

# Diabetes Prediction by Frog Jumping Algorithm and Artificial Neural Network

Roopa Shrivastava, Prof. Rajesh Ku. Nigam, Prof. Rakesh Ku. Tiwari

Dept. of Computer Science and Engineering  
Technocrats Institute of Technology & Science Bhopal, MP, India

**Abstract-** Diabetes is a disease that is continuously increasing in developed and developing countries. The efficient solution to deal with this disease is need of the time. This work has reduced the input dataset size by use of frog leaping genetic algorithm that give high effective features. Selected features will be further used for the training of neural network that give more accurate results. Use of less feature for training and testing directly reduces the evaluation complexity. Experiment was done on real dataset of users feature who might have chance of diabetes. Comparison of proposed model is done with existing techniques on different evaluation parameters. Result shows that proposed model has improved precision, recall, and accuracy of diabetes prediction.

**Keywords-** Cloud service modes, computers, tablets, laptops, or smartphones etc .

## I. INTRODUCTION

Sugar in the blood that is increasingly affecting people all over the world, including those in the most developed nations. Diabetes, by its very nature as a globally problematic disease, necessitates the highest level of commitment from medical professionals, patients, families, and society as a whole. Diabetes is a chronic disease that has significant social, health, and economic consequences [1]. Diabetes is a chronic disease characterised by an increase in glucose or blood sugar levels as a result of the body's inability to produce insulin, or as a result of insufficient insulin production, or as a result of insulin's inability to act on the cells of the body [2, 3]. Medical professionals are still baffled as to why such a thing is occurring, and they have labelled the underlying cause as "x syndrome." Historically, diabetes treatment has focused on treating the symptoms rather than the underlying cause.

Approximately 5% of the world's population is affected by diabetes, according to the World Health Organization, and the number of patients is increasing all the time [1]. A large proportion of diabetics in developed countries are over the age of 65, with the majority of diabetics occurring in this age group. Whereas in developing countries, such as those in which our country is located, the majority of diabetics are between the ages of 45 and 64, type 2 diabetes has become more prevalent in recent years in people between the ages of 30 and 40 [1]. Multiple factors, such as the history of Blood Glucose (BG) values, insulin dosages, physical activity, and carbohydrate intake, can have a direct impact on blood glucose levels. Other factors such as body mass index, stress level, sleep quantity, illnesses and medications as well as menstruation, alcoholism, allergies and altitude can all have an impact on glucose levels in a person's blood. Furthermore, many

technical factors, in addition to the factors that affect BG levels, can influence the predictive performance metrics of a model, including the type of machine learning, data size, prediction horizon (PH), validation approaches, and so on [4].

EHR data are a valuable source of information that can be used to improve predictive medical models by incorporating it into them. The availability of historical data leads logically to the application of data mining techniques for pattern discovery [5, which are described below]. The goal is to develop rules that will aid in the understanding of diabetes and make it easier to diagnose it at an earlier stage. In the field of medicine, there is a great deal of interest in the prevention of diabetes. The use of data mining expedites data analysis, and analysts can examine existing data to identify patterns and trends in the development and progression of diabetes.

## II. RELATED WORK

Using the definition provided by Deng and D. Yu [6, deep learning is "a class of machine learning algorithms that employ multiple layers of nonlinear processing units for feature extraction and transformation." Machine learning algorithms can be divided into two types: supervised learning algorithms and unsupervised learning algorithms. When given a dataset (labelled data), supervised algorithms learn how to map the input to the associated output/target feature(s). Unsupervised algorithms, on the other hand, create internal representations from the structure and distributions of data samples that are not associated with the output/target feature(s) to identify. According to Rodr'guez-Rodr'guez et al. [7], a total of 25 diabetic type 1 patients were monitored over the course of 14 days while going about their daily routines, using continuous glucose monitoring (CGM) systems for their

blood glucose levels and a wearable smart band that monitored their heart rate, physical activity (number of steps), and sleep data. With this study, researchers compared the features used in feature selection and forecasting for machine learning algorithms that are used to predict glycaemia in type 1 diabetic patients [14].

Deepti and Dilip [8] applied algorithms such as Decision Trees, SVMs, and Naive Bayes. To improve performance, a ten-fold cross validation procedure was used. The Naive Bayes algorithm achieved the highest level of accuracy, with a precision of 76.30 percent on average. The Pima Indian Diabetes dataset was used in both of these studies.

Swapna et al. [9] used Deep Learning techniques to predict the onset of diabetes in patients. Using a Multilayer Feed-Forward Neural Network, the former was accomplished. For the purpose of training the model, the back-propagation algorithm was used. For numerical stability, they also used the PIMA Indian dataset, which was normalised before being used in the pre-processing stage. They were able to achieve an accuracy of 82 percent. The latter made use of a dataset called Electrocardiograms, which was used to train two models using CNN and CNN-LSTM, respectively. The dataset consisted of 142,000 samples with eight attributes per each of the samples. With a five-fold cross validation for both models, they were able to achieve an accuracy of 93.6 percent with the CNN model and an accuracy of 95.1 percent with the CNN-LSTM model, respectively.

Kaur and Kumari, 2019 [10] developed five different models for the detection of diabetes using support vector machines, including the linear kernel support vector machine (SVM-linear), the radial basis kernel support vector machine (SVM-RBF), the K Nearest Neighbour (k-NN), Artificial Neural Networks (ANN), and Multifactor Dimensionality Reduction (MDR) algorithms, among other techniques. The dataset's features were chosen with the help of the Boruta wrapper algorithm, which took into account several evaluation criteria, including accuracy, recall, precision, F1 score, and the Area Under the Curve (AUC).

The experimental results revealed that all of the models produced satisfactory results, with the SVM-linear model providing exceptionally high accuracy and precision (0.89 and 0.88, respectively). Based on the findings of this study, it can be concluded that the linear kernel support vector machine (SVM-linear) and the k-NN are the two (2) most accurate predictive models for diabetes when all of the parameters are taken into consideration. The author of [11] presents a model for diabetes prediction that is based on a fusion of machine learning approaches. Support Vector Machine (SVM) and Artificial Neural Network (ANN) models are two types of models that make up the conceptual framework. These models perform an analysis of the dataset in order to determine whether a diabetes

diagnosis is confirmed or denied. The dataset used in this study is divided into two parts: training data and testing data, with a 70:30 split between the two parts. The output of these models is used as the input membership function for the fuzzy model, and the fuzzy logic is used to determine whether or not a diabetes diagnosis is positive or negative at the end of the process.

### III. PROPOSED METHODOLOGY

Detection of diabetic patient is performed in this work by use of Frog Leaping genetic algorithm and artificial neural network. Explanation of proposed model FLNNDP (Frog Leaping and Neural Network based Diabetic Prediction) is done in this section of paper. Whole work was divide into two module first is feature selection and other is training of neural network.

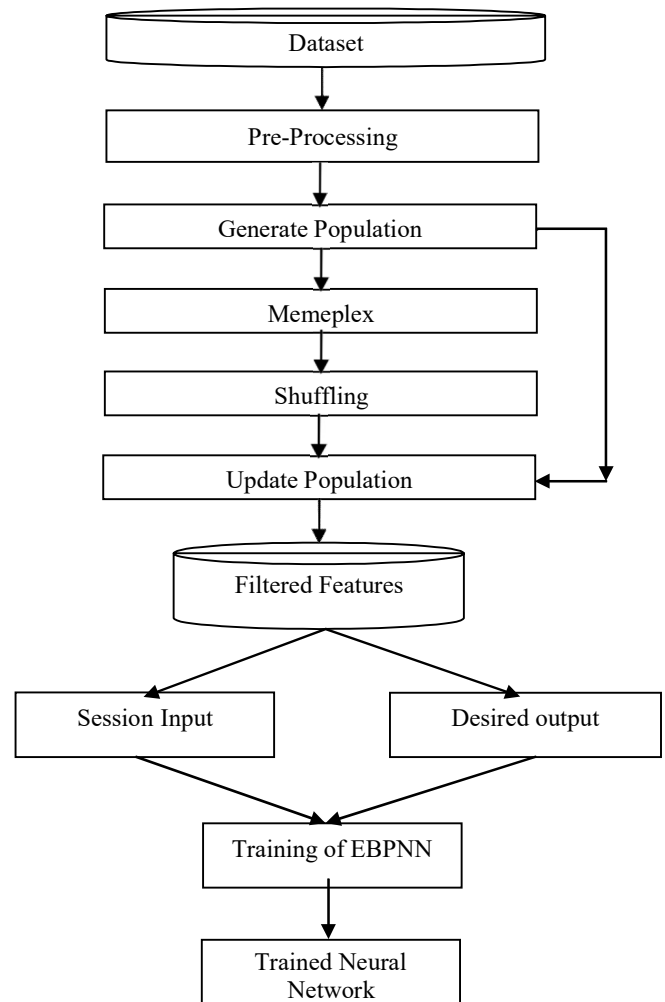


Fig. 1 Block diagram of diabetes prediction.

#### 1. Pre-Processing

In this step of FLNNDP input dataset is pre-process by removing unwanted session features like column name,

row id, etc. Some of text data is transformed into numeric representative values like gender (male / female), weakness (Yes/No), to 1 /0. Numeric values present in data are use as it is without any modification. Class of the data positive / Negative is also transformed into its equivalent numeric value 1/0. This class terms is not feature set but it's a prediction class use in training of neural network.

## 2. Generate population

In this SFLA (Shuffled Frog Leaching Algorithm) set of probable solutions act as frog, so collection of frog is term as population. Hence frog having set of features in form of 1 or 0, where 1 means feature is use and 0 means feature not select in this probable solution (frog) [12]. Assuming  $f_{1-n}$  n number of features, M number of frogs and m minimum number of feature in a set. Hence frog generation function is shown in Eq. 1.

$$P_f \leftarrow \text{Generate\_Frog}(f, M, m) \quad \text{Eq. 1}$$

**3. Fitness Function-** Estimation of fitness value of frog depends upon the input feature training matrix as per frog F. Hence this function perform training of neural network from input feature matrix, than testing of trained neural network for same input feature matrix.

**4. Memplex-** Estimate the fitness value of each frog where accuracy of intrusion detection were sort in descending order. After sorting some set of frogs were cluster which is term as memplex. Hence whole population  $P_f$  is divide into g cluster where each memplex have t number of frog.

$$M = g \times t.$$

**5. Shuffling-** In this step of genetic algorithm crossover of the algorithm was done by selecting one best parent in memplex. So as per fitness value crossover with other set of frogs were performed. So selection of this common parent depends on fitness value. Here best fitness values frog act as common parent in all crossover operation in a memplex. So other set of chromosome undergoes crossover by randomly replacing a feature presence or absence status as per common parent frog set.

**6. Population Updation-** As shuffling changes the frogs of the population so retention of this frog depends on fitness value. This can be understand if new frog have good fitness value as compared to parent frog fitness value than new frog was include in the population, otherwise parent frog will continue in population. Hence in all situation population size will never change from M number.

## 7. Training Error Back Propagation Neural Network

As per SFLA output, features from the input pre-processed dataset is pass as training input vector. As per subclass of

intrusion sessions were target into different class. So output class are shown as positive/Negative (1/0). Here for training input and output are arrange into feature matrix and desired output value.

## 8. Error Back Propagation Neural Network Training:

In this step Error Back Propagation neural network was consider for the training of intrusion detection system, here training vector consist of s selected features value set where if feature is consider for training by SFLA than non zero value is present other 0 present for non selected feature.

- A layer neural network is assume which have three layers.
- Input layer neuron were identified by i, while hidden layer neuron were identified by j. Output layer neuron is identified by k.
- Weights between neuron is represent by  $w_{ij}$ , where i and j are neuron layers.

$$X_j = \sum x_i \cdot w_{ij} \text{-----Eq. 2}$$

where,  $1 \leq i \leq n$ ; n is the number of inputs to node j, and  $b_j$  is the biasing for node j. Hence network will learn the weights between layers. This error need to be correct by adjusting the weight values of each layer. So estimation of error was done by eq. 3 [13].

$$e_k(n) = d_k(n) - y_k(n) \text{---Eq. 3}$$

This error need to be correct by adjusting the weight values of each layer. So here forward movement of the neural network is over and error back propagation starts.

## III. PREDICTION TECHNIQUES COMPARING PARAMETERS

### 1. Experimental Setup

Implementation of proposed genetic algorithm based diabetes prediction model was done on MATLAB software. Implementation of model was done on I3 10<sup>th</sup> generation processor having RAM of 4GB. Experimental dataset have 16 features and 521 patient data for analysis [14].

### 2. Evaluation Parameters

As various techniques evolved different steps of working for segmenting data into appropriate category. So it is highly required that proposed techniques or existing work need to be compare on same dataset. But prediction class which are obtained as output is need to be evaluate on the function or formula. So following are some of the evaluation formula which help to judge the clustering techniques ranking.

$$\text{Precision} = \frac{\text{True\_Positive}}{\text{True\_Positive} + \text{False\_Positive}}$$

$$\text{Re call} = \frac{\text{T rue\_ Positive}}{\text{T rue\_ Positive} + \text{False\_ Negative}}$$

$$F\_Score = \frac{2 * \text{Precision} * \text{Re call}}{\text{Precision} + \text{Re call}}$$

$$\text{Accuracy} = \frac{(\text{True\_Positive} + \text{True\_Negative})}{(\text{True\_Positive} + \text{True\_Negative} + \text{False\_Positive} + \text{False\_Negative})}$$

In above true positive value is obtained by the system when the classified data is same as in actual case or ground truth class. While in case of false positive value it is obtain by the system when the classified data is not of same case as in actual in or ground truth class.

Table 1 Diabetes prediction models precision value based comparison.

Testing Dataset Percentage	Previous Model	FLNNDP
20	0.1096	1
40	0.1792	1
60	0.2331	0.9918
80	0.4225	0.9865
100	0.6101	0.9875

Above table 2 shows that proposed work FLNNDP has improved the precision evaluation parameters values as compared to previous work [11]. Use of frog leaping has optimized features for training. Training of model neural network from selected features has increases the work efficiency.

Table 2 Diabetes prediction models recall value based comparison.

Testing Dataset Percentage	Previous Model	FLNNDP
20	1	1
40	0.9894	0.9895
60	0.9918	0.968
80	0.9954	0.9821
100	0.9968	0.9875

Above table 2 shows that proposed work FLNNDP has improved the recall evaluation parameters values as compared to previous work [11]. Use of sigmoidal function for neural network training has improved the comparing

parameter values. This was enhanced by the frog selected features as well.

Table 3 Diabetes prediction models f-measure value based comparison.

Testing Dataset Percentage	Previous Model	FLNNDP
20	0.1976	1
40	0.3034	0.9947
60	0.3775	0.9798
80	0.5932	0.9843
100	0.7570	0.9875

Above table 3 shows that proposed work FLNNDP has improved the F-measure evaluation parameters values as compared to previous work [11]. Use of frog leaping has optimized features for training. Training of model neural network from selected features has increases the work efficiency.

Table 4 Diabetes prediction models accuracy value based comparison.

Testing Dataset Percentage	Previous Model	FLNNDP
20	0.1113	1
40	0.1804	0.9981
60	0.2342	0.9904
80	0.4261	0.9866
100	0.6142	0.9846

Above table 4 shows that proposed work FLNNDP has improved the recall evaluation parameters values as compared to previous work [11]. Use of sigmoid function for neural network training has improved the comparing parameter values. This was enhanced by the frog selected features as well.

#### IV. CONCLUSIONS

Medical science dependency on medical practioners diagnosis increases day by day. Many of researchers are working to reduce this work load and increase the diagnosis accuracy for identifying disease at very initial stage. This paper has developed a model that predict the diabetic situation of the patient as per different observation features. For prediction out of different features few set of features were select by Frog leaping genetic algorithm. Selected features were used for the training fo neural network that uses sigmoid activation function. Experiment was done on real dataset of diabetic paitentns. Result shows

that proposed model has improved the detection accuracy of the model. In future scholar predict some other disease.

## REFERENCES

- [1] <http://www.ishp.gov.al/wpcontent/uploads/2015/kalendar/Dita%20b%20oterore%20e%20diabetit.pdf>
- [2] <https://www.familjadhshendeti.com/semundja-esheqerit-diabeti-te-femrat/S>.
- [3] Bo He, Kuang-i Shu and Heng Zhang, Machine Learning and Data Mining in Diabetes Diagnosis and Treatment, IOP Conference Series: Materials Science and Engineering, Volume 490, Issue 4, IOP Conf. Series: Materials Science and Engineering 490 (2019) 042049 IOP.
- [4] Ashenafi Zebene Woldaregay, Eirik °Arsand, St°ale Walderhaug, David Albers, Lena Mamykina, Taxiarchis Botsis, Gunnar Hartvigsen Data-driven modeling and prediction of blood glucose dynamics: Machine learning applications in type 1 diabetes. 2019.
- [5] Marcus, Y, Eldor, R, Yaron, M, et al. Improving blood glucose level predictability using machine learning. Diabetes Metab Res Rev. 2020.
- [6] Deng, Li and Dong Yu (2014). „Deep learning: methods and applications“. In: Foundations and trends in signal processing 7.3–4, pages 197–387.
- [7] Rodr´iguez-Rodr´iguez, I.; Rodr´iguez, J.-V.; Woo, W.L.; Wei, B.; Pardo-Quiles, D.- J. Comparison of Feature Selection and Forecasting Machine Learning Algorithms for Predicting Glycaemia in Type 1 Diabetes Mellitus. Appl. Sci. 2021, 11, 1742.
- [8] D. Sisodia, D.S. Sisodia Prediction of diabetes using classification algorithms Procedia Comput. Sci., 132 (2018), pp. 1578-1585
- [9] G. Swapna, K.P. Soman, R. Vinayakumar Automated detection of diabetes using CNN and CNN-LSTM network and heart rate signals Procedia Comput. Sci., 132 (2018), pp. 1253-
- [10] Kaur, H., & Kumari, V. (2019). Predictive modelling and analytics for diabetes using a machine learning approach. Applied Computing and Informatics
- [11] U. Ahmed et al., "Prediction Of Diabetes Empowered With Fused Machine Learning," in IEEE Access.
- [12] Eusuff, Muzaffar & Lansey, Kevin & Pasha, Fayzul. “Shuffled frog leaping algorithm:A memtic meta heuristic for discrete optimization”. Engineering Optimization - ENG OPTIMIZ. 38, 2006.
- [13] L. Haripriya, M.A.Jabbar. "A Survey on Neural Networks and Its Applications" International Journal of Engineering Research in Computer Science and Engineering, Vol 5, Issue 4, April 2018.
- [14] <https://arciv.ics.uci.edu/ml/machine-learning-databases/0059/>