Turkish Online Journal of Qualitative Inquiry (TOJQI)

Volume 12, Issue 3, June 2021: 4461 - 4468

Research Article

Deep learning approach for Diabetes Risk Analysis

Vidyullatha Sukhavasi[1], P Hima[2], Putta Sujitha[3]

Abstract

Diabetes is a chronic disease that is affecting more and more people around the world because of high glucose levels. The prevalence and incidence rates have been very much to increase year-on-year. If left untreated, the health problems associated with diabetes in most of the organs of the body can be devastating. The main purpose of which is to predict in patients with diabetes mellitus. Various classification algorithms such as Support Vector Machine(SVM), K-nearest Neighbor (KNN), Decision tree, Naïve Bayes(NB), can be used on the Pima Indians Diabetes Database (PIDD), downloaded from the UCI repository of machine learning and compared with an approach of deep learning used to predict the development of diabetes mellitus. The accuracy is measured based on machine learning(ML) classification algorithms with a comparison of deep learning(DL) model.

Keywords: Machine Learning; Deep Learning; Diabetes prediction; PIMA Dataset

1.Introduction

According to the International Diabetes Federation (IDF) in 2019, 463 million people diagnosed with diabetes worldwide. About 1 in 11 of the world's population is affected by diabetes. The impact of diabetes, which is incurable, increases rapidly and leads to an increase in mortality. As a result, there will be an increase of 700 million affected people by 2045 [1]. China has the biggest population of diabetics in the world, with 116 million individuals suffering from the disease. The increase in cases of diabetes can be reduced by changing one's diet and daily lifestyle. Sugar in the blood that affects the ability of the body to grow. Insulin is released, causing carbohydrate digestion to become imbalanced and blood sugar levels to rise. Typically, a person with diabetes from high blood glucose levels. Many different methods are used in the medical field, such as diabetes.

[1]Assistant professor, VFSTR (Deemed to be University), Email- vidyullatha.1988@gmail.com
[2]Assistant professor, VFSTR (Deemed to be University), Email- himasyamala137@gmail.com
[3]Assistant professor, VFSTR (Deemed to be University), Email- sujithacse2012@gmail.com

Diabetes is caused by a lack of insulin production that lowers blood glucose levels. A person with diabetes suffers from hunger, thirst and urinary incontinence due to increased sugar levels [2]. Diabetes will be affected by genetic problems and by a decrease in insulin factor. The only solution is to get symptoms early and get the necessary treatment by consulting a physician to avoid complications [3]. Traditionally predicting diabetes by performing tests doctors collect information such as age, body index, diastolic blood pressure, glucose levels etc. to determine the

symptoms and determine whether a person has diabetes or not. But it takes time to get a decision from a doctor because you need to use his knowledge and evaluate past health conditions and current patient outcomes. This causes a great deal of seriousness among researchers; the prediction of the disease prevalent in the body is one of the critical tasks. A support framework is provided using algorithms that are accessible and need to boost the performance of existing approaches to make the prediction process simple. So as support to physicians, there is a potential for predicting diabetic disease.

The remainder of the manuscript, which will be as follows: Section 2 gives a quick rundown of all the researchers who have worked in this area. A comprehensive review of machine learning algorithms and operations is provided in section 3. On the basis of the test results, Section 4 outlines a method for predicting diabetes mellitus. The fifth section comes to an end.

2.Related Work

An analysis is provided by the author Swapna and co-authors [4] with the help of the machine learning which have shown that it is effective and accurate in building a diabetes prediction model using HRV signals throughout the DL method. This work is done to prevent the risk of deaths in our country every year due to diabetes. The author developed a novel speculation model that included a convolution neural network (CNN), long short-term memory (LSTM), and a hybrid technique to determine the nature of HRV input data. The Support vector machine is then added into the segmentation properties. This proposed method could aid doctors in the diagnosis of diabetes.

Author Sajida Perveen [5], and others suggest the use of a new AdaBoost-based technologies. The AdaBoost ensemble of model has been used for the evaluation of diabetic patients, which is similar to the Bags, and J48. The diagnosis and treatment of diabetes mellitus is becoming more and more important for the medical community, [6]. A prediction model is proposed for the improvement of the Canadian population of patients with diabetes in the three age groups. Sacks and bags, J48, and AdaBoost are three of the ensemble of the models that the authors, in the review of the data-to-measure, precision, and performance. In each of these three models, AdaBoost performs well in terms of accuracy, and may also be used for the prediction of hypertension and coronary heart disease.

Stefan, et al. [7] describe the importance of data mining approaches are associated with serious health problems. The model proposed by the authors, it is the recognition of the risk of the disease. They are suited to the range of the strategy in real-world health data, and a comparison of the available clinical data, with the help of an algorithm.

DL seems to be a type of machine learning that differs from standard methods in that it learns from a variety of various types of raw data. On the basis of artificial neural networks (ANNs), this allows you to process and present data at multiple levels of abstraction and in different computational models with varied levels of processing [8].

3.Background study

3.1 Support Vector Machine(SVM): The SVM [9] method is used to address regression and classification issues. The hyperplane with the biggest derivative, which is the space between the two classes of data points, was used to decide the SVM algorithm. SVM makes use of the support available, and to improve the performance of the algorithm. In this article, you will find the three main points that support SVM, which are listed below.

The Linear Kernel can be simply done by performing dot product between two vectors which can be interpreted in Eq. (1).

$$\text{Kernel}(x, y) = x.\ y \quad (1)$$

The Polynomial Kernel is similar to Linear Kernel but consists of a polynomial degree associated with the Eq. (2)

$$\text{Kernel}(x, y) = (x \cdot y + C)^p \quad (2)$$

Radial Bias Kernel is the another kernel which finds a non-linear classifier described in Eq. (3)

$$\text{Kernel}(x, y) = \exp(-\gamma \|x - y\|^2) \quad (3)$$

3.2 K-Nearest Neighbor(KNN): KNN [10] is a supervised learning algorithm, which looks for similarities between the existing and the new data. KNN is primarily used for the regression, but also for the purpose of classification. Evaluate the distance between a point with (x, y) coordinates, and their next-door neighbor. The maximum distance between a point and its neighbor is determined by using the Eq. (4).

$$\text{Euclidean distance of point}(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)2} \quad (4)$$

3.3 Decision tree(DT): The classifier using decision tree uses the internal nodes[11] as symbols of the data set, where decision rules are represented by branches and result represented by leaf node. Decision Node and the Destination Node are the two nodes available in decision tree. The best splitting criteria to choose in decision tree be done with the help of Gini index and Entropy findings.

3.4. Naïve Bayes(NB) Classifier: A probabilistic machine learning model that can be used as a classification problem [12] using Naïve Bayes' classifier. This is the opinion of independent features that do not have an impact on the others. Bayes theorem is based on the probability concept:

$$prob(A|B) = \frac{\text{prob(A)prob(B|A)}}{\text{prob(B)}} \quad (5)$$

where prob (A|B) is A happens given that B occurs
prob(B|A) is B happens given that A occurs
prob(A) is A itself
prob(B) is B itself

4.Proposed Methodology and experiment results: A neural network consists of multiple layers having input and output layers. Networks are made up of the elements: neurons, weights and biases.

The proposed workflow of predicting diabetes using an artificial neural network is shown in Figure 1.

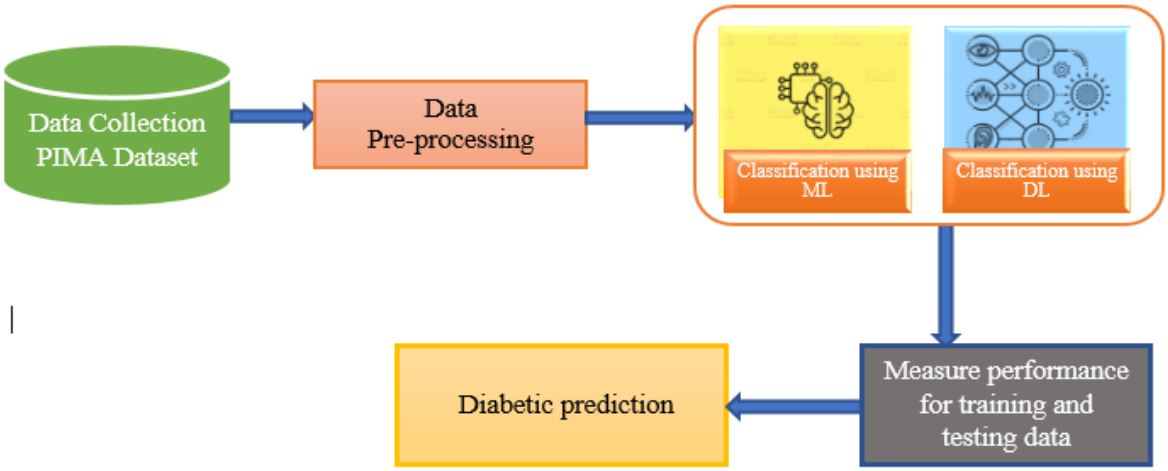

Figure 1 Proposed architecture for prediction of diabetes

4.1 Dataset

The dataset used in the proposed work to predict diabetes is PIMA Indian dataset [13]. The following Table 1 represents the features description in PIMA dataset and Figure 2 represents the histograms of dataset attributes.

Attribute name	Description	Feature values
NPREG	Number of times pregnant	Numeric
Age	Age in years	Numeric
BMI	Body Mass Index	weight in kg
Insulin	Serum insulin	Numeric
Glucose	Plasma Glucose concentration	Numeric
Skin	Skin fold Thickness(mm)	Numeric(mm)
BP	Blood Pressure	Numeric
PED	Pedigree function to know family inheritance	Numeric
Type	Non-diabetic, diabetic	Nominal

Table 1. Features description in PIMA dataset

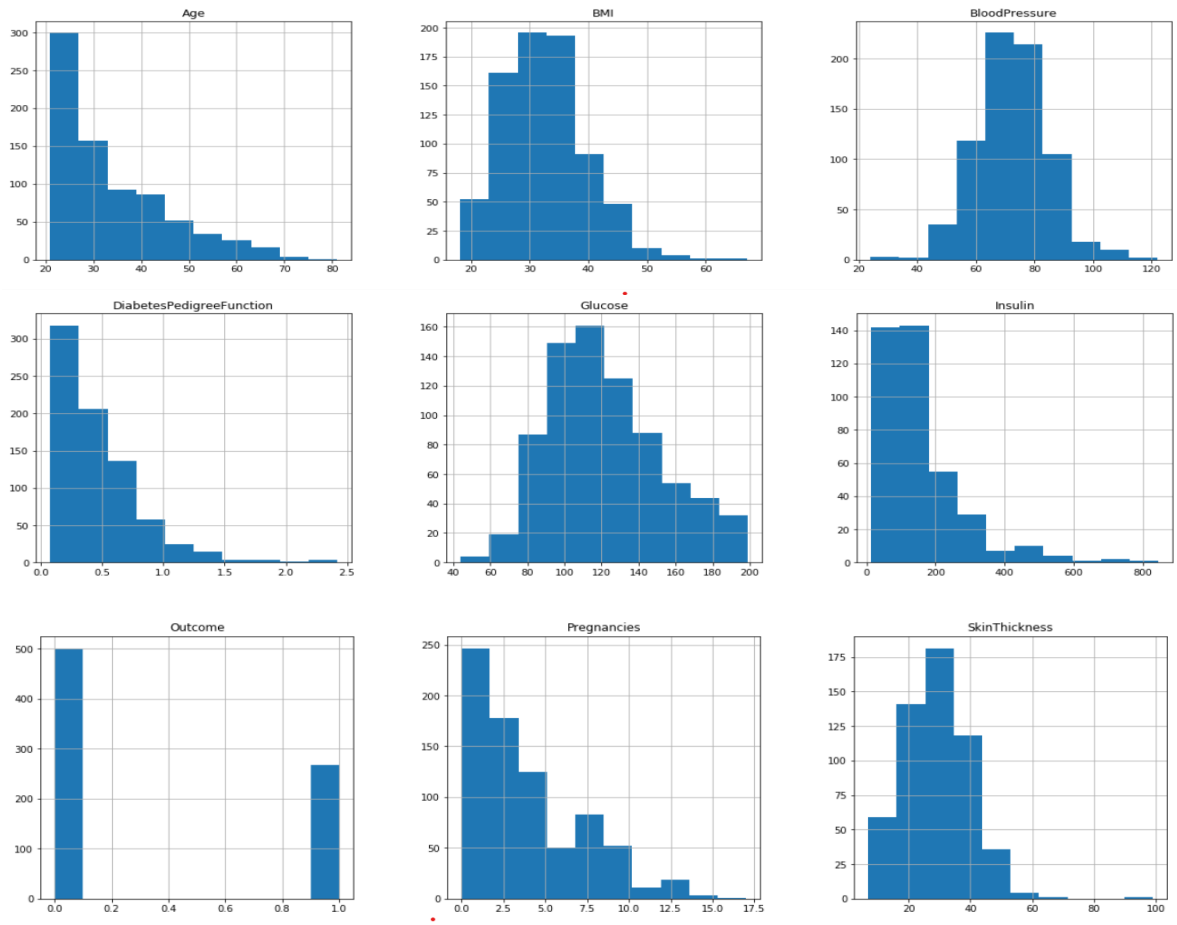

Figure 2. Histograms of dataset attributes

4.2. Methodology

4.2.1 Data Pre-processing

The standard of the data is crucial for the performance of any system. In this phase the instances in the dataset is removed, if the data value present as 0, which considered as a missed data. In the next step data discretization to improve the quality of data for effective classification.

4.2.2 Proposed work

The learning capability of machines is like an intelligence method which learns relation from data without having to plan ahead of time or define past correlations between data items [14]. DL is a method of teaching machines that differs from standard methods [15] in the way it captures from different types of raw data. DL enables various classifiers of neural networks to maintain data at various output forms [16]. The proposed method uses a stochastic gradient descent-trained multilayer-based perceptron feed-based model that supports ANN structures. A network is made up of four layers that contain nodes and neurons and are designed to communicate in one way. Each node has a single link to the next node and hidden layers helps each node to train global

parameters using local features. It also employs many phases to assess the model's contribution from the overall network. Advanced characteristics such as Relu, rectifier, and maxout activation, learning rate, are enabled by the backpropagation model of hidden layer neurons to improve the efficiency.

Table 2. shows the suggested model accuracy in comparison to existing ML models. The decision tree had a training accuracy of 76.23 % as well as testing accuracy of 73.45 %. Naive Bayes received 79.23 % training accuracy and 78.14 % testing accuracy. KNN received a training accuracy of 79.45 % and a testing accuracy of 77.32 %. SVM has a training accuracy of 76.18 % and a testing accuracy of 77.64 %. The proposed Neural Network achieved 82.80% training accuracy and 80.20% testing accuracy.

Classifier	Training Accuracy	Testing Accuracy
Decision tree	76.23	73.45
Naïve Bayes	79.23	78.14
K-nearest neighbour	79.45	77.32
SVM	76.18	77.64
Deep Neural Network	**82.80**	**80.20**

Table 2. Accuracies of different classifiers

Figure 3. depicts a graphical depiction of existing and proposed method accuracies.

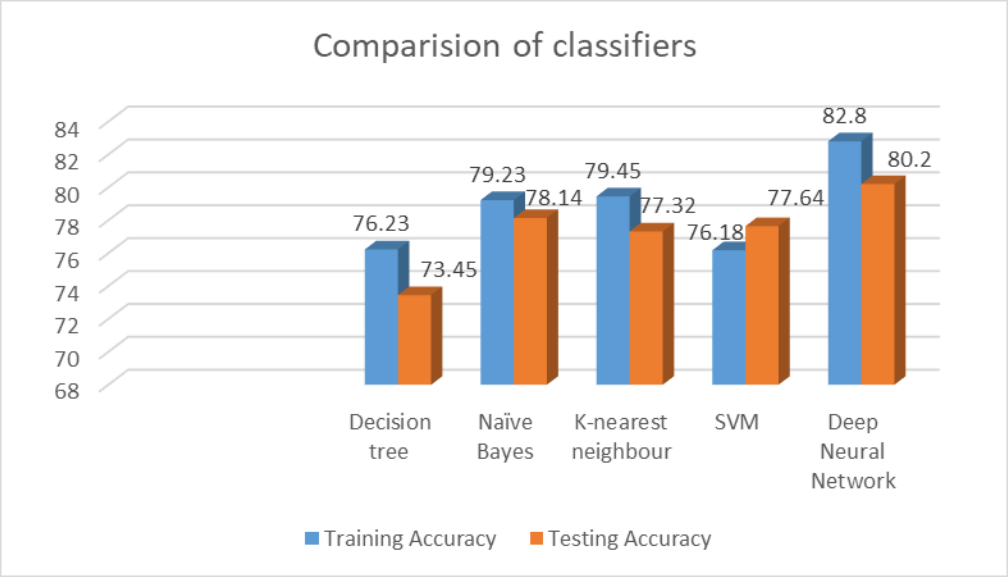

Figure 3. Comparison of different classifiers

5. CONCLUSION

The purpose of this study is to make use of a prediction algorithm to estimate the risk of diabetes. As previously said, diabetes affects the great majority of people. As a result, we used multiple classifiers in the PIMA database in our suggested study and shown that the proposed neural network reduces risk variables and improves the outcome in the terms of accuracy. The result obtained from the PIMA dataset using deep learning model is higher than machine learning algorithms as described in Table 2. The prediction rate accuracy of DT, NB, KNN, and SVM is between 70 and 78 percent. With an accuracy rate of 80.20 percent, the suggested Deep neural network is regarded the most effective and promising in diagnosing diabetes.

6. References

[1] Saeedi, P., Petersohn, I., Salpea, P., Malanda, B., Karuranga, S., Unwin, N., ... & Shaw, J. E. (2019). Global and regional diabetes prevalence estimates for 2019 and projections for 2030 and 2045: Results from the International Diabetes Federation Diabetes Atlas. *Diabetes research and clinical practice*, *157*, 107843.

[2] Kumar, D.A., Govindasamy, R., 2015. Performance and Evaluation of Classification Data Mining Techniques in Diabetes. International Journal of Computer Science and Information Technologies, 6, 1312–1319.

[3] Vijayan, V.V., Anjali, C., 2015. Prediction and diagnosis of diabetes mellitus A machine learning approach. 2015 IEEE Recent Advances in Intelligent Computational Systems (RAICS), 122–127doi:10.1109/RAICS.2015.7488400.

[4] Swapna G, Vinayakumar R, Soman KP. Diabetes detection using deep learning algorithms. ICT Express. 2018;4(4):243–6. https://doi.org/10.1016/j.icte.2018.10.005. Elsevier B.V.
[5] Perveen S, Shahbaz M, Keshavjee K, Guergachi A. Metabolic syndrome and development of diabetes mellitus: predictive modeling based onmachine learning techniques, IEEE Access. IEEE. 2019; 7:1365–75. https://doi.org/10.1109/ACCESS.2018.2884249.

[6] Perveen S, et al. Performance analysis of data mining classification techniques to predict diabetes. Procedia Computer Science.2016;82:115–21. https://doi.org/10.1016/j.procs.2016.04.016
Elsevier Masson SAS.
[7] Ravizza S, Huschto T, Adamov A, Böhm L, Büsser A, Flöther FF, et al. Predicting the early risk of chronic kidney disease in patients with diabetes using real-world data. Nature Medicine. 2019;25(1):57–9. https://doi.org/10.1038/s41591-018-0239-8. Springer US.
[8] Davazdahemami B, Delen D. The confounding role of common diabetes medications in developing acute renal failure: a data mining approach with emphasis on drug-drug interactions. Expert Systems with Applications. 2019; 123:168–77. https://doi.org/10.1016/j.eswa.2019.01.006. Elsevier Ltd.
[9] D. Çalişir and E. Dogantekin, "An automatic diabetes diagnosis system based on LDA-Wavelet Support Vector Machine Classifier," Expert Syst. Appl., vol. 38, no. 7, pp. 8311–8315, 2011.
[10] Tang, H., Xu, Y., Lin, A., Heidari, A.A., Wang, M., Chen, H., Luo, Y., Li, C. (2020). Predicting green consumption behaviors of students using efficient firefly grey wolfassisted K-

nearest neighbor classifiers. IEEE Access, 8: 35546-35562. https://doi.org/10.1109/ACCESS.2020.2973763.
[11] Mantovani RG. An empirical study on hyperparameter tuning of decision trees" arXiv : 1812 . 02207v2 [cs . LG]. 2019.
[12] Kaur, G., Oberoi, A. (2020). Novel approach for brain tumor detection based on Naïve Bayes classification. In: Sharma N., Chakrabarti A., Balas V. (eds) Data Management, Analytics and Innovation. Advances in Intelligent Systems and Computing, vol 1042. Springer, Singapore. https://doi.org/10.1007/978-981-32-9949- 8_31
[13] Naz, Huma, and Sachin Ahuja. "Deep learning approach for diabetes prediction using PIMA Indian dataset." *Journal of Diabetes & Metabolic Disorders* 19.1 (2020): 391-403.
[14] Swapna G, Vinayakumar R, Soman KP. Diabetes detection using deep learning algorithms. ICT Express. 2018;4(4):243–6. https:// doi.org/10.1016/j.icte.2018.10.005. Elsevier B.V.
[15] Wu H, et al. Type 2 diabetes mellitus prediction model based on data mining. Informatics in Medicine Unlocked. 2018;10:100–7. https://doi.org/10.1016/j.imu.2017.12.006. Elsevier Ltd.
[16] Davazdahemami B, Delen D. The confounding role of common diabetes medications in developing acute renal failure: a data mining approach with emphasis on drug-drug interactions. Expert Systems with Applications. 2019;123:168–77. https://doi.org/10. 1016/j.eswa.2019.01.006. Elsevier Ltd.