Advanced Journal of Science, Technology and Engineering
ISSN: 2997-5972
Volume 5, Issue 1, 2025 (pp. 70-84)
www.abjournals.org

COMPARATIVE STUDY OF LINEAR DISCRIMINANT ANALYSIS (LDA), QUADRATIC DISCRIMINANT ANALYSIS (QDA) AND SUPPORT VECTOR MACHINE (SVM) IN DATASET

Owoyi Mildred Chiyeaka[1], Okoh Jophet Ewere[2],

Obukohwo Victor[1], and Olamuyiwa Shola[2]

[1]Department of Mathematics, Dennis Osadebay University, Asaba, Delta State, Nigeria.

[2]Department of Statistics, Dennis Osadebay University, Asaba, Delta State, Nigeria.

[*]Corresponding Author's Email: jophet.okoh@dou.edu.ng

Cite this article:

Owoyi, M. C., Okoh, J. E., Obukohwo, V., Olamuyiwa, S. (2025), Comparative Study of Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA) and Support Vector Machine (SVM) In Dataset. Advanced Journal of Science, Technology and Engineering 5(1), 70-84. DOI: 10.52589/AJSTE-UAZPPMER

Manuscript History

Received: 11 Jan 2025

Accepted: 15 Feb 2025

Published: 10 Mar 2025

ABTRACT: *Classification techniques is an important factor in data analysis. Over the years, different classification method have been proposed for classification of dataset. In this paper, we compared three classifiers (LDA, QDA and SVM) in three imbalanced datasets (Iris, Pima and Glass data) and misclassification rate of the three classification method were compared. The experiments concentrated on analyzing the average misclassification rate among classifiers across the three dataset studied using the misforest imputation method to balance the dataset respectively. The results reveal that for the glass dataset, the QDA classifies the dataset better than the two other classification method studied, while for the iris and glass datasets, the LDA outperformed the other two classifiers studied. The conclusion in this study is that LDA have the least average misclassification error, followed by the QDA and then the SVM with an average misclassification rate of 0.2863.*

KEYWORDS: imbalanced data; misclassification rate; Average misclassification; classification.

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

Advanced Journal of Science, Technology and Engineering
ISSN: 2997-5972
Volume 5, Issue 1, 2025 (pp. 70-84)
www.abjournals.org

INTRODUCTION

In performing classification tasks, a discriminant analysis which aims at finding an independent and identically distributed training dataset and a discriminant function that will predict correctly the new instances is needed. Linear Discriminant Analysis is a supervised machine learning algorithm for classification and dimensionality reduction. It is widely used in multivariate statistical techniques for data analysis. Assumptions in LDA is that variables are assumed to be normally distributed with an equal covariance matrix. When performing discriminant analysis, users can discuss classification methods in which two or more groups and one or more independent variables are placed into one of the measured characteristics. Medical scientists investigate how groups (characterized by blood pressure, blood glucose levels, and age) differ across independent variables. (Fernandez et al, 2006) used discriminant analysis to Obtain the number of patients who had previously suffered a heart attack to classify if the patient would survive based on other variables. Quadratic Discriminant Analysis (QDA) is a variant of LDA in which an individual covariance matrix is estimated for every class of observations. QDA is particularly useful if there is prior knowledge that individual classes exhibit distinct covariance. One weakness of QDA is that it cannot be used as a dimensionality reduction technique. A support Vector machine is a supervised machine learning algorithm which performs classification by finding the optimal line which maximizes the distance between each class. it is used for both classification and regression task.

The LDA and QDA are both very common traditional classifiers. Both method assumes multivariate Gaussian distribution and employ variance- covariance matrix but LDA assume equal covariance matrix unlike the QDA that estimates covariance matrix per class. In QDA, $\sum_k$ is required for each class of $k \in \{1,\cdots,K\}$ rather than assuming $\sum_k = \sum$ as it is done in LDA. (Morrais and Lima, 2018) applied the principal component analysis with LDA and QDA for discriminants between healthy control and cancer samples using the MS data sets. (Nikita and Nikitas, 2021) examined seven classification methods with binary logistic, probability, and cumulative probability regression, LDA, QDA, artificial neural networks, and naıve Bayes classification, to examine skeletal sex estimation. Consequently, LDA may be more preferred in skeletal sex estimation than other methods. (Sarkodie and fergusson-res, 2021) used LDA and QDA to propose a flow regime identification which combine responses from a non-intrusive optical sensor for air and water's vertical upward gas-liquid flow. Different researched articles have used different machine learning classifiers that have been formed in recent years to resolve classification accuracy problems and evaluation metrics (Sarker, 2021; Yu et al.,2020; Alanaza et al., 2021). In ((Bickel and Levna, 2004; Pattison and Gossink, 1999), they wrote on FLD classification error in data space. (Davenport et al, 2007) wrote on the problem associated with establishing a classifier probability of misclassification error. several recent work has researched on efficient learning in low dimensional spaces. such as , in (Calderbank et al, 2009)) they explained that when dimensional data points which is high have a sparse representation in some linear basis, then a soft-margin SVM classifier can be trained on a low dimensional projection of that data while keeping a performance in classification that is same as the result obtained in the original data space.

This work is divided into five sections excluding the introduction. Section 2 explains briefly the procedures and methodology of the three classifiers. Section 3 contains the datasets, Section 4 contains the Results, and comparative performance analysis of the three classifiers while the conclusion is presented in section 5.

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

Advanced Journal of Science, Technology and Engineering
ISSN: 2997-5972
Volume 5, Issue 1, 2025 (pp. 70-84)

www.abjournals.org

Procedure and Experimental Methodology

This section includes a description of the dataset, and evaluation matrices, as well as the process and methodology used in the study.

Linear Discriminant Analysis:

The LDA is given as

$$LDA = G^1 y_{ij} = \left(\mu_1 - \mu_2\right)^T \; ps_{oc}^{-1} y_{ij} \qquad (1.1)$$

Where μ_1 and μ_2 are the mean of eack class

y_{ij} is the independent variable

ps_{oc}^{-1} inverse of the pooled sample covariance matrice

$$\overline{LDA} = \frac{\left(\mu_1 + \mu_2\right)}{2} G \qquad (1.2)$$

Equation 1.1 is the LDA defined by Ronald A. Fisher and 1.2 is the discriminant mean.

Classification Rule:

Classify y_1 to class m_1 if $LDA \geq \overline{LDA}$ and $y_1$to majority class if m_2 if $LDA < \overline{LDA}$

Quadratic Discriminant Analysis (QDA)

The quadratic discriminant function is given as

$$\delta_k = -\frac{1}{2}\log|\sum_k| - \frac{1}{2}(x-\mu_x)^T \sum_k^{-1}(x-\mu_x) + \log \pi_k \qquad (2.26)$$

Since QDA estimates a covariance matrix for each class, it has a greater number of effective parameters than LDA. The quadratic discriminant function is quadratic in nature and contains a second order terms.

The classification rule for the quadratic discriminant function:

$$\hat{G}(x) = arg \max_k \delta_k(x) \qquad (2.27)$$

The classification rule is equally similar to LDA since all that is expected is to find the class k which maximizes the quadratic discriminant function.

QDA Algorithm.

1. Input independent variables $X = (x_1, ..., x_n)$ of p samples
2. Find the prior probability for each class
3. Calculate the covariance matrix for each class

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

Advanced Journal of Science, Technology and Engineering
ISSN: 2997-5972
Volume 5, Issue 1, 2025 (pp. 70-84)

www.abjournals.org

4. compute the QDA function

5. Assign class label

Support Vector Machine (SVM)

The idea of SVM go along with: Input vectors **x** are mapped to a very significant dimension feature space z through some nonlinear map $\phi(\mathbf{x})$, such that $z = \phi(\mathbf{x})$. Thus, an optimal separating hyperplane is constructed. For a given training dataset with n samples, $(x_1,\ y_1), (x_2,\ y_2), \dots, (x_n,\ y_n)$, where x_i is a feature vector in a d-dimensional feature space R^d and $y_i \in \{1,\ +1\}$ is the corresponding class label. The task is to find a classifier with a decision function $F(x, \mu, \mu_0) = \boldsymbol{\mu}^T\mathbf{x} + \mu_0$. The SVM then finds an optimal hyperplane with the maximal margin that separates the data points in both groups, (Musa, 2012). To find the optimal separating hyperplane having maximal margin, we can minimize $\| \mu \|$, that is, minimizing the objective function

Objective Function: $\min \frac{1}{2}\mu^T\mu$

Subject to;

$$y_i(\ \boldsymbol{\mu}^T\mathbf{x} + \mu_0) \geq 1 \qquad for\ \ i = 1, \dots, n \tag{3.11}$$

where **μ** is the normal vector for the ‘‘separating’’ hyperplane, $(\mu, \Phi(x)) + \mu_0 = 0$ this can be transferred within the two fold by reducing the subsequent primal lagrangian

$$\mathrm{L_d}(\mu, \mu_0, \alpha) = \frac{1}{2}\mu'\mu - \sum_{i=1}^{n} \alpha_i\{\ y_i[\mu'\phi(x_i) + \mu_0] - 1\} \tag{3.12}$$

With respect to μ and μ_0 by using $\frac{\partial L_d}{\partial \mu} = 0$ and $\frac{\partial L_d}{\partial \mu_0} = 0$

$$\frac{\partial Ld}{\partial \mu} = 0, \qquad \mu = \sum_{i=1}^{n} \alpha_i y_i \phi(x_i) \tag{3.13}$$

$$\frac{\partial L_d}{\partial \mu_0} = 0, \mu = \sum_{i=1}^{n} \alpha_i y_i = 0 \tag{3.14}$$

Making substitutions using (3.13) and (3.14), gives;

$$L_d(\alpha) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} y_i y_j\, \alpha_i \alpha_j k(x_i, x_j) \tag{3.15}$$

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

Advanced Journal of Science, Technology and Engineering
ISSN: 2997-5972
Volume 5, Issue 1, 2025 (pp. 70-84)

Where $k(x_i,\ x_j) = \Phi'(x_i)\Phi(x_j)$ is a Kernel which permits the evaluation of scalar product between a multi-scale areas beyond specifically understanding the non-linear mapping

$L_d(\alpha)$ is dependent on;

$$\alpha_i \geq 0, \quad i = 1, \dots, n$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{3.15}$$

However, when there are intersection of learning set, that problem is inseparable, the restriction in working out the two lagrangian problem in (3.15) becomes;

$$0 \leq \alpha_i \leq c, \quad i = 1, \dots, n$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{3.16}$$

where $(x_i, \dots, x_j)$ are the weights allocated to the learning set x_i. If $\alpha_i > 0$, x_i is called a support vector. c is known as a regulation parameter used to achieve a trade-off between the learning precision and the complicated design to enable an excellent concept efficacy to be achieved. (Musa, 2012).

Following the lagrangian criteria$(\alpha_i, \dots, \alpha_n)$, the decision function can be formulated as follows:

$$f(\mathbf{x}) = \mu'\mathbf{x} + \mu_0 = \sum_{i=1}^{n} \alpha_i y_i k(x, x_i) + \mu_0 \tag{3.16}$$

where **x** is the d-size vector of the test samples and μ_0 is the SVM predilection expression that rely upon the exert kernel which could be suggested segment of the kernel function, (Musa, 2012).

It was discovered, by fulfilling the conditions that the values of a decision function at the support vectors ought to be given as, y_i, $(y_i = \pm 1)$. $f(x_s) = y_s = \pm 1$.

classification rule:

$$x_p = \begin{cases} 1 \text{ if } x_p > 0 \\ -1 \text{ if } x_p < 0 \end{cases}$$

Performance evaluation:

The confusion matrix will be use to check the performance of the models. The accuracy will be tested as well as the misclassification error rate.

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

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

$$\text{misclassification rate} = \frac{FP+FN}{TP+TN+FP+FN}$$

DATASETS

1. Iris Dataset

Edgar Anderson's Iris Data, often referred to as Fisher's or Anderson's iris dataset, is a classic dataset in the field of machine learning and statistics. It provides comprehensive measurements of floral attributes for three species of iris flowers. The primary purpose of the Iris dataset is to facilitate the study of pattern recognition and classification techniques. By using the measurements of iris flower attributes, researchers and students can explore various machine learning algorithms for distinguishing between the three iris species based on their morphological features.

2. Glass Identification Dataset

The "Glass" dataset is a comprehensive collection of chemical analysis data from 214 observations across 10 variables. This dataset is commonly used for classification tasks where the goal is to predict the type of glass based on its chemical composition. The dataset is provided as a data frame and is available through the mlbench package in R. The primary objective of this dataset is to predict the type of glass based on its chemical composition. This predictive modeling task is essential in fields such as forensic science, where identifying the type of glass found at a crime scene can provide crucial evidence for investigative purposes. The classification of glass types in this dataset was initially motivated by criminological investigations. The ability to correctly classify and identify the type of glass found at a crime scene can significantly aid law enforcement agencies in reconstructing events and identifying potential suspects. Hence, the "Glass" dataset is a valuable resource for studying classification problems related to identifying types of glass based on chemical analysis.

3. Pima Indians Diabetes Dataset

The "Pima Indians Diabetes Database" is a well-known dataset that provides information on various health parameters of Pima Indian women, with the aim of predicting the onset of diabetes. The dataset consists of 768 observations (instances) of Pima Indian women with 500 intances with no diabetes and 268 with diabetes . There are 9 variables recorded for each observation. The main objective of this dataset is to predict whether a Pima Indian woman will develop diabetes based on her health attributes. This prediction task is crucial for early intervention and preventive healthcare strategies. The "Pima Indians Diabetes Database" is widely utilized in the field of medical research and data science for its relevance in predicting diabetes onset based on demographic, clinical, and lifestyle factors. The dataset can be accessible at https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database.

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

Advanced Journal of Science, Technology and Engineering
ISSN: 2997-5972
Volume 5, Issue 1, 2025 (pp. 70-84)
www.abjournals.org

RESULTS

Table 4.1: Result of Principal Component Analysis using the iris dataset

	PC1	PC2	PC3	PC4
Standard deviation	1.729	0.917	0.38617	0.14774
Proportion of Variance	0.747	0.2102	0.03728	0.00546
Cumulative Proportion	0.747	0.9573	0.99454	1

The result presented in table 4.1 shows the summary of the Principal Component Analysis (PCA) performed on the iris dataset. It was found that Principal Component 1 (PC1) captures the majority of the variance in the data (74.7%). This means that the largest variations in the data are along this component. Principal Component 2 (PC2) captures an additional 21.02% of the variance. Together with PC1, it explains 95.73% of the total variance, meaning the first two components capture most of the information in the dataset. Principal Component 3 (PC3) and Principal Component 4 (PC4) capture very little additional variance (3.728% and 0.546% respectively). Thus, they add little additional information beyond what is captured by the first two components. Based on this PCA, one could consider reducing the dataset from four dimensions to two dimensions (PC1 and PC2) without losing much information, as these two components capture over 95% of the total variance.

Summary Result of the LDA , QDA and SVM for the iris dataset using the missForest imputation method

Confusion Matrix and Statistics

		Reference	
Prediction	setosa	Versicolor	Virginica
setosa	*15*	0	0
Versicolor	0	13	2
Virginica	0	2	13

TABLE 4.2: Summary Result of the LDA, QDA and SVM for the iris dataset

	LDA	**QDA**	**SVM**
ACCURACY	0.9111	0.9111	0.8667
95% CL	(0.7878, 0.9752)	(0.7878, 0.9752)	(0.7321, 0.9495)
NO INFORMATION RATE	0.3333	0.3333	0.3333
P-VALUE [ACC > NIR]	$8.46e^{-16}$	$8.46e^{-16}$	$1.905e^{-13}$
KAPPA	0.8667	0.8667	0.8

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

Advanced Journal of Science, Technology and Engineering
ISSN: 2997-5972
Volume 5, Issue 1, 2025 (pp. 70-84)
www.abjournals.org

STATISTICS BY CLASS									
	Setosa	**Versicolor**	**Virginica**	**Setosa**	**Versicolor**	**Virginica**	**Setosa**	**Versicolor**	**Virginica**
SENSITIVITY	1.0000	0.8667	0.8667	1.0000	0.8667	0.8667	1.0000	0.8000	0.8000
SPECIFICITY	1.0000	0.9333	0.9333	1.0000	0.9333	0.9333	1.0000	0.9000	0.9000
POSITIVE PREDICTED	1.0000	0.8667	0.8667	1.0000	0.8667	0.8667	1.0000	0.8000	0.8000
NEGATIVE PRDICTED	1.0000	0.9333	0.9333	1.0000	0.9333	0.9333	1.0000	0.9000	0.9000
PREVALENCE	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333
DETECTION RATE	0.3333	0.2889	0.2889	0.3333	0.2889	0.2889	0.3333	0.2667	0.2667
DETECTION PREVALENCE	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333	0.3333
BALANCED ACCURACY	1.0000	0.9000	0.9000	1.0000	0.9000	0.9000	1.0000	0.8500	0.8500

The result presented in Table 4.2 revealed that the LDA model has an overall accuracy of 91.11%, indicating it correctly classifies most instances. The model perfectly classifies all instances of setosa. The LDA model was found to perform well for versicolor and virginica with some misclassifications, reflected in the slightly lower sensitivity and specificity for these classes. The Kappa value of 0.8667 indicates substantial agreement between the predicted and actual classifications, beyond what would be expected by chance. Hence, the classification model performs very well on this dataset, especially in identifying setosa instances accurately. There are minor misclassifications between versicolor and virginica, which is common given the overlap between these two species. For the QDA model, it achieved an accuracy of 91.11% just like the LDA, indicating it correctly classifies most instances. The model was found to perfectly classifies all instances of setosa. The model performs well for versicolor and virginica, with some misclassifications, reflected in the slightly lower sensitivity and specificity for these classes. The Kappa value of 0.8667 indicates substantial agreement between the predicted and actual classifications, beyond what would be expected by chance. Hence, the QDA model performs very well on this dataset with a minor misclassifications misclassifications between versicolor and virginica just like the LDA. The SVM model achieved an accuracy of 86.67%, Which performed poorly when compared with the result from the LDA and QDA. The model perfectly classifies all instances of setosa. The model performs well for versicolor and virginica, but there are some misclassifications between these two species, reflected in the slightly lower sensitivity and specificity for these classes. The Kappa value obtained was 0.8 which is slightly different from what was obtained in LDA and QDA. Hence, the SVM model performs very well on this dataset, especially in identifying setosa instances accurately.

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

Table 4.3 Result of Misclassification Error Rate for the LDA, QDA and SVM using the Iris dataset

Methods	LDA	QDA	SVM
Misclassification Error Rate	0.0889	0.0889	0.1333

The result obtained in Table 4.3 showed that both LDA and QDA models have the same misclassification error rate of 8.89%, indicating they performed equally well and better than SVM in terms of classification accuracy on the Iris dataset. The SVM model has a higher misclassification error rate of 13.33%, showing it was less accurate than both LDA and QDA. Hence, based on the misclassification error rates, both LDA and QDA outperformed SVM on the Iris dataset, achieving a lower error rate and higher accuracy.

TABLE 4.4. Result of Principal Component Analysis using the Glass dataset

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8	PC9
Standard deviation	1.5898	1.4382	1.1901	1.0566	0.9553	0.7167	0.6066	0.2546	0.1096
Proportion of Variance	0.2808	0.2298	0.1574	0.124	0.1014	0.0570	0.0408	0.0072	0.0013
Cumulative Proportion	0.2808	0.5107	0.6680	0.7921	0.8935	0.9505	0.9914	0.9986	1

The result presented in Table 4.4 showed that the first few principal components (PC1 to PC4) capture the majority of the variance in the Glass dataset (79.21%). It was found that up to PC7 captures over 99% of the variance, suggesting that the remaining components (PC8 and PC9) contribute very little additional information. This information can be used to reduce the dimensionality of the dataset effectively, retaining most of the important information while simplifying the dataset for further analysis or modeling. Hence, the choice of keeping the first 4 to 7 principal components is expected to balance between simplifying the model and retaining as much variance as possible.

Summary Result of the LDA, QDA and SVM for the Glass dataset using the missForest imputation method

Confusion Matrix and Statistics

```
                Reference
Prediction      1  2  3  5  6  7
                1 10  4  2  0  0  0
                2 11 14  3  1  1  1
                3  0  0  0  0  0  0
                5  0  3  0  0  1  0
```


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

6 0 1 0 0 0 0

7 0 0 0 2 0 7

TABLE 4.5: Summary Result of the LDA, QDA and SVM for the Glass dataset

	LDA						QDA						SVM					
ACCURACY	0.5082						0.5738						0.5246					
95% CL	(0.377, 0.6386)						(0.4406, 0.6996)						(0.3927, 0.654)					
NO INFORMATION RATE	0.3607						0.3607						0.3607					
P-VALUE [ACC > NIR]	0.01288						0.0005611						0.006438					
KAPPA	**0.3007**						0.381						0.3387					
STATISTICS BY CLASS																		
	Class1	**Class 2**	**Class 3**	**Class 5**	**Class 6**	**Class 7**	**Class 1**	**Class 2**	**Class 3**	**Class 5**	**Class 6**	**Class 7**	**Class 1**	**Class 2**	**Class 3**	**Class 5**	**Class 6**	**Class 7**
SENSITIVITY	0.4762	0.6364	0.0000	0.00000	0.0000	0.8750	0.90488	0.4091	0.0000	0.00000	0.00000	0.87500	0.6190	0.50000	0.0000	0.33333	0.0000	0.87500
SPECIFICITY	0.8500	0.5641	1.0000	0.9310	0.9831	0.9623	0.55000	0.8718	1.0000	1.00000	0.98310	0.96223	0.7500	0.74336	1.0000	0.89966	1.0000	0.94334
POSITIVE PREDICTIVE	0.6250	0.4516	NAN	0.00000	0.0000	0.7778	0.51355	0.6429	NaN	NaN	0.00000	0.77778	0.5652	0.52380	NaN	0.14290	NaN	0.70000
NEGATIVE PREDICTIVE	0.7556	0.7333	0.9180	0.9474	0.9667	0.9808	0.91667	0.7234	0.9180	0.95088	0.96667	0.98088	0.7895	0.72500	0.9180	0.96629	0.9672	0.98044
PREVALENCE	0.3443	0.3607	0.0819	0.0492	0.0328	0.1311	0.34443	0.3607	0.08197	0.04918	0.03279	0.13111	0.3443	0.36067	0.0819	0.04918	0.0328	0.13111

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

DETECTION RATE	0.1639	0.2295	0.0000	0.00000	0.0000	0.1148	0.31155	0.14755	0.0000	0.00000	0.00000	0.11488	0.2131	0.18003	0.0000	0.0064	0.0000	0.1148
DETECTION PREVALENCE	0.2623	0.5082	0.0000	0.0657	0.0164	0.1475	0.60666	0.22955	0.0000	0.00000	0.01164	0.14755	0.3770	0.34443	0.0000	0.01148	0.0000	0.16639
BALANCED ACCURACY	0.6002	0.5000	0.4656	0.49155	0.9186	0.9186	0.72744	0.64044	0.5000	0.50000	0.49150	0.90918	0.6845	0.62188	0.5000	0.61449	0.5000	0.90092

The result presented in section 4.4 showed that the overall accuracy of the LDA model is 50.82%, which is significantly better than random guessing (No Information Rate of 36.07%). The model performs well for predicting Class 7 with high sensitivity and specificity. The model struggles with Classes 3, 5, and 6, showing 0% sensitivity. The kappa statistic of 0.3007 indicates fair agreement between the predicted and actual classifications.

The result presented in section 4.5 showed that the overall accuracy of the QDA model is 57.38% in predicting the type of glass, which is better than random guessing (No Information Rate of 36.07%). The model was found to perform well for predicting Class 1 and Class 7, with high sensitivity and balanced accuracy. The model struggles with Classes 3, 5, and 6, showing 0% sensitivity. The kappa statistic of 0.381 indicates a fair agreement between the predicted and actual classifications

The result presented in section 4.6 showed that the overall accuracy of the SVM model is 52.46% for predicting the type of glass, which is significantly better than random guessing (No Information Rate of 36.07%). The model performed moderately well for predicting Class 1 and Class 7, with relatively high sensitivity and balanced accuracy. The model struggles with Classes 3 and 6, showing 0% sensitivity.The kappa statistic of 0.3387 indicates a fair agreement between the predicted and actual classifications.

Table 4.4. Result of Misclassification Error Rate for the LDA, QDA and SVM using the Glass dataset

Methods	LDA	QDA	SVM
Misclassification Error Rate	0.4918	0.4262	0.4754

The result presented in Table 4.4 shows the misclassification error rates for three different classification methods: Linear Discriminant Analysis (LDA), Quadratic Discriminant Analysis (QDA), and Support Vector Machine (SVM) when applied to the Glass dataset using the missForest imputation method. The QDA has the lowest misclassification error rate (42.62%),

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

indicating it performed best among the three methods for the Glass dataset. This suggests that the Glass dataset likely contains non-linear relationships that QDA can model effectively. LDA has the highest misclassification error rate (49.18%), implying that its linear decision boundaries were not as effective for this dataset. SVM has an intermediate error rate (47.54%), showing it performed better than LDA but worse than QDA. The choice of method should consider the complexity and nature of the dataset, with QDA being preferable here due to its better handling of non-linear relationships as indicated by its lower misclassification error rate.

Table 4.5. Result of Principal Component Analysis using the PIMA dataset

	PC1	PC2	PC3	PC4	PC5	PC6	PC7	PC8
Standard deviation	1.449	1.3188	1.0132	0.9355	0.8701	0.8285	0.6467	0.6308
Proportion of Variance	0.2625	0.2174	0.1283	0.1094	0.0946	0.0858	0.0523	0.0497
Cumulative Proportion	0.2625	0.4798	0.6082	0.7176	0.8121	0.8980	0.9503	1

The result presented in Table 4.5 showed that the first few principal components capture most of the variance in the PIMA dataset, with the first component alone explaining 26.25% of the variance. It was found that up to the sixth principal component captures nearly 90% of the total variance, suggesting that these components retain most of the information present in the original dataset. This reduction in dimensionality can be useful for visualization, noise reduction, and improving the efficiency of machine learning algorithms.

Summary Result of the LDA for the PIMA dataset using the missForest imputation method

Confusion Matrix and Statistics

		Reference
Prediction		neg pos
	neg	135 49
	pos	15 31

TABLE 4.6: Summary Result of the LDA for the PIMA dataset

	LDA	QDA	SVM
ACCURACY	0.7217	0.6957	0.7043
95% CL	(0.659, 0.7786)	(0.6318, 0.7544)	(0.6408, 0.7625)
NO INFORMATION RATE	0.0148	0.6522	0.6522
P-VALUE [ACC > NIR]	0.0148	0.0932627	0.0543270

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

KAPPA	0.3191	0.2648	0.3387
SENSITIVITY	0.9000	0.8667	0.8800
SPECIFICITY	0.3875	0.3750	0.3750
PREDICTED VALUE (+)	0.7337	0.7222	0.7253
PREDICTED VALUE(-)	0.6739	0.6000	0.6250
PREVALENCE	0.6522	0.6522	0.6522
DETECTION RATE	0.5870	0.5652	0.5739
DETECTION PREVALENCE	0.8000	0.7826	0.5739
BALANCED ACCURACY	0.6438	0.6208	0.6275

The result obtained in section 4.6 show that the LDA model has an accuracy of 72.17%, which is significantly better than the no-information rate of 65.22%. The high sensitivity (90.00%) indicates the model is very good at identifying negative cases, but the low specificity (38.75%) shows it struggles to correctly identify positive cases. The model's Kappa value of 0.3191 suggests fair agreement between the predicted and actual classes. The positive predictive value (73.37%) and negative predictive value (67.39%) indicate the model's reliability in predicting negative and positive classes, respectively. The result indicated that while the LDA model performs well in identifying negative cases, improvements are needed in correctly identifying positive cases to achieve better balanced accuracy. For the QDA model it has an accuracy of 69.57%, which is slightly higher than the no-information rate of 65.22%, but not significantly better as indicated by the p-value (0.0932627). The high sensitivity (86.67%) indicates the model is very good at identifying negative cases, but the low specificity (37.50%) shows it struggles to correctly identify positive cases. The model's Kappa value of 0.2648 suggests fair agreement between the predicted and actual classes. The positive predictive value (72.22%) and negative predictive value (60.00%) indicate the model's reliability in predicting negative and positive classes, respectively. The result underscored that while the model performs well in identifying negative cases, improvements are needed in correctly identifying positive cases to achieve better balanced accuracy. The SVM model has an accuracy of 70.43%, which is slightly higher than the no-information rate of 65.22%, but not significantly better as indicated by the p-value (0.0543270). The high sensitivity (88.00%) indicates the model is very good at identifying negative cases, but the low specificity (37.50%) shows it struggles to correctly identify positive cases. The model's Kappa value of 0.2813 suggests fair agreement between the predicted and actual classes. The positive predictive value (72.53%) and negative predictive value (62.50%) indicate the model's reliability in predicting negative and positive classes, respectively. The result implies that the SVM model performs well in identifying negative cases, but improvements are needed in correctly identifying positive cases to achieve better balanced accuracy.

Table 4.7. Result of Misclassification Error Rate for the LDA, QDA and SVM using the PIMA dataset

Methods	LDA	QDA	SVM
Misclassification Error Rate	0.2782	0.3043	0.2956

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

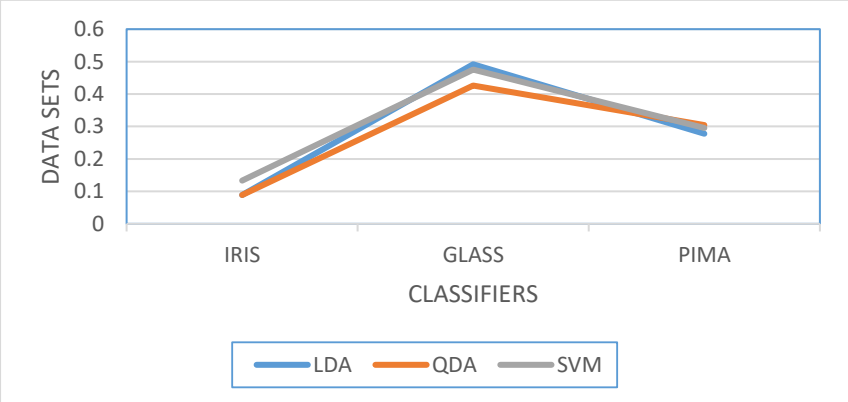
Advanced Journal of Science, Technology and Engineering
ISSN: 2997-5972
Volume 5, Issue 1, 2025 (pp. 70-84)
www.abjournals.org

The result presented in Table 4.7 showed that the LDA has the lowest misclassification error rate (27.82%), which implies that the LDA performs the best among the three methods for this dataset. The QDA was found to record the highest misclassification error rate (30.43%), QDA performs the worst. The SVM with a misclassification error rate of 29.56% shows that the SVM performs better than QDA but worse than LDA. Hence, the LDA is the most accurate method for classifying the PIMA dataset among the three methods tested, followed by SVM, and then QDA. These results suggest that LDA is better suited for this particular dataset and problem.

Table 4.8. Result of Average Misclassification Error Rate for the LDA, QDA and SVM across the various datasets considered in this study.

Dataset	LDA	QDA	SVM
Iris	0.0889	0.0889	0.1333
Glass	0.4918	0.4262	0.4754
Pima	0.2782	0.3043	0.2956
Average Misclassification Error Rate	0.2863	0.2732	0.3014

FIG.1 *Misclassification Error Rate for the LDA, QDA and SVM across the various datasets considered in this study.*



CONCLUSION

The result presented in Table 4.8 shows that on average, QDA (0.2732) has the lowest misclassification error rate across the datasets considered, followed closely by LDA (0.2863). SVM (0.3014) generally has a slightly higher average error rate compared to both LDA and QDA. QDA appears to have a slight edge in terms of average performance across these datasets, but the differences are relatively small. These error rates provide a quantitative measure of how well each method performs on average in classifying instances across different datasets. They are essential for comparing the effectiveness of different classification algorithms in various real-world applications.

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

REFERENCES

Alanazi, E. M., Abdou, A., and Luo, J. (2021). "Predicting Risk of Stroke From Lab Tests Using Machine Learning Algorithms: Development and Evaluation of Prediction Models," *JMIR Form Res*, vol. 5(12), 23440. doi: 10.2196/23440

Bickel P. and Levina, E. (2004). Some theory for Fisher's linear discriminant function, 'na¨ıve Bayes', and some alternatives when there are many more variables than observations. Bernoulli, 10(6):989–1010.

Calderbank, R., Jafarpour, S. and R. Schapire.(2009). Compressed learning: Universal sparse dimensionality reduction and learning in the measurement domain. Technical report, Rice University.

Davenport, M.A., Wakin M.B., and Baraniuk, R.G.. (2007). Detection and estimation with compressive measurements. Technical Report TREE 0610, Rice University

Fernandez, M. A., Rueda, C. and Salvador, B. (2006) "Incorporating additional information to normal linear discriminant rules,"*Journal of the American Statistical Association*, vol. 101(474) pp. 569–577

Nikita E and Nikitas, P. (2022). "Sex estimation: a comparison of techniques based on binary logistic, probability and cumulative probability regression, linear and quadratic discriminant analysis, neural networks, and naive Bayes classification using ordinal variables," International Journal of Legal Medicine, 134 (3),pp. 1213–1225.

Musa, A. B. (2012). "Comparative study on classification performance Between Support vector machine and logistic regression", International Journal of Machine Learning and Cybernetics.

Morrais C. L. and Lima, K. M. (2018)."Principle component analysis with linear and quadratic discriminant analysis for identification of cancer samples based on mass spectrometry,"Journal of the Brazilian Chemical Society, 29(3),pp. 472–481.

Pattison T. and Gossink. D. (1999.) Misclassification Probability Bounds for Multivariate Gaussian Classes.Digital Signal Processing, 9:280–296.

Sarkodie,K. and Fergusson-Rees, A. (2021)."Flow regime identification in vertical upward gas-liquid flow using an optical sensor with linear and quadratic discriminant analysis,"Journal of Fluids Engineering, 143,(2), pp. 1–12.

Sarker, I. H. (2021) "Machine Learning: Algorithms, Real-World Applications and Research Directions," *SN Comput Sci*, vol. 2(3), 160. doi: 10.1007/s42979-021-00592-x.

Yu, J., Park, S., Kwon, S.-H., Ho, C. M. B., Pyo, C.-S, and Lee, H. (2020,) "AI-Based Stroke Disease Prediction System Using Real-Time Electromyography Signals," *Applied Sciences*, 10(19). doi: 10.3390/app10196791.

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