Proceedings of the 3rd International Conference on Computing Innovation and Applied Physics
DOI: 10.54254/2753-8818/13/20240827

Machine learning and deep learning techniques in diabetes prediction

Yousheng Zhang

Faculty of Science and Technology, Beijing Normal University-Hongkong Baptist University United International College, Zhuhai, China.

s230005062@mail.uic.edu.cn

Abstract. Under the joint influence of environment and genes, diabetes mellitus has now become a growing issue with a series of complications, which leads to a low quality of life. Patients with developing diabetes may suffer from limited diets, scheduled medication, physical pain, and mental torment. Considering the high morbidity, it is of great significance to predict whether a person has diabetes and to take action to alleviate the disease effectively. So far, there are many studies focusing on diabetes prediction with high efficiency and accuracy by introducing specific data mining techniques and algorithms, such as Machine Learning and Deep Learning. The practice of Machine Learning techniques including Decision Tree, XG boost, and Random Forest in diabetes prediction, has been illustrated by different authors and compared accordingly. Artificial Neural Network, one of the Deep Learning methods, being used to predict diabetes patients, was put forward by various researchers with the comparison of the accuracy. In this article, the factors that influence the prediction accuracy are discussed. The practical application of these methods is discussed as well, with the aim of obtaining a higher accuracy of diabetes prediction in clinical fields.

Keywords: Diabetes Prediction, Machine Learning, Deep Learning.

1. Introduction

Considering the high pressure of medical diagnosis resulting from the rapidly growing population, it is of great significance and emergency to develop a series of clinical diagnosis techniques, such as data-based prediction, complication-based detection, and the distribution of resources. The technology plays an important role in bridging the connection of patients and medical which helps patients realize their disease as early as possible and provides them better understand of the diseases.

With the property of being incurable, diabetes has become one of the most serious diseases and spread widely throughout the world. Patients suffering from developing diabetes have no approach to completely cure this disease, only can alleviate the disease by either physical or chemical interruptions,which can either produce very little effect or be extremely expensive. Studying the clinical prediction of diabetes can effectively diagnose whether a patient has developing diabetes as early as possible, which provides the patient with a better life quality and minors the probability of complications. In addition, figuring out the most appropriate and suitable algorithm which can completely adapt the application of diabetes prediction can cut down the expenses of diabetes diagnose.

DOI: 10.54254/2753-8818/13/20240827

Diabetes detection is a considerable workload due to the variety of variables such as pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, and diabetes pedigree function. Thus, the crucial method to cope with the large number of variables is to utilize the algorithms of Machine learning and Deep learning. So far, various methods have been developed to adapt the intention of diabetes prediction, such as Decision Tree, XG boost, Random forest, Artificial neural network, Support Vector Machine, and Naive Bayes Classifier based on the algorithms of Machine Learning and Deep Learning. Therefore, in order to review the current literature on specific methods of Machine learning and Deep learning on diabetes prediction, attempts were performed within the framework of this study.

In this study, the application of three types of Machine Learning approaches and one type of Deep Learning method to the practice of diabetes prediction will be illustrated in specific to the organization as follows: section 2 provides Decision Tree, XG boost, and Random forest as Machine Learning methods; section 3 provides Artificial Neural Network as Deep Learning method.

2. Machine learning in diabetes prediction

2.1. Decision tree

Decision tree is one of the common data mining methods of machine learning for the establishment of classification systems based on numerous variables or the development of prediction algorithms for a target variable [1]. With a root node, internal nodes, and leaf nodes, the Decision tree model can deal with large, comprehensive datasets with a clear, visible structure.

Based on the tree-like structure of the decision tree, there are many applications of the decision tree, for example, the evaluation of financial risk, and the medical diagnosis, especially the patients who suffer from developing diabetes can be predicted.

Jarullah used the decision tree model to predict type 2 diabetes patients [2]. The datasets were collected from the Pima Indians Diabetes Data Set, including the information of patients with or without developing diabetes [2]. There are two stages in the study, the first stage is the data pre-processing, and the second stage is constructing a decision tree to predict diabetes patients [2]. The study used plasma glucose, diabetes pedigree function(DPF), and body mass index(BMI) accordingly to test if a patient is suffering from diabetes [2]. Even though this study did not consider other factors including smoking, family history, inactive lifestyle, and so on, the accuracy of this test reached 78.1768% [2].

Dudkina et al. used the same model to predict diabetes patients datasets consisting of 768 patients [3]. With 9 attributes included in the datasets, the researchers conducted three experiments based on different training and testing proportions [3]. According to the results of three experiments, it shows that the larger the datasets are, the more accurate the prediction [3]. In addition, the study shows that 50% for training and 50% for testing is the best method with 0.71 accuracy [3].

Mary Posonia et al. used a decision tree, one of the methods of machine learning, and took out the same datasets of the Pima Diabetes Database with parameters such as BMI, age, and 2-hour serum insulin [4]. Mary Posonia et al also showed that plasma, pedigree, and pregnant times were the most crucial characteristics of this decision tree model [4]. This model was trained to show the result of 91.2% accuracy of total patients [4].

All 3 studies above used the same diabetes datasets, the Pima Indians Diabetes Data Set, to predict patients suffering from developing diabetes. The reason for the three experiences showing different accuracy is different prediction objects, using different methods to pre-process data, and using different training-testing proportions. Jarullah aimed at type 2 diabetes patients, while Dudkina et al. and Mary Posonia et al. aimed at all diabetes patients [1-3]. Dudkina et al. used a 50% training-testing proportion which the other two studies did not mention [4]. In addition, all three studies used their own way to pre-process the datasets.

DOI: 10.54254/2753-8818/13/20240827

2.2. XG boost

Extreme gradient boosting package which is known as XG boost. An effective linear model solver and a tree learning algorithm are included in the package. It provides a number of objective operations, such as ranking, classification, and regression [5]. With high efficiency, the package can solve multiple data science problems within a short time. It has been widely used in medical practical problems, such as chronic kidney disease diagnosis, classification of patients with epilepsy, and diabetes prediction especially.

Due to the large amount of variables that need to be considered during the medical diagnosis, Paleczek et al. developed a new XG boost algorithm to interpret the medical data based on acetone which was considered a diabetes biomarker of the breath stimulation [6]. With the artificial intelligence element mixing in the exhaled human breath, the XGboost algorithm showed great performance and was highly selective to the marked acetone, regardless of small concentration.

Wang et al. used the XGboost algorithm to predict the patients with type 2 developing diabetes and compare it with another advanced algorithm such as random forest, support vector machine, and K-nearest neighbor [7]. Using the questionnaire in the Beijing area, 380 people were asked to take the survey about their personal information [7]. The result of this research showed that XGboost holds the best generalization capacity and prediction precision which reached 0.8909 [7]. This helps to lead to a better estimation of potential diabetes patients and offers novel thoughts on chronic disease prevention.

Midroni et al. used datasets named OhioT1DM datasets to predict the potential type 1 diabetes patients [8]. Midroni et al. demonstrated that the blood glucose level at a 30-minute horizon is a reflection of type 1 diabetes [8]. According to them, XGboost was used as a predictor of blood glucose level to estimate the patients with type 1 diabetes [8]. The research leads to the conclusion that XGboost is a useful and proper predictor of diabetes diagnosis, and the different biomarkers make a difference in the accuracy of the XGboost.

The above 3 studies showed different results due to different purposes and different observing objects. Paleczek et al. used the artificial intelligence element acetone to estimate the outcomes, while Midroni et al. used blood glucose level as an observing object [6,8]. Moreover, Wang et al.aimed at predicting type 2 diabetes patients, and Midroni et al. aimed at patients who suffer from type 1 diabetes by contrast [7,8].

2.3. Random Forest

Random forest technique is an algorithm containing multiple decision trees and is used as an excellent classification and regression method. The datasets are randomly picked out and then put back. At the same time, some characteristics are picked out as input. This approach has many advantages, including high accuracy, dealing with large input variables, evaluating the importance of the variable, and a fast learning process.

According to the advantages of this approach, the random forest model can be utilized in many different fields, such as facial recognition, social network analysis, image classification, financial risk control, and the prediction of medical diagnosis. In this part of the article, the application of the random forest for the prediction of diabetes will be illustrated in detail.

Xu et al. used a random forest algorithm to predict the patients with type 2 developing diabetes whose datasets were re-collected from the University of Virginia [9]. Xu et al. compared the random forest approach to the naive Bayes algorithm, ID3 algorithm, and AdaBoost algorithm [9]. By analyzing the given indicators, such as age, height, waist, and hip, which were included in the datasets, it turns out that, the accuracy and sensitivity of the random forest algorithm are superior to the naive Bayes algorithm, ID3 algorithm, and AdaBoost algorithm. Also, the research found that the random forest model's accuracy is also continuously increasing while maintaining the same quantity of increased data in the same proportion of cases, which indicates that the random forest algorithm is an effective method of predicting the risk of type 2 diabetes.

DOI: 10.54254/2753-8818/13/20240827

Benbelkacem et al. collected the data from the Pima Indians Diabetes data set and split the data into two groups, one group of data was used for training to optimize the number of the decision trees, and the other one was used for testing and comparing the random forest model to other machine learning models [10]. The study shows that with 40 trees in the random forest algorithm, the random forest algorithm has the lowest error rate among C4.5, RepTree, SimpleCart, BFTree, and SVM, which is 0.21. Based on this result, the author suggested that the reason random forest offers a better performance is that random forests are a collective approach to classifiers that can benefit from the synergies of individual classifiers to enhance performance.

Both two studies showed that the random forests algorithm performed better than other machine learning approaches, despite their different datasets. The advantages of using the random forests model can be concluded from the two articles. Firstly, the indicated variables can be easily obtained which can largely reduce the cost of diagnosis [9]. Secondly, based on the satisfactory result, the random forests algorithm can be utilized to help the handling of children's emergencies [10].

3. Deep Learning in diabetes prediction

Artificial neural networks are operational models which were connected by a large number of nodes, also called neurons. Each neuron represents a specific output function, the so-called activation function, and each connection between different neurons represents a weighted value accordingly. There are four basic characteristics of the neural network, which are non-linear, non-limited, non-qualitative, and non-convex [11].

In recent years, with the deepening exploration of the neural network, the artificial neural network, so far, has already made huge progress, especially in the fields of smart robots, automatic controlling, predicted estimation, biology, and medicine. In this article, the practice of neural networks in the prediction of diabetes patients will be elaborated.

Sapon et al. used MATLAB to train the artificial neural network algorithm by using 250 diabetes patients between 25 to 78 years old from Pusat Perubatan Universiti Kebangsaan Malaysia, Kuala Lumpur [12]. This data contains 27 variables including blood pressure, urine PH, and fasting glucose [12]. After the process of the training algorithm, the author used different methods to analyze the correlation between expected results and predicted results. The result shows that the Bayesian regulation performed best and had the best accuracy of 88.8% which indicates that this method is suitable for the task of diabetes prediction [12].

Jerjawi et al. utilized an artificial neural network to forecast the patients with developing diabetes in order to cut down the expense of caring for diabetes patients [13]. By analyzing the data, which was collected from the documentation of Urmia's Association of Diabetics, with the algorithm that offers the neural network's operation, the researchers found that under the environment of just neural network (JNN), the algorithm of an artificial neural network can successfully predict the patients with developing diabetes with the accuracy of 87.3% [13].

The two articles mentioned above both showed high accuracy in predicting diabetes patients by using the artificial neural network.

4. Conclusion

In this study, a number of methods of Machine learning and Deep learning have been discussed, ranging from different techniques such as Decision Trees, XG Boost, and Random Forests to Artificial Neural Networks, in order to address the current problem of diabetes prediction. According to the research results, each algorithm model showed a good performance with high prediction accuracy and clear construction. In addition, some of the algorithms provide more accurate detection results after adjusting the training-testing proportion. Moreover, the study shows that the high accuracy of diagnosis often comes with a large amount of data of diabetes patients. Hence, the researchers suggest that this clinical detection model can be improved specifically by providing the right proportion of training-testing and offering considerable clinical datasets. Through the application of different algorithms such as Decision Trees, XG Boost, Random Forest, and Artificial Neural Networks, the

DOI: 10.54254/2753-8818/13/20240827

researchers, medical practitioners, and clinical data analysts can obtain better outcomes of prediction by altering the training-testing rate and adjusting the number of data, which makes the diagnosis of patients with developing diabetes more accurate, and helps the patients taking the early action to alleviate disease. Although this study has not compared the efficiency between different methods of machine learning and deep learning due to the use of different datasets. The differences in research purpose, and observing variables, this research still has a positive impact on clinical diagnosis such as providing available access to the algorithm options and the reduction of clinical cost because of the use of effective prediction algorithms.

Reference

[1] Song, YY. Lu, Y. 2015, Decision tree methods: applications for classification and prediction. (Shanghai Arch Psychiatry, vol. 27), no. 2, pp. 130-5.
[2] Jarullah, A. A. Al. 2011, "Decision tree discovery for the diagnosis of type II diabetes," 2011 International Conference on Innovations in Information Technology, pp. 303-307.
[3] Dudkina, T. Meniailov, I. Bazilevych K, et al. 2021, Classification and Prediction of Diabetes Disease using Decision Tree Method, pp. 163-172.
[4] Posonia, A. M. Vigneshwari,S. Rani, D. J. 2020, "Machine Learning based Diabetes Prediction using Decision Tree J48," (2020 3rd International Conference on Intelligent Sustainable Systems (ICISS)), pp. 498-502.
[5] Chen, T. He, T. Benesty, M. et al. 2015. Xgboost: extreme gradient boosting. (R package version, vol. 1), no. 4, pp. 1-4.
[6] Paleczek, A. Grochala, D. Rydosz, A. 2021, Artificial Breath Classification Using XGBoost Algorithm for Diabetes Detection. (Sensors, vol. 21), pp. 4187.
[7] Wang, L. Wang, X. Chen, A. Jin, X. Che, H. 2020, Prediction of Type 2 Diabetes Risk and Its Effect Evaluation Based on the XGBoost Model. (Healthcare vol. 8), pp. 247.
[8] Midroni, C. Leimbigler, P. J. Baruah, G. et al. 2018, Predicting glycemia in type 1 diabetes patients: experiments with XGBoost. (Heart, vol. 60). no. 90, pp. 120.
[9] Xu, W. Zhang, J. Zhang, Q. Wei, X. 2017, "Risk prediction of type II diabetes based on random forest model," Communication and Bio-Informatics (AEEICB), pp. 382- 386,
[10] Benbelkacem, S. Atmani, B. 2019, "Random Forests for Diabetes Diagnosis," 2019 International Conference on Computer and Information Sciences (ICCIS), pp. 1-4.
[11] Wu, Y. C. Feng, J. W. 2018, Development and Application of Artificial Neural Network. (Wireless Pers Commun, vol. 102), pp. 1645–1656 .
[12] Sapon, M. A, Ismail, K. Zainudin, S. 2011. Prediction of diabetes by using artificial neural network. Proceedings of the 2011 International Conference on Circuits, System and Simulation, Singapore. vol. 2829, pp. 299303.
[13] Jerjawi, N. S. Abu-Naser, S. S. 2018, Diabetes Prediction Using Artificial Neural Network. (International Journal of Advanced Science and Technology, vol. 121), pp. 54-64.