Data Mining Applications In Healthcare Sector: A Study

M. Durairaj, V. Ranjani

ABSTRACT: In this paper, we have focused to compare a variety of techniques, approaches and different tools and its impact on the healthcare sector. The goal of data mining application is to turn that data are facts, numbers, or text which can be processed by a computer into knowledge or information. The main purpose of data mining application in healthcare systems is to develop an automated tool for identifying and disseminating relevant healthcare information. This paper aims to make a detailed study report of different types of data mining applications in the healthcare sector and to reduce the complexity of the study of the healthcare data transactions. Also presents a comparative study of different data mining applications, techniques and different methodologies applied for extracting knowledge from database generated in the healthcare industry. Finally, the existing data mining techniques with data mining algorithms and its application tools which are more valuable for healthcare services are discussed in detail.

Index Terms: Data Mining, Knowledge Discovery Database, In-Vitro Fertilization (IVF), Artificial Neural Network, WEKA, NCC2.

1. INTRODUCTION

The purpose of data mining is to extract useful information from large databases or data warehouses. Data mining applications are used for commercial and scientific sides [1]. This study mainly discusses the Data Mining applications in the scientific side. Scientific data mining distinguishes itself in the sense that the nature of the datasets is often very different from traditional market driven data mining applications. In this work, a detailed survey is carried out on data mining applications in the healthcare sector, types of data used and details of the information extracted. Data mining algorithms applied in healthcare industry play a significant role in prediction and diagnosis of the diseases. There are a large number of data mining applications are found in the medical related areas such as Medical device industry, Pharmaceutical Industry and Hospital Management. To find the useful and hidden knowledge from the database is the purpose behind the application of data mining. Popularly data mining called knowledge discovery from the data. The knowledge discovery is an interactive process, consisting by developing an understanding of the application domain, selecting and creating a data set, preprocessing, data transformation. Data Mining has been used in a variety of applications such as marketing, customer relationship management, engineering, and medicine analysis, expert prediction, web mining and mobile and mobile computing. In health care institutions leak the appropriate information systems to produce reliable reports with respect to other information in purely financial and volume related statements. Data mining tools to answer the question that traditionally was a time consuming and too complex to resolve. They prepare databases for finding predictive information. Data mining tasks are Association Rule, Patterns, Classification and Prediction, Clustering. Most common modeling objectives are classification and prediction. The reason that attracted a great deal of attention in information technology for the discovery of useful information from large collections is due to the perception that we are data rich but information poor.Some the sample data mining applications are:

- Developing models to detect fraudulent phone or credit-card activity
- Predicting good and poor sales prospectus.
- Predicting whether a heart attack is likely to recur among those with cardiac disease.
- Identifying factors that lead to defects in a manufacturing process.

Expanding the health coverage to as many people as possible, and providing financial assistance to help those with lower incomes purchase coverage [2]. Eliminating current health disparities would decrease the costs associated with the increased disease burden borne by certain population groups.Health administration or healthcare administration is the field relating to leadership, management, and administration of hospitals, hospital networks, and health care systems[1,3]. In the Healthcare sector Government spends more money.

- ➢ Proposal in draft NHP 2001 is timely that State health expenditures be raised to 7% by 2015 and to 8% of State budgets thereafter [21].
- ➢ Health spending in India at 6% of GDP is among the highest levels estimated for developing countries.
- ➢ Public spending on health in India has itself declined after liberalization from 1.3% of GDP in 1990 to 0.9% in 1999. Central budget allocations for health have stagnated at 1.3%ofthe total Central budget. In the States it has declined from7.0% to 5.5% of the State health budget.

- *M. Durairaj, Assistant Professor, Department of Computer Science, Engineering and Technology, Bharathidasan University, India,*
  Mobile No:+919487542202,
  (e-mail: durairaj_m@bdu.ac.in).
- *V. Ranjani, Research Scholar, Department of Computer Science, Engineering and Technology, Bharathidasan University, India,*
  Mobile No: +918056930161,
  (e-mail: ranjanimca2@gmail.com).

This paper mainly compares the data mining tools deals with the health care problems. The comparative study compares the accuracy level predicted by data mining applications in healthcare. Infertility is on the rise across the globe and it needs the sophisticated techniques and methodologies to predict the end results of infertility treatments particulars IVF (in-vitro fertilization) treatments, since the cost of IVF procedure is on the rise. In this study, we have taken this issue and compare the different techniques of data mining applications for predicting the Success rate of IVF treatment with the accuracy level. This comparative study could be useful for aspiring researchers in the field of data mining by knowing which data mining tool gives an accuracy level in extracting information from healthcare data.

2. LITERATURE REVIEW

A literature review is a text written by critical points of current knowledge including substantive find theoretical and methodological contributions to a particular topic. Literature reviews are secondary sources and do not report any new or original experimental work.

HianChyeKoh and Gerald Tan mainly discusses data mining and its applications with major areas like Treatment effectiveness, Management of healthcare, Detection of fraud and abuse, Customer relationship management[1].

JayanthiRanjan presents how data mining discovers and extracts useful patterns of this large data to find observable patterns. This paper demonstrates the ability of Data mining in improving the quality of the decision making process in pharma industry. Issues in the pharma industry are adverse reactions to the drugs [2].

M. Durairaj, K. Meena illustrates a hybrid prediction system consists of Rough Set Theory (RST) and Artificial Neural Network (ANN) for dispensation medical data. The process of developing a new data mining technique and software to assist competent solutions for medical data analysis has been explained. Propose a hybrid tool that incorporates RST and ANN to make proficient data analysis and indicative predictions. The experiments onspermatological data set for predicting excellence of animal semen is carried out. The projected hybrid prediction system is applied for pre-processing of medical database and to train the ANN for production prediction. The prediction accuracy is observed by comparing observed and predicted cleavage rate[20].

K. Srinivas, B. Kavitha Rani and Dr. A. Goverdhan discusses mainly examine the potential use of classification based data mining techniques such as Rule Based, Decision tree, Naïve Bayes and Artificial Neural Network to the massive volume of healthcare data. Using an age, sex, blood pressure and blood sugar medical profiles it can predict the likelihood of patients getting a heart disease[4].

ShwetaKharyadiscussed various data mining approaches that have been utilized for breast cancer diagnosis and prognosis Decision tree is found to be the best predictor with 93.62% Accuracy on benchmark dataset and also on SEER data set[5].

Elias Lemuye discussed the AIDS is the disease caused by HIV, which weakens the body's immune system until it can no longer fight off the simple infections that most healthy people's immune system can resist. Apriori algorithm is used to discover association rules. WEKA 3.6 is used as the data mining tool to implement the Algorithms. The J48 classifier performs classification with 81.8% accuracy in predicting the HIV status[6].

Arvind Sharma and P.C. Gupta discussedData mining can contribute with important benefits to the blood bank sector. J48 algorithm and WEKA tool have been used for the complete research work. Classification rules performed well in the classification of blood donors, whose accuracy rate reached 89.9%[7].

3. DATA MINING

Data mining is the non trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data. With the widespread use of databases and the explosive growth in their sizes, organizations are faced with the problem of information overload. The problem of effectively utilizing these massive volumes of data is becoming a major problem or all enterprises.

Definition

Data mining or knowledge discovery in database, as it is also known, is the non-trivial extraction of implicit, previously unknown and potentially useful information from the data. This encompasses a number of technical approaches, such as clustering, data summarization, classification, finding dependency networks, analyzing changes, and detecting anomalies[8].

Development of data mining

The current evaluation of data mining functions and products is the results of influence from many disciplines, including databases, information retrieval, statistics, algorithms, and machine learning [9] (See Fig. 1).

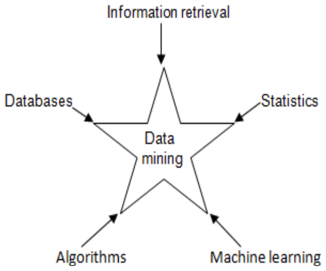


Fig. 1. Historical perspective of data mining

History of Data Base and Data Mining

Data mining development and the history represented in the Fig. 2. The data mining system started from the year of 1960s and earlier. In this, the data mining is simply on file processing. The next stage its Database management Systems to be started year of 1970s early to 1980s. In this OLTP, Data modeling tools and Query processing are worked. From database management system there three broad categories to be worked. First one is Advanced Database Systems, this evaluated year of Mid-1980s to present in this Data models and Application oriented process are worked. The Second part is Data Warehousing and Data Mining worked since the year of the late 1980s to present. The third part is Web based Database Systems this started from 1990s to present and in this Web mining and XML based database systems are included. These three broad categories are joined and create the new process that's called New generation of the Integrated Information system is started in 2000.

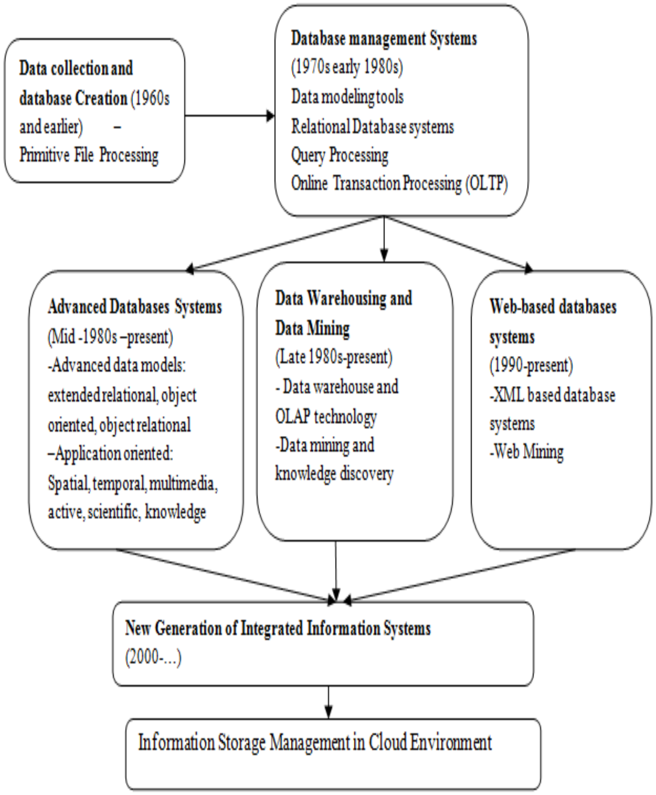

Data collection and database Creation (1960s and earlier) – Primitive File Processing
Database management Systems (1970s early 1980s)
Data modeling tools
Relational Database systems
Query Processing
Online Transaction Processing (OLTP)
Advanced Databases Systems (Mid -1980s –present)
-Advanced data models: extended relational, object oriented, object relational
–Application oriented: Spatial, temporal, multimedia, active, scientific, knowledge
Data Warehousing and Data Mining (Late 1980s-present)
- Data warehouse and OLAP technology
-Data mining and knowledge discovery
Web-based databases systems (1990-present)
-XML based database systems
-Web Mining
New Generation of Integrated Information Systems (2000-...)
Information Storage Management in Cloud Environment

Fig. 2. History of Database Systems and Data Mining

Data Mining Application Areas

Data mining is driven in part by new applications which require new capabilities that are not currently being supplied by today's technology. These new applications can be naturally into two broad categories.

- Business and E-Commerce
- Scientific, Engineering and Health Care Data

Data Mining Tasks

Data mining tasks are mainly classified into two broad categories:

- Predictive model
- Descriptive model

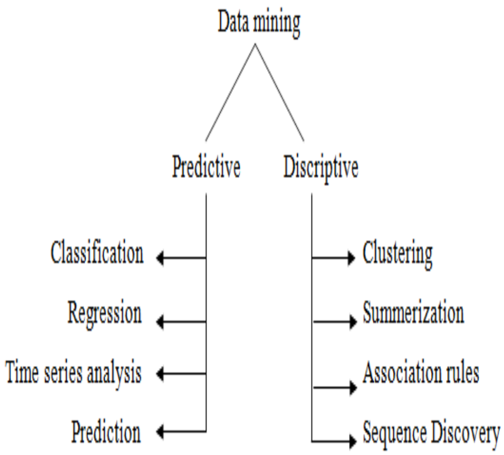


Fig 3.3 Data mining models and tasks

4. DATA MINING APPLICATIONS IN HEALTHCARE SECTOR

Healthcare industry today generates large amounts of complex data about patients, hospital resources, disease diagnosis, electronic patient records, medical devices etc. Larger amounts of data are a key resource to be processed and analyzed for knowledge extraction that enables support for cost-savings and decision making. Data mining applications in healthcare can be grouped as the evaluation into broad categories[1,10],

Treatment effectiveness

Data mining applications can develop to evaluate the effectiveness of medical treatments. Data mining can deliver an analysis of which course of action proves effective by comparing and contrasting causes, symptoms, and courses of treatments.

Healthcare management

Data mining applications can be developed to better identify and track chronic disease states and high-risk patients, design appropriate interventions, and reduce the number of hospital admissions and claims to aid healthcare management. Data mining used to analyze massive volumes of data and statistics to search for patterns that might indicate an attack by bio-terrorists.

Customer relationship management

Customer relationship management is a core approach to managing interactions between commercial organizations-typically banks and retailers-and their customers, it is no less important in a healthcare context. Customer interactions may occur through call centers, physicians' offices, billing departments, inpatient settings, and ambulatory care settings.

Fraud and abuse

Detect fraud and abuses establish norms and then identify unusual or abnormal patterns of claims by physicians, clinics, or others attempt in data mining applications. Data mining applications fraud and abuse applications can highlight inappropriate prescriptions or referrals and fraudulent insurance and medical claims.

Medical Device Industry

Healthcare system's one important point is medical device. For best communication work this one is mostly used. Mobile communications and low-cost of wireless bio-sensors have paved the way for development of mobile healthcare applications that supply a convenient, safe and constant way of monitoring of vital signs of patients[11]. Ubiquitous Data Stream Mining (UDM) techniques such as light weight, one-pass data stream mining algorithms can perform real-time analysis on-board small/mobile devices while considering available resources such as battery charge and available memory.

Pharmaceutical Industry

The technology is being used to help the pharmaceutical firms manage their inventories and to develop new product and services. A deep understanding of the knowledge hidden in the Pharma data is vital to a firm's competitive position and organizational decision-making.

Hospital Management

Organizations including modern hospitals are capable of generating and collecting a huge amount of data. Application of data mining to data stored in a hospital information system in which temporal behavior of global hospital activities is visualized[12]. Three layers of hospital management:

- ➢ Services for hospital management
- ➢ Services for medical staff
- ➢ Services for patients

System Biology

Biological databases contain a wide variety of data types, often with rich relational structure.Consequently multi-relational data mining techniques are frequently applied to biological data[13]. Systems biology is at least as demanding as, and perhaps more demanding than, the genomic challenge that has fired international science and gained public attention.

4. RESULTS OF COMPARATIVE STUDY

This chapter, a comparative study of data mining applications in healthcare sector by different researchers given in detail. Mainly data mining tools are used to predict the successful results from the data recorded on healthcare problems. Different data mining tools are used to predict the accuracy level in different healthcare problems. In this study, the following list of medical problems has been analyzed and evaluated.

- Heart Disease
- Cancer
- HIV/AIDS
- Blood
- Brain Cancer
- Tuberculosis
- Diabetes Mellitus
- Kidney dialysis
- Dengue
- IVF
- Hepatitis C

In the Table 1,the most important healthcare problems specifically in disease side and research results have been illustrated. The diseases are the most critical problems in human. To analyze the effectiveness of the data mining applications for diagnosing the disease, the traditional methods of mathematical / statistical applications are also given and compared. Listed eleven problems are taken for comparison with this work.

TABLE 1. DATA MINING APPLICATIONS IN HEALTHCARE

S.No	Type of disease	Data mining tool	Technique	Algorithm	Traditional Method	Accuracy level(%) from DM application
1	Heart Disease	ODND, NCC2	Classification	Naïve	Probability	60
2	Cancer	WEKA	Classification	Rules. Decision Table		97.77
3	HIV/AIDS	WEKA 3.6	Classification, Association Rule Mining	J48	Statistics	81.8
4	Blood Bank Sector	WEKA	Classification	J48		89.9
5	Brain Cancer	K-means Clustering	Clustering	MAFIA		85
6	Tuberculosis	WEKA	Naïve Bayes Classifier	KNN	Probability, Statistics	78
7	Diabetes Mellitus	ANN	Classification	C4.5 algorithm	Neural Network	82.6
8	Kidney dialysis	RST	Classification	Decision Making	Statistics	75.97
9	Dengue	SPSS Modeler		C5.0	Statistics	80
10	IVF	ANN, RST	Classification			91
11	Hepatitis C	SNP	Information Gain	Decision rule		73.20

Graph chart formed by using this table with the values of health care problems, Data Mining tools and Accuracy Level is as illustrated in Fig. 2. In this chart, the prediction accuracy level of different data mining applications has been compared.

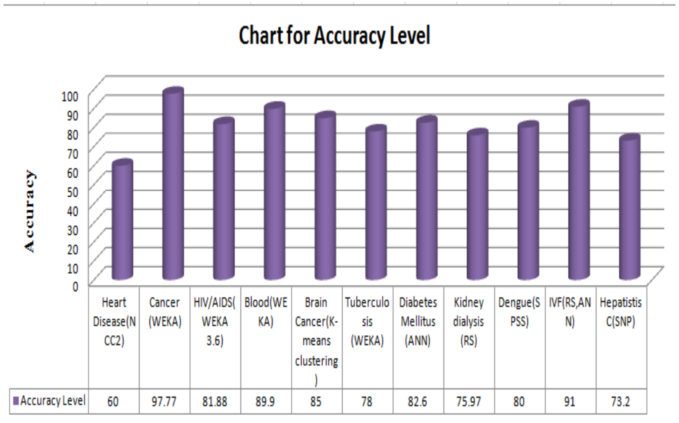


Fig. 4. Chart for Accuracy Level of using Data mining tools for diagnosis

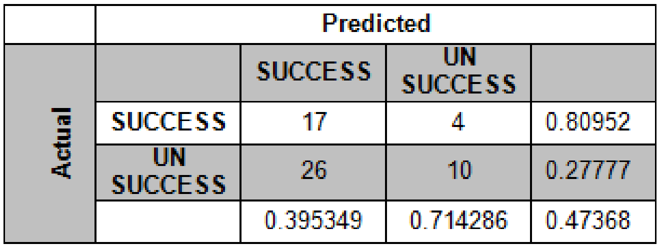
COMPARATIVE STUDY OF IVF SUCCESS RATE PREDICTION

The section deals with the comparative study of three different data mining application for predicting the success rate of IVF treatment. The process of data mining applications, its advantages and results obtained are compared. The detailed study of selected works gives a broad idea about the application of data mining techniques. This study mainly compares the three different data mining applications carried out on the prediction of the IVF treatment success rate.

a) Application of rough set theory for medical informatics data analysis

The research work aims to analyze the medical data by applying Rough Set Theory of data mining approach. The data reduction process has been done using rough set theory reduction algorithm. Rough set is mainly used to reduce the attributes without compromising its knowledge of the original. To analyze the fertilization data, ROSETTA tool kit reduction algorithm is used in this work to produce the optimal reduct set without affecting the original knowledge. The treatment success rate is predicted and tabulated as depicted in Table 2.

TABLE 2. IVF SUCCESS RATE PREDICTED BY ROUGH SET



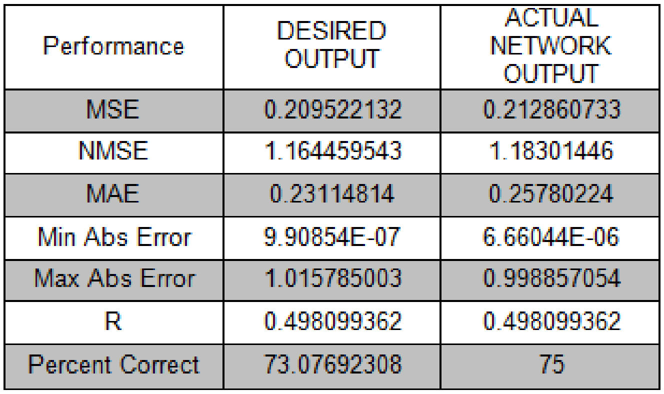
		Predicted		
Actual		SUCCESS	UN SUCCESS	
	SUCCESS	17	4	0.80952
	UN SUCCESS	26	10	0.27777
		0.395349	0.714286	0.47368

The actual and desired outputs are compared with each other. It also depicts that the success rate obtained after reducing the number of attributes is 47%.

b) Artificial neural network in classification and prediction

This research work is mainly aimed to predict and classify the IVF treatment results using Artificial Neural Network (ANN). The artificial neural network is constructed with multi-layer perception and back-propagation training algorithm, and constructed network is trained, tested and validated using patients' sample IVF data. This work finally compares the success rate between desired output which is field recorded data and actual output which is predicted output of neural network. In the Table 3, the comparison between desired and actual output of the neural network is illustrated.

TABLE 3. IVF SUCCESS RATE PREDICTED BY ANN



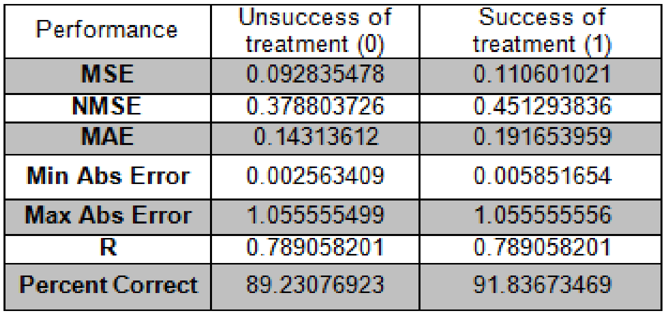
Performance	DESIRED OUTPUT	ACTUAL NETWORK OUTPUT
MSE	0.209522132	0.212860733
NMSE	1.164459543	1.18301446
MAE	0.23114814	0.25780224
Min Abs Error	9.90854E-07	6.66044E-06
Max Abs Error	1.015785003	0.998857054
R	0.498099362	0.498099362
Percent Correct	73.07692308	75

This work finds the actual output using patients' IVF data by applying Artificial Neural Network. By comparing success rate, desired and actual output, the result obtained has a prediction accuracy of 73%.

c) Modeling an integrated methodology of neural networks and rough sets for analyzing medical data

This work is mainly aimed to develop a combined prediction system for analyzing medical data using Artificial Neural Network and Rough Set Theory. Two kinds of rules Deterministic and Non-deterministic are effected in the application of Rough set tool. For the rough set application, the software tool Neuro solution is used to predict the result. The performance of the combined technique of Artificial neural network and rough set theory is described in the Table 4.

TABLE 4. PERFORMANCE OF IVF SUCCESS RATE PREDICTION USING HYBRID TECHNIQUE



Performance	Unsuccess of treatment (0)	Success of treatment (1)
MSE	0.092835478	0.110601021
NMSE	0.378803726	0.451293836
MAE	0.14313612	0.191653959
Min Abs Error	0.002563409	0.005851654
Max Abs Error	1.055555499	1.055555556
R	0.789058201	0.789058201
Percent Correct	89.23076923	91.83673469

The prediction accuracy of this hybrid approach of combined use of ANN and RST is around 90%. These comparison results of three different data mining applications for predicting the success rate of IVF treatments are shown in Table 5 and Fig. 5.

TABLE 5. COMPARISON BETWEEN THREE DIFFERENT DATA MINING APPLICATIONS

	Rough Set	ANN	Rough Set & ANN (Hybrid)
Percentage of Accuracy in Estimating Success	47	73	90

The application of combined Rough Set and Artificial Neural Network yields better result when compared with other techniques. It is observed that the hybrid technique of combined use of two or more machine learning tool yields better results than the use of a single technique for mining information from the database.

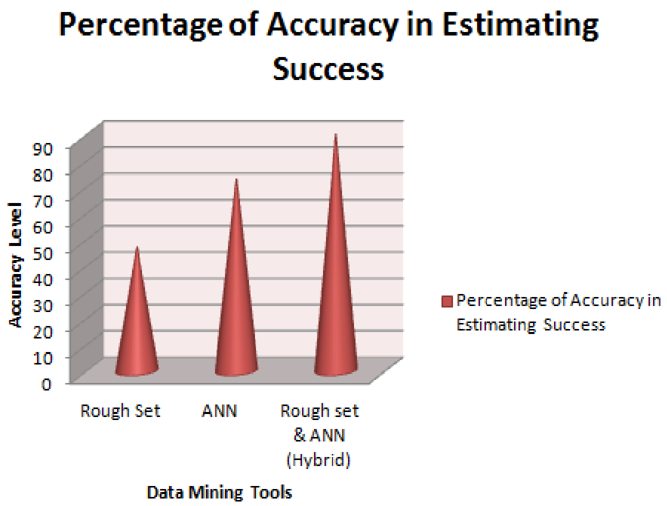


Fig. 5. The Success rate of Rough Set, ANN and Hybrid Technique

5. CONCLUSION

This paper aimed to compare the different data mining application in the healthcare sector for extracting useful information. The prediction of diseases using Data Mining applications is a challenging task but it drastically reduces the human effort and increases the diagnostic accuracy. Developing efficient data mining tools for an application could reduce the cost and time constraint in terms of human resources and expertise. Exploring knowledge from the medical data is such a risk task as the data found are noisy, irrelevant and massive too. In this scenario, data mining tools come in handy in exploring of knowledge of the medical data and it is quite interesting. It is observed from this study that a combination of more than one data mining techniques than a single technique for diagnosing or predicting diseases in healthcare sector could yield more promising results. The comparison study shows the interesting results that data mining techniques in all the health care applications give a more encouraging level of accuracy like 97.77% for cancer prediction and around 70 % for estimating the success rate of IVF treatment.

REFERENCES

[1]. HianChyeKoh and Gerald Tan, "Data Mining Applications in Healthcare", journal of Healthcare Information Management – Vol 19, No 2.

[2]. JayanthiRanjan, "Applications of data mining techniques in pharmaceutical industry", Journal of Theoretical and Applied Technology, (2007).

[3]. RubanD.Canlas Jr., MSIT., MBA , " Data mining in Healthcare: Current applications and issues".

[4]. K. Srinivas , B. Kavitha Rani and Dr. A. Govrdhan, "Applications of Data Mining Techniques in Healthcare and Prediction of Heart Attacks" International Journal on Computer Science and Engineering (2010).

[5]. ShwetaKharya, "Using Data Mining Techniques ForDiagnosis And Prognosis Of Cancer Disease", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), Vol.2, No.2, April 2012.

[6]. EliasLemuye, "Hiv Status Predictive Modeling Using Data Mining Technology".

[7]. Arvind Sharma and P.C. Gupta "Predicting the Number of Blood Donors through their Age and Blood Group by using Data Mining Tool" International Journal of Communication and Computer Technologies Volume 01 – No.6, Issue: 02 September 2012.

[8]. Arun K Punjari, "Data Mining Techniques", Universities (India) Press Private Limited, 2006.

[9]. Margaret H.Dunham, "Data Mining Introductory and Advanced Topics", Pearson Education (Singapore) Pte.Ltd.,India. 2005.

[10]. PrasannaDesikan, Kuo-Wei Hsu, JaideepSrivastava, "Data Mining For Healthcare Management", 2011SIAM International Conference on Data Mining, April, 2011.

[11]. Mobile Data Mining for Intelligent Healthcare Support

[12]. ShusakuTsumoto and Shoji Hirano, "Temporal Data Mining in Hospital Information Systems".

[13]. David Page and Mark Craven, "Biological Applications of MultiRelationalData Mining".

[14]. N. AdityaSundar, P. PushpaLatha and M. Rama Chandra, "Performance Analysis of Classification Data Mining Techniques Over Heart Disease Data Base", International Journal of Engineering Science & Advanced Technology, (2012).

[15]. HardikManiya, Mosin I. Hasan and Komal P. Patel, "Comparative study of Naïve Bayes Classifier and

KNN for Tuberculosis", International Conference on Web Services Computing (ICWSC) 2011 Proceedings published by International Journal of Computer Applications® (IJCA).

[16]. Andrew Kusiak , Bradley Dixonb and ShitalShaha, "Predicting survival time for kidney dialysis patients: a data mining approach", Computers in Biology and Medicine 35 (2005) 311–327.

[17]. B.Renuka Devi, Dr.K.NageswaraRao, Dr.S.PallamSetty and Dr.M.NagabhushanaRao," Disaster Prediction System Using IBM SPSS Data Mining Tool", International Journal of Engineering Trends and Technology (IJETT) - Volume4 Issue8- August 2013ISSN: 2231.

[18]. Leah Passmore, Julie Goodside, Lutz Hamel, LilianaGonzalez, T Ali Silberstein And James Trimarchi, "Assesing Decision Tree Models For Clinical In-Vitro Fertilization Data", Technical Report TR03-296

[19]. SaangyongUhmn, Dong-Hoi Kim , Jin Kim , Sung Won Cho and Jae Youn Cheong, "Chronic Hepatitis Classification using SNP data and Data Mining Techniques", Frontiers in the Convergence of Bioscience and Information Technologies 2007.

[20]. M.Durairaj, K.Meena, "A Hybrid Prediction System Using Rough Sets and Artificial Neural Networks", International Journal Of Innovative Technology & Creative Engineering (ISSN: 2045-8711) VOL.1 NO.7 JULY 2011.

[21]. R. Srinivasan, Health care in India – Vision 2020 Issues and Prospects.

AUTHOR'S PROFILE

1. He is currently working as a Assistant Professor, Dept. of Computer Science, Engineering &Technology, Bharathidasan University, Trichy, Tamilnadu, India. He completed his Ph.D. in Computer Science as a full time research scholar at Bharathidasan University on April, 2011. Prior to that, he received master degree (M.C.A.) in 1997 and bachelor degree (B.Sc. in Computer Science) in 1993 from Bharathidasan University. Prior to this, he was working at National Research Centre on Rapeseed-Mustard (Indian Council of Agricultural Research), Rajasthan, India, and at National Institute of Animal Nutrition and Physiology (ICAR), Bangalore as a Technical Officer (Computer Science) for 12 years. He has published 20 research papers in both national and international journals. His areas of interest include Artificial Neural Network, Soft Computing, Rough Set Theory and Data Mining.

2. She received the Master Degree (M.C.A) in 2012 from Anna University, Chennai and Bachelor degree (B.C.A) in 2009 from BharathidasanUniversity, Trichy. She is currently pursuing asM.Phil Research Scholar in the Department of Computer Science, Engineering and Technology at Bharathidasan University. Her areas of interest are Data Mining, Artificial Neural Network, Rough Set Theory and Network.