2023 3rd International Conference on Technological Advancements in Computational Sciences (ICTACS) | 979-8-3503-4233-8/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICTACS59847.2023.10390518

An Efficient Deep Learning Technique for Diabetes Classification and Prediction Based on Indian Diabetes Dataset

Prateeksha Singh[1], Dr. Sanjay Silakari[2], Dr. Shikha Agrawal[3]
[1,2,3]*Department of Computer Science & Engineering,*
University Institute of Technology, RGPV, Bhopal, India

***Abstract*—Diabetes is a chronic metabolic disorder that affects millions of people worldwide, and early diagnosis and effective prediction of this disease are crucial for its management and prevention of associated complications. Recently, deep_learning shows effective results in various healthcare applications, including disease classification and prediction. This paper presents diabetes classification and prediction using machine learning approach. Proposed system is increased the accuracy of the results by diabetes prediction framework using artificial neural network (ANN) deep learning algorithms. The performance of the overall classification results is improved as a consequence of this. The use of machine learning (ML) algorithms is currently effective for sickness detection.**

Keywords: ANN, Machine Learning, Diabetes Classification, Disease.

I. Introduction

A diabetic person's body has a hard time breaking down blood glucose, often known as blood sugar. Symptomatic of this illness is hyperglycemia, which may result from dysfunction in either insulin secretion or insulin action [1]. Insulin deficiency is the defining feature of type_1 diabetes, often known as T1D. The inability of a diabetic patient to make use of their own insulin is a major contributor to the growth of the illness. Type_2 diabetes (T2D) describes this condition. The incidence of both types of diabetes is rising, although more individuals are getting type 2 diabetes than type_1. Between 90 and 95% of people with diabetes have type_2.

The reason of this disease is insufficient dose of the insulin which is generated by body. Insulin is a important type o body hormone that use to inside glucose cells in body, it is used as energy source. Researchers have identified three distinct types of diabetes mellitus.

The reason of the Type 1 diabetes is to inability of pancreas'. Formerly referred to as juvenile diabetes or insulin-dependent diabetes mellitus, this kind of diabetes has now been renamed. The down of beta cells back to an autoimmune response. Adults are not immune to developing Type 1 diabetes, despite the fact that the majority of persons are diagnosed with the illness in childhood or adolescence. Insulin resistance, a condition in which cells do not respond appropriately to insulin, is the underlying cause of type 2 diabetes. Type 2 diabetes is characterized by insulin levels that are too high. Later on in the course of the disease, there is a possibility that insulin insufficiency may develop. The term "adult-onset diabetes" or "non-insulin-dependent diabetes mellitus" was once used to refer to this specific kind of diabetes. Despite the fact that older individuals make up the bulk of those who are diagnosed with type 2 diabetes, the increasing rates of childhood obesity have led to an increase in cases of type 2 diabetes among younger people. Unhealthy levels of body fat and insufficient physical activity are among the most prevalent risk factors.

The third risk factor for developing gestational diabetes is pregnancy in women who have never had diabetes before. This condition manifests itself in women having excessively high blood sugar levels during their pregnancy. Blood sugar levels in women who have gestational diabetes often return to normal within a short period of time following delivery of their babies. Women who have had gestational diabetes have a significantly increased risk of developing type 2 diabetes in the years to come.

Injections of insulin are the only treatment option for managing type 1 diabetes. A healthy lifestyle that emphasizes eating properly, exercising frequently, maintaining a healthy weight, and avoiding smoking may help prevent type 2 diabetes as well as manage the condition after it has developed. Oral antidiabetic medications are a potential treatment option for patients with type 2 diabetes. Insulin is also an option. The management of symptoms is the primary focus of treatment, which may include blood pressure control, foot care, and eye health maintenance. Hypoglycemia is the medical term for low blood sugar, which may be brought on by taking some oral medications or insulin. Those who are severely obese and have type 2 diabetes may discover that bariatric surgery is helpful in controlling their diabetes. After birth, the majority of women with gestational diabetes no longer have the condition.

Patients who have diabetes are at risk for developing a condition of the retina known as proliferative diabetic retinopathy (PDR). PDR may be identified by its

characteristic symptom, neovascularization, which describes a disease in which abnormal blood vessels form on the retina. If this issue is not detected and treated in a timely manner, it may result in complete and permanent loss of vision. Identifying neovascularization in fundus photos may be accomplished by the use of a variety of image processing algorithms, as recommended by a number of studies. Neovascularization is still difficult to spot due to the unpredictable growth pattern and the minute size of the blood vessels that it creates. As a result of their potential for autonomous feature extraction on objects with difficult attributes, deep learning algorithms are gaining favor in the field of neovascularization detection, which is driving this trend. Through the use of transfer learning, we provide a method for the detection of neovascularization in this particular research [1, 3]. Diabetic retinopathy, often known as diabetic retinopathy (DR), is the most common microvascular complication of diabetes. It may develop without any warning and can ultimately result in total blindness. Even though DR is quite frequent in today's society, it is still difficult to cure [4]. It is essential to do early screenings for the condition in order to reduce the risk of getting type 2 diabetes. There is a possibility that people whose medical profiles are one of a kind will not be excellent matches for the common criteria employed by the majority of prediction models for detection systems. As a consequence of these findings, a method for forecasting the start of type 2 diabetes based on factors that represent an individual's current state of health has been proposed as a result of this study.

II. Literature Survey

N. Bhaskar et al.,[1] presented an automated medical system that can determine if a person has type 2 diabetes based on their exhaled air. Because it includes many of the same gases that are dissolved in the blood, human breath may be utilized as a diagnostic sample for the purpose of diagnosing a wide variety of disorders. Analysis of exhaled breath stands out among the many non-invasive methods of detection due to the fact that it produces more accurate forecasts and offers a number of benefits.

G. Annuzzi et al.,[2] Diabetes type 1, often known as T1D, is an autoimmune condition that affects millions of individuals all over the globe. Patients with type 1 diabetes have a significant challenge when it comes to regulating their postprandial glucose response (PGR), which is accomplished by administering the appropriate amount of insulin bolus injections before meals.

S. T. Himi et al.,[3] presented a latest technology enabled watch to track the health issues and records.

P. Hu et al.,[4] presents the diabetes disease predicting framework based on the new method of machine and deep.

T. Zhu et al.,[5] obstacles include uncertain prediction confidence and a lack of training data for newly diagnosed individuals with type 1 diabetes. In order to address these clinical issues, we present a revolutionary deep learning architecture. In specifically, a recurrent neural network is used in order to train representations from input and then forward a weighted sum of hidden states.

M. Dodek et al.,[6] presented, individuals are able to offer up-to-date information on their health state (for example, data pertaining to insulin), and they are able to visit various healthcare facilities.

R. Marzouk et al.,[7] use of a patient's QR card as part of a diabetic monitoring system that connects patients, physicians, and other medical professionals to the Internet of Things is one option being considered.

Z. Ye et al.,[8] this research suggests that the E-Nose system, which is equipped with machine learning, is a technique that is both effective and accurate in terms of diagnosing diseases in a non-invasive and cost-effective manner.

M. Ismail et al.,[9] research is predicated on the theory that provide the maximum rate of prediction of the diabetes.

R. Tiwari et al.,[10] presented regression model based on the random forest and the decision tree for optimization of parameters from the selected dataset

V. Felizardo et al.,[11] presented a frame that mix a predictive frame with a reasoning model, with the goal of notifying of approaching events and interpreting the existing state in order to effectively advise the diabetic user. This design has the potential to improve the accuracy of diabetic advice.

III. Proposed Methodology

The following sub modules are used to describe the proposed technique:-

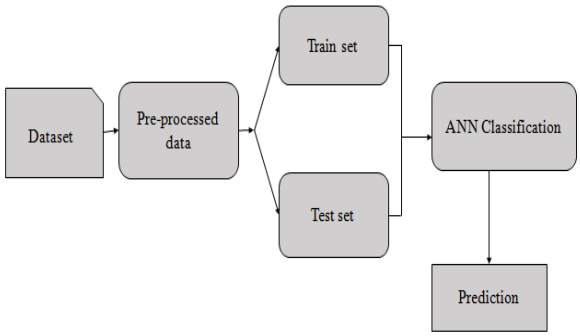

Fig. 1: Flow Chart

1. In the first step, get the diabetes prediction dataset from an online repository of publicly accessible data.
2. It is now time to do the preprocessing of the data, which entails removing null values and sending over any missing data.
3. Next, extract the data features and do an analysis using the dependent and independent variables.
4. At this point, put into practice the ANN classification approach, which is based on machine learning and deep learning.

5. At this point, you should construct a confusion matrix and display all of the classes that were predicted, such as true_positive, false_positive, true_negative, and false_negative.
6. At this point, you will need to determine the performance metrics by making use of the usual formulae in terms of the precision, recall, F_measure, accuracy, and error_rate.

A. Data Selection and Loading

- It is the process of choosing the proper data kind and source, as well as a suitable source to gather data.
- Data loading is the act of transferring data from one location to another.

B. Data Preprocessing

- It is the process of eliminating undesired data from the dataset.
- It covers missing data removal and encoding categorical data.
- It also manages the encoding of categorical data.

C. Splitting Dataset

- It is the process of separating accessible data into several categories.
- One component of the data is utilized to create a predictive model, while the other portion of the data is often used for purposes of cross-validation. And the other to assess the usefulness of the model.

D. Classification

Artificial Neural Networks (ANNs)- It is powerful technique consist structure and function of biological neural networks. They have been widely applied in various domains, including healthcare, for disease prediction and classification tasks.

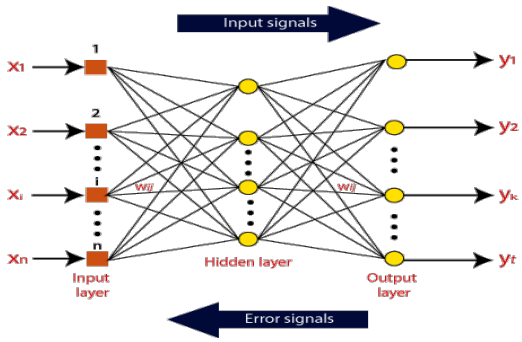

Fig. 2: Layers Representation of ANN

1) After Architecture of an Artificial Neural Network

A neuron is a unit which is used for the computing in the ANN, and these neurons are arranged into layers. In an ANN, the input_layer, hidden_layer(s), and output_layer(s) are the three primary kinds of layers. The input layer is where data is received, whereas the output layer is where the prediction or output is generated. The input is transformed by calculations performed by one or more hidden layers, depending on the model's architecture.

2) Neuron and Activation Function:

An ANN's neurons do a weighted sum of their inputs, then apply a non-linear modification through an activation function. Mathematically, for a neuron j in layer l, the weighted sum of inputs can be expressed as: $Z_j^l = \Sigma(w_{ij}^l * a_i^{(l-1)}) + b_j^l$ where w_{ij}^l represents the weight associated with the connection between neuron i in layer (l-1) and neuron j in layer l, $a_i^{(l-1)}$ is the activation output of neuron i in layer (l-1), and b_j^l is the bias term for neuron j in layer l.

The function of activation controls the range of values that the neuron's output may take and adds non-linearity.

3) Forward Propagation

Next, the activations signal is compute layer wise with starting of input layer and goes upto hidden levels until they reach the output layer, a process known as forward propagation. The projected values are at the output of the last layer.

4) Loss Function and Backpropagation

When training an ANN, the amount of error between the predicted and actual values is measured using a loss function. Mostly using loss functions used in diabetes_prediction include MSE and BCE loss.

The ANN's weights and biases are updated by a process called backpropagation, which takes into account the estimated loss. Iteratively adjusting the weights and biases is done by determining the loss gradient w.r.t. the network's parameters and then utilizing gradient descent optimization methods.

Prediction

This study improved the overall performance of the prediction findings, allowing for accurate prediction of the dataset's data. It was able to make an accurate prediction of the diabetes.

Algorithm:-

Input:- Informationset for Diabetes Prediction is the input to the algorithm.

Included are such basics as blood sugar and pressure, BMI, skin thickness, insulin, diabetes, family history, function, age, and result.

Output:- The highest possible values for the performance parameter

Step:- 1. The dataset is segmented into two parts: the train dataset and the test dataset, with the y and x train dataset and the similar variable y and x for test dataset.

2. Extractions_features, features = {} for diabetes_count: features [diabetes_count] = True

3. Optimize the prediction frame and splitting data

$y_{training}$

$X_{testing}$

4. artificial neural network classfier.

5. T_P, F_P, T_N, and F_N values generate through Confusion matrix.

6. Result calculation.

7. ROC generation.

Evaluation

Accuracy,precision, and recall are the three major metrics that are considered whenever a classification frame is being evaluated.

- The ratio of true_positives to total_positives is used to describe accuracy, whereas the ratio of positives to negatives is used to define recall.
- Accuracy equals [TP + TN] divided by [TP+TN+FP+FN];
- F1-Score equals 2 times (Precision x Recall) divided by (Precision + Recall);
- Classification Error equals 100-Accuracy.

Result Generation

The ultimate result will be created with the help of all of the classification and prediction that was done. The usefulness of the proposed approach is evaluated using a variety of measures, including accuracy, error rate, and others of a similar kind.

IV. Simulation and Results

Python software with Spyder IDE 3.7 is used in order to carry out the aforementioned computation while it is being carried out.

Fig. 3: Dataset

Figure 3 illustrates the dataset within the context of the Python environment. This figure also shows the dataset. The full dataset has 1547 rows and 8 columns in its entirety. The total number of rows and columns is 1547. There is a mention of the characteristic in each of the individual columns.

Fig. 4: Y Test

The y test of the dataset that was provided is shown in Figure 4. The provided dataset has a total of 1044 data that will be used for training and 115 data that will be used for testing.

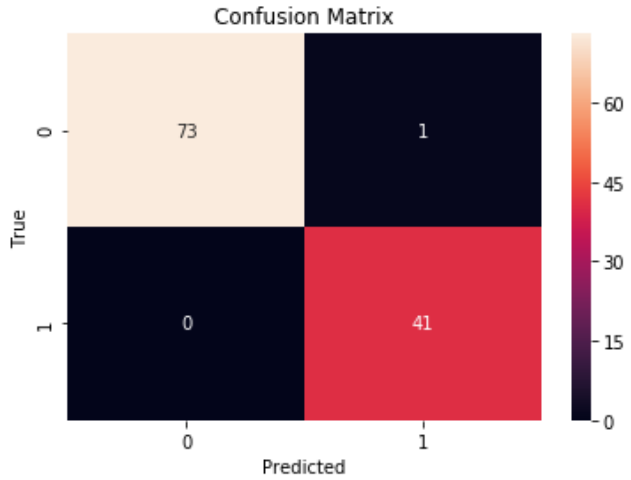

Fig. 5: Confusion Matrix of SVM

The confusion matrix of the ANN technique that has been developed can be seen in Figure 5. There is a value of 73 for a true positive, a value of 1 for a false positive, a value of 0 for a false negative, and a value of 41 for a genuine negative.

Table 1: Result Comparison

Sr. No.	Parameters	Previous_Work [1]	Proposed_Work
1	Method	Fused ML Decision	ANN
2	Accuracy	94.87 %	99.13 %
3	Error Rate	5.13 %	0.87 %
4	Sensitivity	95.52 %	100 %
5	Specificity	94.38 %	97.61 %

V. Conclusion

Diabetes is a term that is highly common in today's globe and is a significant problem in both developed and developing nations. The categorization and prediction of diabetes using a technique based on machine learning is presented in this research. Python Sypder 3.7 is the

program that is used to carry out the simulation. According to the simulated findings, the total accuracy reached by the suggested work is 99.13%, while the accuracy achieved by the prior work is only 94.87%. The proposed technique has an error rate of 0.87% when it comes to categorization, while the strategy that is currently being used has an error rate of 5.13%. The findings of the simulation make it plainly clear that the work that was proposed generates considerably better outcomes than the work that has previously been done. This is due to the fact that the simulation was run using the suggested work.

References

[1] N. Bhaskar, V. Bairagi, E. Boonchieng and M. V. Munot, "Automated Detection of Diabetes From Exhaled Human Breath Using Deep Hybrid Architecture," in IEEE Access, vol. 11, pp. 51712-51722, 2023, doi: 10.1109/ACCESS.2023.3278278.

[2] G. Annuzzi et al., "Impact of Nutritional Factors in Blood Glucose Prediction in Type 1 Diabetes Through Machine Learning," in IEEE Access, vol. 11, pp. 17104-17115, 2023, doi: 10.1109/ACCESS.2023.3244712.

[3] S. T. Himi, N. T. Monalisa, M. Whaiduzzaman, A. Barros and M. S. Uddin, "MedAi: A Smartwatch-Based Application Framework for the Prediction of Common Diseases Using Machine Learning," in IEEE Access, vol. 11, pp. 12342-12359, 2023, doi: 10.1109/ACCESS.2023.3236002.

[4] P. Hu et al., "Prediction of New-Onset Diabetes After Pancreatectomy With Subspace Clustering Based Multi-View Feature Selection," in IEEE Journal of Biomedical and Health Informatics, vol. 27, no. 3, pp. 1588-1599, March 2023, doi: 10.1109/JBHI.2022.3233402.

[5] T. Zhu, K. Li, P. Herrero and P. Georgiou, "Personalized Blood Glucose Prediction for Type 1 Diabetes Using Evidential Deep Learning and Meta-Learning," in IEEE Transactions on Biomedical Engineering, vol. 70, no. 1, pp. 193-204, Jan. 2023, doi: 10.1109/TBME.2022.3187703.

[6] M. Dodek, E. Miklovičová and M. Tárník, "Correlation Method for Identification of a Nonparametric Model of Type 1 Diabetes," in IEEE Access, vol. 10, pp. 106369-106385, 2022, doi: 10.1109/ACCESS.2022.3212435.

[7] R. Marzouk, A. S. Alluhaidan and S. A. El_Rahman, "An Analytical Predictive Models and Secure Web-Based Personalized Diabetes Monitoring System," in IEEE Access, vol. 10, pp. 105657-105673, 2022, doi: 10.1109/ACCESS.2022.3211264.

[8] Z. Ye, J. Wang, H. Hua, X. Zhou and Q. Li, "Precise Detection and Quantitative Prediction of Blood Glucose Level With an Electronic Nose System," in IEEE Sensors Journal, vol. 22, no. 13, pp. 12452-12459, 1 July1, 2022, doi: 10.1109/JSEN.2022.3178996.

[9] M. Ismail et al., "Radiomic Deformation and Textural Heterogeneity (R-DepTH) Descriptor to Characterize Tumor Field Effect: Application to Survival Prediction in Glioblastoma," in IEEE Transactions on Medical Imaging, vol. 41, no. 7, pp. 1764-1777, July 2022, doi: 10.1109/TMI.2022.3148780.

[10] R. Tiwari, R. Sharma, and R. Dubey, "Microstrip Patch Antenna Parameter Optimization Prediction Model using Machine Learning Techniques ", IJRITCC, vol. 10, no. 9, pp. 53–59, Sep. 2022. doi. org/10.17762/ijritcc.v10i9.5691.

[11] V. Felizardo, D. Machado, N. M. Garcia, N. Pombo and P. Brandão, "Hypoglycaemia Prediction Models With Auto Explanation," in IEEE Access, vol. 10, pp. 57930-57941, 2022, doi: 10.1109/ACCESS.2021.3117340.

