.5.
Dan-qing (2012) developed the SPRINT Classification Algorithm for the gastric clinical data. The decision-tree based model was proposed for analyzing gastric cancer from the clinical dataset. This model was developed from the existing effective database or data warehouse. The gastric cancer recurrence was extracted to form a decision tree training data set. By using the SPRINT classification algorithm, the risk factors were analyzed for the gastric cancer recurrence. Based on the examination of all the potential factors affecting the clinical treatment, diagnosis and prognosis were estimated. The SPRINT classification algorithm estimates the primary risk factor for gastric cancer recurrence in the medical dataset. The gastric cancer was not classified accurately by using the SPRINT classification algorithm.

“Rough – Granular Computing in Knowledge Discovery and Data Mining”, (Stepaniuk 2008) at present developed so far rough set methods have shown to be very useful in many real life applications. Rough set based software systems, such as RSES, ROSETTA, LERS and Rough Family have been applied to KDD problems. The patterns discovered by the above systems are expressed in attribute-value languages. There are numerous areas of successful applications of rough set software systems. The method presents applications of rough set and clustering methods for knowledge discovery in real life medical data set. The method has four sub-tasks identification of the most relevant condition attributes, application of nearest neighbor algorithms for rough set based reduced data, discovery of decision rules characterizing the dependency between the values of condition attributes and decision attribute, information granulation using clustering. The nearest neighbor paradigm provides an effective approach to classification and is one of the top ten algorithms in data mining. The k-nearest neighbor (kNN) classification finds a group of k objects in the training set that are closest to the test object and bases the assignment of a decision class on the predominance of a particular class in this neighborhood. There are three key elements of this
approach: a set of labeled objects, e.g., a decision table, a distance or similarity metric to compute the distance between objects, and the value of k, the number of nearest neighbors. To classify the new object, the distance of this object to the labeled objects is computed, its k-nearest neighbors are identified, and the decision class of these nearest neighbors is then used to determine the decision class the object. A major advantage of nearest neighbor algorithms is that they are nonparametric, with no assumptions imposed on the data other than the existence of a metric. However, the nearest neighbor paradigm is especially susceptible to the presence of irrelevant attributes. The rough set approach is used for selection of the most relevant attributes within the diabetes dataset. Next-nearest neighbor algorithms are applied with respect to a reduced set of attributes.

In learning approximations of complex concepts, there is a need to choose a description language. This choice may limit the domains to which a given algorithm can be applied. There are at least two basic types of objects: structured and unstructured. An unstructured object is usually described by attribute-value pairs. For objects having an internal structure first order logic language are often used. In the book, the investigation is for both types of objects. In the former case the propositional language with atomic formulas being selectors (i.e. Pairs attribute=value) is used, in the latter case, the first order language is considered. Attribute-value languages have the expressive power of propositional logic. These languages, sometimes do not allow for proper representation of complex structured objects and relations among objects or their components. The background knowledge that can be used in the discovery process is of a restricted form and other relations from the database cannot be used in the discovery process. Using first-order logic (or FOL for short) has some advantages over propositional logic. First order logic provides a uniform and very expressive means of representation. The background knowledge and the examples, as well as the induced patterns, can
all be represented as formulas in a first order language. Unlike propositional learning systems, the first order approaches do not require that the relevant data be composed into a single relation but, rather can take into account data, which is organized in several database relations with various connections existing among them. First order logic can face problems which cannot be reduced to propositional logics, such as recurrent structures. On the other hand, even if a problem can be reduced to propositional logics, the solutions found in FOL are more readable and simpler than the corresponding ones in propositional logics.

Tu et al (2009) described the medical data classification based on the decision tree and the bagging algorithms. Diagnosing of the heart disease was one of the important issues and many researchers examined to improve medical decision support systems intelligently for helping the physicians. The classification process was one of the essential tasks in the data mining. The input for classification was a set of training records and the training instances. The several attributes were present in the each record. The decision tree C4.5 algorithm and bagging with decision tree C4.5 algorithm were used to classify the medical dataset.

Kharche et al (2014) was compared with different datasets using various classification techniques with Waikato Environment for the knowledge learning (WEKA) application. The classification techniques were used to classify the datasets into a predefined set of classes. Different data sets from Irvine and University of California were compared with different classification techniques. The open Source Application— WEKA was a collection of machine learning algorithms and data preprocessing tools. Different types of the classifiers such as J48, Naive Bayes, Naive Bayes Updatable Classification algorithm, Classification and Regression Testing (CART) were used in the WEKA application. The data mining techniques
such as data preprocessing, feature selection, clustering, and classification were supported by the WEKA application.

The J48 algorithm was constructed the decision tree from a training data by using the Information entropy. The J48 technique was examined the normalized factor from the selected attributes. The decision tree was formed based on the splitting criteria. Bayes Net Classifier depends on the Bayes theorem and this classifier was a directed acyclic graph. In Bayes Network, all the attributes were nominal in nature. The Naive Bayes Updatable algorithm was the improved version of Naive Bayes. The default precision was used for generating the numeric attributes. The CART decision tree was used to construct the tree based on the categorical or numeric data set. Two branches were generated at each node based on the CART decision tree, therefore the binary decision tree was formed for each dataset. The J48 and the Naïve Bayes Updatable algorithm produce high classification accuracy compared to another method.

2.4.2.1 Fuzzy Approach

Garibaldi et al (2012) described a novel approach to incorporate variability within a fuzzy inference system using non-stationary fuzzy sets in order to replicate human variability. The proposed approach is applied to a decision problem concerning the recommendation of post-operative breast cancer treatment; specifically, whether or not to administer chemotherapy based on assessment of five clinical variables: NPI (the Nottingham Prognostic Index), estrogen receptor status, vascular invasion, age and lymph node status. In doing so, the method explored whether such explicit modeling of variability provides any performance advantage over a more conventional fuzzy approach, when tested on a set of 1310 unselected cases collected over a fourteen-year period at the Nottingham University Hospitals NHS Trust, UK. The experimental results show that the standard
fuzzy inference system (that does not model variability) achieves overall agreement in clinical practice around 84.6% (95% CI: 84.1–84.9%), while the non-stationary fuzzy model can significantly increase performance to around 88.1% (95% CI: 88.0–88.2%), p < 0.001. The Conclusion is that non-stationary fuzzy models provide a valuable new approach that may be applied to clinical decision support systems in any application domain.

Huang et al (2012) proposed the data mining procedure involves neural network based clustering using Adaptive Resonance Theory 2 (ART2), and the extraction of fuzzy decision trees for each homogeneous cluster of data records using fuzzy set theory. Besides, another objective of this paper is to examine the effect of the number of membership functions on building decision trees. The experiments confirm that the erroneous classification rates are significantly reduced by processing simplified decision trees. Next, the experimental results show that the discovered decision tree rules from individual cluster are more efficient than those based on all of the data. The number of erroneous clustered patterns in the decision tree of individual cluster is smaller than for the entire tree. Applications of the proposed method are not limited to medical databases. As long as the accuracy and efficiency of data classification are concerned, the presented work is applicable to the multi-attribute databases.

Nguyen et al (2015) developed the Classification method for healthcare data using genetic fuzzy logic system and wavelets. The healthcare played an important role in stimulating the general health and well-being of people around the world. The healthcare data were difficult to classify the information from the medical database. An integration of fuzzy standard additive model (SAM) with Genetic Algorithm (GA), called GSAM, was used to deal with uncertainty and computational challenges in medical database. The GSAM learning process was required the minimum computational costs
for classifying the data and provide the maximum efficiency compared to SAM technique. The GSAM technique does not suitable for small medical database.

Begum & Devi (2011) brought down the concept that one of the most likely applicable fields of fuzzy set theory which Zadeh himself described was medical diagnosis (Steimann 1997; Seising 2006) and since then several methods based on fuzzy knowledge and information has been developed to detect the diseases at its early stage. Fuzzy set theory provides a number of suitable properties for the pattern recognition diagnostic system due to its ability to deal with uncertainties, vagueness, and incompleteness in medical diagnosis and prognosis. It can be used to represent fuzzy objects (both linguistic and/or set of variables) and fuzzy logic (reasoning methods). (Torres & Nieto 2006) presents a review of the current applications of fuzzy logic medicine and bioinformatics. The main reasons for the application of fuzzy set theory in pattern recognition are: (i) its way of representation in linguistic approach with excellent formulation of input feature, (ii) Representation of missing or incomplete knowledge as a degree of membership and (iii) its capability of drawing approximate inferences. Fuzzy set theory help to transfer a qualitative evaluation of the medical data into the algorithmic structure and the focus of this paper is on algorithmic methods for pattern recognition based on fuzzy set theory. Baraldi & Blonda (1999a,1999b) present a survey of fuzzy clustering algorithms for pattern recognition. Literature survey shows that many fuzzy clustering algorithms aim to model fuzzy (i.e., ambiguous) unsupervised (unlabeled) patterns efficiently and is widely used for segmentation of MRI in brain tissue. Following are the some of the fuzzy clustering techniques.
2.4.3 Pattern Mining Techniques

Sreedevi et al (2014) proposed the concept due to large amounts of records and dimensions is increasing in available databases, pattern mining is a challenging problem in data mining research. A good number of parallel and distributed algorithms have been proposed to mine frequent item sets based on a support threshold of item set. But, regularity of item set is essential than the frequency of item set in bank transactions, network transactions, and sensor data, etc. Closed item set has gained lot of attention than frequent item set mining in recent data mining research. Based on these considerations, a novel approach was proposed with the parallel and distributed method for incremental mining of closed regular patterns using the vertical data format. The proposed method is capable of generating local models (each node has its own database summary) as well as a global model of closed regular patterns (node has the summary of whole database). This ability permits our approach to generating high contrast, closed regular item sets, which allows examining how the data is subjective at different nodes.

Long et al (2011) describes the process of finding correlations or patterns among dozens of fields in large relational databases. While Association Rules Mining (ARM) algorithm, especially the Apriori algorithm has been an active research work in recent years. Diverse improvement varies with term of producing more frequent items and also generating further k-length. Based on this idea this research work focused on new approach new approach for ARM based on Multiple Attribute Value within the non-binary search spaces. The proposed algorithm improves the existing frequent pattern mining by generating the most frequent values (item) within the attribute and generate candidate based on the frequent attribute value.
Chapman et al (2012) presents a study to describe the characteristics of the corpus in terms of the frequency of anaphoric relations, the syntactic and semantic nature of the members of the pairs, and the types of anaphoric relations that occur. Understanding how anaphoric reference is exhibited in clinical reports is critical to developing reference resolution algorithms and to identifying peculiarities of clinical text that may alter the features and methodologies that will be successful for automated anaphora resolution.

Khaleel et al (2014) focused on the data mining approaches or methods that help in finding periodically frequent diseases. A pattern which occurs at regular time intervals is said to be periodically-frequent. Periodic frequent pattern discovery can help people concerned to make well-informed decisions. The method surveyed in this paper is FP-Tree, Apriori, and MaxCPF-Tree Pattern Mining. There are two parameters that are to be supplied by domain experts in order to extract very useful periodic frequency patterns that can be used for expert decision making. These parameters are minimum support and periodicity. However, some methods found in the literature make use of both while other methods use only support parameter. The rate item problem is also identified and which can be resolved using low support value and high periodicity, it will not ignore certain values for consideration.

Alghamdi (2011) presents the Frequent Pattern (FP)-Growth algorithm is the major concern for data mining in medical data. FP-growth is an approach based on divide and conquers method. The main purpose of this technique is to produce frequent item sets by using the combination of data attributes. It basically works on to generate frequent item set without candidate set generation.
The major steps of FP-growth are:

(i) First reduces the database with frequent item set into a FP-tree.

(ii) It divides the FP-tree into a set of conditional database and mines each database separately, thus, extract frequent item sets from the FP-tree directly. It consists of one root labeled as null, a set of item prefix sub trees as the children of the root, and a frequent item header table. Each node in the item prefixes sub-tree consists of three fields: item-name, count and node link where, item name registers which item the node represents; count registers the number of transactions represented by the portion of path reaching this node, node-link links to the next node in the FP-tree. Each item in the header table consists of two fields, such as item name and head of the node -link, which points to the first node in the FP-tree carrying the item name.

Sabita Barik et al (2010) describes the difficulties in Apriori algorithm, which generates and tests candidate itemsets (gene sets) level by level. This processing causes iterative database (dataset) scans and high computational costs. Apriori algorithm also suffers from mapping the support and confidence framework to a crisp boundary. This research work proposed Fuzzy FP-growth approach not only outperforms the Apriori with respect to computational costs, but also it builds a tight tree structure to keep the membership values of the fuzzy region to overcome the sharp boundary problem and it also takes care of scalability issues as the number of genes and condition increases. FP-Tree developed for Gene expression clustered data sets.
2.5 EFFICIENT QUERIES RETRIEVAL

Jamsutkar et al (2013) describes the importance of efficient query retrieval. In recent years, with the development of computer network and database technology, distributed database is more and more widely used; with the expanding application, data queries are increasingly complex, the efficiency requests are increasingly high, so query processing is a key issue of the distributed database system. In a distributed database environment, the data stored at different sites connected through network.

Hong et al (2014) discussed the performance evaluation measurement and business computing system optimization based on some experiment researches on the efficiency of database queries. The query efficiency affected the performance of the RDBMS. According to the experiments, several factors (the design of database, the difference between DBMSs, the selection of indexes) could be the reasons of the system performance. They should be considered when implementing the optimization. The database optimization can save the cost. It will significantly improve the user experience and shorten the waiting time of some applications’ responses. According to the experiments, the indexes created on columns appearing in SELECT clauses had not made any effect. Appearing in WHERE clauses, the indexes on columns were useful, For the LIKE operator, it could handle parameter is not starting with a wildcard. Indexes on foreign keys are effective. In some experiments, creating indexes was not the more the better. The index took up some disk space and other resources. Inherent constraints also slowed down INSERT and UPDATE operations on the table. So creating an index should be cautious. Sometimes it could significantly optimize the performance, but the necessity and other conditions should also be considered.

Karde & Thakare (2010) present a paper that gives the results of proposed tree based materialized view selection algorithm for query
processing. In a distributed environment where database is distributed over the nodes on which query should get executed and also plays an important role. This paper also proposes node selection algorithm for fast materialized view selection in a distributed environment. A common method that is used in practice for providing higher information and best response time is the concept of materialized views, where a query is more quickly answered. One of the most important decisions in designing data Warehouse is selecting views to materialize for the purpose of efficiently supporting the decision making. The view selection problem defined is to select a set of derived views to materialize that minimizes the sum of total query response time & maintenance of the selected views. So the goal is to select an appropriate set of views that minimizes total query response time and also maintains the selected views.

Nithya et al (2014) proposed the concept that draws upon the functions relating to allocation, storage and retrieval of multimedia data that is stored at one or more sites in a distributed environment. The multimedia data considered in the current work are texts, images, videos, audios and graphical objects. The user queries and multimedia types can vary to a greater extent depending upon the application and deployment scenario. Online learning has increased considerably in recent times. This places a heavy demand on retrieving the most relevant multimedia information with reasonably good response time. This paper deals with an architectural design for improving the efficiency of multimedia data retrieval in a distributed environment. It includes the development of allocation manager, multimedia storage manager, query evaluation plan generator, query execution manager and query interface manager. The allocation manager allocates multimedia data to one or more sites. The multimedia storage manager controls the indexing mechanism and schema information. The query evaluation plan generator prepares a query plan for user queries. The query execution manager facilitates execution of query plans. The query interface manager handles user input and output.
<table>
<thead>
<tr>
<th>S.No</th>
<th>Author</th>
<th>Contributions</th>
<th>Support for Proposed Research Work</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>El-Sappagh &amp; El-Masri 2014</td>
<td>Distributed Clinical Support System Architecture. Data mining engine fetches distributed sources and helps for timely decisions</td>
<td>Supports knowledge with Distributed Decision Support –Collects patient knowledge base from different sources.</td>
</tr>
<tr>
<td>3</td>
<td>Liu et al 2011</td>
<td>A framework for automated data mining system based on intelligent.</td>
<td>Provides knowledge base about agent system.</td>
</tr>
<tr>
<td>4</td>
<td>Sycara &amp; Zeng 1996</td>
<td>Proposed the distributed system architecture, agent collaboration interactions, and a reusable set of software components for structuring agents. Identified single agent limitations and proposed solutions to use of multi-agents such as Interface, Task and Information agents and its collaborations for decision making and financial management.</td>
<td>Assist in knowing about multi-agent architecture and its collaboration structure applied in distributed system architecture.</td>
</tr>
<tr>
<td>S.No</td>
<td>Author</td>
<td>Contributions</td>
<td>Support for Proposed Research Work</td>
</tr>
<tr>
<td>------</td>
<td>--------</td>
<td>---------------</td>
<td>-----------------------------------</td>
</tr>
<tr>
<td>5</td>
<td>Sharma &amp; Shadabi 2014</td>
<td>Designed multi-agent based distributed data mining for intelligent decision support system. Designed with JADE and WEKA. Database used is MYSQL. Divides data locations into partitions and data mining agent operates on its own partition using decision tree classifier. Compared batch processing and multi-agent mining system and shows the improved process time and achieves higher scalability in agent based system.</td>
<td>Explores knowledge about multi-agent in distributed data mining, JADE, MYSQL and its advantages. Communication overhead reduction.</td>
</tr>
<tr>
<td>6</td>
<td>Laftah Al-Yaseen et al 2015</td>
<td>Proposed three agents, namely coordinator, analysis, and communication agent. Utilized MAS to divide the large datasets into partitions and distributes agents depending on data network size and CPU availability. Used K-Means with C4.5 classification developed in JADE Platform. Reduces upto 70% intrusion detection system processing time and detection accuracy.</td>
<td>Supports with multi-agent knowledge, k-means algorithm, JADE platform advantages for improving accuracy.</td>
</tr>
</tbody>
</table>
### Table 2.1 (Continued)

<table>
<thead>
<tr>
<th>S.No</th>
<th>Author</th>
<th>Contributions</th>
<th>Support for Proposed Research Work</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>Cao et al 2012</td>
<td>Proposed K-modes algorithm for measuring dissimilar criteria in biological data and genetic taxonomy. The cluster feature were extracted based on the attribute values. The new dissimilarities were calculated between the nodes of the cluster object. The cluster scalability was analyzed based on the dissimilarity criteria. Concluded that k-modes was suitable only for the limited number of the datasets.</td>
<td>Helps to understand the clustering of biological data and the limitations of k-modes clustering technique.</td>
</tr>
<tr>
<td>8</td>
<td>Kharche et al 2014</td>
<td>Compared the different datasets with various classification techniques with WEKA. Classification techniques used such as J48, Bayes net, Naive Bayes updatable and CART. Data mining used is data preprocessing, feature selection, clustering and classification. The binary decision tree was formed for each dataset with different algorithms.</td>
<td>Helps to provide knowledge with various datasets and classification algorithms and its efficiency.</td>
</tr>
<tr>
<td>S.No</td>
<td>Author</td>
<td>Contributions</td>
<td>Support for Proposed Research Work</td>
</tr>
<tr>
<td>------</td>
<td>--------</td>
<td>---------------</td>
<td>-----------------------------------</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Compared the accuracy and time efficiency with the classification techniques. J48 and Bayes updatable algorithm produced high classification accuracy with less consuming compared with other algorithms. Maximum accuracy percentage is 96 and it is based on the data set.</td>
<td>Brought down the knowledge about significance of use of fuzzy logic in health care. Comparative results shows that the fuzzy approach gives high accuracy compared to other approaches. Tested with breast cancer data set and heart disease dataset.</td>
</tr>
<tr>
<td>9</td>
<td>Nguyen et al 2015</td>
<td>Developed classification method for healthcare data using fuzzy logic system and wavelets. An integration of fuzzy standard additive model (SAM) with genetic algorithm (GA) called GSAM was used to deal with uncertainty and computational challenges in medical database. GSAM learning process comprises three continual steps: rule initialization by unsupervised learning using the adaptive vector quantization clustering, evolutionary rule optimization by GA and parameter tuning by the gradient descent supervised learning. Wavelet transformation is employed to extract discriminative features for high-dimensional datasets. GSAM becomes highly capable when deployed with small number of wavelet features as its computational burden is remarkably reduced.</td>
<td></td>
</tr>
</tbody>
</table>
Table 2.1(Continued)

<table>
<thead>
<tr>
<th>S.No</th>
<th>Author</th>
<th>Contributions</th>
<th>Support for Proposed Research Work</th>
</tr>
</thead>
<tbody>
<tr>
<td>10</td>
<td>Khaleel 2014</td>
<td>Focused on mining of frequent disease patterns. The method surveyed in this paper is FP-Tree, Apriori, and MaxCPF-Tree Pattern Mining. The two parameters used to support for frequent mining is minimum support and periodicity. Some methods found in the survey used both while other methods use only support parameter.</td>
<td>Gives knowledge about frequent pattern mining algorithms and its features. Comparatively FP –Tree growth algorithm used divide and conquer technique. It uses the combination of data attributes for frequent mining. And it does not generate candidate set generation as like apriori.</td>
</tr>
</tbody>
</table>
2.6 RESEARCH GAP IDENTIFICATION

- Can we perform data mining from distributed medical experts historical database?
- Can use multi-agents with distributed clustering, mining only relevant data to main container and makes fuzzy classification?
- Can fetch only relevant data to the main container?
- Can we find prone disease with the given symptom set with maximum accuracy?
- Can agents move only to the maximum density data locations?

2.7 CONCLUSION

In this chapter the related work of this research work and existing techniques for distributed decision making, Agent-based distributed data mining, Models used for disease prediction, pattern mining and efficient query retrieval are discussed. In medical database, the large volume of data is stored in distributed locations. It is observed that the multi-agents are required for easy retrieval and perform the medical data mining in an efficient way. The data mining techniques such as clustering, association rule mining and fuzzy classification methods are vital for easy diagnosing the medical records. In clustering, the similar disease sets are grouped to form single clusters. The frequent patterns are identified using clustered data. The generated rules may be redundant. An efficient classification method is required for accurate decision making. The probabilistic fuzzy classifies provides high accuracy compared to other existing methods. All the existing methods retrieve all the data from all the locations for diagnosing, which leads to time-consuming and space complexity. The efficiency of accurate classification, retrieving data in optimal locations and minimizing space usage are discussed in the following chapters.