Advanced Journal of Science, Technology and Engineering
ISSN: 2997-5972
Volume 5, Issue 1, 2025 (pp. 115-128)
www.abjournals.org

# ARTIFICIAL NEURAL NETWORKS FOR DETECTION OF DIABETES MELLITUS BY SELECTING SIGNIFICANT RISK FACTORS USING BACKWARD STEPWISE FEATURE SELECTION METHOD

**Kabiru H.[1], Buhari Shehu A.[2], and Idris A. K.[3]**

[1,3]Department of Science Laboratory, Federal Polytechnic Kaura, Namoda.

[2]Department of Statistics, Federal Polytechnic Kaura Namoda.

Emails:

[1]kabiruhashimu@yahoo.com, [2]buharishehukaura@gmail.com, [3]aliyukankara@gmail.com

**Cite this article:**

Kabiru, H., Buhari, S. A., Idris, A. K. (2025), Artificial Neural Networks for Detection of Diabetes Mellitus by Selecting Significant Risk Factors Using Backward Stepwise Feature Selection Method. Advanced Journal of Science, Technology and Engineering 5(1), 115-128. DOI: 10.52589/AJSTE-JZDPAPUK

**Manuscript History**

Received: 29 Jan 2025

Accepted: 9 Mar 2025

Published: 23 Mar 2025



**ABSTRACT:** *Diabetes Mellitus (DM) is a chronic disorder affecting more than 300 million people world-wide and it needs urgent attention. Selecting of significant risk factors (SRFs) and their contributions to the risk of the disease is the key for early detection of the disease. The aim of this paper is to use Backward Stepwise Feature Selection Method (BSFSM) to select the SRFs and their contribution to the risk of DM and Kappa statistic value (KSV) to evaluate the model performance. Dataset consists of 400 patients with demographic, clinical, lifestyle and dietary risk factors collected from General Hospital Kaura Namoda, Zamfara State, Nigeria from 2019 to 2023 by checking the file of patients suffering from DM. The results obtained revealed that BSFSM retained twelve (12) SRFs, namely Blood Glucose level (BGL), High Body Mass Index (BMI), Family History of DM (FHDM), Preference for Sweet Food (PSWF), Age, Lack of Physical Activity (LPA), Blood Pressure (BP), Red Meat (RM), Refined Carbs (RC), Energy Drink (ED), White Rice (WR) and Processed Meat (PM), and removed two (2) Non-SRFs: Sex and Preference for Salty Food (PSF). The SRFs contributed 85.40%, 51.34%, 55.72%, 68.23%, 57.50%, 29.96%, 66.18%, 41.42%, 12.20%, 18.65%, 29.76% and 10.11% respectively to the risk of DM. Similarly, the Non-SRFs contributed 0.98% and 1.16% respectively to the risk of DM. The MLP model detected 98.6% DM patients in the training set, 96.3% in the validation set and 92.9% in test set. There were 97.8% Non-DM patients in the training set, 93.9% in the validation set and 93.8% for the test set. The KSV of the model was 0.94 and it was capable of distinguishing between DM and Non-DM patients. This paper demonstrated that BSFSM was capable of selecting the SRFs and their contributions, and KSV adequately evaluate the performance of the model.*

**KEYWORDS:** Diabetes Mellitus, Backpropagation, Detection, Artificial Neural Network, Kappa Statistic Value.



Article DOI: 10.52589/AJSTE-JZDPAPUK
DOI URL: https://doi.org/10.52589/AJSTE-JZDPAPUK

Advanced Journal of Science, Technology and Engineering
ISSN: 2997-5972
Volume 5, Issue 1, 2025 (pp. 115-128)

www.abjournals.org

## INTRODUCTION

Diabetes Mellitus (DM) is a chronic disorder characterized by abnormally high levels of sugar (glucose) in the blood. In people with DM, blood sugar levels remain high. This might be because insulin is not being produced at all, is not made at sufficient level or not as effective as it should be. DM affects more than 300 million people worldwide. In 2021, it was discovered that 1 in 7 people of age 50 years and above has DM. The highest prevalence (17.9%) was found in American Indians and Alaska natives (Al-shayea, 2011).

Nigeria has the largest population in Africa of more than 220 million and, of this, the 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 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 it 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 parts of Nigeria; its prevalence ranges 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).

Artificial Intelligence (AI) has been shown in various scientific studies as an effective diagnostic tool to identify various diseases in health care services. Conventional diagnostic methods profoundly depend on physicians' experience and professional knowledge, which might lead to a high rate of misdiagnosis and waste a large amount of medical data. Therefore, AI in medicine has the potential to revolutionize the existing disease diagnosis system and create a significant clinical effect in ophthalmic health care service. Diseases such as DM, hypertension, stroke and heart attack were the most common causes of death on a global scale. Therefore, prediction and treatment of them and other disorders through unmanned automated applications is necessary (Naser & Ola, 2008).

Artificial Neural Networks (ANNs) are branches of AI used to build complicated models. Basically, an Artificial Neural Network model contains three layers: input layer, intermediate hidden layer and output layer, with each layer made up of nodes (neurons) and links. The nodes in the input layer are viewed as predicted variables whereas the nodes in the output layer are analyzed as the outcome variables (Bellazi & Zupan, 2011). This paper used a popular ANN Architecture called Multilayer Perceptron (MLP) with back-propagation (i.e., Supervised Learning Algorithm), which is arguably the most commonly used and well-studied ANN architecture. MLP is a feed-forward neural network trained with the standard back-propagation algorithm and it is known to be a powerful function approximator for prediction and classification problems (Xue-Hui Meng *et al.*, 2011). ANNs provide 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-Shayea, 2011).

The paper reviewed works on ANNs for the detection of DM. Sahu and Mantri (2023) used the MLP model for the detection of Diabetes and Principal Component Method (PCM) to


Article DOI: 10.52589/AJSTE-JZDPAPUK
DOI URL: https://doi.org/10.52589/AJSTE-JZDPAPUK

select significant demographic and clinical risk factors in the face of inconsistent results, gaps and data class imbalance. The model achieved an accuracy of 84% relative to the baseline with Area under the Receiver Operating Characteristic (ROC) Curve of 79%. Similarly, the work of Chen *et al.* (2024) observed that ANNs trained using risk factors had better efficacy and facilitated the reduction of harm caused by Type 2 DM combined with Hyperuricaemia. Likewise, Bukhari *et al*. (2021) used information gain to select significant demographic, clinical and lifestyle risk factors to train Artificial Backpropagation Stochastic Gradient Neural Network (ABPSCGNN) algorithm for the detection of DM patients; the ABPSCGNN model achieved 93% accuracy. Also, Pradhan *et al.* (2020) applied MLP model for the detection of Diabetes patients and PCM to select nine (9) features. The model had 85.09% accuracy. Moreover, the work of Setiawan *et al.* (2024) focused on the Neural Network model for the detection of DM patients using clinical data. The result obtained showed that the model had an accuracy of 97% and area under the (ROC) Curve of 93.5%, and this demonstrates the ability of the model in detecting DM patients. Zou *et al.* (2018) used Decision Tree (DT), Random Forest (RF) and Neural Network to detect DM using significant demographic and clinical risk factors selected by PCM. Their results showed that RF had the best accuracy of 80.8% and area under the ROC Curve of 94.8%. Furthermore, Evwiekpaefe and Abdulkadir (2023) employed three (3) models, namely K-Nearest Neighbour (K-NN), DT and Artificial Neural Networks (ANNs) to detect DM in individuals at an early stage. Their work used the information gain method and identified nine (9) clinical and demographic risk factors responsible for DM. On the other hand, Farooqui *et al*. (2023) used clinical, demographic and lifestyle risk factors, and built four models [DT, K-NN, RF and Support Vector Machine (SVM)]. They found that RF achieved better accuracy of 96.9% and area under the ROC Curve of 95.1%. The work of Roobini *et al.* (2020) detected the early stage of DM using different models (DT, K-NN, SVM and RF) and discovered that RF had the highest detection accuracy. Also, Roobini and Lakshmi (2021) trained AdaBoost algorithm using significant demographic and clinical risk factors for the detection of DM. Their work revealed that the model had better accuracy and area under the ROC curve compared to existing models.

However, all the works reviewed used feature selection methods to select SRFs and ROC curve to evaluate the model performance, but none of them used backward stepwise feature selection method (BSFSM) to identify the contributions of the SRFs to the risk of DM and Kappa Statistic Value (KSV) to evaluate the model performance. These are the gaps to be addressed in this paper. Therefore, the aim of this paper is to use backward stepwise feature selection method to select the significant risk factors and their contribution to the risk of DM and KSV to evaluate the model.


Article DOI: 10.52589/AJSTE-JZDPAPUK
DOI URL: https://doi.org/10.52589/AJSTE-JZDPAPUK

Advanced Journal of Science, Technology and Engineering
ISSN: 2997-5972
Volume 5, Issue 1, 2025 (pp. 115-128)

www.abjournals.org

## MATERIAL AND METHODS

### Data Collection

The data used in this paper was obtained from past patient records of patients suffering from DM in General Hospital Kaura Namoda Zamfara State. The Diabetes patient data (400 records) were extracted from the outgoing patient's records from 2019 to 2023. The data records represent the risk factors and diagnosis recommended for these patients by the physician that attended to them. The demographic risk factors were Age and Sex. The clinical risk factors were Family History of DM (FHD), Blood Glucose Level (BGL), Blood Pressure (BP) and Body Mass Index (BMI). The lifestyle variable was Lack of Physical Activity (LPA), and the dietary risk factors were Preference for Sweet Food (PSWF) and Preference for Salty Food (PSF), Red Meat (RM), Refined Carbs (RC), Energy Drinks (ED), White Rice (WR), Processed Meats (PM) and 1 output. The data was cleaned up by filtering out incomplete data and standardized. Table 1 presents the data risk factors and their data formats.

### Data Preprocessing and Preparation

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 were imbalanced because of 236 (59%) assigned to DM class (majority class) and 164 (41%) allocated to non-DM class (minority class). A previous study by Krawczyk (2016) has shown that the classifiers trained with imbalanced datasets have higher accuracy for detecting the majority class and the minority class could not be trained with higher accuracy. To address imbalanced datasets in this paper, the first approach was sampling from data without balancing the class distribution; the second was oversampling from the minority class, and, the third, combining undersampling and oversampling. All were done during training of the model.

**Table 1: Data Risk Factors and Their Format**

| Variable Name | Classification Network Type | Predictive Network Type |
|---|---|---|
| 14 risk factors | Y or N (Character) | 1 or 0 (Continuous) |
| Diagnosis | DM<br>Non- DM | 1<br>0 |

Table 1 shows that the dataset held fourteen (14) risk factors which served as inputs to the network and a diagnosis which indicated whether the patient was DM or Non-DM. Similarly, Yes was a character and used to indicate the presence of the risk factors and No was also a character and used to indicate the absence of the risk factors. For the predictive network type, 1 and 0 are continuous: if the output is 1, it indicates DM, and if it is 0, Non-DM. For symbolic data records, the statistics presented in Table 2 were generated.


Article DOI: 10.52589/AJSTE-JZDPAPUK
DOI URL: https://doi.org/10.52589/AJSTE-JZDPAPUK

Advanced Journal of Science, Technology and Engineering
ISSN: 2997-5972
Volume 5, Issue 1, 2025 (pp. 115-128)
www.abjournals.org

**Table 2: Analysis of Symbolic Data Values**

| S/NO | Risk Factors | Classification Network Type | Code | Count |
|---|---|---|---|---|
| 1 | Age | Integer (Continuous) | | 400 |
| 2 | Sex | Male | 1 | 108 |
| | | Female (Nominal) | 0 | 292 |
| 3 | Family History of DM | Yes | 1 | 330 |
| | | No (Nominal) | 0 | 70 |
| 4 | Blood Glucose Level | Integer (Continuous) | | 400 |
| 5 | Blood Pressure Level | Integer (Continuous) | | 400 |
| 6 | Body Mass Index | Integer (Continuous) | | 400 |
| 7 | Lack of Physical Activity | Yes | 1 | 310 |
| | | No (Nominal) | 0 | 90 |
| 8 | Preference for Sweet Food | Yes | 1 | 328 |
| | | No (Nominal) | 0 | 72 |
| 9 | Preference for Salty Food | Yes | 1 | 165 |
| | | No (Nominal) | 0 | 235 |
| 10 | Red Meat | Yes | 1 | 229 |
| | | No (Nominal) | 0 | 171 |
| 11 | Refined Carbs | Yes | 1 | 88 |
| | | No (Nominal) | 0 | 312 |
| 12 | Energy Drinks | Yes | 1 | 206 |
| | | No (Nominal) | 0 | 194 |
| 13 | White Rice | Yes | 1 | 297 |
| | | No (Nominal) | 0 | 103 |
| 14 | Processed Meat | Yes | 1 | 15 |
| | | No (Nominal) | 0 | 385 |
| Diagnosis | DM | | 1 | 236 |
| | Non-DM | | 0 | 164 |

Table 2 reveals that four (4) risk factors were continuous with integer values and the remaining ten (10) risk factors were nominal with two categories Yes and No, coded Yes as 1 and No as 0. The dataset used had 400 records of patients with 236 DM cases and 164 Non-DM cases.


Article DOI: 10.52589/AJSTE-JZDPAPUK
DOI URL: https://doi.org/10.52589/AJSTE-JZDPAPUK

Advanced Journal of Science, Technology and Engineering
ISSN: 2997-5972
Volume 5, Issue 1, 2025 (pp. 115-128)

www.abjournals.org

**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

$$X_{normalized} = \frac{X - X_{\min}}{X_{\max} - X_{\min}} \tag{2.1}$$

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.

**Feature Selection Method**

BSFSM was used in this paper to remove risk factors that are not significant in training the model and determine the contributions of the risk factors to the risk of DM. The method starts with a full set of the risk factors and iteratively removes one feature at a time based on a predefined criterion. The following criterion to remove least informative risk factors were adopted: (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 have 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 the new dataset. After fitting the model with the new dataset, jump back to (iii). This process continues until you reach a point in (iv) where the highest p-value from all the remaining risk factors in the dataset are less than 0.05. For the risk factor contributions, the paper used a random forest method to determine the contribution of each significant risk factor.


Article DOI: 10.52589/AJSTE-JZDPAPUK
DOI URL: https://doi.org/10.52589/AJSTE-JZDPAPUK

Advanced Journal of Science, Technology and Engineering
ISSN: 2997-5972
Volume 5, Issue 1, 2025 (pp. 115-128)
www.abjournals.org

## Data Splitting

In data splitting, the dataset was divided into training, validation and test subsets. The training set contained 70% (280) data which was used to train the model, the validation set contained 15% (60) data to validate the model, and the test set contained 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 dataset splitting was presented in Table 3.

**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 |

## Hyperparameter Tuning

Hyperparameter tuning was applied using Grid search because it defines a set of parameter values to search over and the algorithm tries all possible combinations. Similarly, the paper employed the 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 searched for the optimal combination of hyperparameters within a predefined set of possible values. The paper used supervised learning algorithms to train the model using the significant risk factors and the library of the model was imported from R computing language, an instance of the model was created and the model trained using the model: fit(x_ train, y_ train) function. During training, the hyperparameters of the model were selected to obtain the best performance and the best classification of the data. The MLP model initially used its default settings so that, as the model was adjusted to the data in the training process, the hyperparameters were also adjusted. After training, the hyperparameters of the model were activation “sigmoid”, alpha “0.05”, hidden layer sizes “42:42”, learning rate 0.8, momentum rate 0.7 and number of iterations 500.


Article DOI: 10.52589/AJSTE-JZDPAPUK
DOI URL: https://doi.org/10.52589/AJSTE-JZDPAPUK

Advanced Journal of Science, Technology and Engineering
ISSN: 2997-5972
Volume 5, Issue 1, 2025 (pp. 115-128)
www.abjournals.org

**Multilayer Perceptron**

MLP was designed using 28 input layers, 42 hidden layers and 1 output. Figure 1 shows a diagrammatic representation of the proposed MLP.

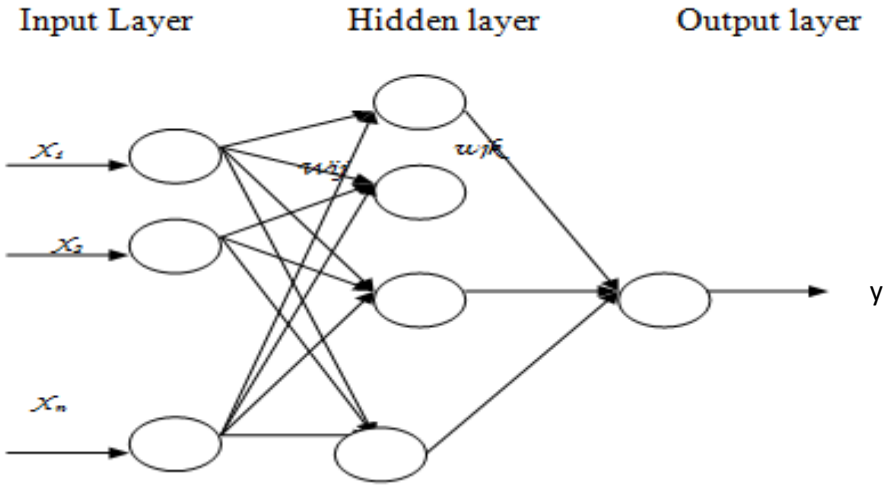


**Figure 1:** Design of the Proposed Multilayer Perceptron

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 DM/Non-DM patients, one output neuron was 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) was 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).

$$a_j = g\left(\sum_i w_{ij} x_i\right) \tag{2.2}$$

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 (2.3) (Irie & Miyake, 2023).

$$y_k = f\left(\sum_i w_{jk} a_j\right) = f\left(\sum_i w_{jk}' g\left(\sum w_{ij}' x_i\right)\right) \tag{2.3}$$

where $w_{ij}'$ and $w_{jk}'$ are weights.

Equations (2.4) and (2.5) showed the calculation formula from input layer (*i*) to hidden layer (*j*), where $O_j$ represents the output of node j, $O_i$ indicates the output of node i, $w_{ij}$ is the weight connected between node i and node j, and $\theta_j$ represents the bias of node j.


Article DOI: 10.52589/AJSTE-JZDPAPUK
DOI URL: https://doi.org/10.52589/AJSTE-JZDPAPUK

Advanced Journal of Science, Technology and Engineering
ISSN: 2997-5972
Volume 5, Issue 1, 2025 (pp. 115-128)

www.abjournals.org

$$O_j = f\left(net_j\right) = \frac{1}{1+e^{-net_j}} \tag{2.4}$$

$$net_j = \sum_i w_{ij} O_i + \theta_j \tag{2.5}$$

Similarly, Equation (2.6) and (2.7) showed computation formula for hidden layer (j) to output layer (k), where $O_k$ is the output of node k, $O_j$ presents the output of node j, $w_{jk}$ indicates the weight connected between node j and k, and $\theta_k$ represents the bias of node k.

$$O_k = f\left(net_k\right) = \frac{1}{1+e^{-net_k}} \tag{2.6}$$

$$et_k = \sum_k w_{jk} O_j + \theta_k \tag{2.7}$$

The network activation function in Equations (2.4) and (2.7) was sigmoid activation function. Moreover, error was calculated using Equation (2.8) to measure the differences between desired output and actual output that had been produced in the feed-forward phase. Error was then propagated backward through the network from the output layer to the input layer, and weights were modified to reduce the error as the error was propagated.

$$Error = \frac{1}{2}\left[Output_{desired} - Output_{actual}\right]^2 \tag{2.8}$$

Based on the error calculated, back propagation was applied from output (k) to hidden (j) as in Equation (2.9) and (2.11).

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

$$\Delta w_{ji}\left(t+1\right) = \eta \delta_k O_j + \alpha \Delta w_{ji}\left(t\right) \tag{2.10}$$

$$\delta_k = O_k\left(1-O_k\right)\left(t_k - O_k\right) \tag{2.11}$$

where $w_{ji}\left(t\right)$ is the weight from node j to node i at time t, $\Delta w_{ji}$ indicates the weight adjustment, $\eta$ represents the learning rate, $\alpha$ is the momentum rate, $\delta_j$ is error at node j, $\delta_k$ 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 $\theta_k$ is the bias of node k (Irie & Miyake, 2023). This process was repeated iteratively until convergence was achieved (targeted learning error) and it was achieved after 500 iterations.



Article DOI: 10.52589/AJSTE-JZDPAPUK
DOI URL: https://doi.org/10.52589/AJSTE-JZDPAPUK

Advanced Journal of Science, Technology and Engineering
ISSN: 2997-5972
Volume 5, Issue 1, 2025 (pp. 115-128)

www.abjournals.org

**Kappa Statistic Value**

This paper used KSV to measure the level of agreement of the model in distinguishing between DM and Non-DM patients (Cohen, 1960). KSV measures the percentage of data values in the main diagonal of the table and then adjusts these values for the amount of agreement that would be expected due to chance. To compute the KSV of the model, this paper calculates the observed level of agreement using Equation (2.12).

$$P_0 = P_{11} + P_{22} \tag{2.12}$$

where $P_O$ is the proportion of observed level of agreement, $P_{11}$ is the proportion of positive cases and $P_{22}$ is the proportion of negative cases.

This value $P_O$ compared to the value that would be expected which was calculated using Equation (2.13).

$$P_e = p_1 p_1 + p_2 p_2 \tag{2.13}$$

where $P_E$ is the proportion of expected level of agreement, $P_1$ is row one proportion total, $P_1$ is column one proportion total, $P_2$ is row two proportion total and $P_2$ is column two proportion total.

Now, KSV could be obtained using Equation (2.14)

$$K = \frac{P_O - P_e}{1 - P_e} \tag{2.14}$$

The paper also used the interpretation given by Altman (1991) to interpret the Kappa value of the model, such as, less than or equal to 0.49 indicates poor agreement, 0.50 to 0.59 fair agreement, 0.60 to 0.69 moderate agreement, 0.70 to 0.79 good agreement, 0.80 to 0.89 very good agreement, and 0.900 to 1.00 excellent agreement.


Article DOI: 10.52589/AJSTE-JZDPAPUK
DOI URL: https://doi.org/10.52589/AJSTE-JZDPAPUK

Advanced Journal of Science, Technology and Engineering
ISSN: 2997-5972
Volume 5, Issue 1, 2025 (pp. 115-128)
www.abjournals.org

## RESULTS AND DISCUSSION

This paper used backward stepwise feature selection method to determine the SRFs and their contributions to the risk of DM using R statistical computing software version 4.33. The result obtained is presented in Table 4.

**Table 4: Risk Factors and Their Contribution**

| Risk Factors | | P- Value | Percentage Contribution to the Risk of DM |
|---|---|---|---|
| | Body Mass Index | *2.235e-05* | *51.34%* |
| | Blood Pressure | *1.110e-08* | *66.18%* |
| | Family History of DM | *7.461e-10* | *55.72%* |
| | Preference for Sweet Food | *3.822e-09* | *68.23%* |
| | Sex | *5.619e-01* | *0.98%* |
| | Age | *6.506e-07* | *57.50%* |
| | Blood Glucose Level | *4.711e-12* | *85.40%* |
| | Lack of Physical Activity | *8.310e-6* | *29.96%* |
| | Preference for Salty Food | *1.450e-01* | *1.16%* |
| | Red Meat | *4.002e-07* | *41.4%* |
| | Refined Curbs | *1.061e-04* | *15.2%* |
| | Energy Drink | *1.105e-05* | *18.6%* |
| | White Rice | *2.030e-08* | *29.7%* |
| | Processed Meat | *5.120e-04* | *10.1%* |

Table 4 shows that BGL, BMI, FHD, PSWF, Age, LPA, BP, RM, RC, ED, WR and PM were the significant risk factors because their P-values were less than 0.05, while Sex and preference for salty food were not significant because their p-value were greater than 0.05. The significant risk factors, namely BGL, BMI, FHDM, PSWF, Age, LPA, BP, RM, RC, ED, WR and PM contributed 85.40%, 51.34%, 55.72%, 68.23%, 57.50%, 29.96%, 66.18%, 41.42%, 12.20%, 18.65%, 29.76% and 10.11% to the risk of DM respectively. Similarly, the non-significant risk factors (Sex and PSF) contributed 0.98% and 1.16% respectively to the risk of DM. The significant risk factors were used to train the MLP model for the detection of DM and Non-DM patients, and it achieved 97.5% training accuracy and 0.12% training loss. Training sample was used to develop the models, validation sample to validate the model and test sample to evaluate the model.

Table 5 indicates that the MLP model detected 98.6% DM patients in the training set, 96.3% in the validation set, and 92.9% for test set. Also, the model detected 97.8% Non-DM patients in the training set, 93.9% in the validation set and 93.8% for the test set.


Article DOI: 10.52589/AJSTE-JZDPAPUK
DOI URL: https://doi.org/10.52589/AJSTE-JZDPAPUK

Advanced Journal of Science, Technology and Engineering
ISSN: 2997-5972
Volume 5, Issue 1, 2025 (pp. 115-128)
www.abjournals.org

**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.

The KSV of the model was 0.94 and its value fell within the interval of 0.900 to 1.00. This indicated that the model had an excellent agreement in distinguishing between DM and Non-DM patients. In another way, it was capable of distinguishing between DM and Non-DM patients. Similarly, the asymptotic 95% confidence interval of the KSV had lower bound of 0.926 and upper bound of 0.959 which means that if the DM and Non-DM patients represent a random sampling from a larger population, one could be 95% sure that the confidence interval contains the true area.

## CONCLUSION

This paper used BSFSM to select SRFs and their contributions to the risk of DM and KSV to evaluate the model performance. The results obtained indicated that the method selected twelve (12) SRFs: BGL, BMI, FHDM, PSWF, Age, LPA, BP, RM, RC, ED, WR and PM, and removed two (2) Non-SRFs: Sex and PSF. The SRFs contributed 85.40%, 51.34%, 55.72%, 68.23%, 57.50%, 29.96%, 66.18%, 41.42%, 12.20%, 18.65%, 29.76% and 10.11% respectively to the risk of DM, and these contributions could help stakeholders in the health sector for early detection of the disease from the grassroot. The MLP model detected 98.6% DM patients in the training set, 96.3% in the validation set and 92.9% in the test set, as well as 97.8% Non-DM patients in the training set, 93.9% in the validation set and 93.8% for the test set. The KSV of the model was 0.94 and it was capable of distinguishing between DM and Non-DM patients. This paper demonstrated that BSFSM was capable of selecting SRFs and their contributions, and KSV adequately evaluated the performance of the model. The paper suggested that for future


Article DOI: 10.52589/AJSTE-JZDPAPUK
DOI URL: https://doi.org/10.52589/AJSTE-JZDPAPUK

research work, different methods of feature selection and evaluation of the model should be compared for the detection of DM and Non-DM patients using ANNs approach.

## ACKNOWLEDGEMENT

This research work was fully sponsored by the Tertiary Education Trust Fund (TETFund) through Institution-Based Research (IBR).

## COMPETING INTERESTS

Authors have declared that no competing interests exist.

## 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.

Altman, D.G. (1991). *Practical Statistics for Medical Researcher*. Third edition, London: Chapman and Hall. Pages 186-190.

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.

Cohen, H. (1960). *Statistics in Medicine*. Second edition, Boston: Little Brown and Company. Pages 234-241.

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.

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.

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.

Naser, D. and Ola, S.M. (2008). The Control of Hypertension in Persons with Diabetes: A Public Health Approach. *Public Health Journal* 102(5): 522-9.


Article DOI: 10.52589/AJSTE-JZDPAPUK
DOI URL: https://doi.org/10.52589/AJSTE-JZDPAPUK

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.

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.


Article DOI: 10.52589/AJSTE-JZDPAPUK
DOI URL: https://doi.org/10.52589/AJSTE-JZDPAPUK