Application of Artificial Neural Networks Models in Diabetes Mellitus Classification

Fatima Zohra HADDAD
Ecole Nationale Supérieure de Statistique et d'Economie Appliquée, Kolea, Tipaza, Algeria
haddad.fatima@enssea.net.

Mohammed Amine BENBOURAS
Laboratory of Materials in Civil Engineering and Environment, Ecole Nationale Polytechnique, El Harrach, Algiers, Algeria, Laboratoire Central des Travaux Publics
mouhamed_amine.benbouras@g.enp.edu.dz.

Abstract— Diabetes mellitus is one of the most worrying chronic diseases. It results from disturbances in blood sugar levels causing many complications and sometimes leads to death. International Diabetes Federation has stated that the number of diabetics is rising and could up to 642 million in 2040. The rapid development of information technology has imposed new advanced methods entitled "learning machine" or "artificial intelligence techniques" which have led to impressive results in the medical field. Based on this background, this study contributes to classify diabetes, by the mean applied of artificial neural networks 'ANN' method. The idea is based on the application of ANN on the pima Indian diabetes database on 3 cases, according to the number and the type of selected features "attributes". The results show high accuracy of the model with all attributes (92.3%) and without Diastolic blood pressure (92.6%). The proposed ANN model composed by two hidden layers ensures better predictability in data learning and yields data prediction values better than the ones published in previous studies.

Index Terms— Diabetes Mellitus, Artificial Neural Networks (ANN), Pima Indian Diabetes (PID), Classification.

I. INTRODUCTION

Chronic diseases are dangerous maladies that persist for a long time and develop slowly, and are several types of them heart diseases, cancer and diabetes mellitus and this latter is the most widespread. Diabetes is a very known disease that comes from glucose disorders in the human body causing a permanent change in the person's internal chemistry and resulting in a significant increase in blood glucose levels due to lack of insulin hormone. The latter is produced by special cells in the pancreas called beta-β cells in the bloodstream. Its function is to reduce the level of glucose in the blood [1]. There are several types of diabetes, most notably; Diabetes Type 1 (T1D) is a chronic disease characterized by the complete destruction of beta-pancreatic (Beta) cells, which produces insulin leading to a complete cessation of insulin production. Diabetes Type 2 (T2D) is an insulin-dependent diabetes caused by a partial lack of insulin, which makes it insufficient to control the normal level of glucose, and it often affects the adults. Another type is known as Gestational Diabetes is hyperglycemia with blood glucose values above normal it affects women during pregnancy, this is type can remain after pregnancy and is considered a type 2 of diabetes [2].

Many people in the world suffer from diabetes whatever their different races and ages are. The number of people Infected by diabetes has increased recently according to the World Health Organization (WHO), which reported an increase of the number of diabetics from 108 million in 1980 to 422 million in 2014, and in 2017 reached 425 million diabetic patient people in the worldwide. Other statistics indicated that one adult from two adults with diabetes are undiagnosed (over 212 million), according to the International Diabetes Federation (IDF). This increase is due to several causes by type of diabetes (T1D and T2D) and many others factors represented at age (> 45 years), obesity or overweight (BMI >=25 kg/m2), sedentarily and familial antecedent. Based on this background, artificial intelligence techniques represent an impressive idea for modeling the important factors to classify individuals of diabetic and Not Diabetic. On the other hand, it has seen the use of the ANN method in classifying diabetes has witnessed a minor development, which allows simulating the capabilities of human thought the formulation of advanced programs based on statistical data and information for a group of individuals patients and non-patients.

Many previous studies have been carried out in order to search for a classification model to diabetes based on Pima Indian diabetes database (PID). Numerous algorithms in the previews studies, study titled "Design of fuzzy classifier for diabetes disease using Modified Artificial Bee Colony algorithm", for the Fayssal BELAIFA et M.A CHIKH in 2013 [3] where proposed a novel Artificial Bee Colony (ABC) and its performance are evaluated through classification rate sensitivity and specificity values using 10-fold cross-validation method and the obtained classification rate of is 84.21%. Several researchers have also relied on other methods as Khyati K.GANDHI et Nilesh B.PRAJAPATI in 2014 [4] presenting a study entitled "Diabetes prediction using feature selection and classification", where F-score method and K-means clustering was used for feature selection and the performance of the SVM classifier is empirically was evaluated on the reduced feature subset of Pima Indian diabetes dataset used for testing data mining algorithms to see their prediction accuracy in diabetes data classification. Mehmet Recep BOZKURT et al [5] in 2014, they presented a study entitled Comparison of different methods for determining diabetes, (Rabindra Nath DAS) [6] in 2014 presented "Determinants of Diabetes Mellitus in the Pima Indian Mothers and Indian Medical Students", This article aims to identify the determinants of diabetes mellitus in the Pima Indian mothers and Indian young medical students. (Mukesh KUMARI et al) [7] in 2014, "Prediction of Diabetes Using Bayesian Network" This paper helps to predict the persons whether diabetic or not by applying technique Bayesian Network classifier. Others studies have used K-means, genetic algorithm and support vector machine by using J.48 and decision tree for diabetes diagnosis, for researchers T.SANTHANAM et M.S PADMAVATHI, in 2015 [8], In this article, K-Means is used for deleting the corrupt data, also used genetic algorithms for finding the optimal set of features with Support Vector Machine (SVM) as for classification. Another study of Sun JIAN, in 2016 [9], This study relied Pima Indian diabetes dataset with the aim of Predict the probability that individual females have diabetes in GLM. D.K Ghoubey et al, in 2015, 2016 and 2017 [10-11-12], they presented research's for aim the Classification diabetes, by many application j48 graft decision tree(j48 GDT), multilayer perceptron neural network (MLP NN), naive bayes (NBs) and reapplication all this methods with genetic algorithm (GA_ j48 GDT, GA_ MLP NN and GA_NBs). Han WU et al, in 2018 [13], proposed a novel model based on data mining techniques for predicting type 2 diabetes mellitus, Based on the improved K-means algorithm and the logistic regression algorithm and Pima Indians Diabetes Dataset. Abir ALHARBI et Munirah ALGHAHTANI, in 2018 [14], proposed An system for diagnosis of type 2-diabetes, by using the Extreme Learning Machine neural network for classification and the evolutionary genetic algorithms, and K. Gholipour et al, in 2018[15], it came with a goal to compare the power of an artificial neural network and logistic regression in identifying type 2 diabetes mellitus risk factors, and study for N.S. El-Jerjawi et S.S. Abu-Naser, in 2018[16], entitled diabetes prediction using artificial neural network where relied on ANN with 3 hidden layers. There are also many other studies that have focused on the classification of diabetes based on others databases than Pima Indian Diabetes. Other studies have stated that ANN method could give superior results. However, the previous studies that used ANN for classifying diabetes modeling have been impaired by some shortcomings. The observed criticism is that they have built a model using few input parameters and, therefore, they have ignored the different parameters that could increase the learning capacity of the network. Moreover, they have used few samples and one hidden layer ANN architecture; although many researchers have stated that, the use of two hidden layers could offer the flexibility required for modeling complex function [17-18]. To improve the predictive capability in classifying the modeling diabetes; the current study aims to propose an alternative approach based on the building of an accurate artificial intelligent model using tow hidden layers.

II. Materials and Methods

In this study, the model of artificial neural networks was applied to distinguish between diabetic patients and non-patients, it will be used a dataset entitled “the database Pima Indian Diabetes (PID)” provided by the UCI Machine Learning Repository [19] (famous repository for machine learning data sets). MATLAB statistical program has been used due to its satisfactory revealed in previous studies. An overview of the ANN method and the database PID will be given below:

A.Definition of Database

The Pima Indian Diabetes women database has been relied upon, which consists of 768 cases, between them 268 positive for diabetes (diabetic) and 500 negative for diabetes (non-diabetic), where it consists of 8 inputs “Attributes or Features”:

- NTP : Number of times pregnant;
- PGC: Plasma glucose concentration;
- DBP: Diastolic blood pressure (mm Hg);
- TST: Triceps skin fold thickness (mm);
- HSI: 2 Hours serum insulin (mu U/ml);
- BMI: Body mass index (Kg/m2);
- DPF: Diabetes predigree function;
- Age (years).

The PID sample was taken after deleting all cases with the missing values, relying on only 392 cases, with the application of neural networks model to the database with 8 attributes and re-application without diastolic blood pressure “this latter was according to the researcher’s perspective where he considered it as a quantitative variable associated with the psychological factors of human at the moment of measurement” and re-application ANN again with only 5 attributes. After deleting both number of times pregnant, plasma glucose concentration and diastolic blood pressure, and rely on Triceps skin fold thickness, 2 Hours serum insulin, Body mass index, Diabetes predigree function and age, this is according to the most important factors causing diabetes of study presented by Rabindra Nath DAS in the year 2014 [6].

B.Artificiel Neural Networks

Artificial neural networks are information processing systems that simulate biological neural systems. It consists of a large number of neurons connected to each other in a complex network “nodes”, and The ANN is divided into layers: input layer, output layer and intermediate layers known as hidden layers [20], as illustrated in the following figure:

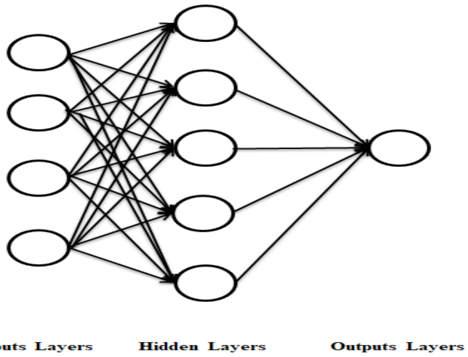


Fig.1. Neural Network

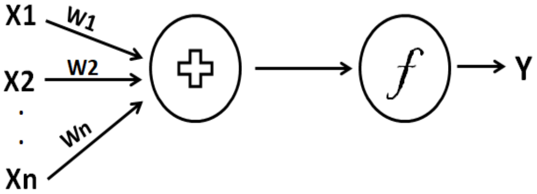


Fig.2. Node

Where: $$Y = \mathbf{f}\ (net) \tag{1}$$

$$net = W^{t*}X \tag{2}$$

$\mathbf{f}$ (net): is Activation function

In the field of artificial neural networks, the activation function “transformation function” is a mathematical function applied to a signal outputs from an artificial neuron. There are many types of neural network activation function, whereas this study was based on three activation functions “Linear Sigmoid and Hyperbolic Tangent” (see TABLE 01). It because that they are the most according to previous studies as they give good results.

TABLE I. EXAMPLE OF TRANSFER FUNCTION

Function Name	Inputs-Outputs Relationship	Curve	Name in Matlab
Linear	Y=X		Purelin
Sigmoid	$Y = \frac{1}{1+e^{-X}}$		Logsig
Hyperbolic tangent	$Y = \frac{e^{X}-e^{-X}}{e^{X}+e^{-X}}$		Tansig

Source: PhD thesis [21]

C.Training and testing

The PID database was divided into 70% for training, 15% for validation and 15% for testing. They are relying on the confusion matrix to evaluate model performance by calculating both Correlation Coefficient (Rall), Mean Squared Error (MSE) sensitivity, Specificity, Precision, Accuracy and F1-Score testing and finally comparing most of these results for both models with 8 attributes, 7 attributes and 5 attributes As well as compared with the results of the previous study which relied on artificial neural networks with a single hidden layer.

1.A confusion matrix: is a technique for summarizing the classification performance of a classifier with respect to some test data. It is a two-dimensional matrix, a confusion matrix, also known as an error matrix, it is a table that allows to measuring the performance of an algorithm, by knowing whether the model is well classified. Where each row in the matrix represents the predicted class and each column represents the actual class [22].

TABLE II. CONFUSION MATRIX

		Actual class	
		Positive	Negative
Predicted class	Positive	TP	FP
	Negative	FN	TN

Where:

TP: True Positive, for correctly predicted event values.

FP: False Positive, for incorrectly predicted event values.

TN: True Negative, for correctly predicted no-event values.

FN: False Negative, for incorrectly predicted no-event values.

and Both TP, FP, TN and FN are called roc value.

2. Performance Measures: The performances of the implemented classifier was evaluated by computing the percentages of sensitivity (Se), specificity (Sp) precision (Pr), accuracy (Ac) and F1-Score, the respective formula of all are as follows:

$$Se\% = \frac{TP}{TP+FN} * 100 \qquad (03)$$

$$Sp\% = \frac{TN}{TN+FP} * 100 \qquad (04)$$

$$Pr\% = \frac{TP}{TP+FP} * 100 \qquad (05)$$

$$Ac\% = \frac{TN+TP}{TN+TP+FP+FN} * 100 \qquad (06)$$

$$F1-Score = \frac{2*Se*Pr}{Se+Pr} \qquad (07)$$

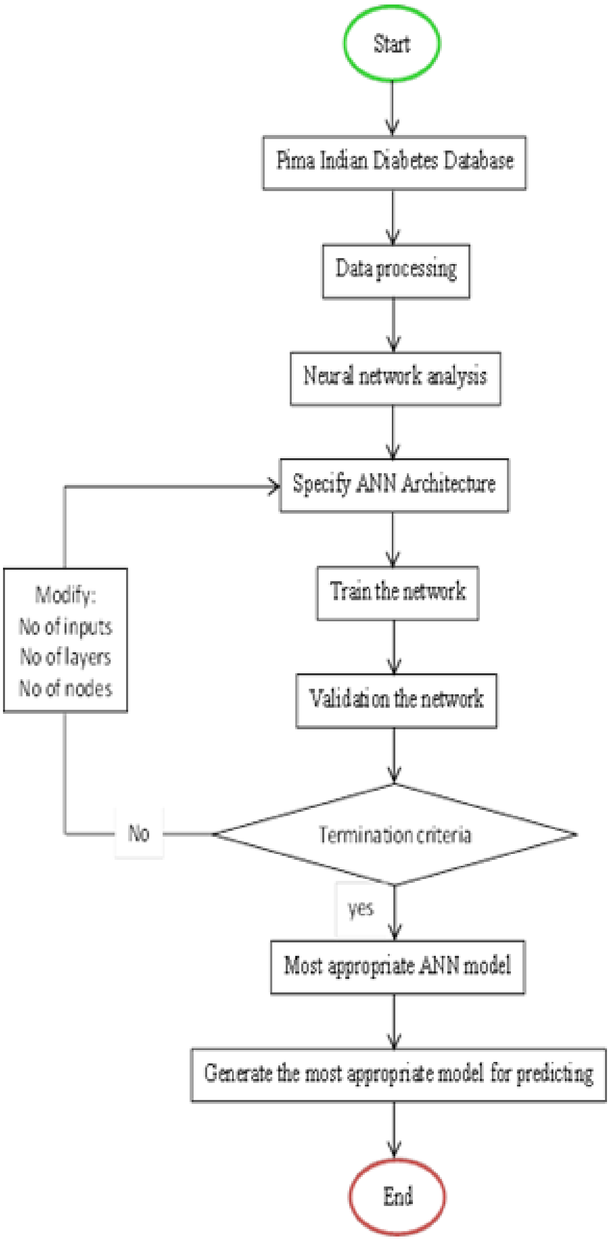


Fig.3. Flowchart of key steps for the research methodology.

III. RESULTS AND DISCUSSION

A. Generation of the database

The database used in this study consisted of sample with

392 individuals (262 Diabetic and 130 Non-Diabetic), with 8 quantitative variables also known as attributes or features obtained from the UCI Machine Learning Repository [19].

According to SPSS treatment, table III and IV shows the descriptive statistics of study variables, such as: minimums maximums, means and standard deviations. All variables (NTP, PGC, DBP, TST, BMI, DPF and age) are homogenies distribution except for the distribution HIS is dispersed somewhat. As for the skewness and kurtosis values show that most of variables are regularly distributed, the results also indicate that the database comprises a wide range of data. Subsequently, this database can be used to develop new empirical equations allows the Discrimination between diabetics and non-diabetics and compare the performance of existing formulae between them.

TABLE III. Descriptive statistics for the database

	N	Minimum	Maximum	Mean	Std. Deviation
NTP	392	0	17	3.30	3.211
PGC	392	56	198	122.63	30.861
DBP	392	24	110	70.66	12.496
TST	392	7	63	29.15	10.516
HSI	392	14	846	156.06	118.842
BMI	392	18.2	67.1	33.086	7.0277
DPF	392	.085	2.420	.52305	.345488
Age	392	21	81	30.86	10.201

TABLE IV. Skewness and kurtosis parameters

	Skewness		Kurtosis	
	Statistic	Std. Error	Statistic	Std. Error
NTP	1.336	.123	1.486	.246
PGC	.518	.123	-.483	.246
DBP	-.088	.123	.795	.246
TST	.209	.123	-.458	.246
HSI	2.165	.123	6.357	.246
BMI	.663	.123	1.557	.246
DPF	1.959	.123	6.367	.246
Age	1.404	.123	1.738	.246

The matrix of correlation between diabetes variables presented in Table V, provides an overview of spearman R correlation and its significance. The scatter plot matrix of studied parameters is presented in Fig.5 in order to provide a descriptive overview of the data distribution. The findings imply the existence of a fairly moderate positive relationship between variable outcome (Diabetics and Non-Diabetics) and other variables. The latter means that an increase in the first parameters tends to proportionally increase variable outcome. Furthermore, Table V shows that the significance of the regression coefficient between variable outcome and other diabetics parameters in the significant levels 1% and 5%, meaning that the correlations are statistically significant. These results indicate that the previous parameters are supposed to have a complex nonlinear correlation with outcome. Based on that, in order to reliably simulate the complex relationships between variable outcome and other diabetics parameters, new artificial intelligent approaches should be used as ANN method.

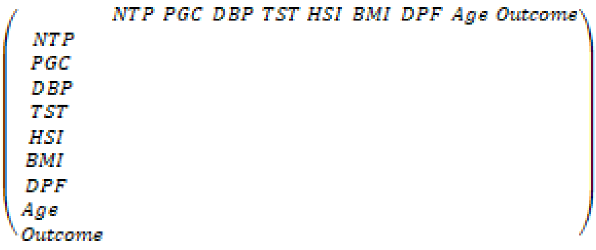


Fig.4. Correlation matrix

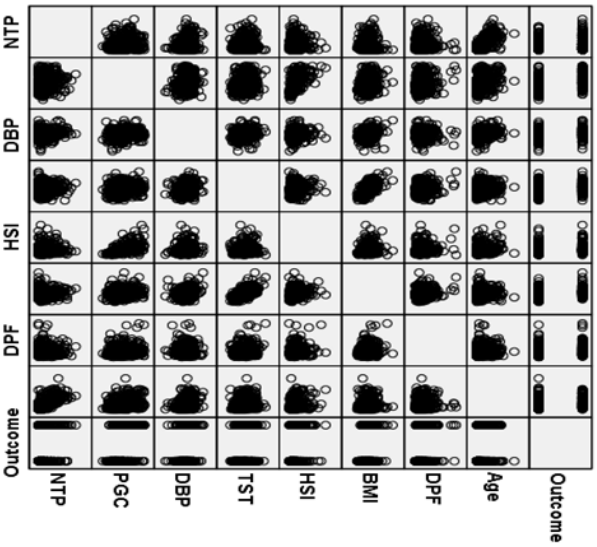


Fig.5. Scatterplot matrix of the diabetics' parameters

TABLE V. Matrix of correlation between determinants of diabetes

		NTP	PGC	DBP	TST	HSI	BMI	DPF	Age	Outcome
NTP	Correlation Coefficient	1.000	.190**	.152**	.055	.123*	-.066	.012	.634**	.200**
	Sig. (2-tailed)	.	.000	.002	.279	.015	.195	.817	.000	.000
PGC	Correlation Coefficient	.190**	1.000	.237**	.216**	.659**	.199**	.089	.350**	.499**
	Sig. (2-tailed)	.000	.	.000	.000	.000	.000	.077	.000	.000
DBP	Correlation Coefficient	.152**	.237**	1.000	.250**	.132**	.317**	-.021	.329**	.198**
	Sig. (2-tailed)	.002	.000	.	.000	.009	.000	.680	.000	.000
TST	Correlation Coefficient	.055	.216**	.250**	1.000	.241**	.674**	.093	.242**	.260**
	Sig. (2-tailed)	.279	.000	.000	.	.000	.000	.066	.000	.000
HSI	Correlation Coefficient	.123*	.659**	.132**	.241**	1.000	.301**	.132**	.261**	.375**
	Sig. (2-tailed)	.015	.000	.009	.000	.	.000	.009	.000	.000
BMI	Correlation Coefficient	-.066	.199**	.317**	.674**	.301**	1.000	.096	.167**	.267**
	Sig. (2-tailed)	.195	.000	.000	.000	.000	.	.057	.001	.000
DPF	Correlation Coefficient	.012	.089	-.021	.093	.132**	.096	1.000	.103*	.198**
	Sig. (2-tailed)	.817	.077	.680	.066	.009	.057	.	.042	.000
Age	Correlation Coefficient	.634**	.350**	.329**	.242**	.261**	.167**	.103*	1.000	.397**
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.001	.042	.	.000
Outcome	Correlation Coefficient	.200**	.499**	.198**	.260**	.375**	.267**	.198**	.397**	1.000
	Sig. (2-tailed)	.000	.000	.000	.000	.000	.000	.000	.000	.

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

B.Evaluation of outcome parameter using ANN

The appropriate model is determined based on ANN analysis by select the appropriate input parameters. However, the optimal number of nodes in selected

by evaluating the outcome based on the minimum MSE and maximum Rall in each hidden layer. Furthermore, the ANN has been applied several times according to the number of hidden layers and the number of nodes in each layer with the three cases under study. The best model was chosen based on several factors such as Correlation Coefficient (Rall), Mean Squared Error (MSE) and Accuracy; this is illustrated in TABLE VI and Fig.6 below:

TABLE VI. Three different feature subsets using ANN

Attributes selected	8 Attributes	7Attributes	5 Attributes
Rall	0.83972	0.82882	0.74832
MSE	0.06756	0.07068	0.10143
Nodes Number	[13, 2]	[13, 2]	[20, 19]
Transformation function	Logsig Tansig	Logsig Tansig	Logsig Tansig
Sensitivity	96.6%	94.7%	89.3%
Specificity	83.8%	88.5%	61.5%
Precision	92.3%	94.3%	82.4%
Accuracy	92.3%	92,6%	80.1%
F1-Score	96.6%	94.5%	85.7%

Source: Prepared by the researcher according to MATLAB outputs

TABLE VII. comparison with other methods

Authors	Year	Methods	Accuracy
F.Beloufa and M.A.Chikh	2013	ABC Algorithm	84.21%
M.R.Bozkurt et al	2014	DTDN	76%
D. K.Ghoubey et al	2015	J48 GDT	76.52%
		GA_j48 GDT	74.78%
	2016	MLP NN	78.26%
		GA_MLP NN	79.13%
	2017	NBs	76.96%
		GA_NBs	78.70%
K. Gholipour et al	2018	ANN	83.9%
		Logistique regression	85.4%
N. El-Jerjawi and S. Abu-Naser	2018	ANN	87.3%
Our study		ANN	92.6%

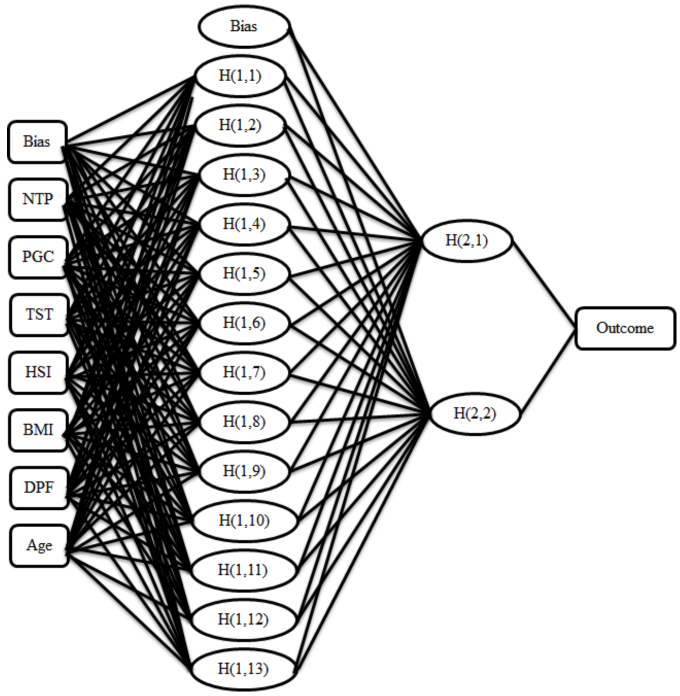


Fig.6. The architecture of the most appropriate ANN model (7-13-2-1).

In this study, the ANN method were applied and relied on two hidden layers with [13-02] nodes and the PID database were used in the classification of diabetes. After applying ANN in three cases with all the features “8 Attributes” without diastolic blood pressure “7 Attributes” and 5 features a number of the results were reached in TABLE VI, this is according to MATLAB outputs of Fig.5 Fig.6, Fig.7, Fig.8 Fig.9 and Fig.10. Where these results show both Correlation Coefficient (Rall), Mean Squared Error (MSE) and as well as various accuracy factors that can be relied upon to determine the best model and how his significant, according to the results shown in Table VI. The findings indicate that in both cases, with 8 attributes and 7 attributes, show model significant, accuracy and clarity where it was Correlation

Coefficient with 8 attributes is approx. (Rall = 84%) and with 7 attributes is about (Rall = 83%), mean squared error with 8 attributes is about (MSE = 7%) and with 7 attributes is about (MSE = 7%). The results indicate that the model is accepted and considered significant and well classifier for diabetes, it also shows the most important coefficient of accuracy with 8

attributes is about (Ac = 92%) and with 7 attributes approx. (Ac = 93%). Both models are accurate, although there are small variation in some accuracy ratios thus, the diastolic blood pressure feature can be dispensed with. The comparison of the results of this study with the others studies published this is according to the accuracy of the models that were used in different previous studies (TABLE VII), as F.Beloufa with ABC Algorithme application (84.21%) M.R.Bozkurt with DTDN application (76%), and the various methods that D. K.Ghoubey relied on in his studies in 2015, 2016 and 2017, Where was the best result with MLP 79.13% accuracy, study of K. Gholipour, where he relied on two applications: Logistique regression (85.4%) and ANN , where he relied in this latter (ANN) on one hidden layer with 83.9% accuracy and the last study of N. el-jerjawi and S. abu-naser relied ANN with 3 hidden layers. it turns out that relying on two hidden layers (this is illustrated in Fig.6 above), improves the accuracy of the model somewhat compared to the accuracy of the previous study there is no need to resort to 3 hidden layers and the complexity of the studied phenomenon, as it is two hidden layers that give us good results, this is what was relied upon in our study.

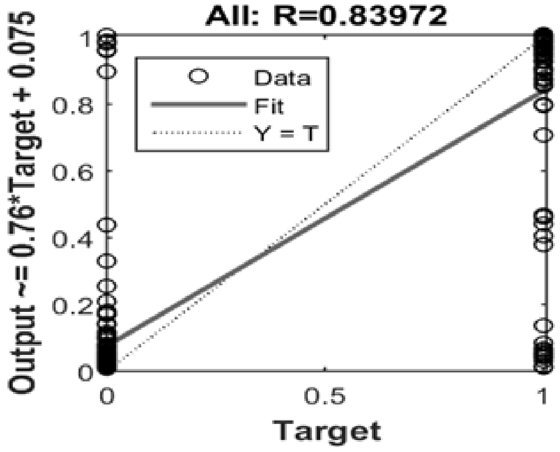


Fig.7. Plot-Regression with 8 Attributes

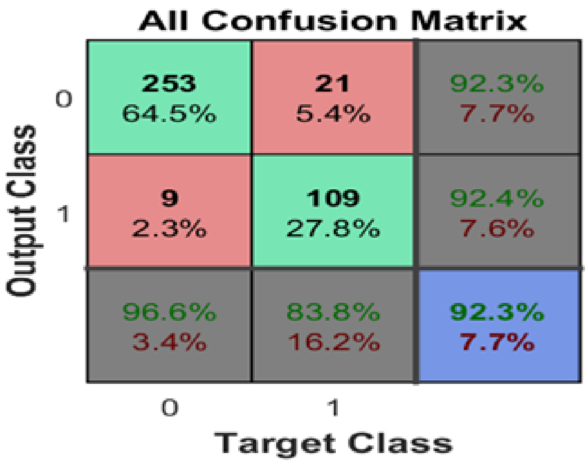


Fig.8. Plot-Confusion with 8 Attributes

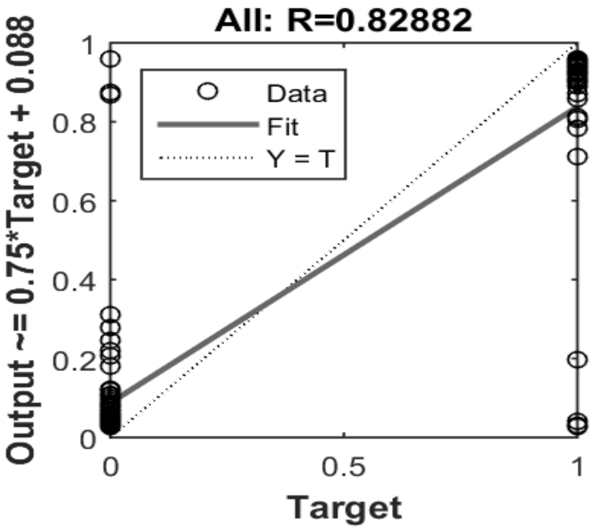
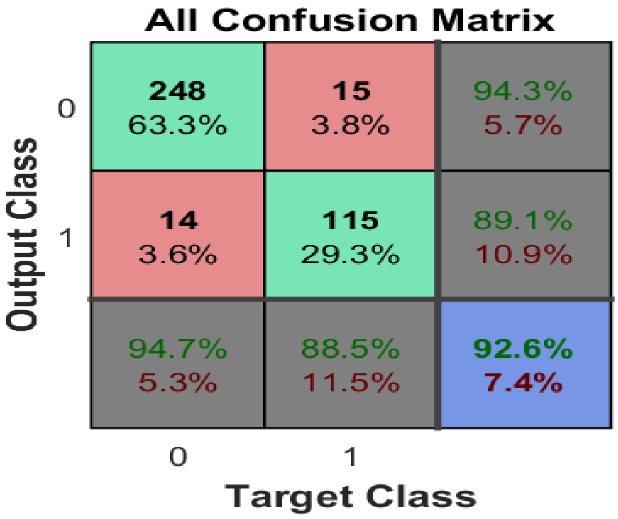


Fig.9. Plot-Regression with 7 Attributes



	0	1	
0	248 63.3%	15 3.8%	94.3% 5.7%
1	14 3.6%	115 29.3%	89.1% 10.9%
	94.7% 5.3%	88.5% 11.5%	92.6% 7.4%

Fig.10. Plot-Confusion with 7Attributes

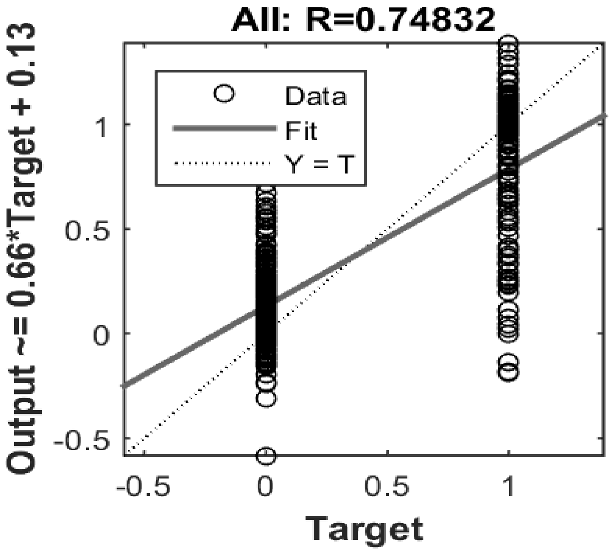
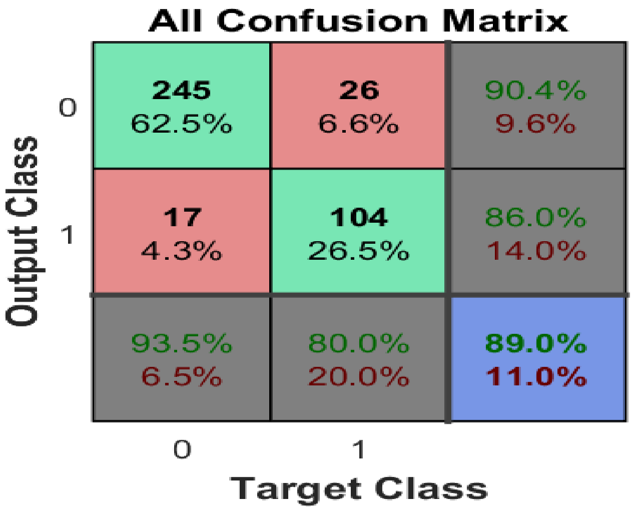


Fig.11. Fig.11: Plot-Regression with 5 Attributes



	0	1	
0	245 62.5%	26 6.6%	90.4% 9.6%
1	17 4.3%	104 26.5%	86.0% 14.0%
	93.5% 6.5%	80.0% 20.0%	89.0% 11.0%

Fig.12. Plot-Confusion with 5 Attributes

IV. CONCLUSION

Diabetes is one of the most prevalent diseases in the world, targeting all age groups. The number of people with diabetes has increased in the recent period, which reached ٤٢٥ million in ٢٠١٧. Despite the high prevalence of diabetes, it remains a non-fatal disease, but if it is neglected and not controlled for a long time, it will result in many complications that pose a threat to the lives of individuals and communities in general. These complications may lead to the risk of premature death, as well as serious injuries difficult to treat, such as loss of vision due to damage to the retina; it may lead to kidney failure and damage to the nerves, loss of sensation, heart attack, stroke and amputation of the leg. For these reasons, our idea is to model the disease using artificial intelligence methods.

The current study contributes to propose an alternative model based on the application of ANN with two layers hidden for ٣ cases with depending on all the features «٨ attributes» and then without diastolic blood pressure «٧ features» and finally with ٥ features.

After applying the neural network model, given the accuracy of the model in all three cases, it was found that the diastolic blood pressure can be dispensed with, because its presence or non-existence does not affect the accuracy of the model and deleting it is better, As for the structure of neural networks, relying on two hidden layers gives more accurate and effective results, compared to previous studies that relied on one hidden layer in its structure, it is advisable in the case of complex databases rely on two hidden layers or more. This work has opened up several questions that need of further investigations in the future studies about the modeling of chronic diseases and we recommend designing a graphical interface for easily using the proposed model in the future Medical examination.

REFERENCES

[1] C. E. Mogensen, "Pharmacotherapy of diabetes. In: Pharmacotherapy of Diabetes: New Developments,". Springer Boston MA, 2007.

[2] World health organization: Diabetes,". oct.30, 2018. Accessed on: July. 01, 2019.[online]. Available: https://www.who.int/news-room/fact-sheets/detail/diabetes.

[3] F. Beloufa and M. A. Chikh, "Design of fuzzy classifier for diabetes disease using Modified Artificial Bee Colony algorithm," Computer methods and programs in biomedicine, vol. 112, no.1, pp. 92-103, 2013.

[4] K. K. GANDHI and N. B. Prajapati, "Diabetes prediction using feature selection and classification," International journal of advance Engineering and Research Development, vol. 1, no.05, 2014.

[5] M. R. Bozkurt, N. Yurtay, Z. Yilmaz, and C. Sertkama, "Comparison of different methods for determining diabetes," Turkish Journal of Electrical Engineering & Computer Sciences, vol. 22, no. 4, pp. 1044-1055, 2014.

[6] R. N. Das, "Determinants of diabetes melitus in Pima Indian Mothers and Indian Medical students," Diabetes Journal, vol. 7, pp 5-13, 2014.

[7] M. Kumari, R. Vohra, and A. Arora "Prediction of diabetes using Bayesian network," International Journal of Computer Science and Information Technologies, Vol. 5, no.4, pp. 5174-5178, 2014.

[8] T. Santhanam and M. S. Padmavathi, "Application of K-means and genetic algorithms for dimension reduction by integrating SVM for diabetes diagnosis," Procedia Computer Science, vol. 47, pp. 76-83, 2015.

[9] S. Jian, "The Study of Pima Indian Diabetes," oct, 2016. Accessed on: july.20, 2019. [online]. Available: https://www.researchgate.net/publication/309679945_The_Study_of_Pima_Indian_Diabetes.

[10] D. K. Choubey, S. Paul, "A Hybrid Intelligent System for Diabetes Disease Diagnosis", International Journal of Bio-Science and Bio-Technology (IJBSBT), SERSC, ISSN: 2233–7849, Vol. 7, No. 5, pp. 135–150, 2015.

[11] D. K. Choubey, S. Paul, "A Hybrid Intelligent System for Diabetes Disease Diagnosis", International Journal of Intelligent Systems and Applications, Vol. 8, No. 1, pp. 49–59, 2016.

[12] D. K. Choubey, S. Paul, S. Kumar, and S. Kumar, "Classification of Pima Indian diabetes dataset using naive Bayes with genetic algorithm as an attribute selection," Communication and Computing Systems: Proceedings, London, 2017, pp. 451-455, Accessed on: july.12, 2019.

[13] H. Wu, S. Yang, Z. Huang, J. He, and X. Wang, "Type 2 diabetes mellitus prediction model based on data mining," Informatics in Medicine Unlocked, vol. 10, pp. 100-107, 2018.

[14] A. Alharbi, and M. Alghahtani, "Using Genetic Algorithm and ELM Neural Networks for Feature Extraction and Classification of Type 2-Diabetes Mellitus," Applied Artificial Intelligence, vol. 33, no. 4, pp. 311-328, 2018.

[15] K. Gholipour, M. A. Jafarabadi, S. Iezadi, S. Jannati, and S. Keshavarz, "Modeling the prevalence of diabetes mellitus risk factors based on artificial neural network and multiple regression," Eastern Mediterranean Health Journal, vol. 24, no. 8, pp. 770-777, 2018.

[16] N.S. El-Jerjawi et S.S. Abu-Naser, 'Diabetes Prediction Using Artificial Neural Network,' International Journal of Advanced Science and Technology, vol. 121, pp. 55-64, 2018.

[17] M.A. Benbouras, R. Mitiche Kettab, H. Zedira, F. Debiche,and N. Zaidi, "comparing nonlinear regression analysis and artificial neural networks to predict geotechnical parameters from standard penetration test,". Urbanism. Architecture. Constructions, vol. 9, no 3, pp. 275-288, 2018.

[18] M.A. Benbouras, R. Mitiche Kettab, H. Zedira, F. Debiche,and N. Zaidi, "comparing nonlinear regression analysis and artificial neural networks to predict geotechnical parameters from standard penetration test,". urbanism. Architecture. Constructions, vol. 9, no 3, pp. 275-288, 2018.

[19] M. A. Benbouras, B. Mitiche Kettab, H. Zedira A. I. Petrisor, N. Mezouar, and, F. Debiche, "A new approach to predict the compression index using artificial intelligence methods," Marine Georesources and Geotechnology, vol. 37, no 6, pp 704-720.

[20] Kaggle.com, "The pima-indians-diabetes-database," Mars, 2016. Accessed on: Avrile.14, 2019. [online]. Available: www.kaggle.com/uciml/pima-indians-diabetes-database.

[21] G. Ciaburro, "MATLAB for Machine Learning: Practical examples of regression, clustering and neural networks," Packt Publishing, 2017.

[22] M. A. Benbouras, "L'utilisation des SIG sur les données d'infrastructure et de fondations : Application dans la zone d'Alger," Ph.D dissertation, University Abbes Laghrour, Khenchela, 2018.

[23] K.M. Ting, "Confusion matrix," Encyclopedia of Machine Learning and Data Mining , pp. 260-260, 2017.