A survey on Data Mining approaches for Healthcare

S. Nivetha[#1], Dr. M. Shanthakumar[*2]
[#]Ph.D Research scholar,
Assistant professor in Computer Science,
Kamban College of Arts and Science,
Sulthanpet.

[*] Assistant professor in Computer Science,
Kamban College of Arts and Science,
Sulthanpet.

***Abstract*— Data Mining is one of the most vital and motivating area of research with the objective of finding meaningful information from huge data sets. In present era, Data Mining is becoming popular in healthcare field because there is a need of efficient analytical methodology for detecting unknown and valuable information in health data. Data mining (DM) tools provide a useful for alternative framework that addresses many problems. Diabetes mellitus (DM), commonly known as diabetes, is a chronic and one of the dramatically increasing metabolic diseases in the world. Data mining is a prominent tool set in medical databases. This promising approach improves sensitivity and/or specificity of disease detection and diagnosis by opening a window of comparatively better resources. Due to the growing unstructured nature of diabetic data form health industry or all other sources, it is necessary to structure and emphasis its size into nominal value with possible solution. With the help of technological developments, it is necessary to combine robust diabetic data sharing and electronic communication systems can facilitate better access to health services at all the levels of patients. This paper surveys the Data Mining approaches for Healthcare, especially for diabetics.**

***Keywords*— Chronic, Diabetes mellitus, Diseases, Metabolic and Mining.**

I. Introduction

Data mining is defined as a non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data. The term process implies that data mining consists of many steps, nontrivial means process is not straight forward and some search or inference is involved. The term pattern is an expression in some language describing a subset of data, or, in general making any high level description of a set of data. Pattern should be novel and potentially useful, that is, it should lead to some benefits to the user or task. Ultimately, pattern should be understandable, if not immediately then at a later stage after some post processing [1].

Data mining is used for a variety of purposes in both the private and public sectors [2]. Industries such as banking, insurance, medicine, and retailing commonly use data mining to reduce costs, enhance research, and increase sales. For instance, the insurance and banking industries use data mining applications to detect fraud and assist in risk assessment (e.g., credit scoring). Using customer data collected over several years, companies can develop models that predict whether a customer is a good credit risk, or whether an accident claim may be fraudulent and should be investigated more closely. Data mining techniques are used in a many research areas, including mathematics, cybernetics, genetics, and marketing.

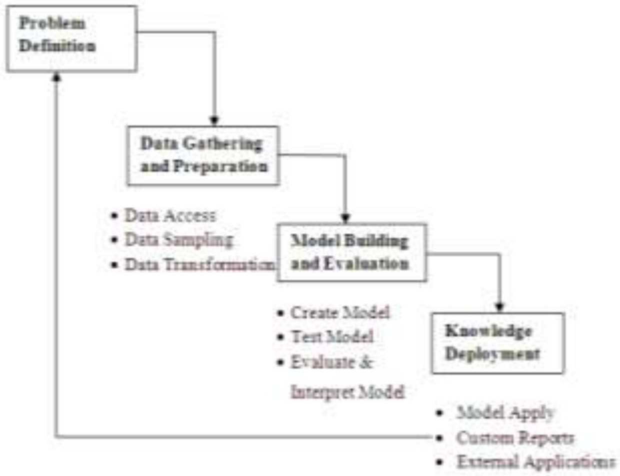

Fig. 1: Datamining Process

The Figure 1.1 illustrates the phases, and the iterative nature, of a data-mining process. The process flow shows that a data mining process does not stop when a particular solution is deployed [3]. The results of data mining trigger new business questions, which in turn can be used to develop more focused models. Section II describes the Analysis and Study of Diabetes using Data mining [4], section III presents the Data-Mining Technologies for Diabetes, section IV discusses on the related work, and finally section V Concludes the paper.

II. Analysis and Study of Diabetes using Data mining

Diabetes is a severe metabolic disorder marked by high blood glucose level [5], excessive urination, and persistent thirst, caused by lack of insulin actions. There are usually three forms of diabetes—Type 1, Type 2, and gestational. It is believed that diabetes is a particularly opportune disease for data mining technology for a number of reasons [6].

Diabetes is considered as one of the deadliest and chronic diseases which causes an increase in blood sugar. Many complications occur if diabetes remains untreated and unidentified. The tedious identifying process results in visiting of a patient to a diagnostic center and consulting doctor. But the rise in machine learning approaches solves this critical problem[8].

Diabetes is one of the main topics for medical research due to the longevity of the diabetes and the huge cost on the health care providers. Early detecting of diabetes ultimately reduces cost on health care providers for treating diabetic patients, but it is a challenging task [6] [7]. For early detecting of diabetes, researchers can take advantage of the patient's health care data to convert raw data into meaningful information and extract hidden knowledge by applying data mining to construct an intelligent predictive model.

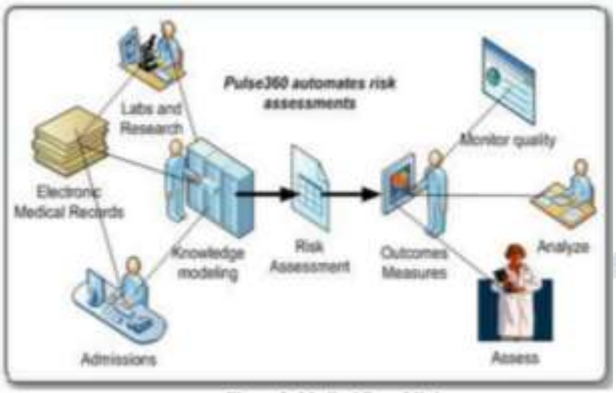

Fig. 2: Outline of Medical Datamining

A. *Disease Prediction System-Prevention and diagnosis*

Data mining technique made prediction system plays a vital role in strategy preparation for prevention of communicable as well as non communicable diseases in located area. Lifestyle related diseases like hypertension, diabetes mellitus, cardiovascular diseases; stroke etc can be easily and accurately classified and possible to locate their etiological area cluster patterns. These techniques are also useful in disease diagnosis. Ms. Is take et al. developed a prototype Intelligent Heart Disease Prediction System (IHDPS) using three data mining modeling techniques, namely, Decision Trees, Naïve Bays and Neural Network. IHDPS can discover and extract hidden knowledge (patterns and relationships) associated with heart disease from a historical heart disease database. It can answer complex queries for diagnosing heart disease and thus assist healthcare practitioners to make intelligent clinical decisions which traditional decision support systems cannot. By providing effective treatments, it also helps to reduce treatment costs. To enhance visualization and ease of interpretation, it displays the results both in tabular and graphical forms [6].

B. *Workout of treatment plan*

The data mining techniques play an important role in treatment plan workout, surgical procedures, rehabilitation, chronic diseases management plan etc. Long term follow up plan may be easily guided and keen supervision is possible. Example, a patient of hypertension can be long term manage and back through record of number of patients will guide in implementing future strategies [6].

C. *Reduction of cost of patient management*

These systems may definitely helpful in reduction of cost of patient management by avoiding unnecessary investigations and patients follow up. These prediction systems will add accuracy and time management.

D. *Discovery of hidden etiological factors*

This is most exciting objective planned by using these data mining systems and its various methods. This will be helpful for confirmation of geographical variations. Most of our health strategies are planned on the basis of data interpretations from developed countries. We can formulate our own systems and can avoid geographical errors. Computer-based patient support systems benefit patients by providing informational support that increases their participation in health care [6].

III. Data-Mining Technologies for Diabetes

The remarkable advances in biotechnology and health sciences have led to a significant production of data, such as high throughput genetic data and clinical information, generated from large Electronic Health Records (EHRs) [8] [9]. To this end, application of machine learning and data mining methods in biosciences is presently, more than ever before, vital and indispensable in efforts to transform intelligently all available information into valuable knowledge. Diabetes mellitus (DM) is defined as a group of metabolic disorders exerting significant pressure on human health worldwide. Extensive research in all aspects of diabetes (diagnosis, etiopathophysiology, therapy, etc.) has led to the generation of huge amounts of data. The following table depicts the Study of Data Mining Technologies for Diabetes [8][9].

TABLE I
Data Mining Technologies for Diabetes – A Study

Author (Year)	**Topic of Research**	**Diabetes Type**	**Data set**	**Data-Mining Methods**
Bellazzi & Abu-Hanna, 2009	Interpretation and prediction of BGL	N/A	Blood glucose home-monitoring data, ICU blood glucose data	Association/ Temporal abstraction, Classification/ Subgroup discovery
Bellazzi *et al.*, 1998	Interpretation of BGL	N/A	Blood glucose home-monitoring data	Association/Temporal abstraction
Breault *et al.*, 2002	Prediction of BGL	N/A	15,902 patients with diabetes	Classification/ CART
Brown *et al.*, 2005	Genomic data analysis	T2DM	LocusLink database	Clustering
Concaro *et al.*, 2009	Healthcare flow	N/A	101,339 health care events	Association/Temporal abstraction
Covani *et al.*, 2009	Genomic data analysis	T2DM	Gene list associated with T2DM, periodontitis, and sinusitis	Clustering/ Hierarchical, k-means
Duhamel *et al.*, 2003	Data preprocessing/cleaning	T2DM	23,601 records of T2DM patients	Clustering/ k-means, Classification/ Decision tree
DuMouch	Adverse	N/A	2.4 million	Classification/

el *et al.*, 2008	drug effect		reports from FDA AERS database	Proportional Reporting Ratio, Bayes Multi-Item Gamma Poison Shrinker, Logistic regression
Gerling *et al*, 2006	Genomic data analysis	T1DM	2D gel proteome data	Clustering/k-means, principal component analysis
Huang *et al*, 2007	Feature selection	T2DM	2064 patient information: 1148 male, 916 female	Classification/ Naïve Bayes, IB1, Decision tree—C4.5
Liou *et al*, 2008	Insurance-fraud detection	N/A	Taiwan's national health insurance database	Classification/ Neural Network, Classification Tree
Miyaki *et al*, 2002	Feature selection	T2DM	165 patient's records	Classification/ CART
Richards *et al*, 2001	Prediction of early mortality	N/A	21,000 patient's clinical records	Classification/ Simulated annealing
Sigurdardottir *et al*, 2007	Feature selection	T2DM	21 articles from Medline, Scopus, and CINAHL	Classification/ Decision tree—C4.5
Toussi *et al*, 2009	Clinical guideline	T2DM	Patient records with missing or incomplete rules in guideline	Classification/ Decision tree—C5.0
Wright *et al*, 2005	Feature selection	T1DM, T2DM	MQIC data warehouse: 50,428 records extracted for this study from 3.6 million records	Classification/ Reconstructibility analysis (RA)
Yamaguchi *et al*, 2006	Predict BGL	T1DM	FBG, metabolic rate, food intake, and physical condition	Classification

IV. Related Work

Diabetes is one of the most deadly, disabling, and costly diseases observed in many of the nations at present, and the disease continues to be on the rise at an alarming rate . The data mining techniques play an important role in treatment plan workout, surgical procedures, rehabilitation, chronic diseases management plan, etc. In this section, we have made a survey and presented below.

H.Vignesh Ramamoorthy [10] and A. Manjula proposed that Data Mining is a broad category of applications and technologies or gathering, storing, analyzing and help to make decisions. Recent Data Mining research has many classification and prediction methods have been proposed by researches in machine learning, pattern recognition, and statistics and medical diagnosis. This proposal entitled ―Diabetes Forecasting using Modified RBF Neural Networks‖ is used to predict the diabetes for the patients. This technique is to find out the information which is hidden in the dataset. Modified Radial basis Functional (MRBF) Neural Networks is the Data Mining technique used to predict the diabetes disease. In this proposal, Modified RBF Neural Networks is a Data Mining technique based classification model for classifying diabetic patients. For achieving better results, genetic algorithm is used for feature selection. This model is trained with Back Propagation algorithm and Genetic Algorithm and classification accuracies are compared. The proposed approaches are evaluated by the Pima Indian Diabetes data sets. The Pima Indian Diabetes data set is a data mining dataset.

Joseph L.[11] et.al., proposed that Diabetes is a major health problem in the United States. There is a long history of diabetic registries and databases with systematically collected patient information. We examine one such diabetic data warehouse, showing a method of applying data mining techniques, and some of the data issues, analysis problems, and results. The diabetic data warehouse is from a large integrated health care system in the New Orleans area with 30,383 diabetic patients. Methods for translating a complex relational database with time series and sequencing information to a flat file suitable for data mining are challenging. This work discusses two variables in detail, a comorbidity index and the HgbA1c, a measure of glycemic control related to outcomes.

Dr Saravana kumar N M [12], et.al, Diabetic Mellitus (DM) is one of the Non Communicable Diseases (NCD), is a major health hazard in developing countries such as India. The acute nature of DM is associated with long term complications and numerous of health disorders. This paper, uses the predictive analysis algorithm in Hadoop/Map Reduce environment to predict the diabetes types prevalent, complications associated with it and the type of treatment to be provided. Based on the analysis, this system provides an efficient way to cure and care the patients with better outcomes like affordability and availability.

Ioannis Kavakiotis [13], et.al, study is to conduct a systematic review of the applications of machine learning, data mining techniques and tools in the field of diabetes research with respect to a) Prediction and Diagnosis, b) Diabetic Complications, c) Genetic Background and Environment, and e) Health Care and Management with the first category appearing to be the most popular. A wide range of machine learning algorithms were employed. In general, 85% of those used were characterized by supervised learning approaches and 15% by unsupervised ones, and more specifically, association rules. Support vector machines (SVM) arise as the most successful and widely used algorithm. Concerning the type of data, clinical datasets were mainly used. The title applications in the selected articles project the usefulness of extracting valuable knowledge leading to new hypotheses targeting deeper understanding and further investigation in DM.

Miroslav Marinov [8], et.al., discussed that The objective of this study is to conduct a systematic review of applications of data-mining techniques in the field of diabetes research. The authors searched the MEDLINE database through PubMed. They initially identified 31 articles by the search, and selected 17 articles representing various data-mining methods used for diabetes research. The main interest was to identify research goals, diabetes types, data sets, data-mining methods, data-mining software and technologies, and outcomes.

Thomas Porter [14], et.al., discussed that Mounting amounts of data made traditional data analysis methods impractical. Data mining (DM) tools provide a useful for alternative framework that addresses this problem. This study follows a DM technique to identify diabetic patients. They have developed a model that clusters diabetes patients of a large healthcare company into different subpopulation. Consequently, the researchers show the value of applying a DM model to identify diabetic patients.

Abdullah A. Aljumah, [15], et.al, research concentrates upon predictive analysis of diabetic treatment using a regression-based data mining technique. The Oracle Data Miner (ODM) was employed as a software mining tool for predicting modes of treating diabetes. The support vector machine algorithm was used for experimental analysis.

Neesha Jothi, [16], et.al, highlighted that the data mining and healthcare industry have emerged some of reliable early detection systems and other various healthcare related systems from the clinical and diagnosis data. In regard to this emerge, here the authors reviewed various papers involved in this field in terms of method, algorithms and results. This review paper has consolidated the papers reviewed inline to the disciplines, model, tasks and methods. Results and evaluation methods are discussed for selected papers and a summary of the finding is presented to conclude the paper.

Han Wu, [17], et.al, work shows that the model attained a 3.04% higher accuracy of prediction than those of other researchers. Moreover, our model ensures that the dataset quality is sufficient. To further evaluate the performance of our model, we applied it to two other diabetes datasets. Both experiments' results show good performance. As a result, the model is shown to be useful for the realistic health management of diabetes.

Harleen Kaur [18], et.al, review is about the Valuable knowledge, can be discovered from application of data mining techniques in healthcare system. In this study, we briefly examine the potential use of classification based data mining techniques such as Rule based, decision tree and Artificial Neural Network to massive volume of healthcare data. In particular we consider a case study using classification techniques on a medical data set of diabetic patients.

Sajida Perveen, [19], et.al, study follows the adaboost and bagging ensemble techniques using J48 (c4.5) decision tree as a base learner along with standalone data mining technique J48 to classify patients with diabetes mellitus using diabetes risk factors. This classification is done across three different ordinal adults groups in Canadian Primary Care Sentinel Surveillance network. Experimental result shows that, overall performance of adaboost ensemble method is better than bagging as well as standalone J48 decision tree.

Divya Tomar and Sonali Agarwal [20], presented a brief introduction of Data Mining techniques and their advantages and disadvantages. This survey also highlights applications, challenges and future issues of Data Mining in healthcare. Recommendation regarding the suitable choice of available Data Mining technique is also discussed in this paper.

Rashedur [21], et.al, analyzed the performance of different classification techniques for a set of large data. A fundamental review on the selected techniques is presented for introduction purpose. The diabetes data with a total instance of 768 and 9 attributes (8 for input and 1 for output) will be used to test and justify the differences between the classification methods. Subsequently, the classification technique that has the potential to significantly improve the common or conventional methods will be suggested for use in large scale data, bioinformatics or other general application.

Messan Komi [22], et.al., explored the early prediction of diabetes via five different data mining methods including: GMM, SVM, Logistic regression, ELM, ANN. The experiment result proves that ANN (Artificial Neural Network) provides the highest accuracy than other techniques.

This section presented a survey on Data mining approaches for Healthcare of various researchers.

V. Conclusions

Data mining is the process of sorting through large data sets to identify patterns and establish relationships to solve problems through data analysis. Data mining tools allow enterprises to predict future trends. Various traditional methods, based on physical and chemical tests, are available for diagnosing diabetes. The methods strongly based on the data mining techniques can be effectively applied for high blood pressure risk prediction. This paper has analysed role of Data mining in Diabetes Healthcare, studied various Data-Mining Technologies for Diabetes and presented literature survey of various researchers.

References

[1] Hemlata Sahu, Shalini Shrma, Seema Gondhalakar, "A Brief Overview on Data Mining Survey", International Journal of Computer Technology and Electronics Engineering (IJCTEE), Volume 1, Issue 3, ISSN 2249-6343.

[2] Shital P. Bora, “Data mining and ware housing”. Electronics Computer Technology (ICECT), 3rd International Conference on Volume:1, IEEE Xplore, Publication Year: 2011 , Page(s): 1 – 5, 10.1109/ ICEC TECH. 2011.5941548.

[3] Chen Hongfei, Wang Xiaoyan, “The applied research on data mining in the financial analysis of university with the analysis of college students, arrears as an example”, Business Management and Electronic Information (BMEI), 2011 International Conference on Volume:2 Digital Object Identifier: 10.1109/ICBMEI.2011.5917992, Publication Year: 2011 , Page(s): 633 - 636.

[4] H.Vignesh Ramamoorthy, ‘An Encrypted Technique with Association Rule Mining in Cloud Environment’ published in International Journal of Computer Applications (IJCA), Foundation of Computer Science, New York, USA, 2012, Page – 5 to 8, ISBN: 973-93-80867-88-1 (Impact factor: 0.821).

[5] Yukai Li, Huling Li and Hua Yao, "Analysis and Study of Diabetes Follow-Up Data Using a Data-Mining-Based Approach ", Computational and Mathematical Methods in Medicine, Volume 2018, Article ID 7207151, 8 pages https://doi.org/10.1155/2018/7207151.

[6] Harleen & Bhambri, "A Prediction Technique in Data Mining for Diabetes Mellitus", Apeejay-Journal of Management Sciences and Technology, 4 (1), October – 2016 ISSN -2347-5005.

[7] DeeptiSisodia, "Prediction of Diabetes using Classification Algorithms", Procedia Computer Science 132 (2018) 1578–1585, Published by Elsevier Ltd., Volume 132, 2018, Pages 1578-1585, https://doi.org/10.1016/j.procs.2018.05.122

[8] Miroslav Marinov, M.S., Abu Saleh Mohammad Mosa, M.S., Illhoi Yoo, Ph.D.,Suzanne Austin Boren, Ph.D., MHA, “Data-Mining Technologies for Diabetes: A Systematic Review” Journal of Diabetes Science and Technology ,Volume 5, Issue 6, November 2011 . J Diabetes Sci Technol 2017;5(6):1549-1556

[9] Ioannis Kavakiotis, Olga Tsave, et. al., "Machine Learning and Data Mining Methods in Diabetes Research", Published online 2017 Jan 8. doi: [10.1016/j.csbj.2016.12.005], Comput Struct Biotechnol J. 2017; 15: 104–116.

[10] H.Vignesh Ramamoorthy, A. Manjula, “Diabetes Forecasting using Modified RBF Neural Networks ”, International Journal of Scientific Research in Computer Science Applications and Management Studies, Volume 3, Issue 5, September 2014, ISSN 2319 – 1953.

[11] Joseph L. Breault, Colin R. Goodall, Peter J. Fos, “Erratum to Data mining a diabetic data warehouse’’, Elsevier Science doi: 10.1016/ S0933-3657(03) 00012-5, Artificial Intelligence in Medicine 27 (2013), 227.

[12] Dr Saravana kumar N M , Eswari T , Sampath P & Lavanya S, “Predictive Methodology for Diabetic Data Analysis in Big Data”, 2nd International Symposium on Big Data and Cloud Computing (ISBCC’15), Procedia Computer Science 50 (2015) 203 – 208.

[13] Ioannis Kavakiotis, Olga Tsave , Athanasios Salifoglou , Nicos Maglaveras ,Ioannis Vlahavas , Ioanna Chouvarda,“Machine Learning and Data Mining Methods in Diabetes Research”, Elsevier B.V, Computational and Structural Biotechnology Journal 15 (2017), 104–116.

[14] Thomas Porter, Barbara Green, “Identifying Diabetic Patients: A Data Mining Approach” AIS Electronic Library (AISeL) ,Americas Conference on Information Systems (AMCIS), Proceedings of the Fifteenth Americas Conference on Information Systems, San Francisco, California August 6th -9th 2009.

[15] Abdullah A. Aljumah, Mohammed Gulam Ahamad, Mohammad Khubeb Siddiqu , “Application of data mining: Diabetes health care in young and old patients”, Elsevier, Journal of King Saud University – Computer and Information Sciences (2013) 25, 127–136.

[16] Neesha Jothi, Nur’Aini Abdul Rashid, Wahidah Husain, “ Data Mining in Healthcare – A Review”, The Third Information Systems International Conference , Elsevier, Procedia Computer Science 72 (2015) 306 – 313.

[17] Han Wu, Shengqi Yang , Zhangqin Huang, Jian He, Xiaoyi Wang, “Type 2 diabetes mellitus prediction model based on data mining”, Elsevier, Informatics in Medicine Unlocked 10 (2018) 100–107.

[18] Harleen Kaur and Siri Krishan Wasan, “Empirical Study on Applications of Data Mining Techniques in Healthcare”, Journal of Computer Science 2 (2): 194-200, 2006, ISSN 1549-3636.

[19] Sajida Perveen, Muhammad Shahbaz, Aziz Guergachi, Karim Keshavjee, “Performance Analysis of Data Mining Classification Techniques to Predict Diabetes”, Symposium on Data Mining Applications, SDMA2016, 30 March 2016, Riyadh, Saudi Arabia, Elsevier, Procedia Computer Science 82 (2016) 115 – 121.

[20] Divya Tomar and Sonali Agarwal, “A survey on Data Mining approaches for Healthcare ”, International Journal of Bio-Science and Bio-Technology Vol.5, No.5 (2013), pp. 241-266, ISSN: 2233-7849, International Journal of Bio-Science and Bio-Technology Vol.5, No.5 (2013).

[21] Rashedur M. Rahman, Farhana Afroz , “Comparison of Various Classification Techniques Using Different Data Mining Tools for Diabetes Diagnosis” Journal of Software Engineering and Applications, 2013, 6, 85-97.

[22] Messan Komi, et.al., “Application of data mining methods in diabetes prediction”, 2017 2nd International Conference on Image, Vision and Computing (ICIVC), June 2017, DOI: 10.1109/ICIVC.2017.7984706, Publisher: IEEE Digital Xplore.