Predicting Diabetes using Deep Belief Network

Sureeluk Ma[1], Marusdee Yusoh[2*], Nitinun Pongsiri

(Received: 15 May,2023 ; Revised: 1 June,2023 ; Accepted 15 June,2023)

Abstract

Diabetes is a common disease affecting millions of people in the United States. The prevalence of diabetes has been steadily increasing over the past few decades. Left untreated diabetes leads to a risk factor for multiple complications such as heart disease, stroke, kidney and nerve damage. However, diabetes can be effectively managed when early diagnosed. Recent clinical data have applied various advanced technologies to diagnose or predict people with diabetes. Therefore, this paper aims to predict diabetes by using a Deep Belief Network (DBN) with Rectified Linear Unit (ReLU) activation function. Learning parameters from data through unsupervised path using minimized contrastive divergent algorithm, followed by supervised path using back-propagation algorithm. Diabetes Dataset from the National Institute of Diabetes and Digestive and Kidney Diseases was used, consisting of female patients aged at least 21 years old of Pima Indian heritage. The result shows that the DBN with ReLU activation function provides an 81% accuracy in diabetes prediction.

keywords: Artificial Neural Network, Deep Learning, Classification, Diabetes Modeling, Pima Indians

---

[1]**Faculty of Science and Technology, Princess of Naradhiwas University, Narathiwat, 96000, Thailand**

[2]**Islamic Sciences Demonstrations School, Prince of Songkla University, Pattani campus, 94000, Thailand**

[3]**Faculty of Science and Technology, Prince of Songkla University, Pattani campus, 94000, Thailand**

***Corresponding Madrusdee Yusoh, E-mail: madrusfee@gmail.com**

Introduction

Diabetes is a chronic disease that occurs when the pancreas does not produce enough insulin or the body can not effectively use the insulin. Insulin is a hormone that regulates blood glucose. The importance of insulin is to move glucose from the blood into body cells. Improper functioning of insulin causes glucose accumulation in the bloodstream. Three types of diabetes are known type 1, type 2 and gestational. An autoimmune reaction causes type 1 diabetes, which stops the body from producing insulin. Type 2 diabetes, usually caused by excess body weight and physical inactivity, results from the body's ineffective use of insulin. Gestational diabetes occurs during pregnancy, which is hyperglycemia characterized by higher than normal blood glucose values. Women with gestational diabetes have an increased risk of pregnancy complications, and their children also have an increased risk of type 2 diabetes (Centers for Disease Control and Prevention, 2023).

Approximately 422 million people worldwide have diabetes, the majority of whom live in low-and middle-income countries (World Health Organization, 2023). The number of cases and the prevalence of diabetes have been increasing dramatically over the past few decades. Each year, 1.5 million deaths are directly attributed to diabetes (World Health Organization, 2023). The Pima are North American Indians who traditionally lived along the Gisa and Salt rivers in Arizona, United States (US). The population of Pima Indians has one of the highest prevalence of diabetes in the US (Pearson, 2015).

Recent healthcare studies have applied various technologies to predict disease dynamics based on clinical data. A variety of diabetes prediction algorithms have been proposed to diagnose diabetes. For instance, Wu, Diao, Li, Fang & Ma (2009) have proposed a semi-supervised learning method based onLaplacian support vector machine (LapSVM) to predict diabetes. The results show that LapSVM can be of great help to physicians in the process of diagnosing diabetes. Sanakal & Jayakumari (2014) have diagnosed diabetes using the prognosis of fuzzy c-means clustering and supper vector machine (SVM). Their findings indicated an

accurate implementation of SVM through SMO algorithm for diseases diagnosis. Dagliati et al. (2018) proposed logistic regression (LR), naive Bayes (NB), SVMs, and Random forest (RF) to predict diabetes. The results showed that the values of AUC are higher for SVMs and RF. Gadekallu et al. (2020) classified the extracted features of diabetic retinopathy dataset using the principal component analysis based deep neural network model using Grey Wolf Optimization (GWO) algorithm. The proposed model is further compared with the traditional machine learning (ML) algorithms such as the SVM, NB Classifier, decision trees (DT) and XGBoost. The results showed that the proposed model offered better performance than the aforementioned algorithms. Faruque, Asaduzzaman & Sarker (2019) employed four popular ML algorithms to predict diabetes mellitus, namely the SVM, NB classifier, K-Nearest Neighbor (KNN) and C4.5 decision tree. The findings from the study showed that the C 4.5 decision tree achieved higher accuracy than other ML models. Zou et al. (2018) used DT, RF and neural network to predict diabetes mellitus with hospital physical examination data in Luzhou, China. It contains 14 attributes. The results showed that prediction with random forest could reach the highest accuracy. Dey, Hossain & Rahman (2018) built a web application based on the higher prediction accuracy of some robust ML algorithms. The artificial neural network (ANN) had the highest prediction accuracy (82.35%).

The examined literature above indicates that a variety of ML models have been applied to predict diabetes among different populations with varying accuracy. In order to predict binary data, this research presented a deep learning neural network called Deep Belief Network (DBN) with Restricted Boltzmann Machine (RBM) to predict diabetes. The main objective was to apply the DBN with Rectified Linear Unit (ReLu) activation function to predict diabetes based on data from the national institute of diabetes and digestive and kidney diseases.

Data

The Diabetes dataset used in this study was obtained from the national institute of Diabetes and digestive and kidney disease. The dataset consists of 767 female patients of Pima Indian heritage and at least 21 years old. There are eight independent variables, including age, pregnancies, glucose, blood pressure, skin thickness, insulin, BMI, and diabetes pedigree function. The dependent variable was whether or not the patient had been diagnosed with diabetes, which was binary and indicated as yes or no. The independent variables were re-categorized as 1 and 0, representing yes and no, respectively.

The data was divided into two sets, training and testing, according to the ratio 8:2.

Method

Restricted Boltzmann Machine (RBM) was introduced by Chen and Murray in 2003. The structure of RBM is made up of two layers, visible and hidden layers. The visible layer consists of visible neurons, while the hidden layer consists of hidden neurons. The architecture of RBN is a complete bipartite graph shown in Fig1.

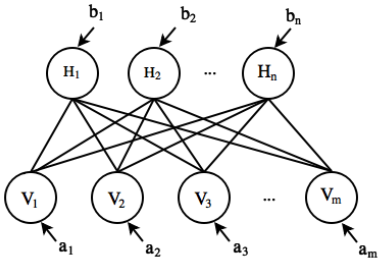

Fig1: Binary Restricted Boltzmann Machine

The visible layer (V_i) consists of m neurons, and each neuron has two values: input and output. Let c_i be the input value and $R(c_i)$ be the output value called Rectified Linear Unit (ReLU) of visible neuron, as follows:

$$c_i = \sum_{j=1}^{m} h_j w_{ij} + a_i$$

$$R(c_i) = \max\left(0, c_i\right)$$

where h_j is the output value of H_j, w_{ij} is the weight value from to V_i to H_j and a_i is the bias value of V_i.

The hidden layer (H_j) consists of n neurons, and each neuron has two values: input and output. Let d_j be the input value and $R\left(d_j\right)$ be the output value called Rectified Linear Unit (ReLU) of hidden neuron, as follows:

$$d_j = \sum_{i=1}^{n} v_i w_{ij} + b_j$$

$$R(d_j) = \max\left(0, d_j\right)$$

where v_i is the output value of V_i, w_{ij} is the weight value from H_j to V_i and b_j is the bias value of H_j.

The deep belief network (DBN) is a stack of RBM, and the structure is shown in Fig2. There are two paths to find the appropriate weight values between each layer: unsupervised and supervised paths. In the unsupervised path, each RBM uses a minimizing contrastive divergence algorithm as follow:

$$\Delta \hat{w}_{ij} = \eta_w \left\langle v_i^{(0)} h_j^{(0)} \right\rangle - \left\langle v_i^{(1)} h_j^{(1)} \right\rangle,$$

$$\Delta \hat{a}_i = \eta_a \left\langle v_i^{(0)} \right\rangle - \left\langle v_i^{(1)} \right\rangle,$$

$$\Delta \hat{b}_j = \eta_b \left\langle h_j^{(0)} \right\rangle - \left\langle h_j^{(1)} \right\rangle$$

where η_w, η_a and η_b are the learning rate for weight parameters, bias value in V_i and bias value in H_j respectively. $v_i^{(0)}$ be the output value of V_i at the initial state, $v_i^{(1)}$ be the output

value of V_i at the one-step Gibb sampling state, $h_j^{(0)}$ be the output value of H_j at the initial state, $h_j^{(1)}$ be the output value of H_j at the one-step Gibb sampling state. The best weight value in the unsupervised path will be the initial value in the supervised path. Back-propagation neuron network algorithm was used to learn the data in the supervised path.

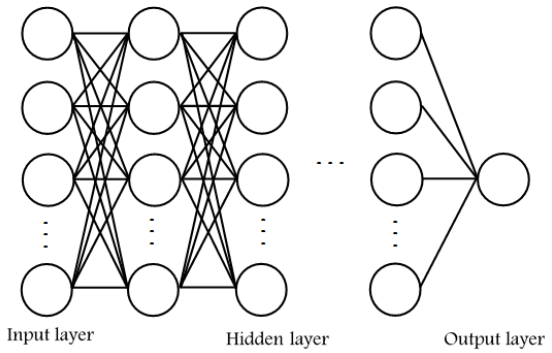

Fig2: Deep Belief Network

There is no mature method to determine the architecture of DBN. Therefore, this study used the experimentation method. This research constructed the architecture of DBN consisting of 2 RBMs stacked together. The MSE MAE and R-squared methods were used to assess the performance of the model.

Result

This research constructed the architecture of DBN consisting of 2 RBMs stacked together. The experimentation results of the first hidden layer are shown in Table 1.

Table1: Effects of first hidden layer

Input node	First hidden layer	MSE	MAE	R-squared
8	5	0.4698	0.2207	0.7792
	10	**0.4413**	**0.1948**	**0.8051**

Input node	First hidden layer	MSE	MAE	R-squared
	15	0.4558	0.2077	0.7922
	20	0.4486	0.2012	0.7987
	25	0.4629	0.2142	0.7857
	30	0.4558	0.2077	0.7922
	35	0.4558	0.2077	0.7922
	40	0.4629	0.2142	0.7857
	45	0.4486	0.2012	0.7987
	50	0.4558	0.2077	0.7922

The smallest MSE and MAE of experimentation in Table 1 occur with 10 neurons in the first hidden layer. A second hidden layer was added to the structure. The experimental results after adding a second hidden layer are shown in Table 2.

Table2: Effects of second hidden layer

Input node	Fist hidden layer	Second hidden layer	MSE	MAE	R-squared
8	10	5	0.4629	0.2142	0.7857
		10	0.4767	0.2272	0.7727
		15	0.4486	0.2012	0.7987
		20	0.4901	0.2402	0.7597
		25	0.4558	0.2077	0.7922
		30	0.4834	0.2337	0.7662
		35	0.448663	0.201299	0.798701
		40	0.455842	0.207792	0.792208
		45	0.455842	0.207792	0.792208
		50	0.469871	0.220779	0.779221

The smallest MSE and MAE of the experimentation after adding second hidden layer had higher errors than the smallest MSE and MAE of adding the first hidden layer. Therefore, the appropriate architecture of diabetes data is 8 input nodes, 10 neurons in first hidden layer and 1 output node. This architecture provides 81 percent accuracy in predicting diabetes.

Conclusion

Deep Belief Network with Restricted Boltzmann machine is a predictive method for binary data. There is no mature method to determine architecture of DBN. Therefore, this research define to construct architecture of DBN consists of 2 RBMs stacked together**,** with this experimentation provides 81 percent accuracy in predicting diabetes. Defining a variety of experiments by adding hidden layer and hidde neuron may achieve high accuracy. Furthermore, diabetes Dataset from national institute of Diabetes and digestive and kidney includes a small number of petient with a total of 767 cases.

References

Centers for Disease Control and Prevention. (2023). Retrieved April 11, 2023, from:https://www.cdc.gov/diabetes/basics/diabetes.html

Dey, S. K., Hossain, A., & Rahman, M. M. (2018, December). Implementation of a web application to predict diabetes disease: an approach using machine learning algorithm. In 2018 21st international conference of computer and information technology (ICCIT) (pp. 1-5). IEEE.

Pearson, E. R. Dissecting the Etiology of Type 2 Diabetes in the Pima Indian Population. *Diabetes* 1 December 2015; 64 (12): 3993–3995.

Faruque, M. F., & Sarker, I. H. (2019, February). Performance analysis of machine learning techniques to predict diabetes mellitus. In *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)* (pp. 1-4). IEEE.

Gadekallu, T. R., Khare, N., Bhattacharya, S., Singh, S., Maddikunta, P. K. R., & Srivastava, G. (2020). Deep neural networks to predict diabetic retinopathy. *Journal of Ambient Intelligence and Humanized Computing*, 1-14.

Sanakal, R., & Jayakumari, T. (2014). Prognosis of diabetes using data mining approach-fuzzy C means clustering and support vector machine. *International Journal of Computer Trends and Technology*, *11*(2), 94-98.

World Health Organization (2023). Retrieved April 5, 2023, from: https://www.who.int/news-room/fact-sheets/detail/diabetes

Wu, J., Diao, Y. B., Li, M. L., Fang, Y. P., & Ma, D. C. (2009). A semi-supervised learning based method: Laplacian support vector machine used in diabetes disease diagnosis. *Interdisciplinary Sciences: Computational Life Sciences*, *1*, 151-155.

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y., & Tang, H. (2018). Predicting diabetes mellitus with machine learning techniques. Frontiers in genetics, 9, 515.