Available online www.ejaet.com

European Journal of Advances in Engineering and Technology, 2021, 8(10):28-35

Research Article ISSN: 2394 - 658X

Improving Credit Scoring Classification Performance using Self Organizing Map-Based Machine Learning Techniques

Shamsuddeen Suleiman[1], Anas Ibrahim[2], Dauda Usman[3], Bala Yabo Isah[4] and Hairullahi Muhammad Usman[5]

[1]Department of Mathematical Sciences, Federal University Dutsin–Ma, Katsina State, Nigeria
[2]Department of Mathematics, Adeyemi College of Education, Ondo, Ondo State, Nigeria
[3]Department of Mathematics and Statistics, Umaru Musa YarAdua University Katsina, Katsina State, Nigeria
[4]Department of Mathematics UsmanuDanfodiyo University, Sokoto, Sokoto State, Nigeria
[5]Department of Science, Mathematics Unit, School of Basic and Remedial Studies, Sokoto, Sokoto State, Nigeria
Corresponding author email: anaseencollection@gmail.com

ABSTRACT

This research usesself-organizing maps (SOM) in order improve the ability of the pattern recognition techniques including neural networks and K-nearest neighbour used to forecast the credit risk of borrowers from Bank of Agriculture (BOA) Sokoto. In this work, a hybrid approach to building the credit scoring model was proposed using the unsupervised learning based on self-organizing map (SOM) to specifically improve the discriminant capabilities of K-nearest neighbour and Neural networks. Within the two-stage scheme, the knowledge (i.e., prototypes of clusters) found by SOM is considered as input to the subsequent pattern recognition models. The results from BOA, Sokoto data indicate that the two-stage models improved the performances of Neural Network and K-nearest neighbour from 96.3% and 95.7% to 97.3% and 96.3% respectively.

Key words: Credit Scoring, Self-Organizing Map, Pattern Recognition, K-nearest neighbour, Neural Network

---

1. INTRODUCTION

Credit evaluation is one of the most crucial processes in bankscredit management decisions. This process includes collecting, analyzing and classifying different credit elements and variables to assess the credit decisions. One of the most important kits, to classify a bank's customers, as a part of the credit evaluation process to reduce the current and the expected risk of a customer being bad credit, is credit scoring [1]. The World Bank Group also uses credit scoring approaches to rate sovereign borrowers, counterparties in financial derivative transactions, investment targets, and private sector borrowers [2]. The advantages of using credit scoring models include reducing the cost of credit analysis, enabling faster credit decisions and diminishing possible risk [3]. A Self-organizing map (SOM) invented by TeuvoKohonen [4] is a type of ANN where the neurons are set along a grid. SOM maps higher dimensional input onto the lower dimensional grid while preserving topological ordering present in the input space. The principal objective of SOM is to transform a complex high-dimensional input space into a simpler low-dimensional (typically two-dimensional) discrete output space by preserving the relationships (i.e. topology) in the data, but not the actual distances. The spatial locations (i.e., coordinates) of the nodes in the output space are indicative of inherent statistical features contained in the input space [5]. K-Nearest Neighbour (KNN) is a standard nonparametric technique used for probability density function estimation and classification [6]. Priyanka D., Mishika S., Gopal P., Amit H. [7] analyzed a detailed comparison between Random Forest and K Nearest Neighbours algorithm, and discussed the speed and accuracy of the two Machine Learning algorithms mentioned after testing them on the UCI Credit Card database. After comparison and finding the gender with maximum debt, both the methods are refined and tuned to obtain better precision.Neural Networks (NN) is a mathematical representation inspired by the human brain and its ability to adapt on the basis of the inflow of new information. Mathematically, NN is a non-linear optimization tool. The NN design called multilayer perceptron (MLP) is especially suitable for classification and is widely used in practice. Suleiman and Badamasi [8] uses Variance inflation

factor (VIF) to detect multicollinearity among some risk factors and Principal component technique (PCA) was employed to remove it, they usedLevenberg-Marquardt (LM) algorithm to train the statistical neural network for the original and principal components data. Their results show that when five (5) hidden neurons architecture is used, the model achieved 99.0% and 93.9% accuracy for training the original and reduced data respectively. Gulumbe, Suleiman, Badamasi, Tambuwal and Usman [9] Considered 100 patients from Ahmadu Bello University Teaching Hospital who have undergone diabetes screening test and 29 risk factors were used. Back propagation algorithm was used to train the artificial neural network for the original and simulated data sets. The results show that the models achieved 98.7%, 57.0%, 73.3%, and 63.0% accuracy for training the original, simulated at 100, simulated at 150 and simulated at 200 data sets respectively. The results also shows that the areas covered under receiver operating curves are 0.997, 0.587, 0.849 and 0.706 for training the original, simulated at 100, simulated at 150 and simulated at 200 datasets respectively. Ali, Ning & Bernardete [10] presented a hybrid approach to building credit scoring model, it illustrates how the unsupervised learning based on self-organizing map (SOM) can improve the discriminant capability of feed forward neural network (FNN). Suleiman and Badamasi [8] uses Variance inflation factor (VIF) to detect multicollinearity among some risk factors and Principal component technique (PCA) was employed to remove it, they usedLevenberg-Marquardt (LM) algorithm to train the statistical neural network for the original and principal components data. Their results show that when five (5) hidden neurons architecture is used, the model achieved 99.0% and 93.9% accuracy for training the original and reduced data respectively.Credit loans and finances have risk of being defaulted. These loans involve large amounts of capital and their non-retrieval can lead to major loss for the financial institution. Therefore, the accurate assessment of the risk involved is a crucial matter for banks and other such organizations. It is not only important to minimize the risks in granting credit, but also, on reducing errors in declining valid clients. This is to save the banks from lawsuits [11]. Suleiman, Gulumbe and Shehu [12] managed loan default predictors in agricultural credit scoring and obtained 96.3% predictive performance for his proposed Neural Network using Genetic Algorithm and compared it with other conventional predictive models, including Discriminant Analysis with 96.0%, K-NN classifier with 95.8%, Logistic Regression with 96.0% and CART with 96.7% classification performance accuracy. Since the focal point of credit risk management is to classify creditworthy applicant accurately, present work suggests SOM+KNN and SOM+NN two-stage models which may result in the enhancement of classification performance accuracy.The aim of this research is to improve the classification ability of KNN and NN models using Self-Organizing Map. This research work improves the work of Suleiman, Gulumbe and Shehu, [12]. The scope is to enhance the performance of KNN and NN model as proposed by Suleiman, Gulumbe and Shehu, [12] by adding Self Organizing Map (SOM) in credit risk management.

2. METHODOLOGY

A. SELF ORGANIZING MAP (SOM)

A SOM invented by TeuvoKohonen [4] is able to reduce the amount of data and simultaneously project the data nonlinearly onto a lower dimensional array (see Figure 1). In each iteration of the training process, the reference vectors are updated in such a way that the best-matching neuron and its neighbours on the grid are dragged toward the input. As a result, the neurons are topologically ordered on the grid, where instances that have similar features in the input space will be projected to the neurons located close to each other in the grid space Ali, Ning and Bernardete [10].

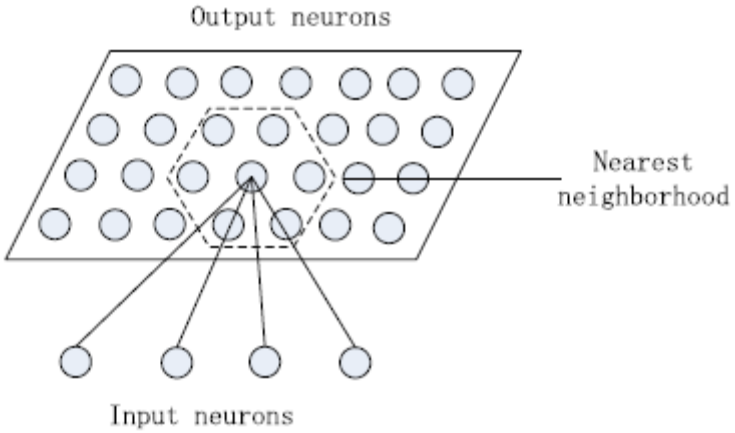

Fig. 1 Example of self-organizing map composed of 4 input neurons and an output grid [10]

SOM Algorithm (Training Process) steps are as follows:

1. Initialize Neural Network weights
2. Randomly select an input
3. Select the winning neuron using Euclidean Distance:

$$d_j = \sqrt{\sum_{i=1}^{n} \left(x_i - w_{k,i} \right)^2} \qquad 1$$

Where d_j represent the Euclidean distance for difference between each input and each neuron, *x* represent the input, *w* represent neuron weight, k = 1, …, m number of neurons, and *i* = 1, …, n number of inputs.

4. Update neuron weight, using weight update formula (use the weight of the winning neuron to update the weight of the same winning neuron and neurons around it, using weight update formula):
$\Delta w_{j,i} = \eta(t) * T_{j,I(x)}(t) * (x_i - w_{j,i})$ 2
Where $\eta(t) = \eta_0 exp\left(-\frac{t}{\tau_\eta}\right)$; is the **learning rate** which determines how quickly is the weight update, $T_{j,I(x)}(t) = exp\left(-\frac{S^2_{j,i(X)}}{2\sigma(t)^2}\right)$is the **Topological Neighborhood**which defines the extent to which the neighboring neurons and the winning neuron update their weights, for $S_{j,I} = \| w_j - w_i \|$ is the **Lateral Distance Between Neurons**, which is essentially the same as Euclidean distance , but applied to two different neurons rather than a neuron and an input and $\sigma(t) = \sigma_0 \exp\left(-\frac{t}{\tau_0}\right)$ is the **Neighborhood size**, while $(x_i - w_{j,i})$is the difference between the input and the weight.
5. Go back to 2 until done training.

B. K-Nearest Neighbor (KNN) Classifier

The K-nearest neighbors (KNN) algorithm is a type of supervised machine learning algorithm that can be used for classification and regression classification problems. KNN is a lazy learning algorithm since it does not have a specialized training phase and uses all the data for training while classification. It is also a non-parametric learning algorithm because it does not assume anything about the underlying data. The KNN algorithm uses feature similarity to classify the values of the latest data points which can be assigned a worth supported by how closely it matches the points within the training set. The output of k-NN depends on its use of classification: Class membership is obtained as the result of *k*-NN classification. An object is known by a plurality vote of its neighbours, with the object being assigned to the class most common among its k nearest neighbours "where K is a positive integer, typically small" [7]. K value is chosen by taking square root of the number of observations (i.e. $\sqrt{n}$) and the value should be approximated to the most nearest odd number from the obtained value of root *n*. When performing the *k*-NN methodology, a very important step is the choice of the metric used. Henley and Hand [6] describe the choice of the metric and the choice of the number of nearest neighbours to consider [12]. A commonly used metric is the standard Euclidean distance given by:

$$d = \sqrt{\sum_{i=1}^{n}\left(x_i - y_j\right)^2} \quad 3$$

Where *x* and *y* are data points, for $i = 1,...,n$ number of inputs and $j = 1,...,m$ number of compared inputs.

C. Neural Networks

A neural network (NNW) is a mathematical representation inspired by the human brain and its ability to adapt on the basis of the inflow of new information. Mathematically, NNW is a non-linear optimization tool. Many various types of NNW have been specified in the literature. The NNW design called multilayer perceptron (MLP) is especially suitable for classification and is widely used in practice (Suleiman, Gulumbe and Shehu, 2012). The network consists of one input layer, one or more hidden layers and one output layer, each consisting of several neurons. Each neuron processes its input and generates one output value that is transmitted to the neurons in the subsequent layer. Each neuron in the input layer (indexed *i*=1, …, *n*) delivers the value of one predictor (or the characteristics) from vector *x*. When considering default/non-default discrimination, one output neuron is satisfactory. In each layer, the signal propagation is accomplished as follows. First, a weighed sum of inputs is calculated at each neuron: the output value of each neuron in the proceeding network layer times the respective weight of the connection with that neuron. A transfer function *g(x)* is then applied to the to this weighted sum to determine the neurons output value. So, each neuron in the hidden layer (indexed *j*=1, …, q) produces the so-called activation [12]:

$$\alpha_j = g\left(\sum_{i=1}^{n} w_{ijx_i}\right) \quad 4$$

The neurons in the output layer (indexed *k* =1, …, *m*) behave in a manner similar to the neurons of the hidden layer to produce the output of the network:

$$y_k = \left(\sum_{j=1}^{n} w_{jk}^1 a_j\right) = f\left[\sum_{j=1}^{n} w_{jk}^1 g\left(\sum_{j=1}^{n} w_{jk}^1 x_i\right)\right] \quad 5$$

Where w_{jk}^1 and w_{ij}^1 are weights.

D. Classifiers' Evaluation Measures

In order to evaluate a binary decision task, we defined the following three performance metrics:

$Accuracy = \frac{tp + tn}{tp + fn + tn + fp}$ denotes the proportion of correct predictions out of the total samples.

$Sensitivity = \frac{tp}{tp + fn}$ denotes the fraction of true positives that are actually positive.

$Specificity = \frac{tn}{tn + fp}$ denotes the fraction of true negatives that are actually negative.

Where tp refers to true positive, tn true negative, fp is the false positive and fn refers to false negative.

3. RESULTS AND DISCUSSIONS

A real world credit dataset is used in this research. The dataset is extracted from the application forms of **Agricultural and Rural Development Bank Sokoto**. The dataset is referred to as "Credit Dataset". After preparing or cleaning the dataset, it is used in the subsequent sections for conducting the analysis with Self-Organizing Map (SOM), k-Nearest Neighbor (KNN) and Neural Network (NN).The Dataset contains 300 cases, 164 applicants were considered as "Creditworthy" and the rest 136 were treated as "Non-Creditworthy". The dataset was described as follows:

Table -1 Dataset Description

S/No.	Variable	Type	Scale	Description
1.	Attribute1	Input Variable	Scale	Age of the Applicant
2.	Attribute2	Input Variable	Nominal	Sex of the Applicant
3.	Attribute3	Input Variable	Nominal	Marital Status of the Applicant
4.	Attribute4	Input Variable	Ordinal	Job of the Applicant
5.	Attribute5	Input Variable	Nominal	Purpose of the loan
6.	Attribute6	Input Variable	Scale	Credit Amount of the loan
7.	Attribute7	Input Variable	Scale	Estimated Annual Salary of the Applicant
8.	Attribute8	Input Variable	Nominal	Repayment Plan of the loan
9.	Attribute9	Input Variable	Nominal	Application type
10.	Attribute10	Input Variable	Nominal	Application Period
11.	Attribute11	Output Variable	Nominal	Status of the Credit Applicant

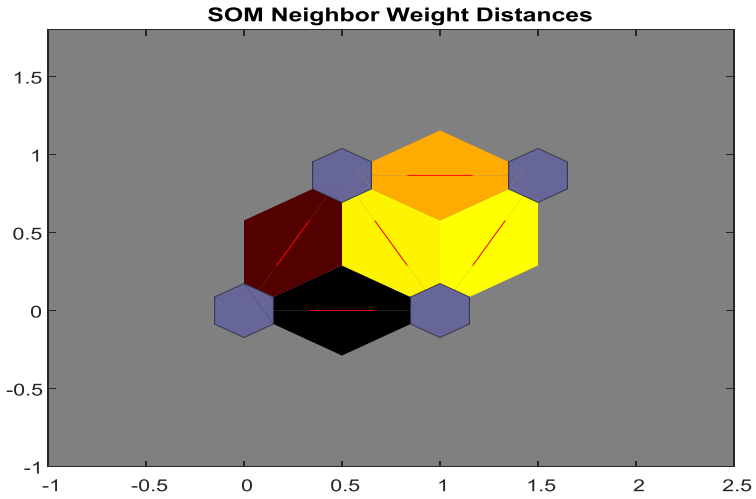

Fig. 2 SOM Neighbor Weight Distances

Figure 2 indicates the distances between neighboring neurons, it uses the following color coding:

- The blue hexagons represent the neurons.
- The red lines connect neighboring neurons.
- The colors in the regions containing the red lines indicate the distances between neurons.
- The darker colors represent larger distances.
- The lighter colors represent smaller distances.

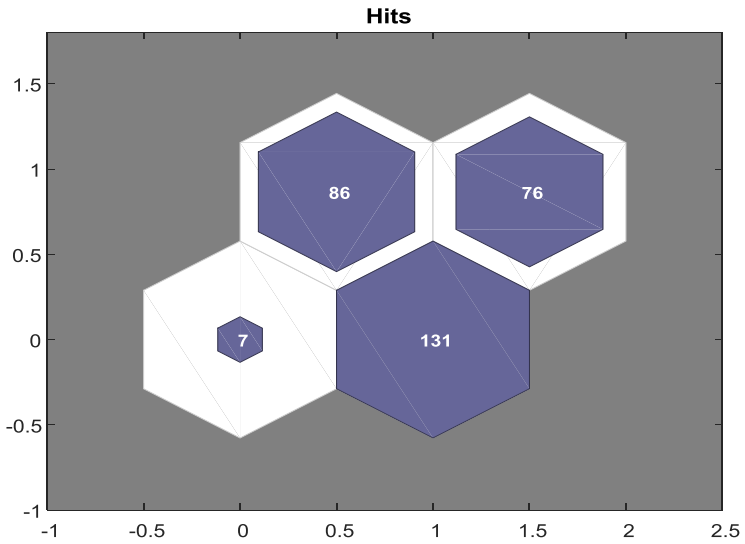

Fig. 3 SOM Sample Hits

Figure 3 shows data points associated with each neuron. It is best if the data are fairly evenly distributed across the neurons. In this example, the data are concentrated a little more in the lower-right neurons, but overall the distribution is fairly even.

Table -2 Credit Scoring Classifiers Performance Evaluation

Models	Accuracy (%)	Sensitivity (%)	Specificity (%)
KNN	95.7	97.8	93.9
SOM+KNN	96.3	97.8	95.1
NN	96.3	97.8	95.1
SOM+NN	97.3	97.8	97.1

Table 2 indicates that KNN model has 95.7% accuracy, that is, the model gives 0.957 chance of correctly classifying the applicants in to their respective groups. Similarly, KNN model has 97.8% Sensitivity, that is, the model gives 0.978 probability of correctly classifying the applicant as defaulter. In other words, sensitivity corresponds to the proportion of non-creditworthy that are correctly classified by the classification model. Finally, KNN has 93.9% specificity, that is, the model gives 0.939 probability of wrongly classifying applicant as a defaulter in other words, specificity represents the proportion of good payers that are correctly classified by the classification model.

SOM+KNN model has 96.3% accuracy, that is, the model gives 0.963 chance of correctly classifying the applicants in to their respective groups. Similarly, SOM+KNN model has 97.8% Sensitivity, that is, the model gives 0.978 probability of correctly classifying the applicant as defaulter. In other words, sensitivity corresponds to the proportion of non-creditworthy that are correctly classified by the classification model. Finally, SOM+KNN has 95.1% specificity, that is, the model gives 0.951 probability of wrongly classifying applicant as a defaulter in other words, specificity represents the proportion of good payers that are correctly classified by the classification model.

NN model has 96.3% accuracy, that is, the model gives 0.963 chance of correctly classifying the applicants in to their respective groups. Similarly, NN model has 97.8% Sensitivity, that is, the model gives 0.978 probability of correctly classifying the applicant as defaulter. In other words, sensitivity corresponds to the proportion of non-creditworthy that are correctly classified by the classification model. Finally, NN has 95.1% specificity, that is, the model gives 0.951 probability of wrongly classifying applicant as a defaulter in other words, specificity represents the proportion of good payers that are correctly classified by the classification model.

SOM+NN model has 97.3% accuracy, that is, the model gives 0.973 chance of correctly classifying the applicants in to their respective groups. Similarly, SOM+NN model has 97.8% Sensitivity, that is, the model gives 0.978 probability of correctly classifying the applicant as defaulter. In other words, sensitivity corresponds to the proportion of non-creditworthy that are correctly classified by the classification model. Finally, SOM+NN has 97.1% specificity, that is, the model gives 0.971 probability of wrongly classifying applicant as a defaulter in other words, specificity represents the proportion of good payers that are correctly classified by the classification model.

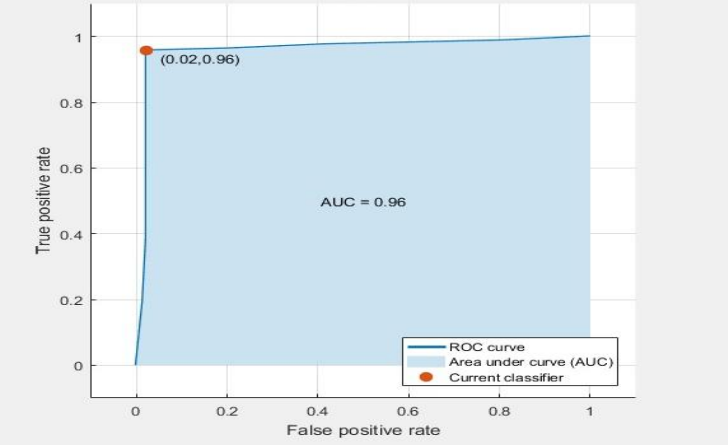

Fig. 4 KNN ROC Curve

Figure 4 shows the trade-off between sensitivity (or TPR) and specificity (1 – FPR) of KNN classifier. Since area covered under the curve is 0.96 and this is closer to the top-left corner of the curve indicating a better performance by the model. As a baseline, a random classifier is expected to give points lying along the diagonal (FPR = TPR). The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

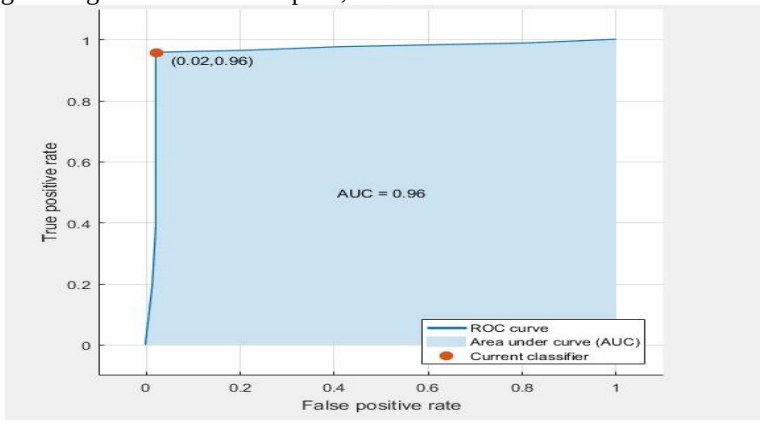

Fig. 5 SOM+KNN ROC Curve

Figure 5 shows the trade-off between sensitivity (or TPR) and specificity (1 – FPR) of SOM+KNN classifier. Since area covered under the curve is 0.96 and this is closer to the top-left corner of the curve indicating a better performance by the model. As a baseline, a random classifier is expected to give points lying along the diagonal (FPR = TPR). The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

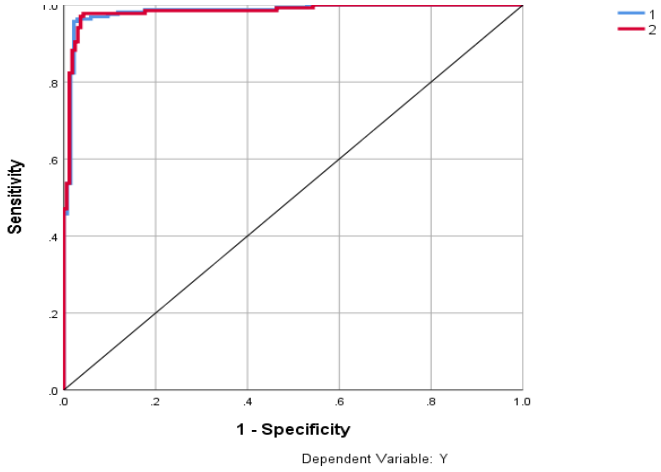

Fig. 6 NN ROC Curve

Table -3 NN Area Under the Curve

		Area
Y	1	0.983
	2	0.983

Figure 6 shows the trade-off between sensitivity (or TPR) and specificity (1 – FPR) of NN classifier. Since area covered under the curve is 0.98 in Table 3 and this is closer to the top-left corner of the curve indicating a better performance by the model. As a baseline, a random classifier is expected to give points lying along the diagonal (FPR = TPR). The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

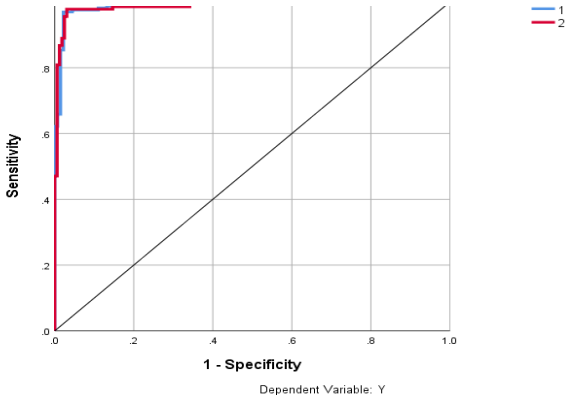

Fig. 7 SOM+NN ROC Curve

Table -4 SOM+NN Area under the Curve

		Area
Y	1	0.988
	2	0.988

Figure 7 shows the trade-off between sensitivity (or TPR) and specificity (1 – FPR) of SOM+NN classifier. Since area covered under the curve is 0.99 in Table 4 and this is closer to the top-left corner of the curve indicating a better performance by the model. As a baseline, a random classifier is expected to give points lying along the diagonal (FPR = TPR). The closer the curve comes to the 45-degree diagonal of the ROC space, the less accurate the test.

4. CONCLUSION

Since credit loans and finances have risk of being defaulted, it is crucial for financial institutions to develop reliable credit scoring systems in order to classify creditworthiness of the borrowers. A variety of pattern recognition techniques including neural networks and K-nearest neighbour have been applied to predict whether the borrowers should be considered a good or bad credit risk. In this present work, a two-stage approach to building the credit scoring model using unsupervised learning based on self-organizing map (SOM) was proposed to improve the discriminant capability of neural networks and K-nearest neighbour that were used by Suleiman *et al* [12] to classify agricultural credit defaulters for Bank of Agriculture (BOA), Sokoto. Thus, the inputs of the two-stage models were the observations clustered in to four (4) groups by the SOM algorithm. The results indicate that the discrimination accuracy of the KNN model can be improved from 95.7% to 96.3%. Similarly, the discrimination accuracy of the NN model can be improved from 96.3% to 97.3%. The results therefore indicate that the integration of SOM algorithm in to these techniques made neural network to outperform KNN.

REFERENCE

[1]. Abdou, H. & Pointon, J. (2011) "Credit scoring, statistical techniques and evaluation criteria: a review of the literature". *Intelligent Systems in Accounting, Finance & Management*, 18 (2-3), pp. 59-88.

[2]. Fritz, V. (2016). From form to function to sustainable solutions? Reforming public sectors in low income countries: New approaches and available evidence of "what works". *Public Administration and Development*, *36*(5), 299-312.

[3]. Dahiya S., Handa S.S., and Singh N.P. (2015). "Credit Scoring Using Ensemble of Various Classifiers on Reduced Feature Set". *Manav Rachna International University (MRIU), Department of Computer Science and Engineering., Faridabad, India*.

[4]. Kohonen T (1982). "Self-organized formation of topologically correct feature maps". *Biological Cybernetics 43:59–69.*

[5]. Asan U. and Ercan S. (2012). "An Introduction to Self-Organizing Maps*". Department of Industrial Engineering, Istanbul Technical University, 34357, Macka, Istanbul, Turkey.*

[6]. W.E. Henley and D.J. Hand "A k-nearest neighbour classifier for assessing consumer credit risk". *Royal Statistical Society. The Statistician, Volume 45, Issue 1 (1996), 77-95.*

[7]. Priyanka D., Mishika S., Gopal P. & Amit H. (2020). "Credit Scoring: A Comparison between Random Forest Classifier and K- Nearest Neighbours for Credit Defaulters Prediction". *International Research Journal of Engineering and Technology (IRJET).*

[8]. S. Suleiman and S. Badamsi (2019) "Effect of Multicollinearity in Predicting Diabetes Mellitus Using Statistical Neural Networks". *European Journal of Advances in Engineering and Technology, 2019, 6(6):30-38.*

[9]. S.U. Gulumbe, S. Suleiman, S. Badamasi, A.Y. Tambuwal and Umar Usman (2019). "Predicting Diabetes Mellitus Using Artificial Neural Network through a Simulation Study". *Machine Learning Research*. Vol. 4, No. 2, 2019, pp. 33-38.

[10]. Ali A., Ning C. and Bernardete R. (2016) "Improve credit scoring using transfer of learned knowledge from self-organizing map", *The Natural Computing Applications Forum 2016.*

[11]. A.A. Sawant & P.M. Chawan (2013). "Study of Data Mining Techniques used for Financial Data Analysis*". International Journal of Engineering Science and Innovative Technology, Vol. 3, Issue 3, May 2013.*

[12]. S. Suleiman, S.U. Gulumbe and N. Shehu (2012) "Default predictors in Agricultural Credit Scoring: Evidence from Bank of Agriculture (BOA) Data". *Nigerian Statistical Association (NSA) 2012 Conference Proceedings on the Theme Statistics: A Tool for National Transformation Held at Royal Choice In. Limited, Makurdi, benue State p32-43"*.