Prediction of Type 2 Diabetes Using Machine Learning Algorithms

Dr. P.Pandi selvi MCA.,M.Phil.,Ph.D.,
Assistant Professor, Department of Computer Science,
Mangayarkarasi College of Arts and Science for Women, Paravai, Madurai

P.Abirami
PG Student
Mangayarkarasi College of Arts and Science for Women, Paravai, Madurai

Abstract

Diabetes is one of the top 10 diseases in the world, which causes death globally. Predicting type 2 diabetes is important for providing prognosis or diagnosis support to allied health professionals, and aiding in the development of an efficient and effective prevention plan. Several works proposed machine-learning algorithms to predict type 2 diabetes. In this paper, the authors proposed a machine learning (ML) Model to predict T2D occurrence by using two different algorithms such as, SVM and XG Boost. As a first step, patients data set was collected with type 2 diabetes and established the prediction model for future risks of (Diabetic retinopathy) DR based on a machine learning (ML) algorithm. The given input image was first pre-processed and then segmented to extract the infected part. After Segmentation the given input dataset was split into test and train dataset. Classification was then performed with the following algorithms, SVM and XG Boost. The performances of the two algorithms were analysed and their accuracy was measured based on sensitivity and specificity. It is evident from the result that the performance of XG boost model was found to be best when compared to SVM classifier.

Keywords: Machine learning, T2D, Diabetic Retinopathy (DR).

1.Introduction

Diabetes is one of the major health problems in both developed and developing countries [1]. As stated by the National Diabetes Statistics Report 2020, 34.2 million people (or 10.5%) of U.S. population are suffering from diabetes. There are 26.9 million people who are diagnosed with diabetes and 7.3 million people are unaware of this condition (21.4% people who have diabetes are unaware) [2]. In 2019, around 77 million people were diagnosed with diabetes in India, which was ranked as second country with highest number of diabetic people in the world [3]. Diabetes is a chronic disease and causes long-term and short-term complications where short-term complications include dehydration and diabetic coma and long-term complications include heart attack, blindness, kidney failure, stroke and foot ulcers, etc.

Generally, diabetes is classified into three different types which include type 1 diabetes, type 2 diabetes and gestational diabetes, Type 1 diabetes is a condition in which body is incapable of producing insulin for the proper functioning of the body. It is an autoimmune disease in which β-cells of the body are destroyed which result in the lack of insulin, β-cells are liable for the storage and release of the insulin. Type2 diabetes is state in which the body is unable to produce enough insulin or there is insulin but body is not able to use it, this condition is known as insulin resistance. It is the most prevalent type of diabetes, which is detected in 90% of the cases.

(Diabetic retinopathy) DR is the most common microvascular complication of diabetes mellitus. It has been demonstrated to be a leading cause of preventable blindness in the working-age population in most countries (1).The American Academy of Ophthalmology (AAO), in 2019,stated that the prevalence of DR among diabetic patients worldwide is about 34.6%. 10.2% (28 million) diabetic patients suffer from vision-threatening DR (2).

Effective management of DR requires a deep understanding of the predisposing factors, early diagnosis, and timely therapeutic intervention. Early identification of patients at risk of developing DR is the key to effective intervention, which is significant in reducing the progression of DR and thereby reducing the risk of blindness (3). Moreover, individual patients can be stratified according to different risk levels and get optimal treatment. Due to no typical symptoms in the early stage of the disease, however, most patients with DR may not seek medical evaluation until progression to the proliferative stage, resulting in irreversible visual damage (4). Therefore, methods for accurate prediction of the risk of DR are in urgent need.

At present, several DR risk prediction models based on cross-sectional studies have been developed (5–9). Deep learning algorithms were also applied (10). Although these models can predict the occurrence of DR at the index date, they cannot predict DR occurrence and development of the same patient at designated time points in the future. This will obviously restrict their clinical application. Similarly, based on the clinical characteristics related to the occurrence or development of DR, several models have been developed for optimization of the screening interval in DR screening (11,12). However, due to the small number of cases in the studies, the proposed model shave not been fully validated so far.

On the other hand, several studies investigated the pathogenesis and risk factors of DR to provide guidance for DR management. Epidemiological studies have shown that age, course of diabetes, haemoglobin A1c(HbA1c), fasting blood glucose (FBG), blood pressure, blood lipids, body mass index (BMI), smoking, proteinuria, and several others are all risk factors for DR (13,14). Among them, duration of diabetes antihyperglycemic were demonstrated as strong risk factors for the occurrence and development of DR (15,16). However, patients without DR were hardly unusual among those suffering from diabetes for a long time (17). The influences of other factors in DR occurrence also need to be proven. Further studies are required to elucidate the correlation and thus construct standard procedures for the management of this disease.

In this retrospective cohort study, we collected electronic health record data from hospitalized patients with type 2 diabetes and established the prediction model for future risks of DR based on a machine learning (ML) algorithm. To our knowledge, this is the first study to predict the occurrence of DR at each follow-uptime point in up to 10 years. We also explore the risk factors that may affect the occurrence of DR and hope this work can provide a basis for further studies concerning the prevention and management of DR.

2.Literature Review

2.1 Related Work

A Neural Network for DR Diagnosis Patients with long-term diabetic conditions are most likely to develop diabetic retinopathy (DR), leading to blindness. The disease can be prevented or delayed by early detection of biomarkers and effective treatment. To better understand the prevalence and progression of DR, researchers have looked into several biomarkers. The existence of microaneurysms, exudates, haemorrhages, and the like in the patients' retinas all contributed to the disease. A strategy for preventing blindness has been suggested by looking at iris images from time to time. This study used a DR dataset and various machine learning classification algorithms to predict the occurrences of the DR in this study (Valorie et al., 2019). The main focus of this paper is on the use of data mining and fuzzy system techniques in the diagnosis of diabetes. (Thakkar et al., 2021). To detect type II diabetes, data mining algorithms such as Naïve Bayes classifier, RBF System, and J48 are described in that article.

To facilitate early detection of T2D, numerous research studies employing ML techniques have been conducted. These studies include the development of screening, diagnosis, and prediction tools to detect the occurrence of the disease and the likelihood of its onset [5,21]. Screening methods for prediabetes using ML models for the South Korean population are presented in [5], which developed an intelligence-based screening model for prediabetes using a dataset from the Korean National Health and Nutrition Examination Survey (KNHANES) [22]. The KNHANES 2010 dataset, with 4685 instances, was used to train SVM and artificial neural network (ANN) based models, and the KNHANES 2011dataset was used for validation. The authors claimed that the SVM model performed better than the ANN model, with an area under curve (AUC) value of 0.73. The study was limited to identifying a prediabetes condition only.

A model for predicting the onset of type 2 diabetes in non-diabetic patients with cardiovascular disease is presented in [21]. The study reported a T2D prediction model to forecast the occurrence of the disease within the follow-up period. The electronic health records (EHRs) for the study were collected from Korea University Guro Hospital (KUGH). The total number of features was 28, with 8454 subjects over five years of follow-up. The authors claimed that they had achieved a value of 78.0 in AUC measure for the logistic regression (LR) model. In this study, the dataset included only individuals with cardiovascular risks.

A comprehensive study on machine learning techniques for diabetes identification is presented in [23]. The study analysed two essential data processors: PCA (Principal Component Analysis) and LDA (Linear Discriminant Analysis) for various machine learning algorithms. Through an experiment, they identified the best data

Preprocessors for each algorithm and conducted parameter tuning to find the optimum performance. Pima Indian data set was utilized to examine the performance of the algorithms. The highest accuracy obtained among the employed five algorithms (neural Network, Support Vector Machine, Decision tree, Logistic regression, and Naïve Bayes) was 77.86% using 10-fold cross-validation.

Diabetes is a long-term condition that occurs when the body does not make enough insulin or is unable to utilize the insulin properly. Naive Bayes and SVMs, two of the most popular Machine Learning techniques, have been used to classify data in most systems. (Parthiban & K. Srivatsa, 2012) Data mining algorithms and their implementations were examined in this paper. Self-organizing maps outperformed Random Forest as well as other data mining algorithms, such as naive Bayes, decision trees, SVMs, and MLPs, in the evaluation of their ability to accurately diagnose diabetes. (Alawa, 2019).

Based on eye fundus images, this article aims at developing a classification algorithm that can identify diabetic retinal disease in patients. (Abreu et al., 2021). The CNN model for detecting plants and flowers, was discussed in this article (Wasson et al., 2021). Li et al. devise a scoring model to help patients with type 2 diabetes distinguish between diabetic nephropathy (DN) and non-diabetic renal disease (NDRD)(Li et al., 2020). In this study, the authors compare and contrast various machine learning-based classifiers. The COVID-19 pandemic tweet datasets were used in experiments by the author. The author used seven machine learning-based classifiers (Wisteria et al., 2021).

2.2 Type 2 Diabetes (T2D)

Diabetes mellitus is a group of metabolic abnormality identified by hyperglycaemia resulting from defects in insulin secretion, insulin action, or both [1]. According to the American Diabetes Association (ADA) guidelines, T2D is defined by fasting plasma glucose (FPG) levels above 125 mg/dL; the normal (non-diabetic) range is below 100 mg/dL [25].

It is highly affected by lifestyle activities, such as drinking, exercise, and dietary habits.T2D diminishes quality of life and lowers life expectancy. Several studies have shown that a combination of lifestyle improvement and medication intervention can prevent complications from the disease. Both early diagnosis and treatment of T2D are thus critical in preventing serious and potentially life-threatening complications in patients [21]. In this study, T2D was diagnosed according to the ADA guidelines. T2D is defined by FPG levels above 125 mg/dl; the normal range is below 100 mg/dL and between 100 and 125 mg/dL is considered prediabetes.

3.Methodology

The main motive is to develop a prediction model to forecast the occurrence of T2D in patients at an early stage. Various steps involved to generate the model is as follows, data pre-processing, Segmentation, feature selection, training, testing, classification and recognition.

3.1 Dataset

Sample iris data from infected patients were collected and a data set was created.

3.2 Data Pre processing

Once after collecting the dataset, it needs to be preprocessed to remove any unwanted data in the given input image.

3.3 Segmentation

. The infected from the given image was then segmented and then processed to the next stage.

3.4 Feature Selection

The desired features from the segmented part of the image were then extracted.

3.5 Testing and Train dataset

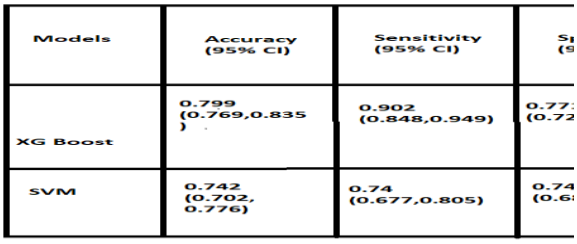
The given dataset was split into test and train dataset, and it was then trained with the given set of extracted features.

3.6 Classification

Once after training, the system was then tested with the test dataset and the given image was then classified based on the occurrence of the disease. Classification was performed with SVM and XG Boost. Finally, the performance of the algorithm was also compared and tabulated.

4.Results

The experimental results of the proposed models were assessed in terms of predefined evaluation metrics and ROC (Table 1 and Figure 1).



Models	Accuracy (95% CI)	Sensitivity (95% CI)	Sp (9
XG Boost	0.799 (0.769,0.835)	0.902 (0.848,0.949)	0.77 (0.72
SVM	0.742 (0.702, 0.776)	0.74 (0.677,0.805)	0.74 (0.6

TABLE 1: The performance metrics of the cross-validated machine learning algorithms on the test data

The ROC curve of the above metrics is as shown in below figure 1.

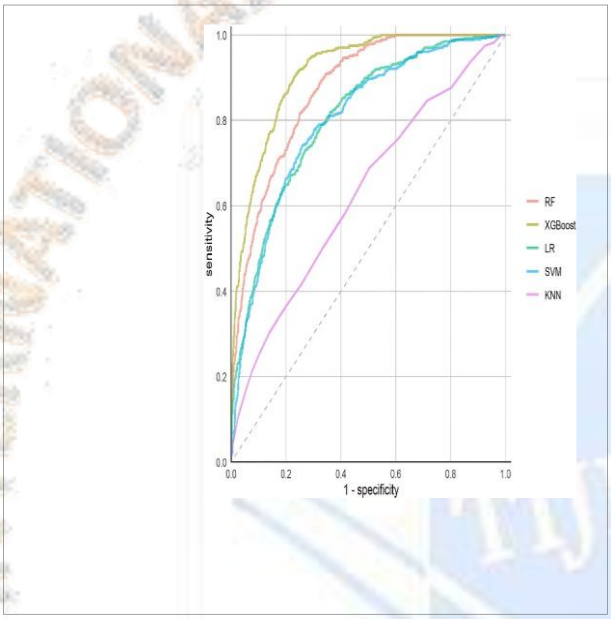


Figure1: Receiver Operating Characteristics (*ROC*) curves of the ML models.

It is evident from the results that the performance of the XG BOOST was found to be very effective when compared to SVM classifier

5.Conclusion

In this paper, the authors developed and evaluated the ML-based model for predicting the risk of type 2 diabetes. As a first step, the given input image was preprocessed to remove any unwanted information. It then undergoes segmentation to extract the infected part of the image. Once after segmentation classification of the presence of disease was carried out with XG BOOST and SVM classifier. With the results, it is evident that the performance of XG BOOST was found to be very high when compared to SVM classifier.

6.References

[1]. Alberti KG, Zimmet PZ. Definition, Diagnosis and Classification of Diabetes Mellitus and its Complications. Part 1: Diagnosis and Classification of Diabetes Mellitus Provisional Report of a WHO Consultation. Diabetes Med(1998)15(7):539–53. Doi:10.1002/(SICI)10969136(199807)15:7<539:AID-DIA668>3.0.CO;2-S

[2]. Breiiman L. Random Forests. Mach Learn (2001) 45(1):5–32. Doi: 10.1023/A:1010933404324

[3]. Bora A, Balasubramanian S, Babenko B, Virmani S, Venugopalan S, Mitani A.Predicting the Risk of Developing Diabetic Retinopathy Using Deep Learning. Lancet Digit Health (2021) 3(1):e10–9. doi: 10.1016/S2589-7500(20)30250-8

[4]. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. KDD '16: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. (2016), 785–94. Doi: 10.1145/2939672.2939785

[5]. Dagliati A, Marini S, Sacchi L, Cogni G, Teliti M, Tibollo V. Machine Learning Methods to Predict Diabetes Complications. J Diabetes SciTechnol(2018)12(2):295–302.doi: 10.1177/19322968177063751

[6]. Flaxel CJ, Adelman RA, Bailey ST, Fawzi A, Lim JI, Vemulakonda GA.Diabetic Retinopathy Preferred Practice Pattern. Ophthalmology (2020) 127(1):66–145. doi: 10.1016/j.ophtha.2019.09.025

[7]. Jampol LM, Glassman AR, Sun J. Evaluation and Care of Patients With Diabetic Retinopathy. N Engl J Med; (2020) 382(17):1629–37. doi: 10.1056/NEJMra1909637

[8]. Kempen JH, O'Colmain BJ, Leske MC, Haffner SM, Klein R, Moss SE. The Prevalence of Diabetic Retinopathy Among Adults in the United States. ArchOphthalmol (2004) 122(4):552–63. doi: 10.1001/archopht.122.4.552

[9]. Keenan HA, Costacou T, Sun JK, Doria A, Cavellerano J, Coney J. ClinicalFactors Associated With Resistance to Microvascular Complications in Diabetic Patients of Extreme Disease Duration: The 50-Year Medallists. Diabetes Care (2007) 30(8):1995–7. Doi: 10.2337/dc06-2222

[10]. Liew G, Michaelides M, Bunce C. A Comparison of the Causes of Blindness Certifications in England and Wales in Working Age Adults (16-64 Years),1999-2000 With 2009-2010. BMJ Open (2014) 4(2):e004015. doi: 10.1136/bmjopen-2013-004015

[11]. Levey AS, Stevens LA, Schmid CH, Zhang Y, Castro AFIII, Feldman HI. ANew Equation to Estimate Glomerular Filtration Rate. Ann Internal Med(2009) 150(9):604–12. Doi: 10.7326/0003-4819-150-9-200905050-00006

[12]. Mo R, Shi R, Hu Y, Hu F. Nomogram-Based Prediction of the Risk of Diabetic Retinopathy: A Retrospective Study. *J Diabetes Res* (2020) 2020:7261047. doi: 10.1155/2020/7261047

[13]. Mehlsen J, Erlandsen M, Poulsen PL, Bek T. Individualized Optimization of the Screening Interval for Diabetic Retinopathy: A New Model. Exophthalmos (2012) 90(2):109–14. Doi: 10.1111/j.1755-3768.2010.01882.x.

[14]. Nathan DM, Genuth S, Lachin J, Cleary P, Crofford O, Davis M. The Effect of Intensive Treatment of Diabetes on the Development and Progression of Long-Term Complications in Insulin-Dependent Diabetes Mellitus. N Engl JMed (1993) 329(14):977–86. Doi: 10.1056/NEJM1993093032914

[15]. Niroomand M, Afsar J, Hosseinpanah F, Afrakhteh M, Farzaneh F, Serahati Comparison of the International Association of Diabetes in Pregnancy Study Group Criteria With the Old American Diabetes Association Criteria for Diagnosis of Gestational Diabetes Mellitus. Int J Endocrinol Meta (2019) 17(4):e88343. Doi: 10.5812/ijem.88343

[16]. Oh E, Yoo TK, Park E-C. Diabetic Retinopathy Risk Prediction for Fundus Examination Using Sparse Learning: A Cross-Sectional Study. BMC Med Indecision making (2013) 13:106. doi: 10.1186/1472-6947-13-106

[17]. Ogunyemi O, Kermah D. Machine Learning Approaches for Detecting Diabetic Retinopathy From Clinical and Public Health Records. AMIAAnnu Symp Proc (2015) 2015:983–90.

[18]. Ogunyemi OI, Gandhi M, Tayek C. Predictive Models for Diabetic Retinopathy From Non-Image Teleretinal Screening Data. AMIA Summits Trans Sci Proc (2019) 2019:472–7.

[19]. Tsao H-Y,Chan P-Y, Su EC-Y. Predicting Diabetic Retinopathy and Identifying Interpretable Biomedical Features Using Machine Learning Algorithms. BMCBioinf (2018) 19(Suppl 9):283. doi: 10.1186/s12859-018-2277-0

[20]. UK Prospective Diabetes Study (UKPDS) Group. Intensive Blood-Glucose Control With Sulphonylureas or Insulin Compared With Conventional Treatment and Risk of Complications in Patients With Type 2 Diabetes (UKPDS 33). Lancet (1998) 352(9131):837–53. Doi: 10.1016/S0140-6736(98)07019-6.

[21]. Varma R, Macias GL, Torres M, Klein R, Peña FY, Azen SP. Biologic RiskFactors Associated With Diabetic Retinopathy: The Los Angeles Latino Eye Study. Ophthalmology (2007) 114(7):1332–40. Doi: 10.1016/j.ophtha.2006.10.023

[22]. Wilkinson CP, Ferris FL3rd, Klein RE, Lee PP, Agardh CD, Davis M.Proposed International Clinical Diabetic Retinopathy and Diabetic MacularEdema Disease Severity Scales. Ophthalmology (2003) 110(9):1677–82. doi:10.1016/S0161-6420(03)00475-5

[23]. Williams B, Mancia G, Spiering W, Agabiti Rosei E, Azizi M, Burnier M. 2018ESC/ESH Guidelines for the Management of Arterial Hypertension'.EurHeart J (2018) 39(33):3021–104. Doi: 10.1093/Earhart/ehy339.