Advances in Informatics, Management and Technology in Healthcare
J. Mantas et al. (Eds.)

doi:10.3233/SHTI220752

Early Diabetes Prediction: A Comparative Study Using Machine Learning Techniques

Tahmina Nasrin POLY[a,c,d], Md Mohaimenul ISLAM[b,c,d] and Yu-Chuan (Jack) LI[a,b,c,d,e,1]
[a] *Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei 110, Taiwan*
[b] *AESOP Technology, Songshan District, Taipei 105, Taiwan*
[c] *International Center for Health Information Technology (ICHIT), Taipei Medical University, Taipei 110, Taiwan*
[d] *Research Center of Big Data and Meta-Analysis, Wan Fang Hospital, Taipei Medical University, Taipei 116, Taiwan*
[e] *Department of Dermatology, Wan Fang Hospital, Taipei 116, Taiwan*

Abstract. Most screening tests for Diabetes Mellitus (DM) in use today were developed using electronically collected data from Electronic Health Record (EHR). However, developing and under-developing countries are still struggling to build EHR in their hospitals. Due to the lack of HER data, early screening tools are not available for those countries. This study develops a prediction model for early DM by direct questionnaires for a tertiary hospital in Bangladesh. Information gain technique was used to reduce irreverent features. Using selected variables, we developed logistic regression, support vector machine, K-nearest neighbor, Naïve Bayes, random forest (RF), and neural network models to predict diabetes at an early stage. RF outperformed other machine learning algorithms achieved 100% accuracy. These findings suggest that a combination of simple questionnaires and a machine learning algorithm can be a powerful tool to identify undiagnosed DM patients.

Keywords. Diabetes, machine learning, random forest, early-stage prediction

1. Introduction

Diabetes Mellitus (DM) is one of the major global public health concerns, imposing an immense financial burden on public health. The diabetes burden has been increased over the time both in developing and developed countries due to the complex reason [1,2]. According to estimates from the International Diabetes Federation (IDF) in 2021, approximately 537 million adults are living with DM, and the total number of DM patients is expected to rise 643 million by 2030 [2,3].

Diabetes is associated with premature deaths from both communicable and non-communicable diseases. Therefore, early prediction of DM has great potential to reduce financial burden as well as increase life expectancy[4]. Machine learning is a growing field in healthcare for its promising performance in term of early prediction, diagnosis,

[1] Corresponding Author, Yu-Chuan (Jack) LI, Graduate Institute of Biomedical Informatics, College of Medical Science and Technology, Taipei Medical University, Taipei 110, Taiwan; Email: jack@tmu.edu.tw / jaak88@gmail.com.

classification, and risk stratification of diseases[5]. This study proposes a diabetes prediction model that offers an earlier prediction of DM using simple variables.

2. Methods

2.1. Dataset

We collected data from the UCI repository diabetes database to evaluate the performance of machine learning model for earlier diabetes prediction. A direct questionnaire was used to collect patient information for a tertiary hospital in Bangladesh. Dataset was comprised of 16 variables with 320 diabetes and 200 non-diabetes patients.

2.2. Preprocessing and feature section

There were no missing value in the dataset. We just converted nominal data to numeric data, such as male and female was converted into 0 and 1. To evaluate the variables that have the most influence on the outcome prediction, SHAP (SHapley Additive exPlanations) was used to get the potential variables. The idea behind SHAP feature importance is straightforward: variables with large absolute shapley values are important. Figure 1 shows the summary plot that combines feature importance with feature effects. However, we considered only the top ten variables in our model.

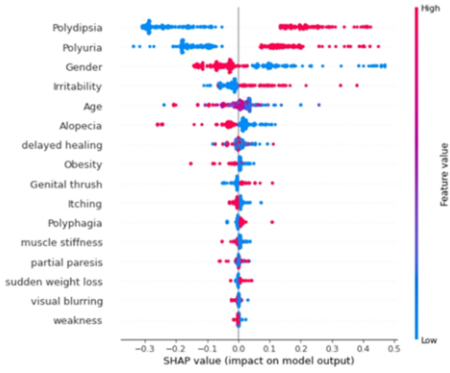

Figure 1. SHAP summary plot

We used six different classification models such as random forest, support vector machine, K-nearest neighbor, logistic regression, support vector machine, and multi-layer perceptron. The dataset was randomly divided into a derivation set (70%) used to train the model and a validation set (30%) used to test the accuracy of the model. The accuracy, precision, recall, and F1 score were used to measure the performance of each model.

3. Results

3.1. Patient Characteristics

The mean age of diabetes patients was 49.07, and non-diabetes patients were 46.36. Male comprised 63.07% of the study population. Approximately 86.5% of female patients had diabetes. Of the 320 diabetes patients, 243 (81%) patients had polyuria, 225 (75%) patients had polydipsia, 188 (62.67%) patients had sudden weight loss, 218 (68.12%) patients had weakness.

3.2. Importance variable for machine learning model

In predicting the primary outcome (diabetes), the ten most significant predictors for the SHAP were age, gender, weakness, itching, polyuria, delayed healing, visual blurring, polyphagia, polydipsia, and partial paresis. Figure 2 shows the variable's importance in the prediction of diabetes by SHAP.

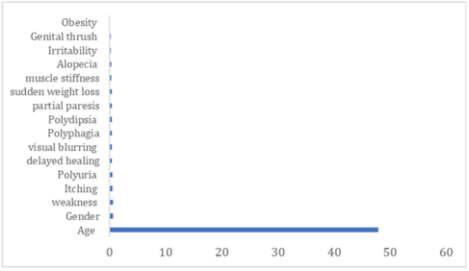

Figure 2: Variable importance by SHAP

3.3. Performance of machine learning

The performance of the machine learning models in the final testing set are summarized in Table 1. The random forest (RF) model showed high areas under the receiver operating characteristics curve, high accuracy, precision, and recall for prediction of diabetes patients early (Figure 3).

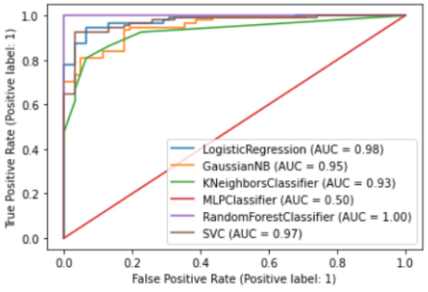

Figure 3. Area under the receiver operating characteristic curves is shown for all six machine learning models

The accuracy of LR, SVM, NB, MLP, and KNN was 0.94, 0.93, 0.92, 0.92, 0.85, respectively.

Table 1. Performance of machine learning algorithms

Model	Accuracy	Precision	Recall	F1-score
Random forest (RF)	1	1	1	1
Logistic regression (LR)	0.94	0.94	0.94	0.94
Support vector machine (SVM)	0.93	0.93	0.93	0.93
Naïve Bayes (NB)	0.92	0.92	0.92	0.92
Multi-layer perceptron (MLP)	0.92	0.92	0.92	0.92
K-nearest neighbor (KNN)	0.85	0.86	0.85	0.85

4. Discussion

We have developed machine learning models that accurately predict DM early using simple variables. Among the six classification models, random forest showed high prediction capability in terms of accuracy, precision, recall, and F-score. Random forest model outperformed both the previously published models [6,7], had the best predictive ability to identify patients at high risk for DM, and may be used to risk stratify patients with regard to their risk of DM development.

Although prior studies have developed predictive models for DM, but they did not used any feature selection techniques. In the clinical practice, fewer variables are better to predict disease because it is hard to use many variables to stratify patients. Our study is the first to use feature selection techniques and used only ten variables to predict DM. As DM is often unidentified at its early stage due to its asymptomatic characteristics[8]. Therefore, it's really important to diagnose DM at an early stage to minimize the severity and mortality of this disease [9]. Our study shows very promising performance to identify patients at high risk of DM earlier. Implementing our model in the clinical setting can help physicians to correctly stratify patients, better disease management, and minimize comorbidity through appropriate treatment on time.

5. Conclusions

These findings suggest that RF was the most accurate for predicting diabetes at early stage by using only 10 features. This study demonstrates a less time consuming and cost-effective screening model for predicting DM in a rural place.

References

[1] Cho N, Shaw J, Karuranga S et al. IDF Diabetes Atlas: Global estimates of diabetes prevalence for 2017 and projections for 2045. Diabetes research and clinical practice 2018;138:271-281.

[2] Wang L, Gao P, Zhang M et al. Prevalence and Ethnic Pattern of Diabetes and Prediabetes in China in 2013. JAMA 2017;317:2515-2523.

[3] Saeedi P, Petersohn I, Salpea P et al. Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas, 9(th) edition. Diabetes Res Clin Pract 2019;157:107843.

[4] Alshammari R, Atiyah N, Daghistani T et al. Improving Accuracy for Diabetes Mellitus Prediction by Using Deepnet. Online journal of public health informatics 2020;12:e11.

[5] Davenport T, Kalakota R. The potential for artificial intelligence in healthcare. Future healthcare journal 2019;6:94-98.
[6] Islam M, Ferdousi R, Rahman S et al. Likelihood prediction of diabetes at early stage using data mining techniques. In, Computer Vision and Machine Intelligence in Medical Image Analysis: Springer; 2020:113-125.
[7] Refat MAR, Al Amin M, Kaushal C et al. A Comparative Analysis of Early Stage Diabetes Prediction using Machine Learning and Deep Learning Approach. In, 2021 6th International Conference on Signal Processing, Computing and Control (ISPCC): IEEE;2021:654-659.
[8] [Anonymous]. Diagnosis and classification of diabetes mellitus. Diabetes care 2009; 32 Suppl 1: S62-67
[9] Yang JJ, Yu D, Wen W et al. Association of Diabetes With All-Cause and Cause-Specific Mortality in Asia: A Pooled Analysis of More Than 1 Million Participants. JAMA Network Open 2019;2:e192696-e192696.