ISSN No: 2350-1278 (Print)
Peer Reviewed & Refereed Journal
National Research Journal

AN AI-BASED INTELLIGENT HYBRID MODEL FOR DIABETES MELLITUS PREDICTION

Parminder Kaur

Associate Professor, Department of Computer Science, Khalsa College for Women, Civil Lines, Ludhiana

Kiranjit Kaur

Assistant Professor, Department of Computer Science, Guru Nanak College, Moga

ABSTRACT

Artificial Neural Networks (ANN) are widely employed for prediction in several arenas. ANNs are highly capable of interpreting and extracting meaningful insights from complex medical data sets with great precision and have thus been found to be immensely successful in disease prediction, diagnosis, and facilitating customized treatment approaches. The capability of Machine learning techniques to make accurate predictions at high speeds has resulted in their successful application in the healthcare sector as predictive models in almost all the departments of the hospitals.

Diabetes Mellitus, more commonly known as Diabetes is a widespread disease across the globe. Early prediction of diabetes is essential for timely interventions and lifestyle modifications. It also reduces the risks of complications facilitating more effective management strategies for enhanced long-term health outcomes.

This paper explores the possibility of combining the benefits of predicting Diabetes Mellitus using Artificial Neural Networks based models with the Nature Inspired Optimisation Algorithm, namely, Bacteria Foraging Optimisation Algorithm to propose a hybrid model for predicting the Diabetes Mellitus disease with higher accuracy.

Keywords: ANN, Artificial Neural Network, Bacteria Foraging Optimisation, Diabetes Mellitus, Hybrid model.

1. INTRODUCTION

Diabetes is a pervasive illness that affects millions of people across the world, thus posing momentous health challenges. Around 537 million adults aged between 20 and 79 years are currently affected with diabetes which constitutes 1 in 10 individuals. This number is predicted to rise to 643 million by 2030 and 783 million by 2045 (International Diabetes Federation, 2021). Diabetes is also considered to be responsible for 6.7 million deaths in 2021 at the rate of 1 in every 5 seconds. Diabetes related health expenditures have risen to as much as USD 966 billion dollars which is nearly 316% increase over the last 15 years (International Diabetes Federation, 2021).

Diabetes mellitus is a persistent ailment characterized by raised blood glucose levels. It stems from either inadequate insulin production by the pancreas or reduced cellular sensitivity to insulin's glucose-lowering effects (World Health Organization, n.d.). Secondary causes may include pancreatic disorders, genetic syndromes like myotonic dystrophy, or medication, such as glucocorticoids. Gestational diabetes, a temporary condition during pregnancy, involves increased blood glucose levels that typically normalize post-delivery (National Centre for Biotechnology Information, n.d.). Diabetes is a major cause of blindness, kidney failure,

ISSN No: 2350-1278 (Print)
Peer Reviewed & Refereed Journal
National Research Journal

heart attacks, stroke and lower limb amputation. In 2019, diabetes and kidney disease due to diabetes caused an estimated 2 million deaths (World Health Organization, n.d.).

The study is noteworthy in the perspective that it combines the advantages of ANN and Nature Inspired Optimisation algorithm to propose a hybrid prediction model for prediction of Diabetes mellitus with greater precision. The paper has five sections. Section 1 presents the research problem along with Artificial Neural Networks and Bacteria Foraging Optimization Methods used for the study. Section 2 discusses the previous research done in the area through a literature survey. The dataset used and the research methodology followed for this study are discussed in Section 3. In Section 4, the results achieved during the study are recorded while Section 5 discusses the conclusions drawn from the study.

1.1 Artificial Neural Networks (ANNs)

Artificial Neural Networks (ANN) is the AI technique that replicates the behavior and functioning of human brain. These networks mimic working of neurons of the human brain (Peterson and Rognvaldsson, 1992). The structure and working of ANNs have been described in a number of studies by different researchers from time to time (Maren et.al (1990), Hecht-Nielsen (1990), Zurada (1992), Fausett (1994), Ripley (1996)). These networks learn from examples. The network weights are suitably adjusted in these networks to relate the input variables with the output variables. The ability of Artificial Neural Networks to model complex problems without any prior information about the nature of relationship between the network variables give them an edge over the other statistical techniques used for prediction purposes (Hubick, 1992). Different types of ANN architectural models are available.

This paper uses Feed Forward Back Propagation Artificial Neural Network (FFBP-ANN) based model for hybridization with the BFO model for forecasting Diabetes mellitus. The FFBP ANN network architecture uses three kinds of layers namely, input, hidden and output layer. All the neurons in one layer are connected to every other neuron of the subsequent layer. For training the network, Back propagation learning algorithm is used. Using the current values of the inputs and the weights, the output is obtained. The network error value is by comparing the forecasted diabetes indicators with the actual values. This value is propagated backward in the network to update and adjust the network weights. This helps in further improving the network's prediction accuracy.

1.2 Bacteria Foraging Optimisation Algorithm

Passino (2002), proposed this nature inspired algorithm. This algorithm is centered on the foraging behavior of a swarm of bacteria, E.coli. While foraging, these bacteria attempt to maximize their energy intake (E) per unit time that they spend on foraging (T). Atasagun Y & Kara Y, 2014), (Singh G & Walia BS, 2016), (Dhaliwal BS and Pattnaik SS, 2016), besides others researchers have implemented BFOA. BFOA algorithm makes use of four steps (Passino, 2002), Das et al, 2009), (Coelho, 2010), (Gazi and Passino, 2010), (Brownlee, 2011), (Atasagun and Kara, 2014).

1.2.1 Chemotaxis

This step is based on the two types of movements of bacteria, E.coli: swimming or tumbling. Random movements with little displacement takes place during tumbling while movement along a single direction takes place during the swimming movement. For the p^{th} bacteria at q^{th} chemotactic step, r^{th} reproductive step and l^{th} elimination dispersal step, the model can be can mathematically defined as (Passino, 2002):

$$\theta^i(q+1,r,l) = \theta^i(q,r,l) + C(p)\emptyset(q)$$

ISSN No: 2350-1278 (Print)
Peer Reviewed & Refereed Journal
National Research Journal

$where\ \theta^i(q,r,l)$ denotes the position of p^{th} bacteria for the given values of parameters q, r and l. The step size is represented by *C (p)* which also represents the unit run-length of the E-coli bacteria. For swimming movement, $\theta^p(q+1,r,l)>\theta^p(q,r,l)$ and for the tumbling movement of the bacteria, $\theta^p(q+1,r,l)>\theta^p(q,r,l)$.

1.2.2 Swarming

In this step, swarms of bacteria move in concentric paths towards high nutrition regions (Passino, 2002) on release an attractant called aspartate. To signal the other bacteria to maintain a minimum distance from each other, a repellant is also released by the bacteria. Mathematically this step can be modeled as:

$$J_c(\theta, P(j,k,l)) = \sum_{i=1}^{S} \quad J_c^i(\theta, \theta^i(j,k,l))$$

$$= \sum_{i=1}^{S} \quad [J_c^i(\theta, \theta^i(j,k,l))$$

$$= \sum_{i=1}^{S} \quad [-dexp(-\omega_{attr} \sum_{m=1}^{P} \quad (\theta_m - \theta_m^i)2)_{attr}]$$

Here S denotes the sum count of all the bacteria present in the swarm;

and P represents the total number of variables that need to be optimized; $\theta = [\theta_1, \theta_2, \theta_{3,\ldots\ldots}, \theta_p]^T$ characterizes a point in a search space of dimension *P;*

d_{attr}, ω_{attr}, h_{repl}, ω_{attr} represent the coefficients that signify the respective quantity as well as diffusion rates of the attractant along with the repellant (Das et al., 2009).

$J_c(\theta, P(j,k,l))$ signifies a time varying objective function. The sum value of this objective function and actual objective function is computed.

1.2.3 Reproduction

The cost function denoted by *J (i, j, k, l)* is used to calculate the health or fitness value of each bacteria. (J_{health}) which gives the health or the fitness value of the i^{th} bacterium can be represented as,

$$J_{health}^i = \sum_{i=1}^{N_c} \quad J^i(j,k,l)$$

Where N_c is the count of the chemotaxis steps.

The bacteria are then organized in the ascending order of their fitness values, after which they are divided into two groups on the basis of their respective fitness values. Subsequently, the group possessing lower fitness values dies. And the remaining half of the bacteria are then allowed to reproduce asexually by splitting into two identical bacteria (Okaeme and Zanchetta, 2013), thus keeping the swarm size fixed (Dasgupta, 2009).

1.2.4 Dispersal and Elimination

This step of Dispersal and Elimination helps in preventing the bacteria from being confined in local optima and helps it obtain a global optimum value. For elimination, a bacterium with a predefined probability p_{ed} is selected while another bacterium is chosen for dispersal.

ISSN No: 2350-1278 (Print)
Peer Reviewed & Refereed Journal
National Research Journal

2. DATA AND METHODOLOGY

The dataset selected for this study is the Pima dataset which is well-known for its application to diabetes research (Smith et al., 1998). Compiled by the National Institute of Diabetes and Digestive and Kidney Diseases, the Pima dataset which is an open-source dataset includes 768 observations related to females only, providing a wide-ranging pool of data for analysis. This dataset includes essential features such as Age, Glucose levels, Insulin levels, Blood Pressure readings, Pregnancy history, Skin Thickness measurements, BMI calculations, and the Diabetes Pedigree Function score, each contributing to the study of diabetes-related patterns and tendencies. The "Outcome" variable in the dataset serves as the target variable, signifying whether an individual exhibits the absence (0) or presence (1) of diabetes.

Table 1 describes the variables used in the Pima dataset used for the study.

Table 1: Attributes of the dataset

Variable Name	Description
Age	Age of the individual in years
Glucose	Amount of glucose present in the blood after two hours (2-hour postprandial blood sugar level)
Insulin	Insulin level in the blood (measured in μU/ml)
Blood Pressure	Blood pressure measured against artery walls as it passes through the body (mm Hg)
Pregnancies	Total number of pregnancies the woman has carried
Skin Thickness	Thickness of the triceps skinfold (mm)
BMI	Body Mass Index calculated as (weight in kg) / (height in m) ^2
Diabetes Pedigree Function	A scoring system assigning a score to the chance of developing diabetes based on family medical history
Outcome	Target variable indicating the absence (0) or presence (1) of diabetes in the patient (Class Variable)

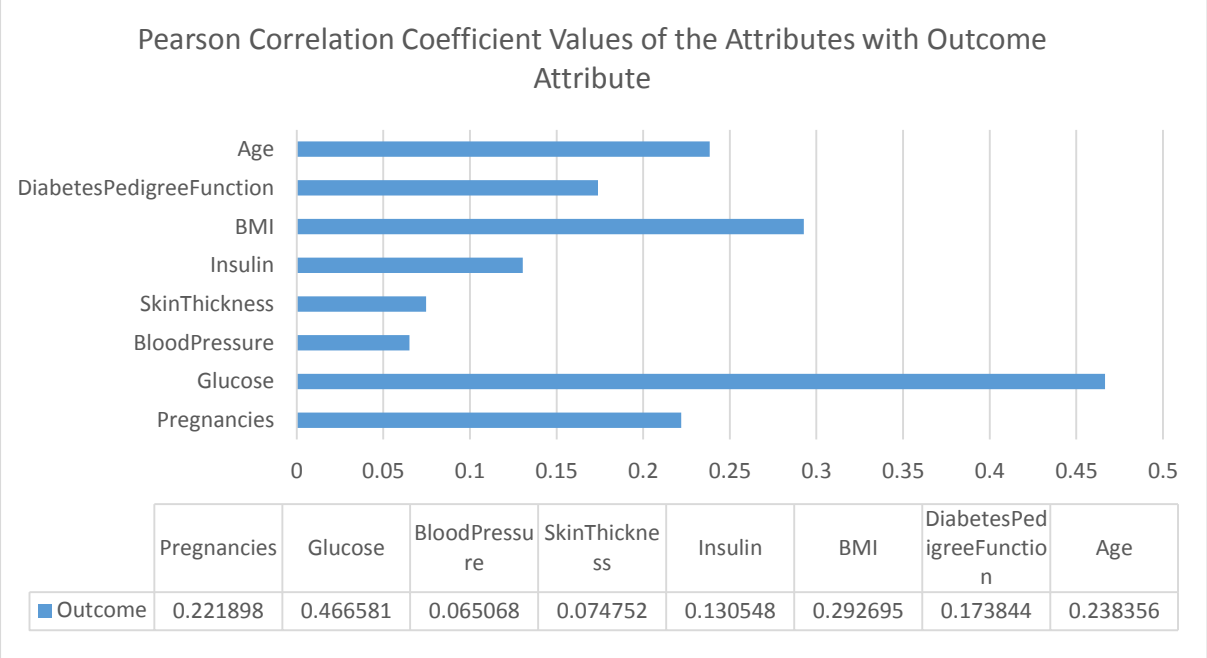
The descriptive statistics of the Pima Indian Diabetes dataset is given in Table 1.

Pearson Correlation Coefficient values of the different attributes with the Outcome attribute were computed to find the influence of each attribute on the Outcome variable. These correlation values depict the amount and direction of influence of each attribute on the likelihood of occurrence of diabetes. This is useful in providing useful insights for further analysis. Fig 2 provides the various correlation coefficient values of the attributes of the dataset with the Outcome variable

Table 1: Descriptive Statistics of the Pima Indian Diabetes Dataset

	Pregnancies	Glucose	Blood Pressure	Skin Thickness	Insulin	BMI	Diabetes Pedigree Function	Age	Outcome
COUNT	768	768	768	768	768	768	768	768	768
MEAN	3.845052	120.894531	69.105469	20.536458	79.799479	31.992578	0.47188	33.2409	0.34896
STD	3.369578	31.972618	19.355807	15.952218	115.244002	7.88416	0.33133	11.7602	0.47695
MIN	0	0	0	0	0	0	0.078	21	0
25%	1	99	62	0	0	27.3	0.24375	24	0
50%	3	117	72	23	30.5	32	0.3725	29	0
75%	6	140.25	80	32	127.25	36.6	0.62625	41	1
MAX	17	199	122	99	846	67.1	2.42	81	1

ISSN No: 2350-1278 (Print)
Peer Reviewed & Refereed Journal
National Research Journal



	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunction	Age
Outcome	0.221898	0.466581	0.065068	0.074752	0.130548	0.292695	0.173844	0.238356

Fig 1: Pearson's Correlation Coefficient Values of the Attributes of the dataset with the Outcome Variable

To visualize the distribution of the attributes of the dataset and to identify the outliers in the dataset, Box Plots of the dataset was plotted.

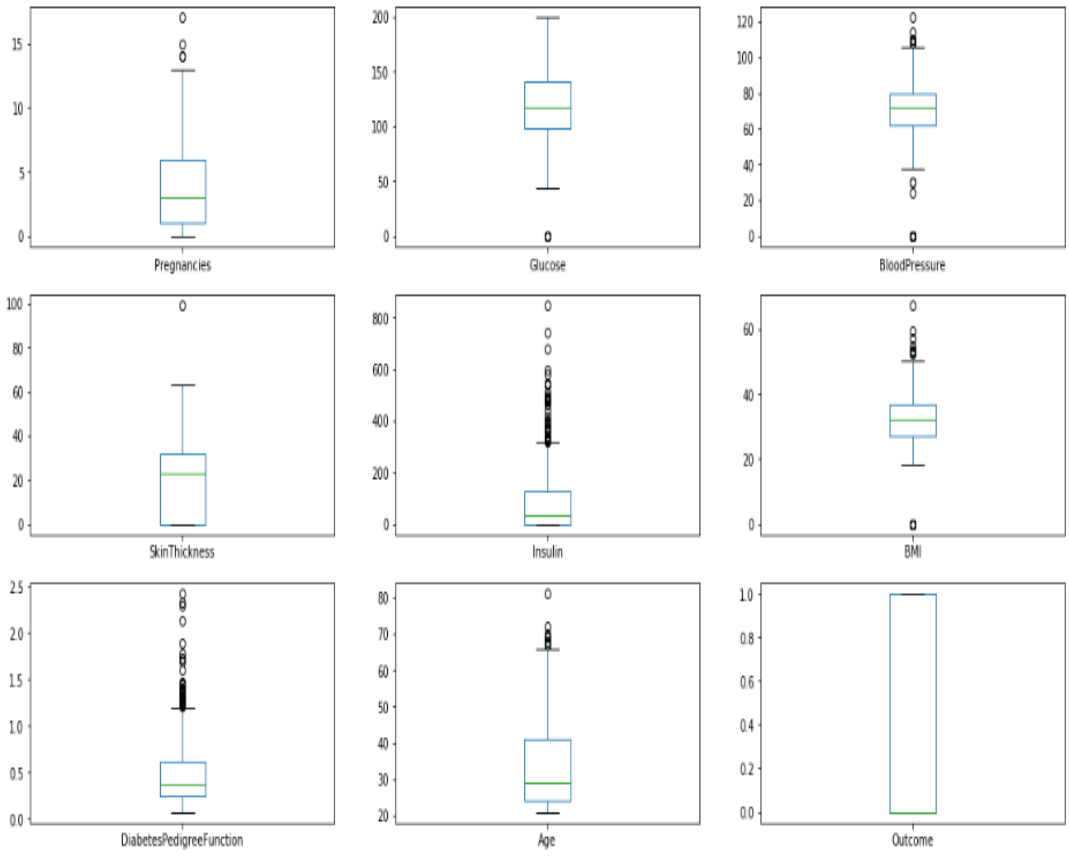


Fig 2: Boxplot of the attributes of the dataset

ISSN No: 2350-1278 (Print)
Peer Reviewed & Refereed Journal
National Research Journal

Three models were developed. The first model used FFBP ANN to predict occurrence of diabetes. In the second model the BFO model was used for forecasting while in the third model, BFO was hybridized with FFBP ANN model.

60% of the dataset was utilized for training the network, while 20% was used for testing the network. For validation purposes, the remainder 20% of dataset was utilized. Matlab 8.1 software was used for the BFO and the ANN model. Training of the networks was carried out for a constant number of epochs. The optimum count of hidden neurons was established experimentally by varying the network design and implementing the training procedure numerous times until optimum performance level of the network was achieved.

3. RESULTS

A three-layer network MLP structure (with one layer each of input, hidden and output) was chosen with tan-sigmoid training function for the first layer of ANN and Purelin training function was employed for the second layer of ANN. Levenberg Marquardt training algorithm was used to train the ANN. ANN architecture 11-18-1 with 11input layer neurons, 18 hidden layer neurons and 1 output layer neuron was found to be the most optimum one. The learning rate employed was 0.8.

The correlation coefficient values obtained for the three models-FFBP ANN only, BFO only and FFBP ANN+ BFO hybrid model for forecasting the occurrence of diabetes mellitus are listed in Table 2.

Table 2: Values of Correlation Coefficients obtained before and after hybridization of BFO model with FFBPANN for forecasting Diabetes Mellitus.

Sr. No.	Model	Coeff. of Correlation
1	**FFBPANN**	0.9872
2	**BFO only**	0.9881
3	**BFO+FFBPANN**	**0.9899**

4. CONCLUSION

An AI-based hybrid model using the Bacteria Foraging Optimisation method and Feed-Forward Back-Propagation Artificial Neural Network was developed for forecasting the occurrence of diabetes mellitus. The hybrid model showed an improved correlation between the Outcome variable values and the actual outcome values over the models that employed only either FFBP ANN or BFO only methods. The hybrid model obtained a correlation coefficient value of 0.9899 which was more than the coefficient correlation coefficient value of 0.9872 obtained through FFBP ANN model and the BFO model which resulted in correlation coefficient value of 0.9881. Hence, it can be concluded that the hybridization of BFO optimization model with FFBP ANN model improves the forecasting efficiency of FFBP ANN model.

REFERENCES

Atasagun, Y., Kara, Y., (2014), "Bacterial Foraging Optimization Algorithm for Assembly Line Balancing," Neural Comput & Applic, 25(1), pp. 237-250.

Brownlee, J., (2011), Clever Algorithms: Nature Inspired Programming Recipes, Lulu Enterprises, Australia, pp. 257–264.

Coelho, L. S., Silveira, C. C, Sierakowski, C. A., Alotto, P., (2010), "Improved Bacterial Foraging Strategy Applied to TEAM Workshop Benchmark Problem," IEEE Trans Magn, 46(8), pp. 2903–2906.

Das, S., Dasgupta, S., Biswas, A., Abraham, A., (2009), "On Stability of the Chemotactic Dynamics in Bacterial-Foraging Optimization Algorithm," IEEE Trans Syst Man Cybern, 39(3), pp. 670–679.

ISSN No: 2350-1278 (Print)
Peer Reviewed & Refereed Journal
National Research Journal

Dhaliwal, B. S., Pattnaik, S. S., (2016), "BFO–ANN Ensemble Hybrid Algorithm to Design Compact Fractal Antenna for Rectenna System," Neural Comput & Applic, 28(S1).

Fausett, L. V., (1994), "Fundamentals Neural Networks: Architecture, Algorithms and Applications," Prentice-Hall, Inc., Englewood Cliffs, New Jersey.

Gazi, V., Passino, K. M., (2010), Swarm Stability and Optimization, Springer, New York, pp. 233–249.

Hecht-Nielsen, R., (1990), Neurocomputing, Addison-Wesley Publishing Company.

International Diabetes Federation. (2021). IDF Diabetes Atlas, 10th edn. Brussels, Belgium. Retrieved from https://www.diabetesatlas.org

Maren, A., Harston, C., Pap, R., (1990), Handbook of Neural Computing Applications, Academic Press, Inc., San Diego, California.

National Centre for Biotechnology Information. (n.d.). Diabetes Mellitus. In StatPearls [Internet]. Retrieved March 7, 2023 from https://www.ncbi.nlm.nih.gov/books/NBK279128/

Okaeme, N. A., Zanchetta, P., (2013), "Hybrid Bacterial Foraging Optimization Strategy for Automated Experimental Control Design in Electrical Drives," IEEE Trans Ind Informat, 9(2), pp. 668–678.

Passino, K. M., (2002), "Biomimicry of Bacterial Foraging for Distributed Optimization and Control," IEEE Control System Management, 22, pp. 52–67.

Peterson, C., Rognvaldsson, T., (1992), "An Introduction to Artificial Neural Networks," In: CERN Yellow Report, pp. 113–170.

Ripley, B. D., (1996), Pattern Recognition and Neural Networks, Cambridge University Press.

Singh, G., Walia, B. S., (2017), "Performance Evaluation of Nature-Inspired Algorithms for the Design of Bored Pile Foundation by Artificial Neural Networks," Neural Comput & Applic, 28(S1), pp. 289–298.

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1998). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. In Annual Symposium on Computer Applications in Medical Care (pp. 261–265).

World Health Organization. (n.d.). Diabetes. Retrieved Retrieved March 7, 2023 from https://www.who.int/health-topics/diabetes#tab=tab_1

Zurada, J. M., (1992), Introduction to Artificial Neural Systems, West Publishing Company, St. Paul.