IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 11, November 2016

Data Mining Techniques for Diagnosing Diabetes and Hepatitis Disease

Ms. N. Gayathri [1], Ms. K. Yemuna Rane M.Sc., M. Phil., M.Sc (App. Psy) [2]

M.Phil Research Scholar, Dept of Computer Science, Kongunadu Arts and Science College, Coimbatore[1]

Assistant Professor, Dept of Computer Applications, Kongunadu Arts and Science College, Coimbatore[2]

Abstract: Data mining tools are giving successful result for disease diagnosis where it is one of the applications. Widely used in classification and prediction technology in the field of bioinformatics. In this paper diseases such as diabetes and hepatitis has been analyzed and compared using data mining applications. It also summarizes some techniques on medical field (diagnosis and prognosis). To develop disease estimation process using data mining technique it is focused on current research which is being carried out.

Keywords: Data mining techniques, Data mining applications, Hepatitis, Diabetes, Disease diagnosis, Classification, prediction.

I. INTRODUCTION

Data mining is the process of evaluating data from various surfaces and summarizing it into useful information. Data mining refers to separating or mining the knowledge from large amount of data. Data collection and storage technology has made it possible for organizations to collect huge amount of data at lower cost. Data mining sometimes called as data or KDD (knowledge discovery in databases). Data are any facts, numbers, or text that can be handled by a computer. The core purpose of data mining is applying various techniques to identify nuggets of information for decision making knowledge in bodies of data.

The field of data mining has been increased day by day in the areas of human life with different combinations and upgrading in the fields of statistics, databases, machine learning, pattern reorganization, artificial intelligence and computation capabilities etc. In future it is going to deliver more tedious scientific and research fields, social networking, medical diagnosis, web and business using cloud computing and multi-agent technologies. The advancement of data mining established when business data was first stored in computers and technologies were accomplished to allow users to operate through the data in real time. Data mining takes this execution process.

A. WHAT IS DATA MINING AND WHY USE DATA MINING?

- Data mining refers to the finding of applicable and efficient information.
- Data mining is the process of exploration and analysis, by automatic or semiautomatic means of large abundance of data in order to identify meaningful patterns and rules.
- Data mining is also about solving problems by analysing data already present in databases.
- According to Gartner Group "Data mining is the process of discovering meaningful new interrelationship patterns and trends by shifting huge amount of data stored in repositories using pattern recognition techniques as well as statistical and mathematical techniques".
- "Data mining is the analysis of large observational data sets to find unsuspected relationships and to summarize the data in novel ways that are both understandable and useful to the data owner".
- Data mining is to extract information from large amount of data base. There are two main reasons to use data mining as rapidly increase demands of data.

They are:

- Very high data and very low information.
- There is need to remove the useful information from the data and to explain the data.

B.ISSUES OF DATA MINING

- Security and social issues
- User interface issues
- Mining methodology issues
- Performance issues
- Data source issues

Security and social issues

Today, security is a serious controversy with any data collection that is contributed and is expected to be used for vital decision-making. Hefty amounts of emotional and personal information about self or companies are collected and stored when data is collected for customer profiling, user performance understanding, interacting personal data with other information, etc., due to the value of data, database content are sold and because of that some important information could withheld while other

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 11, November 2016

information could be distributed widely and used without control.

User interface issues
Good data visualization helps users to better understand their needs and reduce the interpretation of the data mining results. Data exploratory tasks are powerfully eased by the capacity to see data in proper visual presentation. There are hefty of visualization ideas and suggestions for effective data graphical presentation. In order to obtain good visualization tools for big datasets that could be used to display and manipulate mined knowledge.
Major issues related to user interface and visualization is:

- "screen real-estate"
- Information rendering
- Interaction

Mining methodology issues
Mining different kinds of knowledge in databases to cover a broad range of knowledge. Interactive mining of knowledge at multiple levels of abstraction will allows the user to focus the search for patterns, contributing and clarifying data mining requests based on the returned results. Background knowledge can be used to guide and express the discovery process and discovered patterns.

Performance issues
In order to effectively remove the information from hefty amount of data in databases data mining algorithm must be good and scalable. Huge size of databases, wide distribution of data, and elaboration of data mining methods persuade the development of parallel and distributed data mining algorithms. The incremental algorithms update databases without mining the data again from scratch.

Data source issues
Different types of data are storing in a different repositories so it is very tedious to predict a data mining system to acquire good mining results in all types of data sources.

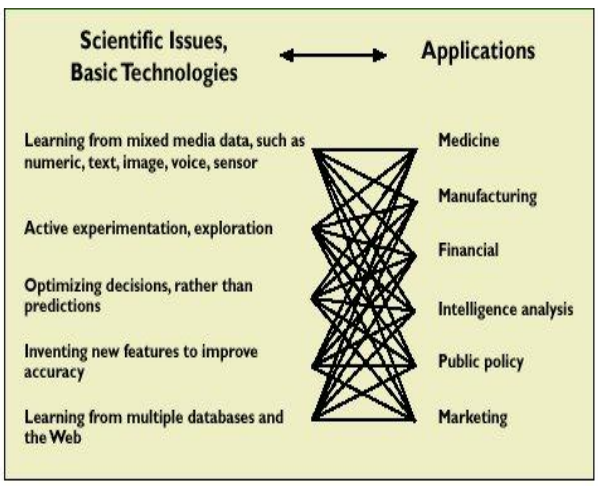

Fig 1: Applications of data mining

Various types of data and sources may lack different algorithms and methodologies. For all kinds of data accomplished data mining tool will not be pragmatic.

II. DATA MINING IN MEDICAL SECTOR

Health care industries are using data mining to analyze huge data of medical research, patient, staff and doctors records, biotech and medicines and compound pharmaceutical industries. Thus it enables the discovery of relationship between diseases, research of new drugs, treatments effectiveness, and genetic network analysis and market activities in drug transport services.

Health care industries now a day's produces hefty amount of tedious data about patients, hospital resources, disease diagnosis, electronic patient records, medical devices etc. Heavy amounts of data are a key resource to be processed and analyzed for knowledge extraction that enables platform for cost-savings and decision making.

A. HEALTHCARE IN BROAD CATEGORIES

- Treatment value
- Healthcare organization
- Patron association management
- Scam and exploitation
- Medical gadget industry
- Pharmaceutical production
- Hospital administration

Treatment value
Data mining applications can evolve to estimate the efficiency of medical treatments. Data mining can produce an analysis of which course of action proves efficiency by comparing and contrasting causes, symptoms, and courses of treatments.

Healthcare organization
Data mining applications is used to identify and track chronic disease states and high-risk patients. Reduce the number of hospital admittance and pretends to aid healthcare management. Data mining used to analyse enormous contents of data and statistics to search for patterns that might intimate an attack by bio-terrorists.

Patron association management
It is a mass access to handling the communication between economic organizations-typically banks and retailers-and their customers.

Scam and exploitation
Scam and exploitation establishes norms and then identifies rare and unexpected patterns of claims by physicians, clinics, or others attempt in data mining applications. Fraud and abuse applications identify and emphasize the inapplicable prescriptions or referral and fraudulence insurance and medical claims.

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 11, November 2016

Medical gadget industry
Medical device is one of the important points in healthcare system. It is used only for communication work. Mobile healthcare applications playing a vital role in day to day life.

It is very convenient and safe for monitoring important symptoms of patients.

Pharmaceutical production
Data mining techniques are very helpful for the pharmaceutical firms to handle their inventions and to build up the new products and services.

Pharmacy data hidden knowledge is important for decision making in the organizations.

Hospital administration
In large hospitals health care information systems are introduced to store the records of patients, laboratory, clinical and radiological actions. Storing data using data mining is not only for decision making also for hospital management.

Three layers of hospital management:

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

B. USING DATA MINING ANALYSING MEDICAL PROBLEMS

Data mining tools are used to identify the good results from the data report on healthcare problems. Different mining tools are used to identify the accuracy level in different healthcare problems. The following medical problems are:

- Heart disease
- Cancer
- HIV/AIDS
- Tuberculosis
- Diabetes Mellitus
- Kidney dialysis
- Dengue
- IVF
- Hepatitis C

C. DATA MINING APPLICATIONS IN HEALTHCARE

The diseases are the most dangerous problems in human. To analyse data mining applications for diagnosing the disease, mathematical / statistical applications are also given and compared. Eleven problems are taken for comparison with this work.

TABLE 1 COMPARISON BETWEEN DISEASES WITH DIFFERENT DATA MINING TECHNIQUES

S.no	Type of disease	Data mining tool	Technique	Algorithm	Accuracy level(%) from DM application
1	Heart disease	ODND,NCC2	Classification	Naïve	60
2	Cancer	WEKA	Classification		97.77
3	HIV/AIDS	WEKA 3.6	Classification Association Rule mining	Rule decision table	81.8
4	Blood Bank sector	WEKA	Classification	J48	89.9
5	Brain cancer	K-means clustering	Clustering	J48	85
6	Tuberculosis	WEKA	Naïve Bayes classifier	MAFIA	78
7	Diabetes mellitus	ANN	classification	KNN	82.6
8	Kidney dialysis	RST	classification	C4.5 algorithm	75.97
9	Dengue	SPSS Modeler		Decision making	80
10	IVF	ANN,RST	classification	C5.0	91
11	Hepatitis C	SNP	Information gain	Decision rule	73.20

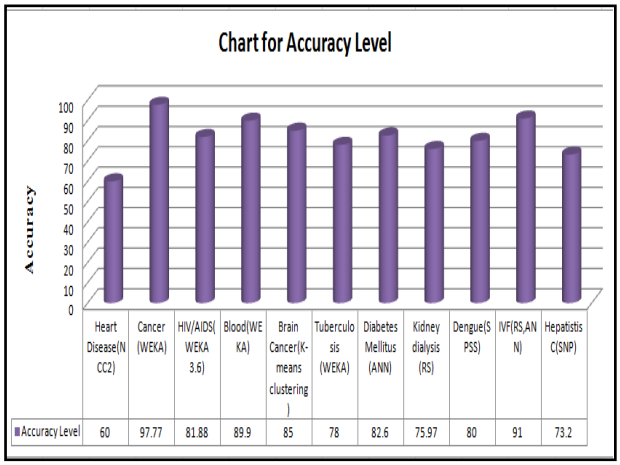

CHART 1: ACCURACY LEVEL OF VARIOUS DISEASES

III.COMPARATIVE STUDY OF DIABETES MELLITUS AND HEPATITIS C

A. DIABETES MELLITUS

Diabetes is known as diabetes mellitus comes from Greek word “siphon”. It is a metabolic disease where the person will affect by high blood glucose (blood sugar). It is either because insulin production is inadequate or because the body cells won’t respond properly to insulin or both. If a patient is affected by high blood sugar will experience polyuria (frequent urination) then it will increases polydipsia (thirsty) and polyphagia (hungry).

1. TYPES OF DIABETES

- Type 1 diabetes
- Type 2 diabetes
- Gestational diabetes

DOI 10.17148/IJARCCE.2016.51135

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 11, November 2016

Type 1 diabetes
Type 1 diabetes is referred to as insulin-dependent diabetes also it is used to be called as juvenile diabetes or early onset diabetes. Mainly begins in childhood. It is an autoimmune condition so it attacks the pancreas with anti-bodies. The damaged pancreas won't produce insulin. This type may be caused by genetic predisposition and results in faulty beta cells in the pancreas that normally produce insulin.

It damages

- Eyes (diabetic retinopathy)
- Nerves(diabetic neuropathy)
- Kidneys(diabetic nephropathy)

If it is more serious the patients will have heart disease and stroke.

Type 2 diabetes
Type 2 diabetes is referred to as non-insulin – independent diabetes and it is also used to be called as adult-onset diabetes.95% of diabetes cases the adults with the epidemic of obese and overweight kids. Pancreas produces insulin but the amount of producing is not enough for the body need. Type 2 also causes health problems such as eyes, nerves, kidneys if more risk it will increases to have heart disease and stroke. Diabetes can't be cure but it can be controlled by weight management, nutrition and exercise.

Gestational diabetes
Gestational diabetes is referred to as gestational diabetes mellitus which occurs during the pregnancy of women. It is caused when insulin receptors do not function properly. Only few symptoms are there and it is diagnosed by screening. Undiagnosed or uncontrolled gestational diabetes can increase the risk of complications during child birth.

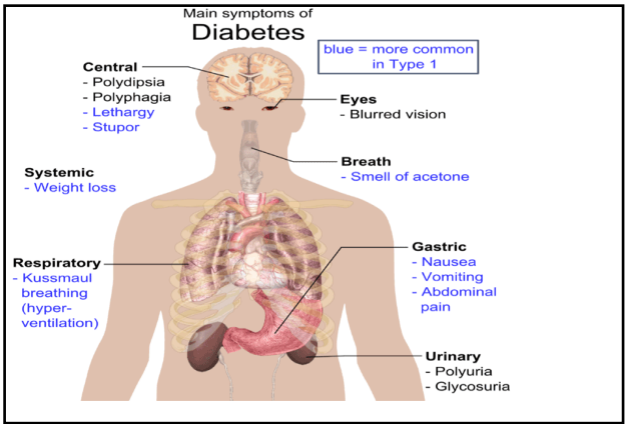

Fig 2: Symptoms of Diabetes

2. RISK FACTORS OF DIABETES IN INDIA ARE

- Age
- Family hereditary
- Central obesity
- Physical inactivity and sedentary living
- Insulin resistance

Age
Indian peoples having diabetes at very young age at least in the age of 10-15 years. Due to the early occurrence it will take more than enough time to develop the chronic complications of diabetes. The life span of Indian's were increased but number of peoples are affected by diabetes are detected.

Family heredity
Diabetes increases due to family heredity problem. More than 50 percent the child are affected by diabetes through parent heredity. Diabetes is the high genetic risk for Indians which is observed in Asian Indians which then spread to other countries

Central obesity
Weight range, weight gain increases the risk of diabetes. Extra body fat within the abdomen has an increased risk of diabetes. Indians waist circumference should be 90 cm for males and 80 cm for females.

Physical inactivity and sedentary living

- Physical inactivity may increase the risks of certain cancers.
- Physical inactivity may contribute to distress and sadness.
- Physical inactivity has been shown to be a risk factor for certain cardiovascular diseases.
- People who engage in many physical activities are less likely to develop coronary heart disease.
- People who are further active are less likely to be overweight or obese.
- Sitting too much may cause a decrease in skeletal muscle mass.
- Physical inactivity is associated to high blood pressure and elevated cholesterol levels.

Insulin resistance
Insulin is a hormone that is produced by the beta cells of the pancreas. The insulin produced is released into the blood stream and spreads throughout the body. Insulin is an important hormone that has many actions within the body. Most actions of insulin are heading for metabolism (control) of carbohydrates (sugars and starches), lipids (fats), and proteins. Insulin is critical for the body's use of glucose as energy.

It consists of:

- Abnormal fats
- High blood pressure
- Obesity
- Abnormal glucose level

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 11, November 2016

TABLE 2 DATA SET FOR DATA MINING TECHNOLOGIES

Author (Year)	Study Purpose	Group/Topic of Research	Diabetes Type	Data set	Data-Mining Methods	Software	Outcome
Bellazzi & Abu-Hanna, 2009[7]	Patient need	Interpretation and prediction of BGL	N/A	Blood glucose home-monitoring data, ICU blood glucose data	Association/Temporal abstraction, Classification/ Subgroup discovery	N/A	Trends and daily cycles of BGL, predict high levels of BGL
Bellazzi *et al.*, 1998[8]	Patient need	Interpretation of BGL	N/A	Blood glucose home-monitoring data	Association/Temporal abstraction	N/A	Trends and daily cycles of BGL
Breault *et al.*, 2002[9]	Science research	Prediction of BGL	N/A	15,902 patients with diabetes	Classification/ CART	CART software by Salford Systems	Best predictor and rules to predict glycemic control
Brown *et al.*, 2005[10]	Science research	Genomic data analysis	T2DM	LocusLink database	Clustering	ExQuest	Candidate genes that contribute to diabesity
Concaro *et al.*, 2009[11]	Science research	Healthcare flow	N/A	101,339 health care events	Association/Temporal abstraction	N/A	Temporal association rules on sequence

B.HEPATITIS

The word hepatitis comes from the ancient Greek word hepar meaning 'liver' Hepatitis refers to an injury to a liver with inflammation of the liver cells. This is a self-limiting or can progress to fibrosis, cirrhosis or liver cancer. There are 5 types of hepatitis virus they are

- Hepatitis A virus (HAV)
- Hepatitis B virus (HBV)
- Hepatitis C virus (HCV)
- Hepatitis D virus (HDV)
- Hepatitis E virus (HEV)

Hepatitis A virus

HAV caused through infected food or water which is affected by a virus. HAV patients will get full recovery they won't lead to chronic disease.

Hepatitis B virus

HBV is a sexually transmitted disease. It spreads by contact with infected blood, semen, and some other body fluids.

Hepatitis C virus

HCV spread through direct contact with blood of a person who is infected by the disease. Due to this liver will get damaged and can swell. In this type only 20% of hepatitis C patient will get cirrhosis.

Hepatitis D virus

HDV will affect only when the person already infected with hepatitis B. infected through:

- Infected blood
- Unprotected sex
- Perforation of the skin with infected needles

Hepatitis E virus

HEV caused through drinking water and the liver starts to swell but not for long-term.

1. HEPATITIS AND LIVER

Largest gland in the human body is liver.3 lb(1.36kg) is the approximate weight. It is reddish brown in color and separated into 4 lobes of different sizes and lengths. Largest internal organ is below the diaphragm on the right in the thoracic region of the abdomen. Through the hepatic artery and the portal vein blood reaches the liver. While the hepatic artery carries oxygen-rich blood from the aorta the portal vein carries blood containing digested food from the small intestine. Liver is made up of thousands of lobules and each lobules consist of many hepatic cells.

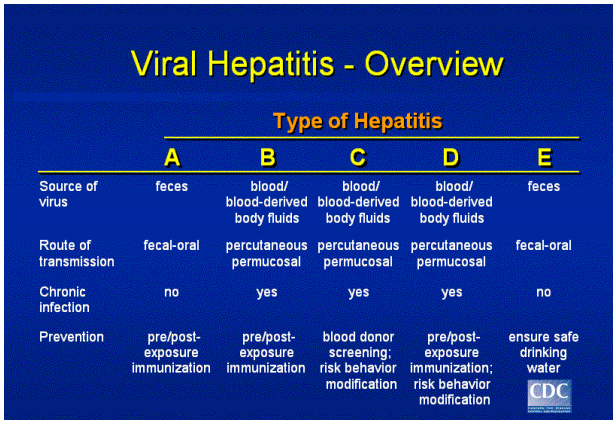

Fig 3**:** Over View of Hepatitis Virus

2. HEPATITIS SYMPTOMS

- jaundice (a yellowing of the skin and eyes)
- abdominal pain
- loss of appetite
- nausea and vomiting
- diarrhoea
- fever

TABLE 3 VARIOUS TECHNIQUES USED FOR HEPATITIS VIRUS

S.No	Researcher	Publication	Technique	Performances
1	Yilmaz Kaya	Elsevier - 2013	RS-ELM [Rough set - Extreme Learning Machine]	The classification accuracy was 96.49% using RS-ELM model.
2	Javad Salimi Sartakhti	Elsevier 2011	Using support vector machine (SVM) and simulated annealing (SA).	Obtained classification accuracy of 96.25%
3	Duygu et. Al.	Elsevier - 2011	Compared with LS-SVM classifiers and PCA-LSSVM.	Accuracy is 96.12%
4	G. SathyaDevi	IEEE 2011	Use of decision tree C4.5 algorithm, ID3 algorithm and CART algorithm	CART should the accuracy rate of 83.2%
5	A.H.Roslina	IEEE 2010	SVM and Wrapper method to remove noise feature before classification. [Used without feature selection & with feature selection]	Accuracy rate is 74.55%
6	Fadl Mutaher	IJSER 2013	Using Classification algorithms like Naive Bayes, PT Tree, KStar, J4.8	Naive Bayes showed the accuracy of 96.52%

DOI 10.17148/IJARCCE.2016.51135

IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 5, Issue 11, November 2016

TABLE 4 RISK FACTORS OF HEPATITIS

Viral factors	Host factors	Environmental factors
High viral load	Advanced age	Aflatoxin exposure
Genotype (HBV-C > HBV-B; HBV-D > HBV-A)	Male gender	Alcohol consumption
Basal core promoter mutation	Genetic alterations	Cigarette smoking
Pre-S deletion	Family history of HCC	Concurrent infection with hepatitis C or
	Ethnicity (Asian > Caucasian)	D virus or with HIV
		Diabetes mellitus
		Obesity
		Metabolic syndrome

3. TECHNIQUES USED FOR HEPATITIS

- Classification
- Support vector machine
- Naive Bayesian

Classification

Classification is a dependent variable and is used to classify data into predefined class labels. Classification algorithm is used to identify hepatitis and predicts based on the symptoms and health condition. It consists of two processes training and testing.

- **Training**

By examining trained data it builds a classification model.

- **Testing**

Using text data it analyzes the classifier for accurate trained data.

Classification algorithms are:

- Naive Bayes
- FT Tree
- KStar
- J48
- Neural network

Support vector machine

SVM has multilayer perceptrons and radial- basis function networks. Mainly it is used for pattern classification. SVM algorithm is the inner-product kernel between support vector and the vector from inside space. It is a small subset of training data extracted by the algorithm. SVM algorithm used to construct three types of learning machines such as

- Polynomial learning machines,
- Radial-basis function networks,
- Two-Layer perceptrons.

Naive bayesian

A Naive Bayesian is a probabilistic statistical classifier. The "naive" reduces complexity to a simple multiplication of probabilities. Advantage of the Naive Bayesian classifier is its rapidity of use. It can handle a data set with many attributes due to its simplicity. To develop proper parameter it needs only small set of training data.

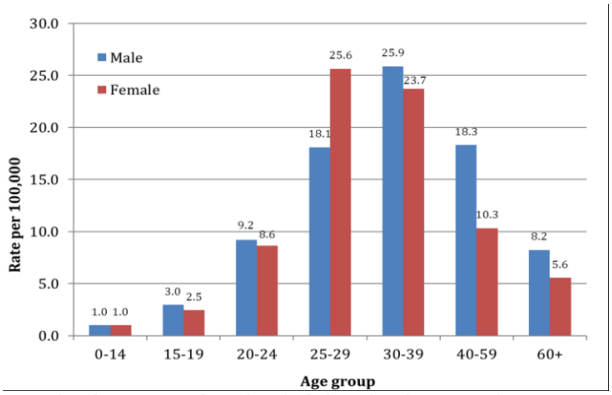

CHART 2: VARIOUS AGE GROUP PEOPLE AFFECTED BY HEPATITIS

IV. CONCLUSION

This paper summarizes the different classifications of medial sector and the comparison of data mining application in healthcare sector. Data mining application is the tedious and challenging task to examine the diseases. Reduce the human effort. By applying data mining techniques in the diagnosis of diabetes and hepatitis disease shown promising result, so by applying this technique in choosing the appropriate treatment for diabetes and hepatitis patients need further investigation.

REFERENCES

[1] Neelamadhab Padhy, Dr. Pragnyaban Mishra and Rasmita Panigrahi, "The Survey of Data Mining Applications and Feature Scope, International Journal of Computer Science, Engineering and Information Technology (IJCSEIT)", vol.2, no.3, June

[2] Xingquan Zhu, Ian Davidson, "Knowledge Discovery and Data Mining:Challenges and Realities", ISBN 978-1-59904-252, Hershey, New York, 2007.

[3] Joseph, Zernik, "Data Mining as a Civic Duty – Online Public Prisoners Registration Systems", International Journal on Social Media: Monitoring, Measurement, Mining, vol. - 1, no.-1, pp. 84-96, September2010.

[4] Heikki, Mannila, "Data mining: machine learning, statistics and databases," IEEE, 1996.

[5] Fayadd, U., Piatesky -Shapiro, G., and Smyth, P, From Data Mining To Knowledge Discovery in Databases", The MIT Press, ISBN 0–26256097–6, Fayap, 1996.

[6] Jing He, Advances in Data Mining: History and Future, Third international Symposium on Information Technology Application, 978-0-7695- 3859-4, IEEE, 2009.

[7] Bellazzi R, Abu-Hanna A. Data mining technologies for blood glucose and diabetes management. J Diabetes Sci Technol. 2009;3(3):603–612. [PMC free article] [PubMed]

[8] Bellazzi R, Magni P, Larizza C, De Nicolao G, Riva A, Stefanelli M. Mining biomedical time series by combining structural analysis and temporal abstractions. Proc AMIA Symp. 1998:160–164. [PMC free article] [PubMed]

[9] Breault JL, Goodall CR, Fos PJ. "Data mining a diabetic datawarehouse." Artif Intell Med. 2002;26(1-2):37–54. [PubMed]

[10] Brown AC, Olver WI, Donnelly CJ, May ME, Naggert JK, Shaffer DJ, Roopenian DC. Searching QTL by gene expression: analysis of diabesity. BMC Genet. 2005;6:12. [PMC free article] [PubMed]

[11] Concaro S, Sacchi L, Cerra C, Bellazzi R. "Mining administrative and clinical diabetes data with temporal association rules". Stud Health Technol Inform. 2009; 150:574–578. [PubMed]