Diabetes Risk Prediction using Machine Learning: Prospect and Challenges

Shankaracharya*

The University of Texas MD Anderson Cancer Center, Houston, USA

*Corresponding author:** Shankaracharya, Postdoctoral fellow, University of Texas MD Anderson Cancer Center, Houston, USA; [Tel] - 17134976813; E-mail- FShankaracharya@mdanderson.org

Introduction

According to first WHO Global report on diabetes published on world health day April 7, 2016, the number of adults living with diabetes has almost increased four times since 1980 to 422 million. It caused about 1.5 million deaths in 2012. This threatening numbers necessitates the development of effective and accurate diagnosis tools that may reach to the table of clinicians. Diabetes diagnosis is based on various epidemiological and genetic factors. Epidemiological risk factors include smoking status, dietary habits, physical activities, BMI etc. whereas genetic factors are the inherited causative genes from parents. Hence it is very necessary to consider all factors collectively to get the accurate prediction and diagnosis of the disease. Many factors including lack of experience or fatigue of an expert may lead to incorrect diagnosis. Therefore computational approach may provide a strong alternative for diabetes prediction and diagnosis. These computational tools may help clinicians to make accurate diagnosis. Also, it will help individuals to get acquainted about their health status and future possible diabetic condition so that they can get chance to adopt better lifestyle to prevent the disease. This editorial aims to present the prospect and challenges of diabetes risk prediction using supervised machine learning methods.

World is moving towards the revolutionized application of computational methods for the prediction of many common and complicated diseases like diabetes and cancer. Machine-learning methods are the most popular and effective tool that has the capacity to improve the accuracy of the prediction and diagnosis of diabetes diagnosis. Several machine-learning tools have already been applied for the developement of risk prediction models[1]. The methods producing high accuracy in prediction includes (but are not limited to) Artificial Neural Net-

work (ANN), Mixture of Experts (ME), Random Forest (RF), Regression Tree (RT), and Support Vector Machine (SVM)[2,3]. Some of these current methods have shown their potential in the prediction of diabetes with high accuracy. Along with the selection of methods machine learning model accuracy also depends on the quality and quantity of data used for training. The prediction with the data having less noise and more samples leads to more accurate prediction of diabetes risk. Recently the google team is able to successfully demonstrate the role of machine learning[4] in the detection of diabetic retinopathy and macular edema using retinal fundus photographs with high sensitivity (about 90%) and specificity (more than 98%). They use the deep learning approach, a recently developed ANN based supervised machine learning methods that helps to learn machine about the pattern to classify the images of cases and controls. This kind of research has started the new era of effective and accurate diagnosis and prediction of diabetes outcome. The availability of genetic data has also empowered the method of diabetes risk prediction using machine learning methods[5].

The machine learning models are complicated in nature compared to general statistical regression methods. In general machine learning (ML) model structure consists of applied parameters and their learned values for the training dataset. Hence the availability of clean and large dataset plays an important role in the development of accurate and reliable model. Also, disease risk prediction models are usually specific for a particular population and the single model may not apply for all populations. Therefore, different risk prediction models are needed for different population datasets. The major challenges in the disease risk prediction modeling with the machine learning methods include the lack of reproducibility and external validation. This is primarily due to the unavailability of models generated from the

research and the program objects used to make the model. Thus, there is a need of development of tool that can accommodate the option of using most of the machine learning methods and can work with large amount of datasets. Moreover, the tool should have the ability to perform the cross validation of the generated model to get confidence on the model and should be able to predict diabetes in the next successive years. The prediction of diabetes in next 5-year, 10-year or any number of years in a population will help in making effective plans to combat the disease.

Conclusion

Machine learning has the great ability to revolutionize the diabetes risk prediction with the help of advanced computational methods and availability of large amount of epidemiological and genetic diabetes risk dataset. The technique may also help researchers to develop an accurate and effective tool that will reach at the table of clinicians to help them make better decision about the disease status.

References

1. Shankaracharya, et al. Computational intelligence in early diabetes diagnosis: A review. (2010). Rev. Diabetic Stud 7(4): 252-262.
2. Shankaracharya et al. Java-based diabetes type 2 prediction tools for better diagnosis. (2011) Diabetes Technol Ther14 (3): 251–256.
3. Shankaracharya et al. Computational intelligence-based diagnosis tool for the detection of pre diabetes and type 2 diabetes in India. (2012) Rev Diabet Stud 9(1): 55-62.
4. Gulshan et al. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fund us photographs. (2016) JAMA.
5. Keating, B.J. Advances in Risk Prediction of Type 2 Diabetes: Integrating Genetic Scores with Framingham Risk Models. (2015) Diabetes 64(5): 1495-1497.