

## Research Article

## Open Access

## Effects of Correlation and Outliers on Data Reduction and Classification of High Dimensional Data

Emmanuel Idowu Olamide\* and Oluwatosin Olabode Olawumi

Department of Statistics, Federal University of Technology Akure, Nigeria

### ABSTRACT

In this study, Maximum depth and Depth distribution classification methods were considered for high dimensional data that are highly correlated in order to determine the effective method of classification in dimensional reduction using variable combinations and variable selection approaches with applications to some real-life data and two different simulation studies. Four levels of correlation coefficients were considered in order to examine the degree of correlation that exists in the high dimensional data set. The performances of depth functions in Maximum depth and depth distribution classifications were investigated for both the high dimensional dataset and the now reduced dataset when t and F tests were applied and the optimal depth function in Maximum depth classification and depth distribution classification was identified. The depth function with the highest mean proportion of correct classification was identified to be the optimal depth function. Moreover, the performances of this optimal depth function were evaluated for the two simulations and the real-life dataset.

### \*Corresponding author

: Emmanuel Idowu Olamide, Department of Statistics, Federal University of Technology Akure, Nigeria.

**Received:** September 25, 2026; **Accepted:** October 13, 2026, **Published:** October 19, 2026

**Keywords:** Classification, High Dimensionality, Maximum Depth Function, Optimal Depth Function, Variable Selection, Spatial Depth Function

### Introduction

The complexity of a dataset lies in its dimensionality. The higher the dimension of a dataset the more complex it becomes. Hence, the solution becomes more complex too and thus, a need for special methods for handling such dataset arises. That is, there have to be ways of reducing the dimensionality before they can be processed for analysis, because without the dimension reduction, analysis becomes difficult and brings a less-desiring result. Dimension reduction does not decrease the quality of the data nor diminishes it, but rather transform it to a new data set for easy, accessible and analyzable data.

High dimensionality in dataset occurs when the number of attributes of the dataset, that is, what constitutes a particular dataset is greater in number than the whole number of observations or sample sizes. An instance of this is the blood sample collected from 100 patients in Hospital, each sample from each patient contains over thousands of attributes, and the problem arises when there is a need to analyze such data with its attributes. Hence the problems of high dimensional data arise and the ways to go about these problems arise too. Some groups of the dataset may share similar characteristics while other groups from the same data set also share some similarities. Grouping the dataset in these groups that have common characters is called classification and the algorithm that classifies this dataset into their various groups is called classifiers.

Classifiers can be a model or an algorithm depending on the types of classification as there are parametric and non-parametric methods of classification. When a classifier makes use of a parametric model or a particular distribution in classification, then the method of a classification is said to be parametric while it is non-parametric when the classifier makes use of an algorithm or method that are distribution free in the classification process.

One of the major problems high dimensional data poses is termed “Curse of dimensionality”. This problem causes low accuracy, inefficiency of classification method or algorithms and the term was first coined by which simply means most classifiers are not practically possible on high dimensional data. Another challenge of high dimensionality is the problem of what is described as “poor generalization ability”. This arises when a small portion of the high dimensional data set cause the classification model to overfit to the data, thus causing a poor generalization ability for the model [1].

One of the major and cogent solutions to these problems is to reduce the dimensionality of the dataset. This is referred to as “dimension reduction”. Dimension reduction is just the transformation of data from high dimensional space to low dimensional space so that the transformed dataset still retains the intrinsic properties of the old dataset (i.e. still retain the properties of the high dimensional data set). Several methods can be applied for dimension reduction. Some of the common ones are principal component analysis, linear discriminant analysis, canonical correlation analysis, amongst others. The most commonly used method to reduce the dimensionality is principal component analysis.

Anderson defined classification problem as the problem of decision making; where a decision is made as to which population does a particular new data set belongs. That is, it is a problem of classifying new objects into the classes of existing objects. This can however lead to a problem of misclassification (error committed when a particular data set is misclassified). It should be noted that in classification, misclassification error should be as slow as possible because high error of misclassification leads to low relative efficiency of the method or approach employed. Classification method may be parametric or non-parametric [2].

Parametric method of classification is based on strong assumption about the form of the mapping function with a prior knowledge. A learning model that summarizes data with a set of parameters of fixed size is known as parametric model. Bayes (1934) developed a classifier which is referred to as Naive Bayes (a parametric method of classification). It is widely used in the classification problem but some parametric methods have been proposed over the years that have been found to have Bayes' equivalence, but Bayes' is still widely use today. Examples of parametric methods of classifications are linear discriminant analysis (LDA), quadratic discriminant analysis (QDA). Linear discriminant analysis and quadratic discriminant analysis are well known in literature for classification of multivariate outcomes. Both are parametric because they involve knowledge or assumption of distribution of dataset. Fisher (1936) proposed that if a two population have the same covariance matrix and are both multivariate normally distributed, classification of observations into either population is optimal if using linear discriminant analysis. Also, if the two populations have dissimilar covariance matrix, classification exercise is of optimal performance if quadratic discriminant analysis is employed.

The importance of data depth lies in nonparametric approach for analyzing multivariate data. It provides one possible way of ordering the multivariate data. This is called a centre-outward ordering. Data depth measures the depth or outlyingness of a particular multivariate sample with respect to its underlying distribution. Thus, in general, any function which provides any reasonable centre-outward ordering of points in multidimensional space is termed as a depth function

Tukey (1975) first proposed the notion of data depth as a tool for visualizing bivariate data set graphically. Donoho and Gasko (1992) extended the concept of data depth to dimension greater than two [3,4].

The definition of depth function was formally formulated by Zuo and serfling. Depth function is defined as the degree of centrality of a point in respect to a probability distribution. The concept of data depth in multivariate context has been extended to functional data [5].

## Materials and Methods

### Classification Methods

#### Maximum depth classification

Suppose  $Y$  is a random observation in  $R^d$  having a distribution  $F$  and  $D(F, Y)$  is a depth function of  $Y$  with respect to distribution  $F$ . One can define the distribution function of  $D(F, Y)$ , following Makinde and Fasoranbaku (2018), as,  $CF(a) = P(D(F, Y) \leq a)$ .

A related notion to  $CF(a)$  is the distribution of the spatial outlyingness function. Makinde (2019) considered distribution function of data depth in function space and proposed a classification rule for functional data [6,7].

Consider  $J$  distributions;  $F_1, F_2, \dots, F_J$  the classification rule based on the depth distribution for any random vector  $Y$  is defined as

assign  $Y$  to  $F_k$  if  $CF_k(ak) = \max CF_j(aj); 1 \leq j \leq J$ ,

where  $aj = D(F_j, y)$  is the depth value of  $y$  with respect to  $F_j$ ,  $j = 1, 2, 3, \dots, J$

#### Maximum Depth Classification Rule

Ghosh and Chaudhuri (2005) proposed a maximum rule based on simple idea of assigning a new observation to any of the  $J$  competing classes, for which it attains maximal depth value. It is expressed as;

$$D(FK, Y) = \arg \max D(Fj, Y); \quad 1 \leq j \leq J,$$

Where  $F_k$  is the distribution of  $k^{\text{th}}$  class and  $1 \leq j \leq J$ . Let us consider two classes for simplicity. In two-class problem, the classification rule based on maximum depth classification rule can be expressed as;

assign  $Y$  to  $F$  if  $D(F, Y) > D(G, Y)$  and to  $G$  if otherwise [8].

#### Depth Distribution Classification

Suppose  $Y$  is a random observation in  $R^d$  having a distribution  $F$  and  $D(F, Y)$  is a depth function of  $Y$  with respect to distribution  $F$ . One can define the distribution function of  $D(F, Y)$ , following Makinde and Fasoranbaku (2018), as  $CF(a) = P(D(F, Y) \leq a)$ .

A related notion to  $CF(a)$  is the distribution of the spatial outlyingness function. Makinde (2019) considered distribution function of data depth in function space and proposed a classification rule for functional data.

Consider  $J$  distributions;  $F_1, F_2, \dots, F_J$  the classification rule based on the depth distribution for any random vector  $Y$  is defined as

assign  $Y$  to  $F_k$  if  $CF_k(ak) = \max CF_j(aj); 1 \leq j \leq J$ ,

where  $aj = D(F_j, y)$  is the depth value of  $y$  with respect to  $F_j$ ,  $j = 1, 2, 3, \dots, J$

#### Variable Selection

##### F test

An  $F$ -test is any statistical test in which the test statistic has an  $F$ -distribution under the null hypothesis. It is most often used when comparing statistical models that have been fitted to a data set, in order to identify the model that best fits the population from which the data were sampled. Exact " $F$ -tests" mainly arise when the models have been fitted to the data using least squares. The name was coined by George W. Snedecor, in honour of Sir Ronald A. Fisher. Fisher initially developed the statistic as the variance ratio in the 1920's.

$$F \text{ is given as } \frac{\text{Between-group variability}}{\text{Within-group variability}}$$

##### t-test

t test or student t distribution is a test statistics used for null hypothesis when the number of observation is less than 30. A t test helps in determine when a mean is different from a particular value. t- test is defined as

$$t = \frac{\bar{x} - \mu}{\sqrt{s/n}}.$$

## Results and Discussions

### Simulation Studies

Two simulated examples are presented to illustrate the effects of outliers and correlation on high dimensional data classifiers. For each of the simulation examples, the correlation coefficient  $\rho$  is taken to be 0.9, 0.75, 0.50, and 0.25.

The experiment is performed and the simulation size is taken to be 500. Analysis was done using R programming.

- Let  $X$  and  $Y$  be the two random vectors, such that  $X=A+B$  and  $Y=A+C$ , where  $A \sim N(\mu, \Sigma)$ ,  $B \sim \text{Exp}(0.5)$ ,  $C \sim \text{Exp}(1.5)$ . If  $1 \leq j \leq 100$  and  $C_j \sim \text{Exp}(0.5)$ , otherwise,  $\mu_j = 0$  for all  $j$ ,  $\Sigma_{ij} = \rho |i - j|$ , for all values of  $i$  and  $j$ . A random training sample

of size 100 is selected from each of  $X$  and  $Y$  groups. Similarly, a random test sample of size 100 is selected from each of the two groups.

- Let  $X$  and  $Y$  be random vectors distributed as  $N(\mu_X, \Sigma)$  and  $N(\mu_Y, \Sigma)$ , where,  $\mu_X = 0$ , if  $1 \leq j \leq 100$ ; and

$$\mu_X = 2 \text{ if } j > 100 \text{ and } \mu_Y = 2, \Sigma = \begin{pmatrix} 1 & \rho & \dots & \rho^{98} & \rho^{99} \\ \rho & 1 & \dots & \rho^{97} & \rho^{98} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \rho^{99} & \rho^{98} & \dots & \rho & 1 \end{pmatrix}$$

A random training sample of size 100 and a random test sample of size 100 are selected from each of the  $X$  and  $Y$  groups

**Table 1: Descriptive Statistics of Simulations 1 at 0.90, 0.75, 0.50 and 0.25 Level of Correlation Coefficients**

| Descriptive statistics of simulation 1 for rho=0.90 |               |                 |                 |               |                 |                 |                |                  |                  |                |                  |                  |
|-----------------------------------------------------|---------------|-----------------|-----------------|---------------|-----------------|-----------------|----------------|------------------|------------------|----------------|------------------|------------------|
| Descriptive Statistics                              | Maxdepth+ SPD | Maxdepth +SPD+t | Maxdepth +SPD+F | Maxdepth+ MBD | Maxdepth+ MBD+t | Maxdepth+ MBD+F | Depthdist+ SPD | Depthdist+ SPD+t | Depthdist+ SPD+F | Depthdist+ MBD | Depthdist +MBD+t | Depthdist+ MBD+F |
| Minimum                                             | 0.415         | 0.5             | 0.495           | 0.575         | 0.575           | 0.575           | 0.05           | 0.495            | 0.5              | 0.495          | 0.48             | 0.46             |
| 1st quarter                                         | 0.465         | 0.53            | 0.53            | 0.63          | 0.63            | 0.63            | 0.125          | 0.52             | 0.52             | 0.575          | 0.55             | 0.55             |
| Median                                              | 0.475         | 0.545           | 0.545           | 0.65          | 0.645           | 0.645           | 0.1575         | 0.53             | 0.53             | 0.595          | 0.57             | 0.57             |
| 3rd quarter                                         | 0.49          | 0.555           | 0.555           | 0.665         | 0.66            | 0.66            | 0.205          | 0.54             | 0.54             | 0.615          | 0.59             | 0.59             |
| Maximum                                             | 0.545         | 0.6             | 0.63            | 0.74          | 0.73            | 0.73            | 0.415          | 0.6              | 0.58             | 0.69           | 0.65             | 0.65             |
| Mean                                                | 0.47771       | 0.5447          | 0.54479         | 0.64911       | 0.64655         | 0.64655         | 0.16911        | 0.531231         | 0.53196          | 0.59708        | 0.56906          | 0.56952          |
| Standard Deviation                                  | 0.0187379     | 0.0185742       | 0.0193457       | 0.0244756     | 0.02232587      | 0.02232587      | 0.0620359      | 0.0157267        | 0.01513          | 0.0298086      | 0.028663         | 0.02859905       |
| Descriptive statistics of simulation 1 for rho=0.75 |               |                 |                 |               |                 |                 |                |                  |                  |                |                  |                  |
| Minimum                                             | 0.455         | 0.495           | 0.495           | 0.605         | 0.64            | 0.64            | 0              | 0.495            | 0.495            | 0.55           | 0.5              | 0.64             |
| 1st quarter                                         | 0.485         | 0.51            | 0.51            | 0.68          | 0.69            | 0.69            | 0.055          | 0.51             | 0.505            | 0.64           | 0.6              | 0.69             |
| Median                                              | 0.495         | 0.52            | 0.515           | 0.695         | 0.705           | 0.705           | 0.085          | 0.515            | 0.515            | 0.655          | 0.625            | 0.705            |
| 3rd quarter                                         | 0.5           | 0.525           | 0.525           | 0.715         | 0.72            | 0.72            | 0.12           | 0.525            | 0.525            | 0.68           | 0.645            | 0.72             |
| Maximum                                             | 0.59          | 0.565           | 0.565           | 0.785         | 0.775           | 0.775           | 0.41           | 0.57             | 0.56             | 0.77           | 0.72             | 0.775            |
| Mean                                                | 0.49311       | 0.51963         | 0.51866         | 0.6794        | 0.70405         | 0.70405         | 0.09919        | 0.5163           | 0.51588          | 0.65819        | 0.62422          | 0.70405          |
| Standard Deviation                                  | 0.0132487     | 0.0126781       | 0.011878        | 0.0279697     | 0.02424713      | 0.02424713      | 0.0670559      | 0.0119852        | 0.0117004        | 0.0329111      | 0.031905         | 0.02424713       |
| Descriptive statistics of simulation 1 for rho=0.50 |               |                 |                 |               |                 |                 |                |                  |                  |                |                  |                  |
| Minimum                                             | 0.47          | 0.5             | 0.5             | 0.635         | 0.68            | 0.68            | 0              | 0.495            | 0.495            | 0.565          | 0.565            | 0.585            |
| 1st quarter                                         | 0.495         | 0.5             | 0.5             | 0.72          | 0.73375         | 0.73375         | 0.02           | 0.505            | 0.505            | 0.675          | 0.655            | 0.0655           |
| Median                                              | 0.5           | 0.505           | 0.505           | 0.74          | 0.745           | 0.745           | 0.04           | 0.04             | 0.51             | 0.7            | 0.675            | 0.675            |
| 3rd quarter                                         | 0.5           | 0.51            | 0.51            | 0.76          | 0.765           | 0.765           | 0.07625        | 0.515            | 0.515            | 0.725          | 0.695            | 0.7              |
| Maximum                                             | 0.545         | 0.545           | 0.545           | 0.845         | 0.83            | 0.83            | 0.37           | 0.55             | 0.545            | 0.815          | 0.77             | 0.77             |
| Mean                                                | 0.49872       | 0.50783         | 0.50783         | 0.7389        | 0.74831         | 0.74831         | 0.06009        | 0.50977          | 0.5108           | 0.70143        | 0.67445          | 0.67693          |
| Standard Deviation                                  | 0.0073396     | 0.0076654       | 0.0076654       | 0.033221      | 0.025579337     | 0.02557934      | 0.0595148      | 0.0087149        | 0.0093077        | 0.0390404      | 0.031339         |                  |
| Descriptive statistics of simulation 1 for rho=0.25 |               |                 |                 |               |                 |                 |                |                  |                  |                |                  |                  |
| Minimum                                             | 0.48          | 0.5             | 0.5             | 0.615         | 0.685           | 0.685           | 0              | 0.495            | 0.495            | 0.56           | 0.605            | 0.605            |
| 1st quarter                                         | 0.5           | 0.5             | 0.5             | 0.72875       | 0.75            | 0.75            | 0.01           | 0.5              | 0.5              | 0.695          | 0.685            | 0.685            |
| Median                                              | 0.5           | 0.505           | 0.505           | 0.75          | 0.77            | 0.77            | 0.025          | 0.505            | 0.505            | 0.72           | 0.705            | 0.705            |
| 3rd quarter                                         | 0.5           | 0.505           | 0.505           | 0.775         | 0.79            | 0.79            | 0.055          | 0.51             | 0.51             | 0.75           | 0.73             | 0.73             |
| Maximum                                             | 0.53          | 0.545           | 0.53            | 0.84          | 0.86            | 0.86            | 0.34           | 0.555            | 0.55             | 0.82           | 0.815            | 0.8              |
| Mean                                                | 0.50027       | 0.50433         | 0.50415         | 0.75171       | 0.76931         | 0.76931         | 0.04544        | 0.50744          | 0.50792          | 0.72129        | 0.70647          | 0.70597          |
| Standard Deviation                                  | 0.004617      | 0.005211        | 0.005097        | 0.036766      | 0.0276261       | 0.0276261       | 0.055664       | 0.008628         | 0.007816         | 0.040366       | 0.03204          | 0.032458         |

Table 2: Descriptive Statistics of Simulations 2 at 0.90, 0.75, 0.50 and 0.25 Level of Correlation Coefficients

| Descriptive statistics of simulation 2 for rho=0.90 |               |                 |                 |               |                 |                 |                |                  |                  |                |                  |                  |
|-----------------------------------------------------|---------------|-----------------|-----------------|---------------|-----------------|-----------------|----------------|------------------|------------------|----------------|------------------|------------------|
| Descriptive Statistics                              | Maxdepth+ SPD | Maxdepth +SPD+t | Maxdepth +SPD+F | Maxdepth+ MBD | Maxdepth+ MBD+t | Maxdepth+ MBD+F | Depthdist+ SPD | Depthdist+ SPD+t | Depthdist+ SPD+F | Depthdist+ MBD | Depthdist +MBD+t | Depthdist+ MBD+F |
| Minimum                                             | 0.965         | 0.965           | 0.965           | 0.96          | 0.95            | 0.95            | 0.925          | 0.86             | 0.88             | 0.91           | 0.765            | 0.775            |
| 1st quarter                                         | 0.985         | 0.99            | 0.99            | 0.98          | 0.985           | 0.985           | 0.965          | 0.93             | 0.93             | 0.96           | 0.86             | 0.865            |
| Median                                              | 0.99          | 0.995           | 0.995           | 0.99          | 0.99            | 0.99            | 0.975          | 0.945            | 0.945            | 0.97           | 0.885            | 0.89             |
| 3rd quarter                                         | 0.995         | 0.995           | 0.995           | 0.99          | 0.995           | 0.995           | 0.985          | 0.955            | 0.955            | 0.98           | 0.91             | 0.91             |
| Maximum                                             | 1             | 1               | 1               | 1             | 1               | 1               | 1              | 0.955            | 0.99             | 1              | 0.975            | 0.96             |
| Mean                                                | 0.99115       | 0.99202         | 0.9918          | 0.9851        | 0.98814         | 0.98814         | 0.97384        | 0.94162          | 0.94252          | 0.97035        | 0.88451          | 0.88574          |
| Standard Deviation                                  | 0.00670951    | 0.006309        | 0.006507        | 0.008391      | 0.00765883      | 0.0076588       | 0.0146453      | 0.0198409        | 0.018456         | 0.0148448      | 0.0376875        | 0.035298         |
| Descriptive statistics of simulation 2 for rho=0.75 |               |                 |                 |               |                 |                 |                |                  |                  |                |                  |                  |
| Descriptive Statistics                              | Maxdepth+ SPD | Maxdepth +SPD+t | Maxdepth +SPD+F | Maxdepth+ MBD | Maxdepth+ MBD+t | Maxdepth+ MBD+F | Depthdist+ SPD | Depthdist+ SPD+t | Depthdist+ SPD+F | Depthdist+ MBD | Depthdist +MBD+t | Depthdist+ MBD+F |
| Minimum                                             | 0.995         | 0.995           | 0.995           | 0.99          | 0.995           | 0.99            | 0.905          | 0.9              | 0.885            | 0.9            | 0.75             | 0.675            |
| 1st quarter                                         | 1             | 1               | 1               | 1             | 1               | 1               | 0.965          | 0.94             | 0.945            | 0.97           | 0.84             | 0.84             |
| Median                                              | 1             | 1               | 1               | 1             | 1               | 1               | 0.98           | 0.955            | 0.955            | 0.985          | 0.87             | 0.87             |
| 3rd quarter                                         | 1             | 1               | 1               | 1             | 1               | 1               | 0.99           | 0.97             | 0.965            | 0.99           | 0.895            | 0.895            |
| Maximum                                             | 1             | 1               | 1               | 1             | 1               | 1               | 1              | 0.995            | 0.99             | 1              | 0.97             | 0.97             |
| Mean                                                | 0.99993       | 0.99991         | 0.99995         | 0.99972       | 0.99983         | 0.99982         | 0.97678        | 0.95427          | 0.95393          | 0.97898        | 0.86629          | 0.86745          |
| Standard Deviation                                  | 0.00058804    | 0.000665        | 0.000498        | 0.001194      | 0.00090705      | 0.0009847       | 0.0175513      | 0.0180796        | 0.017067         | 0.016189       | 0.0411555        | 0.040689         |
| Descriptive statistics of simulation 2 for rho=0.50 |               |                 |                 |               |                 |                 |                |                  |                  |                |                  |                  |
| Descriptive Statistics                              | Maxdepth+ SPD | Maxdepth +SPD+t | Maxdepth +SPD+F | Maxdepth+ MBD | Maxdepth+ MBD+t | Maxdepth+ MBD+F | Depthdist+ SPD | Depthdist+ SPD+t | Depthdist+ SPD+F | Depthdist+ MBD | Depthdist +MBD+t | Depthdist+ MBD+F |
| Minimum                                             | 1             | 1               | 1               | 1             | 1               | 1               | 0.855          | 0.88             | 0.87             | 0.88           | 0.7              | 0.66             |
| 1st quarter                                         | 1             | 1               | 1               | 1             | 1               | 1               | 0.955          | 0.945            | 0.945            | 0.965          | 0.815            | 0.815            |
| Median                                              | 1             | 1               | 1               | 1             | 1               | 1               | 0.97           | 0.96             | 0.955            | 0.975          | 0.845            | 0.85             |
| 3rd quarter                                         | 1             | 1               | 1               | 1             | 1               | 1               | 0.98           | 0.97             | 0.97             | 0.99           | 0.875            | 0.88             |
| Maximum                                             | 1             | 1               | 1               | 1             | 1               | 1               | 1              | 1                | 0.99             | 1              | 0.95             | 0.955            |
| Mean                                                | 1             | 1               | 1               | 1             | 1               | 1               | 0.96708        | 0.95688          | 0.95496          | 0.97349        | 0.84424          | 0.84591          |
| Standard Deviation                                  | 0             | 0               | 0               | 0             | 0               | 0               | 0.0218501      | 0.0171217        | 0.018428         | 0.0197497      | 0.0446386        | 0.046776         |
| Descriptive statistics of simulation 1 for rho=0.25 |               |                 |                 |               |                 |                 |                |                  |                  |                |                  |                  |
| Descriptive Statistics                              | Maxdepth+ SPD | Maxdepth +SPD+t | Maxdepth +SPD+F | Maxdepth+ MBD | Maxdepth+ MBD+t | Maxdepth+ MBD+F | Depthdist+ SPD | Depthdist+ SPD+t | Depthdist+ SPD+F | Depthdist+ MBD | Depthdist +MBD+t | Depthdist+ MBD+F |
| Minimum                                             | 1             | 1               | 1               | 1             | 1               | 1               | 0.85           | 0.895            | 0.89             | 0.89           | 0.675            | 0.66             |
| 1st quarter                                         | 1             | 1               | 1               | 1             | 1               | 1               | 0.94           | 0.945            | 0.945            | 0.95875        | 0.795            | 0.79             |
| Median                                              | 1             | 1               | 1               | 1             | 1               | 1               | 0.96           | 0.96             | 0.955            | 0.975          | 0.83             | 0.82             |
| 3rd quarter                                         | 1             | 1               | 1               | 1             | 1               | 1               | 0.98           | 0.97             | 0.97             | 0.985          | 0.855            | 0.855            |
| Maximum                                             | 1             | 1               | 1               | 1             | 1               | 1               | 1              | 0.995            | 0.995            | 1              | 0.945            | 0.925            |
| Mean                                                | 1             | 1               | 1               | 1             | 1               | 1               | 0.95621        | 0.95711          | 0.95586          | 0.96909        | 0.82622          | 0.81921          |
| Standard Deviation                                  | 0             | 0               | 0               | 0             | 0               | 0               | 0.0274571      | 0.0174788        | 0.017406         | 0.0209882      | 0.0469737        | 0.046975         |

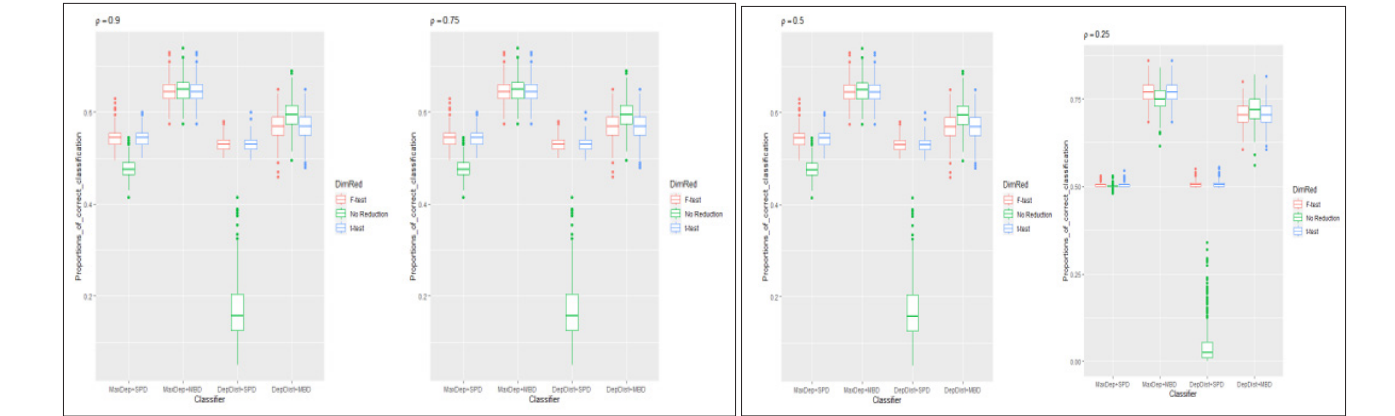


Figure 1: Boxplot of Simulation 1at 0.90, 0.75, 0.50 and 0.25 Level of Correlation Coefficients.

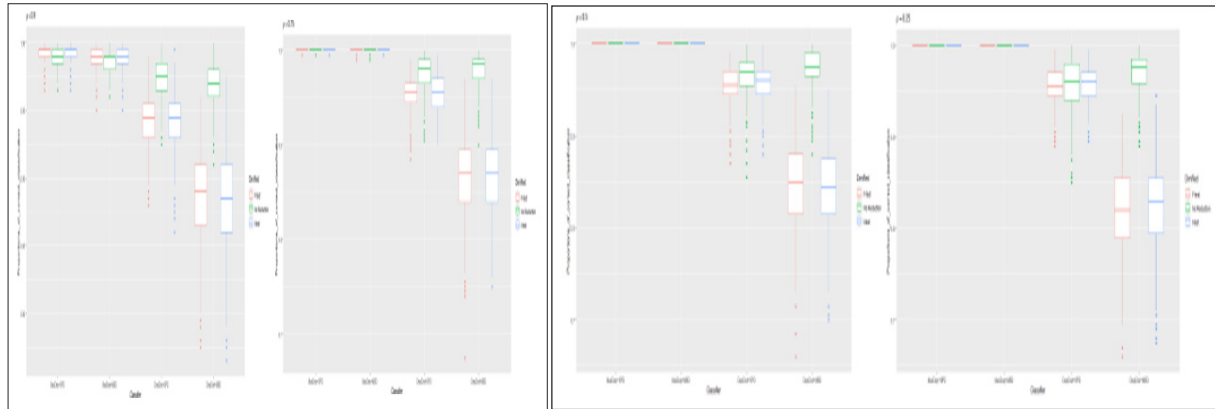


**Figure 2:** Boxplot of Simulation 2 at 0.90, 0.75, 0.50 and 0.25 Level of Correlation Coefficients.

### Applications to Real Life Data

We analyse two benchmark data sets to illustrate the performances of the classification methods. These datasets are Madelon data and Prostate Cancer data

#### Madelon Data

Madelon data is artificial dataset, which consists of two groups with 500 continuous input variables. The groups are positive and negative (+1 and -1), a random training sample of size 200 is selected from each of the groups. The experiment was repeated 500 times. In each replication, p-value is estimated and power is determined as the proportion of p-values less than level of significance.

We analyze Madelon data. Madelondata, available at <https://archive.ics.uci.edu/ml/machine-learning-databases/Madelon/madelon-train.data>

#### Prostate Cancer Data

Prostate Cancer data is available at <https://archive.ics.uci.edu/ml/machine-learning-databases//Prostatecancer/PROSTATECANCER-train.data>

**Table 3: Descriptive Statistics for Madelon Data**

| Descriptive statistics of real data-Madelon |               |                 |                 |               |                 |                 |                |                  |                  |                 |                  |                  |
|---------------------------------------------|---------------|-----------------|-----------------|---------------|-----------------|-----------------|----------------|------------------|------------------|-----------------|------------------|------------------|
| Descriptive Statistics                      | Maxdepth+ SPD | Maxdepth+ SPD+t | Maxdepth+ SPD+F | Maxdepth+ MBD | Maxdepth+ MBD+t | Maxdepth+ MBD+F | Depthdist+ SPD | Depthdist+ SPD+t | Depthdist+ SPD+F | Depthdist+ +MBD | Depthdist+M BD+t | Depthdist+ MBD+F |
| Minimum                                     | 0.49          | 0.515           | 0.5125          | 0.5           | 0.48            | 0.4825          | 0.5025         | 0.51             | 0.5275           | 0.4775          | 0.2725           | 0.235            |
| 1st quarter                                 | 0.559375      | 0.57            | 0.5725          | 0.5425        | 0.555           | 0.555           | 0.55           | 0.57             | 0.57             | 0.525           | 0.325            | 0.3275           |
| Median                                      | 0.575         | 0.5875          | 0.585           | 0.5575        | 0.5725          | 0.5725          | 0.565          | 0.585            | 0.585            | 0.54125         | 0.3475           | 0.3525           |
| 3rd quarter                                 | 0.5925        | 0.6             | 0.6             | 0.575         | 0.5875          | 0.5875          | 0.58           | 0.6025           | 0.600625         | 0.555           | 0.37             | 0.3725           |
| Maximum                                     | 0.6675        | 0.645           | 0.6475          | 0.62          | 0.6625          | 0.6525          | 0.6525         | 0.66             | 0.6675           | 0.6075          | 0.425            | 0.4425           |
| Mean                                        | 0.575365      | 0.58598         | 0.584595        | 0.55917       | 0.5715          | 0.571345        | 0.56517        | 0.58519          | 0.5855           | 0.54011         | 0.34752          | 0.349245         |
| Standard Deviation                          | 0.0259731     | 0.0218637       | 0.021015        | 0.0231194     | 0.0246382       | 0.0248569       | 0.023196       | 0.0250423        | 0.0238837        | 0.02315         | 0.03041092       | 0.03186689       |

**Table 4: Descriptive Statistics for Prostate Cancer Data**

| Descriptive statistics of real data-Prostate Cancer |               |                 |                 |               |                 |                 |                |                  |                  |                 |                  |                  |
|-----------------------------------------------------|---------------|-----------------|-----------------|---------------|-----------------|-----------------|----------------|------------------|------------------|-----------------|------------------|------------------|
| Descriptive Statistics                              | Maxdepth+ SPD | Maxdepth+ SPD+t | Maxdepth+ SPD+F | Maxdepth+ MBD | Maxdepth+ MBD+t | Maxdepth+ MBD+F | Depthdist+ SPD | Depthdist+ SPD+t | Depthdist+ SPD+F | Depthdist+ +MBD | Depthdist+M BD+t | Depthdist+ MBD+F |
| Minimum                                             | 0.4047619     | 0.6428571       | 0.6428571       | 0.4285714     | 0.547619        | 0.547619        | 0.404762       | 0.5952381        | 0.5952381        | 0.380952        | 0.5238095        | 0.5238095        |
| 1st quarter                                         | 0.6666667     | 0.6428571       | 0.6428571       | 0.5952381     | 0.547619        | 0.547619        | 0.619005       | 0.5952381        | 0.5952381        | 0.547619        | 0.5238095        | 0.5238095        |
| Median                                              | 0.7142857     | 0.6428571       | 0.6428571       | 0.6190476     | 0.547619        | 0.547619        | 0.666667       | 0.5952381        | 0.5952381        | 0.595238        | 0.5238095        | 0.5238095        |
| 3rd quarter                                         | 0.7619048     | 0.6428571       | 0.6428571       | 0.6904762     | 0.547619        | 0.547619        | 0.714286       | 0.5952381        | 0.5952381        | 0.619048        | 0.5238095        | 0.5238095        |
| Maximum                                             | 0.9285714     | 0.7619048       | 0.8571429       | 0.8571429     | 0.7380952       | 0.6666667       | 0.857143       | 0.5952381        | 0.6904762        | 0.833333        | 0.714285         | 0.5238095        |
| Mean                                                | 0.7059048     | 0.6435714       | 0.6432857       | 0.6354286     | 0.548           | 0.5478571       | 0.661191       | 0.5952381        | 0.5954286        | 0.58581         | 0.5241905        | 0.5238095        |
| Standard Deviation                                  | 0.0765793     | 0.0920289       | 0.0095831       | 0.0721748     | 0.0085184       | 0.005324        | 0.07109        | 0                | 0.0042592        | 0.066815        | 0.00851835       | 0                |

### Applications to Real Life Data

We analyse two benchmark data sets to illustrate the performances of the classification methods. These datasets are Madelon data and Prostate Cancer data

## Madelon Data

Madelon data is artificial dataset, which consists of two groups with 500 continuous input variables. The groups are positive and negative (+1 and -1), a random training sample of size 200 is selected from each of the groups. The experiment was repeated 500 times. In each replication, p-value is estimated and power is determined as the proportion of p-values less than level of significance.

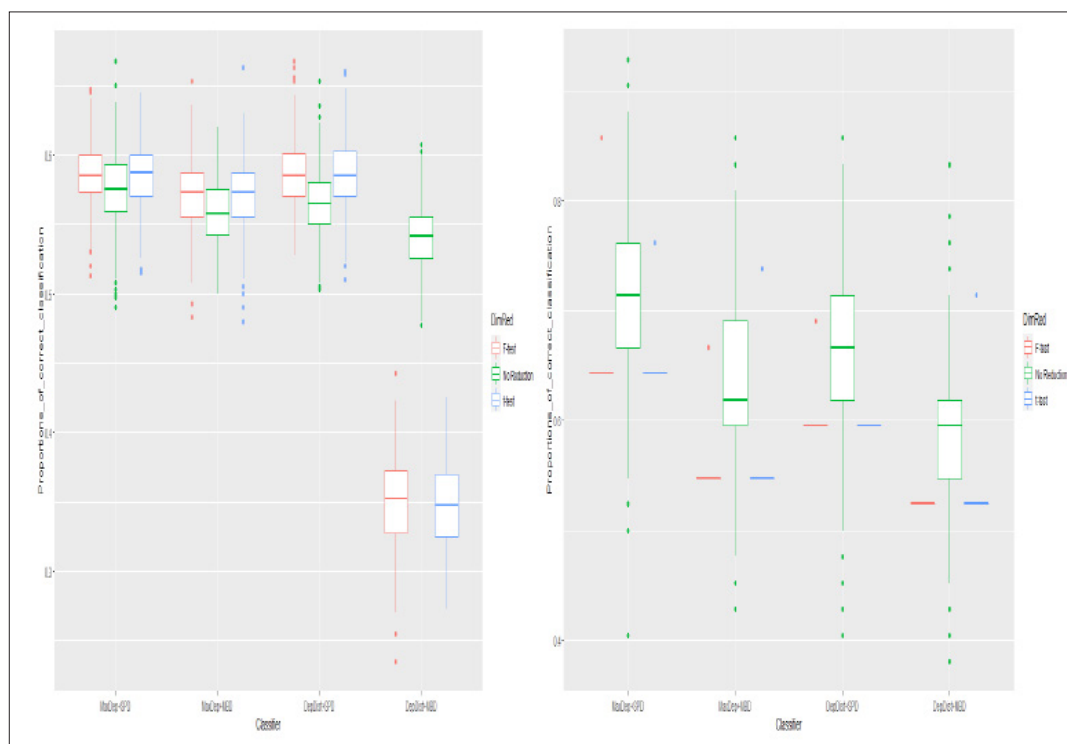


**Figure 3:** Boxplots of Madelon Data and Prostate Cancer Data

## Conclusion

The optimal depth function from the analysis of the two simulations and the real dataset (Madelon and Prostate cancer) was found for high dimensional data (when there is and no data reduction). The spatial depth function performed optimally (by achieving the highest mean proportion of correct classification) among the depth functions considered for depth-oriented F-test and t-test. In this study, both the F-test and t-test employed in the dimension reduction performed almost at the same rate in the dimension reduction process for various values of correlation coefficient,  $\rho$ . Though, modified band depth performed well at some point across the four levels of correlation coefficients that were used, spatial depth function mostly achieved the highest proportion of correct classifications. For the two simulations and well as the two real datasets (Madelon and Prostate Cancer), Maximum depth classifier achieved the highest rate of proportion as a classifying method between the two classifiers that were used.

## Competing Interest

The authors declare they have no conflicts of interest.

## Acknowledgements

The authors acknowledge the support of the Department of Statistics, Federal University of Technology Akure, Nigeria.

## Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

## References

1. Bellman RE (1961) Adaptive Control Processes. Princeton University Press Princeton NJ <https://www.scrip.org/>

[reference/referencespapers?referenceid=3610272](https://www.scrip.org/reference/referencespapers?referenceid=3610272).

2. Anderson TW (1972) Asymptotic evaluation of the probabilities of misclassification by linear discriminant function. Stanford University <https://dtic.minsky.ai/index/AD0749972>.
3. Tukey J (1975) Mathematics and the picturing of data. Proceedings of the International Congress of Mathematicians <https://www.scrip.org/reference/referencespapers?referenceid=4304989>.
4. Donoho DL, Gasko M (1992) Breaking properties of location estimates based on half space depth and projected outlyingness. Annals of Statistics 20: 1803-1827.
5. Serfling R, Zuo Y (2000) General notions of statistical depth function. Annals of Statistics 28: 461- 482.
6. Makinde OS, Fasoranbaku OA (2018) On maximum depth classifiers: Depth distribution approach. Journal of Applied Statistics 45: 1106-1117.
7. Makinde OS (2019) Some classification rules based on distribution functions of functional depth. Statistical Papers 60: 629-640.
8. Ghosh AK, Chaudhuri P (2005) On maximum depth and related classifiers. Scandinavian Journal of Statistical Association 32: 327-350.

**Copyright:** ©2026 Emmanuel Idowu Olamide. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.