Review on Prediction of Diabetes using Data Mining Technique

Vrushali Balpande[1], Rakhi Wajgi[2]

[1]*M. Tech., Computer Science and Engineering, Yeshwantrao Chavan College of Engineering, Nagpur.*
[2]*Computer Science and Engineering, Yeshwantrao Chavan College of Engineering, Nagpur.*

Abstract **- Data mining plays an efficient role in prediction of diseases in health care industry. Diabetes is one of the major global health problems. According to WHO 2014 report, around 422 million people worldwide are suffering from diabetes. Diabetes is a metabolic disease where the improper management of blood glucose levels led to risk of many diseases like heart attack, kidney disease, eye etc. Many algorithms are developed for prediction of diabetes and accuracy estimation but there is no such algorithm which will provide severity in terms of ratio interpreted as impact of diabetes on different organs of human body. This paper gives detailed review of existing data mining methods used for prediction of diabetes. It also gives future direction for severity estimation of diabetes**

Keywords **- Data mining, Diabetes, Body Mass Index, OGTT, FPG etc**

I. INTRODUCTION

Data mining is described as the process of discovering correlations, patterns and trends to search through a large amount of data stored in repositories, databases, and data warehouses. Humans in that sensitivity are limited by data overload so there are new tools and techniques are being progress to solve this problem through automation. Data mining adopts a series of pattern recognition technologies and statistical and mathematical techniques to discover the possible rules or relationships that govern the data in the databases. Data mining must also be known as a process that requires goals and objectives to be specified. Diabetes is a chronic disease and a major public health challenge worldwide. It happens when a body is not able to react or outgrowth properly to insulin, which is needed to maintain the rate of glucose. Diabetes can be controlled with the help of insulin injections, a healthy diet and regular exercise but there is no whole cure is available. Diabetes leads to much other disease such as blindness, blood pressure, heart disease, and kidney disease and nerve damage. There are three prime types of diabetes mellitus: Type 1 Diabetes Mellitus results from the body's failure to produce insulin. This form was previously referred to as insulin-dependent diabetes mellitus.Type2 Diabetes Mellitus conclusion from insulin resistance which is a condition in which cells fail to use insulin properly, although for sometimes also with an absolute insulin deficiency. This type was previously referred to as non insulin-dependent diabetes mellitus. Gestational diabetes is the third main form and occurs when a pregnant women previously seems diagnosis of diabetes develop a high blood glucose level. In order to automate the overall process of diabetes prediction and severity estimation, diabetic database is needed. This repository of diabetic database helps in identification of impact of diabetes on various human organs. More the accuracy of prediction, more the chances of accurate severity estimation. Therefore this paper has presented different prediction methods of diabetes.

II. RELATED WORK

Gyorgy J. Simon, Pedro J.Caraballo, *et al.*, [1] proposed the method of distributional association rule mining to identify sets of risk factors and the corresponding patient subpopulations that are significantly increased risk of progressing to diabetes. And to discover sets of risk factor, here uses bottom up summarization algorithm which produces most suitable summary that describes subpopulations at high risk of diabetes. The Subpopulation identified by this summary covered most high risk of patients, had low overlap and were at very high risk. This method is used for when the patient having high risk. Dr. Zuber khan, shaifali sing, *et al.*, [2] worked on the concept of Diabetes Mellitus using k-Nearest Neighbor algorithm which is most Important technique of Artificial Intelligence. The accuracy rate is showing that how many outputs of the data of the test dataset are same as the output of the data of different features of the trained dataset. The error rate is sighting that how many outputs of the data of the test dataset are not same as the output of the data of different features of the training dataset. The result they showed that as the value of k increases, accuracy rate and error rate will increase. K-Nearest Neighbour algorithm is one of the most important techniques of AI which is used widely for diagnostic purposes. Through KNN more Accurate results can be obtain. This method is very effective for the training data set which is very large. Mukesh kumari and Dr. Rajan Vohra [3] worked on the concept of data mining is to extract knowledge from information stored in dataset and generate clear and understandable description of patterns. The techniques are attributes selection, data normalization and then classifier is applied on data set to construct Bayesian model. Bayesian network classifier was proposed for the prediction of person weather diabetic or not. By using Bayesian classifier patient is undergoing classified in classes of Pre-diabetic, Non-diabetic,

Diabetic according to the attributes selected. The techniques they applied as preprocessing attribute identification and selection, data normalization. And then classifier is applied to the modified data set to construct the Bayesian model. The Bayesian network has a benefit of it encodes all variables, missing data entries can be handled successfully. Dr.Pramanand Perumal and Sankaranarayanan [6] proposed an idea about diabetes mellitus its diagnosis using data mining with minimum number of attributes applied to classification algorithms. They worked on Apriori and FP-growth techniques. In FP-growth the novel data structure frequent pattern tree is being implemented for storing compressed crucial information about frequent pattern. It is observed that both of the techniques generate the same number of frequent sets as a importance same number of rules for the same known dataset under the same constraints. with the help of data Apriori and FP-growth algorithms, the computation cost decreases and also the classificationperformance increases. Satyanarayana Gandi and Amarendra Kothalanka [7] worked on the initial training data set to the optimal process to extract the optimal data set, on that optimal dataset they applied classification with Bayesian classifier. Bayesian classifier methods is uses getting training data set and convert it into classified data. Initially they extract the optimal feature set from existing training data and calculates the positive and negative probability, until the new data set if formed with same size and forwards the current generated dataset for classification their it classifies the testing dataset with new feature. Sanchita paul and Dilip kumar Choubey [9] proposed an approach for feature selection, classification and used Genetic Algorithm, Multilayer Perceptron Neural network on diabetes data set. With features selection methodology using Genetic algorithm they improve the accuracy but achieved slightly less ROC. With feature Selection methodology genetic algorithm improved accuracy but achieved less ROC by applying GA,MLP NN methodology classification ROC is also improved. Ramkrishnan Shrikant and Rakesh Agrawal [8] proposed a systematic framework of building a risk prediction model for type-2 diabetes disease. The GBRE algorithm identifies the best set of indicators that can predict risk level of diabetes and then multiple classifiers are trained and their accuracy are measured. Alan J. Garber,MD and Martin J.Abrahamson *et al.,* [10] developed case study includes Evalution for Complications and staging, Lifestyle Modifications, Algorithm for adding/Intensifying insulin, CVD Risk factor algorithm, Profiles of anti-diabetic Medications. Principles of the AACE Algorithm for the treatment of type 2 diabetes. Rohit Prasad Bakshi and Sonali Agrawal [16]proposed a systematic framework of building a risk of prediction model for type-2 diabetes disease. The GBRE algorithm finds the best set of that can found risk level of diabetes and then multiple classifiers are trained and their accuracy are being measured. The classifier has been selected by voting policy technique. The suggested approach can be applied significantly in prediction modeling of other diseases. S.Sapna and Dr.A.Tamilarasi[17] proposed a concept of Genetic Algorithm and Fuzzy system on chromosomes. To Obtained the accuracy of chromosome and to evaluate the diabetes in diabetic patient GA is implemented. The connection between fuzzy system and genetic algorithm is bidirectional. Genetic Algorithms are utilized to deal with various optimization problems involves fuzzy system. Using GA optimization of chromosome is obtained and based on the rate of old population diabetes can be restrained in new population to get chromosomal accuracy. Srideivanai Nagarajan and R.M. Chandrasekaran[18] proposed a method for improvement of diagnosis of gestational diabetes with data mining techniques. Also they Analyse the performance of ID3, Naïve Bayes, C4.5, and Random tree i.e. the algorithm for supervised Learning. They used the data set of Pregnant Womens. The results they found that Random tree served to be the best one with higher accuracy and least error rate. Veena V.Vijayan and Aswathy Ravikumar[19] discussed the main data mining algorithm, K-Means Algorithm, Amalgam KNN algorithm and ANFIS algorithm They proposed the study of Expectation Maximization algorithm used for sampling to determine and maximize the expectation in successive iteration cycles. K- Nearest Neighbor Algorithm is used for classification of objects and used for prediction of labels based on some closest training examples in the feature space. K-Means algorithm follows partition methods based on some input parameters on the dataset of n objects. They discussed about Amalgam Algorithm combines both the feature of K-Nearest Neighbor and K-Means with some additional processing and the Adaptive Neuro Fuzzy Inference System which combines the Features of Adaptive Neural Network and Fuzzy Inference System. They choose the dataset from PIMA Indian Diabetic Set from University of California. K.Rajesh and V.Sangeetha[20] proposed that data mining relationship for efficient classification they applied data mining techniques to classify diabetes clinical data and predict the patient being affected with diabetes or not. They presented a system which gave training data on that data feature relevance analysis is done then comparison of classification algorithm, Selecting classifier then improved classification algorithm is applied and then found out the evaluation that compared with training data. They applied C4.5 Algorithm gave classification rate of 91%.Dr. B .L. Shivkumar and S. Alby[21]presents a survey paper for data mining methods that have been commonly applied to diabetes data analysis and prediction of disease. They done an analysis of various presentations and studies done by other researches. From the analysis of different research papers it is evident that the occurrence of diabetes is having strong relation with diseases like Wheeze Edema, Oral diseases, Female Pregnant and increase of age. Using data mining techniques the chance of diabetes can be predicted which is helpful for early detection of the disease. Carlos Fernandez_Llatas and Antonio Martinez_Millanu[22] proposed the use of Interactive Pattern Recognition techniques for the iterative design of protocols and analyzing the problems of using process mining to infer care flows and how to cope the resulting spaghetti Effect. Below figure shows the brief description about methods which previously used for prediction analysis

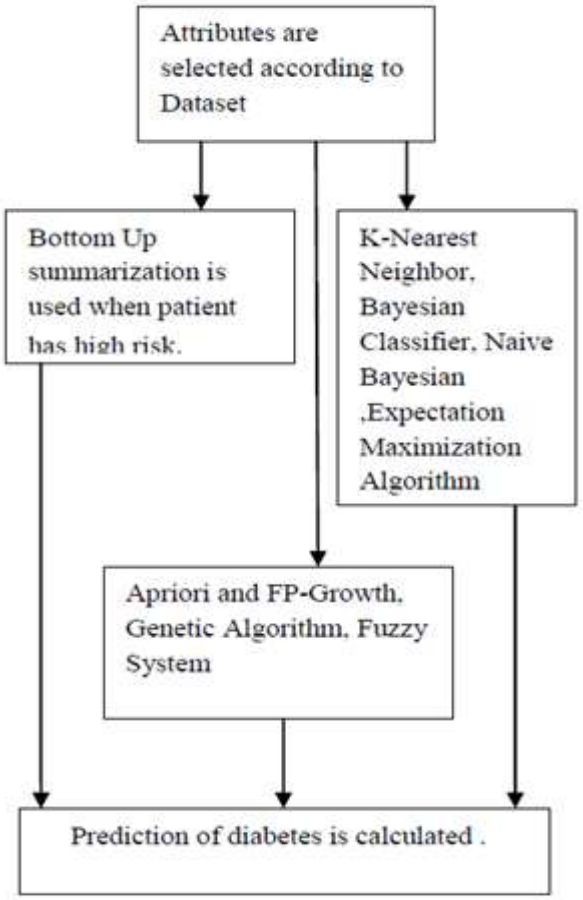

Figure: Different Algorithms Used for Prediction of diabetes.

III. EFFECT OF DIABETES

Cardiovascular disease includes blood vessel disease, heart attack and stroke. The risk is greater for people with diabetes, who have progressed cholesterol and blood pressure levels. If family has smoking history also increases heart problems. To reduce the risk and pick up any problems early: Have the blood pressure checked at least every six months, but more often if person have high blood pressure or are taking medication to lower this. Have the test of HbA1c checked at least every year it may need to be checked three to six monthly.

Have the cholesterol checked at least yearly. Further pathology tests such as an electrocardiogram (ECG) or exercise stress test may also be recommended by doctor. Heart disease and blood vessel disease are common problems for many people who don't have their diabetes under control. Blood vessel damage or nerve damage may also cause foot problems that, in rare cases, can lead to amputations. People having diabetes are ten times occurred to have their feet and legs removed than those without the disease.

Peripheral diabetic neuropathy can cause pain or a loss of feeling in feet. It commonly starts with toes. It can also severe for hands and other body parts. Autonomic neuropathy branch from damage to the nerves that control internal organs. Symptoms include sexual problems, digestive issues, trouble sensing when the bladder is full, dizziness and fainting, or not knowing when blood sugar is low.

- Retinopathy – with this condition, the blood vessels in the retina become damaged and eventually this can affect vision. Retinopathy has various stages. During early stages there are usually no symptoms, so having a full diabetes eye check is essential to detect earlier. Regular eye checks help detect any changes and allow for early treatment where needed to prohibit further damage.
- Macular oedema – The macula is some part of the retina and helps us to see things clearly. Swelling of this area can happen when the blood vessels in the retina are ruin and cause fluid to build up. This can lead to the macula being ruin and vision may become blurry. Treatment is available and Early detection is important.
- Cataracts – The lens of the eye becomes cloudy and can account vision to become cloudy, distorted or sensitive to glare. People with diabetes can develop cataracts at an earlier age.
- Glaucoma – The pressure of the fluid within the eye builds up to a greater level than is healthy. This burden can damage the eye over time. Glaucoma occurs in people with and without diabetes but is more typical in people with diabetes. Most damage to the eyes is free of symptoms in the earlier stages, there are convinced symptoms that may occur and these need urgent review.
- Regular eye checks - All Persons with diabetes should have a professional eye examination by an ophthalmologist when they are first diagnosed, and then at least every two years after that. It is essential that to inform the person checking of eyes who has diabetes. If retinopathy or another abnormality is found, eye tests will be required every year, or more regularly if advised by ophthalmologist. Below figure shows that the prevalence of Type 2 diabetes mellitus.

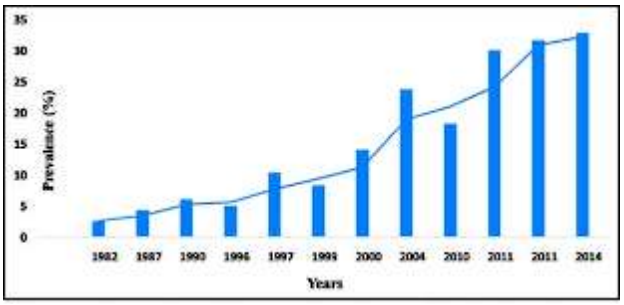

Figure2: Prevalence of Type 2 diabetes mellitus

IV. CONCLUSION

The Amount of Research work has been done for Prediction of diabetes using data mining technique. The bottom up summarization technique uses when patient has high risk of diabetes. The K-Nearest Neighbor Algorithm, Bayesian Classifier, Naïve Bayesian Classifier, Artificial Neural Network, Bayesian Network, Association Rule Mining all methods used for prediction of diabetes which gives patients condition of Normal, Pre-diabetes, Diabetes. In K-Nearest neighbor algorithm always need to determine the value of K. All above methods used to predict diabetes. But if Patient is detected as diabetes firstly there is a need of finding Control and Un-control condition of diabetes. Because if Patient has diabetes in Un- control condition, may be the patient has severe effect on Patient's Organ like Heart, Eye, Kidney etc. So there is need of finding early Severity which may be help patient for reducing the Severity on Organ or Halting the Severe Effect on Organ.

REFERENCES

[1]. GyorgyJ.Simon,Pedro J.Caraballo,Terry M. Therneau,Steven S. Cha, M. Regina Castro and Peter W.Li "Extending Association Rule Summarization Techniques to Assess Risk Of Diabetes Mellitus," *IEEE Transanctions on Knowledge and Data Engineering* ,vol 27,No.1,January 2015

[2]. Dr.Zuber khan, shaifali singh and Krati Sexena," Diagnosis of Diabetes Mellitus using K- Nearest Neighbor Algorithm," *in proceeding of International Journal of Computer Science Trends and Technology,* vol.2 , July-Aug 2014

[3]. Mukesh kumari and Dr. Rajan Vohra, "Prediction of Diabetes Using Bayesian Network,"*in proceeding of International Journal of Computer Science and Information Technologies,* vol. 5 , 2014

[4]. Jianchao Han, Juan C. Rodriguze and Mohsen Beheshti , "Diabetes Data Analysis and Prediction Model Discovery Using RapidMiner,"*in proceeding of Second International Conference on Future Generation Communication and Networking*, vol.2,2008

[5]. Wang ZuoCheng and XUE Li Xia ,"A Fast Algorithm for Mining Association Rules in Image," *in proceeding of International Conference on Data Engineering,* vol.5, 2008

[6]. Dr. Pramanand Perumal and Sankaranarayanan, "Diabetic prognosis through Data Mining Methods and Techniques," *in proceeding of International Conference on Intelligent Computing Applications,* vol. 2, 2014

[7]. Satyanarayana Gandi and Amarendra Kothalanka,"An Efficient Expert System For Diabetes By Naïve Bayesian Classifier," *in proceeding of International Journal of Engineering Trends and Technology* ,vol. 4 ,Issue 10 , Oct 2013

[8]. Ramkrishnan Shrikant and Rakesh Agrawal,"Fast Algorithms for mining association rule," *in proceeding of IEEE International Conference on Data Engineering,*vol.16,2007

[9]. Dilip Kumar Choubey and Sanchita Paul,"GA_MLP NN: A Hybrid Intelligent System for Diabetes Disease Diagnosis*",in proceedings of I.J.Intelligent System and Applications,* vol.1,pp.49-59,2016

[10]. Alan J. Garber,MD and Martin J.Abrahamson, Case study on "AACE/ACE Comprehensive Diabetes Management Algorithm"

[11]. H. S. Kim, A. M. Shin, M. K. Kim, and N. Kim, "Comorbidity study on type 2 diabetes mellitus using data mining," *in proceedings of Korean J. Intern. Med.*, vol. 27, no. 2, pp. 197–202, Jun. 2012

[12]. Kawita Rawat and Kawita Bhurse" A Comparative Approach for Pima Indians Diabetes Diagnosis using LDA-Support Vector Machine and Feed Forward Neural Network,"*in proceedings of International Journal of Advanced Research in Computer Science and Software Engineering*, vol.4, Nov. 2014

[13]. G. S Collins, S. Mallett, O. Omar, and L.-M. Yu, "Developing risk prediction models for type 2 diabetes: A systematic review of methodology and reporting,"in proceedings of BMC Med., 9:103,Sept. 2011

[14]. R. Agrawal and R. Srikant, "Fast algorithms for mining association rules," in Proceedings of 20th VLDB, Santiago, Chile, 1994

[15]. M. A. Hasan, "Summarization in pattern mining," in proceedings of Encyclopedia of Data Warehousing and Mining, 2nd ed. Hershey, PA, USA:Information Science Reference, 2008

[16]. R.P.Bakshi and S.Agrawal,"Modeling Risk of Prediction of Diabetes - a preventive Measure," in proceedings of BMC Med.,2012.

[17]. S.Sapna and Dr.A.Tamilarasi,"Implementation of Genetic algorithm in Predicting Diabetes" in Proceedings of International journal of Computer science ,vol.9,Issue.1,N0.3,Jan-2012.

[18]. S. Nagarajan and R.M.Chandrasekaran,"Data Mining Techniques for Performance Evaluation of Diagnosis in Gestational Diabetes" in proceedings of International Journal of Current Research and academic Review, vol. 2,No. 10,pp. 91-98.

[19]. V.Vijayan and A.Ravikumar," Study of data mining algorithms for Prediction and diagnosis of diabetes mellitus," in proceedings of International Journal of Computer Application, vol. 9,No. 17, June 2014.

[20]. J.Tuomilehto, "Prevention of type 2 diabetes mellitus by changes in lifestyle among subjects with impared glucose tolerance",in proceedings of International Journal of Medical Research, vol. 344,no. 18,pp. 1343-1350, 2001.

[21]. K.Rajesh and V.Sangeetha,"Application of Data Mining Methods and Techniques for Diabetes Diagnosis," in proceedings of International journal of Engineering and Innovative Technology, vol.2, Issue 3, September 2012.

[22]. B.L. Shivkumar and S Alby,"A Survey on Data Mining Technologies for Prediction and Diagnosis of Diabetes," in proceedings of International Conference on Intelligent Computing Application, 2014.

[23]. Carlos Fernandez_Llatas and Antonio Martinez_Millanu, "Diabetes care related process modelling using Process Mining Techniques Lessons Learned in the Application of Interactive Pattern Recognition :Coping with the Spaghetii Effect, 2015.