

Distance-based robust linear discriminant analysis using mom-Qn and wmom-Qn estimators

¹Yerima, T. B., ¹Rasheed, B. A., ¹Adamu, A. and ¹Buba, C. P.

¹Department of Mathematical Sciences, Gombe State University, Gombe State, Nigeria
arasheedbello@yahoo.com, abumalala@gsu.edu.ng, bchajire@gsu.edu.ng

Paper History

Received: 20th Jan, 2026

Accepted: 31st Jan, 2026

Published: January, 2026

Abstract:

Linear Discriminant Analysis (LDA) is a cornerstone of multivariate classification but its performance significantly degrades when assumptions of multivariate normality, group covariance homogeneity are violated. However, this situation becomes worst when outliers are present in the datasets. There is a need for more reliable classification that will address the problem under such conditions. This research aims to address these limitations by adapting a distance-based robust LDA framework designed to maintain high predictive accuracy and stability. The study integrates the Modified One-Step M-estimator (MOM) and its Winsorized variant (WMOM) with the high-breakdown point Qn scale estimator. The study employed Robust Mahalanobis distance to effectively handle multivariate outliers. Performance was evaluated using simulation and a real-life dataset of clinical application distinguishing between Type 1 and Type 2 diabetes. The validation was performed using Leave One Out Cross Validation (LOOCV) and the PRESS-Q statistic. Simulation results under both balanced and unbalanced sample conditions showed that the MOM-Qn model consistently yielded the lowest mean misclassification rates, while Classical LDA recorded the highest misclassification rates. In the real-life dataset MOM-Qn estimator achieved a superior classification accuracy of 98.72% with a misclassification error of only 1.28%. These findings indicate that the adapted robust estimators are significantly less sensitive to both sample conditions, increase in dimensions, contamination, shift in location and shift in shape compare to classical. By incorporating high-breakdown estimators, the framework remains resilient even when standard statistical assumptions are violated. In conclusion, the MOM-Qn and WMOM-Qn framework provide a more stable and efficient alternative to classical LDA. This adapted approach is highly recommended for application in medical diagnostics, where ensuring reliability in the presence of data anomalies is critical for accurate patient classification.

Corresponding author
 Yerima, T. B.

tbyerima@gmail.com

Keywords: Diabetes, LDA, LOOCV, Robust estimators, Robust Mahalanobis distance

1. Introduction

Linear Discriminant Analysis (LDA) is a multivariate technique which is apt when the dependent variable is a categorical variable and the predictor variables are numerical variables. It focuses on separating distinct sets of objects into two or more groups and allocates new observations to previously defined groups. Generally, LDA is the process of constructing rules to assign a new individual observation point into one of the known populations via discriminant rules. This discriminant rules are constructed based on information (such as variables and groups) in the training data. Classification is done by allocating new observations into this discriminant rule and obtaining the group membership to which the new observation belongs [1]. A good discriminant rule is when it can provide low misclassification error rates. The Linear Discriminant Rule (LDR) performs well for data that follow normal distribution with identical population covariance

matrices. However, this rule is deemed unstable when any of these assumptions is violated [2]. This is due to the fact that the classical estimators, the mean and covariance, are known to be sensitive to deviation from the assumptions. The performance of the classical estimators can be dramatically affected if the data deviate from normality [3]. Unfortunately, ideal data set having normal distribution is hardly attainable in real life situation. To circumvent this problem, some works that are related to the robustness issues of LDA are addressed by several authors.

A number of Robust LDRs (RLDRs) which can deal with non-normality have been developed by replacing the classical mean and covariance matrix with some robust estimators of location and scatter respectively. Robust estimators such as Coordinate-wise Trimming (CT) estimators [3], M-estimators [4], S-estimators [5, 6, 7], Minimum Covariance Determinant (MCD) estimators [5, 8, 9], Minimum Volume Ellipsoid (MVE) estimators

[10], Feasible Solution Algorithm (FSA) [11] and automatic trimmed mean estimators [12] were used to alleviate the sensitivity problem of discrimination analysis rules. However, these robust estimators cannot guarantee the precision of the performance in all situations. Some estimators are good on certain conditions only but perform badly on other conditions. Winsorization is a strategy that pays more attention to the central portion of a distribution by transforming the tails of the distribution [13]. This winsorization approach was chosen based on their great performance in constructing robust Hotelling's T2 control chart [14]. While past research has successfully integrated robust estimators like MOM and WMOM with scale estimators such as Sn, Qn in a coordinate-wise manner [15,16], their potential within a distance-based LDA framework using the highly efficient Qn scale estimator remains underexplored as well as use of LOOCV [17] for

model evaluation. A distance-based approach by [18] uses Mahalanobis distance [19], in this study robust mahalanob distance is used with the robust estimators to handles multivariate outliers to offer a more holistic robustness.

2. Materials and method

2.1 Data

Secondary Data for this study is collected from the medical records department of Specialist Hospital Gombe, Nigeria. The data include key clinical variables such as age(A), body mass index (BMI), C-peptide level (CL), glucose level (GL), and insulin level (IL). The collection will focus exclusively on female patients. simulated data is also used using some specific simulation conditions as shown in Table 1.

Table 1: Simulation conditions

Variables	Description
Number of groups(K)	2
Dimensions of variables(p)	5,10,20
Sample size per group (n_1, n_2)	Balance: (50,50), (100,100), (200,200) Unbalance: (80,100), (145,150), (180,150)
Percentage contamination (ε)	(0.5,20)%
Shift in location of the population (μ)	0,2,5
Shift in shape of the population (κ)	1,3,5
Data distribution	Multivariate normal

2.2 Classical linear discriminant rule (CLDR)

This classical discriminant rule is built to be optimal in classifying the new observation x_0 under the assumptions that T_1 and T_2 are both multivariate normal distributions with different location but having identical covariance matrix [20]. In particular, T_1 and T_2 are $Nd(\mu_1, \Sigma_1)$ and $Nd(\mu_2, \Sigma_2)$ respectively and under the assumption $\Sigma_1 = \Sigma_2 = \Sigma$. If there are outliers in the training data, then the estimators of mean (μ) and covariance (Σ) can be affected. To alleviate this problem, a robust estimator will be adapted by replacing the classical mean and covariance with robust estimators. Equation 1 is the general form of LDA, we got equation 2 by substituting the five variables used in this study

$$D = b_0 + b_1x_1 + b_2x_2 + \dots + b_gx_g \quad (1)$$

Where D is the discriminant score, b_0 is the constant term, $b_1, b_2, \dots, b_g$ are the discriminant coefficients or weights for each predictor variables $x_1, x_2, \dots, x_n$

$$D = b_0 + b_1x_A + b_2x_{BMI} + b_3x_{CL} + b_4x_{GL} + b_5x_{IL} \quad (2)$$

The assignment of an observation to one of the groups can be based on the discriminant score D . Allocate X_0 to T_1 if equation 3 is satisfied

$$(\mu_{T_1} - \mu_{T_2})^T \Sigma_{pooled}^{-1} \left\{ X_0 - \frac{1}{2}(\mu_{T_1} + \mu_{T_2}) \right\} \geq \ln \left(\frac{P_2}{P_1} \right) \quad (3)$$

Allocate X_0 to T_2 otherwise the costs of misclassification are assumed identical in this study. Where P_1 and P_2 are the prior probabilities that an observation

comes from population T_1 and T_2 respectively and is equal for the real-life data both group has equal size

2.3 Modified one-step m-estimator with Qn (MOM-Qn)

The algorithm to combine these two techniques is described using equations 4 to 8

Step 1: Initial Trimming Compute the median $\hat{M}_j$ and MAD_n , use them to identify and trim outliers from each group.

$$\hat{\theta}_j = \frac{\sum_{i=i_1+1}^{n_j-i_2} Y_{(i)j}}{n_j - i_1 - i_2} \quad (4)$$

$$\hat{\Sigma}^{(k)} = \rho_{gh}^{(k)} \times Q_{n,g}^{(k)} \times Q_{n,h}^{(k)} \quad (5)$$

$$Q_{n,g}^{(k)} = d\{ |x_{ig}^{(k)} - x_{jg}^{(k)}|; i < j \}_{(q)}$$

$$q = \binom{s}{2}, s = \left\lfloor \frac{n}{2} \right\rfloor + 1 \quad (6)$$

$$D_R^2(X_i, K) = (X_i - \theta^{(k)})^T (\Sigma^{(k)})^{-1} (X_i - \theta^{(k)}) \quad (7)$$

$$D_i^2 \text{ or } D_R^2(X_i, K) > \chi_{p,\alpha}^2 \quad (8)$$

Step 2: Compute $\hat{\theta}_j$ MOM Estimate (Mean of Trimmed Data) using equation 4

Step 3: Compute Q_n Scale Estimator (Robust Spread) from equation 6

Step 4: Replace the default scale estimator MAD_n in step 2 with the Q_n estimator to obtain i_1 as the number of

observations Y_{ij} such that $\hat{\theta}_j - 2.24(Q_{nj})$ and i_2 is the number of observations $\hat{\theta}_j + 2.24(Q_{nj})$

Step 5: Recompute MOM $\hat{\theta}_j$ with Q_n Based Trimming

Step 6: compute robust covariance using equation 5

Step 7: use equation 7 to detect multivariate

Any observation that satisfies equation 8 will be flagged as multivariate outlier [11]

2.4 Winsorized modified one-step M-Estimator with Q_n (WMOM- Q_n)

Given below is the general approach of the WMOM algorithm and also insight is given on how to merge this estimator with the Q_n scale estimator using equations 5 to 10.

Step 1: Replace outliers identified in MOM by winsorizing them i.e replacing them with the next non-outlying value.

Step 2: compute the winsorized MOM for j -th variable (i.e the mean of the winsorised data) and the winsorized covariance matrix between w_i and w_j variables respectively represented by equation 9:

$$\bar{W}_j = \frac{1}{m_j} \quad (9)$$

and the vector of winsorized MOM estimator is obtained from equation 10:

$$\bar{W} = \begin{bmatrix} \bar{W}_1 \\ \vdots \\ \bar{W}_p \end{bmatrix} \text{ and}$$

$$S_{WMAD_n}(W_i, W_j) = \frac{1}{m-1} [\sum_{k=1}^m W_{ki} W_{kj} - m \bar{W}_i \bar{W}_j] \quad (10)$$

Step 3: inverse the covariance matrix from step 2, thus the inverse of the matrix winsorised covariance is $S_{WMAD_n}^{-1}$ computed using MAD the default scale estimator for the criterion on trimming.

Step 4: use equation 7 to detect multivariate outliers by substituting winsorized mean and inverse of the winsorized covariance matrix into the equation

Step 5: Replace the default scale estimator MAD_n in Step 3 with the Q_n estimator in equation 6.

Step 6: compute robust covariance using equation (5)

Step 7: use equation (7) to detect multivariate outliers

Any observation that satisfies equation (8) will be flagged as multivariate outlier

2.5 Model evaluation

To robustly assess the performance of the proposed robust model LOOCV and Press-Q founded by [21] was employed.

$$LOOCV \text{ Error} = \frac{1}{n} \sum_{i=1}^n L(y_i \neq \hat{y}_{-i}) \quad (11)$$

$\hat{y}_{-i}$ is the predicted class for observation using the model without observation i , y_i is the true class label, L is the loss function indicator it assigns a loss of 0 if the prediction is correct and assigns a loss of 1 otherwise

Press-Q Statistic

$$Q = \frac{[N - (n_{correct})K]^2}{N(K-1)} \quad (12)$$

If $Q > \chi^2_{\alpha,1}$ it can be concluded that the classification is accurate.

3. Results

3.1 Simulation results

Figures 1-5 shows the results for the different simulation conditions as stated in Table 1:

From Figure 1 the MOM- Q_n method is the most accurate classifier, consistently exhibiting the lowest misclassification rate across nearly all dimensions and sample conditions, In Figure 2 the classical model is clearly the least classifier, showing a sharp increase in misclassification error as the contamination level rises. Both the MOM- Q_n and WMOM- Q_n models exhibit strong robustness to outliers with the former outperforming the later, in Figure 3.

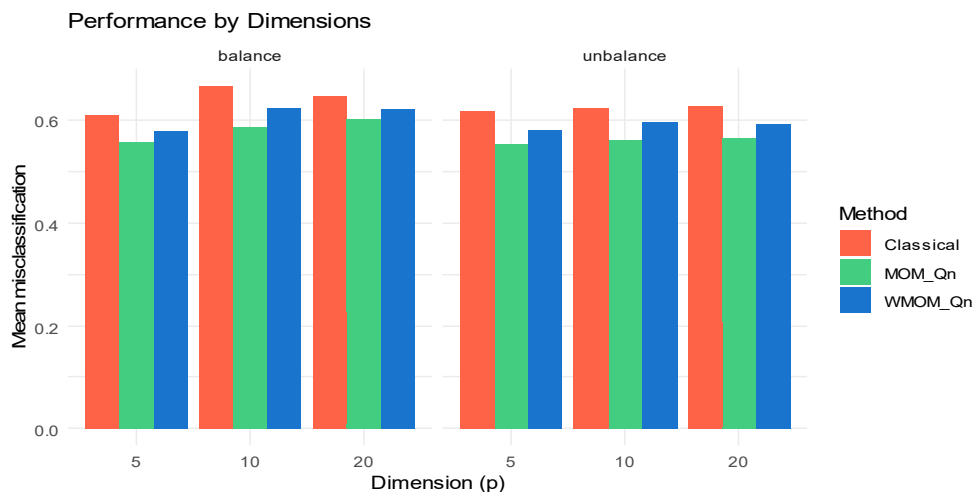


Figure 1: Performance of the three models by change in Dimensions (P)

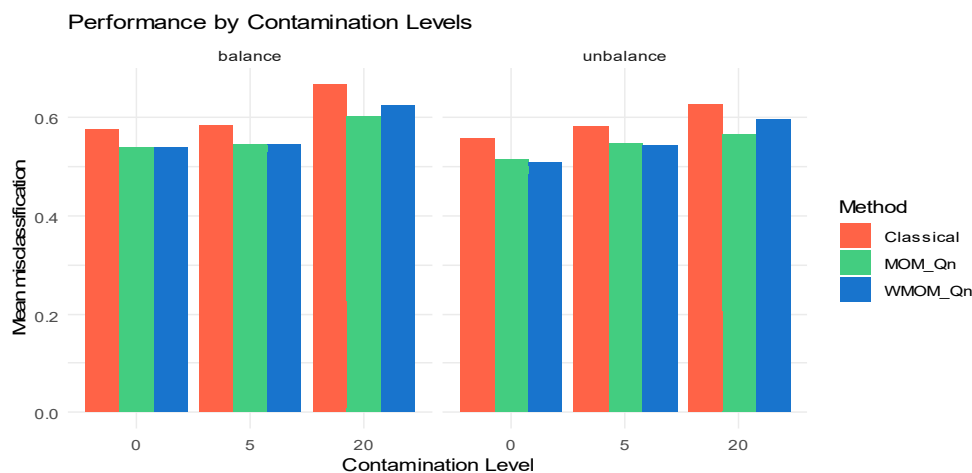


Figure 2: Performance of the three models by change in contamination levels

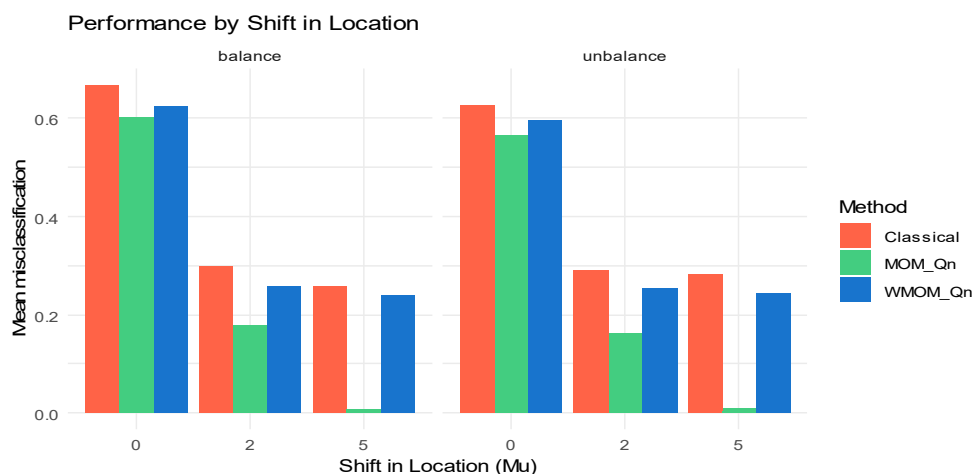


Figure 3: Performance of the three models by shift in location



Figure 4: performance of the three models by shift in shape

The MOM_Qn model is the most effective, achieving near-perfect classification (error approaching zero) at the $\mu=5$ location shift, The Classical model is the least effective across all shift conditions. In Figure 4 all models experience a general degradation in performance as the shape shift

increases; The MOM_Qn model proves to be the most accurate and relatively robust method, maintaining the lowest mean misclassification error across all levels of shape shift and data condition.

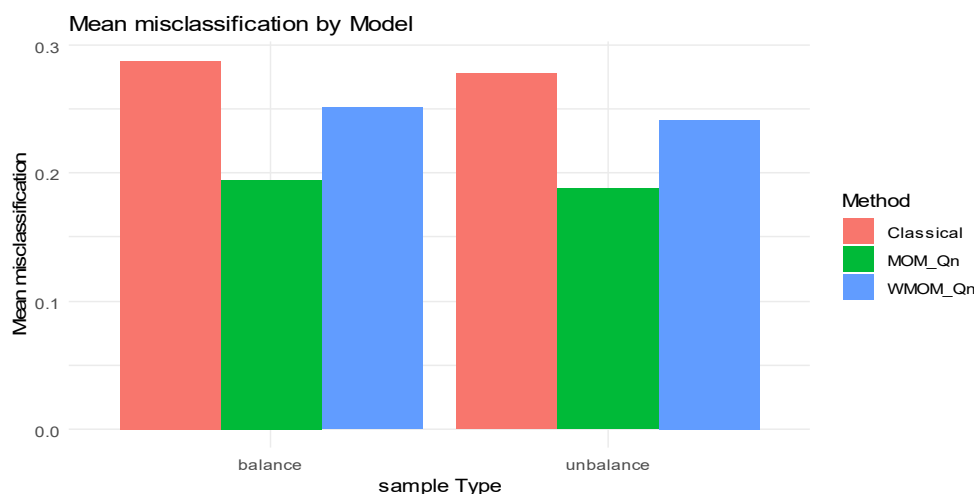


Figure 5: Mean Misclassification by model

Figure 5 shows that for both balanced and unbalanced samples the MOM_Qn model consistently records the lowest mean misclassification rate, followed by the WMOM_Qn model, while the Classical model shows the highest misclassification rate, making it the least accurate across both sample types.

Table 2: Summary of mean misclassification rates for the three models

Sample Type	Classical	MOM_Qn	WMOM_Qn
Balance	0.288	0.194	0.251
Unbalance	0.278	0.188	0.242

Table 2 shows that both in balanced and unbalanced sample the MOM_Qn model has the lowest mean misclassification rate, followed by WMOM_Qn, while the Classical model has the highest. This indicates that MOM_Qn provides the most accurate and robust classification across both sample types, with a slight improvement in performance under the unbalanced condition.

3.2 Real data results

3.2.1 Classical L.D.A model

Equation 13 is the classical LDA model that produces the confusion matrix in Table 3

$$D(x)_{CLASSICAL} = 1.110223e^{-1} + 2.0339_{AGE} + 0.3012_{BMI} + 0.0446_{IL} - 0.3355_{GL} + 0.6663_{CL} \quad (13)$$

Table 3 shows that the classical model exhibits a good performance achieving a very high accuracy of 97.8% and a low misclassification error of only 2.2%. The model's classification ability is statistically better than random chance, as confirmed by the Press-Q Statistic (456.968) significantly exceeding the critical value (3.841).

Table 3: Confusion matrix

Actual/predicted	Type 1	Type 2
Type 1	240	10
Type 2	1	249

3.2.2 MOM-Qn L.D.A model

Equation 14 is the MOM-Qn LDA model that produces the confusion matrix in Table 4

$$D(x)_{MOM-Qn} = 0.0984 + 2.3627_{AGE} + 1.4718_{BMI} - 0.1708_{IL} - 0.1934_{GL} + 0.3287_{CL} \quad (14)$$

Table 4: Confusion matrix

Actual/predicted	Type 1	Type 2
Type 1	234	6
Type 2	0	230

From Table 4, the MOM_Qn LDA Model demonstrates outstanding classification performance, achieving an exceptional accuracy of 98.72% and a remarkably low misclassification error of 1.28%. The model's classification ability is statistically better than random chance, as confirmed by the Press-Q Statistic (446.31) significantly exceeding the critical value (3.841).

3.2.3 WMOM-Qn MODEL

Equation 15 is the WMOM-Qn LDA model that produces the confusion matrix in Table 5

$$D(x)_{WM-Qn} = -0.01730959 + 2.017604000_{AGE} + 0.356986300_{BMI} + 0.002390079_{IL} - 0.448733175_{GL} + 0.629155332_{CL} \quad (5)$$

Table 5: Confusion matrix

Actual/predicted	Type 1	Type 2
Type 1	241	9
Type 2	1	249

The confusion matrix in table 5 indicates a high accuracy of 98% and a misclassification rate of 2%, showing excellent model performance, also since the PRESS-Q statistic (460.8) is greater than the critical value (3.841).

4. Conclusion

This study concludes that MOM-Qn and WMOM-Qn based robust LDA methods significantly outperform the Classical LDA in terms of classification accuracy, robustness and stability. The MOM-Qn estimator in particular demonstrated superior performance across both simulated and real datasets. The robust methods proved capable of maintaining reliable classification even when the assumptions of normality and covariance homogeneity were not satisfied. Therefore, the MOM-Qn-based robust LDA can be recommended as a more dependable alternative to the Classical LDA. The robust discriminant approaches successfully address key limitations of classical LDA, providing a more stable and accurate framework for classification tasks involving non-ideal or contaminated datasets.

References

- [1]. Maharaj, E. A. and Alonso, A. M., (2012). Discriminant analysis of multivariate time series using wavelets, *Statistics and Economics Series* (03), Working Papers 12-06.
- [2]. Croux, C. Filzmoser, P. and Joossen, K., (2008). Classification efficiencies for robust linear discriminant analysis, *Statistica Sinica*, 18(2), 581-599.
- [3]. Sajobi, L. M., Lix, T. T., Dansu, B. M., Laverty, W. and Li, L., (2012). Robust descriptive discriminant analysis for repeated measures data, *Computational Statistic and Data Analysis*, 56(9), 2782-2794.
- [4]. Wang, D. and Romagnoli, J. A., (2005). Robust multi-scale principal components analysis with applications to process monitoring, *Journal of Process Control*, 15(8), 869-882.
- [5]. Lim, Y. F., Yahaya, S. S. S., Idris, F., Ali, H. and Omar, Z., (2014). *Robust Linear Discriminant Models to solve Financial Crisis in Banking Sector*, in Proceedings of the 3rd International Conference on Quantitative Sciences and Its Applications, AIP Conference Proceedings 1635, Kedah, 2014, pp: 794-798.
- [6]. Croux, C. and Dehon, C., (2001). Robust linear discriminant analysis using S-estimators. *Canadian journal of statistics*, 29(3), 473-493.
- [7]. He, X. and Fung, W. K., (2000). High breakdown estimation for multiple populations with applications to discriminant analysis. *Journal of Multivariate Analysis*, 72(2), 151-162.
- [8]. Alrawashdeh, M. J., Muhammad Sabri S. R. and Ismail, M. T., (2012). Robust linear discriminant analysis with financial ratios in special interval, *Applied Mathematical Sciences*, 6(121) 6021-6034.
- [9]. Hubert, M. and Driessen, K., (2004). Fast and robust discriminant analysis. *Computational Statistics and Data Analysis*, 45, 301-320.
- [10]. Chorl, C. Y. and Rousseeuw, P. J., (1992). Integrating a high-breakdown option into discriminant analysis in exploration geochemistry, *Journal of Geochemical Exploration*, 43(3), 191-203.
- [11]. Wina, D., Herwindiati, E. and Isa, S. M., (2014). *Robust discriminant analysis for classification of remote sensing data*, International Conference on Advanced Computer Science and Information Systems, IEEE, Jakarta, 2014, pp: 454-458.
- [12]. Syed Yahaya, S. S., Lim, Y. F., Ali, H. and Omar, Z., (2016). Robust linear discriminant analysis with automatic trimmed mean, *Journal of Telecommunication, Electronic and Computer Engineering*, 8(10), 1-3.
- [13]. Haddad, F. S., Syed-Yahaya S. S. and Alfaro, J. L., (2013). Alternative Hotelling's T² charts using winsorized modified one-step M estimator, *Quality and Reliability Engineering International*, 29, 583-593.
- [14]. Haddad, F. S. Alfaro J. L. and Alsmadi, M. K., (2015). Hotelling's T² charts using winsorized modified one-step M-estimator for individual non normal data, *journal of theoretical and applied information technology*, 72(2), 215-226.
- [15]. Melik, H. N. (2017). *Robust linear discriminant analysis using MOM-Qn and WMOM-Qn estimators: Coordinate – wise approach*. Master's thesis, Universiti Utara Malaysia, UUM Repository. <http://etd.uum.edu.my/6810/>.
- [16]. Tjahaya, M. S., Raupong, G. M. and Tinungki, G. M., (2022). Analisis diskriminan linear robust dengan metode Winsorized Modified One-Step M-Estimator (WMOM) menggunakan skala Sn pada data diabetes dan pradiabetes periode Desember 2016–Januari 2017. *Estimasi: Journal of Statistics and Its Application*, 3(1), 1-13
- [17]. Lumumba, V. W., Kiprotich, D., Lemasulani Mpaine, M., Grace Makena, N., and Daniel Kavita, M. (2024). Comparative analysis of Cross-Validation techniques: LOOCV, K-folds Cross-Validation, and repeated K-folds Cross-Validation in machine learning models. *K-folds Cross-Validation, and Repeated K-folds Cross-Validation in Machine Learning Models* (June 01, 2024).
- [18]. Lim, Y. F., (2021). *Robust linear discriminant rules (RLDRs) using coordinate-wise and distance-based approaches*, [Doctoral dissertation, Universiti Utara Malaysia]. UUMRepository. <https://etd.uum.edu.my/8799/>
- [19]. Mahalanobis, P. C., (1936). On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*, 2, 49-55.
- [20]. Croux, C., Filzmoser, P. and Joossens, K.(2008). Classification efficiencies for robust linear discriminant analysis, *Statistica Sinica*, 18, 581-599.
- [21]. Press, S. J., (1972). *Applied multivariate analysis*. Holt, Rinehart and Winston.