Study on Prediction and Analysis of Diabetic Data using various Machine Learning Techniques

M. K. Prakash[a,*] and Dr. A. V. Ramani[b]

[a]*Assistant Professor, Department of BCA, Sri Ramakrishna Mission Vidyalaya College of Arts and Science, Coimbatore, Tamil Nadu*

[b]*Head and Associate Professor, Department of Computer Science, Sri Ramakrishna Mission Vidyalaya College of Arts and Science, Coimbatore, Tamil Nadu*

Abstract. The Diabetes is a disease which caused 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. In the present living scenario Diabetes is considered as one of the deadliest and chronic diseases. It is a major public health challenge, in worldwide. If Diabetes is unidentified and untreated then it may cause many other complications. The constant hyperglycemia of diabetes is related to long-haul harm, brokenness, and failure of various organs, particularly the Eyes, Kidneys, Nerves, Heart, and Veins. The Earlier detection of diabetes may control its seriousness and cause to other diseases can be considerably avoided. Now Machine Learning, Artificial Intelligence and statistical methods are used in medical analysis to enhance and speed the process of diagnosing diseases. The aim of the present study is to conduct a systematic review of the applications of Machine Learning, data mining techniques and tools in the field of diabetes research. This work highlights the issues involved in detection of diabetes using Machine Learning algorithms.

Keywords: Diabetes disease, Prediction, Data mining, Statistics, Classification, Machine Learning algorithm, Artificial Intelligence

1. Introduction

Diabetes Mellitus (DM) is a collection of metabolic infections in which a human being has elevated blood sugar, either for the reason that the pancreas does not generate sufficient Insulin, or because cells don’t react to the insulin that is generated. This elevated Blood sugar makes the conventional signs of polyuria (regular urination), ploydipsia (increased need for liquids) and polyphagia (increased starvation).

Diabetes Mellitus (DM) which generally referred as diabetes is a kind of chronic illness produces a group of disorders characterized with High blood sugar levels over a prolonged period [1]. The symptoms of frequent urination, increased hunger and thirst, if it is not untreated then diabetes may cause many complications [2]. Diabetes leads to disfunction of various tissues specifically Eyes, Heart, Kidney, Blood vessels and nerves [3]. Diabetes disease is classified into two types, patients with type 1 diabetes are normally younger and mostly less than 30 years old.

The symptoms of type 1 Diabetes are increased Thirst, High blood glucose levels and frequent urination [4] and this type of diabetes cannot be cured effectively just by taking oral medicines they should also needs to use insulin therapy. Middle aged and elderly people are more commonly suffer from Type 2 Diabetes which is related with the occurrence of Hypertension, Obesity, Arteriosclerosis, Dyslipidemia and other diseases [5].

According to the report of WHO [6] around worldwide, the number of adults with type 2 Diabetes is expected to rise by more than a fifth from 406 million in 2018 to 511 million in 2030. As a result, it has significantly increased mortality in patients.

1.1. *Causes and Effects of diabetes*

Diabetes is influenced by different parts of the body which incorporates some of the following effects:

[*] Corresponding author. E-mail: editorial@iospress.nl. Check if the checkbox in menu *Tools/Options/Compatibility/Lay out footnotes like Word 6.x/95/97* is selected if you make a footnote for the corresponding author.

1.1.1 Loss of Vision:

Retinopathy retina is a condition where the retina, optic nerve, the focal point is harmed. A result of finish night visual impairment issues, swelling in the region of the retina, lessening the contact the mind may happen. A Diabetic individual should deal with eye vision through a few tests and pharmaceutical at the beginning times [11]. The treatment incorporates visual sharpness testing, tonometry, student enlargement, and Optical Coherence Tomography (OCT). Different medicines incorporate Anti-VEGF infusion therapy, focal/lattice macular laser medical procedure, corticosteroid (It is used to provide relief for inflamed areas of the body).

1.1.2 Kidney neuropathy:

Chronic kidney infection or Diabetic Neuropathy [12] is where the high sugar level in blood harms the vessels in the kidney. The usefulness of the kidney is to channel the waste and abundant water in the blood. Because of hypertension and sugar level in Kidney endeavors to have overhead to clean the blood this may prompt kidney disappointment or successive dialysis of blood is required. The treatment may incorporate kidney substitution treatment, kidney and pancreas transplant.

1.1.3 Liver Problems

Liver assumes an indispensable job in adjusting the blood glucose level in blood through starch digestion by methods neoglucogenesis and glycogenosis's [13]. Sort 2 diabetes expands the danger of liver issues. Fatty liver assumes the stipulate job in creating a liver tumour. The difficulties incorporate Renal debilitation, modified metabolism, Insulin opposition and Hyperglycaemia, malnutrition. Affect individual needs to experience different anti-toxin drugs [14] and administration of liver incorporates other treatment [15] like the way of life alteration, Pharmacological treatment, Insulin secretagogues, Biguanides, α-glucosidase inhibitors, TZDs, weight to decrease.

1.1.4 Heart Problems

According to American heart affiliation, 68% of individuals will experience the ill effects of heart issues to driving even to Death, Heart stroke, Atherosclerosis or solidifying of the supply routes, stress and load on the heart make individual to death. Because of high sugar level, blood conveys greater thickness, it adheres to the veins, supply routes and veins put more strain to proceed onward. Persistently it harms the vessels and nerves prompting disappointment of circulatory framework or organ disappointment in person. Hazard for creating cardiovascular illness incorporates Hypertension, unusual cholesterol and high triglycerides, corpulence, the absence of physical activity. The effect of different clinical parameters like poor glycaemic control, insulin opposition of diabetes greatly affects heart issues [16].

*1.2. **Machine Learning***

Machine Learning is the scientific field dealing with the ways in which machines learn from experience. For many scientists, the term "Machine Learning" is identical to the term "artificial intelligence", given that the possibility of learning is the main characteristic of an entity called intelligent in the broadest sense of the word.

The purpose of Machine Learning is the construction of computer systems that can adapt and learn from their experience. A more detailed and formal definition of Machine Learning is given by Mitchel: A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E, with the rise of Machine Learning approaches we have the ability to find a solution to this issue, we have developed a system using data mining which has the ability to predict whether the patient has diabetes or not, Furthermore, predicting the disease early leads to treating the patients before it becomes critical.

Data mining has the ability to extract hidden knowledge from a huge amount of diabetes-related data. Because of that, it has a significant role in diabetes research, now more than ever. The aim of this research is to develop a system which can predict the diabetic risk level of a patient with a higher accuracy. This research has focused on developing a system based on some classification methods, Data Mining Techniques and Artificial Neural Network algorithms

2. Literature Survey

In 2017, National Diabetes Statistic Report [7] for Center Disease Control and Prevention (CDC), gives the facts give an account of the United States that 30.3 million individuals have diabetes, among that 23.1 are analyzed and 7.2 million are undiscovered individuals [8]. In 2018, the American Diabetes Association models of therapeutic care [9] in diabetes discharges a report about "Order and finding of diabetes" which incorporates the arrangement of diabetes, diabetes care, treatment objectives, criteria for conclusion test ranges and dangers esteems, chance engaged with diabetes.

In 2017, Global provides details regarding Diabetes by world wellbeing association [10], it expresses the weight of diabetes, hazard components and inconveniences of diabetes. Likewise, gives the data about counteracting diabetes in individuals with high hazard and overseeing diabetes at beginning times with fundamental solutions to be taken.

3. Comparative Study on Predicting Diabetes Mellitus with Machine Learning Approaches

Perveen et al., [19] explained the act of ensemble Machine Learning approaches namely Adaboost and Bagging to improve the J48 decision tree for classifying diabetes Mellitus and patients as diabetic or non-diabetic, based on diabetes risk factors. Results achieved after the experiment proves that, Adaboost Machine Learning ensemble technique outperforms well comparatively bagging as well as a J48 decision tree.

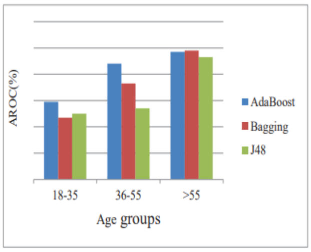

Fig. 1. Comparison of ensembles and J48 decision tree across three different age groups in CPCSSN dataset

Orabi et al., [20] designed a system for diabetes prediction, whose main aim is the prediction of diabetes a candidate is suffering at a particular age. The system is designed based on the concept of Machine Learning, by applying decision tree. Obtained results were satisfactory as the designed system works well in predicting the diabetes incidents at a particular age, with higher accuracy using Decision tree.

Pradhan et al., [21] adapted Genetic programming (GP) for the training and testing of the database for prediction of diabetes by employing Diabetes data set which is sourced from UCI repository. Results achieved using Genetic Programming gives optimal accuracy as compared to other implemented techniques. It shows significant improvement in accuracy by taking less time for classifier generation. It proves to be useful for diabetes prediction at low cost.

Rashid et al. [22] designed a prediction model with two sub-modules to predict diabetes-chronic disease. ANN (Artificial Neural Network) is used in the first module and FBS (Fasting Blood Sugar) is used in the second module.

Nai Arun et al [23] applied an algorithm which classifies the risk of diabetes mellitus. To fulfill the objective author has employed four following renowned Machine Learning classification methods namely Decision Tree, Artificial Neural Networks, Logistic Regression and Naive Bayes. For improving the robustness of designed model Bagging and Boosting techniques are used. Experimentation results shows the Random Forest algorithm gives best results among all the algorithms employed.

Calisir and Dogantekin [24] projected LDA–MWSVM, a system for diabetes diagnosis. The system performs feature extraction and reduction using the Linear Discriminant Analysis (LDA) method, followed by classification using the Morlet Wavelet Support Vector Machine (MWSVM) classifier.

Gangji and Abadeh [25] developed an Ant Colony-based classification system to extract a set of fuzzy rules, named FCS-ANTMINER, for diabetes diagnosis, deals with glucose prediction as a multivariate regression problem utilizing Support Vector Regression (SVR).

Agarwal et al. [26] utilized semi-automatically labeled training sets to create phenotype models via Machine Learning methods.

El-Sappagh et al [27] developed a fuzzy ontology-based Case-based Reasoning (CBR) framework, mimicking expert thinking, further tested on diabetes diagnosis problems. Fong e al [28], authors performed an evaluation of Stream Mining Classifiers for Real-time Clinical Decision Support Systems

Ioannis Kavakioetis et al., [29] presented Research work on study of review in Machine Learning, data mining techniques and tools in the field of diabetes research with respect Prediction and Diagnosis, Diabetic Complications, Genetic Background and Environment, and Health Care and Management with the first category appearing to be the most popular. A wide range of Machine Learning algorithms were employed. In general, 85% of those used were characterized by supervised learning approaches and 15% by unsupervised ones, and more specifically, association rules. Support vector machines (SVM) arise as the most successful and widely used algorithm. Concerning the type of data, clinical datasets were mainly used.

Quan Zou1 et al [30] examined a five-fold cross validation to verify the universal applicability of the approaches, to select few methods that have the best performance for independent test experiments. Arbitrarily nominated 68994 healthy people and diabetic patients' data, respectively as training set. Due to data unbalance, they randomly extracted five times data. And the result is the average of these five experiments. In this study, they used principal component analysis (PCA) and minimum redundancy maximum relevance (mRMR) to reduce the dimensionality. The results showed that prediction with random forest could reach the highest accuracy when all the attributes were used.

Deepti Sisodia and Dilip singh [31] they designed a model which can prognosticate the likelihood of diabetes in patients with maximum accuracy. Therefore, three Machine Learning classification algorithms namely Decision Tree, SVM and Naive Bayes are used in this experiment to detect diabetes at an early stage. Results obtained show Naive Bayes outperforms with the highest accuracy comparatively other algorithms. These results are verified using Receiver Operating Characteristic (ROC) curves in a proper and systematic manner.

4. Methodologies of Data Mining Approaches in Prediction of Diabetes Disease

There is a close relationship between Machine Learning and data mining, with the latter being more generic. Thus, often in scientific literature, Machine Learning methods are called data mining methods.

One of the emerging research areas is computation health informatics which comprised of different kinds of sciences like nursing, medical, biomedical, statistics and information technology [17]. To infer the hidden knowledge about patterns and existing of relationship among data warehouses, the technique which integrates statistical analysis, database technology and Machine Learning is known as data mining.

To control the diabetes, diagnosing it quickly and accurately is very important. Many researchers are conducting experiments for diagnosing the diseases using various classification algorithms of Machine Learning approach. Machine Learning can help people make a preliminary judgment about diabetes mellitus according to their daily physical examination data, and it can serve as a reference for doctors [18]. For Machine Learning method, how to select the valid features and the correct classifier are the most important problems. Recently, numerous algorithms are used to predict diabetes, including the traditional Machine Learning method.

The remaining of this paper reports the survey on many existing Machine Learning approaches involved in prediction of diabetes.

5. Problem Statement

Even though there are many research works are in existence for prediction of the diabetes disease they also have several issues. The ultimate objective of this work is to investigate the existing issues and develop a novel approach which handles these problems more precisely.

5.1. The diabetes dataset under investigation is voluminous which consist of more features which may be either relevant or irrelevant for the prediction process. There is a need to decrease the dimensionality of the dataset by determining optimal feature subset by introducing feature selection approach.

5.2. After discovering the diabetes, the other risk factors which leads to heart disease, kidney failure, liver failure can be analyzed by developing supervised learning models

5.3. As the real time diabetes data are vague and inconsistent to handle by the standard classification or clustering models so uncertainty handling algorithms has to be proposed to improve the quality of accuracy in prediction of diabetes.

Conclusion

It performs the detailed analysis of existing research work done on the prediction of diabetes using classification model, clustering model, decision making system and fuzzy models. The problems related to the process of diabetes prediction are also highlighted in this work.

This paper discusses about the importance of diabetes prediction in early stages by conducting detailed study about the type of diabetes, effects of diabetes. This study explains the importance of the mining approaches in prediction of diabetes.

Use single line spacing throughout the document. Keep the abstract, running text and long captions justified; the chapter title, author’s name, affiliation, the table text, section headings – aligned left. Indent the first line of each paragraph by 0.37 cm.

References

1. About diabetes, World Health Organization. Archived from the original on 31 March 2014. Retrieved 4 April 2014.
2. Diabetes Fact sheet N°312", WHO. October 2013. Archived from the original on 26 August 2013. Retrieved 25 March 2014.
3. Krasteva, A., Panov, V., Krasteva, A., Kisselova, A., and Krastev, Z. (2011). Oral cavity and systemic diseases—Diabetes Mellitus. Biotechnol. Biotechnol. Equip. 25, 2183–2186.
4. Iancu, I., Mota, M., and Iancu, E. (2008). “Method for the analysing of blood glucose dynamics in diabetes mellitus patients,” in Proceedings of the IEEE International Conference on Automation, Quality and Testing, Robotics, Cluj-Napoca, 2008
5. Robertson, G., Lehmann, E. D., Sandham, W, and Hamilton, D, Blood glucose prediction using artificial neural networks trained with the AIDA diabetes simulator: a proof-of-concept pilot study. J. Electronic. Computing Engineering, 2011.
6. https://www.indiatoday.in/education-today/gk-current-affairs/story/98-million-indians-diabetes-2030-prevention-1394158-2018-11-22
7. Centres for Disease Control and Prevention. National Diabetes Statistics Report. Atlanta: Centers for Disease Control and Prevention, US Department of Health and Human Services; 2017.
8. http://care.diabe tesjo urnal s.org.
9. Classification and Diagnosis of Diabetes: Standards of Medical Care in Diabetes—2018 American Diabetes Association Diabetes Care 2018; 41(Supplement 1): S13–S27.
10. World Health Organization Global Report on Diabetes 2017, http://www.who.int/diabetes/publications/grd-2016/en/.
11. Patil S, Kumaraswamy Y. Intelligent and effective heart attack prediction system using data mining and artificial neural networks. Eur J Sci Res. 2009;31(2009):642–56.
12. Picardi A, D’Avola D, Gentilucci UV, Galati G, Fiori E, Spataro S, et al. Diabetes in chronic liver disease: from old concepts to new evidence. Diabetes Metab Res Rev. 22:274–83, 2006
13. Gangopadhyay KK, Singh P. Consensus statement on dose modifications of antidiabetic agents in patients with hepatic impairment. Indian J EndocrMetab, 21:341–54, 2017
14. Tolman KG, Vivian F, Anthony D, Meng H. Tan, Spectrum of liver disease in type 2 diabetes and management of patients with diabetes and liver disease. Diabetes Care. 2007;30(3):734–43.
15. Scott MG, Ivor JB, Gregory LB, Alan C, Robert HE, Barbara VH, William M, Sidney CS, James RS. Diabetes and cardiovascular disease a statement for healthcare professionals from the American Heart Association. Circulation, 1999;100(10):1134–46.
16. de Mattos Matheus AS, Tannus LR, Cobas RA, Sousa Palma CC, Negrato CA, de Brito Gomes M. Impact of diabetes on cardiovascular disease: an update. Int J Hyperten;65, 2013
17. Emrana Kabir Hashi, Md. Shahid Uz Zaman and Md. Rokibul Hasan, “An Expert Clinical Decision Support System to Predict Disease Using Classification Techniques”, International Conference on Electrical, Computer and Communication Engineering (ECCE), pp. 396-400, February 16-18, 2017.
18. Lee, B. J., and Kim, J. Y. (2016). Identification of type 2 diabetes risk factors using phenotypes consisting of anthropometry and triglycerides based on Machine Learning. IEEE J. Biomed. Health Inform. 20, 39–46.
19. Perveen, S., Shahbaz, M., Guergachi, A., Keshavjee, K., 2016. Performance Analysis of Data Mining Classification Techniques to Predict Diabetes. Procedia Computer Science 82, 115–121.
20. Orabi, K.M., Kamal, Y.M., Rabah, T.M., 2016. Early Predictive System for Diabetes Mellitus Disease, in: Industrial Conference on Data Mining, Springer. Springer. pp. 420–427
21. Pradhan, Bamnote, Tribhuvan, Jadhav, Chabukswar, Dhobale, A Genetic Programming Approach for Detection of Diabetes, International Journal of Computational Engineering Research, 2,91–94, 2012.
22. Tarik A. Rashid, S.M.A., Abdullah, R.M., Abstract, 2016. An Intelligent Approach for Diabetes Classification, Prediction and Description, Advances in Intelligent Systems and Computing 424, 323–335.
23. Nai-Arun, N., Moungmai, R, Comparison of Classifiers for the Risk of Diabetes Prediction. Procedia Computer Science 69, 132–142, 2015
24. Duygu Çalişir, Esin Doğantekin, An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier, Expert Systems with Applications, 38(7):8311-8315, July 2011
25. Ganji MF, Abadeh MS. A fuzzy classification system based on ant colony optimization for diabetes disease diagnosis. Expert Syst Appl 2011;38(12):14650–9.
26. Agarwal V, Podchiyska T, Banda JM, Goel V, Leung TI, Minty EP, et al. Learning statistical models of phenotypes using noisy labeled training data. J Am Med Inform Assoc May 12 2016.

27. El-Sappagh S, Elmogy M, Riad AM. A fuzzy-ontology-oriented case-based reasoning framework for semantic diabetes diagnosis. Artif Intell Med Nov 2015;65(3): 179–208.

28. Fong S, Zhang Y, Fiaidhi J, Mohammed O, Mohammed S. Evaluation of stream mining classifiers for real-time clinical decision support system: a case study of blood glucose prediction in diabetes therapy. Biomed Res Int 2013;

29. Ioannis Kavakiotis, Olga Tsave , Athanasios Salifoglou , Nicos Maglaveras, Ioannis Vlahavas, Ioanna Chouvarda, Machine Learning and Data Mining Methods in Diabetes Research, Computational and Structural Biotechnology Journal 15 (2017) 104–116

30. Quan Zou, Kaiyang Qu, Yamei Luo, Dehui Yin , Ying Ju and Hua Tang, Predicting Diabetes Mellitus With Machine Learning Techniques, Frontiers in Genetics, Volume 9 , Article 515, pp 1-10, 2018

31. Deepti Sisodia, Dilip Singh Sisodia, Prediction of Diabetes using Classification Algorithms, International Conference on Computational Intelligence and Data Science (ICCIDS 2018)