American Journal of Applied Mathematics and Statistics, 2018, Vol. 6, No. 5, 186-200
Available online at http://pubs.sciepub.com/ajams/6/5/3

DOI:10.12691/ajams-6-5-3

Modelling Diabetes Mellitus among Adult Kenyan Population Using Artificial Neural Network

Pius Miri Ng'ang'a[1,*], Antony Waititu Gichuhi[2], Antony Wanjoya[2], Thomas Mageto[2]

[1]Population and Social Statistics, Kenya National Bureau of Statistics, Nairobi, Kenya
[2]Department of Statistics and Actuarial Sciences, Jomo Kenyatta University of Agriculture and Technology, Nairobi, Kenya
*Corresponding author: piusmiri@gmail.com

Received August 10, 2018; Revised September 17, 2018; Accepted October 07, 2018

Abstract Artificial Neural Network (ANN) is a parallel connection of a set of nodes called neurons which mimic biological neural system. Statistically, ANN represents a class of non-parametric models which is capable of approximating a non-linear function by a composition of low dimensional ridge functions. This study aimed at modeling diabetes mellitus among adult Kenyan population using 2015 stepwise survey data from Kenya National Bureau of Statistics. Data analysis was carried out using R statistical software version 3.5.0. Among the input variables Age, Sex, Alcoholic status, Sugar consumption, Physical Inactivity, Obesity status, Systolic and Diastolic blood pressure had a significant relationship with diabetic status at 5% level of significance. A multi layered feed-forward neural network with a back propagation algorithm and a logistic activation function was used. Considering a parsimonious model, the model selected had the eight input variables with two neurons in the hidden layer since it gave a minimum MSE of 0.0580 reported. 75% of data was used for training while 25% was used for testing. The sensitivity of the trained network was reported as 75% while specificity was 94.29%. The overall accuracy of the model was 84.64% . This implied that the model could correctly classify an individual as either diabetic or not with an accuracy rate of 84.64%. A 10-fold cross validation was carried out and an average MSE of 0.0686 reported. Kolmogorov-Smirnov test of normality was carried out and at 5% level of significance, for most parameter estimates, we failed to reject the null hypothesis and concluded that the network parameter estimates were asymptotically normal and consistent. With a good choice of risk factors for diabetes, neural network structures could be successfully used to accurately model diabetes melitus among Kenyan adult population.

Keywords: *artificial neural network, diabetes mellitus, activation function, asymptotic normality, accuracy*

Cite This Article: Pius Miri Ng'ang'a, Antony Waititu Gichuhi, Antony Wanjoya, and Thomas Mageto, "Modelling Diabetes Mellitus among Adult Kenyan Population Using Artificial Neural Network." *American Journal of Applied Mathematics and Statistics*, vol. 6, no. 5 (2018): 186-200. doi: 10.12691/ajams-6-5-3.

1. Introduction

Artificial Neural Networks have recently received a great deal of attention in many fields of study. This is due to the fact that ANN attempts to model the capabilities of human brain. They have been used in a variety of applications where statistical methods are traditionally employed. Globally, an estimated 422 million adults were living with diabetes in 2014, compared to 108 million in 1980. The global prevalence (age-standardized) of diabetes has nearly doubled since 1980, rising from 4.7% to 8.5% in the adult population. This reflects an increase in associated risk factors such as being overweight or obese. Over the past decade, diabetes prevalence has risen faster in low- and middle-income countries than in high-income countries [1]. Diabetes caused 1.5 million deaths in 2012. Higher-than-optimal blood glucose caused an additional 2.2 million deaths, by increasing the risks of cardiovascular and other diseases. Forty-three percent of these 3.7 million deaths occur before the age of 70 years. The percentage of deaths attributable to high blood glucose or diabetes that occurs prior to age 70 is higher in low- and middle-income countries than in high-income countries [1].

Diabetes can be classified as type 1(which requires insulin injections for survival) and type 2(where the body cannot properly use the insulin it produces). The majority of people with diabetes are affected by type 2 diabetes. This used to occur nearly entirely among adults, but now occurs in children too.

Sophisticated laboratory tests are usually required to diagnose diabetes. To complement this, researchers are nowadays turning to use of computer based diagnoses which sometimes can be more accurate than the clinical diagnosis. One such computer based diagnosis is the use of Artificial Neural Network. The neural network, firstly developed in 1943, is a part of artificial intelligence developed to predict a model outcome. When the output of the network is discrete, then this is a classification and when the output has continuous values it is performing prediction [2]. This is a suitable and powerful tool to help doctors in the medical field with several advantages such

as the ability to deal with a great amount of data and reduced time of diagnoses. The ability of neural networks to produce good prediction results in classification and regression problems has motivated its use on data related to health outcomes such as death or illness diagnosis [3], [4]. In such studies, the dependent variable of interest is a class label, and the set of possible explanatory variables which are the inputs to the neural networks may be binary or continuous. In this study, ANN was used to classify the individual either as diabetic or non-diabetic based on input variables. The input variables were the physical risk factors for diabetes (Age, Sex, Smoking behavior, Alcoholic status, Salt consumption, Sugar consumption, Physical Inactivity) and secondary risk factors (Obesity status, Systolic and Diastolic blood pressure).

Diagnostics of diseases is broad and challenging area. Its task is to detect a disease that patient with the symptoms have. This process is very complicated, because not all disease's symptoms are specific to only one disease and often the symptoms overlaps. Errors caused by human factor are not rare in this process. To eliminate human error, in modern medicine, different technologies are used nowadays. Some of them are clinical decision support systems. Using information about a patient's condition in the mathematical model, the probable diagnosis can be determined. These mathematical models include Artificial Neural Networks. Artificial Neural Networks (ANNs) play a vital role in the medical field in solving various health problems like validating clinical diagnosis of various diseases.

The main objective of this study was to apply artificial neural network in diagnosing diabetes mellitus among Kenyan adult population. Specifically, the study aimed at: (1) determinining the relationship between diabetes mellitus status and various risk factors, (2) exploring the asymptotic properties of Artificial Neural Network parameter estimates and (3) ascertaining the best Artificial Neural Network models for diagnosing diabetes.

2. Literature Review

Generally, in Kenya not much of the study has been carried out to diagnose diabetes mellitus using ANN for adult Kenyan population. However, a lot of research has been done using ANN in medical diagnosis worldwide

2.1. Classical Models Versus ANN models for Prediction and Diagnosis

Some of the classical Statistical tools applied for prediction and diagnosis in many disciplines are Discriminant analysis [5,6]; Logistic regression [7]; Bayesian approach [8] and Multiple Regression [9,10,11,12]. All these models have been proven to be very effective for solving relatively less complex statistical problems [13]. On the other hand, real world problems are very complex in nature and as such classical models rely heavily on priori assumptions.

To overcome this problem, Artificial Neural networks are increasingly becoming important due to the following reasons. First, as opposed to the classical model-based methods, ANNs are data-driven self-adaptive methods in that there are few a priori assumptions about the models for problems under study. They learn from examples and capture very complex functional relationships among the data even if the underlying relationships are unknown or hard to describe [14]. Second, ANNs can generalize. After learning the data presented to them (a sample), ANNs can often correctly infer the unseen part of a population even if the sample data contain noisy information. Third, ANNs are universal functional approximators. It has been shown that a network can approximate any continuous function to any desired accuracy [15,16,17,18,19]. ANNs have more general and flexible functional forms than the traditional statistical methods can effectively deal with. Due to these properties, ANN is increasingly becoming popular as compared to traditional statistical models.

2.2. Artificial Neural Network in Medical Diagnosis

Artificial neural networks provide a powerful tool to help doctors to analyze, model and make sense of complex clinical data across a broad range of medical applications. Most applications of artificial neural networks to medicine are classification problems; that is, the task is on the basis of the measured features to assign the patient to one of a small set of classes [20].

There are several reviews concerning the application of ANNs in medical diagnosis. The concept was first outlined in 1988 in the pioneering work of [21] and since then many papers have been published. In his work, [22] used artificial neural networks to find potent combination of key variables which accurately identified specific analytes and their level of toxicity. He found that ANN can find potent biomarkers embedded in any type of expression data, mainly proteins which systematically identify the treatment classes of interest with a near 100% accuracy. Whether these proteins are useful in actual diagnosis is tested by presenting the computer model with unknown classes.

Reference [23] developed one of the most successful application of ANN in clinical diagnosis of myocardial infarction. He trained ANN on a group of 356 patients with and without acute myocardial infarction in a cardiac intensive care setting. Using a multi-layer feed forward network trained using a back propagation algorithm, the ANN had unprecedented sensitivity of 92% and a specificity of 96%

2.3. Artificial Neural Network in Diabetic Mellitus Diagnosis

Application of Artificial Neural Network in diagnosing diabetes mellitus has been extensively used by various authors specifically using the Pima Indian data set taken from the UCI machine learning repository. This database has a well validated data resource for exploring the prediction and classification of diabetes mellitus. The data set has eight attributes i.e Number of times pregnant, Plasma glucose concentration (a 2 h in an oral glucose tolerance test), Diastolic blood pressure (mm Hg), Triceps skin fold thickness (mm), 2-h serum insulin (lU/ml), Body mass index (weight in kg/(height in metres)^2), Diabetes pedigree function and Age (years).

Various researchers have used different algorithms and techniques to compare the various classification accuracies obtained. [24] applied neural network classification to Pima Indian diabetes dataset. Using various combinations of pre-processing and missing value techniques, the experimental system achieved an excellent classification accuracy of 99% which is among the best.

Reference [25] applied artificial neural network using Levenberg-Marquardt (LM) algorithm and a probabilistic neural network(PNN) structure to pima Indian data set to diagnose diabetes. They obtained an accuracy of 82.37% and 78.13% using Multi-Layer Neural Network (LM algorithm) and PNN respectively.

Reference [26] used the same Pima data set for diagnosing diabetes onset. They used multilayer feed-forward neural network with back propagation training algorithm to classify patients as diabetic and not diabetic. Using a sigmoid transfer function for the hidden and the output layer and a momentum rate of 0.66 and a learning rate of 0.33, they obtained a classification accuracy of 82%. Comparing this classification accuracy to other algorithms, multilayer feed-forward trained with back-propagation algorithms was higher than other algorithms like nearest neighbor with backward sequential selection of feature.

Reference [27] in their work developed Artificial Neural Network models using both classification and predictive neural networks for the rapid diagnosis of diabetes mellitus. They used a dataset with 465 records which were divided into 440 training data sets and 25 testing data sets. The classification network which was trained using Genetic learning had 19 input variables and the target output variables was the "Diagnosis". The classification results for the training data set showed that 88.41% of the data was correctly classified while 76% of the test set was correctly classified. Generally, both neural network models were able to learn the problem with the predictive network giving a better performance of 84% correctly classified records as opposed to 76% achieved by the classifier network on the same data set.

Reference [28] proposed a method to predict diabetes mellitus using back propagation algorithm of Artificial Neural Network. They treated the problem of diagnosing diabetes as a binary classification i.e those predicted to be diabetic falling under category 1 and non-diabetic under category 0. They used the supervised multilayer feed-forward network architecture with back propagation algorithm. The input parameters used were: Random Blood Sugar test result, Fasting Blood Sugar test result, Post Plasma Blood Sugar test, age, sex and occupation. They measured the performance of the network in terms of absolute error calculated between network response and desired target. The network achieved a classification accuracy of 92.5%. i.e the model was able to predict whether a person was diabetic or not at 92.5% accuracy.

As in [29], used neural network based rule discovery system to determine the presence of hypoglycemic episodes based on the type 1 diabetic patients' physiological parameters, rate of change of heart rate, corrected QT interval of electrocardiogram signal and rate of change of corrected QT interval. He used a sample size of 420 patients with 320 data sets used to develop the neural network based rule discovery system and 100 data sets used to validate its performance. The sensitivity and specificity were found as 79.30% and 60.53% respectively which are considered to be reasonable and better than the ones found by the commonly used methods, statistical regression, genetic programming and fuzzy regression.

3. Methodology

The study utilized secondary data from 2015 Kenya Stepwise survey for Non Communicable Diseases risk factors. Artificial Neural network was used to classify diabetic and non diabetic patients using several input variables (diabetes risk factors). More specifically, a multi layered feed-forward neural network with logistic activation function was used. a 10-fold Cross validation was carried out to validate the model.

3.1. Study Area

The study was carried out in all the forty seven counties of Kenya as shown in Figure 2. A nationally representative sample was selected from the fifth National Sample Survey and Evaluation Programme (NASSEP V) Frame.

3.2. Study Subjects

The recommendation for STEPs was to draw sample population from the targeted population by use of age-sex groups. The age groups used intervals of 12 years of individuals aged 18 years to 69 years. The population covered by the 2015 Kenya STEPS survey was defined as the universe of non-institutionalized population of men and women aged 18 - 69 years. A sample of households was selected and one person identified within the age groups of interest in the households was eligible for interview and measurements [30].

3.3. Sample Size Determination

Following the recommendations detailed in STEP-wise approach to surveillance (STEPS) manual, the survey drew sample population from the targeted population by use of age-sex groups. The age groups used intervals of 12 years of population age 18 years to 69 years, resulting into eight groups.

The sample size was calculated using the formula;

$$n = \frac{Z^2 P(1-P)}{e^2},$$

where

n = Sample size,

Z = Level of confidence,

P = Baseline label of selected indicator,

e = Margin of error.

Using the values, $P = 0.5$, $Z = 1.96$ (95 percent confidence Interval), $P = 50$ percent (as recommended by WHO for countries who have not conducted a STEPS survey before) and $e = 0.05$, the initial estimated sample size was 384. Further adjustments that included multiplication of the sample by 1.5 (design effect to cater for complex survey),

8 (the number of 12 year age-sex groups and 1.25 (to cater for 20 percent non-response) yielded a sample of 5,760. The sample was further adjusted to ease allocation into various strata.

The sample was allocated into all the 92 strata in the NASSEP V frame, ensuring that a minimum of two clusters were selected per strata. This was achieved using power allocation method.

The sample size for 2015 Kenya STEPS survey was 6,000 individuals selected from a total of 200 clusters (100 in urban and 100 in rural) with a uniform sample of 30 individuals per cluster [30].

3.3.1. Sample Inclusion and Exclusion Criteria

The inclusion criteria was:

i). Individuals aged between 18 and 69 years.

ii). Willing and able to provide informed consent for participation.

The exclusion criteria was:

i). Individuals not aged between 18 and 69 years.

ii). Unable or unwilling to provide informed consent or assent.

3.3.2. Sampling Strategy

a) Sample Frame

Administratively, Kenya is divided into 47 Counties. In turn, each county is subdivided into Sub- Counties. Prior to the enactment of the current constitution in 2010, the sub-counties had not been created but similar units were the districts. Each district was divided into divisions, each division into locations and each location into sub-locations. In addition to these administrative units, prior to the 2009 population census, each sub-location was subdivided into census enumeration areas (EAs) i.e. small geographic units with clearly defined boundaries. A total of 96,251 EAs were developed. The list of EAs is grouped by administrative units and includes information on the number of households and population. This information was used in 2010 to design a master sample known as the fifth National Sample Survey and Evaluation Programme (NASSEP V) with a total of 5,360 selected EAs [30].

The NASSEP V master frame follows a two-stage stratified cluster sample format. The first stage involved selection of Primary Sampling Units (PSUs) which were the EAs using probability proportional to size (PPS) method, with the measure of size being the households from 2009 census. The second stage involves the selection of households for various surveys. The frame was designed in a multi-tied structure with four sub-samples (C1, C2, C3 and C4), each consisting of 1,340 EAs that can serve as independent frames. The NASSEP V frame used the counties as the first level stratification and further sub divided into rural and urban sub domains. The sampling was done independently within rural - urban sub domains. Each sampled EA was developed into a cluster and undergone listing and mapping process and clusters are within measure of size of average of 100 households (between 50 households and 149 households) [30].

b) Sample Selection

The 2015 Kenya STEPS survey sample was selected in three stages. Stage one involved selection of PSUs (i.e. clusters), households and individuals.

c) Selection of PSUs

The selection of clusters was done using the Equal Probability Selection Method (EPSEM). The clusters were selected systematically from NASSEP V frame with equal probability independently within the urban-rural domains. The process involved ordering the clusters by county, then by urban/rural, and finally by unique geocode. The resulting sample retained properties of PPS as used in creation of the frame.

d) Household selection

Using the total number of households from each sampled cluster available from the NASSEP V, a uniform sample of 30 households per cluster was selected using systematic sampling method. This procedure of selecting the sample households with a random start was done by the following criteria:

Let L be the total number of households listed in the cluster;

Let R be a random number between (0, 1);

Let n be the number of households selected in the cluster;

Let $I = L/n$ be the sampling interval.

1. The first selected sample household is k (k is the serial number of the household in the listing) if and only if:

$$\frac{k-1}{L} < R \le k/L$$

2. The subsequent selected households are those having serial numbers: $k+(j-1)*I$ (rounded to integers) for $j = 2,3,...,n$. Random numbers were different and independent from cluster to cluster [30].

e) Individual selection

All the selected clusters and corresponding households were loaded into Personal Digital Assistants (PDAs). During interviews, all the eligible household members were listed down and PDA used to randomly select one for interviews using the inbuilt Kish Grid method [30].

3.4. Statistical Model

Artificial Neural Network (ANN) was used to classify individuals as either diabetic or not based on physical and behavioural characteristics as input variables. Since secondary data was used in this study, it will be first cleaned by checking missing data and outliers. Outliers will be excluded in the final analysis for the model. Chi square test will be carried out to determine the relationship between diabetes mellitus status and various risk factors.

At the inferential stage, a multi layered feed-forward neural network with logistic activation function model will be used to fit the data. Schwarz information Criterion (SIC), will be used for model selection. Classification Accuracy rate and Mean squared error (MSE) will be reported. To validate our diagnosis model, a 10 fold cross validation will also be carried out.

3.4.1. Chi-Square Test of Independence

In order to determine the relationship between diabetes mellitus status and various risk factors, Chi-square test of independence/no relationship was carried out. Two variables are said to be statistically independent if the population conditional distributions of Y are identical at

each level of . When two variables are independent, the probability of any particular column outcome j is the same in each row. Statistical independence is, equivalently, the property that all joint probabilities equal the product of their marginal probabilities, $\pi_{ij} = \pi_{i+}\pi_{+j}$ for $i = 1,...,I$ and $j = 1,...,J$; that is, the probability that X falls in row i and Y falls in column j is the product of the probability that X fall in row i with the probability that Y falls in column j [31].

Consider the null hypothesis (H_0) that cell probabilities equal certain fixed values $\{\pi_{ij}\}$. For a sample of size n with cell counts n_{ij}, the values $\{\mu_{ij} = n\pi_{ij}\}$ are expected frequencies. They represent the values of the expectations $\{E(n_{ij})\}$ when H_0 is true. To judge whether the data contradict H_0, we compare $\{n_{ij}\}$ to $\{\mu_{ij}\}$. If H_0 is true, n_{ij} should be close to μ_{ij} in each cell. The larger the difference $\{n_{ij} - \mu_{ij}\}$, the stronger the evidence against H_0. The Pearson test statistic is used to make such comparisons and it has large-sample chi-squared distributions [31].

The Pearson chi-squared statistic for testing H_0 is:

$$\chi^2 = \sum \frac{(n_{ij} - \mu_{ij})^2}{\mu_{ij}}.$$

This statistic takes its minimum value of zero when all $n_{ij} - \mu_{ij}$. For a fixed sample size, greater differences $\{n_{ij} - \mu_{ij}\}$ produce larger χ^2 values and stronger evidence against H_0. Since larger χ^2 values are more contradictory to H_0, the P-value is the null probability that χ^2 is at least as large as the observed value. The χ^2 statistic has approximately a chi-squared distribution, for large n. The P-value is the chi-squared right-tail probability above the observed χ^2 value. The chi-squared approximation improves as μ_{ij} increase, and $\{\mu_{ij} \geq 5\}$ is usually sufficient for a decent approximation as discussed in [31].

The chi-squared distribution is concentrated over nonnegative values. It has mean equal to its degrees of freedom df, and its standard deviation equals $\sqrt{2df}$. As df increases, the distribution concentrates around larger values and is more spread out.

3.4.2. Relative Risk

This is a ratio of two proportions. For 2×2 tables, the relative risk is the ratio,

$$\text{Relative Risk} = \frac{\pi_1}{\pi_2}.$$

A relative risk of 1 occurs when $\pi_1 = \pi_2$ i.e when the response is independet of the group. [31]

3.4.3. Introduction to Neural Network

An artificial neural network (ANN) is a parallel connection of a set of nodes called neurons which mimic biological neural system. Statistically, ANN represents a class of non parametric models which is capable of approximating a non linear function by a composition of low dimensional ridge functions [33]. It represents a function of explanatory variables which is composed of simple building blocks and which may be used to provide an approximation of conditional expectations or, in particular, probabilities in regression [34]. ANN is widely used in classification, regression and statistical pattern recognition problems.

3.4.4. Definition of the ANN

Consider a feed-forward net with $d+1$ input nodes, one layer of m hidden nodes, one output node and an activation function $\psi(x)$. The input and hidden layer nodes are connected by weights W_{hj} for $h \in \{1,...,m\}$ and $j \in \{0,...,d\}$. The hidden and output layers are connected by weights α_h for $h \in 0,...,m$ where α_0 is the weight from the bias node to the output node [34]. Considering an input vector $\mathbf{X} = (x_1,...,x_d) \in R^d$, then the input $\upsilon_h(\mathbf{X})$ to the h^{th} hidden node is the value

$$\upsilon_h(\mathbf{X};\mathrm{q}) = W_{h0} + \sum_{j=1}^{d} W_{hj} x_j. \quad (1)$$

The output $\phi_h(\mathbf{X};\mathrm{q})$ of the h^{th} hidden node is the value

$$\phi_h(\mathbf{X};\mathrm{q}) = \psi(\upsilon_h(\mathbf{X};\mathrm{q})). \quad (2)$$

The net input to the output node is the value

$$O_m(\mathbf{X};\mathrm{q}) = \alpha_0 + \sum_{h=1}^{m} \alpha_h \phi_h(\mathbf{X};\mathrm{q}). \quad (3)$$

Finally, the output $g(\mathbf{X};\mathrm{q})$ of the network is the value

$$g(\mathbf{X};\mathrm{q}) = \psi(O_m(\mathbf{X};\mathrm{q})) \quad (4)$$

We note that q stands for all the parameters $\alpha_0,...,\alpha_m$ and $W_{hj}, h = 1,...,m, j = 0,...,d$ of the network [34]. We also write $\mathrm{a} = (\alpha_0,...,\alpha_m)^t$ and

$$\mathbf{W} = (W_{hj}, h = 1,...,m, j = 0,...,d)$$

denoting them as vectors. In prediction and classification problems, the activation function $\psi(x)$ is usually chosen to be symmetric sigmoidal function i.e fixed bounded continuous non decreasing function.

$$\psi(x) = \begin{cases} 1, & x \to \infty \\ 0, & x \to -\infty \\ \psi(x) + \psi(-x) = 1 \end{cases} \quad (5)$$

The most appropriate choice of the activation function above is the logistic function given as

$$\psi(x) = \frac{exp(a(x-b))}{1+exp(a(x-b))} = \frac{1}{exp(-a(x-b))},$$

where α is the learning rate while b is called the bias.

In this study, we assumed a statistical model that relates Y and $g(\mathbf{X};\mathrm{q})$ as follows:

$$Y = g(\mathbf{X};\mathrm{q}) + \varepsilon, \tag{6}$$

and $\varepsilon \sim N(0,\sigma^2)$ is the error term.

The network is trained on the dataset

$$(Y_1, X_1),...,(Y_n, X_n);$$

i.e these data are used to come up with an estimator $\hat{\mathrm{q}}$ for θ [34].

3.4.5. Training the Network

There are two types of network training i.e Supervised and unsupervised learning. In this study supervised training will be used. The supervised training of a neural net requires the following:

1. A sample of n input vectors, $X = X_1,..,X_n \in R^d$ of size d each and an associated output vector

$$Y = Y_1,...,Y_n \in R.$$

2. The selection of an initial weight set.
3. A repetitive method to update the current weights to optimize the input-output map.
4. A stopping rule

The maximum likelihood method is used to find the optimal estimator $\hat{\mathrm{q}}$ for the network [34].

The task here is to minimize the error in equation (6). The conditional density of Y_i given $X_i = x$ is given as:

$$f(\mathbf{Y} \mid \mathbf{X}) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\{\frac{-(\mathbf{y} - g(\mathbf{X};\mathrm{q}))^2}{2\sigma^2}\},$$

so that the log-likelihood function is given by

$$L = \frac{-\sum_{i=1}^{n}(Y_i - g(\mathbf{X};\mathrm{q}))^2}{2\sigma^2} - \frac{n}{2}ln(2\pi) - \frac{n}{2}ln(2\sigma^2). \tag{7}$$

The second and the third term of the above equation is independent of the weights q and therefore can be omitted so that maximizing equation (7) is equivalent to minimizing

$$S(\mathbf{Y},\mathbf{X};\mathrm{q}) = \frac{1}{2}\sum_{i=1}^{n}(Y_i - g(\mathbf{X};\mathrm{q}))^2. \tag{8}$$

The weights are then adjusted in such a way that the error function in equation (8) is minimized. However, this study is on classification and the target variable is binary. The probability weights of Y_i given $X_i = x$ are

$$\pi(\mathbf{Y} \mid \mathbf{X}) = g(\mathbf{X};\mathrm{q})^Y (1 - g(\mathbf{X};\mathrm{q}))^{1-Y}; Y = 0,1 \tag{9}$$

and the likelihood of equation (9) is given by

$$L = \prod_{i=1}^{n} g(\mathbf{X};\mathrm{q})^{Y_i} (1 - g(\mathbf{X};\mathrm{q}))^{1-Y_i} \tag{10}$$

and the negative of the log likelihood is given as

$$S(\mathbf{Y},\mathbf{X};\mathrm{q}) = -\sum_{i=1}^{n}\{Y_i \ln(g(\mathbf{X};\mathrm{q})) + (1-Y_i)\ln(1 - g(\mathbf{X};\mathrm{q}))\} \tag{11}$$

where

$$g(\mathbf{X};\mathrm{q}) = \psi(O_m(\mathbf{X};\mathrm{q})) \tag{12}$$

$\hat{\mathrm{q}}$ is the value of q that maximizes the equation above i.e

$$\hat{\mathrm{q}} = arg\min_{\mathrm{q}\in\Theta_0} S(\mathbf{Y},\mathbf{X};\mathrm{q}). \tag{13}$$

In equation (11), the weights are adjusted in such a way that the error between the targets Y and the actual output $g(\mathbf{X};\mathrm{q})$, is minimized. The goodness of the network approximation can be evaluated using a penalty function, π, that measures how well network output $g(\mathbf{X};\mathrm{q})$ matches the "target" output y corresponding to given inputs x. Since the output is binary, negative entropy is a good penalty [35]. Performance as a function of q for given x and y can be measured as $q(y,x,\theta) \equiv \pi(y, g(X;\theta))$. A measure of overall network performance is given by the expected penalty, $Q(\theta) \equiv E(q(Y,X,\theta))$, where the random target/input pair (Y,X) is drawn from the population distribution governing the phenomenon of interest. Choosing q to solve $\min_{\mathrm{q}\in\Theta_0} Q(\mathrm{q})$, yields a network producing the smallest average penalty, given an input randomly drawn from the operating environment. This provides an objective way to choose the "best" approximation and formalizes the requirement that the network "generalizes" well. There are various methods of minimizing equation (8). These include Backpropagation, Quasi-Newton method and Simulated annealing method.

In this study, back propagation method was used to minimize the error.

3.4.6. Back Propagation Method

This is a kind of coordinate wise gradient descent method. The goal is to find a set of weights $\mathrm{a} = (\alpha_0,...,\alpha_m)^t$ and $\mathbf{W} = (W_{hj}, h = 1,...,m, j = 0,...,d)$ that minimizes our objective function, equation (11). Therefore, the partial derivative of the objective function with respect to a weight represents the rate of change of the error function with respect to that weight (it is the slope of the objective function). Moving the weights in a direction down this slope will result in a decrease in the objective function. This intuitively suggests a method to iteratively find values for the weights. We evaluate the partial derivative of the objective function with respect to the weights, and then move the weights in a direction down the slope, continuing until the error function no longer decreases [36]. Mathematically, the weights are adjusted as follows, taking a unipolar activation function $\psi(x)$:

$$\mathbf{W}^{r+1} = \mathbf{W}^{r} + \Delta\mathbf{W}$$

$$\mathrm{a}^{r+1} = \mathrm{a}^{r} + \Delta\mathrm{a}.$$

Taking individual weights, we have the r^{th} iteration weight as

$$\alpha_h^{r+1} = \alpha_h^r - \lambda_1\{\frac{\partial S(\mathbf{Y},\mathbf{X};\mathrm{q}^{(\mathbf{r})})}{\partial \alpha_h}\}$$

for $i = 1,...,n$ and $h = 1,...,m$.

Similarly,

$$W_{hj}^{(r+1)} = W_{hj}^{r} - \lambda_2\{\frac{\partial S(\mathbf{Y},\mathbf{X};\mathrm{q}^{(\mathbf{r})})}{\partial W_{hj}}\}$$

for $i = 1,...,n$ and $h = 1,...,m$ and $j = 1,...,d$, with λ_1 and λ_2 representing the step gain [34].

3.4.7. Parameter Estimation

We first discuss the concept of existence of the estimator $\hat{\mathrm{q}}$. Existence of a solution to equation (13) is guaranteed by the following lemma with assumption that Θ is compact [34].

Lemma 1. *Assume* (11) *and* (12) *holds, then there exists a solution of the maximum likelihood equation* (13).

Proof. By our choice of $\psi(.)$ and $O_m(.)$, $g(\mathbf{X};\mathrm{q})$ given by (12) is continuous in X and θ, and $0 < g(\mathbf{X};\mathrm{q}) < 1$ for all $\mathbf{X},\mathrm{q}$. Therefore, $S(\mathbf{Y},\mathbf{X}\mathrm{q})$ is continuous in q for all $\mathbf{Y},\mathbf{X}$, and it assumes it minimum on compact sets.

Next we discuss the concept of the model irreducibility/Redudancy.

We say that a neural network (with a fixed set of parameters) is "redundant" if there exists another network that represents exactly the same relationship function $g_{\mathrm{q}}(.)$. A related definition is the reducibility of q stated by [37] as follows.

Definition: For ψ satisfying equation (5),

$$\mathrm{q} = (\alpha = (\alpha_0,...,\alpha_m)^t$$

and

$$\mathbf{W} = (W_{hj}, h = 1,...,m, j = 0,...,d))$$

is called reducible if one of the following three conditions holds for $h \neq 0$ and $j \neq 0$.

a) $\alpha_h = 0$ for some $h = 1,...,m$

b) $W_{hj} = 0$ for some $h = 1,...,m$ or

c) $(W_{hj}, W_{h0}) = \pm(W_{jh}, W_{j0})$ for some $i \neq j$, where $\mathbf{0}$ denotes the zero vector of the appropriate size.

A reducible q with symmetric sigmoidal f leads to a redundant network because it gives a $g_{\mathrm{q}}(.)$ function that can be represented by another network by deleting the h^{th} neuron, where is described in the conditions above. For condition (a), it is obvious. For (b), delete the h^{th} neuron and replace α_0 by $\alpha_0 + \alpha_h f(W_{h0})$. In (c), if $(W_{hj}, W_{h0}) = (W_{jh}, W_{j0})$, then we can delete the h^{th} neuron and replace α_j by $\alpha_h + \alpha_j$. On the other hand, if $(W_{hj}, W_{h0}) = -(W_{jh}, W_{j0})$, then we can replace α_j and α_0 by $\alpha_j - \alpha_h$ and $\alpha_0 + \alpha_h$ because

$$\alpha_h f(W_{hj}\mathbf{X} - W_{h0}) = \alpha_h - \alpha_h f(-W_{hj}\mathbf{x} - W_{h0}). \quad (14)$$

3.5. Model Identifiability

This is a fundamental problem in neural network. The parameters are not unique since we have a different set of parameters with an identical distributions of (Y, X) [38]. Let the weights be represented as follows:

$$\alpha_0 \text{ and } \mathrm{b}_{\mathbf{i}} = (\alpha_i, \mathbf{W}_{\mathbf{i}} \text{ for } i = 1,...,m) \quad (15)$$

where $\mathbf{W}_{\mathbf{i}} = (W_{i0}, W_{i1},..., W_{id})$.

At this point we note two kinds of transformations that make the input-output map invariant:

i) The function is unchanged if we permute $\beta_{i'}s$. For example if β_1 and β_2 are interchanged, $g(\mathbf{X};\mathrm{q})$ remains unchanged.

ii) Equation (14), can be used to establish that the parameters $(\alpha_0, \mathrm{b}_1,..., \mathrm{b}_i,..., \mathrm{b}_m)$ and

$$(\alpha_0 + \alpha_i, \beta_1,..., \mathrm{b}_i,..., \mathrm{b}_m)$$

gives exactly the same value of $g(\mathbf{X};\mathrm{q})$ and hence the same distribution of Y.

The transformations described above generate a family with $2^m m!$ elements. Call this family of transformations λ. For all transformation λ in this family,

$$g(\mathbf{X};\mathrm{q}) = g_{\lambda}(\mathbf{X};\mathrm{q}).$$

Each transformation can be characterized as being composite function of $\{\lambda_1,..,\lambda_m\}$, where

$$\lambda_1((\alpha_0, \beta_1,..., \beta_i,..., \beta_m)) = (\alpha_0 + \alpha_i, -\beta_1, \beta_2,..., \beta_m)$$
and
$$\lambda_1((\alpha_0, \beta_1,..., \beta_i,..., \beta_m)) \quad (16)$$
$$= (\alpha_0, \beta_i, \beta_2,..., \beta_{i-1}, \beta_{i+1},..., \beta_m) \text{ for } i = 2,...,m.$$

The following two conditions must be satisfied by the activation functions.

1. Condition A: The class of functions

$$\{f(bx + b_0), b > 0\} \cup \{f \equiv 1\}$$

is linearly independent. More precisely, for any positive integer m and any scalars a_0, a_i, b_{i0} and $b_i > 0, i = 1,...,m$ with $(b_i, b_{i0}) \neq (b_j, b_{j0})$ for every $i \neq j$, the condition

$$a_0 + \sum_{i=1}^{m} a_i f(b_i x + b_{i0}) = 0 \qquad \forall x \in \mathbb{R}$$

Implies that

$$a_0 = a_1 = ... = a_m = 0$$

2. Condition B: Assume that f is differentiable and f' is its derivative. The class of functions

$$\{f(bx+b_0), b>0\}\cup\{f'(bx+b_0), b>0\}\cup$$
$$\{xf'(bx+b_0), b>0\}\cup\{f\equiv 1\}$$

is linearly independent.

As a result of the above two conditions and assuming models (4), (5) and (6) with a continuous function f satisfy condition A. (NB: $f\equiv\psi$). Suppose that q is irreducible. Also assume that the distribution of x has the support $\mathbb{R}^d$. Then the following apply as discussed by [38]:

a) q is **identifiable** up to the family of transformations generated by (16). That is, if there exists another q^* such that $g(\mathbf{X};\mathrm{q}^*)=g(\mathbf{X};\mathrm{q})$, then there exist a transformation generated by (16) that transforms q^* to q .

b) Under further assumption that f is continuously differentiable and satisfies condition B, the matrix $\mathrm{S}=E\{[\nabla_\theta g_\theta(x)][\nabla_\theta g_\theta(\mathbf{x})]^t\}$ is non singular. Here $\nabla_\theta g_\theta(\mathbf{x})$ is a column vector and denotes the gradient of $g_\theta(\mathbf{x})$ hence S is a square matrix. Also, the expectation E is taken with respect to the random vector $\mathbf{x}$.

Any non decreasing symmetric sigmoidal function that satisfies condition B also satisfies condition A. [38]. Also any non decreasing function satisfying the first two properties of equation (5) must be a cumulative distribution function (cdf) of a one dimensional random variable. Condition A says that $\{f(bx+b_0), b>0\}$ are independent, which is equivalent to the mixture probability density functions being **identifiable**.

3.5.1. Consistency and Asymptotic Normality of Network Parameter Estimates

Assume that the set $(Y_i,\mathbf{X}_i), i=1,...,n$ are i.i.d with conditional probability distribution

$$\pi(Y_i\mid\mathbf{X}_i=\mathbf{x})=B(1,p(\mathbf{x})) \tag{17}$$

we will fit a neural network output function $g(\mathbf{X};\mathrm{q})$ to $p(\mathbf{x})$ by minimizing the negative log likelihood equation (11) multiplied by $1/n$.

$$S(\mathbf{Y},\mathbf{X};\mathrm{q})=-\frac{1}{n}\sum_{i=1}^{n}\{Y_i\ln(g(\mathbf{X};\mathrm{q}))+(1-Y_i)\ln(1-g(\mathbf{X};\mathrm{q}))\}. \tag{18}$$

Let $S_0(\mathrm{q})=E[S(\mathrm{q})]$ denote the expectation of the target function $S(\mathrm{q})$. Since $(Y_i,\mathbf{X}_i)$ are i.i.d, we have

$$S(\mathrm{q})=-E\begin{bmatrix}Y_1\ln(g(\mathbf{X_1};\mathrm{q}))\\+(1-Y_1)\ln(1-g(\mathbf{X_1};\mathrm{q}))\end{bmatrix}=-E\begin{bmatrix}p(x_1)\ln(g(\mathbf{X_1};\mathrm{q}))\\+(1-p(x_1))\ln(1-g(\mathbf{X_i};\mathrm{q}))\end{bmatrix} \tag{19}$$

Assume that $S_0(\mathrm{q})$ has a unique minimum if q ranges over a given compact set Q . Then this minimum is characterized by

$$\mathbf{0}=\nabla S_0(\mathrm{q})=E\left\{\frac{p(x_1)}{g(\mathbf{X_1};\mathrm{q})}-\frac{1-p(x_1)}{1-g(\mathbf{X_1};\mathrm{q})}\right\}\nabla g(\mathbf{X_1};\mathrm{q}) \tag{20}$$

By the fact that equation (4) is continuous in $\mathbf{x}$ and continuously differentiable in q , we may interchange expectation and differentiation.

Since in this study we are dealing with classification problem, the correctly classified case where $p(\mathbf{x})=g(\mathbf{X};\mathrm{q}_0)$ for some $b\mathrm{q}_0\in\Theta$, equation (20) is solved for $\mathrm{q}=\mathrm{q}_0$. i.e $S_0(\mathrm{q})$ is minimized at the true parameter value q_0 . In general if there is no true value, we may define q_0 as

$$\mathrm{q}_0=arg\min_{\mathrm{q}\in\mathrm{Q}}S_0(\mathrm{q}) \tag{21}$$

By minimizing equation (18), we get the estimator $\hat{\mathrm{q}}$. Consistency of this estimator $\hat{\mathrm{q}}$ therefore means that $\hat{\mathrm{q}}$ converges in probability to q_0 as the sample size tends to infinity [34].

Next, we discuss the asymptotic normality of the network parameters. In a classical context, our model can be written as follows:

$$Y_i=p(\mathbf{X}_i)+\varepsilon_i, i=1,...,n. \tag{22}$$

From the above equation, the residuals ε_i can therefore be expressed as;

$$\varepsilon_i=Y_i-p(\mathbf{X}_i). \tag{23}$$

Since $(\mathbf{X}_i,Y_i)$ are i.i.d and

$$P(Y_i=1\mid\mathbf{X}_i)=E(Y_i=1\mid\mathbf{X}_i)=P(\mathbf{X}_i),$$

the residuals ε_i are also i.i.d implying that $E(\varepsilon_i)=0$ and

$$\begin{aligned}var(\varepsilon_i)&=E(Y_i-p(\mathbf{X}_i))^2\\&=E\{E[Y_i-p(\mathbf{X}_i))^2\mid\mathbf{X}_i]\}\\&=E[p(\mathbf{X}_i)(1-p(\mathbf{X}_i))^2+(1-p(\mathbf{X}_i))p^2(\mathbf{X}_i)]\\&=E[p(\mathbf{X}_i)(1-p(\mathbf{X}_i))]=\sigma_i^2<\infty.\end{aligned} \tag{24}$$

Also,

$$var(\varepsilon_i\mid\mathbf{X}_i=\mathbf{x})=\sigma_\varepsilon^2=p(x)(1-p(x)).$$

We note that $var(\varepsilon_i)$ does not depend on q .

Since the residuals ε_i are not only i.i.d but also bounded in absolute value by 1, their assumptions reduce to,

A1). The activation function ψ is bounded and twice continuously differentiable with bounded derivatives.

A2). $S_0(\mathbf{q})$ has a global minimum at $\mathbf{_0}$ lying in the interior of Q and with a positive definite Hessian

$$A(\mathrm{q}_0)=\left(\frac{\partial^2}{\partial \mathrm{q}_k \partial \mathrm{q}_t} S_0(\mathrm{q})\right)=\nabla^2$$

A3). Let Θ be chosen such that for some $\Delta>0$, we have $\Delta \le g(\mathbf{X};\mathrm{q}) \le 1-\Delta$, for all $x \in \Re^d, \theta \in \Theta$.

A4). $(Y_i, \mathbf{X}_i), i=1,...,n$ be i.i.d with unknown density $H(x)$ whose support is $\Re^d$.

A5). $p(x)$ is continuous in $\mathbf{x}$ and $0<\delta \le p(x) \le 1-\delta$ for some $\delta>0$.

Having discussed the necessary theory, we now discuss the asymptotic normality of the network parameter estimates.

Let $(Y_i, X_i), i=1,...,n$ be i.i.d with

$$\pi(Y_i \mid \mathbf{X}_i = \mathbf{x}) = B(1, p(\mathbf{x})).$$

Suppose that assumptions A1 to A5 are satisfied. Then, for $n \to \infty$, with $\hat{\mathrm{q}}, \mathrm{q}_0$ as above

$$\sqrt{n}(\hat{\mathrm{q}}-\mathrm{q}_0) \to^d N(0, \Sigma_1+\Sigma_2)$$

where

$$\Sigma_1 = A^{-1}(\mathrm{q}_0)B_1(\mathrm{q}_0)A^{-1}(\mathrm{q}_0)$$

$$\Sigma_2 = A^{-1}(\mathrm{q}_0)B_2(\mathrm{q}_0)A^{-1}(\mathrm{q}_0)$$

with

$$B_1(\mathrm{q}_0)=E\left\{\frac{p(\mathbf{X}_1-g(\mathbf{X}_1;\mathrm{q}_0))^2}{g^2(\mathbf{X}_1;\mathrm{q}_0)(1-(\mathbf{X}_1;\mathrm{q}_0))^2}\right\}$$

$$\nabla g(\mathbf{X}_1;\mathrm{q}_0).\nabla^T g(\mathbf{X}_1;\mathrm{q}_0)$$

$$B_2(\mathrm{q}_0)=E\left\{\frac{p(\mathbf{X}_1)(1-p(\mathbf{X}_1))}{g^2(\mathbf{X}_1;\mathrm{q}_0)(1-(\mathbf{X}_1;\mathrm{q}_0))^2}\right\}$$

$$\nabla g(\mathbf{X}_1;\mathrm{q}_0).\nabla^T g(\mathbf{X}_1;\mathrm{q}_0)$$

and

$$A(\mathrm{q}_0)=\left(\frac{\partial^2}{\partial \mathrm{q}_k \partial \mathrm{q}_t} S_0(\mathrm{q})\right)=\nabla^2 S_0(\mathrm{q}).$$

We note that the asymptotic covariance matrix, $\Sigma_1+\Sigma_2$ reflects the two sources of error in $B_1(\mathrm{q}_0)$ and $B_2(\mathrm{q}_0)$. $B_1(\mathrm{q}_0)$ contains the squared modeling bias $(p(\mathbf{X}_1)-g(\mathbf{X}_1;\mathrm{q}_0))^2$ which vanishes in the correctly specified case while $B_2(\mathrm{q}_0)$ contains $p(\mathbf{X}_1)(1-p(\mathbf{X}_1))=Var(Y_1 \mid X_1)$ which reflects the randomness in the response variable Y_1 [34].

The asymptotic normality of network parameter estimates was determined by use of normal quantile-quantile (qq) plots.

3.5.2. Normality Test

Kolmogorov-Smirnov test of normality will be used to test our hypotheis. Suppose we have an i.i.d sample $X_1, \ldots, X_n$ with some unknown distribution P and we would like to test our hypotheis that P is equal to a normal distribution P_0.

Lets denote by $F(x)=P(X_1 \le x)$ a c.d.f of a true underlying distribution of the data. We define an empirical c.d.f by $F_n(x)=P(X_1 \le x)=\frac{1}{n}\sum_{i=1}^n I(X_i \le x)$ that counts the proportion of the sample points below level x. For any fixed point $x \in \Re$, the law of large numbers implies that

$$F_n(x)=\frac{1}{n}\sum_{i=1}^n I(X_i \le x) \to EI(X_1 \le x)=F(x),$$

i.e the proportion of the sample in the set $(-\infty,x]$ approximates the probability of this set. It is easy to show that from here that this approximation holds uniformly over all $x \in \Re, \sup|F_n(x)-F(x)| \to 0$. i.e the largest difference between F_n and F goes to 0 in probability [39]. The key observation in Kolmogorov-Smirnov test is that the distribution of this supremum does not dependon the 'unknown' distribution P of the sample if P is continous distribution.

For a fixed point x, the central limit theorem implis that,

$$\sqrt{n}\left(F_n(x)-F(x)\right) \xrightarrow{d} N\left(0, F(x)\left(1-F(x)\right)\right)$$

because $F(x)\left(1-F(x)\right)$ is the variance of $I(X_1 \le x)$, it turns out that $\sqrt{n}\sup|F_n(x)-F(x)|=D_n$, which is the KS statistics [39].

3.5.3. Model Selection and Complexity Regularization

A network model with sufficiently large number of hidden units can approximate any unknown function. When a training sample is fixed, a complex network with a large number of hidden units may over fit the data. Thus, there is a trade off between approximation capability and over-fitting while implementing ANN models. One easy approach to regularizing the network complexity is to use model selection criteria. Two such criteria are the Schwarz Information criterion (SIC) proposed by [25] and Predictive Stochastic Complexity criterion (PSC) introduced by [40].

In this study, we used the Schwarz Information Criterion (SIC) which is given as;

$$SIC(h)=\ln(\hat{\sigma}^2)+(h(2+d)+1)\frac{\ln(n)}{n}. \qquad (25)$$

The first term is the goodness of fit measure (Regression Mean Squared Error) while the second term penalizes model complexity. The Mean Squared Error (MSE) is given by;

$$MSE=\frac{1}{n}\sum_{i=1}^n (y_i-g(\mathbf{X};\mathrm{q}))^2.$$

This MSE was also used to determine the number of hidden neurons but comparison was made with SIC. Using the SIC criterion, we started with a single hidden neuron and determined SIC(1). Then the second hidden neuron was added and SIC(2) determined. The process continued until an extra hidden neuron did not improve the SIC. We therefore estimated $h+1$ models in order to choose a model with h neurons [34].

3.5.4. Cross Validation

Cross-validation is a process that can be used to estimate the quality of a neural network. When applied to several neural networks with different free parameter values (such as the number of hidden nodes and back-propagation learning rate), the results of cross-validation can be used to select the best set of parameter values. The initial data set is divided into k subsets of approximately equal size. The model is then estimated k times, each time leaving out one of the subsets. A series of Mean squared error is computed on the basis of the omitted subset. This method is called leave out one cross validation [41].

3.5.5. Model Assessment

Inorder to assess the fitness of the model, Accuracy, Sensitivity and Specificity were reported. The accuracy of a diagnostic test is often assessed with two conditional probabilities: Given that a subject has the disease, the probability the diagnostic test (prediction) is positive is called Sensitivity [31]. Given that the subject does not have the disease, the probability that the test is negative is called Specificity. The overall accuracy of the model is the average of specificity and sensitivity.

Consider a 2 × 2 table with notation,

		Predicted		
		Diabetic	Non Diabetic	Total
Actual	Diabetic	A	B	A+B
	Non Diabetic	C	D	C+D
	Total	A+C	B+D	A+B+C+D

The sensitivity, Specificity and Accuracy are calculated as follow;

$$Sensitivity = \frac{A}{(A+C)}$$

$$Specificity = \frac{D}{(B+D)}$$

$$Accuracy = \frac{1}{2}(Sensitivity + Specificity).$$

4. Results and Presentations

4.1. The Data

The study utilized secondary data from 2015 Kenya Stepwise survey for Non Communicable Diseases risk factors. The input variables were the physical risk factors i.e. Age, Sex, Smoking behavior, Alcoholic status, Salt consumption, Sugar consumption, Physical activity/Inactivity, Obesity status, Systolic and Diastolic blood pressure, while the output variable was diabetic status (diabetic or not diabetic). An obese person in this study is any person whose Body Mass Index was greater than or equal to 30 while a diabetic person is someone whose fasting glucose was greater than or equal to 6.1mmol/l.

The table below summarizes the variables and its measurements.

Table 1. Input and Output Variables

No	Variable	Description	Measurement	Value
1.	age	Age(years)	Scale	Numeric
2.	sex		Nominal	0=Female
				1=Male
3.	smoke	Smoking behaviour	Nominal	0=Don't smoke
				1=Smoke
4.	alcohol	Alcohol drinking	Nominal	0=Don't drink
				1=Drink
5.	salt	Excess salt consumption	Nominal	0=No
				1=Yes
6.	sugar	Excess sugar consumption	Nominal	0=No
				1=Yes
7.	inactive	Physically inactive	Nominal	0=No
				1=Yes
8.	sbp	Systolic blood pressure (mmHg)	Scale	Numeric
9.	dbp	Diastolic blood pressure (mmHg)	Scale	Numeric
10.	obese	Obese	Nominal	0=No
				1=Yes
11.	diabetic	Diabetic status	Nominal	0=Not Diabetic
				1=Diabetic

4.2. Descriptive Statistics

The table below gives descriptive statistics for continuous variables

Table 2. Descriptive Statistics for Continous Variables

Variable	N	Minimum	Maximum	Mean	Standard Deviation
Age	4115	18	69	37.77	13.438
SBP	4115	80	218	126.60	18.318
DBP	4115	48	129	82.05	11.402

From Table 2, its clear that the age of respondents was well within the survey inclusion criteria. The minimum age was 18 years and the maximum age was 69 years while mean age of respondents was approximately 38 years. The average systolic blood pressure was 126.6mmHg while the average diastolic pressure was 82.05. The SBP ranged from 80mmHg to 218 mmHg while DBP ranged from 48mmHg to 129 mmHg.

Table 3 shows frequency distribution of the categorical input variables. The results from the study showed that 91.2% of respondents did not smoke or had never smoked. It is also clear that, of all the respondents, 10.3% were obese. Only 7.0% of the respondents were diabetic.

Table 3. Frequency Distribution for Categorical Input Variables

Variable	Category	Frequency	Percent
Sex	Female	2414	58.7
	Male	1701	41.3
Smoke	No	3751	91.2
	Yes	364	8.8
Drink Alcohol	No	2735	66.5
	Yes	1380	33.5
Excess Salt Consumption	No	3680	89.4
	Yes	435	10.6
Excess Sugar Consumption	No	3478	84.5
	Yes	637	15.5
Physically Inactive	No	2883	70.1
	Yes	1232	29.9
Obese	No	3682	89.5
	Yes	433	10.3
Diabetic	No	3826	93
	Yes	289	7.0

4.3. Relationship between Diabetes Mellitus and Various Risk factors

Inorder to find the relationship between diabetic status and the various input variable, a cross tabulation was carried out and summary results presented in the table below.

At 5% level of significance, Sex of respondent, Alcohol consumption, sugar consumption, physical inactivity and Obesity were significant while Smoking and Salt consumption were not significant. This implies that, there is a strong relationship between diabetic status and the significant factors while there is no relationship or association between diabetic status and smoking or salt consumption.

From this analysis, all the significant variables have a relative risk greater than one. The risk of having diabetes is atleast 29% higher for females as compared to males. For those who consume alcohol, the risk of diabetes is at least 33% higher as compared to those who do not consume alcohol. Those who consume excess sugar are 2.2 times likely to have diabetes as compared to those who do not consume excess sugar. It is also evident that, those who are physically inactive have a 73% higher risk of having diabetes as compared to those who are physically active. Those who are Obese are 2.18 times likely to have diabetes as compared to those who are not obese.

Table 4. Cross classification of diabetic risk factors and diabetic status

		Diabetic						
Variables	**Category**	**Yes**	**No**	**Total**	**χ^2 Value**	**P-Value**	**Proportion**	**Relative Risk**
Sex	Female	187	2227	2414	4.680	0.031*	0.0775	1.29
	Male	102	1599	1701			0.0600	
Smoke	Yes	23	341	364	0.303	0.582	0.0632	0.89
	No	266	3485	3751			0.0709	
Alcoholic	Yes	116	1264	1380	6.079	0.014*	0.0840	1.33
	No	173	2562	2735			0.0633	
Excess salt Cons	Yes	26	409	435	0.815	0.367	0.0598	0.836
	No	263	3417	3680			0.0715	
Excess Sugar Cons	Yes	83	554	637	41.644	0.000***	0.1303	2.20
	No	206	3272	3478			0.0592	
Inactive	Yes	123	1109	1232	23.606	0.000***	0.0998	1.73
	No	166	2717	2883			0.0576	
Obese	Yes	59	374	433	32.309	0.000***	0.1363	2.18
	No	230	3452	3682			0.0625	

Inorder to fit the Neural network model, smoking status and salt consumption will not be considered since they do not have any significant relationship with diabetes mellitus.

4.4 Model Selection

The model with the least MSE was selected as per the Table 5 below.

Table 5. Model Selection Using MSE

Nodes	Error	SIC	Threshold	Steps	MSE
1	103.02	294.429	0.00921	40	0.0609
2	97.03	362.783	0.00981	401	0.0580
3	95.85	440.779	0.00939	3488	0.0586
4	93.18	515.760	0.00978	4175	0.0604
5	90.41	590.575	0.00993	5334	0.0661
6	86.85	663.810	0.00994	16612	0.0720
7	83.73	737.908	0.00979	46966	0.0777

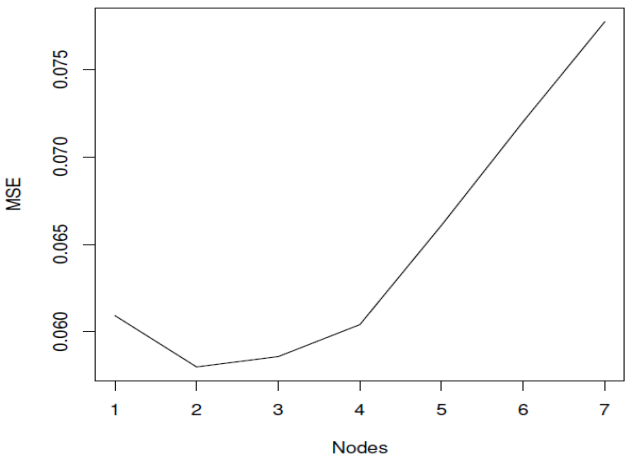

Figure 1. Plot of MSE Against Number of Hidden Nodes

From Figure 1, its very clear that the MSE increases with increase in number of hidden nodes. The MSE is minimum at nodes 2 implying that in order to regularize the network, a model with two hidden nodes should be chosen.

We now train our model with eight input variable and two hidden nodes.

4.5. Network Training

Before training the network, the data set was split into two i.e training set and test set. 75% of data set was for training the network while 25% was for testing and validating. A plot of the network with weights is a shown in Figure 2.

The trained network had twenty one weights. The training process needed 401 steps until all absolute partial derivatives of the error function were smaller than 0.01 (the default threshold). The estimated weights range from -39.2000 to 3.6194. For instance, the intercepts of the first hidden layer are -1.3032 and 3.0729 and the four weights leading to the first hidden neuron are estimated as 3.6194, -27.3465, -35.4901, 0.2369, -0.7540, -8.7735 and 0.9512 for the covariates age, sex, alcohol, sugar, inactive, sbp, dbp and obese, respectively.

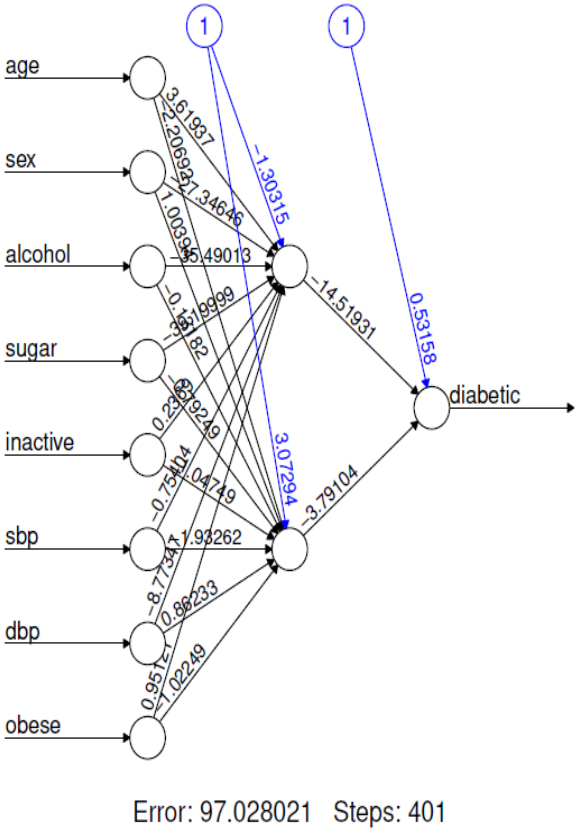

Figure 2. Neural Network Plot

A summary table for weights is as in Table 6.

4.6. Trained Network Assessment

To assess the fitness of the model, a cross classification of the actual data and the predicted outcome using test data set was reported. Table 7 below shows the results of the confusion matrix.

Table 6. Trained Neural Network Results

Description	Weight
error	97.0280
reached.threshold	0.0098
steps	401.0000
aic	236.0560
bic	362.7833
Intercept.to.1layhid1 (bias1)	-1.3032
age.to.1layhid1 (w11)	3.6194
sex.to.1layhid1 (w21)	-27.3465
alcohol.to.1layhid1 (w31)	-35.4901
sugar.to.1layhid1 (w41)	-39.2000
inactive.to.1layhid1 (w51)	0.2369
sbp.to.1layhid1 (w61)	-0.7540
dbp.to.1layhid1 (w71)	-8.7735
obese.to.1layhid1 (w81)	0.9512
Intercept.to.1layhid2 (bias2)	3.0729
age.to.1layhid2 (w12)	-2.2069
sex.to.1layhid2 (w22)	1.0039
alcohol.to.1layhid2 (w32)	-0.1318
sugar.to.1layhid2 (w42)	-0.7925
inactive.to.1layhid2 (w52)	-1.0475
sbp.to.1layhid2 (w62)	-1.9326
dbp.to.1layhid2 (w72)	0.8623
obese.to.1layhid2 (w82)	-1.0225
Intercept.to.diabetic (alhpa0)	0.5316
layhid.1.to.diabetic (alpha1)	-14.5193
layhid.2.to.diabetic (alpha2)	-3.7910

Table 7. Confusion Matrix

		Predicted		
		Diabetic	Non Diabetic	Total
Actual	Diabetic	9	58	67
	Non Diabetic	3	959	962
	Total	12	1017	1029

The sensitivity of the trained network was reported as 75% while specificity was 94.29%. This implied that the overall accuracy rate was 84.64%. This implied that the model could correctly classify an individual as either diabetic or not with an accuracy rate of 84.64%. These results are consistent with other neural network models for binary classification.

After the model was trained, a 10 fold cross validation was carried out inorder to test the generalization of the model. The MSE for each fold was reported as in Table 8. The average of these results gives the test accuracy of the algorithm. From this study, it is clear that the validated average MSE was 0.0686 or the error rate was 6.86%.

Table 8. Cross Validation MSE

Fold	CV Error
1	0.0685
2	0.0525
3	0.0783
4	0.0499
5	0.0780
6	0.0635
7	0.0681
8	0.0850
9	0.0813
10	0.0608

4.7. Asymptotic Normality of ANN Parameter Estimates

4.7.1. Kolmogorov-Smirnov Test of Normality

In order to test our hypothesis, Kolmogorov-Smirnov test of normality was used. The hypothesis stated that;

H_0 : The Artificial Neural Network parameter estimates are asymptotically normal and consistent.

Table 9 gives the results of the test for the various parameters.

At 5% significance level, we do not reject the null hypothesis for all the parameters except w71. We therefore conclude that most parameter estimates did not have a significant departure from normality. Its only the estimator, w71 that have a significant departure from normality at 5% since it has a P-Value of 0.0325.

4.7.2. Normal Q-Q Plot

The Normal Q-Q plot, or Normal quantile-quantile plot, is a graphical tool used to assess if a set of data plausibly came from some Normal theoretical distribution. It allows one to see at-a-glance if the assumption of normality is plausible and if not, how the assumption is violated and what data points contribute to the violation. If both sets of quantiles came from the same distribution, the points should form a line that is roughly straight. A qq- plot to study the behavior of the ANN parameter estimates with a simulation of large sample shows that the parameter estimates aligned themselves in a straight line. Clearly showing that the ANN parameter estimates had a normal distribution and thus no violation of normality assumption. This is demonstrated in Figure 3 and Figure 4.

Table 9. Kolmogorov-Smirnov Test

Estimator	K-S Statistics	P-Value
alpha0	0.0083	0.8811
alpha1	0.0154	0.1865
alpha2	0.0130	0.3667
bias1	0.0130	0.3667
bias2	0.0104	0.6518
w11	0.0116	0.5116
w12	0.0061	0.9923
w21	0.0093	0.7801
w22	0.0159	0.1595
w31	0.0130	0.3667
w32	0.0192	0.5012
w41	0.0177	0.0872
w42	0.0121	0.4569
w51	0.0092	0.7912
w52	0.0106	0.6280
w61	0.0171	0.1074
w62	0.0157	0.1699
w71	0.0203	0.0325*
w72	0.0147	0.2301
w81	0.0128	0.3857
w82	0.0148	0.2234

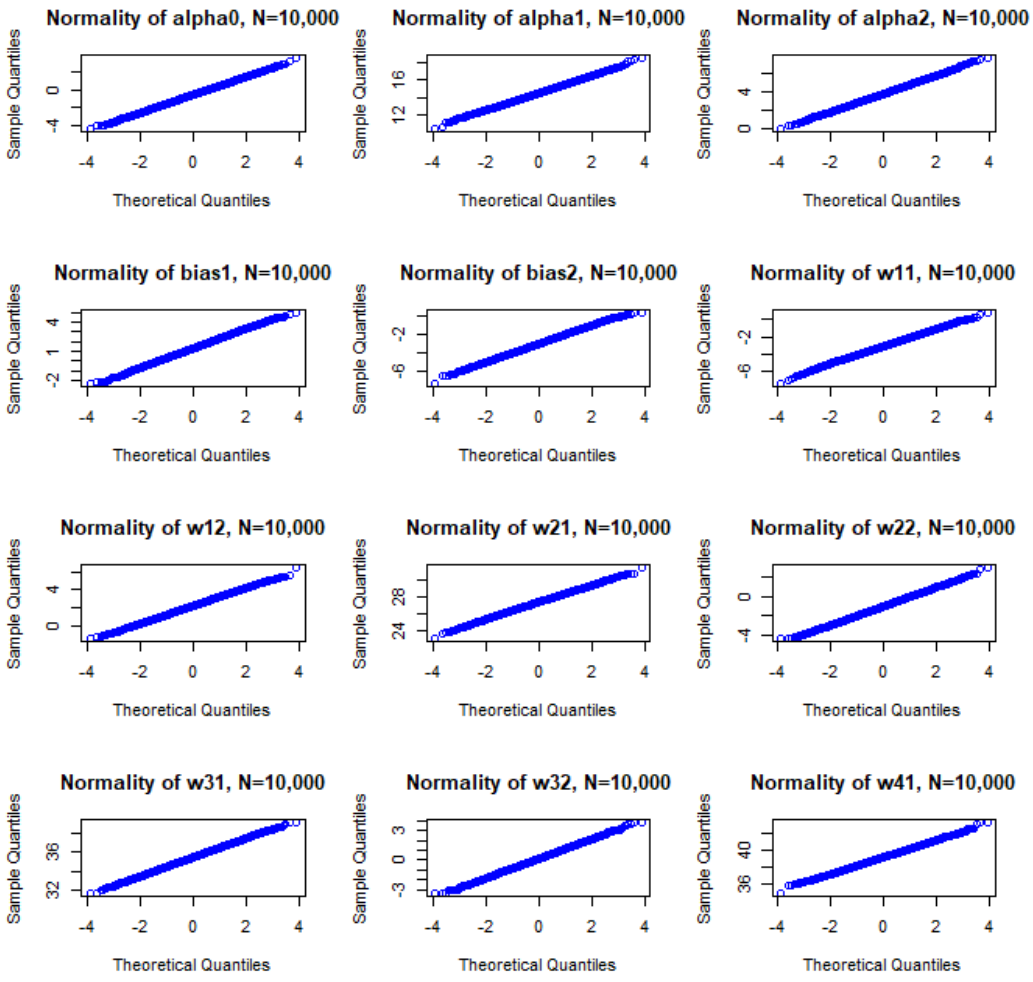

Figure 3. QQ plot for Neural Network weights

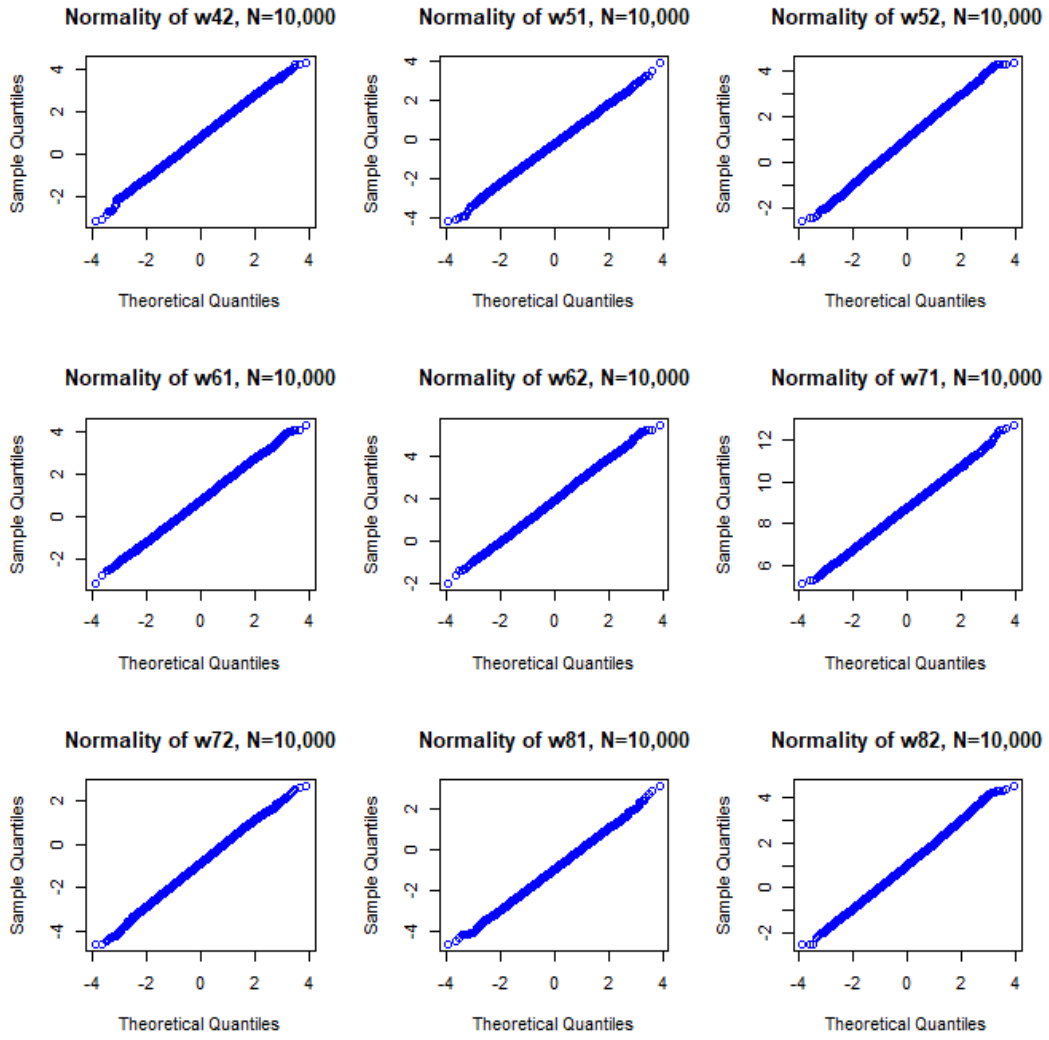

Figure 4. Cont' QQ plot for Neural Network weights

5. Conclusions

Advancement in modern computing has lead to the use of artificial neural networks which mimics the human brain. Combined with the statistical analysis, artificial neural networks are used to identify complex patterns among data. This study aimed at modeling diabetes mellitus using ANN. This combined with clinical diagnoses can greatly assist the clinicians and doctors in correctly diagnosing the underlying disease. The accuracy obtained from the trained model is a good indication that, with good investment and further research in this field, the classification accuracy can be improved and hence, the model can be used in future.

In this study the Diagnosis of Diabetes Mellitus has been modeled using neural network classifier. In order to come up with a network architecture, chi square test of statistics was first carried out inorder to establish the input variables that had a significant relationship with diabetes mellitus. For variables that were continous, a stepwise model building was carried out and the networks MSE did not increase indicating that they were also significant. Inorder to determine the appropriate number of hidden nodes (size of the network), MSE and SIC were used to determine the parsimonious model. The model with more than seven nodes did not converge. Within the models that converged, the model with two hidden node had the minimumm MSE and thus was chosen as the model that could fit the data well.

The ANN network had 9 inputs neurons, one hidden layer with two neurons and the output layer had one neurons. The hidden and output layers used the sigmoid transfer function and were trained using the back propagation algorithm. The data was split into training set, test set and validation set. A 10 fold cross validation was carried out inorder to test the classification accuracy of the model. The sensitivity of the trained network was reported as 75% while specificity was 94.29%. The overall accuracy of the model was 84.64%. As conclusions, It was seen that with a good choice of risk factors for diabetes, neural network structures could be successfully used to help diagnose diabetes disease among Kenyan adult population.

6. Recommendations

This study sets a precedent in modeling diabetes mellitus using artificial neural network among adult Kenyan population. With increasing interest in artificial intelligence, I would recommend that future research be focused in embracing this new field of statistical computing. More important would be to integrate clinical/medical diagnosis with artificial intelligence. More complex machine learning algorithms like support

vector machine, self organizing maps should be applied in diagnosing diabetes.

Acknowledgements

I thank the Almighty God for enabling me come this far. Without His grace and favour, i would not have made it. Special regards to my Supervisors Prof. Anthony G. Waititu, Dr. Anthony Wanjoya and Dr. Thomas Mageto for their guidance throughout this work. Special Thanks to my entire family for their moral support which kept me going. I sincerely thank my wife Beth and my daughter Salome for their unending support. My gratitudes goes to my employer, Kenya National Bureau of Statistics (KNBS) for fully sponsoring me to undertake my Masters studies. Finally, I thank Alexander Kasyoki Muoka for encouraging me to finish this project and all who supported me in one way or the other, God bless you richly.

References

[1] World Health Organization. *Global Report on Diabetes*, WHO Press, Geneva, 2006, page 6.

[2] Zainab A, et al. Using Neural Network to predict the Hypertension. *International Journal of Scientific Development and Research*, 2(2): 35-38, 2017.

[3] Ripley, B. D. *Pattern Recognition and Neural Networks,* Oxford Press, London, 1996.

[4] Robert, S. "Artificial Intelligence: its use in Medical Diagnosis", *The Journal of Nuclear Medicine*, 34 (3): 510-514, 1993.

[5] Flury, B., & Riedwyl, H.*Multivariate statistics: A practical approach*. London: Chapman and Hall,1999.

[6] Press, S. J.,&Wilson, S. "Choosing between logistic regression and discriminant analysis", *Journal of the American Statistical Association*, 73(364): 699-705, 1978.

[7] Hosmer, D. W., & Lemeshow, S. *Applied logistic regression.* New York: Wiley Series, 1989.

[8] Buntine, W. L., & Weigend, A. S. "Bayesian Back-propagation", *Complex Systems*, 5(6):603-643, 1991.

[9] Menard, S. Applied logistic regression analysis, series: *Quantitative applications in the social sciences.* Thousand Oaks, CA: Sage, 1993.

[10] Myers, R. H. *Classical and modern regression with applications (2nd edition).* PWS-KENT Publishing Company, Boston, Massachusetts, 1990.

[11] Neter, J., Li, W., Nachtsheim, C.J., & Kutner, M. H. *Applied linear statistical models (5th edition)*, McGraw-Hill/Irwin, New York, 2005.

[12] Snedecor, G. W., & Cochran, W. G. *Statistical methods (7th edition)*. Ames, IA: The Iowa State University Press, 1980.

[13] Razi M.A.,& Athappilly, K. . A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models. *Expert Systems with Applications* , 29(1): 69-74, 2005.

[14] Zhang, G., Patuwo B.E., & Hu, M.Y. "Forecasting with artificial neural networks: The state of the art", *International Journal of Forecasting*, 14(1): 35-62, 1998.

[15] Cybenko, G. "Approximation by superpositions of a sigmoidal function". *Mathematics of Controls Signals and Systems*, 2(4): 303-314, 1989.

[16] Funahashi, K. "On the approximate realization of continuous mappings by neural networks". *Neural Networks*, 2(3): 183-192, 1989.

[17] Hornik, K., Stichcombe, M., & White H. "Multilayer feedforward networks are universal approximators". *Neural Networks*, (2): 359-366, 1989.

[18] Hornik, K. "Approximation capabilities of multilayer feed-forward networks". *Neural Networks*, 4(2): 251-257, 1991.

[19] Irie, B., & Miyake, S. "Capabilities of three-layered perceptrons*,"* In*: Proceedings of the IEEE Second International Conference on Neural Networks*, July 1988, San Diego, California USA.

[20] Dybowski, R., & Gant,V. *Clinical Applications of Artificial Neural Networks*, Cambridge University Press, London, 2007.

[21] Szolovits, P., Patil,S., & Schwartz, W. "Artificial Intelligence in Medical Diagnosis.", *Annals of Internal Medicine*, 108(1): 80-87, 1988.

[22] Bradley, B. "Finding Biomarkers is Getting Easier", *Ecotoxicology*, 21(3): 631-636, 2012.

[23] Baxt, W.G. "Use of an artificial neural network for the diagnosis of myocardial infarction". *Annals of Internal Medicine*, 115(11): 843-848, 1991.

[24] Jayalakshmi, T., & Santhakumaran A. "A novel classification method for classification of diabetes mellitus using artificial neural networks", In: *International Conference on Data Storage and Data Engineering*, February, 2010, Bangalore, India.

[25] Swanson, N. R., & White, H. "A model-selection approach to assessing the information in the term structure using linear models and artificial neural networks", *Journal of Business Economic Statistics*, 13(3): 265-275, 1995.

[26] Olaniyi E. O, & Adnan K. "Onset Diabetes Diagnosis Using Artificial Neural Network", *International Journal of Scientific & Engineering Research*, 5(10): 754-759, 2014.

[27] Adeyemo, A & Akinwonmi, A. "On the Diagnosis of Diabetes Mellitus Using Artificial Neural Network Models". *African Journal of Computing & ICT*, 4(1):1-8, 2011.

[28] Rajib D, V. Bajpai, G. Gandhi & B. Dey. "Application of artificial neural network technique for diagnosing diabetes mellitus", In:*2008 IEEE Region 10 Colloquium and the Third ICIIS*, 8-10 December, 2008, Kharagpur, INDIA.

[29] Chan K, Ling S, Dillon T, & Nguyen H. "Diagnosis of hypoglycemic episodes using a neural network based rule discovery system", *Expert System Applications*. 38(8): 9799-9808, 2011.

[30] Kenya National Bureau of Statistics. *Kenya STEPwise Survey For Non Communicable Diseases Risk Factors* KNBS, MOH and WHO, Nairobi, page 137-140, 2015.

[31] Agresti, A. *An Introduction to Categorical Data Analysis (2nd edition)*. New York: Wiley Series, 2007.

[32] Amato F, et al. "Artificial neural networks in medical diagnosis". *Journal of Applied Biomedicine*, 11(2): 47-58, 2013.

[33] Intrator, O., & Intrator, N. "Interpreting Neural Networks Results: A simulation study", *Computational Statistics and Data Analysis,* 37(3): 373-393, 2001.

[34] Waititu, A.G. *Nonparametric Change point Analysis for Bernoulli Random Variables Based on Neural Networks*, Phd Thesis, Kaiserslautern University, Germany.(https://kluedo.ub.uni-kl.de), 2008.

[35] White, H. "Some Asymptotic Results for Learning in Single Hidden-Layer Feedforward Network Models", *Journal of the American Statistical Association*, 84(408): 1003-1013, 1989.

[36] Warner, B., & Misra, M. "Understanding neural networks as statistical tools", *The American Statistician*, 50(4), 284-293, 1996.

[37] Sussmann, H. J. "Uniqueness of the weights for minimal feed-forward nets with a given input-output map",*Neural Networks*, 5(4): 589-593, 1992.

[38] Hwang, J. T., & Ding, A. A. "Prediction intervals for artificial neural networks". *Journal of American Statistical Association*, 92(438): 748-757, 1997.

[39] Berger, V., & Zhou, Y. "Kolmogorov Smirnov test: Overview". In *Wiley statsref: Statistics reference online*. New York: John Wiley & Sons, Ltd.

[40] Rissanen, J. "Stochastic complexity and modeling", *Annals of Statistics*, 14(3): 1080-1100, 1986.

[41] Haykin, S. *Neural Networks and Learning Machines (3rd edition)*. New Jersey: Pearson Education, 2009.

[42] Muhammad, A. R., & Kuriakose, A. "A comparative predictive analysis of neural networks (NNs), nonlinear regression and classification and regression tree (CART) models", *Expert Systems with Applications*, 29(1): 65-74, 2005.

[43] Temurtas, H., Yumusak, N., & Temurtas F. "A comparative study on diabetes disease diagnosis using neural networks", *Expert System Applications*, 36(4): 8610-8615, 2009.

[44] White, H. *Artificial neural networks: Approximation and learning theory*. Oxford: Basil Blackwell, 1992.