International Journal of Computer Sciences and Engineering Open Access
Survey Paper Vol.-7, Issue-1, Jan 2019 E-ISSN: 2347-2693

Diabetes Mellitus and Data Mining Techniques: A survey

Mirza Shuja[1*], Sonu Mittal[2], Majid Zaman[3]

[1,2] School of Computer and System Sciences, Jaipur National University, Jaipur, India.
[2] Directorate of IT&SS, University of Kashmir, Srinagar, India.

[*]*Corresponding Author: info.shuja@yahoo.in.*

Available online at: www.ijcseonline.org

Accepted: 10/Dec/2018, Published: 31/Jan/2019

***Abstract*—** Data has become an integral part of almost every organization. This data contains interesting and vital information that is often hidden to naked eye but is in the greater interest to an organization, this reason has led researchers for finding a special interest in extracting the hidden knowledge that is accumulated within it, with some researchers terming it as goldmine of data. In this scenario data mining has found a special place in the healthcare sector. Data mining has been found to be quite successful in healthcare sector in finding out the hidden patterns that are useful for disease prognosis. These data mining techniques have been successfully applied for prognosis of diabetes. Diabetes mellitus commonly known as diabetes is a metabolic disorder condition which is characterized by high level of sugar in blood. Numerous data mining techniques have been used for designing of the model that could aid physicians in predicting diabetes. In this paper the main focus is to make present detailed survey of various data mining techniques and approaches that have been put to use for prognosis of diabetes. The research presented here is a survey focused mainly on evaluation of various computer based tools designed for prognosis of diabetes**.**

***Keywords*—** *Diabetes, Data mining, Decision tree, Dataset, Prognosis, SVM.*

I. INTRODUCTION

For an active and efficient working of a human body energy is required which is provided by glucose (an important source of energy for body cells) which is produced through a process called as glycolysis. For haulage of glucose into the body cells, it requires a hormone called 'Insulin'. The insulin is an important hormone in our body that is produced by the β-cells of a gland called pancreas. As the level of glucose increases in blood, the β-cells of pancreas are stimulated and insulin is released into the blood, the insulin acts as catalyst that aids the glucose in blood to enter into the cells for providing energy. At times due to malfunctioning of insulin it leads to a serious condition called as diabetes and is characterized by increase in the level of glucose in blood.
The diabetes has a very hazardous effect on human body and has led to some extreme medical conditions in some cases. The diabetes has been divided into two major types: Type 1 (Juvenile-onset diabetes), Type 2 (Adult-onset diabetes). Type 2 diabetes is the major type of diabetes that is found in about 90% of the total number of patients suffering from diabetes. Till present time no cure has been possible for its treatment although changes in lifestyle and basic exercises have proven to be effective to some extent for managing the sugar level.

The serious challenge posed by diabetes has tempted researchers for application of new techniques for aiding remedy finding. Data mining which is considered as an emerging field by both industry as well as scholastics has found a special attention in both. Numerous data mining techniques have been applied for designing of predictive or diagnostic models for diabetes with a reasonable success. Some of the popular algorithms like Decision trees, Naïve Bayes, Support Vector Machines, Neural Networks have been used extensively.

II. DIABETES

Diabetes mellitus commonly known as diabetes is often attributed to changing lifestyle of humankind; it is a metabolic disorder that is characterized by hyperglycemia- a condition of high concentration of glucose in blood. It is the consequence of defective insulin secretion or defective insulin action or in some cases both and affects metabolism of body resulting in raised sugar level in blood. The chronic hyperglycemia has a hazardous effect on body which leads to dysfunctioning and failures of several organs. Neuropathy, Nephropathy, Retinopathy, Cardiac Disorder and Blood Vessels compression being some of its main illnesses caused by it [1].

The chronic condition of diabetes has been broadly classified in to major types: A) Type I diabetes B) Type II diabetes. As per WHO statistics 0f 2014, about 422 million people globally were suffering from diabetes which is four times the increase since 1980. 1.5 million Deaths were directly caused by diabetes or its related ailments worldwide. International Diabetes Federation (IDF) has noted that every 1 in 2 adults with diabetes are often remain undiagnosed and has anticipated that 1 in every 10 individuals would be diabetic by year 2040.

III. DATA MINING

As per [2] data mining is "The process of extraction of meaningful information, new relations and discovery of new trends and patterns from repositories containing vast amount of data, for extraction purposes both pattern recognition as well as mathematical and statistical techniques are used". "To analyze a vast amount of operational dataset for extraction of unspecified relation and summarizing them into way that is novel in understanding and usefulness to owner of data" is defined as data mining [3].

"Data mining is an emerging field and has emerged as interdisciplinary in nature that has bought together techniques from machine learning, pattern recognition, statistics, databases and visualization for extraction of useful and specific information from large databases" [4]. For data mining the phrase "knowledge discovery in databases (KDD)" is often synonymously used. Data are any raw facts that can be in the form of numbers or texts which can be processed by computer system into useful information.

At present organizations are generating vast amount of data which accumulates in different formats and different databases making it a huge Data Repository. For better and an efficient decision making, a proper and efficient mechanism for information extraction is needed. Researchers throughout have successfully applied data mining techniques for prognosis and/or diagnosis of diabetes with every new study proving better than previous.

IV. LITERATURE REVIEW ON APPLICATION OF DATA MINING TECHNIQUES FOR DIABETES PROGNOSIS

As data have become an integral part of any organization and to analyze it for discovering hidden knowledge has become inevitable for improvement in services, same is true for medical field where predictive data mining is used for prognosis of disease at an early stage to pre-empt its effects and to aid physicians in developing contingency plan. The available literature reveals majority of the work that has been carried out on diabetes has focused mainly on developing the methods for prognosis or diagnosis of type II diabetes to reduce its complications, in majority of the cases Pima India dataset has been used for experimentation though methods and tool used have varied.

To detect diabetes at initial stage a multi agent system was designed by [5] called 'CoLe' the system was not just a combinational agent but was having multiple data miners. The intermingling of learning having depiction of information with alternate view point was its fundamental goal and was achieved to large accuracy.

To predict diabetes [6] applied associative rule mining to discretinizing continuous valued attributes an equal interval binning technique was used and for diabetic classification Apriori algorithm was applied and at the end association rules were generated for understanding relationship among measured fields used in prediction.

In [7] Fuzzy ID3 in combination with EM (Estimation Maximization) was used for diabetic prediction the model was called Hybrid Classification System, it was a 2-phasic system in initial phase EM algorithm was fed with cleaned data for clustering data and in second phase adaption rules essential for diabetic prognosis were obtained using ID3 algorithm the model's accuracy was about 91.32%.

An expert system for diabetic prognosis was developed by [8] the system used extended classifier system (XCS)-a learning agent in artificial intelligence have greater accuracy rate than others, the system composed a simple set of 'if-else' rules the achieved accuracy was 91.3%.

In their research work [9] used novel artificial bee algorithm to predict diabetes, to improve prediction they used an artificial mutation operator the model achieved an accuracy of 84.21%.

A joint implementation of two mostly widely used algorithms Support Vector Machines and Naïve Bayes was used by [10] for prognosis of diabetes the system had an accuracy of 97.6%.

Hybrid prediction model for prognosis of type II diabetes was proposed by [11], the system was the combination of two data mining classifiers K-means clustering and C4.5 decision tree algorithm, K-fold cross validation was applied for validation, the system achieved an accuracy of 92.38%.
To predict the risk of heart attack in a diabetic patient [12] applied Naïve Bayes classifier on diabetic data with an accuracy of 74%.
Ensembling methods were combined with J48 decision tree algorithm for designing a technique that could be applied on diabetic data set for predicting the chances a diabetic patient has for diabetic neuropathy, diabetic nephropathy and cardiovascular disease [13].

In their research work [14] used three data mining techniques *viz* C4.5, SVM, KNN classify diabetes with an accuracy of 86% C4.5 decision tree proved to be best among the three and was used further for designing a predictive model for diabetic prognosis.

A diabetic dataset was trained on Bayesian network algorithm for diabetic prognosis by [15]; the proposed model had an accuracy of 90.4%.

For prognosis of diabetes a model was designed by [16] by combining Genetic algorithm with fuzzy logic. The model achieved an accuracy of 80.5%.

In their work [17] used the following techniques for designing predictive model for diabetic prognosis, the techniques were EM (Estimation Maximization), K-Nearest Neighbour (KNN), K-means, Amalgam KNN and ANFIS. With accuracy rate of 80% Amalgam-KNN and ANFIS emerged as best models.

Multi layer Perceptron, J48 decision tree, and Naïve Bayes algorithms were applied by [18] on Pima India diabetic dataset for prediction of diabetes, the model designed using Naïve Bayes had an accuracy of 76.30%.

A combination of OLAP and two data mining algorithms C4.5 and ID3 decision tree were used to develop decision support system for diabetic prediction with an accuracy of 74% by [19].

For diabetic prognosis a hybrid predictive model was designed by [20] using data mining classification algorithms SMO, J48, Bagging, J48, Naïve Bayes, Random Forest and AdaBoost were used. K-means clustering was used combined with these techniques for prediction of Positive and Negative cases of diabetes.

With the aim of increasing the predictive accuracy [21] combined class imbalance reducing algorithm SMOTE with J48 Decision Tree for designing decision support system for diabetes prognosis the designed model achieved an accuracy of 94.70%.

V. DISCUSSIONS

Before any of the learning frameworks is made useful and put to use, its exactness needs to be assessed thoroughly. The process of evaluating precision is a troublesome errand due to restricted accessibility of data. For a machine learning systems perfection picking a decent assessment methodology is a vital step. Due to privacy concerns medical data is not easily available and is the main factor behind making research in medicinal field weighty. The analysis of work done on diabetes that is presented in this paper has been carried on Pima India dataset which is available online; the dataset consists of nine attributes with 768 instances which are considered to be main factors in diabetes. The attribute list is given in table 1.

Table 1. Pima India Attributes.

Sr. No.	Attribute Name
1.	Diastolic Blood Pressure
2.	Plasma Glucose Concentration
3.	Number of Times Pregnant
4.	Body Mass Index
5.	2-Hr Serum Insulin
6.	Triceps Skin Fold Thickness
7.	Age
8.	Diabetic Pedigree Function
9.	Class (yes or no)

All but eight attributes are directly related to diabetes while last attribute is used to differentiate between positive and negative cases of diabetes. The majority of the reviewed research presented in this survey used the same dataset for validating their models. The dataset has some of the limitations as it contains the data related to only female population while data related to male population is overlooked, some of the vital attributes like HbA1c value are not considered. There are many missing values in almost all attributes. Majority of the models have used classification techniques although the prediction accuracy has varied.

VI. CONCLUSION

The focal point of this paper was to present the various data mining techniques that have been applied for diabetic data mining for designing predictive models for prognosis of diabetes. Application of data mining techniques for prognosis of diabetes is task that puts up a great challenge but it has reduced the human effort drastically with increase in prognostic precision. Development of efficient data mining applications has led to reduction of both constraints *viz* cost and time in terms of human resources and expertise. A careful study of various data mining techniques was carried in this study and it could be concluded Decision tree, Support Vector Machines, Naïve Bayes and K-NN were used by researcher in majority of cases in solo or some have used combined techniques inorder to increase the predictive accuracy.

REFERENCES

[1]. American Diabetes Association. "Diagnosis and classification of diabetes mellitus." Diabetes care 37.Supplement 1 (2014): S81-S90.

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

[3]. 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.

[4]. Piatetsky-Shapiro, Gregory, The Data-Mining Industry Coming of Age,"IEEE Intelligent Systems, 2000.
[5]. Gao, Jie, Jörg Denzinger, and Robert C. James. "CoLe: A Cooperative Data Mining Approach and Its Application to Early Diabetes Detection." ICDM. 2005.
[6]. Patil, B. M., R. C. Joshi, and Durga Toshniwal."Association rule for classification of type-2 diabetic patients." Machine Learning and Computing (ICMLC), 2010 Second International Conference on. IEEE, 2010.
[7]. Adidela, D. R., et al. "Application of fuzzy ID3 to predict diabetes." Int J Adv Comput Math Sci 3.4 (2012): 541-5.
[8]. Afrand, P. O. U. Y. A., et al. "Design and implementation of an expert clinical system for diabetes diagnosis." Global Journal of Science, Engineering and Technology (2012): 23-31.
[9]. Beloufa, Fayssal, and Mohammed Amine Chikh."Design of fuzzy classifier for diabetes disease using Modified Artificial Bee Colony algorithm." Computer methods and programs in biomedicine 112.1 (2013): 92-103.
[10]. Tafa, Zhilbert, Nerxhivane Pervetica, and Bertran Karahoda. "An intelligent system for diabetes prediction." Embedded Computing (MECO), 2015 4th Mediterranean Conference on. IEEE, 2015.
[11]. Patil BM, Joshi RC, Toshniwal D. Hybrid prediction model for type-2 diabetic patients. Expert Systems with Applications 2010;37(12):8102–8.
[12]. Parthiban G, Rajesh A, Srivatsa SK. Diagnosis of heart disease for diabetic patients using naive bayes method. International Journal of Computer Applications 2011;24(3):7–11.
[13]. Huang F, Wang S, Chan CC. Predicting disease by using data mining based on healthcare information system. In: Granular computing (GrC), 2012 IEEE international conference on (pp. 191–4). IEEE; August 2012.
[14]. P. Radha, Dr. B. Srinivasan, " Predicting Diabetes by consequencing the various Data mining Classification Techniques", International Journal of Innovative Science, Engineering & Technology, vol. 1 Issue 6, August 2014, pp. 334-339.
[15]. Mohtaram Mohammadi, Mitra Hosseini, Hamid Tabatabaee, "Using Bayesian Network for the prediction and Diagnosis of Diabetes" , MAGNT Research Report, vol.2(5), pp.892-902.
[16]. Sudesh Rao, V. Arun Kumar, "Applying Data mining Technique to predict the diabetes of our future generations", ISRASE eXplore digital library, 2014.
[17]. Veena vijayan, Aswathy Ravikumar, " Study of Data mining algorithms for prediction and diagnosis of Diabetes Mellitus", International Journal of Computer Applications (0975-8887) vol. 95-No.17, June 2014.
[18]. Murat Koklu and Yauz Unal, " Analysis of a population of Diabetic patients Databases with Classifiers", International Journal of Medical,Health,Pharmaceutical and Biomedical Engineering", vol.7 No.8, 2013.
[19]. Rupa Bagdi, Prof. Pramod Patil," Diagnosis of Diabetes Using OLAP and Data Mining Integration", International Journal of Computer Science & Communication Networks,Vol 2(3), pp. 314-322.
[20]. P. Hemant and T. Pushpavathi, "A novel approach to predict diabetes by Cascading Clustering and Classification", In Computing Communication & Networking Technologies (ICCCNT), 2012 Third International Conference on IEEE, (2012) July, pp. 1-7.
[21]. Mirza, S., Mittal, S., & Zaman, M. (May 2018). Decision Support Predictive model for prognosis of diabetes using SMOTE and Decision tree. International Journal of Applied Engineering Research, 13(11), 9277-9282.

Authors Profile

Mr. Shuja Mirza, is a research scholar currently pursuing Ph.D. in Computer Sciences from Jaipur National University, Jaipur. Prior to that he received Master degree in computer application (MCA) from university of Kashmir. He is currently working on predictive application of data mining in diagnosis of diabetes. He has over three year's research experience. His areas of interest are data mining, data analytics and machine learning.

Dr.Sonu Mittal, is currently working as Associate Professor, Department of Computer Science & Engineering at Jaipur National University. He received Ph,D. SGV University Jaipur. Prior to that he received his Masters Degree (M.Tech) from IGNOU. He has over 14 years of teaching and research experience. His areas of interest include Data science and Machine Learning.

Dr. Majid Zaman, is currently working as Scientist D in Directorate of IT&SS Univeristy of Kashmir. He received his Ph.D from University of Kashmir, prior to that he received his Masters of Science (MS) from BITS Pilani. He has over 15 years of teaching and research experience. He has published over 50 papers in reputed journals both at national and international level. His areas of interest include data mining, data analytics, big data.

