A Survey Paper on Evolving Techniques for the Prediction of Type 2 Diabetes.

RATNA NITIN PATIL
Computer Science and Engineering
Vishwakarma Institute of Technology,
Pune (INDIA)
ratna.patil@vit.edu

Dr. SHARVARI CHANDRASHEKHAR TAMANE
Computer Science and Engineering
Jawaharlal Nehru Engineering College,
Aurangabad (INDIA)
sharvaree73@yahoo.com

***Abstract*—Diabetes Mellitus (DM) is a metabolic disease where the person will have high blood sugar due to the pancreas unable to produce sufficient insulin or the cells which are not responding to the insulin produced. There are three types of diabetes. They are Type 1 diabetes, Type 2 diabetes and Gestational diabetes. Type 1 diabetes is mostly occurring in children. Type 2 diabetes is called adult-onset diabetes which is common in adults. Gestational diabetes is only in women during pregnancy. Disease diagnosis is one of the applications where machine learning algorithms are giving successful results. This paper identifies gaps in the research on Type 2 diabetes disease diagnosis and treatment. Different classifiers can be used to explore patients' data and extract a predictive model. The importance of early diagnosis associated with the appropriate treatment is to decrease the chance of developing further complications like nerve damage, kidney failure, heart disease, diabetic retinopathy. Machine learning algorithms can provide reliable performance in determining diabetes mellitus. The focus of this paper is to study the recent algorithms used for diagnosis of Type 2 diabetes and find the research gaps.**

Keywords- Diabetes mellitus; early diagnosis; GA; machine learning; classification.

I. INTRODUCTION

Diabetes is a deadly disease and a major public health challenge worldwide. The number of diabetics in India is doubled from 32 million in 2000 to 63 million in 2013 and the figure is projected to further increase to 102.2 million in the next 15 years. This is the latest assessment by the World Health Organization, raising an alarm over the need to treat the condition. The annual spend on diabetes treatment in India is pegged at Rs1.5 lakh crore, which is 4.7 times the Centre's allocation of Rs32000 crore for health. This cost is projected to rise by 20-30% every year.

Diabetes disease diagnosis via proper interpretation of the Diabetes data is an important classification problem. There are several methodologies available on classification of diabetes disease. But less work has been done on early detection of the disease.

This work will help to develop a predictive model based on set of attributes collected from the patients to develop a mathematical model. It is essential to find a way that can help in early detection with high accuracy and less complexity.

MATERIAL

A. DATASET

In the machine learning research community, a work is going on to solve the classification problem. Pima Indian Dataset (PIMA) has been used to test the classification performance by most of the scholars. It is publicly available in the machine learning dataset UCI. All the instances in this dataset are Pima Indian women of at least 21 years old and living near Phoenix, Arizona, USA. The data is a collection of 768 records.

B. Risk Factors

The following are the parameters which contributes to the development of diabetes. The prevalence of Type 2 diabetes is increasing at a fast pace due to obesity, physical inactivity and unhealthy dietary habits.

- Age – Indians develop diabetes earlier than western population. An early occurrence gives abundant time for the development of prolonged complications of diabetes. The incidence of diabetes increases with an age.
- Family History – The occurrence of diabetes increases with a family history of diabetes. A high incidence of diabetes is seen among the first degree relatives.
- Lifestyle – Deskbound lifestyle is an independent factor for the growth of Type2 diabetes.
- Obesity – There is a close association of obesity with Type2 diabetes. Increase in weight increases Body Mass Index (BMI).
- Stress – The impact of physical and mental stress along with lifestyle changes has an effect of incidence of Type 2 diabetes from persons in a strong genetic background

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

III LITERATURE REVIEW

Table 1 Literature Review

Paper title	Methodology	Dataset	Tool	Advantages	Limitations	Accuracy
Intelligible Support Vector Machines For Diagnosis of Diabetes Mellitus IEEE transaction (2010)	Sequential Covering Approach for Rule Extraction (SQR ex-SVM), and the Eclectic method have been proposed for rule extraction to enable SVMs to be more intelligible means simply turns the "black box" model of an SVM to an intelligible representation of the SVMs diagnostic.	Data from 4682 subjects of age 20 years and above Was collected using a questionnaire regarding demographic data, history, and anthropometric measures.	-----	The extracted rules are comprehensible, simple, and medically sound.	The eclectic method rule is not consistent with accepted medical standard.	94% (Rule performance)- SQRex- SVM. 93% (Rule performance)- Electic method
Feature generation using genetic programming with comparative partner selection for diabetes classification ELSEVIER (SCIENCE DIRECT 2013)	Genetic Programming- K-Nearest Neighbour (GPKNN), Genetic Programming- Support Vector Machines (GPSVM) have been used, in which KNN and SVM tested the new features generated by GP for performance evaluation.	UCI Machine learning Repository.	-----	GP improves the performance and it also reduces the eight input dimensions to a single dimension.	Ignoring the Missing values Without giving any details which values were ignored.	80.5% (GPKNN). 87.0% GPSVM).
A Computational intelligence approach for a better diagnosis of diabetic patients ELSEVIER (SCIENCE DIRECT 2014)	Gini index - Gaussian Fuzzy decision tree algorithm for the diagnosis Gini index - Gaussian fuzzy decision tree algorithm	Pima Indian Diabetes (PID)	MATL AB	minimizes split point and Gini index calculations,	accuracy of the model can be improved by using better fuzzy membership functions which are applicable to diabetes clinical data.	75%
Medical Diagnosis on Pima Indian Diabetes Using General Regression Neural Networks	General Regression Neural Network (GRNN) has been used to classify a medical data in which the optimum spread values were found by trial and error and used for training and the classification of test data.	Pima Indian Diabetes (PID)	MATL A B 5.3, Neural Network Toolbox	GRNN is a simple and practical method to classify medical data.	GRNN require More memory space to store the model.	82.99% (prediction of training set). 80.21% (prediction of test set). 82.29% (Mean total prediction).
Design of a Diabetic Diagnosis System Using Rough Sets VERSITA (Bulgarian academy of sciences) 2013.	Rough set technique is used for the design and developed a diabetic diagnosis system	…….	Java, JSP, Oracle 11	It can be used to develop real time intelligent Systems and the Methodology can be extended for other disease Diagnosis.	The present System knowledge base is designed only for one disease and will need pathological interpretation for other disease diagnosis.	76% (Prediction)

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Paper title	Methodology	Dataset	Tool	Advantages	Limitations	Accuracy
Application of K-means & Genetic algorithms for dimension reduction by integrating SVM for Diabetes Diagnosis ELSEVIER (SCIENCE DIRECT 2015)	K-means is used for removing noisy data GA's for finding the optimal set of features with SVM as classifier	Pima Indian Diabetes (PID)	-----	The proposed method K-means+GA+SVM has given 98.82% accuracy	Standard Deviation can be used to replace missing values, box plot for outlier detection, PCA for feature selection & to experiment classifiers from statistical, neural fuzzy & tree families.	98.82% Out of 768 instances, K-means selected 511 samples, & 257 samples were detected as outliers.
A novel algorithm to diagnosis T2DM based on Association rule mining using MPSO-LSSVM with outlier detection method. Indian Journal of Science & Technology (2015).	CFP- growth algorithm for finding frequent patterns of the input data set. The resulting association rules from above algorithm are discovered using MPSO-LS-SVM with the integration of outlier detection	Pima Indian Diabetes (PID)	-----	Speedy convergence Requires less computational time than other algorithms. Eradicates the effect of unavoidable outliers in investigation sample on a scheme's performance.	Instead of Chi Merge Discretization method for preprocessing other algorithms like Swarm Optimization based ABC can be utilized and classification rate can be investigated.	95%
Detection of the Onset of Diabetes Mellitus by Bayesian Classifier Based Medical Expert System Transaction on Machine Learning and Artificial Intelligence DOI: 10.14738/tmlai.44.1962 Publication Date: 19th July, 2016.	Bayesian classification approach	Pima Indian Database	Java running on NetBeans IDE version 8.0.2.	findings show a good promise compared to previous works reported in the literature, requires a small amount of training data,	Due to the limitation of data, i.e. deficiency of samples, disease prediction does not seem to be much robust, reduction of the number of features are required for the improvement of classification.	more than 87% accuracy.
An Intelligent Type-II Diabetes Mellitus Diagnosis Approach using Improved FP-growth with Hybrid Classifier Based Arm Research Journal of Applied Sciences, Engineering and Technology 11(5): 549-558, 2015	Hybrid Enhanced Artificial Bee Colony-Advanced Kernel Support Vector Machine (HEABC-AKSVM-IFP Growth) classification based Association Rule Mining (ARM)	Pima Indian Database	----	offers better convergence performance as compared to standard ABC-LSSVM, Rules are extracted using HEABC-AKSVM-IFP Growth based Association Rule Mining can be used by physician to diagnose	Sometimes rules are trivial and poorly understandable and they describe only the relationships in the dataset.	Average number of rules generated is less as compared to existing algorithms, as number of rules generated increases the accuracy is increasing

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

IV Proposed Methodology

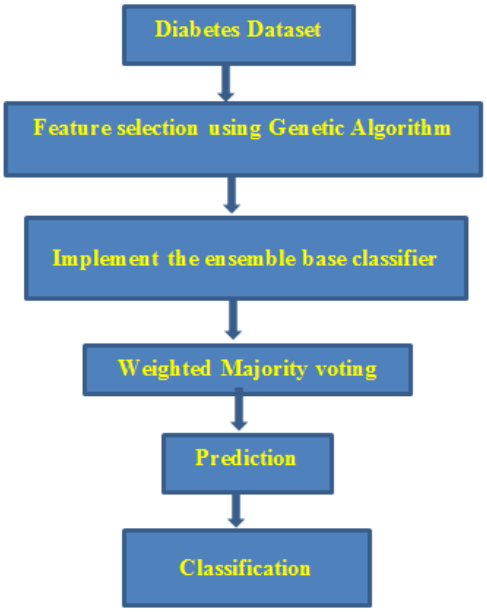

Figure 1 Proposed Classifier

V Comparison of Techniques employed

Table 2 Comparative Study

S. No.	Algorithm	Advantages	Limitations	Prime areas of Applications
1	Naïve Bayes	Easy to implement, requires a small amount of training data, simplicity, robustness.	Loss of accuracy, dependencies exists among variables	Text classification, spam filtering
2	Decision Tree	Do not require any assumptions of linearity in the data, easy to understand and generate rules	Overfitting, pruning is necessary, requires more number of training examples, computationally expensive to train	Agriculture, biomedical engineering, medicine
3	Support Vector Machine (SVM)	Training is easy, scales well to high dimensional data	Need for a good kernel function, not able to provide comprehensible justification for the classification decisions.	Image classification, Hand writing analysis
4	Neural Network (NN)	Mapping capabilities, fault tolerance, parallel and high speed information processing.	Needs training to operate, requires high processing time for large NN, not able to provide justification for the decision.	Pattern recognition, risk assessment, forecasting
5	Fuzzy Logic	Gives justification of decisions, robust, accepts vague data, simulates human control logic	System is complex, automatically rules can't be learned.	Bioinformatics, prediction systems, Business
6	Genetic Algorithm (GA)	Can be easily used in parallel machines, Global search methods, adaptive population based optimization technique	Can't find the exact solutions, no absolute assurance for a global optimum, can't assure constant optimization response times	Bioinformatics, quality control, Scientific and research analysis

VI Conclusion

Diabetes mellitus is a deadly disease which is one of the topmost public challenges around the world. It is a fact that 80% of Type 2 diabetes complications can be prevented by early identification of people at risk. Early detection of this disease has become an essential issue to improve the overall clinical efficiency of the diagnosis process. Motivated by the world-wide increasing mortality of diabetes disease patients each year and the availability of huge amount of data researchers are using Machine Learning techniques for classification. Machine Learning algorithms in the medical field extracts different hidden patterns from the medical data. They can be used for the

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

analysis of important clinical parameters, prediction of various diseases, forecasting tasks in medicine, extraction of medical knowledge, therapy planning support and patient management. This survey has taken various classification methods and ensemble them to give the new model in the search of finding the better result in terms of accuracy, specificity and sensitivity. Further, the work will be extended for diabetes by gathering the information from several hospitals and provide the more precise and general prescient model. The performance can be studied for different parameters for effective diabetes diagnosis. The work can be extended and improved for the automation of diabetes analysis.

REFERENCES

[1] Nahla H. Barakat, Andrew P. Bradley, Senior Member, IEEE, and Mohamed Nabil H. Barakat "Intelligible Support Vector Machines for Diagnosis of Diabetes Mellitus" IEEE Transactions on Information Technology In Biomedicine, Vol. 14, No. 4, July 2010 Digital Object Identifier10.1109/TITB.2009.2039485 July 2010.

[2] Muhammad Waqar Aslam, Zhechen Zhu, Asoke Kumar Nandi "Feature generation using genetic programming with comparative partner selection for diabetes classification" Expert Systems with Applications 40(2013) 5402-5412

[3] Kamadi V.S.R.P. Varma, Allam Appa Rao, T. Sita Maha Lakshmi, P.V. Nageswara Rao "A computational intelligence approach for a better diagnosis of diabetic patients" Computers and Electrical Engineering 40 (2014) 1758–1765

[4] T. Santhanam, M.S Padmavathi "Application of K-Means and Genetic Algorithms for Dimension Reduction by Integrating SVM for Diabetes Diagnosis" Procedia Computer Science 47 (2015) 76 – 83.

[5] Kamer Kayaer, Tulay Yildirim "Medical Diagnosis on Pima Indian Diabetes Using General Regression Neural Networks" Yildiz Technical University, Department of Electronics and Comm. Eng. Besiktas, Istanbul 34349 TURKEY.

[6] Margret Anouncia S., Clara Madonna L. J., Jeevitha P., Nandhini R. T. "Design of a Diabetic Diagnosis System Using Rough Sets" Cybernetics and Information Technologies Volume 13, No 3 DOI: 10.2478/cait-2013-0030.

[7] Mostafa Fathi Ganji, Mohammad Saniee Abadeh "Using fuzzy Ant Colony Optimization for Diagnosis of Diabetes Disease" Proceedings of ICEE 2010, May 11-13, 2010 978-1-4244-6760-0/10/$26.00 ©2010 IEEE

[8] Chang-Shing Lee, Senior Member, IEEE, and Mei-Hui Wang "A Fuzzy Expert System for Diabetes Decision Support Application" IEEE Transactions on Systems, Man, and Cybernetics- Part B: Cybernetics, Vol. 41, No. 1, February 2011 Digital Object Identifier 10.1109/TSMCB.2010.2048899.

[9] Siva Sundhara Raja Dhanushkodi, and Vasuki Manivannan "Diagnosis System for Diabetic Retinopathy to Prevent Vision Loss" Applied Medical Informatics *Soft Computing Approaches for Diabetes Disease Diagnosis: A Survey* 11725 Original Research Vol. 33, No. 3 /2013, pp: 1-11Licensee SRIMA, Cluj-Napoca, Romania.

[10] Polat, K., Gunes, S. An expert system approach based on principal component analysis and adaptive neuro-fuzzy inference system to diagnosis of diabetes disease. Digital Signal Processing, 17(4), 702-710, 2007

[11] Karthikeyini.V., Pervin begum.I., "Comparison a performance of data mining algorithms (CPDMA) in prediction of Diabetes Disease", *International journal of Computer Science and Engineer-ing*, Vol.5, No. 03, March 2013, pp. 205-210.

[12] Karthikeyini.V., Pervin begum.I., Tajuddin.K., Shahina Begum, "Comparative of data mining classification algorithm (CDMCA) in Diabetes Disease Prediction", *International journal of Computer Applications*, Vol.60, No. 12, Dec. 2012, pp. 26-31.

[13] D.S. Kumar, G. Sathyadevi, S.Sivanesh Decision, "Support Sys-tem for Medical Diagnosis Using Data Mining ", *International journal of computer applications*, Vol. 4, No. 5, 2011.

[14] Oliver Faust & Rajendra Acharya U. E. Y. K. Ng Kwan-Hoong Ng Jasjit S. Suri "Algorithms for the Automated Detection of Diabetic Retinopathy Using Digital Fundus Images: A Review" J Med Syst (2012) 36:145–157 DOI 10.1007/s10916-010-9454-7

[15] Fayssal Beloufa, Chikh MA. Algeria: s.n. Automatic fuzzy rules-base generation using modified particle swarm optimization. In: 2nd International symposium on modeling and implementation of complex systems constantine 2012. p. 1–6.

[16] Han Jiawei, Kamber Micheline, Pei Jian. "Data Mining Concepts and Techniques." Waltham: Morgan Kaufmann; 2012.

[17] IDF Diabetes Atlas. The Economic Impacts of Diabetes. Available from http://www.diabetesatlas.org/content/economic-impacts-diabetes. Accessed April 1, 2011.

[18] Xue-Hui Meng, Yi-Xiang Huang, Dong-Ping Rao, Qiu Zhang, Qing Liu "Comparison of three data mining models for predicting diabetes or prediabetes by risk factors." SciVerse ScienceDirect Elsevier 2013.

[19] Md. Mozaharul Mottalib, Md. Mokhlesur Rahman, Md. Tarek Habib and 4Farruk Ahmed "Detection of the Onset of Diabetes Mellitus by Bayesian Classifier Based Medical Expert System" Transaction on Machine Learning and Artificial Intelligence DOI: 10.14738/tmlai.44.1962 Publication Date: 19th July, 2016.

[20] T. Karthikeyan, K. Vembandasamy, RaghavanAn Intelligent Type-II Diabetes Mellitus Diagnosis Approach using Improved FP-growth with Hybrid Classifier Based Arm Research Journal of Applied Sciences, Engineering and Technology 11(5): 549-558, 2015. DOI: 10.19026 /rjaset.11 .1860ISSN: 2040-7459; e-ISSN: 2040-7467.

AUTHOR'S PROFILE

Ratna Nitin Patil, has obtained her Masters in Computer Engineering from Thapar Institute, Patiala, Punjab in 2001. She is pursuing PhD from Babasaheb Ambedkar Marathwada University, Aurangabad, India. She has worked as an Associate Professor at Kanpur Institute of Technology, Kanpur, St Peters College of Engineering, Chennai, Sriram College of Engineering, Chennai, Mar Baselios College of Engineering, Trivandrum. Currently she is working at VIT, Pune. She has 21 years of teaching experience. Her research interest includes Machine Learning, Natural Language Processing and Analysis of Algorithms. Email: ratna.patil@vit.edu

Dr. Sharavari Chandrashekhar Tamane, has obtained her PhD in Computer Science and Engineering from Babasaheb Ambedkar Marathwada University, Aurangabad, India. She has 15 international publications. She is working as an associate professor in Jawaharlal Nehru Engineering College, Aurangabad. She is designated as a chairperson of Second International Conference on Internet of Things, Data and Cloud Computing (ICC'17), Cambridge university, United Kingdom Her research area includes Cloud Computing, Analysis of Algorithms, Network Security. Email: sharvaree73@yahoo.com