A survey on diagnosis of diabetes using various classification algorithm

Ms. Nilam chandgude (Author)
Department of computer engineering
Trinity College of engineering and research
Pune, India.
chandgude.nilam19@gmail.com

Prof. Suvarna pawar
Department Information Technology
Trinity College of engineering and research
Pune, India.
pawar.suvarna@gmail.com

Abstract**—** Diabetes is worldwide problem. It occurs when pancreas does not produce sufficient insulin, or body can not sufficiently use insulin it produces. Diabetes person has increase blood glucose in the body. People with diabetes may develop serious problems such as heart disease, stroke, kidney failure, blindness, and premature death. WHO reported, in 2013 it was found that over 382 million people throughout the world had diabetes and mostly occurred in women than men due to improper food habit or low quality of food. Early diagnosis of diabetes is an important challenge. This survey present various classification are used for diagnosis of diabetes such as artificial neural network, support vector machine, naïve bayes, decision tree. PIMA Indian dataset are chosen for diagnosis of diabetes. The research hopes to propose a quicker and more efficient technique of diagnosing the disease, leading to timely treatment of the patients.

Keywords*- Diabetes Mellitus, Classification, Artificial neural network, support vector machine, Naïve Byes, Decision tree J48, C4.5, and Cart.*

*****

I. INTRODUCTION

Diabetes, often referred to by doctors as **diabetes mellitus**, describes a group of metabolic diseases in which the person has increase glucose in blood , because of insulin production is not sufficient, or because cells in body do not respond properly to insulin which is produce , or both. This high sugar level produces lot of symptoms of polyuria, polydipsia and polyphagia. In recent years; the number of diabetic patients has increased largely because of more population and irregular western food habits or less exercise. Mostly Type 1 diabetes occurs due to Genetic inheritance.

The recent WHO report shown 2013 it was estimated that over 350 million people throughout the world had diabetes. The occurrence of diabetes in India is larger. Probably it is shown largely in women because of improper diet.

There are various types of diabetes which effect on overall body of human

- **Type1 Diabetes** – This type of diabetes called juvenile-onset because it's occurring a very young age of below 20 years. It also called insulin dependent because of human body does not produce sufficient insulin. Near about 10% of all diabetes cases are found in types1.Injection of insulin along with frequent blood test and dietary restriction has to be followed by patient suffering from type 1 diabetes.
- **Type2 Diabetes** – This type of diabetes called adult onset. It also called non-insulin dependent. The human body does not produce sufficient insulin for proper function in the body. Near about type 2 have 90% of all cases of diabetes in worldwide. Obesity, Being overweight, being physically inactive can lead to type2 diabetes.

Gestational Diabetes – Gestational diabetes affects females during pregnancy. most symptoms which occur in human body such as frequent urination, Intensive Thirst, Intensive hunger, Weight gain, unusual weight loss, male sexual dysfunction, tingling in hands and feet.

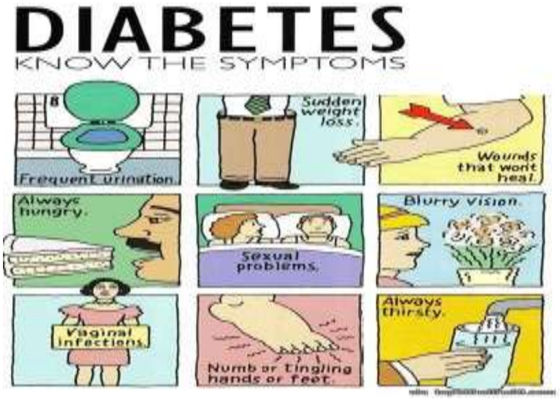


Fig 1. Symptoms of diabetes[28].

There are commonly conducted test for determining whether person has diabetes or not.

- A1C test
- Fasting Plasma Glucose (FPG) test.
- Oral Glucose Tolerance test.

Table 1: Values of diabetes.

Type	**A1c(percent)**	**Fasting plasma Glucose(mg/dl)**	**Oral Glucose Tolerance Test(mg/dl)**
Diabetes	6.5 or above	126 or above	200 or above
Pre-diabetes	5.7 to 6.4	100 to 125	140 to 199
Normal	About 5	99 or below	139 below

II. DATA SOURCES

Now days there are number of data sources are available for diagnosis of diabetes. Diagnosis of disease is very challenging task. So there are many data sources are available such as PIMA Indian diabetes dataset from UCI repository, Pub Med, WebMD, Medline.

A) PIMA Indian diabetes dataset.

PIMA Indian diabetes dataset are from UCI repositories which consist of 338 dataset. PIMA Indian dataset are provide diabetes database which used for easily diagnosis diabetes.

Attribute information in PIMA dataset

1) Number of times pregnant
2) 2 hours plasma glucose concentration in an oral glucose tolerance test
3) Diastolic blood pressure (mm Hg)
4) Triceps skin fold thickness (mm)
5) 2-Hour serum insulin (mu U/ml)
6) Body mass index
7) Pedigree functions of diabetes.
8) Age of person
9) Class Variable (0 or 1)

B) Pub med:

It is developed by the National Center of Biotechnology information (NCBI).It is free full-text archive of biomedical and life sciences journal literature.

It comprises more than 25 million citations for biomedical literature from Medline.

C) WebMD:

WebMD is one of the most popular consumer health websites. It high readability and also user-friendly. It contains various set of document. That document contains blogs written by patient. The part of blog described disease, conditions or treatment.

III. LITERATURE REVIEW

Data mining is process of extracting data from huge database.In medical system large amount of data are present. There are many properties of data mining as Automatic discovery of patterns, Prediction of likely outcomes, Creation of actionable information, Focus on large data sets and databases. Various data mining technique such as clustering, decision tree, and association rule all these are used in data base for extracting medical data. Using such techniques diagnosis of disease is very easy task.

In existing system many classifications technique are used to diagnosis of diseases such as neural network, naive Bays, support vector machine, decision tree.

A)Neural Network for Diagnosis Diabetes Mellitus

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. The key element of this paradigm is the novel structure of the information processing system. It is composed of a large number of highly interconnected processing elements (neurons) working in unison to solve specific problems.

Various neural networks are used to diagnosis of disease in medical field such as diabetes, cancer, attacks. Using such type of network diagnosis of disease is very easy task.

Scott M. Pappada [5] present feed forward neural network. Using such technique easily prediction of glucose in blood within 75 min. Only 10 patients are assed using NNM model but it not included in training data set. Various input are given to neural network model such as CGM value, insulin dosage metered glucose value, nutritional intake, lifestyle, and emotional factors.

This system gives output as real time prediction of glucose. Using such technique processing time is reduced than time lagged ff. The model predicates 88.6%of normal glucose.

Kamer Kayaer[14] Present different types of neual network structure such as Multilayer perceptron(MLP),radial basis function(RBF) and general regression nural network(GRNN). Here PIMA indian diabetes dataset are used.These structure were applied to PIMA indians Diabetes(PID).Shows that performance of radial basis function was worse than Multilayer perceptron. General regression nural network (GRNN) , Multilayer perceptron (MLP) gives 80.21%,77.08% classification accuracy respectively.This paper shows using GRNN gives good and accuracy to classify medical data.

Hasan Temurtas[6]Paper present diagnosis of diabetes using multilayer neural network and probabilistic network of PIMA Indian diabetes database. Diagnosis of diabetes PIMA Indian diabetes dataset is used. Multilayer neural network are trained by Levenberg (LM) algorithm. Two classes used in PIMA Indian dataset and these classes contain 768 samples. Class1 contain 500 samples and class2 contain 268 samples. And eight attribute are consider. Using such techniques diagnosis of diabetes is easy.

Kamer Kayaer , Tülay Yildirim [14] Paper present feed forward neural network. Using such type of network diagnosis of problem is easy. Paper represents not only diagnosis of diabetes but also diagnosis of breast cancer. In this paper PIMA Indian diabetes dataset are used that database contain 768 samples taken from patients. Each sample is described by 8 features. Here taken 500 samples from patients who do not have diabetes & 268 samples that have diabetes.

From above data, they randomly selected 345 samples for training, 39 samples for cross validation and 384 samples for testing. Using such type of data they diagnosis of diabetes easily. Network is trained with an augmented cross entropy error function.

Main advantages are its reduced risk of data over fitting and reduced cost of future data acquisition. Disadvantage is that its time consuming.

Poonguzhali.E [7] Paper present algorithm is back propagation neural network for diagnosis of type 2 diabetes. Neural network is designed, trained and tested in mat lab. Aim of this paper is to using neural network system can improve the strategy to highest level where artificial metaplasticity on perceptrons is implemented [7].That is diabetes of type 2 is diagnosis using neural network with high level of performance. Few numbers of nodes are given to network such as thirst, hungry, nausea, fatigue, vomiting etc.

Procedure of implementing such type of network is data collection, post processing, training system, and testing system. In dataset collection 100 dataset are used which contains people from different age, weight, height, gender etc...Input node contains 13 symptoms as parameter that is 100*13=1300 node is given as input. Output also contains 100 parameter.

Ebenzer obaloluwa Olaniyi[15] Paper present multilayer feed forward neural network which is trained by back propagation algorithm. Sigmoid transfer function was used in hidden layer and also output layer. Output give 82% recognition rate and compare to other algorithm it is good.

Advantage is that they give highest success rate than other algorithm such as c4.5, EM propagation.

SonuKumari and Archana Singh [8] paper present an effective technology for the automated diagnosis of Diabetes. From sitting at home user can diagnose of diabetes whether he/she is suffering or not [8]

2) Decision tree and Naïve Bays

Naïve bayes is machine learning technique for constructing classifier. It is simple probabilistic classifier based on bayes theorem with strong independent assumption**.** It is highly scalable, requiring a number of parameters linear in number variable. It's called also independence bayes. Naive bayes classifier is also used to diagnosis of diabetes in early as possible.

Aiswarya Ilyer [27] Paper present two methodologies decision tree and naïve Bays. It gives simpler solution to the problem for diagnosis of diabetes especially in women. Here used J48 decision tree. Input is PIMA Indian diabetes dataset in csv format. Classification type of data mining has been applied to PIMA Indian diabetes dataset and preprocessing are done using weka tool.

Decision tree is a tree structure, which is form of flow chart. Using nodes and internodes classification and prediction are done. Roots and internodes are used as test cases that separate the instances with different features. Internal nodes are result of attribute cases. Leaf nodes denote the class variable. Class variable determine if person has diabetes or not. Output of decision tree gives either tested-positive or tested negative.

Naïve bays are sequential in natures. Bays algorithm is applied for overcome limitation of existing system. It is applied on larger dataset in real time. Using Naïve bays and decision tree diagnosis of diabetes is efficient way.

G. Parthiban [9] Paper present is to predict chances of diabetic patient getting various diseases like heart disease. In this paper, here applying Naïve Bayes data mining classifier technique which produces an optimal prediction model using minimum training set. Using such attribute such as age, sex, blood pressure and blood sugar and find the diabetic disease like heart disease.

AkaraSopharak [10] presented investigates and proposes a set of optimally adjusted morphological operators to be used for exudates detection on diabetic retinopathy patients' non-dilated pupil and low-contrast images [10]. Using naïve bayes technique they diagnosis of diabetic retinopathy patient. Naive bayes classifier requires small amount of training data for classification. It can be used for both binary and multi class classification problems. [10]

Disadvantage of naïve bayes is iteration are numerous, binning of continuous arguments and high computational time.

3) Support vector machine.

V. Anuja Kumari [11] presented Support Vector Machine (SVM), that machine learning method as the classification technique for diagnosis of diabetes with high level of performance. SVM focuses on classification of diabetes disease from high dimensional medical dataset. Data are trained by using SVM supervised learning.

Advantage of support vector machine they give flexibility to diagnosis of disease. They also give unique solution.

Disadvantage of using support vector machine is lack of transparency of result.

4) ID3, C4.5, Cart

Saba Basher el. At [12] presented best ensemble-based classification technique for diabetes datasets rather than single method. Paper present three types of decision tree such as ID3, C4.5and CART are used as base classifier. Proposed work Shows better performance as compared to single as well as other ensemble techniques**.**

Table 2: Performance of algorithm

Algorithm/Technique	Accuracy %
Neural Network[1]	80.88
Svm[3]	78
Naïve bays[4]	75
C4.5 [2]	71.4
Id3[2]	64.8
Cart[2]	83.2

In table 2 shows performance accuracy in various algorithms. Here give performance accuracy in percentage such as Id3, C4.5, Naïve Bayes, SVM, Neural network and Cart.Table2 and fig. 2 shows the Cart having highest performance accuracy for diagnosis of diabetes.

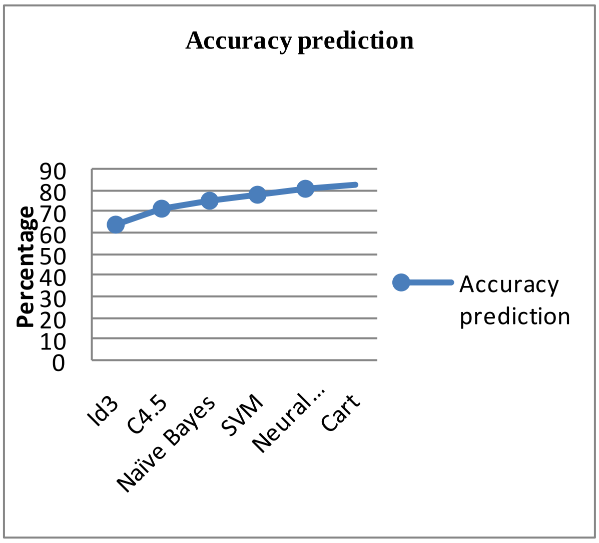
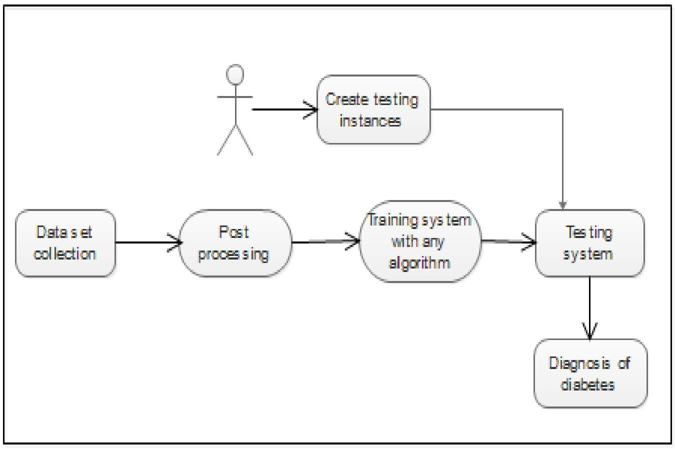


Fig 2. Accuracy prediction

IV. BASIC SYSTEM ARCHITECTURE



1. Dataset collection:
Various types of data set are available which contains different attribute of person such as weight, Height, Gender, Blood group, Symptoms according to considering data it gives as input to system.

2. Post Processing:
In dataset contains various null values or incomplete information. So post processing is done in original data set and null value removed.

3. Training system:
In training all below function are performed.
a)Read all input data set.
b)All activation function and derivatives selection are performed by system.
c) Particular algorithm performed such as back propagation, feed forward neural network, naive bayes, svm etc…
d) Create such type of function that generates network error.
e) Implement the train function

4. Testing System:
Testing system means to check according to symptoms system Diagnosis of disease properly or not. User gives query that is testing instance is created. And that test instances gives to testing module. Considering symptoms testing module diagnosis of diabetes. They diagnosis of diabetes in Yes or No format. If diabetes occurs then give yes otherwise no.

V. CONCLUSION

Diagnosis of diabetes is real world important problem in medical field. This paper shows how to classify techniques such as Neural Network, Naive Bayes, SVM, C4.5, CART, ID3which are used for diagnosis of diabetes. Then accuracy of classification techniques are compared and plot the graph shows in (fig.[2])according to accuracy prediction.

In future there will be planning for diagnosis of diabetes with high accuracy and less time. Also recommend treatment will be providing to the patient according to type of diabetes which is diagnosed.

ACKNOWLEDGMENT

I express great many thanks to Prof. Suvarna Pawar for their great effort of supervising and leading me, to accomplish this work. To college and department staff, they were a great source of support and encouragement.

REFERENCES

[1] Jaafar, S.F.B. and Ali, D.M., Diabetes mellitus forecast using artificial neural network (ANN)‖, IEEE, 2005.

[2] D.Senthil Kumar, G.Sathyadevi and S.Sivanesh Decision Support System for Medical Diagnosis Using Data Mining ,IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 1, May 2011

[3] Saba Bashir, Usman Qamar, Farhan Hassan Khan, M.Younus Javed An Efficient Rule-based Classification of Diabetes Using ID3, C4.5 & CART Ensembles

[4] G.Visalatchi1, S.J Gnanasoundhari2, Dr.M.BalamuruganG.Visalatchi et al A Survey on Data Mining Methods and Techniques for Diabetes Mellitus, International Journal of Computer Science and Mobile Applications, Vol.2 Issue. 2, February- 2014, pg. 100-105 ISSN: 2321-8363

[5] Scott M Pappada,“ Neural Network-Based Real-Time Prediction of Glucose in Patients with Insulin-Dependent Diabetes”, http://www.researchgate.net/publication/49801426, FEBRUARY 2011

[6] Hasan Temurtas a, Nejat Yumusak b, Feyzullah Temurtas,”A comparative study on diabetes disease diagnosis using neural networks”, Expert Systems with Applications 36 (2009) 8610–8615

[7] Poonguzhali.E1, Sabarmathi Kabilan2, Sandia Kannan3, Sivagami.P4,“Diagnosis of Diabetes Mellitus Type 2 using Neural Network”, Website: www.ijetae.com (ISSN 2250-2459, ISO 9001:2008 Certified Journal, Volume 4, Issue 2, February 2014).

[8] SonuKumari and Archana Singh, “A Data Mining Approach for the Diagnosis of Diabetes Mellitus”, 978-1-4673-4603-0112/$31.00 ©2012 IEEE.

[9] G. Parthiban, A. Rajesh Professor,S.K.Srivatsa,” Diagnosis of Heart Disease for Diabetic Patients using Naive Bayes Method”, International Journal of Computer Applications (0975 – 8887) Volume 24– No.3, June 2011 7

[10] Akara Sopharak,Bunyarit Uyyanonvara, Sarah Barmanb, Thomas H. Williamson c”Automatic detection of diabetic retinopathy exudates from non-dilated retinal images using mathematical morphology methods” , Computerized Medical Imaging and Graphics 32 (2008) 720–727.

[11] V. Anuja Kumari, R.Chitra “Classification of Diabetes Disease Using Support Vector Machine”. Vol. 3, Issue 2, March -April 2013, pp.1797-1801.

[12] K.H. Anders. A Hierarchical Graph-Clustering Approach to find Groups of Objects. Proceedings 5th ICA Workshop on Progress in Automated Map Generalization, IGN, Paris, France, 28{30 April, 2003.

[13] Centers for Disease Control and Prevention. National diabetes fact sheet: national estimates and general information on diabetes and prediabetes in the United States, 2011. Atlanta, GA: U.S. Department of Health and Human Services, Centers for Disease Control and Prevention. (Accessed April 5, 2011).

[14] Kamer Kayaer , Tülay Yildirim “Medical diagnosis on Pima Indian diabetes using general regression neural networks”, Yildiz Technical University JANUARY 2003.

[15] Ebenzer obaloluwa Olaniyi, ”Onset diabetes diagnosis using artificial neural network”. International Journal of

scientific and engineering research,volume 5,issue 10,oct2014.

[16] Zhang Y, Dall T, Mann SE, Chen Y, Martin J, Moore V, Baldwin A, Reidel VA, Quick WW. “The economic costs of undiagnosed diabetes”. Popul Health Manag. 2009;12(2):95–101

[17] Harleen Kaur and Siri Krishan Wasan, “Empirical Study on Applications of Data Mining Techniques in Healthcare”,Department of Mathematics, Jamia Millia Islamia, New Delhi-110 025, India. Journal of Computer Science 2 (2): 194-200, 2006 ISSN 1549-3636 © 2006 Science Publications

[18] Stanfford, G.C., P.E. Kelley, J.E.P. Syka, W.E.Reynolds and J.F. Todd, 1984. Recent improvements in and analytical applications of advanced ion-trap technology. Intl. J. Mass Spectrometry Ion Processes, 60: 85-98.

[19] a1rk Hudson Beale, Martin T. Hagan and Howard B. Demuth, “Neural Network Toolbox™ User’s Guide”2009

[20] Murali Shankar,” Using neural network to predict the onset of diabetes Mellitus”. 1996

[21] Aiswarya Iyer, S. Jeyalatha and Ronak Sumbaly, “CLASSIFICATION MINING TECHNIQUES” , International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol.5, No.1, January 2015.

[22] Asha Rajkumar1,Mrs. G.Sophia Reena,”Diagnosis Of Heart Disease Using Datamining Algorithm”. Page 38 Vol. 10 Issue 10 Ver. 1.0 Sepetember 2010.

[23] Nitin Bhatia,Vandana,‖ Survey of Nearest Neighbor Techniques‖ (IJCSIS) Vol. 8, No. 2, 2010, ISSN 1947-5500.

[24] Charanjeet Kaur, ―Association Rule Mining using Apriori Algorithm: A Survey‖, IJARCET Volume 2, Issue 6,June 2013.

[25] V.Karthikeyani, I.Parvin Begum, I.Shahina Begam K.Tajudin, “Comparative of Data Mining Classification Algorithm (CDMCA) in Diabetes Disease Prediction” , volume 60- No.12 December 2012

[26] K. R. Lakshmi and S.Prem Kumar, “Utilization of Data Mining Techniques for Prediction of Diabetes Disease Survivability”, International Journal of Scientific & Engineering Research, Volume 4, Issue 6, June-2013 ISSN 2229-5518

[27] Aiswarya Ilyer , “Diagnosis of diabetes using classification mining techniques”, international Journal of data mining & knowledge management process vol.5,no.1 ,January 2015.

[28] https://www.google.co.in/search?hl=en&site=imghp&tbm=isch&source=hp&biw=993&bih=636&q=symptoms+of+diabetes&oq=symptoms+of+&gs_l=img.1.5.0l10.82.3875.0.6558.11.11.0.0.0.0.305.1665.0j1j4j2.7.0....0...1.1.64.img..4.7.1661.WbILfn-TzzU#imgrc=yrmwOlTWU-t8vM%3A