International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-9 Issue-3, February 2020

Diagnosis of Type-2 Diabetes using Classification and Mining Techniques

Sankar Padmanabhan, Manjunath K M, Madhurima V

Abstract*: Around two hundred and fifty million individuals, with a major part of them being ladies influenced by diabetes. This number may ascend to 380 million by another decade. The sickness has been named as the fifth deadliest illness in the world with not a single inevitable fix to be seen. With the ascent of data innovation and proceeding with an approach into the restorative and medicinal services part, the instances of diabetes and their side effects all around are archived. Information mining is a buzz word separating concealed data from an enormous arrangement of database. It assists scientists in building large database in the area of biomedical engineering. The Pima Indian diabetes database was used for investigation purpose. In this paper an attempt has been made to study the effect of various classification and mining Techniques like Decision Tree, Naïve Bayes, SVM, Regression etc on the diagnosis of Type-2 diabetes.

Keywords: Algorithms, Heart rate variability, J48, Regression, SVM

I. INTRODUCTION

The ECG analysis has a vital significance in analysing the various heart disorders namely, Arrhythmias, Myocardial Infarction (MI), Coronary Heart Disease (CHD), Ischemia, Cardiomyopathy, Heart attack, Aortic Aneurysm etc. The hermitian basis function and the Discrete Wavelet Transform (DWT) were some of the ECG feature extraction methods used, with which the R-R intervals were extracted for HRV analysis. The HRV analysis serves as an additional clinical and research tool for the cardiologists and researchers. In 1965 ,Professor Hon Lee has initiated the importance of HRV while doing research in the monitoring of fetal distress in women [1]. HRV finds its importance in patients with MI, Diabetes Mellitus, CHD, Chronic Heart Failure (CHF), Hypertension etc and so on. In addition to that it has been analyzed in swimmers, athletes, cyclists, foetuses and children of various age groups. Amongst all such people, HRV analysis has a marked significance in Type-2 Diabetic patients as Diabetes mellitus (DM) is major, fast growing and one of the health issues encountered by the world community. In 2030, Diabetes will be one of the leading causes of death and Type 2 Diabetes Mellitus comprises 90% of it. Hence HRV became a non-invasive tool for investigating the autonomic dysfunction related to Type-2 diabetes.[2]. Cardiology Society of European Union and the Society of Pacing and Electrophysiology in North America combine developed a task force for giving aid to these researchers. This task force developed up some standards for the measurement of Heart rate variability [3]. Based on these guidelines, this paper has been formulated by surveying and consolidating various methods, and the applications used in the HRV analysis in the current decade.

The information obtained from patient records from hospitals was used to make inferences by applying several mining techniques. This mining technique helps the detection of diabetes in women at an early stage. After applying several algorithms on the database the results are obtained, compared and tabulated. The organization of the paper is given in various sections.

Revised Manuscript Received on February 05, 2020.

* Correspondence Author

Sankar.P*, ECE, SV Engineering College, Karakambadi, Tirupati, AP Email: sankar.padmanaban@svcolleges.edu.in

Manjunath.K.M, ECE, SV Enguneering College, Karkambadi road, Tirupathi, AP. Email: manjunath.km@svcolleges.edu.in

Madhurima. V, ECE,SV Engineering college, Karkambadi road, Tirupathi, AP . Email: madhurima.v@svcolleges.edu.in

II. HEART RATE VARIABILITY MEASUREMENTS

The linear and non linear are the two methods used to analyze the Heart rate variability. the linear method always assume that the R-R interval is stationary. The R-R interval can also be mentioned as NN Interval which means that, all the processed beats are normal, without ectopic beat (skipped or extra heartbeats as compared to that of the normal ones). Due to physical or postural activities and small disturbances like Premature Ventricular Contraction heart rate can go random in nature while checking. non-linear methods are more effective than the linear methods of HRV analysis, For correct prediction in the variations of heart non linear methods are effective than linear ones

A. Linear methods

The time domain analysis, frequency domain analysis and geometrical analysis are categorized as linear methods.

B. Time domain measures of HRV

The time domain measures of HRV Consist of SDNN SDANN, RMSSD, SDSD and NN 50 Count this standard deviation route mean square values are measured in the NN interval.

C. Geometrical measures of HRV:

- Triangular Index: The ratio between the total NN intervals and the histogram height of all NN intervals in the time scale of milliseconds.;
- Poincare plot: The R-R interval is normally plotted as an instance of the previous interval.

- Differential Index: NN height intervals (at 100 and 1000 sample levels) are computed to which the histogram is applied and the differences are studied.

D. Frequency measures :

The frequency domain parameters for 5 minutes duration will give the low frequency normal values ,high frequency normal values and the total power obtained.

E. Non-linear HRV Methods

- Approximate Entropy (ApEn) *:* ApEn expresses or measures the quantity of the irregular heart rate time series considering parameters like the length of the data (N), threshold(r) and the embedding dimension (m). The advantage of this analysis method is that, it requires only small data samples (n<50) to analyse the HRV signal. The comparison of ApEn values between two datasets must be higher for all conditions tested, which is not satisfactory, and consistent [4] [5].
- Correlation Dimension (CD) *Analysis*: The CD technique calculates the space dimensionality occupied by a set of random points, which in turn, varies with pathological condition of the patient. It is very useful in indicating the various disorders such as congestive heart failure, and ventricular tachycardia. The main limitation of this method is its inability to distinguish chaos from noise [6][7].
- Higuchi's Fractal Dimension (HFD*):*Fractals are infinite and complex patterns, which are self-similar in various time scales. The HFD algorithm is one of the methods used in measuring fractals of the discrete time series sequences. This algorithm is simpler and faster. It also requires less number of points calculate FD estimate and is highly robust even in the presence of noise. During epileptic seizures the Higuchi's algorithm produces varying results of a particular type (increased or decreased value) [8].
- Detrended Fluctuation Analysis *(DFA):*The DFA method assesses and measures the correlation properties of fractals in a non-stationary time series. Advantage of the method is that, it detects the internal self-similarity features in an apparently in non-stationary time. The DFA method requires many data points and N-N inter-beat intervals [9].
- Symbolic Dynamics (SD):- SD characterizes the coarse-grained dynamics of HRV. They are efficient in describing the short-term features of beat-to-beat variability. Symbol strings are lost due to ectopic beats and noise in the HRV signal [10].
- MultiScale Entropy (MSE):- The heart signals contain information on spatiotemporal time scales. Based on that, the method of entropy calculation involved at different time scales is termed as MSE. The method is applied to both physical and physiologic input datasets. Ectopic beats affect the entropy values. As the data points decrease, the MSE consistency is lost [11].
- Poincare Plot:- It is a pictorial method, which depicts the underlying structure of time series of R-R interval into a phase space. The time series of R-R interval gives a point in the pictorial representation which indicates a pair of continuous elements. This is plotted in the graph. The Poincare plot differentiates the heart rate complexity patterns, which is not possible by the use of time domain measures like the SDNN [12]. A Complex Correlation measure predicts and distinguishes between various shapes of Poincare plot when compared to SD1 and SD2 [13].

III. RELATED WORKS

The prediction models designing for diabetes diagnosis has been the area of research for the last decade. Artificial neural networks and clustering algorithms are used extensively for modeling as given by many references. In another work the authors have employed three techniques namely: J48 algorithm, Naive Bayes and K-means clustering algorithms are applied on the diabetic patients. It proves that K-means outperforms all algorithms [14]. When the hidden knowledge from a database is used for diagnosis and prescribing the medicines, health of many diabetic patients was able to be improved considerably [15]. In these researchers used the Pima Indian database for the diabetes diagnosis. This research has been carried out in finding cost effective methods for the diagnosis of Diabetic patients. Still a model is needed for establishing a relationship between diabetic parameters [16].

IV. METHODOLOGIES

The present works addresses the diabetic conditions prevailing in women and suggest a model for classification which provides an easier solution to the diagnosis problem. Several methods are used for analysis of results obtained

V. KNOWLEDGE DISCOVERY (KD)

Mining is mainly used for knowledge discovery. This knowledge discovery process involves the finding of the raw data, pre-processing it into a suitable form, transforming and obtaining the relevant data by extracting out of it. Fig 1 shows the KD process which consists of the previous mentioned processes.

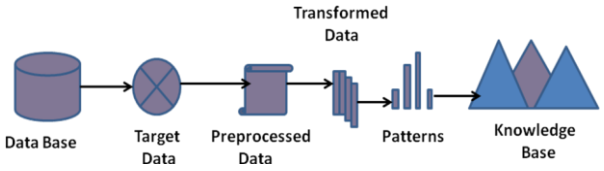

Fig. 1. KD Process

Mining of data is used extensively after the popularity of internet among the common masses. But the requirement for various tools and systems made mining a greater challenge to the society. Now mining is used in telecom, finance, biological, retail business, scientific and medical applications.

VI. MACINE LEARNING ALGORITHMS

There are several machine learning algorithms that can be used for classification and prediction. Some of them are logistic regression decision trees, SVM, Naïve Bayes, and so on..Such types of algorithms are explained in this section.

A. Decision Trees

A tree structure needed for method of prediction and classification inter nodes and nodes in order to separate instances with the different features, the text case which are the internal nodes and root are used. The result of attribute test cases is given by the internal nodes. Class variables are denoted by the leaf nodes A decision tree structure shown in figure.2

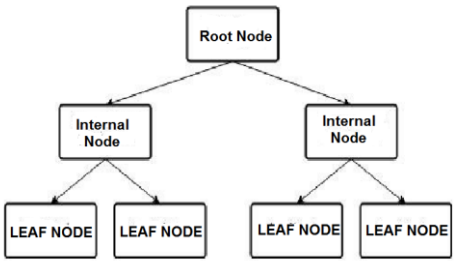

Fig. 2.Decision Tree Structure

This provides powerful tool for diabetic diagnosis the following all some decision tree classification algorithms frequently used by reseachers. They are ID3,C45,C5,J48,CART and CHAID to estabilish a model J48[16] has been chosen decision tree algorithm.the height gain is calculated for the attributes of each node. if any ambiguity is found in any attributes the branch is terminated and assign a taget value

B. Logistic Regression

Applications in which data is linear in nature and adequate for the order approximation, linear regression finds it self useful in many applications linear regression is not appropriate an alternate regression technique is logistic regression.

- Linear Model:

  Assume that for every data set X,Y there are binary outcomes. in a single experiment x_i let y_i Be an outcome with either 1 or 0.if the outcome is one it is set belong to positive class otherwise negative class logistic function is given in equation1.

$$\alpha(x,\gamma) = \exp(\gamma T\, x)/1+\exp(\gamma T\, x) \quad (1)$$

where,. T is expected value and γ is vector of parameters

Thus the regression model is given by equation2

$$y = \alpha(x,\gamma) + € \quad (2)$$

where, € is error with one of the two values if y=1 then it gives €=1- α(x) else €=α(x).

C. SVM (Support vector machine)/ SMO

SVM solves the Quadratic Programming (QP) optimization problem which emerges during the training of data. It does so by breaking down the QP problem into smaller optimization sub-problems. SMO is faster than the other SVM algorithms and has better scaling than any other SVM algorithms, since it uses the smallest possible QP problem. It consists of two parts, namely an analytical solution to a QP problem of the two Lagrange multipliers, and a set of heuristics designed to efficiently choose which multipliers to optimize. SVM is a supervised learning approach SVM super imposes the data for training in to a higher space and splits the instance in to various categories by dividing them in to linear and non linear. SVM tries to keep separation boundary between two different categories (classes) as wide as possible. The perpendicular bisector of the shortest line connecting the two classes is called hyper plane. The hyper plane is far from both classes. The training instances closest to the hyper plane are called support vectors. The support vectors are very important, because they determine the hyper plane, while the other instances might be forgotten. After drawing the hyper plane, the test instances are mapped into the same training space. A class value is determined for each test instance by SVM model. Fig.3 shows SVM for linearly separable data.

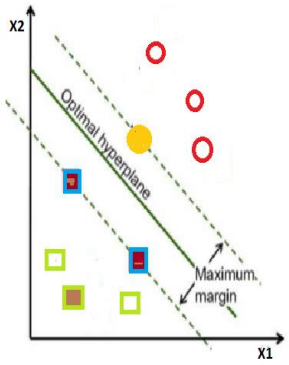

Fig. 3. SVM for linearly separable data

D. Naïve Bayes

Naïve Bayes is probabilistic sequential algorithms it has several steps of execution estimation prediction and classification There are several data mining algorithms which give solution for the relationship between symptoms medicines and diseases. But this algorithms has got limitations such as high computation time,number of iterations etc. Naïve Bayes over comes this limitations and can applied and large data sets it works on the formula as per equation 3,

$$PP = L*CPP/PPP \quad (3)$$

where, PP=Posterior Probability

L= Likelihood

CPP=Class Prior Probability

PPP = Predictor Prior Probability

VII. RESULT AND DISCUSSION

This paper summarizes the characteristics of diabetes, types of measurements of Heart rate variability (HRV) and its effect on various machine learning algorithms when applied to a diabetes data set. Data set are sorted and extensively used in the machine learning algorithms. Additional exploration has been carried out on determining the diabetes seen on women population using machine learning algorithms like Naïve Bayes decision tree the main intension is to predict whether the patient is effected by diabetes are not for this PIMA Indian diabetes database has been extensively used for simulation Transformation and pre processing of the data base is carried out using WEKA Software [17].

Replacement of mining values and normalization of it are the two steps involved in the process As the variables are in the range of 0 to 1,the later makes it easy to use the data base. The statistics of dataset are presented in Table I. The parameters are normalized and it ranges from 0 to 1 .

Table- I: Data set attributes

Parameter	Min	Max	Mean	Std deviation
Plas	0	1	0.608	0.161
Mass	0	1	0.477	0.117
Pedi	0	1	0.168	0.141
Age	0	1	0.204	0.196

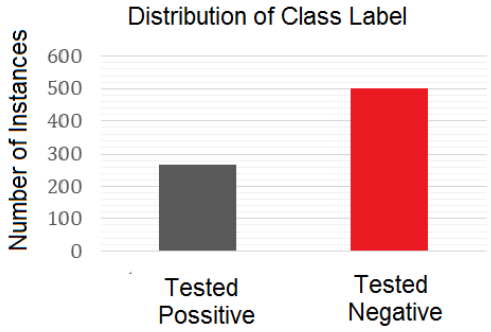

Fig. 4. The distribution of various attributes

The database consists of both diabetic and non–diabetic patients. After applying this to the weka software it indicates that the number of people who are not having diabetes is found to be around 498 and those who are tested positive comes around to 270.This distribution is shown in Fig.4.

Finally the prominent algorithms are applied on the Pima Indian database using python and the results are tabulated as given in the Fig.5

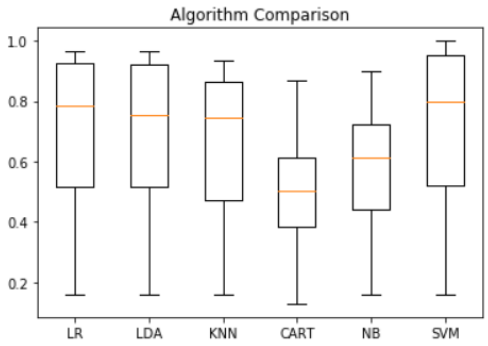

Fig. 5. The Comparison of various algorithms

Table- II: Comparative Scores

S,No	Algorithms	Parametric Accuracy
1	Logistic Regression	0.686705 (0.273225)
2	Support Vector machines	0.701341 (0.283481)
3	K Nearest Neighbor classifier	0.650058 (0.259353)
4	Naïve Bayes	0.580390 (0.232160)
5	Linear Discriminant analysis	0.678269 (0.268841)
6	Classification and regression Trees	0.501769 (0.221501).

The comparative scores of the algorithms was found to be with (a)Logistic Regression,(b)Support Vector machines (c)K Nearest Neighbor classifier (d) Naïve Bayes (e) Linear Discriminant analysis (f) Classification and regression Trees .It is given in Table II. Finally, it has been observed that SVM gives a score of 70 % and good for parametric optimization.

ACKNOWLEDGMENT

The authors would like to thank the unknown reviewers who reviewed the work.

REFERENCES

1. Hon EH, Lee ST. Electronic evaluations of the fetal heart rate patterns preceding fetal death, further observations. Am. J. Obstet Gynec. 1965; 87: 814-826.
2. Wild S, Roglic G, Green A, Sicree R, King H. Global prevalence of diabetes: estimates for the year 2000 and projections for 2030. Diabetes Care. 2004; 27 (5):1047–1053.
3. Marek M. Guidelines: Heart rate variability- standards of measurement, physiological interpretation, and clinical use. Eur Heart J. 1996; 17: 354–381.
4. Francesco B, Grazia B, Emanuele G, Valentina F, Sara C, Chiara F, Riccardo M, Francesco F. Linear and nonlinear Heart Rate Variability Indexes in Clinical Practice. Comput Math Methods. 2012; Article ID 219080,
5. Renu Madhavi CH and Ananth AG. Analysis and Characterisation of Heart Rate Variability (HRV) Data of Different Sets of Subjects Using Nonlinear Measure (Approximate Entropy). Int. J Comp. Theory Engg. 2010; 2 (4): 619-623.
6. Asha VT, Rohini RM. HRV Analysis of Arrhythmias Using Linear – Nonlinear Parameters. Int. J Comp Appl. 2010; 1(12): 71-77.
7. Ramesh K Sunkaria. Recent Trends in Nonlinear Methods of HRV Analysis: A Review. World Academy Sci. Engg. Technol. 2011; 51: 444-449.
8. Carlos G, Angela M, Roberto H,Daniel A and Alberto F. Use of the Higuchi's fractal dimension for the analysis of MEG recordings from Alzheimer's disease patients. Med Eng Phys. 2009; 31(3):306–313.
9. Polychronaki GE,Ktonas PY,Gatzonis S, Siatouni S,Asvestas PA, Tsekou H, Sakas D and Nikita KS. Comparison of fractal dimension estimation algorithms for epileptic seizure onset detection. J Neural Eng. 2010; 7(4):046007.
10. Andreas voss, Steffen S, Rico S, Mathias B and Pere C. Methods derived from nonlinear dynamics for analysing heart rate variability. Phil. Trans. R. Soc. A. 2009; 367 (1887): 277–296.
11. Harish K, Kamaldeep K and Gurpreet K. Heart variability analysis by using non-linear techniques and their comparison. Int J Comp Appl. 2013; 65(20): 33-6.
12. Chandan KK, Ahsan HK, Andreas V and Marimuthu P. Sensitivity of temporal heart rate variability in Poincare plot to changes in parasympathetic nervous system activity. Biomed Eng Online; 2011; 10:17.
13. Renu Madhavi CH and Ananth AG. A Review of heart rate variability and it's association with Diseases, International Journal of Soft Computing and Engineering. 2012; 2(3): 86-90.
14. Sankaranarayanan.S and Dr Pramananda Perumal.T, Predictive Approach for Diabetes Mellitus Disease through Data Mining Technologies, World Congress on Computing and Communication Technologies, 2014, pp.231-233

15. T.Jayalakshmi and Dr.A. Santhakumaran, "A Novel Classification Method for Diagnosis of Diabetes Mellitus Using Artificial Neural Networks", International Conference on Data Storage and Data Engineering, 2010, pp.159-163
16. Neeraj Bhargava, Girja Sharma, Ritu Bhargava and Manish Mathuria, Decision Tree Analysis on J48 Algorithm for Data Mining. Proceedings of International Journal of Advanced Research in Computer Science and Software Engineering, 2013, 3(6).
17. Rossen Dimov, Weka: Practical machine learning tools and techniques with Java implementations, AI Tools Seminar University of Saarland April 30, 2007,pp 1-20.

AUTHORS PROFILE

Sankar. P , received his B.Sc degree in Mathematics from the University of Kerala in 1986, B.E in Electronics and Telecommunication from the University of Poona, India in 1994 and M.E in Communication Systems from Madurai Kamaraj University in 1997. He received his Ph.D. degree in Information and Communication from Anna University,Chennai, India.He is currently working as Professor in the Department of Electronics and Communication Engineering, SV Engineering College For Women, Tirupati, AP, India. He has published papers in national and international journals. He is a member of the IEEE, Fellow of Institution of Engineers (India), and the life member of the ISTE. He has published books on Multimedia and Integrated Electronics His areas of interests include Multimedia information systems, Image and video processing and Computer Networks.

Manjunath. K. M, received his Bachelor's degree with First class honors in Electronics and Communication Engineering in 2005, and his Master's degree in Computer and Communication Engineering 2011, from Jawaharlal Nehru Technological University Kakinada. He is currently working as an Assistant Professor in the Department of Electronics and Communication Engineering,,SV Engineering college,Tirupathi,AP. he is pursuing PhD in VIGNAN'S University He is working Research area in image processing using Machine learning algorithms applying for autism spectrum disorder (ASD) children.

Madhurima. V, received her Bachelor's degree with First class honors in Electronics and Control Engineering in 2004, and her Master's degree in Digital Systems and Computer Electronics 2012, from Jawaharlal Nehru Technological University Ananthapuramu. She is currently working as an Associate Professor in the Department of Electronics and Communication Engineering, SV Engineering College.Tirupati,AP,India. She is pursuing PhD in Jawaharlal Nehru Technological University,Kakinada,AP,India.Her areas of interests include VLSI, Image and Video Processing.