RESEARCH ARTICLE

Prediction of Risk Factors Leading to Diabetes Using Neural Network Analysis

Ahed J. Alkhatib[1,2], Amer Mahmoud Sindiani[3], Eman Hussein Alshdaifat[4]

[1]Department of Legal Medicine, Toxicology and Forensic Medicine, Jordan University of Science and Technology, Jordan, [2]International Mariinskaya Academy, Department of Medicine and Critical Care, Department of Philosophy, Academician Secretary of Department of Sociology, Jordan, [3]Department of Obstetrics and Gynecology, Faculty of Medicine, Jordan University of Science and Technology, Jordan, [4]Department of Obstetrics and Gynecology, Faculty of medicine, Yarmouk University, Jordan

ABSTRACT

This study presents a prediction of diabetes type 2 using neural network analysis. A dataset of diabetes posted in Kaggle was used for analysis. The dataset included one dependent variable, outcome of disease, with two numbers, "0" means no disease, and "1" means disease. There were other eight risk factors for diabetes prediction, glucose, body mass index (BMI), insulin, skin thickness, no. of pregnancies, diabetes pedigree function, and age. We constructed the model using SPSS version 21. The results showed that the model prediction was 78.3% for training and 76.9% for testing. The model showed that the most significant predictors of diabetes were: glucose, BMI, diabetes pedigree function, no. of pregnancies, age, blood pressure, insulin, and skin thickness. Taken together, this model was effective and leads to a good prediction rate.

Key words: Covariates, dependent variable, diabetes type 2, neural network analysis, outcome

INTRODUCTION

In 2017, estimates of diabetes have shown the existence of 425 million diabetics in the world, and this number is expected to reach to 625 million by 2045 (IDF, 2017; Li *et al*., 2019).[1-3]

Diabetes mellitus is a category of endocrine disorders associated with impaired glucose absorption that occurs as a consequence of the "Insulin" hormone's absolute or relative insufficiency. Condition is characterized by a chronic path and a violation of all forms of metabolism.[4]

Diabetes is usually divided into four categories: Type 1 diabetes (T1D), type 2 diabetes, gestational diabetes mellitus, and, due to other factors, particular types of diabetes. T1D and T2D are the two most common forms of the condition (T2D). The former is caused by the degradation of the beta cells of the pancreas, which results in insulin deficiency, whereas the latter is caused by the inadequate transport of insulin to the cells.[5] Life-threatening complications such as strokes, heart attacks, chronic renal failure, diabetic foot syndrome, antipathy, neuropathy, encephalopathy, hyperthyroidism, tumors of the adrenal gland, liver cirrhosis, glucagonoma, and transient hyperglycemia and many other complications can contribute to both forms of disease.[4] For all people who are predisposed to diabetes, the prediction and early detection of diabetes are therefore important. Using artificial intelligence techniques, multiple diseases can currently be diagnosed, and deep neural networks have achieved the best results in classification problems (Miotto etal., 2017; Liu *et al*., 2019).[3,6,7]

Data mining was used to predict diabetes.[5,8,9] Ali *et al*. used neural network analytics to build a model for classifying diabetic people based on past medical history and patient control level. The author developed a new predictive model for data mining methods that would identify the level of control of diabetic patients centered on medical reports that are historical. The analysis was carried out with the use of three techniques

Address for correspondence:
Ahed J. Alkhatib, Department of Legal Medicine, Toxicology and Forensic Medicine, Jordan University of Science and Technology, Jordan. Mobile: 00962795905145.

of data processing, which are Logistic, Naïve Bayes, and J48. The analysis was carried out using the WEKA program. The outcome showed that logistics data mining algorithm gave an average accuracy of 0.73, a recall of 0.744, a metric of 0.653, and a precision of 74.4%. Naïve Bayes gave an average accuracy of 0.717, 0.742 recall, 0.653 F-measure, and 74.2% precision. J48 has given an average accuracy of 0.54, recall of 0.735, F-measure of 0.623, and precision of 0.54 and 73.5 percentage points. This showed that the logistic algorithm was more precise than the other two algorithms.

Study objectives

The main objective of this study was to identify the risk factors associated with diabetes and to determine their relative importance.

METHODOLOGY

The present study was based on a dataset posted in Kaggle.[3] The dataset of diabetes taken from India was analyzed in this study using a neural network. The data consisted of 768 females.

Study variables

The dataset involved one dependent variable, outcome indicating having diabetes (1) or no diabetes (0). There are 8 covariates or independent variable: Glucose, insulin, age, blood pressure, skin thickness, no. of pregnancies, and diabetes pedigree function.

Statistical analysis

Neural network analytics of data was carried out using SPSS version 21.

RESULTS

As demonstrated in Table 1, the study sample included 768 cases, 526 (68.5%) in training section, and 242 (31.5%) in testing section. All cases were included in the study.

Network information

As demonstrated in Table 2, network information included the following parameters: Input layer with 8 covariates: Pregnancies number, glucose, blood pressure, skin thickness, insulin, body mass index (BMI), diabetes pedigree function, and age. Number of units is 8 and rescaling method for covariates is standardized. The second parameter was about the hidden layer (s), there was one hidden layer with five units in hidden layer, and the activation function was hyperbolic tangent. The third parameter was the output layer. There was one variable, the output. It has 2 number of units. Activation function was softmax and the error function was cross-entropy.

Model summary

We constructed the model that had the following characteristics [Table 3]: For training, cross-entropy error was computed as

Table 1: Case processing summary

	n	Percent
Sample		
Training	526	68.5
Testing	242	31.5
Valid	768	100.0
Excluded	0	
Total	768	

Table 2: Network information

Input layer	Covariates	1	Pregnancies No.
		2	Glucose
		3	Blood pressure
		4	Skin thickness
		5	Insulin
		6	BMI
		7	Diabetes pedigree function
		8	Age
	Number of units[a]		8
	Rescaling method for covariates		Standardized
Hidden layer(s)	Number of hidden layers		1
	Number of units in hidden layer 1[a]		5
	Activation function		Hyperbolic tangent
Output layer	Dependent variables	1	outcome
	Number of units		2
	Activation function		Soft max
	Error function		Cross-entropy

BMI: Body mass index

Table 3: Model summary

Training	**???**
Cross-entropy error	241.342
Percent incorrect predictions	21.7
Stopping rule used	1 consecutive step(s) with no decrease in error[a]
Training time	0:00:00.16
Testing	
Cross-entropy error	111.363
Percent incorrect predictions	23.1

241.342, percent incorrect prediction 21.7%, stopping rules used were 1 consecutive step(s) with no decrease in error[a], and training time was 0:000:00.16. For testing, cross-entropy error was 111.363 and percent incorrect prediction was 23.1%. Dependent variable was outcome.

Classification

As seen in Table 4, in this study, the outcome variable implied two values, "0": Non-diabetic, and "1": Diabetic. In the training part, for 0 output, the system determined 345 cases, 47 cases were classified by mistake in "0" output, but the system put them in "1" output. Percent correction was computed as 86.4%. In "1" output, 181 cases were included, among which 67 cases were predicted to be in "0" output. Percent correction was computed as 63%. The overall percent was 78.3%. In the testing part, 155 cases were included in the output "0," among which 23 cases were predicted to have the disease. The percent correction was 85.2%. For the output "1", 87 cases were included, of which 33 cases were predicted to be in the output "0." Percent correction was found to be 62.1%. The overall percent was 76.9%.

Table 4: Classification

Sample	Observed	Predicted		
		0.00	1.00	Percent correct
Training	0.00	298	47	86.4
	1.00	67	114	63.0
	Overall percent	69.4	30.6	78.3
Testing	0.00	132	23	85.2
	1.00	33	54	62.1
	Overall percent	68.2	31.8	76.9

Dependent variable: Outcome

Independent variable importance

As seen in Table 5 and Figure 1, the importance of glucose level was 0.276 (100%), BMI 0.244 (88.4%), diabetes pedigree function 0.145 (52.6%), no. of pregnancy 0.101 (36.5%), age 0.08 (29.1%), blood pressure 0.071 (25.7%), insulin 0.052 (18.8%), and skin thickness 0.030 (10.7%).

DISCUSSION

The results of this study identified diabetes risk factors among females who had diabetes type 2. Using neural network analysis usually uses technical language that is not easily understood by readers. The problem of these studies is being made by non-medical professional, including those working in computer engineering or software engineers. According to this context, the author integrated technical and medical languages to help the reader to be able understand the topic of diabetes.

According to our model, covariates or independent variables, glucose level was the first important predictor of diabetes type 2 among females, and its importance 100%. BMI was able to predict 88.4% of diabetes cases [Figure 1]. The importance of such a figure is that it pointed to points that we need to control. Another important point is the number

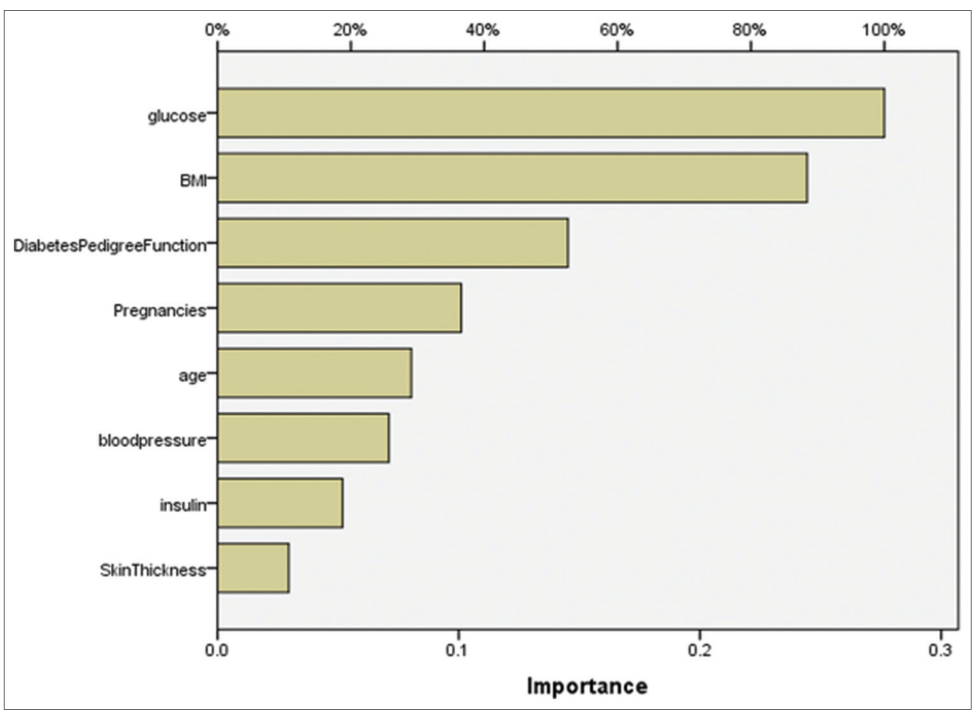

Figure 1: Normalized importance of covariates

Table 5: Independent variable importance

Independent variable	Importance	Normalized importance %
No. of pregnancies	0.101	36.5
Glucose level	0.276	100.0
Blood pressure	0.071	25.7
Skin thickness	0.030	10.7
Insulin	0.052	18.8
BMI	0.244	88.4
Diabetes pedigree function	0.145	52.6
Age	0.080	29.1

BMI: Body mass index

of pregnancies that predicted about 37% of diabetic cases among females.

CONCLUSIONS

The results of the present study pointed to two important considerations, the use of neutral network analytics is effective in the prediction of diabetes and identifying risk factors with their relative importance.

REFERENCES

1. Zhou H, Myrzashova R, Zheng R. Diabetes prediction model based on an enhanced deep neural network. EURASIP J Wirel Commun Netw 2020;148:1621.
2. Li G, Peng S, Wang C, Niu J, Yuan Y. An energy-efficient data collection scheme using denoising autoencoder in wireless sensor networks. Tsinghua Sci Technol 2019;24:86-96.
3. Liu H, Kou H, Yan C, Qi L. Link prediction in paper citation network to construct paper correlated graph. EURASIP J Wirel Commun Netw 2019;233:2352.
4. Available from: https://www.kaggle.com/uciml/pima-indians-diabetes-database. [Last accessed on 2020 Dec 10].
5. American Diabetes Association. Classification and diagnosis of diabetes: Standards of medical care in diabetes-2018. Diabetes Care 2018;41:13-27.
6. Ali R, Siddiqi MH, Idris M, Kang BH, Lee S. Prediction of diabetes mellitus based on boosting ensemble modeling. In: International Conference on Ubiquitous Computing and Ambient Intelligence. Berlin, Germany: Springer International Publishing; 2014. p. 25-8.
7. Miotto R, Wang F, Wang S, Jiang X, Dudley T. Deep learning for healthcare: Review, opportunities and challenges. Brief Bioinform 2017;19:1236-46.
8. International Diabetes Federation. IDF Diabetes Atlas. 8th ed. Brussels, Belgium: International Diabetes Federation; 2017.
9. El_Jerjawi NS, Abu-Naser SS. Diabetes prediction using artificial neural network. Int J Adv Sci Technol 2018;121:55-64.

How to cite this article: Alkhatib AJ, Sindiani AM, Alshdaifat EH. Prediction of Risk Factors Leading to Diabetes Using Neural Network Analysis. Clin Res Diabetes Endocrinol 2020;3(2):1-4.