APPLICATION OF ARTIFICIAL NEURAL NETWORKS FOR EARLY DETECTION OF DIABETES MELLITUS: INSIGHTS FROM A CASE STUDY IN KAURA NAMODA NIGERIA

Abstract

Diabetes Mellitus (DM) is a chronic disorder which needs urgent attention. Detection of the disease from the grass root using patient risk factors is the key for early prevention of the disease. The aim of this paper is to use Artificial Neural Network (ANN) for early detection of DM. Datasets were collected from past patient records of patients suffering from DM in General Hospital Kaura Namoda, Nigeria between 2019 to 2023. The hospital and region are selected because most of the cases of DM were reported in the hospital and the prevalence of the disease in the region is about 8% to 10%. The datasets consists of two sample size of 400 patients each, the first sample dataset consist of 400 patients with demographic, clinical and lifestyle risk factors and second sample dataset consist of 400 patients with demographic, clinical, lifestyle and dietary risk factors. Backward stepwise feature selection method was employed to eliminate the least informative features and the method retained Six (6) risk factors for the first sample dataset age, family history of DM, blood glucose level, blood pressure level, body mass index, physical activity, and removes one risk factor sex. For the second sample dataset, the method retained twelve (12) risk factors age, family history of DM, blood Glucose level, blood pressure, body mass index, physical activity, preference for sweet food, red meats, refined carbs, energy drinks, white rice, processed meat and remove sex and preference for salty food. The results of the analysis showed that MLPNN model demonstrated high accuracy in detecting DM and non-DM patients, with improved performance when dietary risk factors were included. The paper concludes that in order to detect DM and Non-DM accurately, dietary risk factors must be included apart from demographic, clinical and lifestyle risk factors.

Key word: Diabetes Mellitus, Backpropagation, Detection, Artificial Neural Network, Receiver Operating Characteristic Curve.

1. Introduction

Diabetes Mellitus (DM) is a chronic disorder characterized by abnormally high levels of sugar (glucose) in the blood. For people with DM, blood sugar levels remain high and this might be because insulin is not being produced at all or is not made at sufficient level or not as effective as it should be. DM affects more than 300 million people world-wide (WHO, 2022). In 2016, it was discovered that 1 in 5 people of age 50 years and the above had DM. The highest prevalence (17.9%) was found in America Indians and Alaska natives. DM cases increases rapidly all over the world but it was severe in African continent and less in European continent due to their tireless struggle to fight against the disease (WHO, 2022). The misconception that DM is “a disease of the wealthy” is still held by some people around the world; but the evidence published in the Diabetes Atlas of the International Diabetes Federation (IDF, 2021) disproves the assertion. 80% of people with DM live in low and middle-income countries and socially disadvantaged countries are the most vulnerable to the disease. DM was becoming rampant in countries in the Middle East, Western Pacific, South-East Asia where economic development has transformed lifestyles of the people and these rapid transitions bringing previously unheard rates of obesity and DM in the area. Developing countries were facing a firestorm of ill health with inadequate resources to protect their population. Thus, it is necessary to increase awareness of the importance of a healthful diet and physical activity, especially for children and adolescents. Conducive environments have to be created that lay the foundations for healthy living (NIH, 2021).

Nigeria has the largest population in Africa of more than 220 million and out of this adult population aged 20–79 years, is approximately 140 million. One third of all the cases of DM are in the rural communities, while the rest are in the urban centres. About 5 million of the cases of DM in Nigeria are undiagnosed, deaths related to DM in Nigeria in 2023 were estimated to be two hundred and fifteen thousand one hundred and thirty seven (215, 137) and the current prevalence of DM in Nigeria is roughly from 8% to 10%. Of the four classes of DM, two types are frequently found in Nigeria and these are type 1 DM and type 2 DM. Also, among the two, type 2 DM is the most common and accounts for about 90% to 95% of all cases of DM. The prevalence of type 1 DM is not known but there are few reports from various part of Nigeria its prevalence range from 0.1/1000 to 3.1/ 10000 and 1 out of every 17 adults are having the disease, National Institute of Health (NIH, 2021). Moreover, the pooled prevalence of DM in the six geopolitical zones of Nigeria were 3.0% in the North- West, 5.9 in the North-East, 3.8% in the North- Central, 5.5% in the South-West, 4.6 % in the South – East and 9.8% in the South-South (NIH, 2021).

Today many techniques have been developed for data mining, and there is an art to selecting and applying the best method for a particular situation. Methods for analyzing data can be divided into two groups: supervised learning and unsupervised learning. Supervised learning requires input data that has both independent variables or input variables and a dependent variable or output variable whose value is to be estimated. By various means, the process learns how to predict the value of the output variable based on the input variables. Decision Trees (DT), Regression Analysis (RA) and Artificial Neural Networks (ANNs) are examples of supervised learning. Unsupervised learning does not identify output variable, but rather treats all of the variables equally. In this case, the goal is not to predict the value of a variable but rather to look for patterns, groupings or other ways to characterize the data that may lead to understanding of the way the data interrelates. Cluster Analysis (CA), Correlation, Factor Analysis (FA), Principal Components Analysis (PCA) and statistical measures are examples of unsupervised learning (Bellazi and Zupan, 2011; Al-Shaye, 2011).

Bellazi and Zupan (2011), ANNs are popular data mining tool used to build complicated models. Basically an Artificial Neural Network Model contains three layers: input layer, intermediate hidden layer and output layer. Also, each layer being made up of nodes (neurons) and links. The nodes in input layer are viewed as predicted variables whereas the nodes in output layer are analyzed as the outcome variables. The paper used a popular ANN Architecture called Multilayer Perceptron Neural Network (MLPNN) with back- propagation (i.e. Supervised Learning Algorithm) which is arguably the most commonly used and well – studied ANN architecture. MLPNN is feed- forward neural network trained with the standard back-propagation algorithm and they are known to be a powerful function approximator for prediction and classification problems (Xue-Hui Meng *et al.,* 2011). Artificial Neural Network provides a general way of approaching problems. When the output of the network is categorical it is performing prediction and when the output has discrete values it is doing classification (Al-Shaye, 2011). The paper reviewed work on ANN for prediction of Diabetes such as Sahu and Mantri (2023) used MLPNN model for prediction of Diabetes using demographic and clinical risk factors in the face of inconsistent results, gaps and data class imbalance. The model achieved prediction accuracy of 84% relative to baseline. The work of Chen *et al.* (2024) observed that ANNs trained using risk factors had better efficacy and facilitate the reduction of harm caused by type 2 DM combined with Hyperuricaemia. Bukhari *et al*. (2021) used demographic, clinical and lifestyle risk factors to train Artificial Backpropagation Stochastic Gradient Neural Network (ABPSCGNN) algorithm for prediction of Diabetes patients, the ABPSCGNN model achieved 93% prediction accuracy. Also Pradhan *et al.* (2020) applied MLPNN model for prediction of Diabetes patients using nine (9) features. The model had 85.09% prediction accuracy. Moreover, the work of Setiawan *et al.* (2024) focused on Neural Network model for prediction of Diabetes patients using clinical data. The result obtained showed that the model had accuracy of 97% and this demonstrates the ability of the model in predicting diabetes patients. Zou *et al.* (2018) used Decision Tree (DT), Random Forest (RF) and Neural Network to predict DM using demographic and clinical risk factors. Their results showed that RF had the best accuracy of 80.8%. Furthermore, Evwiekpaefe and Abdulkadir (2023) developed three (3) ML models namey K-Nearest Neighbour (K-NN), DT and Artificial Neural Networks (ANNs) to predict DM in individuals at an early stage. Their work identified nine (9) clinical and demographic risk factors were responsible for DM. In the other hand, Farooqui *et al*. (2023) used clinical, demographic and lifestyle risk factors and built four ML models [DT, K-NN, RF and Support Vector Machine (SVM)] and they found that RF achieved better accuracy of 96.89%. Roobini *et al.* (2020) their work predicted early stage of DM using different ML techniques (DT, K-NN, SVM and RF) and discovered that RF had highest prediction accuracy. Also, Roobini and Lakshmi (2021) trained AdaBoost algorithm using demographic and clinical risk factors for prediction of DM. Their work revealed that the model had better accuracy compared to existing ML models. However all the work reviewed used clinical risk factors or demographic and clinical risk factors or combination of demographic, clinical and lifestyle risk factors, and the gap to fill is inclusion dietary risk factors which was not focused on before. Thus, the aim of this paper is to assess the accuracy of predictive model in detecting DM and non-DM based on patient risk factors.

2.0 Materials and Methods

2.1 Data Collection and Preprocessing

First and second sample datasets used in this paper consists of four hundred (400) patients each and are obtained from patients' record suffering from DM in General Hospital Kaura Namoda, Nigeria between 2019 to 2023. Data preprocessing and preparation was conducted and it was divided into two main categories: data cleaning and balanced sampling. Data cleaning steps applied are outlier detection and removal, missing value handling, data normalization and one-hot coding. The datasets was imbalance because 211(52.8%) of the considered patients belong to DM class (majority class) and 189(47.2%) of the patients are assigned to non-DM class (minority class) in the first sample dataset, in the second sample dataset 236(59%) assigned to DM class (majority class) and 164(41%) allocated to non-DM class (minority class). Previous study by Krawczyk (2016) have shown that the classifiers trained with imbalance datasets have higher accuracy for predicting the majority class and minority class could not be trained with high accuracy. To address imbalance datasets in this paper, first approach was sampling from data without balancing the class distribution, second was over sampling from the minority class and third combining under sampling and over sampling which make the data balance.

The first sample dataset consist of 7 risk factors namely 2 demographic risk factors (age and sex), 4 clinical risk factors (family history of DM, blood glucose level, blood pressure level and body mass index) and 1 lifestyle risk factor (physical activity). While, the second dataset contain 14 risk factors which are 2 demographic risk factors (age and sex), 4 clinical risk factors (family history of DM, blood glucose level, blood pressure level and body mass index), 1 lifestyle risk factor (physical activity) and 7 dietary risk factors (preference for sweet food, preference for salty food, red meat, refined carbs, energy drinks, white rice and processed meats) and 1 output (i.e. diagnosis recommended for these patients by the physician that attended to them. That is, the first sample dataset have demographic, clinical and lifestyle risk factors only while second sample dataset consist of demographic, clinical, lifestyle and dietary risk factors. Moreover, there was no any rationale for selecting the risk factors because they were only the risk factors found in the patients' files. The risk factors and their formats are presented in Table 1.

Table 1: Risk Factors and their Format

Dataset	**Variable Name**	**Classification Network Type**	**Predictive Network Type**
First	**7** risk factors	Y or N (Character)	1 or 0 (Continuous)
	Diagnosis	DM Non- DM	1 0
Second	14 risk factors	Y or N (Character)	1 or 0 (Continuous)
	Diagnosis	DM Non- DM	1 0

2.2 Data Normalization

Data normalization was performed because firstly DM datasets have risk factors that differ in range and unit, this would reduce the models performance and accuracy. Secondly, prevent features with larger scales from dominating the learning process. Since the assumption was that ML algorithms are trained in such a way that all features contributed equally to the learning process. There are two major techniques for normalization namely min-max scaling and z-score normalization. But this paper used min-max technique because it transforms risk factors of the datasets to a specified range, usually between zero (0) and one (1) and maintains the interpretability of the original values within the specified range. The min-max scaling formula used was given by

$$X_{normalized} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{1.1}$$

where X is a random risk factor value that is to be normalized, $X_{\min}$ is the minimum risk factor value in the dataset and $X_{\max}$ is the maximum risk factor value.

When X is minimum value, the numerator is zero ($X_{\min}$ - $X_{\min}$) and hence, the normalized value is 0. When X is maximum value, the numerator is equal to the denominator ($X_{\max}$ - $X_{\max}$) and the normalized value is 1. Moreover, when X is neither minimum nor maximum, the normalized value is between 0 and 1.

2.3 Feature Selection Method

Backward stepwise feature selection method was used to remove risk factors that are not important in building the model. The method starts with full set of the risk factors and iteratively removes one feature at a time based on a predefined criterion. The paper used the following steps to remove least informative risk factors (i) select a significant level or select the p-value usually 0.05 (ii) fit the model with all the risk factors selected (iii) identify risk factors which has the highest p-value (iv) if the p-value of this risk factor is greater than 0.05, the risk factor is removed from the dataset. However, if the p-value of this risk factor is less than 0.05, the risk factor is retained (v) remove risk factors with p-value greater than 0.05 from the dataset and fit the model again with new dataset. After fitting the model with the new dataset, jump back to (iii). This process continues until reach a point in (iv) where the highest p-value from all the remaining risk factors in the dataset are less than 0.05. Six (6) risk factors were retained for the first sample dataset age, family history of DM, blood glucose level, blood pressure level, body mass index, physical activity and sex was removed. In the second sample dataset, twelve (12) risk factors were selected age, family history of DM, blood Glucose level, blood pressure, body mass index, physical activity, preference for sweet food, red meats, refined carbs, energy drinks, white rice and processed meat, sex and preference for salty food are eliminated.

2.4 Data Splitting

Data splitting, first and second sample datasets are divided into training, validation and test subsets. The training set contain 70% (280) data which was used to train the model, validation set contain 15% (60) data to validate the model and test set contain 15% (60) to evaluate the model performance. The paper experimented with multiple data splits such as 60:20:20, 80:10:10 and found that the ratio 70:15:15 consistently provided the best result in terms of model stability and accuracy. The result of first and second datasets splitting was presented in Table 2 and Table 3.

Table 2: Splitting of First Sample Dataset into Training, Validation and Test Subsets

	TRAINING SET			VALIDATION SET			TEST SET		
	DM STATUS			DM STATUS			DM STATUS		
	DM	Non-DM	Total	DM	Non-DM	Total	DM	Non-DM	Total
Count	122	158	280	41	19	60	37	23	60
Percentage	43.6	56.4	100.0	68.3	31.7	100.0	61.7	38.3	100.0

Table 3: Splitting of Second Sample Dataset into Training, Validation and Test Subsets

	TRAINING SET			VALIDATION SET			TEST SET		
	DM STATUS			DM STATUS			DM STATUS		
	DM	Non-DM	Total	DM	Non-DM	Total	DM	Non-DM	Total
Count	145	135	280	27	33	60	28	32	60
Percentage	51.8	48.2	100.0	45.0	55.0	100.0	46.7	53.3	100.0

Also, the paper used supervised learning algorithm and trained MLPNN model by using significant risk factors of the two datasets. Then library of the model was imported from R computing language, instance of the model was created and the model trained using model. Fit (…) function.

2.5 Hyperparameter Tuning

Hyperparameter tuning was applied using Grid search because it defines set of parameters values to search over and the algorithms tries all possible combination. Similarly, the paper employed model-centric approach because it focused on the characteristics of the model itself such as the structure of the model or the types of algorithms used. The approach also searches for the optimal combination of hyperparameters within a predefined set of possible values.

During training with first sample dataset, hyperparameters of the model was selected to obtain the best performance and best classification of the data. The MLPNN model initially used it default settings so that, as the model was adjusted to the data in the training process, the hyperparameters were also adjusted. After training, the hyperparameter of the model are activation “sigmoid”, alpha “0.05”, hidden layer sizes “25:25”, learning rate “constant” and momentum rate 0.1. For the second sample dataset activation “sigmoid”, alpha “0.05”, hidden layer sizes “42:42”, learning rate “constant” and momentum rate 0.1.

2.6 Design of Multilayer Perceptron Neural Network

MLPNN was designed for both first and second datasets, for the first sample dataset the network had 14 input layers, 25 hidden layers and 1 output layer. For the second sample dataset, the network had 28 input layers, 42 hidden layers and 1 output. Figure 1 showed diagrammatical representation of the proposed Neural Network (NN). Artificial neural network design called Multilayer Perceptron Neural Network (MLPNN) was especially suitable for prediction and was widely used in practice.

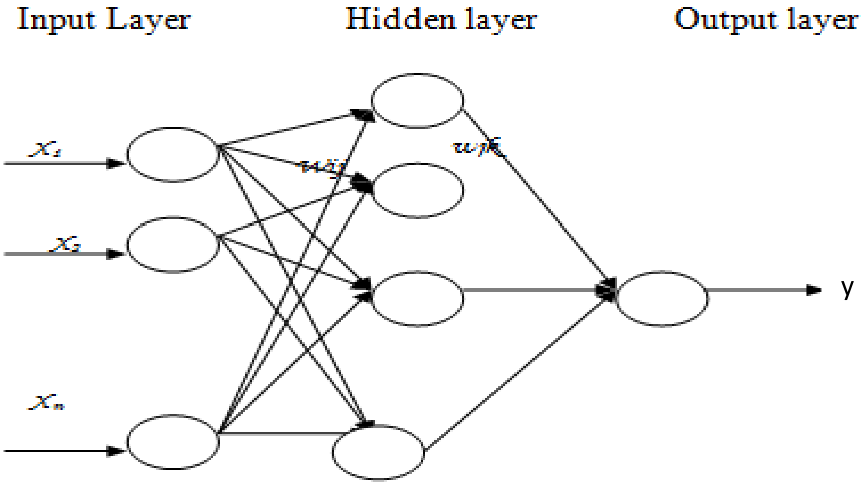


Figure 1: Design of Multilayer Perceptron Neural Network

The network consists of one input layer; one or more hidden layers and one output each consisting of several neurons. Each neuron processes its inputs and generates one output value that is transmitted to the neurons in the subsequent layer. Each neuron in the input layer delivers the value of one predictor from vector x. When considering normal/abnormal patient one output neuron is satisfactory. In each layer, the signal propagation was accomplished as follows: first, a weight sum of inputs was calculated at each neuron; the output value of each neuron in the proceeding network layer times the respective weight of the connection with that neuron. A transfer function g(x) is then applied to this weighted sum to determine the neuron`s output value. So, each neuron in the hidden layer produces the so-called activation (Frank, 2022).

$$a_j = g\left[\sum_i w_{ij} x_i\right] \tag{2.1}$$

The neurons in the output layer behave in a manner similar to the neurons of the hidden layer to produce the output of the network as shown in equation (2.2) (Irie and Miyake, 2023).

$$y_k = f\left[\sum_i w_{jk} a_j\right] = f\left[\sum_i w_{jk}' g\left(\sum w_{ij}' x_i\right)\right] \tag{2.2}$$

Where w_{ij}' and w_{jk}' are weights.

Equation (2.3) and (2.4) showed the calculation formula from input layer (*i*) to hidden layer (*j*), where O_j is the output of node j, O_i is the output of node i, w_{ij} is the weight connected between node i and node j, and θ_j is the bias of node j.

$$O_j = f\left(net_j\right) = \frac{1}{1+e^{-net_j}} \tag{2.3}$$

$$net_j = \sum_i w_{ij} O_i + \theta_j \tag{2.4}$$

Similarly, Equation (2.) and (2.6) showed computation formula for hidden layer (*j*) to output layer (*k*), where O_k is the output of node k, O_j is the output of node j, w_{jk} is the weight connected between node j and k, and θ_k is the bias of node k.

$$O_k = f\left(net_k\right) = \frac{1}{1+e^{-net_k}} \tag{2.5}$$

$$et_k = \sum_k w_{jk} O_j + \theta_k \tag{2.6}$$

The network activation function in Equations (2.3) and (2.5) was Sigmoid Activation Function. Moreover error is calculated using Equation (2.7) to measure the differences between desired output and actual output that had been produced in feed forward phase. Error was then propagated backward through the network from output layer to input layer and weights are modified to reduce the error as the error was propagated.

$$Error = \frac{1}{2}\left[Output_{desired} - Output_{actual}\right]^2 \tag{2.7}$$

Based on the error calculated, back propagation was applied from output (k) to hidden (j) as in Equation (2.8) and (2.10)

$$w_{ji}\left(t+1\right) = w_{ji}\left(t\right) + \Delta w_{ji}\left(t+1\right) \tag{2.8}$$

$$\Delta w_{ji}\left(t+1\right) = \eta \delta_k O_j + \alpha \Delta w_{ji}\left(t\right) \tag{2.9}$$

$$\delta_k = O_k\left(1-O_k\right)\left(t_k - O_k\right) \tag{2.10}$$

where $w_{ji}\left(t\right)$ is the weight from node j to node i at time t, Δw_{ji} is the weight adjustment, η is the learning rate, α is the momentum rate, δ_j is error at node j, δ_k is error at node k, O_i is the actual network output at node i, O_j is the actual network output at node j, O_k is the actual network output at node k, w_{kj} is the weight connected between node j and k, and θ_k is the bias of node k. This process was repeated iteratively until convergence achieved (targeted learning error).

2.7 Evaluation of the Model Performance

MLPNN model was evaluated in terms of its accuracy, sensitivity/recall and specificity using Fogarty and Bamber (2005) formula. Accuracy measures the proportion of cases (DM and non-DM patients) correctly classified, sensitivity/recall measures the fraction of positive cases (DM patients) that are classified as positive and specificity measures the fraction of negative cases (non-DM patients) that are classified as negative

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \tag{2.11}$$

$$Sensitivity / \mathrm{Re}call = \frac{TP}{TP + FN} \tag{2.12}$$

$$Specificity = \frac{TN}{TN + FP} \tag{2.13}$$

where TP, TN, FP, FN, denotes True Positive, True Negative, False Positive, False Negative, respectively. The model would be considered adequate if it has better accuracy, sensitivity/recall and specificity.

After evaluation of the ML models performance, Receiver Operating Characteristic (ROC) Curve was used to determine the discriminatory ability of the models in distinguishing between eye disease and non- eye disease patients. This was done by plotting the true positive rate (sensitivity) against the false positive rate (1-specificity) at different cut-off points. Each point on the ROC plots represents a sensitivity/specificity pair corresponding to a particular decision threshold. The models have perfect discrimination when the ROC plots passes through the upper left corner. The closer the ROC plot was to the upper left corner, the higher the overall accuracy of the models (Zweigh and Campbell, 1993).

Similarly, interpretation given by Traditional Academic Point System (TAPS, 2005) was used to interpret the Area under the Receiver Operating Characteristic (AUROC) Curve of the models, area less than equal to 0.59 indicate poor discrimination, 0.60 to 0.69 indicate fair discrimination, 0.70 to 0.79 indicate good discrimination, 0.80 to 0.89 very good discrimination and 0.90 to 1.00 excellent discrimination.

3.0 Results and Discussions

For the first sample dataset which consist of 400 patients with demographic, clinical and lifestyle risk factors, the trained MLPNN model was used to detect DM and Non-DM patients in the training, validation and test sets. The result in Table 4 indicated that the model detected 97.5% DM patients in the training set, 94.9% in the validation set and 88.9% for the test set. The model detected 94.9% Non-DM patients in the training sample, 82.6% in the validation set and 84.6% for the test set.

Table 4: Detection of DM and Non-DM Patients using MLPNN Model

Observed	**Detected Patients**

		DM	Non-DM	Total	Percent Correct
Training Set	*DM*	*119*	*3*	*122*	*97.5*
	Non-DM	*8*	*150*	*158*	*94.9*
	Total	***127***	***153***	***280***	
Validation Set	*DM*	*39*	*2*	*41*	*95.1*
	Non-DM	*3*	*16*	*19*	*84.2*
	Total	***42***	***18***	***60***	
Test Set	*DM*	*34*	*3*	*37*	*91.9*
	Non-DM	*2*	*21*	*23*	*91.3*
	Total	***36***	***24***	***60***	

Out of 122 DM patients in the training set, the model detected 119 DM patients and 3 non-DM patients. In the validation set out of 41 patients the model detected 39 DM patients and 2 non-DM patients and in the test set out of 37 patients the model detected 34 DM patients and 3 non-DM patients. Likewise, out of 158 Non-DM patients in the training set, the model detected 150 non-DM patients and 8 DM patients, in the validation set out of 19 non- DM patients the model detected 16 non-DM patients and 3 DM patients and in the test set out of 23 patients the model detected 21 non-DM patients and 2 DM patients.

In the second sample dataset which consist of 400 patients with demographic, clinical, lifestyle and dietary risk factors, Table 5 showed that the MLPNN model detected 99.2% DM patients in the training set, 100% in the validation set and 100% for test set. Also, the model detected 98.7% Non-DM patients in the training set, 95.7% in the validation set and 100% for the test set.

Table 5: Detection of DM and Non-DM Patients using MLPNN Model

Observed		**Detected Patients**			
		DM	**Non-DM**	**Total**	**Percent Correct**
Training Sample	DM	*143*	*2*	*145*	*98.6*
	Non-DM	*3*	*132*	*135*	*97.8*
	Total	**146**	**144**	**280**	
Validation Sample	DM	*26*	*1*	*27*	*96.3*
	Non-DM	2	*31*	*33*	*93.9*
	Total	**58**	**22**	**60**	
Test Sample	DM	*26*	*2*	28	92.9
	Non-DM	*2*	*30*	32	93.8
	Total	**28**	**32**	**60**	

Out of 145 DM patients in the training set, the model detected 143 DM patients and 2 non-DM patients. In the validation set out of 27 patients, the model detected 26 DM patients and 1 non-DM patient and in the test set out of 28 patients the model detected 26 DM patients and 2 non-DM patients. Likewise, out of 135 Non-DM patients in the training sample, the model detected 132 non-DM patients and 3 DM patients, in the validation set out of 33 non-DM patients the model detected 31 non-DM patients and 2 DM patients and in the test set out of 32 patients the model detected 30 non-DM patients and 2 DM patients.

MLPNN Model was evaluated in term of its accuracy, sensitivity and specificity for detection of DM and non-DM. The results of training, validation and test sets for first and second sample datasets are presented in Table 6 and 7 respectively. In the first sample dataset, the model achieved 96.1% training accuracy, 93.7% sensitivity and 98.0 % specificity. In the validation set the model achieved 91.7% accuracy, 92.9 % sensitivity and 88.9% specificity and in the test set the model showed accuracy, sensitivity and specificity of 91.7%, 94.4% and 87.5% respectively.

Table 6: Evaluation of Model Performance for First Sample Dataset

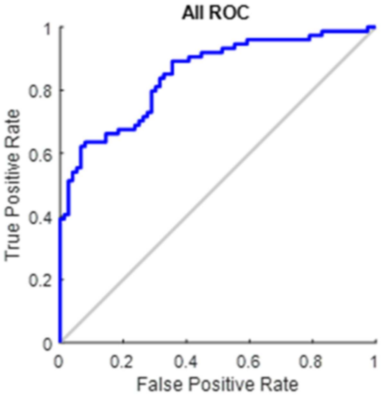
Indices	Training Set	Validation Set	Test Set
Accuracy (%)	96.1	91.7	91.7
Sensitivity (%)	93.7	92.9	94.4
Specificity (%)	98.0	88.9	87.5

In the second sample dataset the model achieved 98.2% training accuracy, 97.9% sensitivity and 98.5% specificity. In the validation set the model achieved 95.0% accuracy, 92.2% sensitivity and 96.9% specificity and in the test set the model achieved 93.3% accuracy, 92.9% sensitivity and 93.8% specificity.

Table 7: Evaluation of Model Performance for Second Sample Dataset

Indices	Training Sample	Validation Sample	Test Sample
Accuracy (%)	98.2	95.0	93.3
Sensitivity (%)	97.9	92.2	92.9
Specificity (%)	98.5	96.9	93.8

Figures 2 and 3 showed ROC Curves of first and second sample datasets. The Model had AUROC Curve of 0.96 for the first sample dataset with 95% confidence interval (0.81 to 0.99) and second sample had AUROC Curve of 0.99 with 95% confidence interval (0.85 to 1.00). The two AUROC Curve fall within 0.90 to 1.00, this indicated excellent discrimination and the model had the ability to discriminate between DM and Non-DM patients. But despite its discriminatory ability, the second sample dataset which used demographic, clinical, lifestyle and dietary risk factors, it AUROC Curve was larger than the first sample dataset that used demographic, clinical and lifestyle risk factors and this was attributed to the inclusion of dietary risk factors in the second sample dataset.



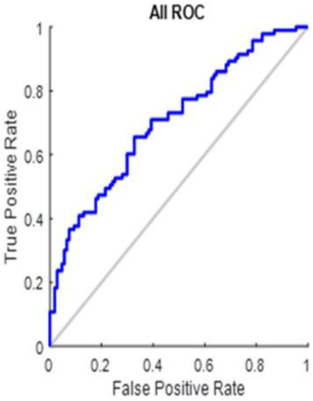


Figure 2: ROC Plot of First Set of Data

Figure 3: ROC Plot of Second Set of Data

The study relies on medical records from a single hospital (General Hospital Kaura Namoda). This may not represent the larger population of Nigeria or other regions. Limited geographic and demographic diversity may introduce bias and affect the generalizability of the findings.

4. Conclusion

This paper presents the development of a Multi-Layer Perceptron Neural Network (MLPNN) model aimed at the early detection of diabetes mellitus (DM) by utilizing two risk factors. The first set includes demographic, clinical, and lifestyle factors, while the second set adds dietary risk factors to the mix.

The results show that the backward stepwise feature selection method identified six significant risk factors in the first dataset: age, family history of DM, blood glucose level, blood pressure level, body mass index, and physical activity. It also removed one insignificant risk factor: sex. For the second dataset, the method retained twelve significant risk factors: age, family history of DM, blood glucose level, blood pressure, body mass index, and physical activity, preference for sweet foods, red meats, refined carbohydrates, energy drinks, white rice, and processed meats. It removed two insignificant risk factors: sex and preference for salty foods.

The retained significant risk factors were utilized to train the model, enabling it to differentiate between DM and non-DM patients. The MLPNN model demonstrated high accuracy in detecting both groups of patients, with improved performance noted when dietary risk factors were included. This indicates that dietary risk factors and demographic, clinical, and lifestyle factors are essential for accurately detecting DM and non-DM cases.

The paper suggests that future research should compare various artificial neural network techniques for the early detection of DM using all four risk factor categories to evaluate whether the MLPNN model can achieve even higher accuracy.

ACKNOWLEDGEMENTS

This research was funded by Tertiary Education Trust Fund (Tetfund) grant under Institute Based Research (IBR)

Ethical Approval

Ethical approval was given for data collection from the ethical committee of the hospital after checking the proposal work; also datasets collection was highly confidential and carried out

COMPETING INTERESTS

Authors have declared that no competing interests exist

Disclaimer (Artificial intelligence)

Option 1:

Author(s) hereby declare that NO generative AI technologies such as Large Language Models (ChatGPT, COPILOT, etc.) and text-to-image generators have been used during the writing or editing of this manuscript.

References

Al-Shayea, Q.K. (2011). Artificial Neural Network in Medical Diagnosis. *International Journal of Computer Science,* 8(2):150-154. Doi: 10.1011/ijcs.2011.150154.

Bellazi, R. and Zupan, B. (2011). Predictive Data Mining in Clinical Medicine: Current Issues and Guidelines. *International Journal of Medical Information,* 8 (77): 81-97. Doi: 10.1024/ijmi.2011.8197.

Bukhari, M., Alkhamees, B.F., Hussain, S., Gumaei, A., Assiri, A. and Ullah, S.S. (2021). An Improved Artificial Neural Network for Effective Diabetes Prediction. *Hindawi Complexity,* doi.org/10.1156/2021/hc.552527/4409-1101.

Chen, Q., Hu, H., She, Y., He, Q., Huang, X., Shi, H., Cao, X. and Zhang, X. (2024). An Artificial Neural Network Model for Evaluating the Risk of Hyperuricaemia in Type 2 Diabetes Mellitus. *Journal of Science Report,* 14: doi.org/10.1038/jsr41598-024-52550-1.

Evwiekpaefe, A.E. and Abdulkadir, N. (2023). A Predictive Model for Diabetes Mellitus using Machine Learning Techniques (A Study in Nigeria). *African Journal of Informative System,* 15(1): https://digitalcommons.kennesaw.edu/ajis/vol 15 1/1.

Farooqui, N.A., Mehra, R. and Tyaqi, A. (2023). Prediction Model for Diabetes Mellitus using Machine Learning Techniques. International *Journal of Computer Science and Engineering,* 6(3): doi.org/10.26438/ijcse/v6i3.

Forgatty, E. and Bamber, D. (2005). Area above the Ordinal Dominance Graph and area Below the Receiver Operating Characteristic Graph. *Journal of Maths Psychology,* 12: 387-415.

Frank, G. (2022). Data Mining Techniques to Predict Hospitalization of Hemodialysis Patients. *Journal of Decision Support System,* 50(8):439-48. doi.org/10.1037/jdss.2022/8805-22.

International Diabetes Federation (2021). Diabetes Prevalence in 2021. Retrieved from diabetesatlas.org

Irie, B. and Miyake, S. (2023). Capabilities of Three-Layered Perceptrons. *International Journal on Neural Networks,* 15(2):641-648. Doi.org/10.1036/ijnns. 1178-023-7.000-8.

Krawczyk, B. (2016). Learning from Imbalanced Data: Open Challenges and Future Directions. *Journal of Progress in Artificial Intelligence*, 5(4): 221-32. Doi : 10.1310/jpai.54221-32

National Institute of Health (2021). Diabetes Mellitus. Retrieved from https://www.ncbi.nlm.nih.gov www.ncbi.nlm.gov

Pradhan, N., Rani, G., Dhaka, V.S. and Poonia, R.C. (2020). Diabetes Prediction using Artificial Neural Network. *Journal of Science Direct,* 18(5): doi.org/10.1016/jsd.8978-0-12-819061-6.00014-8.

Roobini, M.S., Satwick, Y.S., Reddy Kumar, A.A., Lakshmi, M., Deepa, D. and Ponraj, A. (2020). Predictive Analysis of Diabetes Mellitus using Machine Learning Techniques. *Journal of Computational and Theoretical Nanoscience,* 17(8): 3449-3452. Doi: 10.1166/jctn. 2020.9207-52.

Roobini, M.S. and Lakshmi, M. (2021). Predictive Supervised Learning Model for Classification of Type 2 Diabetes Mellitus. *Journal of Research Square,* 1. Doi: 10.21203/rs.3.rs-1009663/v1.

Sahu, P. and Mantri, J.K. (2023). Artificial Neural Network Based Diabetes Prediction Model and Reducing Impact of Class Imbalance on its Performance. *Journal of SSRN*, 15: doi.org/10.2139/ssrn.4538967.

Setiawan, H. (2024). Enhancing the Accuracy of Diabetes Prediction using Feed Forward Multilayer Perceptron Neural Networks. *Journal of Brilliance Research of Artificial Intelligence*, 4(1): doi.org/1047709/brilliance.v4i1.3888.

Traditional Academic Point System (2005). Interpretation of Area Under the Receiver Operating Characteristic Curve. *Journal of Maths Psychology* 18: 156-189.

World Health Organization (2022). Action Plan for the Global Strategy of Prevention and Control of Non-Communicable Disease. *World Health Organization Journal* 15(4):17-19

Xue- Huimeng, G., Yi- Xiang, H., Dong-Ping, R., Qiu -Zhang, H. and Qing- Liu, K.(2013). Comparison of Three Data Mining Models for Predicting Diabetes or Pre Diabetes by Risk Factors. *Kaohsiung Journal of Medical Sciences,* 29: 93-99.

Zou, Q., Qu, K., Luo, Y., Yin, D., Ju, Y. and Tang, H. (2018). Predicting Diabetes Mellitus with Machine Learning Techniques. *Journal of Computational Genomics,* 9: doi.org/10.3389/fgene.2018.00515.

Zweig, P. and Campbell, G. (1993). Advances in Statistical Methodology for Evaluation of Diagnostic and Laboratory Tests. *Journal of Medical Science,* 13: 499-508.