International Journal of Distributed Systems and Technologies
Volume 13 • Issue 1

Development of a Diabetes Diagnosis System Using Machine Learning Algorithms

Victor Chang, Aston University, UK*
https://orcid.org/0000-0002-8012-5852

Keerthi Kandadai, Teesside University, UK

Qianwen Ariel Xu, Teesside University, UK

Steven Guan, Xi'an Jiaotong-Liverpool University, China

ABSTRACT

This paper describes how to develop diabetes diagnosis through the combined use of the support vector machine, the decision tree, naive bayes, k-nearest, and finally, random forest (RF) algorithms. These methods are useful to predict diabetes jointly. The appropriateness of ML-depended techniques to tackle this issue has been revealed. This diabetes diagnosis system using machine-learning algorithms is used to review papers. This project was based on developing Python-based code for machine learning algorithms to perform large scale diabetes analysis. The hardware requirement of machine learning is RAM that is 128 GB DDR4 2133 MHz and 2 TB Hard Disk and needs 512 GB SSD. One standard library is NumPy, which is used to support multi-dimensional arrays objects, various components, and matrices. The random forest prediction representing the pictorial visualization of the model and the accuracy for the data analysis using the random forest is 76%.

KEYWORDS

Diabetes Diagnosis, Machine Learning for Healthcare, Random Forest Prediction

1. INTRODUCTION

Doctors are trained based on traditional knowledge for treatment purposes with insufficient studies, and they acquire deep implicit knowledge after years of studies. However, the demands for diabetes mellitus are high and possibly outside the knowledge from books. Diabetes mellitus patients are faced with metabolic health problems, neuropathy, skin condition and many more. An individual diagnosed with diabetes patients is uncovered to hyperglycaemias for many years and out coming is at higher risk for diabetes difficulties. In addition, this traditional process takes a long time. Diabetes prediction using machine learning helps to identify the patterns much earlier. According to Alghamdi et al. (2017), machine learning is a difficult subject with popularity because it can serve as a general-purpose programming language and its adoption in both scientific computing and artificial intelligence. Often this study requires the support of use cases and their applications in different domains or disciplines.

DOI: 10.4018/IJDST.296246

*Corresponding Author

Machine learning is used for predicting diabetes and getting preferable outcomes. Decision trees are a well-liked machine learning technique in the medical area and effectively provide probabilities of key decisions. In this regard, Neural network algorithms are currently a common machine learning approach that performs better. The support vector machine, the Decision Tree, Naive Bayes algorithm and K-nearest algorithm, and the Random Forest (RF) algorithm have been discussed in this paper. The random forest model assumes a large number of decision trees, and those are used to predict diabetes. Here, this report used a diabete.csv dataset that helps to predict the diabetes rate.

The research aim of the project is to design a high-accuracy diabetes prediction application through machine learning algorithms, which can help identify patterns earlier. The primary goals of this study are as follows:

- To construct a web application that will allow users to evaluate whether or not a patient has diabetes.
- To use the algorithms of SVM, Decision Tree and Random Forest to manipulate the data.
- To critically visualize the data set and find out the correlation between several features of the data set.
- To develop a general workflow of the predicted outcomes To evaluate the outcomes of the end products critically.

This paper is organized as follows: Section 2 reviews the literature on diabetes prediction and machine learning algorithms. Section 3 introduces the methods used in this research. Section 4 reports the results after conducting the algorithms, followed by the discussion on these results in section 5. Finally, we conclude our research in Section 6.

2. LITERATURE REVIEW

2.1 Diabetes Prediction and AI Techniques

Numerous research has attempted to deal with the issue of diabetes prediction with the support of machine learning or artificial intelligence techniques, including support vector machine (SVM), Decision Tree, Gradient Boosting Decision Tree, Naive Bayes, ANN (Artificial Neural Network), etc.

The early diagnosis of diabetes is the focus of the studies on diabetes prediction as in the early stage of diabetes. Suitable treatment can largely minimize the expenditure and mortality in the subsequent phases. Fitriyani et al. (2019) developed a disease prediction model for the early diagnosis of hypertension and diabetes. Their model employed iForest to remove outliers, SMOTETomek to balance data distribution and an ensemble learning approach to make the prediction, and it was proved to have high accuracy. Samant and Agarwal (2018a; 2018b) and Rashid et al. (2021) carried out a comparative analysis on the performance of Random Forest with other ML algorithms. In their research, Random Forest showed better ability in making care decisions. Khanam and Foo (2021) used seven algorithms of machine learning algorithms, data mining and neural network methods to predict diabetes. They used the Pima Indian Diabetes data set for research and found that classifiers based on Support Vector Machines and Logistic Regression effectively predict diabetes. Moreover, they observed that the neural network with two hidden layers achieved an accuracy of 88.6%. In addition to studying diabetes prediction, Ahmad et al. (2021) also compared the role of glycosylated hemoglobin, HbA1c, and FPG as input features. They used Decision Trees, Support Vector Machines, Logistic Regression, Ensemble Majority Voting, and Random Forest to build classifiers and employed feature elimination through feature replacement and hierarchical clustering. They found that SVM performed best on the HbA1c-labeled data set, while RF performed best on the FPG-labeled data set. Different from other scholars, Chong et al. (2021) conduct research on diabetes management from another perspective. They explored different methods of topic modeling, mainly the latent Dirichlet

Utilizing an Augmented Reality System to Address Phantom Limb Syndrome in a Cloud-Based Environment

Afeef Sheikh (2017). *International Journal of Grid and High Performance Computing (pp. 14-24).*

www.irma-international.org/article/utilizing-an-augmented-reality-system-to-address-phantom-limb-syndrome-in-a-cloud-based-environment/181033