EARLY PREDICATION OF DIABETES USING ARTIFICIAL NEURAL NETWORK

Prosanjeet Jyotirmay Sarkar[1*], Dr. Santosh Pawar[2]

Abstarct
Nowadays, the diabetes mellitus disease is increasing rapidly across the world. It is a chronic metabolic disorder, where patients unable to manage their blood glucose levels will lead to heart attack, blindness, renal disease, kidney problems etc. and that causes a lot of deaths per year. Patients need to pay a lot of money to practitioners and medicine for the maintenance of blood glucose level. So, it is necessary to develop a smart and intelligent system based on artificial neural network that can help network early detection of diabetes and be helpful to the practitioner for the early treatment so it can stop the progress of the disease. In this work, a dataset from Pima, India was taken to predict the possibility of diabetes. The objective of the paper was to minimize the error function in neural network training using a neural network model. The average error rate decreases during training and gave an accuracy of 95.93% for the predication of whether a person is diabetic or not.

Keyword:- Artificial Neural Network, Diabetes, Predication, Design, Blood glucose level.

[1]*Department of Electronics and Communication Engineering, Dr. A. P. J. Abdul Kalam University, Indore, Madhya Pradesh, India
[2]Department of Electronics and Communication Engineering, Dr. A. P. J. Abdul Kalam University, Indore, Madhya Pradesh, India

***Corresponding Author:** Prosanjeet Jyotirmay Sarkar
***Department of Electronics and Communication Engineering, Dr. A. P. J. Abdul Kalam University, Indore, Madhya Pradesh, India

DOI: - 10.48047/ecb/2023.12.si10.00250

Introduction

Many nationals at present are facing problems with diabetes mellitus. Diabetes mellitus is one of the untreatable chronic diseases that develop due to abnormal glucose in the blood. Diabetes mellitus is caused for two reasons: when the pancreases are unable to produce an adequate amount of insulin (lower blood glucose level) or the body cells cannot utilize the produced insulin (high blood glucose). Diabetes mellitus is caused due to the deficiency of pancreatic beta cells, resulting in insulin deficiency and latter consequences are the ineffective supply of insulin into the body. Insulin plays an important role in regulating the glucose level in the blood and if the body fails to utilize glucose to produce energy, it builds up in the blood, resulting in hyperglycemia. This can lead to life threatening complications such as heart attack, cardiovascular disease, ketoacidosis, eye blindness, nerve damage, adrenal gland tumour and other complications. There are three types of diabetes: Type-1 (T1D), Type-2 (T2D) and Type-3 (gestation). In the Type-1 diabetes chronic disease where pancreases stop the production of insulin due to which imbalance of glucose levels in the blood and approximately 8%- 10% of all diabetes cases in the worldwide. In Type-2 diabetes, there does not produce enough production of insulin to regulate the blood glucose level, which causes high blood glucose levels in the blood. These generally occur in people in the range of 25-70 years because of unhealthy diet, stressful lifestyle, no exercise etc. and there are approximately 90%- 95% of all diabetes cases worldwide. Type-3 is also known as gestational diabetes. It is generally found in pregnant women and can be cured by proper treatment and diet.

Diabetes mellitus is the world's third largest common disorder and, according to the World Health Organization (WHO), will be 643 million people worldwide suffering from diabetes by 2030 and it also estimates that diabetes is the cause of death for more than 1.5 million individuals per year. India is the second largest diabetes patient country, with approximately 77 million diabetes patients. doctors, medic, nutritionists and current research confirmed that there is no established cure for diabetes but, disease is discovered at an early stage, the chances for recovery will be greater. It can be controlled by regular treatment, proper nutrition and regular exercise.

In this research paper, we have developed an artificial neural network (ANN) model in order to correctly classify new as well as present instances of diabetes patients. The artificial neural network consists of many hidden layers, each hidden layer having a large number of neurons and all hidden layers are fully connected to each other. This present large number of neurons in the hidden layer gives a more accurate solution to the problem. In this research paper we used three hidden layer ANN with hyper parameter tuning, a model which gives a high practitioners a model which is beneficial and valuable to medical practitioners, clinicians and nutritioners for proper diagnosis of the patient.

RELATED WORK

There are several artificial intelligence and machine learning models developed for early predication of diabetes mellitus and some approaches are discussed below.

Rajiv Dey, Vaibhav Bajpai, Gagan Gandhi and Barnali Dey proposed a technique to predict diabetes mellitus using a back propagation algorithm of an artificial neural network (ANN). The ANN architecture used for tasks is a supervised multilayer feed-forward network with back propagation learning techniques. This technique improved the model classification accuracy by 92.5 %, which can be considered as a good result.

Suyash Srivastava, Lokesh Sharma and Vijeta Sharma et. al. proposed an artificial neural network model for predication of diabetes. This technique, we used Pima India diabetes dataset and speeded up the pre-processing by reconstructing the missing value by means of the column feature. The back propagation technique is used to update the weight and the model obtained an accuracy of 92%, which can be considered as a good result.

Rasool Jader and Sadegh Aminifar proposed a method for fast and accurate recognition of diabetes mellitus. In this paper, compare two and three hidden layer neural networks and compare the performance of both the models with epoch 100 and the low learning rate. They found that the three hidden layer neural network was high accuracy of 91% a which can be considered as a good result.

Nesreen Samer El_Jerjawi and Samy S. Abu-Naser proposed an an model for reliably detecting diabetes in patients. This recommended technique used 1004 cases over 8 input features and one output feature. An model consists of one input layer, 3 hidden layers and one output layer. All layers are fully connected to each other. The suggested approach would employ back propagation techniques to update the weight to reduce errors training errors and the model accuracy was 87.3 %.

Manaswini Pradhan and Dr. Ranjit kUmar Shau experimented and suggested a novel artificial neural network for predication of diabetes disease. In this paper, training with back propagation and genetic algorithms and classification accuracy compared. Found that the novel algorithm model performed better than the general model and model accuracy was 73.4 %.

The proposed method aims to avoid the patient going to going to a practitioner and hospital staff for various medical check-up like blood test, blood pressure, weight etc. This algorithm is simply designed for patients who can enter all physical condition parameters and track the diabetes stage.

3. Methods

The proposed model used early prediction of diabetes mellitus using deep learning algorithms and data mining, which is called artificial neural network. Following figure 1, a flowchart involves data collection, data pre-processing, and implementation of neural network and result analysis, which can be used to propose a model for early prediction of diabetes mellitus.

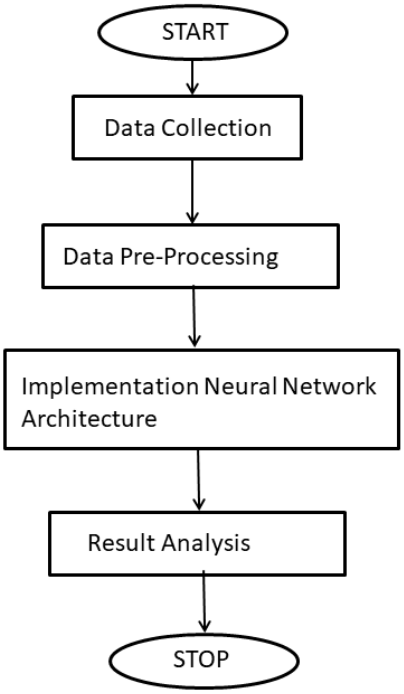

Figure 1:- Flow chart for proposed model

3.1 Data collection

In the modern world, the patient data has been stored is electronic health care equipment and available for research purposed in dataset format. The data that were used to train the model is the Pima India Diabetes dataset. In this dataset 768 samples, 8 feature attributes and one output attribute is a binary class distribution. Class 1 represents the diabetes test positive and class 0 represents the diabetes test negative. The negative class instances 500 i.e 65.1% of the dataset and the positive class instances 268 i.e. 34.9% of the dataset.

```
 #   Column                    Non-Null Count  Dtype
---  ------                    --------------  -----
 0   Pregnancies               768 non-null    int64
 1   Glucose                   768 non-null    int64
 2   BloodPressure             768 non-null    int64
 3   SkinThickness             768 non-null    int64
 4   Insulin                   768 non-null    int64
 5   BMI                       768 non-null    float64
 6   DiabetesPedigreeFunction  768 non-null    float64
 7   Age                       768 non-null    int64
 8   Outcome                   768 non-null    int64
dtypes: float64(2), int64(7)
memory usage: 54.1 KB
```

Figure 2:- Dataset all feature attributes

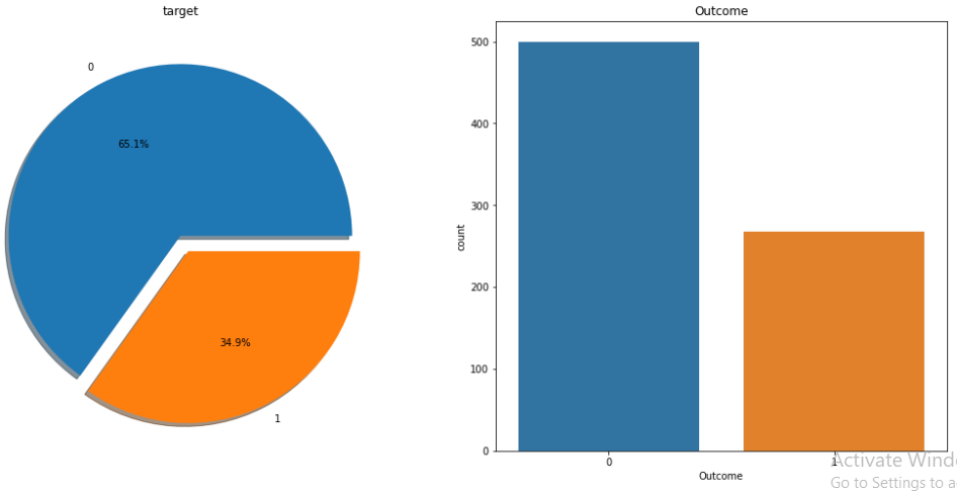

Figure 3:- number of diabetes positive and negative instances

3.2 Data Pre-Processing

In data pre-processing we have to do the following steps.

1. To remove redundant information features from the dataset. This will reduce the size of the dataset and increase the processing operation.
2. To check any missing value or NA in the feature dataset. If we have any missing value or NA, we can replace mean or mead in the same sample feature column.
3. Data normalization is used to change unintelligible data into understandable data. In the dataset feature, numeric values are in different ranges. To get the best accuracy from the model, we need to convert all features numeric values into a single range. The z-score method, which is used to randomize every feature value in the dataset. The mathematical formula for normalizing the z-score, mean and standard deviation can be written in equations 1, 2 and 3 respectively.

$$z = \frac{(Xi - Mean)^2}{SD} \quad (1)$$

Where,

Xi is the i^{th} sample instances

$$\text{Mean} = \frac{\sum_{i=1}^{N} Xi}{N} \quad (2)$$

$$\sqrt{\frac{\sum_{i=1}^{N}(Xi - Mean)^2}{N-1}} \quad (3)$$

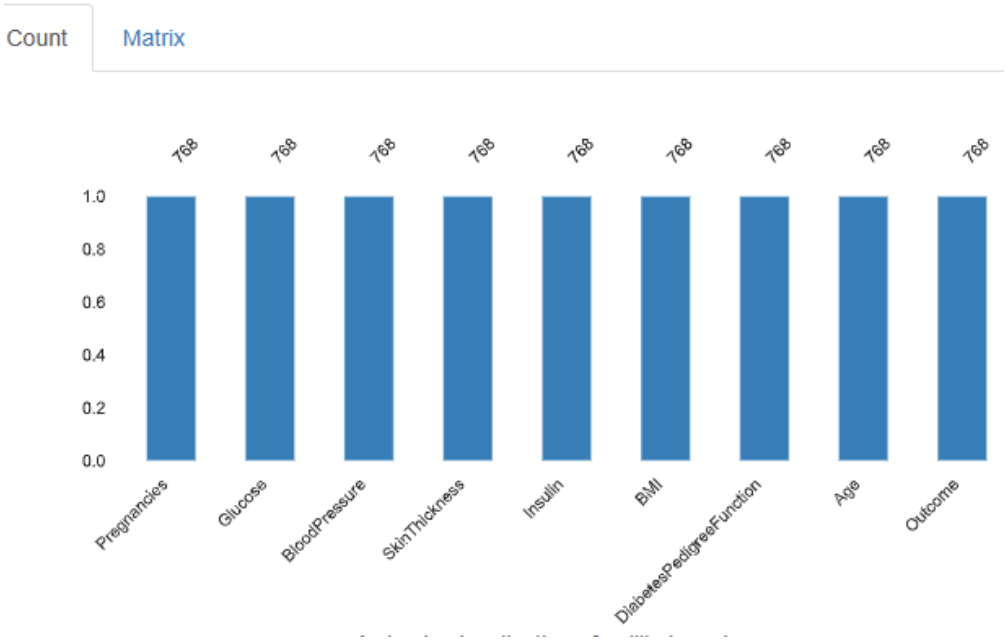

Figure 4:- Feature sample no missing value

3.3 Implementation of Artificial Neural Network

Artificial Neural Network model developed based on biological neural network and he mimics the human brain jobs. Artificial neural network consist of multilayer neurons and each layer neurons are interconnected to each other, which exchange the information. Each connection has numeric weight and bias that can be tuneable based on neuron error. This make neural network adaptive to inputs feature and capable of learning and predict the feature output.

In this we have experimentally recommended three hidden layer neural network and sigmoid activation function. If you increase the hidden layer than computation time and space is going to increase. We get the best outcome accuracy and fast computation when we used three hidden layer, activation function sigmoid, back propagation technique and tuning hyper parameter. The following figure 5 shows the neural network architecture.

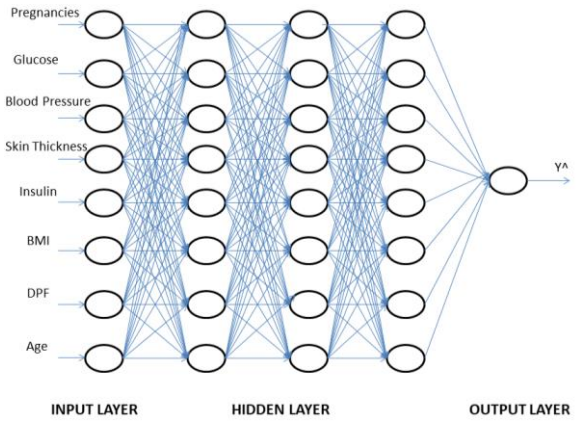

Figure 5:- Artificial Neural Network Architecture

In this artificial neural network we used back propagation technique and hyper parameter tuning for reduce the loss (error) and maximise the accuracy. The following algorithm used to design ann model.

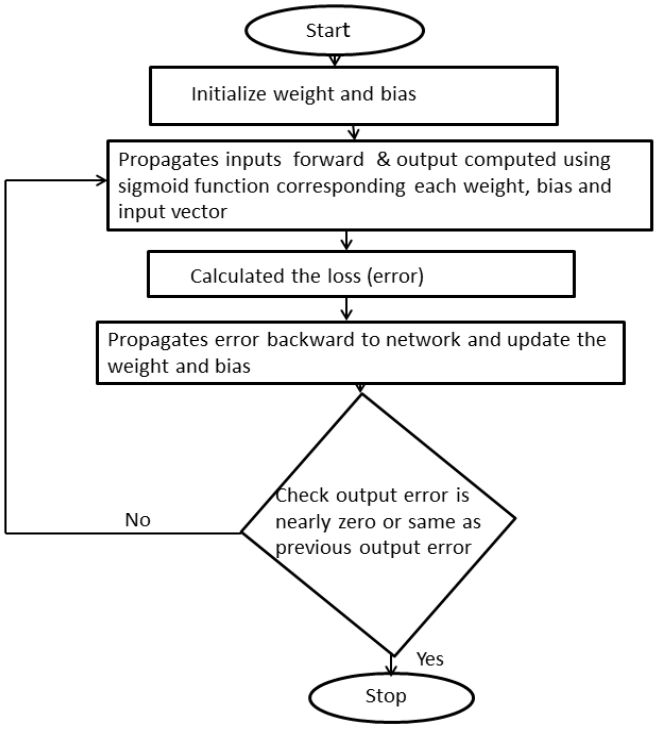

Figure 6:- ANN Algorithm design

Here we consider only one neuron for the understanding of operation of the model.

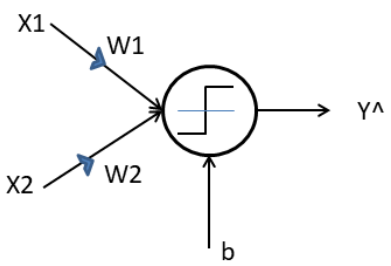

Figure 7:- Single neuron compute output corresponding weight, bias and input vector.

1. Update the weight and bias
Initially the network weight and neuron bias value to be initialize very small random number.

2. Propagate inputs to forward
The input training feature samples feed to the input layer of the network. Each layer neuron compute output based on input vector, weight and bias, pass the output to the next layer neurons. The net input computed to neuron in the layer is linear combination of the inputs and weight as shown mathematically in equation 4.

$$X = X1*W1+X2*W2+b \qquad (4)$$

We used sigmoid function or logistic function for activation of the neuron. Sigmoid function is used for the binary classification and mathematically represented as shown in equation 5.

$$Y\hat{} = \frac{1}{1+e^{-X}} \qquad (5)$$

3. Back propagate the error for updating the weight and bias
The output loss (error) is back propagating to update the weight and bias of the network for minimising the loss. Loss function mathematically shown in equation 6.

$$\text{L (MSE)} = \frac{\sum_{i=1}^{N}(Y-Y\hat{})^2}{N} \qquad (6)$$

Where,
Y^ is the predicated output of the neuron
Y is the actual output of the dataset.
Back propagation techniques used gradient descent method to search for a weight that fit the training data so as to minimise the mean squared error. The new weight update represented in equation 7.

$$\text{Wnew} = \text{Wold} - \partial \frac{dL}{dW\ old} \qquad (7)$$

Bias update represents the following equations 8.

$$\text{bnew} = \text{bold} - \partial \frac{dL}{db\ old} \qquad (8)$$

Where,
∂ is learning rate, helps to reach the global minima fast and also helps to avoid stuck at local minima.

3.4 Hyper-parameter tuning
Hyper-parameter tuning is used to get maximum accuracy of the model; following are the few parameters which are used for increase the performance of the model.

1. To change the learning rate of the model to reach maximum accuracy.
2. To increase the number of epoch the model will reach maximum accuracy.
3. To increase the number of batches the model will reach maximum accuracy.

3.5 Statistical tools for evaluation of performance
The confusion matrix is used to measure the performance of the various classification models.

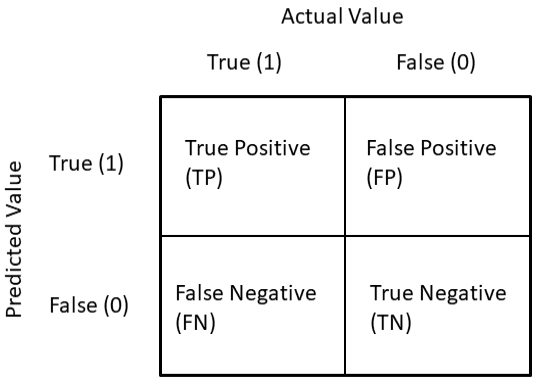

Figure 8:- Confusion Matrix

The confusion matrix is calculated following parameter.

1. Accuracy
Accuracy can be measured by divide the summation of TP and TN against the whole population. It measure weather model correctly represent number of positive or negative sample in dataset. The formula is shown in equation 9.

$$Accuracy = \frac{TP+TN}{TP+TN+FN+FP} \qquad (9)$$

2. Sensitivity
Sensitivity for the matrix can be calculated by taking summation of main diagonal to the true positive value. It is used to measure any disease in classification technique; it may be represented as equation 10.

$$Sensitivity = \frac{TP}{TP+FN} \qquad (10)$$

3. Specificity
It is used to measure the true negative rate and is given in equation 11.

$$Specificity = \frac{TN}{TN+FP} \qquad (11)$$

4. Precision
It is used to measure the number of correct positive result divided by the number of positive result predicated and is given in equation 12.

$$Precision = \frac{TP}{TP+FP} \qquad (12)$$

5. Error rate

It can be calculated by summation of FP and FN against the whole population. The formula is shown in equation 13.

$$Error\ rate = \frac{FP+FN}{TP+TN+FN+FP} \qquad (13)$$

6. F1- Score

It is used to measure the test accuracy, F1 score obtained by dividing the true positive against whole positive class value predicated. The range of F1 score is between [0-1]. It tells how precise your classification is as well as how robust it is. The mathematically is represented in equation 14.

$$F1 - Score = \frac{1}{\frac{1}{Precision}+\frac{1}{recall}} \qquad (14)$$

4. Result Analysis

The model was developed and trained on Anaconda Navigator studio and using popular frame work Tensor Flow open. The hardware that was used for the experiment was Dell laptop with Intel core i3 and 7th generation.

Confusion matrix of diabetes predication result using ANN is shown in figure 9.In this matrix is showing no false negative classification of the model which is best part of the model. Few are false positive classification and this person can go higher test for the confirmation of dieses.

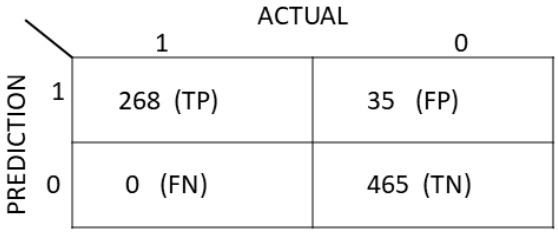

Figure 9:- Dataset classification result in confusion matrix

The following table 1, shows the performance of the evaluation matrix of the proposed model. The designed model was giving accuracy 95.93% which is consider best model for the early predication diabetes mellitus.

Accuracy	Sensitivity	Specificity	Precision	Error Rate	F1-Score
0.9593	1	0.93	0.8844	0.045	0.936

Table 1:- Evaluation matrix for the proposed model

5. Conclusion

Diabetes is an uncontrolled chronic disease and the majority of the people are affected. Uncontrolled disease can create many complications and hence early detection helps for the right treatment and proper management of diabetes. The main objective in this research paper is to find the diabetes at an early stage for the right treatment. In this study, it implemented a new ANN model with hyper-parameter tuning for early prediction of diabetes and found model accuracy is 95.93%. The proposed system will be helpful to medical practitioners and nutritionists for early treatment of diabetes.

6. References

1. Al-Zebari, A. and Sengur, A. (2019) 'Performance Comparison of Machine Learning Techniques on Diabetes Disease Detection', 1st International Informatics and Software Engineering Conference: Innovative Technologies for Digital Transformation, IISEC 2019 - Proceedings, pp. 2–5. doi: 10.1109/UBMYK48245.2019.8965542.
2. Aminifar, S., 2014. Design and implementation of fuzzy controllers for handling uncertainty in an industrial application (Doctoral dissertation, Universiti Sains Malaysia)
3. Alehegn, M., Joshi, R. and Alehegn, M. (2017) 'Analysis and prediction of diabetes diseases using machine learning algorithm: Ensemble approach.', International Research Journal of Engineering and Technology, 4(10), pp. 426–436. Available at: www.irjet.net.
4. D. K. Choubey, P. Kumar, S. Tripathi, and S. Kumar, "Performance evaluation of classification methods with PCA and PSO for diabetes," Network Modeling Analysis in Health Informatics and Bioinformatics, vol. 9, no. 1, Dec. 2019, doi: 10.1007/s13721-019-0210-8.
5. Conway, D. L. (2012) 'Gestational Diabetes Mellitus', Queenan's Management of High-Risk Pregnancy: An Evidence-Based Approach: Sixth Edition, 26, pp. 168– 173. doi: 10.1002/9781119963783.ch20.
6. Devi, M. Renuka, and J. Maria Shyla. "Analysis of Various Data Mining Techniques to Predict Diabetes Mellitus." International Journal of Applied Engineering Research 11.1, pp. 727-730, 2016.
7. Y. LeCun, Y. Bengio, G. E. Hinton, "Deep learning," Nature, vol. 521, pp. 436–444, May 2015.
8. R. Unnikrishnan and A. Misra, "Infections and diabetes: Risks and mitigation with reference

to India," Diabetes & Metabolic Syndrome: Clinical Research & Reviews, vol. 14, no. 6, pp. 1889–1894, Nov. 2020, doi: 10.1016/j.dsx.2020.09.022

9. Mujumdar, A. and Vaidehi, V. (2019) 'Diabetes Prediction using Machine Learning Algorithms', Procedia Computer Science, 165, pp. 292–299. doi: 10.1016/j.procs.2020.01.047.
10.] Saravana Kumar, N. M. et al. (2015) 'Predictive methodology for diabetic data analysis in big data', Procedia Computer Science, 50, pp. 203–208. doi: 10.1016/j.procs.2015.04.069.
11. Chung, J. and Teo, J. (2022) 'Mental Health Prediction Using Machine Learning: Taxonomy, Applications, and Challenges', Applied Computational Intelligence and Soft Computing. Edited by A. Minutolo, 2022, p. 9970363. doi: 10.1155/2022/9970363.
12. M. M. Islam, H. Iqbal, M. R. Haque, and M. K. Hasan, "Prediction of Breast Cancer using Support Vector Machine and K-Nearest Neighbors," in Proc. IEEE Region 10 Humanitarian Technology Conference (R10-HTC), pp 226-229, Dhaka, 2017.
13. M. M. Islam, H. Iqbal, M. R. Haque, and M. K. Hasan, "Prediction of Breast Cancer using Support Vector Machine and K-Nearest Neighbors," in Proc. IEEE Region 10 Humanitarian Technology Conference (R10-HTC), pp 226-229, Dhaka, 2017.
14.. Lekha and S. M, "Real-Time Non-Invasive Detection and Classification of Diabetes Using Modified Convolution Neural Network," IEEE Journal of Biomedical and Health Informatics, vol. 22, no. 5, pp. 1630–1636, Sep. 2018, doi: 10.1109/jbhi.2017.2757510.
15. Sonar, P. and Jaya Malini, K. (2019) 'Diabetes prediction using different machine learning approaches', Proceedings of the 3rd International Conference on Computing Methodologies and Communication, ICCMC 2019, (Iccmc), pp. 367–371. doi: 10.1109/ICCMC.2019.8819841.
16. Nagabushanam, P., George , S. T., S. Raveena Judie Dolly, D., Radha, S., "Artifact Cleaning of Motor Imagery EEG by Statistical Features Extraction using wavelet Families", International Journal of Circuit Theory and Applications, 2219- 41,Vol 48, Issue 12, 2020
17. L. B. Goncalves and M. M. Bernardes, "Inverted Hierarchical Neuro-Fuzzy BSP System: A Novel NeuroFuzzy Model for Pattern Classification and Rule Extraction in Databases," in IEEE Transactions on Systems, Man, and Cybernetics, vol. 36, no. 2, pp. 236-248, Mar. 2006.