Detection of Diabetes Using Multilayer Perceptron

By Athira Kannamparambil Asakumar, Sid : 12612035, kannampara@coventry.ac.uk

& Megha Manoj, Sid : 12711693, manojm@uni.coventry.ac.uk

Abstract

Diabetes mellitus is a chronic condition that frequently results in severe complications. Diabetes is diagnosed and predicted using machine learning algorithms. The decision-making process for disease diagnosis and prognosis is significantly aided by learning-based algorithms. This paper investigates machine learning using neural networks and existing classification techniques for diabetes data. For the nearest K neighbours, Naive Bayes, complement shafts, decision shafts, radial base functions, and multilayer perceptron algorithms, we also assess various performance techniques with various characteristics. aids in the assessment of upcoming diabetic patients. The perceptron multilayer algorithm, which has an MSE value of 0.19 and the maximum prediction accuracy, is demonstrated in this work. MLP has the biggest area under the 86 percent curve and the fewest false positives and false negatives.

Keywords- multilayer perceptron, Diabetes detection, artificial neural network

1. Introduction

One of the aberrant metabolic illnesses is diabetes mellitus (also known as diabetes mellitus). elevation of blood sugar. When carbohydrates are digested, they are converted into glucose, which is then distributed throughout the body via the bloodstream. Hormone insulin, which aids in transporting glucose from the blood to cells. The pancreas generates little to no insulin in this chronic disease, and occasionally the insulin that is generated is rejected by the cells. Insulin resistance is what causes this.

Diabetes is currently regarded as one of the most deadly diseases in the world, and it affects a large population. There are 422 million instances of diabetes worldwide, and diabetes is responsible for 1.6 million fatalities annually. The frequency and number of cases of diabetes have been continuously increasing during the past several decades.

DM can be classified as type 1, type 2, or gestational. When the pancreas produces little or no insulin, it has Type 1 diabetes. Type 2 diabetes is caused by an inability of cells to produce or absorb insulin (T2D). To name just a few of the diseases that can be aggravated or exacerbated by pregnancy-related high blood sugar or sugar levels: hearing loss and dementia; heart and stroke disease; stroke; depression; blindness; retinopathy; and neuropathy. Increases the risk of such complications. Early detection of the disease plays an important role in diagnosing the disease. It is one of the main causes of cardiovascular disease and requires a lot of medical guidance when making decisions. A wide range of machine learning approaches have been employed by numerous researchers with various medical diagnoses.

Most experimenters pay attention to medical expert systems, and this area has undergone significant thought. Constant collaboration between data judges and medical professionals improves the accuracy and practicality of this approach. Recent World Health Organization inspections revealed a sharp rise in diabetic patients, with diabetes being blamed for every single death. In light of this, experimenters and medical interpreters are quite concerned about early diabetes diagnosis. Many computer-based discovery techniques for assessing and predicting diabetes have been developed and laid out. Normal diabetes relationship building takes time. This challenge can be solved, though, with the advancement of machine literacy.

To accurately predict a patient's condition, a model that can accurately detect diabetes using input features is needed. It is possible to increase the effectiveness of opinion with a solid model and precise discovery methods. Medical interpreters might imagine biomedical opinion using engineering techniques that can automatically adapt to any unanticipated future conditions based on their interpretations of the immunisation. Planning and provisioning may benefit greatly from a long-term vaticination algorithm. Intelligence systems can adapt to new stimuli and changes in functional connectivity by learning, adapting, and adjusting their functional dependencies.

2. Methodology

2.1 Proposed Method

Multi-layered Perceptron (MLP) network is the commonly used and simple technique of neural network. Back propagation learning technique is used as classification method of instances in Multi-layered Perceptron (MLP) network. Below Figure describes the proposed work.

Figure 1
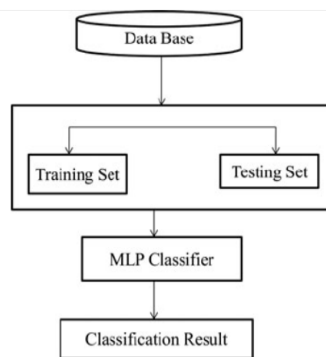
2.2 Multilayer Perceptron (MLP)

The multilayer perceptron is widely regarded as one of the most effective classification algorithms for feed-forward neural networks. It is composed of several distinct layers. Single-Layer Perceptron (SLP) is able to find solutions for issues that are linearly separable, but it is unable to do so for nonlinear problems. MLP is utilised in order to address these challenging problems. A feed-forward neural network is the name given to this method. A feed-forward neural network, as shown in Fig. 2, contains one or more hidden layers. MLP is frequently used to classify input patterns, reorganise input patterns, and anticipate the outcome.

Figure 2 MLP architecture
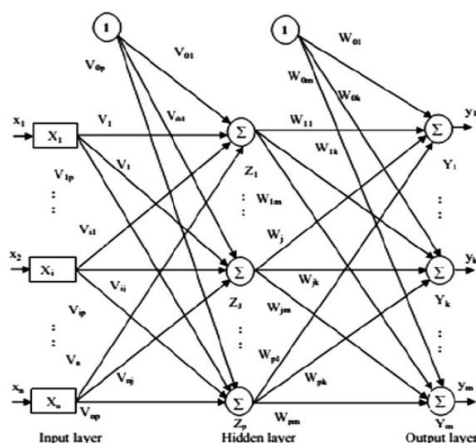
It's possible to use random selection for the weights before the MLP network is trained. The training set, in this case a set of tuples, is subsequently used by the neurons to conduct their study (x1,x2,t).

The network's expected output is t, and its inputs are x1 and x2. The weighted sum of all of the neuron's outputs determines whether y is the neuron's actual output and can be represented as follows:
y = x1w1 + x2w2 --------------------(1)

The network has a nonlinear activation function in its hidden layer. The network's output can be expressed as follows:

X = f(s) = W $\emptyset$ (As + p) +b

where X may be an output vector and s may be an input vector.

A and p are the vectors of the first layer. This first layer's bias vector and weight matrix, p, are referred to here. Bis is the bias vector and W is the weight matrix for the second layer.

The nonlinear component of the second layer, whose output is not visible in the output but is related to the inputs of additional neurons in the hidden layer, has a weight matrix of W and a bias vector of bis.
$\emptyset$ is the nonlinear element. The hidden layer's output, which is not visible, is coupled to the inputs of additional neurons. The output is designated as 0, where 0 denotes non-diabetes and 1 denotes diabetes.

3. Experiment Result

3.1 Data Pre-processing and Cleaning

The National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK) generously provided the diabetes dataset that was used for this work. The purpose of this endeavour is to completely and unerringly base a diagnosis of diabetes on a patient on the basis of diagnostic characteristics. The dataset under examination was cleaned utilising data cleaning and pre-processing techniques and the resulting dataset was employed in a variety of experiments using different classification algorithms.
The database includes information about various diabetes patient types and their status (Non-Diabetes and Diabetes). These patients' data are analysed and utilised to forecast diabetes in the general population.
768 records make up the chosen Diabetes dataset. The characteristics of the patients in this dataset include: 1) the frequency with which they become pregnant, 2) The plasma glucose level, blood pressure in the diastole, 4) Triceps skinfold thickness, in millimetres 5) Insulin 2-hour serum, Body mass index 7) Pedigree function for diabetes Age in years, and a class variable, are both 8. (0 or 1)

Table .1 - cleaned sample record dataset

| Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedig	Age	Outcome
6	148	72	35	0	33.6	0.627	50	1
1	85	66	29	0	26.6	0.351	31	0
8	183	64	0	0	23.3	0.672	32	1
1	89	66	23	94	28.1	0.167	21	0
0	137	40	35	168	43.1	2.288	33	1
5	116	74	0	0	25.6	0.201	30	0
3	78	50	32	88	31	0.248	26	1
10	115	0	0	0	35.3	0.134	29	0
2	197	70	45	543	30.5	0.158	53	1
8	125	96	0	0	0	0.232	54	1
4	110	92	0	0	37.6	0.191	30	0
10	168	74	0	0	38	0.537	34	1
10	139	80	0	0	27.1	1.441	57	0
1	189	60	23	846	30.1	0.398	59	1
5	166	72	19	175	25.8	0.587	51	1
7	100	0	0	0	30	0.484	32	1

It is possible to eliminate missing and anomalous data values from a dataset using the imputation-mean approach. The pre-processed dataset contains 722 records with the three necessary pertinent patient information attributes. In the dataset, there are 248 cases of "diabetes" and 474 cases of "non-diabetes," with 46 records lacking the necessary fundamental parameters.

The input attributes are from the dataset and are drawn from six different features, while the output attribute is drawn from just one of the features. A table.1 that contains information on the patient is presented here.

Figure 3 Visualisation of dataset
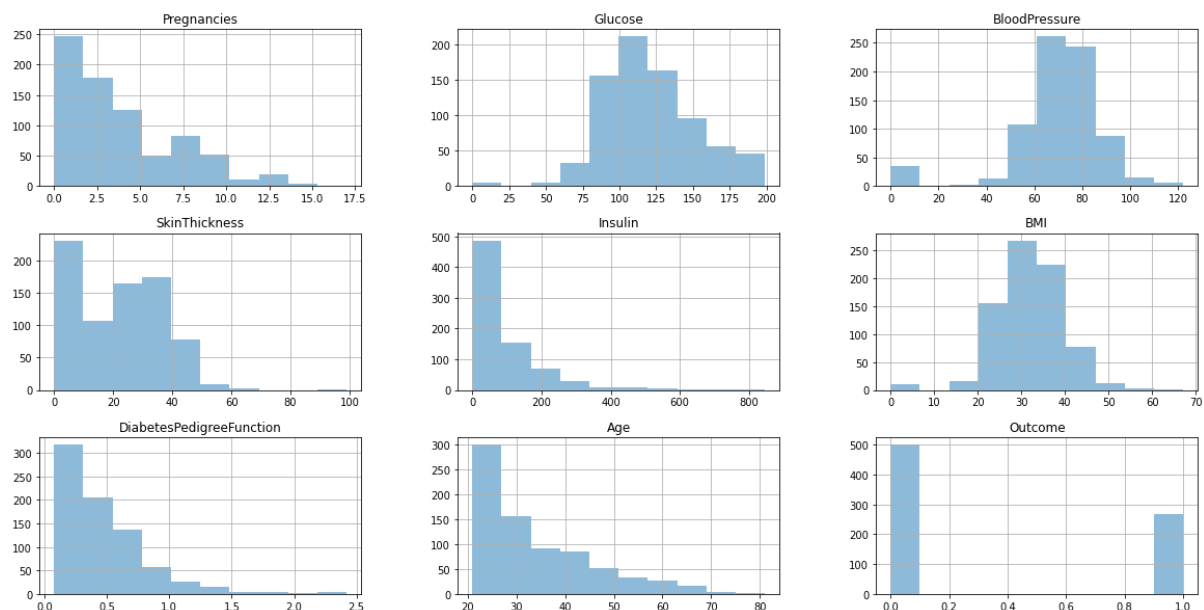
Figure 4 Distribution of dataset
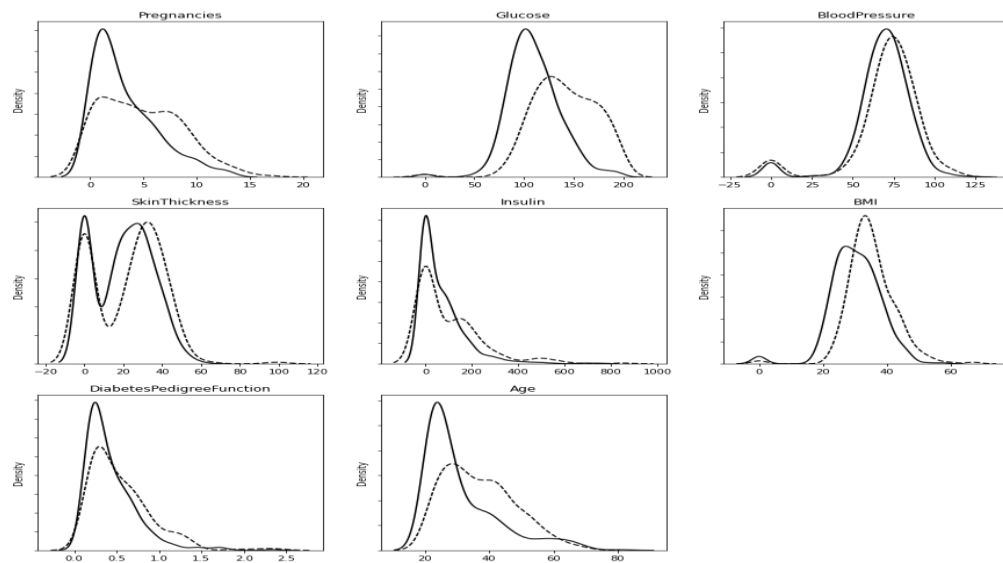
Standardizing the dataset is the next stage. The process of transforming numerical variables so that each has a mean of 0 and a variance of 1 is known as data normalisation.

The measures of insulin and diabetes pedigree function are very different. Insulin has a maximum value of 846 while diabetes pedigree function has a maximum value of just 2.42.

Data is divided into training records, which make up 80%, and testing records, which make up 20%. 218 instances have been taken for testing, while 550 instances have been used for training. The eight feature attributes and a class attribute are both present in both data sets.

There are several steps: first, the original recordings are divided into training (80%) and testing (20 percent). The second division of the training set is into the training set, validation set, and testing set. Data splitting must be done at randomThe dataset should first be split into X (input features) and Y (output features) (target variable).

Due to the eight columns in X train, the first hidden layer will contain 32 nodes and an input dimension of 8. For training, there are 3 different parameters. metrics, a loss function, and an optimizer. Here, the well-liked Keras optimizer Adam is being used. The adam optimizer will function well for the majority of datasets with little adjustment. Since the topic at hand is a binary classification problem, we will utilise accuracy (i.e., the percentage of properly classified samples) as our evaluation metric and binary crossentropy as our loss function.

The following assessment criteria were employed to gauge classifier performance: true positive (TP), true negative (TN), false positive (FP), and false negative (FN).
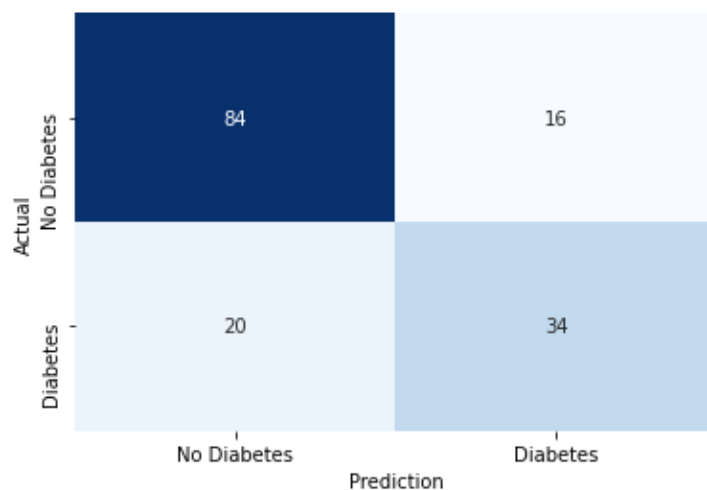
True positive rate (TPR) = True positive / (True positive + false negative)

False positive Rate (FPR) = False positive / (True Negative + False Positive)

Confusion Matrix:

Confusion matrix is defined by four key terms. Performance measurements are calculated using these four factors.

Figure 5

True negative: The real class is negative (has no diabetes), as predicted by the model

False positive : Diabetes is not present in the actual class, despite what the model anticipated (diabetes)

False negative: Although the model anticipated diabetes, the actial class is positive (no diabetes)

True positive: The model predicted positive results, but the actual class is positive (diabetes) (diabetes)

After calculating accuracy got Training Accuracy as 93.89% and Testing Accuracy as 76.62%.

Result
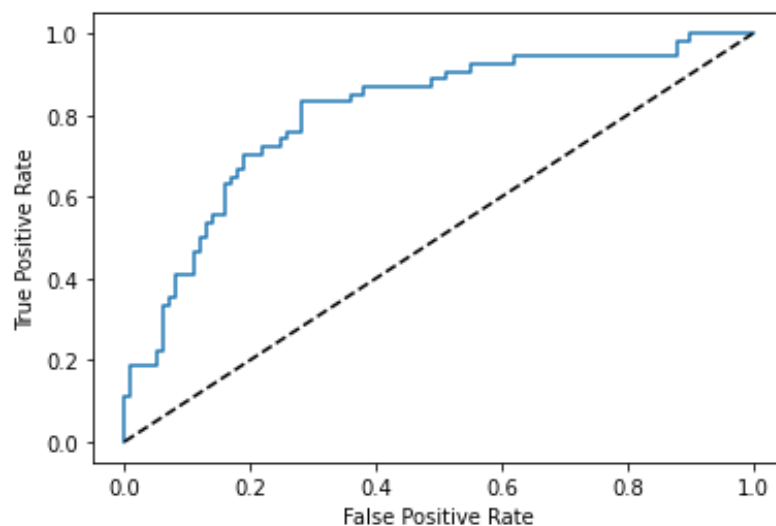
4. Conclusion

On the basis of diabetic statistics, the classification result is efficient despite being of less quantity. Additionally, unique tools could be explored for higher accuracy. It is important to classify this disease accurately in order to help doctors and patients diagnose it effectively. When it comes to data mining and huge data, MLP has proven to be a useful technique for data analysis. It has the ability to handle a variety of challenging tasks in artificial intelligence research. For patients with diabetes, the MLP technique has been used in this paper. The Kaggle data collection is provided for neural network training. In the future, more models might be utilised to get results that are even more precise, and optimization techniques might be applied for comparison.

5. Appendix

The python codes are available in the following link MLP_Diabetes_project

6. Bibliography

1. Lebovitz, H. E. (1999). Type 2 diabetes: an overview. Clinical chemistry, 45(8) - 1339-1345
2. Haykin, S. (2009) Neural Network and Learning Machines. New Jersey, USA: Pearson Education.
3. R. D. Canlas, "Data mining in healthcare: Current applications and issues. School of Information Systems & Management", Carnegie Mellon University, Australia, 2009.
4. https://www.researchgate.net/publication/327545512_Detection_of_Diabetes_Using_Multilayer_Perceptron_Proceedings_of_ICICA_2018