Roopa Shrivastava, 2022, 10:2
ISSN (Online): 2348-4098
ISSN (Print): 2395-4752

A Survey on Various Diabetes Disease Prediction Techniques

Roopa Shrivastava, Prof. Rajesh Ku. Nigam, Prof. Rakesh Ku. Tiwari
Dept. of Computer Science and Engineering
Technocrats Institute of Technology & Science
Bhopal, MP,India

Abstract- Medical data diagnosis for detection of various diseases is depends on machine learning models developed by researchers. Out of different health issues judgment of diabetes out of different symptoms. This paper has given detailed introduction of diabetes related condition around the globe with its impact and types of diabetic disease. Various works done by the researcher for prediction of patient diabetic condition were discussed in the paper with implementing techniques. As per the behavior of users different feature values were store in the training dataset, so extraction of patterns form the data was done by data mining approach and learning models were list in the paper. As per the prediction developed models were compared on evaluation parameters, hence paper detail formulas of such models.

Keywords- diabetic disease, data mining etc.

I. INTRODUCTION

The disease or condition which is continual or whose effects are permanent is a chronic condition. These types of diseases affected quality of life, which is major adverse effect. Diabetes is one of the most acute diseases, and is present worldwide.

A major reason of deaths in adults across the globe includes this chronic condition. Chronic conditions are also cost associated. A major portion of budget is spent on chronic diseases by governments and individuals [1,2]. The worldwide statistics for diabetes in the year 2013 revealed around 382 million individuals had this ailment around the world [3]. It was the fifth leading cause of death in women and eight leading cause of death for both sexes in 2012.

Higher income countries have a high probability of diabetes [4]. In 2017, approximately 451 million adults were treated with diabetes worldwide. It isprojected that in 2045, almost 693 million patients with diabetes will exist around the globe and half of the population will be undiagnosed. In addition, 850 million USD were spent on patients with diabetes in 2017 [5]. Research on biological data is limited but with the passage of time enables computational and statistical models to be used for analysis.

A sufficient amount of data is also being gathered by healthcare organizations. New knowledge is gathered when models are developed to learn from the observed data using data mining techniques. Data mining is the process of extracting from data and can be utilized to create a decision making process with efficiency in the medical domain [6]. Several data mining techniques have been utilizedfor disease prediction as well as for knowledge discovery from biomedical data [7,8].

Introduction 21.0 million is the estimated number of people that have been diagnosed with diabetes, and another 8.1 million people with diabetes have not been diagnosed in the United States in 2012 [3] [1] 1.1. This numbers correspond to 9.3% and 27.8% of total United States population respectively. People who have been diagnosed with diabetes represents 25% of the united stated population over 65years old. It is well known that daily diabetes care is mainly handled by patients and/or their families.

The recommendation is to conduct frequent blood glucose (BG) for people with diabetes by their health care professionals in order to achieve a specific level of glycemic control and to reduce the risk of hypoglycemia. The objective of BG monitoring is to collect detailed information about blood glucose

levels at many time points in order to maintain of a more propriate glucose level by more precise regimens.

Facilitating the development of an individualized blood glucose profile is being aided by self monitoring of BG levels. This individual profile can then used as a guide by the healthcare professionals in treatment planning for an individualized diabetic regimen. The individualized diabetic regimen is able to give people with diabetes and their families the ability to make appropriate day-to-day treatment choices in diet and physical activity as well as ininsulin or other agents.

The ability of the patients‟ recognition of hypoglycemia or severe hyperglycemia can be improved with a personalized diabetic regimen. Further, this will enhance patient education and patient empowerment regarding the effects of lifestyle and pharmaceutical intervention on glycemic control. Individuals can adjust their dietary intake, physical activity, and insulin doses to improve glycemic control on a day-to-day basis by using the adjusted therapeutic regimen.

II. TYPES OF DIABETES

There are three types of diabetes [8], namely:

1. Type 1:

Diabetes mellitus (T1DM) or sometimes called juvenile diabetes is the type of diabetes that results from stopping the insulin generation by the pancreasbeta cells. It is the most severe type of diabetesamong all of the other ones. It is prevalent among children and requires several insulin injections each day to bring the patient‟s glucose levels under control.

2. Type 2:

Diabetes mellitus (T2DM), or the so-called adult-onset diabetes, is the most common type of diabetes, where it compromises 90% of diabetic population worldwide. People can develop T2DM at any age. This form of diabetes usually starts with insulin resistance, which eventually leads to the loss of the pancreas ability to produce enough insulin in response to food intake.

3. Type 3:

Gestational diabetes [9] is the type of diabetes affecting some women during pregnancy only.

III. RELATED WORK

Uloko et al., 2018 [12], conducted a research on the prevalence of risk factors for diabetes mellitus in Nigeria. In conducting this research work, a total of 23 studies (n = 14,650 persons) were considered. In estimating the pooled prevalence of DM, a random effects model was implemented and subgroup specific DM prevalence was used to account for inter-study and intra-study heterogeneity. From the results achieved they concluded that, the prevalence of DM in Nigeria has been on the increase in all regions of the country affected, with south-south with the having the highest prevalence seen in the geopolitical zones. Urbanization, physical inactivity, aging, and unhealthy diet are key risk factors for DM among Nigerians. They recommended the urgent need for a national diabetes care and preventionpolicy scheme.

Chawan, 2018 [13] conducted a research aimed at developing a system which can predict diabetes at an early stage in patients with a high accuracy by combining the results of different machine learning techniques. The research predicts diabetes using two (2) different supervised machine learning methods including SVM and Logistic Regression. It considered seven (7) features of the patients. They reached a conclusion that SVM showed a better performance with accuracy of seventy-nine percent (79%) compared to logistic regression which had a performance accuracy of seventy-eight percent (78%).

Sneha & Gangil, 2019 [14] conducted a research that was aimed at selecting the attributes that aid in early detection of diabetes mellitus using WEKA which is a predictive analysis tool. They were able to reach a conclusion which shows that decision tree algorithm and Random Forest Algorithm has the highest predictive analysis by 98.20% and 98.00%respectively. While Naïve Bayesian outcomes states the best in performance accuracy with 82.30%.

Modern, 2019 [15] stated that, there are an enormous amount of data available in the world today, but very few are there for the analysis of it because of which nowadays many new fields are emerging starting from Data Science to Bioinformatics and Cheminformatics. It can be assured that this world of AI is going to benefit a lot to humanity, converting the toughest jobs to the

simplest ones. Machine learning has led to minimizing the errors Involved with the co-relation of different kinds of attributes. Most importantly, it has transformed the approach of hit and trial method into a way with full of logic and simulations. Today, using various simulations several required properties and the after effects of many materials can be predicted, which has led us to the maximization of a lot of resources. In this review article, they presented the machine learning types, different algorithms and along with their uses in several in different ways.

Kaur & Kumari, 2019 [16] developed five different models for the detection of diabetes using, linear kernel support vector machine (SVM-linear), radial basis kernel support vector machine (SVM-RBF), K Nearest Neighbour (k-NN), Artificial Neural Networks (ANN) and Multifactor Dimensionality Reduction (MDR) algorithms. Feature selection of dataset was done with the help of Boruta wrapper algorithm, considering some evaluation criteria namely; accuracy, recall, precision, F1 score, and Area under the Curve (AUC). The experimental results indicated that all the models achieved good results with SVM- linear model providing a very good accuracy of 0.89 and precision of 0.88. From the results of this study, it can be concluded that on the basis of the entire parameters linear kernel support vector machine (SVM-linear) and k-NN are the two (2) most accurate predictive models for diabetes.

In [17], the authors have analyzed comparatively the results of classification and clustering of spam management of SMS by using two different environments, named as Rapid Miner and WEKA. They performed an experiment by using the same dataset in both environments and the simulations gave the same results and SVM was considered the best classifier in both environments. Machine Learning offers classification and regression methods that can be used to address diagnosis problems in various medical fields. A comparative study of breast cancer prediction has used Artificial Neural Networks(ANN), k-Nearest Neighbors (KNN) and Bayesian Network Classifiers for comparison and based on the results of the analysis, Artificial Neural Networks are better than KNN and Bayesian Classifiers in classification with 97.4% accuracy.

In this study [18], the writers introduced RALE lung sound library support vector machine (SVM) and KNN machine learning algorithms for the analysis of respiratory diseases using pulmonary acoustic signals. And demonstrated that the KNN classifier's generalization capacity is higher compared to SVM's. KNN's precision was calculated to be 98.26% relative to SVM's 92.19%.

The authors in paper [19] proved that, compared to other algorithms such as NB, C4.5 and Repeated Incremental Pruning to Produce Error Production (RIPPER), Support Vector Machine (SVM) gives 93%, the highest accurate value in their developed spam filtering framework.

In [20] author, presents a model using a fused machine learning approach for diabetes prediction. The conceptual framework consists of two types of models: Support Vector Machine (SVM) and Artificial Neural Network (ANN) models. These models analyze the dataset to determine whether a diabetes diagnosis is positive or negative. The dataset used in this research is divided into training data and testing data with a ratio of 70:30 respectively. The output of these models becomes the input membership function for the fuzzy model, whereas the fuzzy logic finally determines whether a diabetes diagnosis is positive or negative.

IV. MACHINE LEARNING TECHNIQUES

1. Misuse identification: [9]
Utilizes examples of the definitely known attacks or the framework's delicate spots to coordinate and distinguish intrusions. For example, in the event that somebody tries to figure a secret key, a mark manage for this sort of conduct could be that 'excessively numerous fizzled login attempted in indicated time' and occasion of his compose may bring about rising an alarm. Misuse recognition observed to be not proficient against the not known attacks that have no coordinated guidelines or examples yet.

2. Anomaly location:
A logging session that is being observed on the off chance that it has altogether lower or higher frequencies an abnormality ready will be raised. Anomaly identification is a successful strategy for discovering perfect or not referred to attacks as the learning is never required with respect to the intrusion attacks. Be that as it may, in the meantimeit tends to raise a larger number of cautions than misuse identification since whatever occasion occurs

in a session, ordinary or unusual conduct, if their frequencies are significantly separate among the threshold found the middle value of frequencies of the client it will raise an alarm.

3. Supervised learning:
In conceptual terms the supervised learning can be seen as a teacher having knowledge of the environment derived from input-output examples. The teacher provide consultancy to the neural network telling it what is normal and abnormal traffic pattern, in the sense of what is classified as malicious and non-malicious.

Basically the supervised learning operates as a portion of network connection is to be analyzed and labeled with the help of the teacher [2, 11]. Afterwards the labeled training data is used by the learning algorithm to generalize the rules. Finally the classifier uses the generated rules to classify newnetwork connections and gives alert if a connection is classified to be malicious.

4. Unsupervised learning:
Unlike the supervised learning, unsupervised learning does not have a teacher to tell what is a "good" or "bad" connection. It has the ability to learn from unlabeled data and create new classes automatically. In with the use of a clustering algorithm it is illustrated how unsupervised learning operates [8]. First, the training data is clustered using the clustering algorithm. Second, the clustered weight vectors can be labeled by a given labeling process,for example by selecting a sample group of the data from a cluster and label that cluster center with the major type of the sample. Finally, the labeled weight vectors can be used to classify the network connections.

5. Genetic Algorithm:
Genetic algorithms are unsupervised search procedures often used for optimization problems. Genetic algorithm is based on the principles of evolution and natural selection of chromosomes. An initial population of chromosomes is generated randomly where each chromosome represents a possible solution to the problem (a set of parameters).

The evaluation function is used to calculate the "goodness" of each chromosome. In evaluation, two operators, crossover and mutation, are used to generate the new population or rules. Then, the best individual or chromosome is selected as the finalresult once the optimization criteria in met

6. Decision tree:
It is another approach to perform characterization. Decision tree [21] is a classifier that is displayed as various hierarchical decay of data space. The tree structure contains two sorts of nodes: leaf node (contains the estimation of the objective quality, for example normal or malicious in twofold order assignment) and choice node (contains a condition on one of the properties for space division).

The division of the information space is done recursively in hierarchical structure of decision tree.

V. PREDICTION TECHNIQUES COMPARING PARAMETERS

As various techniques evolved different steps of working for segmenting data into appropriate category. So it is highly required that proposed techniques or existing work need to be compare on same dataset. But prediction class which is obtained as output is needed to be evaluating on the function or formula.

So following are some of the evaluation formula which help to judge the clustering techniquesranking.

$$\text{Precision} = \frac{True\_Positive}{True\_Positive + False\_Positive}$$

$$\text{Recall} = \frac{\text{True\_Positive}}{\text{True\_Positive} + \text{False\_Negative}}$$

$$F\_Score = \frac{2 * \text{Precision} * \text{Recall}}{\text{Precision} + \text{Recall}}$$

Accuracy = (True_Positive + True_Negative)/(True_Positive + True_Negative+ False_Positive + False_Negative)

In above true positive value is obtained by the system when the classified data is same as in actual case or

ground truth class. While in case of false

positive value it is obtain by the system when the classified data is not of same case as in actual in or ground truth class.

VI. CONCLUSIONS

Large amount of health related data in the digital state attracts people to get some pattern forprediction of disease. Researchers are working in field of diabetes prediction from last few decadesand proposed many model, algorithms to predictpatient diabetic status.

This paper has summarized various techniques of machine learning that can improve the prediction accuracy. Paper has detailed the scholars work done in this field and it was obtained that feature pre-processing technique should be optimized to increase the learning of mathematical model. In future scholars can propose a model that can detect predict with less feature set values.

REFERENCES

[1] D. Falvo, B.E. Holland Medical and psychosocial aspects of chronic illness and disability Jones & Bartlett Learning (2017).

[2] J.S. Skyler, G.L. Bakris, E. Bonifacio, T. Darsow, R.H. Eckel, L. Groop, et al. Differentiation of diabetes by pathophysiology, natural history, and prognosis Diabetes, 66 (2017), pp. 241-255.

[3] Z. Tao, A. Shi, J. Zhao Epidemiological perspectives of diabetes Cell Biochem Biophys, 73 (2015), pp. 181-185.

[4] W.H. Organization World health statistics 2016: monitoring health for the SDGs sustainable development goals World Health Organization (2016).

[5] N. Cho, J. Shaw, S. Karuranga, Y. Huang, J. da Rocha Fernandes, A. Ohlrogge, et al. IDF Diabetes Atlas: global estimates of diabetes prevalence for 2017 and projections for 2045.

[6] S. Diwani, S. Mishol, D.S. Kayange, D. Machuve, A. Sam Overview applications of data mining in health care: the case study of Arusha region|| Int J Comput Eng Res, 3 (2013), pp. 73-77.

[7] T.M. Alam, M.J. Awan Domain analysis of information Extraction Techniques Int JMultidiscip Sci Eng, 9 (2018), pp. 1-9

[8] WHO. "Diabetes Programme", Internet: http:// www.who.int/diabetes/en/> [Sep. 27, 2014].

[9] American Diabetes Association. "Gestational Diabetes." Internet: http://www. Diabet es.org /diabetes-basics/gestational/ [Apr. 19, 2014].

[10] Masaki Yamaguchi, S. Kanbe, Karin Wårdell, Katsuya Yamazaki, Masashi Kobayashi, Nobuaki Honda, Hiroaki Tsutsui, and Chosei Kaseda. "Trend estimation of blood glucose level fluctuations based on data mining," in The 7th world multiconference on systemics, cybernetics and informatics, 2003, pp. 86-91.

[11] Meri Raffetto ,"Glycemic Index Diet." Internet: http://www.dummies.com/howto/content/how-to-measure-your-metabolic-rate.html [Apr. 19, 2014].

[12] Chawan, P. M. (2018). Logistic Regression and Svm Based Diabetes. International Journal For Technological Research In Engineering, 5(6), 4347–4350.

[13] Uloko, A. E., Musa, B. M., Ramalan, M. A., Gezawa, I. D., Puepet, F. H., Uloko, A. T., Borodo, M. M., & Sada, K. B. (2018). Prevalence and Risk Factors for Diabetes Mellitus in Nigeria: A Systematic Review and Meta-Analysis. Diabetes Therapy, 9(3), 1307–1316.

[14] Sneha, N., & Gangil, T. (2019). Analysis of diabetes mellitus for early prediction usingoptimal features selection. Journal of Big Data, 6(1), 1–19.

[15] Modern, S. (2019). A critical review on machine learning algorithms and their applications in pure sciences. Research Journal of Recent Sciences, 8(1), 14–29.

[16] Kaur, H., & Kumari, V. (2019). Predictive modelling and analytics for diabetes using machine learning approach. Applied Computing and Informatics, xxxx, 1–6.

[17] 8 Zainal, K., Sulaiman, N., and Jali, M.: „An analysis of various algorithms for text spam classification and clustering using RapidMiner and Weka", International Journal of Computer Science and Information Security, 2015, 13, (3), pp. 66

[18] 10 Palaniappan, R., Sundaraj, K., and Sundaraj, S.: „A comparative study of the svm and k-nn machine learning algorithms for the diagnosis of respiratory pathologies using pulmonary acoustic signals", BMC bioinformatics, 2014, 15, (1), pp. 223.

[19] 11 Rafique, M.Z., Alrayes, N., and Khan, M.K.: „Application of evolutionary algorithms in detecting SMS spam at access layer", in Editor (Ed.)^(Eds.): „Book Application of evolutionary

algorithms in detecting SMS spam at access layer" (2011, edn.), pp. 1787-1794.

[20] U. Ahmed *et al.*, "Prediction of Diabetes Empowered with Fused Machine Learning," in IEEE Access, doi: 10.1109/ACCE SS.2022.31 42097.