Jurnal Matematika dan Statistika serta Aplikasinya, Vol. 13, No. 2, 212-219, 2025
DOI:https://doi.org/10.24252/msa.v13i2.59043

Comparison of Linear and Robust Discriminant Analysis Methods in the Classification of Malignant and Benign Breast Cancer

Erina Laila Sulaiman[1], Ria Amanda[2], Nur Rahma Wulandari[3], Mawaddah[4], Khalilah Nurfadilah[5]*
[1,2,3,4,5]*Mathematics Study Program, Universitas Islam Negeri Alauddin Makassar, Indonesia*
***Corresponding author**: khalilah@uin-alauddin.ac.id

***Submission date**: 04 July 2025, **Revision**: 02 August 2025, **Accepted**: 22 December 2025

ABSTRACT
Breast cancer is one of the most common types of cancer in women worldwide. According to data from the World Health Organization (WHO), breast cancer accounts for about 25% of all cancer cases in women. Early diagnosis has a very important role in determining the patient's survival rate. The purpose of this study is to determine which method is more effective in classifying breast cancer. The methods used in this study are using linear discriminant analysis and robust discriminant. The results showed that the proportion of errors using linear discriminant analysis was 3, 34% while the proportion of errors using robust discriminant analysis was 12, 3% so it can be concluded that the linear discriminant analysis method is more effective in classifying malignant and benign breast cancer.

KEYWORDS

1. INTRODUCTION

Comparing the effectiveness of linear discriminant analysis (LDA) and robust linear discriminant analysis (RLDA) methods in the classification of malignant and benign breast cancer is a crucial research focus [1]. Recent studies have shown that these two methods give different results depending on the characteristics of the data and the number of outliers in the dataset, so the selection of the right method largely determines the accuracy of the diagnosis which can affect medical decisions and patient safety [2]. Linear discriminant analysis is a classical statistical technique that has long been used in medical classification, but it is sensitive to the presence of outliers and violations of the assumption of multivariate normality. On the other hand, robust linear discriminant analysis was developed to overcome these weaknesses by using robust estimators that can handle data containing outliers [3]. In the context of breast cancer classification, tumors in breast cancer are divided into two, namely benign or commonly called benign and malignant or commonly called malignant [4], so classification accuracy becomes very important to determine the right medical action.

Various studies in Indonesia have explored the use of machine learning for breast cancer classification, with data mining there are many concepts and models, one of which is the concept of Decision Tree [5]. A systematic comparison between LDA and RLDA still requires further research to provide practical recommendations for medical practitioners. Challenges in breast cancer classification are increasingly complex because breast cancer is one of the global health challenges that results in high mortality rates, where in 2020, in Indonesia alone there are more than 22 thousand death cases [6].

Medical datasets often contain noise, missing values, and outliers that can affect the performance of conventional classification algorithms. The development of machine learning and statistical learning technology in the field of medicine has opened up great opportunities to improve the accuracy of disease diagnosis. Discriminant analysis as one of the oldest classification

techniques in statistics continues to evolve, from classical methods to robust approaches that are more resistant to data noise. In the era of big data and precision medicine, an in-depth understanding of the advantages and limitations of each method is fundamental to the development of a reliable, efficient and dependable diagnosis system to support better clinical decision-making in breast cancer treatment

2. LITERATURE REVIEW

2.1 Multivariate Analysis

Multivariate analysis is a statistical approach used to analyze data involving multiple variables simultaneously in order to understand the relationships and structural interdependencies among variables. This method is an extension of univariate analysis, which focuses on a single variable, thereby enabling multivariate analysis to provide a more comprehensive understanding of complex phenomena. Multivariate analysis is particularly relevant when variables in the dataset are correlated and cannot be analyzed independently [1].

Multivariate analysis encompasses various methods, including multivariate regression analysis, factor analysis, discriminant analysis, cluster analysis [7], biplot analysis, and multidimensional scaling (MDS). These methods are employed for classification, dimensionality reduction, and modeling relationships among variables in complex datasets. Therefore, multivariate analysis serves as an important theoretical foundation in this study, particularly in the application of discriminant analysis as a classification method that utilizes multiple variables simultaneously.

2.2 Discriminant Analysis

Discriminant analysis is one of the methods in multivariate analysis that aims to classify objects into predefined groups based on a set of predictor variables. This method seeks linear combinations of independent variables that maximize the separation between groups. Discriminant analysis is widely used in classification problems such as disease identification, pattern recognition, and decision analysis. One of the most commonly used forms of discriminant analysis is Linear Discriminant Analysis (LDA), which assumes that the data follow a multivariate normal distribution and that the variance–covariance matrices are equal across groups [8]. However, in practice, data often contain outliers or violate these assumptions, leading to the development of robust linear discriminant analysis approaches.

2.3 Cancer

Cancer is a disease that occurs when cells in the body grow uncontrollably and can spread to other parts of the body. Cancer cells can damage surrounding normal tissue and disrupt organ function. Cases of cancer deaths in Indonesia continue to show an increase every year. Based on the data, cancer ranks third as the most common cause of death after heart disease and stroke. This high mortality rate is related to the low recovery rate of cancer patients, which is generally caused by delays in detecting this disease, so that it is only discovered when it has entered an advanced stage and is difficult to treat effectively. Some of the factors that exacerbate this condition include the lack of fruit and vegetable consumption, excessive smoking and alcohol consumption, lack of physical activity, high body mass index, and the limitations of relatively expensive treatment costs. In low- and middle-income countries, viral infections such as hepatitis B, hepatitis C and the legal papillomavirus (HPV) are also significant contributors to cancer deaths. In this context, cancer is closely related to tumors, which are abnormal tissue growths in the body. Tumors themselves are divided into two, namely benign and malignant. Benign tumors usually do not spread and are considered harmless, while malignant tumors are aggressive, able to spread to other body tissues, and damage surrounding healthy cells [9].

2.4 Breast Cancer

Breast cancer is a malignant disease that can pose a high risk to women's health. In recent years, breast cancer cases have been increasing every year, so early detection efforts are needed to be more accurate and effective. One way to support the early detection process is through the application of classification methods based on cancer cell characteristic data.

Cell characteristic data can be obtained from microscopic laboratory examination results which are then processed into numerical variables. These variables represent the shape, size, texture, and surface structure of cells [10]. Breast cancer is one of the leading causes of cancer deaths in women, after lung cancer, especially in the United States. About 25% of the total

cancer cases experienced by women are breast cancer, making it the most common type of cancer globally. The incidence of this cancer is much higher in developed countries and cases are recorded to affect women up to 100 more often than men [11].

2.5 WBCD Dataset

The Wisconsin Breast Cancer Dataset (WBCD) is one of the benchmark datasets in breast cancer classification analysis. This dataset provides various numerical variables obtained from the extraction of cell tissue microscopy images, which represent the morphology of cancer cells. There are ten main features analyzed, namely *radius_mean* (average distance from the center to the edge of the cell), *texture_mean* (variability of gray level), *perimeter_mean* (length of cell perimeter), *area_mean* (size of cell area), *smoothness_mean* (slipperiness of cell contour), *compactness_mean* (density of cell shape), *concavity_mean* (depth of indentation), *concave_points_mean* (number of concave points), *symmetry_mean* (degree of symmetry), and *fractal_dimension_mean* (complexity of cell structure). The transformation of these image features into numerical form is essential for enabling biological data to be analyzed using statistical approaches or classification algorithms. Each of these variables contributes to distinguishing between benign and malignant cells, given the striking differences in their morphological characteristics. Therefore, these features are highly relevant for use in an automated and more accurate breast cancer classification process.

2.6 Outlier

Outliers are data that are statistically distant from the center of the data or the general pattern. In multivariate analysis, outliers are commonly recognized using the Mahalanobi distance. However, this method is less effective in identifying outliers that mask

Table 1. Box's M Test for Homogeneity of Covariance Matrices

Test Statistic	Chi-Square	Degrees of Freedom	p-value
Box's M Test	1649.3	55	$< 2.2 \times 10^{-16}$

each other (masking effect). Therefore, a robust distance approach with Minimum Covariance Determinant (MCD) estimation is used as an alternative to mean and covariance estimation.

A data is considered as an outlier if it meets the mismatch with the data center based on the following robust distance formula:

$$(x_i - \mu_{MCD})' \sum_{MCD}^{-1} (x_i - \mu_{MCD}) > x^2_{p,\alpha} \tag{1}$$

Description:
x_i is the i-th observation vector with dimension $p \times 1$
μ_{MCD} is the robust average of the dimensionless MCD estimation result $p \times 1$
MCD^{-1} is the inverse of the dimensionless robust covariance matrix $p \times p$
$x^2_{p,\alpha}$ is the quantile of the chi-square distribution at a given significance level and degree of freedom p.

2.7 Multivariate Normality Test

The multivariate normality test aims to ascertain whether the data follows a multivariate normal distribution pattern. One of the methods used is the Q-Q Plot, which compares the actual data with the theoretical distribution. Data is said to be close to normal distribution if the points in the Q-Q Plot mostly form a straightline pattern (more than 50%).

2.8 Variance-Covariance Matrix Equality Test

To test whether several groups of data have the same variance-covariance structure, Box's M Test method is used [12]. In this test, an approach is taken through the chi-square (χ^2) or F distribution based on the M test statistic.
Hipotesis:
H_0: Between-group covariance matrix (based on diagnosis variable) $(\Sigma_1 = \Sigma_2 = \cdots = \Sigma_k)$
H_1: There is at least a difference in the covariance matrix between groups $(\Sigma_1 \neq \Sigma_2 \neq \cdots \neq \Sigma_k)$

Based on the test results in **Table 1**, the test statistic value Chi-Square estimate of 1649.3 degrees of freedom (df) of 55, p-value is $<$ 2.2e-16 (very small), Since p-value < 0.05, reject H_0. This means that there is a significant difference in the variance-covariance matrix structure between diagnosis groups in the data. In other words, the data from each diagnosis group does not have the same distribution structure (variance and covariance), so you cannot assume homogeneity of covariance, which is especially important for methods like LDA, which assumes similarity of the covariance matrix between classes.

2.9 Linear Discriminant Analysis

The main objective of linear discriminant analysis is to find a linear combination of independent variables that maximally discriminates between groups. This method produces a discriminant function that can be written as:

$$\bar{y} = (\bar{x}_1 - \bar{x}_2)^T S_{pooled}^{-1} x = a^T x \tag{2}$$

where:
$\bar{y}$: the value of the linear discriminant function
a : vector of coefficients
x : vector of independent variables

After obtaining the discriminant function, the next step is to determine the classification of an object into one of the groups. This is done by comparing the object's discriminant value against the center discriminant value of each group. The center discriminant value can be calculated using the following formula:

$$\bar{m} = \frac{1}{2}(\bar{x}_1 - \bar{x}_2)^T S_{pooled}^{-1}(\bar{x}_1 - \bar{x}_2) = \frac{1}{2}(y_1 - y_2) \tag{3}$$

where:
$y_1 = (\bar{x}_1 - \bar{x}_2)^T S_{pooled}^{-1} \bar{x}_1$
$y_2 = (\bar{x}_1 - \bar{x}_2)^T S_{pooled}^{-1} \bar{x}_2$

As a basis for classification, if $\bar{y} \geq \bar{m}$ then the object is classified to group 1, and if $\bar{y} \leq \bar{m}$, then the object is classified to group 2.

2.10 Robust Linear Discriminant Analysis

The robust linear discriminant approach is a development of the conventional linear discriminant method, where the calculation process is performed using the robust Minimum Covariance Determinant (MCD) estimation. In this case, the mean value μ_k and covariance matrix Σ in the discriminant function are replaced by the robust mean $\mu_{MCD,k}$ and the robust covariance matrix Σ_{MCD}. The robust linear discriminant function for each group k can be formulated as follows:

$$y_{MCD,k}(x) = x^T \sum_{MCD}^{-1} \mu_{MCD,k} - \frac{1}{2}\mu_{MCD,k}^T \sum_{MCD}^{-1} \mu_{MCD,k} + \ln(p_k) \tag{4}$$

where:
x is the vector of observations,
$\mu_{MCD,k}$ is the robust mean of the kth group,
Σ_{MCD} is the robust covariance matrix shared across all groups (assumed homogeneous).
p_k is the prior probability for each group k=1,2.

The classification step is performed by selecting the largest robust discriminant function value. An object will be classified into group p_k if it satisfies:

$$y_{MCD,k}(x) = max\{y_{MCD,k}(x) : k = 1,2\} \tag{5}$$

Table 2. Confusion Matrix

	Predicted $\boldsymbol{\pi}_1$	Predicted $\boldsymbol{\pi}_2$	Total
Actual $\boldsymbol{\pi}_1$	n_{11} (correct)	n_{12} (incorrect)	$n_1 = n_{11} + n_{12}$
Actual $\boldsymbol{\pi}_2$	n_{21} (incorrect)	n_{22} (correct)	$n_2 = n_{21} + n_{22}$
Total			$n_1 + n_2$

2.11 Apparent Error Rate (APER)

Apparent Error Rate (APER) is a measure that shows the percentage of misclassification against training data. In other words, APER calculates how much data is misclassified by the discriminant model. The APER value is obtained from the confusion matrix, which is a table that compares the actual membership with the predicted membership [12].

Description:

n_{11}: number of observations in group π_1 that are correctly classified into π_1,
n_{12}: number of observations in group π_1 that are incorrectly classified into π_2,
n_{21}: number of observations in group π_2 that are incorrectly classified into π_1,
n_{22}: number of observations in group π_2 that are correctly classified into π_2,
n_1: total number of original observations in group π_1,
n_2: total number of original observations in group π_2.
Formula:

$$APER = \frac{n_{12} + n_{21}}{n_1 + n_2} \tag{6}$$

That is, APER is the proportion of total misclassification compared to the total amount of data. For example, out of 100 students, 60 are from group A and 40 from group B. If the model misplaces 5 students from A to B and 3 from B to A, then:

$$APER = \frac{5+3}{60+40} = \frac{8}{100} = 0.08$$

This indicates that 8% of the data were misclassified by the model.

3. METHODOLOGY

The data used is secondary data derived from the UCI website dataset of Breast Cancer Wisconsin (Diagnostic) Data Set. **Table 3** and **4** shows the research variables and operational definition of the variables.

Table 3. Data Type of Research Variables

No	Variable Name	Data Type
1	*radius_mean* (X_1)	Numerical
2	*texture_mean* (X_2)	Numerical
3	*perimeter_mean* (X_3)	Numerical
4	*area_mean* (X_4)	Numerical
5	*smoothness_mean* (X_5)	Numerical
6	*compactness_mean* (X_6)	Numerical
7	*concavity_mean* (X_7)	Numerical
8	*concave_points_mean* (X_8)	Numerical
9	*symmetry_mean* (X_9)	Numerical
10	*fractal_dimension_mean* (X_{10})	Numerical
11	*diagnosis* (Y)	Categorical (M = malignant, B = benign)

Table 4. Research Variables and Operational Definition of Variables

No	Variable	Operational Definition
1	*radius_mean* (X_1)	Average distance from the center to the cell perimeter
2	*texture_mean* (X_2)	Standard deviation of gray-scale intensity values
3	*perimeter_mean* (X_3)	Average cell perimeter length
4	*area_mean* (X_4)	Average cell area size
5	*smoothness_mean* (X_5)	Local variation in cell radius length
6	*compactness_mean* (X_6)	$(\text{perimeter}^2/\text{area}) - 1.0$
7	*concavity_mean* (X_7)	Average depth of concave portions of the cell contour
8	*concave_points_mean* (X_8)	Number of concave points in the cell contour
9	*symmetry_mean* (X_9)	Degree of cell symmetry
10	*fractal_dimension_mean* (X_{10})	Complexity of the cell's fractal structure
11	*diagnosis* (Y)	Classification target (M = malignant, B = benign)

3.1 Analysis Procedure

The steps of linear discriminant analysis of breast cancer patients are described as follows:

1. Create descriptive statistics on the research data clusters.
2. Performing data standardization on the cluster
3. Randomly divide the data into two subsets, namely 80% for training data (training set) and the remaining 20% for testing data (testing set) in the research data cluster.
4. Identifying outliers in the research data clusters using robust distance MCD.
5. Test the assumptions of discriminant analysis
6. Analyzing data using linear discriminant analysis method.
7. Analyzing data using robust discriminant analysis method.
8. Comparing the results of linear discriminant analysis with robust discriminant analysis.

4. RESULT & DISCUSSION

Before applying discriminant analysis, a preliminary assumption test was conducted. Based on the Q-Q Plot of Mahalanobis Distance, it can be seen that the data deviates from the normal line, indicating a violation of the assumption of multivariate normality. The number of outliers is 38 or 6.68%. In addition, the Box's M Test results show a value of the p-value is 2.2e-16 ($p < 0.05$), which means that the assumption of homogeneity of the covariance matrix is not met. Therefore, in addition to the Linear discriminant analysis (LDA) approach, the Robust Discriminant Analysis method is considered so that the classification results remain valid despite the assumption violation.

The estimation model of linear discriminant analysis is as follows:

$$\begin{aligned}\hat{y} = {} & 0.598\,(radius\ mean) - 0.004\,(texture\ mean) + 2.746\,(perimeter\ mean) - 16.252\,(area\ mean) \\ & + 0.161\,(smoothness\ mean) - 13.409\,(compactness\ mean) + 0.065\,(concave\ points\ mean) \\ & + 14.831\,(symmetry\ mean) + 12.106(fractal\ dimension\ mean)\end{aligned}$$

whereas, the estimation model of robust discriminant analysis is obtained as follows:

$$\begin{aligned}\hat{y} = {} & 446.371\,(radius\ mean) - 5.539\,(texture\ mean) - 282.093\,(perimeter\ mean) + 5667.2\,(area\ mean) \\ & + 6.502\,(smoothness\ mean) - 624.03\,(compactness\ mean) + 1.45\,(concave\ points\ mean) \\ & - 57.337\,(symmetry\ mean) - 501,3(fractal\ dimension\ mean)\end{aligned}$$

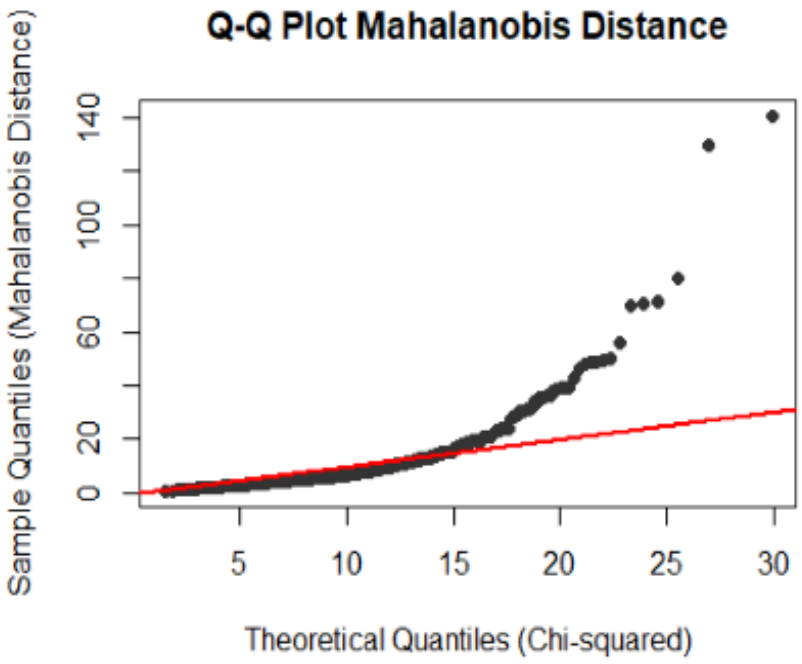

Figure 1. Plot of Outlier

5. CONCLUSION

Based on the proportion of classification errors, linear discriminant analysis shows better performance with an error proportion of 3.34% compared to robust discriminant analysis which has an error proportion of 12.3%, with the number of outliers of 38 or 6.68%. Therefore, it can be concluded that the linear discriminant analysis method is more effective in classifying malignant and benign breast cancer.

REFERENCES

[1] H. Hayati, W. Alwi, and A. Sauddin, "Faktor-Faktor yang Memepengaruhi Tingkat Stres Mahasiswa Prodi Matematika Fakultas Sains dan Teknologi Universitas Islam Negeri Alauddin Makassar dalam Menyelesaikan Tugas Akhir Menggunakan Confirmatory Factor Analysis (CFA)," *msa*, vol. 9, no. 1, p. 37, Jun. 2021.

[2] M. Hubert, J. Raymaekers, and P. J. Rousseeuw, "Robust discriminant analysis," *WIREs Computational Stats*, vol. 16, no. 5, p. e70003, Sep. 2024.

[3] R. A. Johnson and D. W. Wichern, *Applied multivariate statistical analysis*, sixth edition ed., ser. Pearson modern classics. Upper Saddle River, New Jersey: Pearson, 2019.

[4] M. S. Tjahaya, R. Raupong, and G. M. Tinungki, "Analisis Diskriminan Linear Robust Dengan Metode Winsorized Modified One-Step M-Estimator," *ESTIMASI: Journal of Statistics and Its Application*, vol. 3, no. 1, pp. 1–13.

[5] P. K. Illahi, A. R. Viana, M. Permata, and M. Y. Pratama, "Penerapan Algoritma Decision Tree Dan Regresi Linear Untuk Klasifikasi Kanker Payudara," in *SENTIMAS: Seminar Nasional Penelitian dan Pengabdian Masyarakat*, vol. 2. Institut Riset dan Publikasi Indonesia (IRPI), Aug. 2023, pp. 86–92.

[6] K. K. R. Indonesia, "Rencana Kanker Nasional 2024-2034," 2024.

[7] Ahmad Zikir, Khalilah Nurfadilah, Irwan, and Adiatma, "Perbandingan Metode Clustering Dengan Menggunakan Metode Average Linkage dan Metode K-Means Pada Industri Kecil dan Menengah Di Kabupaten Wajo," *MSA*, vol. 10, no. 2, pp. 57–62, Dec. 2022.

[8] I. Irwan and A. Sauddin, *Statistika Multivariat*. Makassar: Alauddin University Press, 2021.

[9] N. S. Lestari, K. A. Nugroho, and N. Ngadikun, "Deteksi Tumor Payudara (Breast Benign Diseases) Berdasar Interaksi Eritrosit Akibat Perubahan Ion-Ion Dalam Darah-Edta Menggunakan Spektrofotometer Uv-Vis," *Diff.: J. Phy. Ed. App. Phy.*, vol. 5, no. 1, pp. 47–61, Jul. 2023.

[10] L. Li, H. Deng, X. Ye, Y. Li, and J. Wang, "Comparison of the diagnostic efficacy of mathematical models in distinguishing ultrasound imaging of breast nodules," *Sci Rep*, vol. 13, no. 1, p. 16047, Sep. 2023.

[11] S. M. Malakouti, "Use of Machine Learning to Diagnose Benign and Malignant Breast Tissues with the Best Degree of Accuracy and in the Shortest Amount of Time," *Chemotherapy: Open Access*, vol. 10, no. 6, pp. 1–5, 2022.

[12] A. C. Rencher and W. F. Christensen, *Methods of Multivariate Analysis*, 1st ed., ser. Wiley Series in Probability and Statistics. Wiley, Jul. 2012.