JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

CLASSIFICATION OF DIABETES USING MLP CLASSIFIER

[1]Marisetti.Niharika , [2] Mohammad.Gousia , [3] Mulagada.Gowri Priya , [4] Ommi.Pavani Kalyani , [5]Rachapalli.Vasavi, [6]Dangety.Sowjanya

[1,2,3,4,5] B.Tech Student,[6] Faculty

[1,2,3,4,5,6] Department of Computer Science and Systems Engineering ,

[1,2,3,4,5,6] Andhra University College of Engineering for Women, Visakhapatnam,India

Abstract :Diabetes Mellitus is commonly known as Diabetes, a metabolic disease that is caused due to improper production of blood glucose, which is not produced by body on regular basis. Some of the major complications of diabetes are blindness, kidney damage and heart attack. This disease is identified as the second most common disease with around 77 million cases observed only in India, which makes it the second most affected in the World, after China. Some of the traditional approaches implemented by the health care professionals for predicting the diabetes are Fasting Plasma Glucose (FPG) Test, A1C Test and Random Plasma Glucose (RPG) Test, which are time taking procedures. Therefore to automate these processes, advanced technologies like Deep Learning and Artificial Intelligence are correlated with medical data to enhance this disease detection in Health Care Industry. The main objective of this model is to predict diabetes resulting in early detection of disease, ultimately decreasing the rate of diabetes patients. This model is developed using MLP classifier performing Binary Classification. The attributes in the dataset are considered from the FPG test reports which is primarily undergoing pre-processing technique. Later, the filtered dataset is loaded into the classifier to extract features used for further classification. The overall accuracy achieved by this model is about 77%.The classification reports of this model are to be displayed using Confusion Matrix and Data Visualization done using Bar Graphs and ROC Curve.

***Index Terms*–Diabetic Prediction, Deep Learning (DL), Fasting Plasma Glucose Test(FPG),A1C Test, Random Plasma Glucose Test(RPG),PIMA dataset,MLP Classifier, Binary Classification, Type 1 Diabetes, Type2 Diabetes, Gestational Diabetes,Confusion Matrix, Receiver Operator Characteristic (ROC).**

I. INTRODUCTION

Diabetes is a metabolic disorder characterized by high blood sugar levels observed over a prolonged period of time. Diabetes is the root cause of many cardiac and kidney related problems. It is caused due to the deficiency of insulin in the human body and is said to be a silent killer as it boosts other diseases in the human body. Diabetes is a chronic disease that depends on glucose levels in the blood. A decrease in glucose levels results in the development of insulin released by the pancreas. Inadequate creation of insulin causes diabetes that is categorized into three types i.e. Type 1 , Type 2 and Gestational diabetes. Previously, these are named as Insulin-Dependent and Non-Insulin-Dependent diabetes. In Type 1 Diabetes, the human body can't produce enough insulin whereas in Type 2 Diabetes, the human body can't create or utilize already created insulin. Gestational Diabetes is mostly observed during in women during pregnancy [1].Hence, to overcome this disease one should maintain a healthy diet and exercise which results in maintaining normalized blood sugar levels. The below Table 1 displays the readings of Diabetes, Prediabetes and Normal conditions of humans with respect to A1C, FPG and RPG test.

Table1: Ranges of Various Categories of Diabetes

	A1C (PERCENT)	FPG Test(mg/d)	RPG Test(mg/d)
DIABETES	6.5 or Above	126 or Above	200 or Above
PREDIABETES	5.7 to 6.4	100 to 125	140 to 199
NORMAL	Below 5.7	Below 100	Below 136

Therefore to automate the disease prediction, Machine Learning is implemented which is a subfield of Artificial Intelligence. It is defined as the capability of a machine to imitate intelligent human behavior. This approach can improve the understanding and specificity of disease recognition and also helps in reducing the expenses of laboratory services enforced on a common man.Various machine learning techniques have been implemented for the Diabetes classification. Most commonly used models are designed based on implementing Decision Tree,

Support Vector Machine, and Naive Bayes[2].Along with these techniques neural networks are also preferred for diabetes disease classification. A neural networks have been successfully applied in an extensive mixture of both supervised and unsupervised learning applications [3].In this Model, an MLP Classifier is used for the prediction of diabetes. Multilayer Perceptron (MLP) Network is one of the popular technique for classification problem. It can classify the disease by taking different inputs from the patient. In this Model, MLP Classifier is used for Binary classification .Proposed technique has been applied on PIMA diabetes dataset .In Section2 Related work is explained, Section3 Methodology is described and then the Results and Conclusion are explained in Section 4 and Section5.

II. Related Work

A lot of research has happened on the non-invasive automated detection of diabetes using Machine Learning and Deep learning techniques. In Machine learning, feature extraction, feature selection and classification are involved. In this paper[4], the author proposed a strategy for the diagnosis of diabetes using Deep Neural Network implementing five-fold and ten-fold cross validation technique. As a result, this developed system obtained an accuracy of 98.35%,F1 score of 98, and MCC of 97 for five-fold cross-validation. Additionally, accuracy of 97.11%, sensitivity of 96.25%, and specificity of 98.80% are obtained for ten-fold cross-validation and thus the proposed system provides promising results in case of five-fold cross-validation. In this journal [5], a useful and efficient prediction model is developed for diabetes detection using HRV signals. The author developed a new predictive model using a Convolution Neural Network (CNN), Long Short-Term Memory (LSTM) and extracted the characteristics from input data. Then Support Vector Machine(SVM) has been applied to those extracted characteristics for classifying the data with 95.7% of accuracy .This may vary while extracting dynamic characteristics from the input data. In the paper[6] a prediction model is proposed for diabetes identification using ANN (Artificial Neural Network) which is very useful for healthcare officials and practitioners. This model is developed using the ANN technique for minimizing the error function. So the average error function calculated was 0.01% and the accuracy attained through ANN was 87.3%. It was further observed that the performance of the above algorithms is not up to the acceptable level when compared to artificial intelligence problems like speech recognition and object recognition mainly because the dimension of the data handled is high. The limitation of machine learning boosted the deep learning research. In the journal [7],they used deep learning techniques to detect diabetes from the input HRV data with an accuracy value of 88%, which closely matches the maximum accuracy achieved for automated diabetes detection. In this proposed paper [8],they implemented a prediction framework for the diabetes mellitus using deep learning approach where the overfitting is diminished by using the dropout method. The system is applied to the Pima dataset and the highest accuracy obtained by the system is 88.41%.In the journal [9] a comparative study is presented between the Deep Neural Network (DNN) and several machine learning techniques utilized for disease prediction. These models are evaluated on various performance metrics such as accuracy, specificity, sensitivity, precision, and F1-score.This model attained high F1-score for the proposed DNN method with 99.75% accuracy.

III. Proposed Model

The Dataset considered in this model is obtained through FPG Test. It consists of 768 samples with 9 attributes of features. They are 8 input attributes which are categorized as pregnancies, Glucose, Blood pressure, Skin Thickness, insulin, BMI, Diabetes Pedigree Function and Age. It is used for Binary Classification i.e, Diabetes and Non Diabetes. The flowchart in Fig-1 describes the structure of the proposed model.

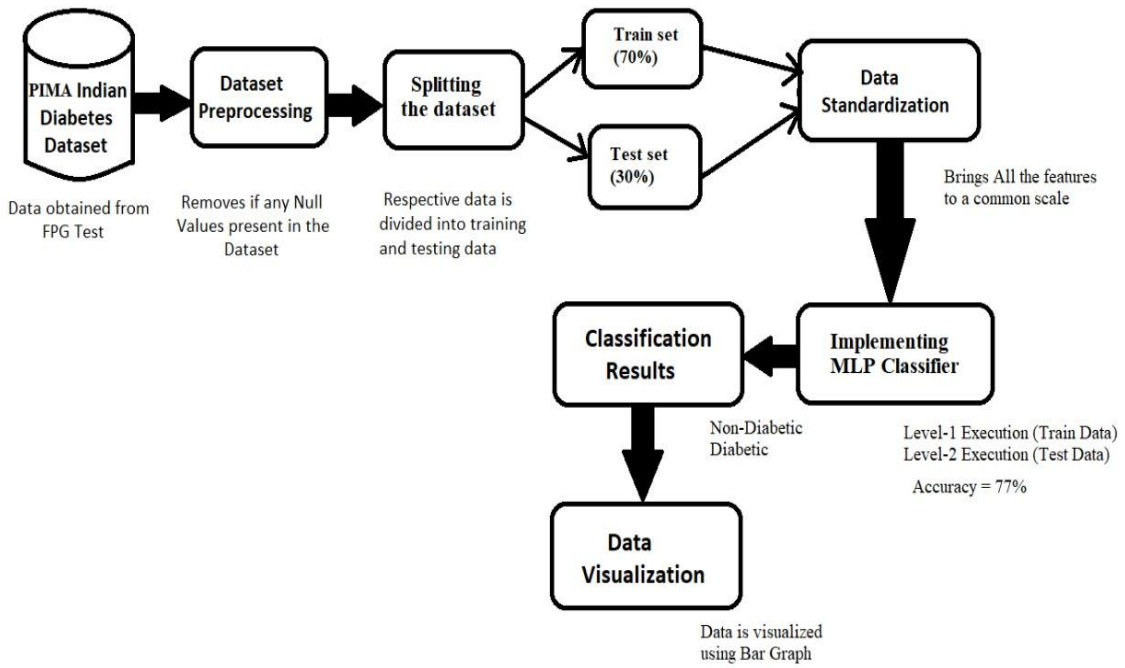

Fig- 1: Structure of Proposed Model

The initial stage is Data Preprocessing, it is used for removing the Unwanted/Unrelated data. Usually dataset consists of noise and missing values, therefore the dataset is filtered by applying preprocessing step that eliminates the redundant, empty, or any other ambiguous data. Then the preprocessed dataset is divided into 70% train dataset and 30% test dataset. 537 instances are taken as training and 231 instances have been taken for testing the dataset. The below Table 2 displays the dataset used for implementing this model.

Table 2: Dataset with Two-Level Classification

Classification	Train Set (70%)	Test Set (30%)
Diabetic	187	81
Non Diabetic	350	150

Data standardization is the process of rescaling the attributes so that they have mean of 0 and variance of 1.The ultimate goal to perform standardization is to bring down all the attributes to a common scale without altering the differences in the range of the values. Here in this model, we used Standard Scaler method to scale the dataset. Then we performed Hyper Parameter tuning for dataset with GridSearchCV to obtain the optimal parameters. Then model is implemented using MLP Classifier on the standardized data.

Multilayer Perceptron (MLP) is a deep learning method, it is one of the feed-forward neural network that generates a set of outputs from a set of inputs. It consists of at least three layers of nodes. They are input layer, a hidden layer and output layer. It is characterized by many layers of input nodes connected as a directed graph between the input layer and output layer. Apart from the input nodes,each node is an neuron that uses a nonlinear activation function. The Single Layer Perceptron (SLP), is only preferred to solve linearly separable problems so we opted for MLP Classifier. This technique has one or more hidden layers as described in Fig- 2. MLP is generally used for pattern recognition, classification of inputs and predicts the result. The below Fig 2 shows the neural network for diabetes prediction. The considered parameters taken as inputs and they are implemented in Neural network, finally the outcome will be given by the Output layer i.e., either 0 or 1.

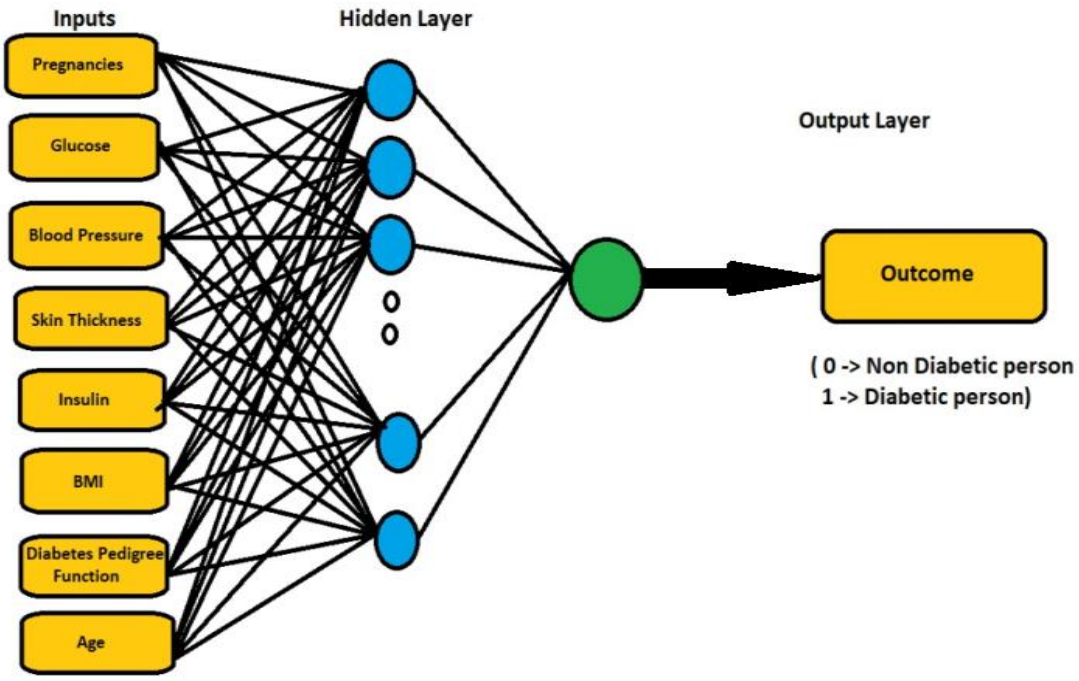

Fig- 2: Neural Network for Diabetic Prediction

Here, the dataset is trained by applying the MLP Classifier algorithm and passing the optimal parameters that are generated from the Hyper parameter Tuning process. Then the trained model is applied on the tested dataset to obtain the accuracy and the classification result. Now ,the obtained classification results are visualized using Confusion Matrix, Bar Graph and Receiver operator characteristic (ROC) Curve.

IV. RESULTS

Finally, the classification accuracy is measured by taking the ratio of the correctly classified case to the total number of cases. Classification result to be displayed using Confusion Matrix. A confusion matrix is a like a summary of the number of correct and incorrect predictions made by a classifier. It is used to evaluate the performance of a classification model by calculating the performance metrics like accuracy, precision, recall, and F1-score .This gives us a holistic view of our classification model. Precision is calculated by True Positive(TP)/(True Positive(TP)+False Positive(FP))formula, Recall is calculated using True Positive(TP)/ (True Positive(TP) + False Negative(FN))formula, F1 score is a weighted average score of the true positive (recall) and precision .

The performance of classification algorithm can be determined using confusion matrix. It is a table displaying different combinations of both predicted and actual values. The actual target values are compared with the values that are predicted by the model.The Confusion matrix is displayed in Fig-3.It is a 2 x 2 matrix with 4 values they are True Positive(TP) the prediction is positive and its actual value is true, True Negative(TN) the prediction is negative and its actual value is true, False Positive(FP) the prediction is positive and it's actual value is false, False Negative(FN) the prediction is negative.The Below Confusion matrix describes the classification result .We observed that there are 136 True Negative, 14 False Negative, 42 False Positive and 39 True Positive.

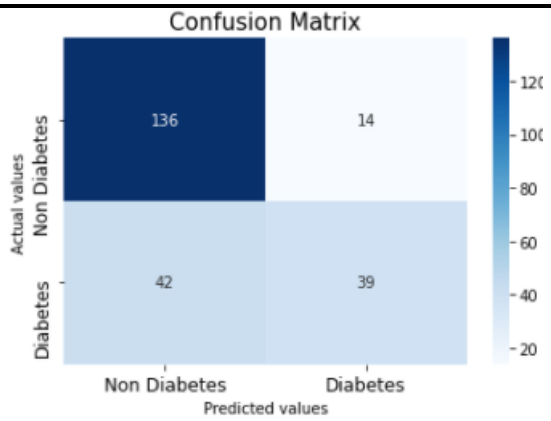

Fig- 3: Confusion matrix

Data Visualization done using Bar Graph. It is a graphical representation of data in which we can highlight the category with specific shapes like a rectangle. X- axis shows the specific categories being compared and Y-axis represents a measured values. Therefore, the visualized data is presented in the above displayed Fig. 4 classifying both Non Diabetes and Diabetes patients extracted from the Dataset.

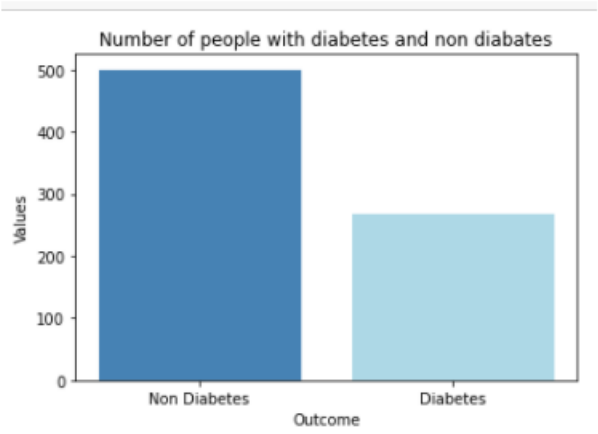

Fig 4: Outcome Image

In above bar graph, the steel blue represents count of Non Diabetes patients and sky-blue represents count of Diabetes patients. Total 768 patients are considered for this experiment, it is found that 268 are suffering with diabetes and the rest 500 cases are Non-Diabetic.

The Receiver Operator Characteristic (ROC) curve is graphical plot used to show diagnostic ability of binary classification problems. It is a probability curve that plots the TPR against FPR at various threshold values and essentially separates the 'signal' from the 'noise'.It is constructed by plotting the True Positive Rate(TPR) against the False Positive Rate(FPR). The true positive rate is proportion of observations that were correctly predicted to be positive out of all positive observations (TP/(TP+FN)).Similarly, the false positive rate is the proportion of observations that are incorrectly predicted to be positive out of all negative observations (FP/TN+FP)).

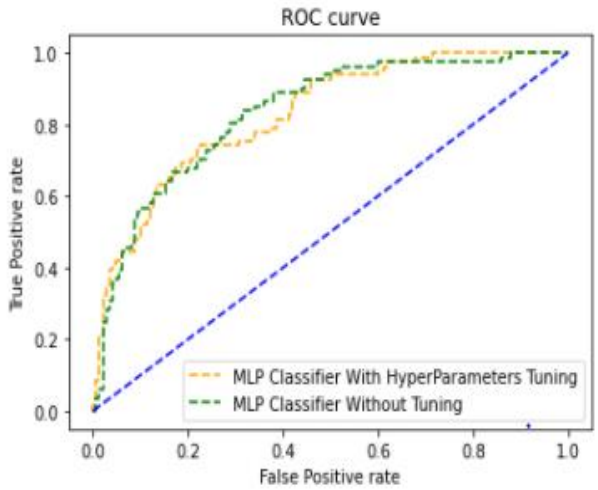

Fig- 5: ROC Curve

The above Fig-5 shows the difference between the accuracy and classification results of MLP Classifier. By applying the hyper parameters to the MLP Classifier, we observed that the classifier is well fitted with trained data and predicted the classification result when compared to MLP Classifier without hyper parameters.

V.CONCLUSION:

In recent years, tremendous research has been going on Diabetes Prediction as it is considered as a universal disorder. Diabetes is a serious life-threatening disease and has caused high mortality rate throughout the world. Therefore, automatic prediction of this disease is very much important for early diagnosis of this disease. Therefore in our proposed model, diabetes prediction is done using MLP classifier. In this model, features are extracted and a two-level binary classification is done i.e. either Diabetic or Non-Diabetic. This MLP classifier is implemented as it is considered as the most efficient and promising technique to analyze diabetes. The developed model has achieved an accuracy of about 77% displaying the classification report in the form of Confusion matrix associated with it's respective Bar Graph and ROC Curve. It is analyzed that if this model is implemented in Medical Industry, it may be supportive for medical professionals for automatic detection of this disease.

SCOPE: In future, the model can be tested on large set of diabetic dataset and performing comparative analysis for analyzing the efficiency of the model. Simultaneously, various deep learning techniques can also be adopted to develop an advanced model using data mining not only for identifying the disease but also the type of diabetes.

REFERENCES

[1] Bala Manoj Kumar P, Srinivasa Perumal R, Nadesh R K, Arivuselvan K,Type 2: Diabetes mellitus prediction using Deep Neural Networks classifier,International Journal of Cognitive Computing in Engineering,Volume 1,2020,Pages 55-61,ISSN 2666-3074,https://doi.org/10.1016/j.ijcce.2020.10.002.(https://www.sciencedirect.com/science/article/pii/S2666307420300073)

[2] Fregoso-Aparicio, L., Noguez, J., Montesinos, L. *et al.* Machine learning and deep learning predictive models for type 2 diabetes: a systematic review. Diabetol Metab Syndr **13,** 148 (2021). https://doi.org/10.1186/s13098-021-00767-9

[3] Mathew, Amitha & Arul, Amudha & Sivakumari, S.. (2021). Deep Learning Techniques: An Overview. 10.1007/978-981-15-3383-9_54.

[4] Safial Islam Ayon, Md. Milon Islam, "Diabetes Prediction: A Deep Learning Approach", International Journal of Information Engineering and Electronic Business(IJIEEB), Vol.11, No.2, pp. 21-27, 2019. DOI: 10.5815/ijieeb.2019.02.03[CrossRef] [Google Scholar]

[5] Swapna G, Vinayakumar R, Soman KP. Diabetes detection using deep learning algorithms.ICTExpress.2018;**4**(4):243–246. doi: 10.1016/j.icte.2018.10.005. [CrossRef] [Google Scholar]

[6] El-Jerjawi NS, Abu-Naser SS. Diabetes prediction using artificial neural network. International Journal of Advanced Science and Technology. 2018;**121**:55–64. doi: 10.14257/ijast.2018.121.05. [CrossRef] [Google Scholar]

[7] T. Zhu, K. Li, P. Herrero and P. Georgiou, "Deep Learning for Diabetes: A Systematic Review," in IEEE Journal of Biomedical and Health Informatics, vol. 25, no. 7, pp. 2744-2757, July 2021, doi: 10.1109/JBHI.2020.3040225 . [CrossRef] [Google Scholar]

[8]Ashiquzzaman, A. *et al.* (2018). Reduction of Overfitting in Diabetes Prediction Using Deep Learning Neural Network. In: Kim, K., Kim, H., Baek, N. (eds) IT Convergence and Security 2017. Lecture Notes in Electrical Engineering, vol 449. Springer, Singapore. https://doi.org/10.1007/978-981-10-6451-7_5[CrossRef] [Google Scholar]

[9] Beghriche, Tawfik & Mohamed, Djerioui & Youcef, Brik & Bilal, Attallah & Brahim Belhaouari, Samir. (2021). An Efficient Prediction System for Diabetes Disease Based on Deep Neural Network. Complexity. 2021. 1-14. 10.1155/2021/6053824. [CrossRef] [Google Scholar]