Vol.4, No.4, Page 1967- 1976, 2026 https://doi.org/10.59628/jast.v4i4.2555 ISSN 2958-9568 JAST

SUN PUBLISHER

مجلة جامعة صنعاء للعلوم التطبيقية والتكنولوجيا

Sana'a University Journal of Applied Sciences and Technology

https://journals.su.edu.ye/index.php/jast/

# The Discriminate Analysis with Applied on Diabetes non- Diabetes in Sana'a, Yemen for the Year 2025

**Amtalteef M.A. Al-Hamzi[1] * and Amani M.A. Al-Hamzi[2]**

[1]**Department of Mathematical, Faculty of Education, Sana'a University, Sana'a, Yemen,**
[2]**Department of Pediatric, Faculty of Medicine, Sana'a University, Sana'a, Yemen.**

*Corresponding author: alhmzi2018@gmail.com.

## Abstract

This study aimed to identify the most significant factors distinguishing between diabetes and non-diabetes, using linear discriminant analysis LDA. A sample of 368 individuals (225 diabetes mellitus, 143 non- diabetes mellitus) was analyzed based on data collected through via personal interviews and Al-Thawrah General Hospital in the capital Sana'a records, for the year 2025. The stepwise discriminant analysis led to the building of a statistically significant model. Results revealed that the variables ($x_5$) body mass index BMI, ($x_{10}$) Cholesterol, ( $x_6$) Gout, ( $x_9$) Blood pressure, and ( $x_2$) Age, had the highest discriminative power, as indicated by standardized coefficients and Wilks' Lambda values ($p < 0.05$). These variables significantly contributed to classifying cases, while the remaining variables did not demonstrate a significant effect. The overall correct classification rate for diabetes mellitus DM reached 87.3%, when the overall classification accuracy of the model reached 89.9%, confirming its strong predictive power.

## 1. INTRODUCTION

The International Diabetes Federation in 2021 indicates that the global number of diabetes cases has climbed to 537 million adults aged 20-79 years worldwide and threatening 860 million adults worldwide due to impaired glucose tolerance and impaired fasting glucose, which are commonly known as prediabetes, thus, diabetes has become a severe global public health concern [1, 2]. Diabetes mellitus (DM) is a combination of metabolic disorders characterized by elevated blood glucose levels over a prolonged duration [3]. Multivariate analysis is a key statistical approach used in various fields, especially when dealing with many variables. Among its purposes is identifying group characteristics, such as health outcomes [4, 5]. Linear Discriminant Analysis (LDA) is one of the most commonly used methods for classifying individuals into groups based on measured indicators, particularly in health research. It is effective when the dependent variable is categorical, helping to determine the best combination of predictors for group separation. The method maximizes between-group variance while minimizing within-group variance, allowing for the development of a statistically sound and interpretable predictive model. In medical contexts, it is commonly applied in disease classification, identification of risk factors, and prediction of health outcomes, thereby supporting evidence-based clinical decisions [6]. Diabetes mellitus the growing prevalence of infectious diseases, there remains a need to identify and quantify the most influential factors associated with diabetes. This study employed multivariate statistical techniques, particularly the linear discriminant analysis technique was used to fit a predictive equation based on the measured variables for classifying new individuals, and to re-classify the original data diabetes from those without, to enable the interpretation of the predictive equation for better understanding of the relationships that may exist among the variables.

## 2. METHODOLOGY AND METHODS

### 2.1. Study Hypotheses

This study seeks to examine the following two hypotheses:

**1.** There were statistically significant differences between diabetes and without non-diabetes individuals attributed to the independent variables, which are considered discriminant factors with an effective ability to predict infection.

**2.** The linear discriminant model has high classification accuracy in assigning individuals to their correct group.

©2026 The Author(s). This article is licensed under CC By 4.0

### 2.2. Study Goal:

We aim through this study to achieve the following two points:

**1.** To examine the statistical significance of the studied factors as valid discriminant variables for predicting infection and classifying individuals into diabetes and non-diabetes groups

**2.** To provide a basis for controlling and managing the most statistically significant factors affecting infection

### 2.3. Importance of study

The importance of this study lies in highlighting the increasing prevalence of diabetes mellitus in Sana'a City, with the objective of drawing the attention of relevant authorities to the need for implementing interventions to mitigate this phenomenon. Moreover, it fills the existing gap resulting from the scarcity of medical and statistical studies on diabetes mellitus in Sana'a and Yemen in general.

### 2.4. Sample size

The study sample consisted of 368 persons, collected from individuals attending Al-Thawra General Hospital, Sana'a Capital Municipality, including 225 with diabetes and 143 without non-diabetes, during the period from January to May 2025.

### 2.5. Statistics methods:

A structured questionnaire was designed to collect personal and health-related information to identify the most influential factors associated with diabetes. We were granted access to the individuals and medical files after acquainting the hospital administration with the objectives and importance of the study, ensuring that the data required for the questionnaire would only be used for research purposes. The questionnaire data were analyzed using SPSS 27. Discriminant analysis was used to identify the most contributing factors in distinguishing and classifying diabetes from non- diabetes, with statistical significance ($p$<0.05). Ten potential factors (age, gender, marital status, immunohistochemistry, body mass index, gout, smoking, sports, and blood pressure) influencing the dependent variable (diabetes non diabetes) were proposed.

## 3. THEORETICAL ASPECT OF DISCRIMINANT ANALYSIS AND THE DISCRIMINANT FUNCTION

### 3.1. Discriminate Analysis DA

DA is a key multivariate statistical method used to classify populations into overlapping groups based on their shared characteristics. It involves a qualitative dependent variable (binary or multiple) and quantitative independent variables that help distinguish between the groups. The analysis builds a discriminant function (DF) to classify new cases with minimal error. There are three types of regression analysis: direct (all variables entered at once), hierarchical (based on a set order), and stepwise (based on statistical criteria). Discriminant functions can be linear (e.g., Fisher's) or nonlinear, with wide applications in various fields [7]. The aim is to construct a linear combination of the most influential variables that best distinguish between groups of the dependent variable, assess the classification accuracy, and apply a quantitative rule to assign new observations to the correct group with minimal classification error [8].

### 3.2. Mathematical theoretical aspect of the LDF

The LDF is a basic classification model derived from random samples of two populations used to assign observations to the correct group. It represents a simple form of discrimination based on specific statistical assumptions:

**1.** The distribution of independent (explanatory) variables is normal [9].

**2.** The variances are equal for all sums (variance and covariance matrices), i.e. accepting the null hypothesis when testing the hypothesis:

$$H_0: \Sigma_1 = \Sigma_2 = \ldots = \Sigma_k VS \quad H_1: \Sigma_1 \neq \Sigma_2 \neq \ldots \neq \Sigma_k$$

$\sum$ *is variance and covariance matrices*

k: Number of totals. And test using M Box's.

**3.** Selecting samples randomly and classifying their observations $n_1, n_2$ into groups accurately.

**4.** The discriminant function is a linear function.

#### *3.2.1. Formulating DA*

Let two groups (samples) of size $n_1, n_2$ were chosen from two populations, and assume that the observation values for m independent variables $x_1, x_2, \ldots, x_m$ which are relied upon by classification. To formulate the following linear equation:

discrimination function:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m \qquad (1)$$

$\beta_i$: Discriminant coefficients for the explanatory variables were used for the classification process; X: Vector of random variables [10].

#### *3.2.2. Find the DF*

There are various methods to construct a linear discriminant function, such as using the probability density function for each group or applying the Lagrange multiplier method, as used in this study. These approaches aim to

©2026 The Author(s). This article is licensed under CC By 4.0

select coefficient estimates ($\beta_i$) that minimize the within-group variance while maximizing the between-group variance. The ratio of between-group to within-group variation is denoted by the symbol q, which measures the distance between the two populations [11]:

$$q = \frac{\text{Between-group variation}}{\text{Within-group variation}} = \frac{(\vec{Y}^{(1)}-\vec{Y}^{(2)})^2}{\sum_i^{n_1}(Y_i^{(1)}-\vec{Y}^{(1)})^2+\sum_j^{n_2}(Y_j^{(2)}-\vec{Y}^{(2)})^2} \quad (2)$$

Choosing values ($\beta_i$) increases the value of (q) [12], where:

$\vec{Y}^{(1)}, \vec{Y}^{(2)}$ : mean observations values for the first and second group

$Y_j^{(1)}, Y_i^{(2)}$ : observation values (i) and (j) in the first and second group,

The larger the value of q, the greater is the difference between the two populations.

### 3.2.3. Estimation of the parameters DF

First, the means of the variables in both the first and second groups were calculated as follows:

$$\bar{X}_1^{(1)} = \frac{\sum_{i=1}^{n_1} x_{1i}}{n_1}, \bar{X}_2^{(1)} = \frac{\sum_{i=1}^{n_1} x_{2i}}{n_1} \ldots, \bar{X}_k^{(1)} = \frac{\sum_{i=1}^{n_1} x_{ki}}{n_1} \quad ; \bar{X}_i^{(1)} \to i = 1,2,\ldots k$$

$$\bar{X}_1^{(2)} = \frac{\sum_{i=1}^{n_2} x_{1i}}{n_2}, \bar{X}_2^{(2)} = \frac{\sum_{i=1}^{n_2} x_{2i}}{n_2} \ldots, \bar{X}_k^{(2)} = \frac{\sum_{i=1}^{n_2} x_{ki}}{n_2} \quad ; \bar{X}_i^{(2)} \to i = 1,2,\ldots k$$

where $\bar{X}_i^{(1)}, \bar{X}_i^{(2)}$ are the averages of variable (i) in the first and second groups, respectively.

Second: Calculate the difference between the means of each variable as follows:

$$d = \begin{bmatrix} d_1 \\ d_2 \\ . \\ d_k \end{bmatrix} = \begin{bmatrix} \bar{X}_1^{(1)} - \bar{X}_1^{(2)} \\ \bar{X}_2^{(1)} - \bar{X}_2^{(2)} \\ . \\ \bar{X}_k^{(1)} - \bar{X}_k^{(2)} \end{bmatrix} \quad (3)$$

Third: Calculate the sum of the squares of each variable and the sum of the product of every two variables in each group, the sum of the squares of variable (i) in the first group:

$$S_{ii}^1 = \sum_{i=1}^{n_1} x_{ii}^2 - \frac{(\sum_{i=1}^{n_1} x_{ii})^2}{n_1} \quad ; i = 1,2,\ldots n_1$$

the sum of the squares of variable (i) in the second group:

$$SS_{ii}^2 = \sum_{i=1}^{n_2} x_{ii}^2 - \frac{(\sum_{i=1}^{n_2} x_{ii})^2}{n_2} \quad ; i = 1,2,\ldots n_2$$

To calculate the product of the two variables, we use the following:

$$S_{ii} = \sum x_{ij} - \frac{(\sum x_i)(\sum x_j)}{n}$$

Fourth, the variance and covariance matrices for the two groups are found as follows:

$$S_{ii} = \frac{S_i^{(1)} + S_i^{i2)}}{n_1 + n_2 - 2} \qquad S_{ij} = \frac{S_{ij}^{(1)} + S_{ij}^{(2)}}{n_1 + n_2 - 2} \quad (4)$$

The covariance and covariance matrix within the two groups can be determined as follows:

$$\underline{S} = \frac{(n_1 - 1)\underline{S}^{(1)} + (n_2 - 1)\underline{S}^{(2)}}{n_1 + n_2 - 2} \quad (5)$$

$$S = \begin{bmatrix} s_{11} & s_{21} & \ldots & s_{1k} \\ s_{21} & s_{22} & \ldots & s_{2k} \\ . & . & . & . \\ s_{k1} & s_{k2} & \ldots & s_{kk} \end{bmatrix} \quad (6)$$

Using matrices, the following formula is:

$$Y = X'S^{-1}(\bar{X}^{(1)} - \bar{X}^{(2)}) = X'\beta \quad (7)$$

where: $\beta = S^{-1}(\overline{X}^{(1)} - \overline{X}^{(2)})$

Thus, it can be found $\vec{Y}^{(1)}, \vec{Y}^{(2)}$ as following [12]:

$$\bar{Y}^{(1)} = \underline{\bar{X}'}^{(1)}\beta = \beta_1\bar{X}_1^{(1)} + \beta_2\bar{X}_2^{(1)} + \ldots + \beta_m\bar{X}_m^{(1)}$$
$$\bar{Y}^{(2)} = \underline{\bar{X}'}^{(2)}\underline{\beta} = \beta_1\bar{X}_1^{(2)} + \beta_2\bar{X}_2^{(2)} + \ldots + \beta_m\bar{X}_m^{(2)} \quad (8)$$

$$\because \left(\vec{Y}^{(1)} - \vec{Y}^{(2)}\right)^2 = \left(\underline{\bar{X}}^{(1)}\underline{\beta} - \underline{\bar{X}}^{(2)}\underline{\beta}\right)^2 = \left(\underline{\beta}'(\underline{\bar{X}}^{(1)} - \underline{\bar{X}}^{(2)})'(\underline{\bar{X}}^{(1)} - \underline{\bar{X}}^{(2)})\underline{\beta}\right)$$

$$\sum_{i=1}^{n_1}(Y_i^{(1)} - \bar{Y}^{(1)})^2 = \underline{\beta}'\left(\sum_{i=1}^{n_1}(X_i^{(1)} - \underline{\bar{X}}^{(1)})'(X_i^{(1)} - \underline{\bar{X}}^{(1)})\right)\underline{\beta} = \underline{\beta}'(n_1 - 1)\underline{S}^{(1)}\underline{\beta}$$

$$\sum_{i=1}^{n_2}(Y_j^{(2)} - \bar{Y}^{(2)})^2 = \underline{\beta}'\left(\sum_{i=1}^{n_2}(X_i^{(2)} - \underline{\bar{X}}^{(2)})'(X_i^{(2)} - \underline{\bar{X}}^{(2)})\right)\underline{\beta} = \underline{\beta}'(n_2 - 1)\underline{S}^{(2)}\underline{\beta}$$

$$\sum_{i=1}^{n_1}(X_i^{(1)} - \underline{\bar{X}}^{(1)})'(X_i^{(1)} - \underline{\bar{X}}^{(1)}) = (n_1 - 1)\underline{S}^{(1)}$$

$$\sum_{i=1}^{n_2}(X_i^{(2)} - \underline{\bar{X}}^{(2)})'(X_i^{(2)} - \underline{\bar{X}}^{(2)}) = (n_2 - 1)\underline{S}^{(2)}$$

$$\therefore \sum_{i=1}^{n_1}(Y_i^{(1)} - \bar{Y}^{(1)})^2 + \sum_{j=1}^{n_2}(Y_j^{(2)} - \bar{Y}^{(2)})^2 = \underline{\beta}'(n_1 - 1)\underline{S}^{(1)}\underline{\beta} + \underline{\beta}'(n_2 - 1)\underline{S}^{(2)}\underline{\beta} \quad (9)$$

Where $\underline{S}^{(2)}, \underline{S}^{(1)}$ are the two variance matrices within the first and second groups, respectively. Assuming that $\underline{s}$ represents the joint variance of the two samples together, it is defined by Eq. (5):

$$\therefore (n_1 + n_2 - 2)\underline{S} = (n_1 - 1)\underline{S}^{(1)} + (n_2 - 1)\underline{S}^{(2)}$$

$$\therefore \sum_{i=1}^{n_1}(Y_i^{(1)} - \vec{Y}^{(1)})^2 + \sum_{i=1}^{n_2}(Y_i^{(2)} - \vec{Y}^{(2)})^2 = (n_1 + n_2 - 2)\underline{\beta}'\underline{S}\underline{\beta} \quad (10)$$

By substituting into Eq. (2)

©2026 The Author(s). This article is licensed under CC By 4.0

$$q = \frac{\underline{\beta}'(\bar{\underline{X}}^{(1)} - \bar{\underline{X}}^{(2)})'(\bar{\underline{X}}^{(1)} - \bar{\underline{X}}^{(2)})\underline{\beta}}{(n_1 + n_2 - 2)\underline{\beta}'\underline{S}\underline{\beta}}$$

$$= \frac{1}{(n_1+n_2-2)} q^* \qquad (11)$$

Where $q^*$ : ratio to be enlarged.

When assuming the variance within the groups is a fixed amount, that is, $\underline{\beta}'\ \underline{S}\ \underline{\beta}$ working to enlarge the variance between the groups, the Lagrange factorial coefficient will be used and it becomes as follows:

$$L(\beta, a) = \underline{\beta}'(\bar{\underline{X}}^{(1)} - \bar{\underline{X}}^{(2)})'(\bar{\underline{X}}^{(1)} - \bar{\underline{X}}^{(2)})\underline{\beta} - a(\underline{\beta}'\underline{S}\underline{\beta} - 1)$$

$$\therefore \frac{\delta L}{\delta \underline{\beta}} = 2(\bar{\underline{X}}^{(1)} - \bar{\underline{X}}^{(2)})(\bar{\underline{X}}^{(1)} - \bar{\underline{X}}^{(2)})\underline{\beta} - 2a\underline{S}\underline{\beta} = 0$$

$$2(\bar{\underline{X}}^{(1)} - \bar{\underline{X}}^{(2)})(\bar{\underline{X}}^{(1)} - \bar{\underline{X}}^{(2)})\underline{\beta} - 2a\underline{S}\underline{\beta} = 0$$

$$(\bar{\underline{X}}^{(1)} - \bar{\underline{X}}^{(2)})(\bar{\underline{X}}^{(1)} - \bar{\underline{X}}^{(2)})'\underline{\beta} = (\bar{\underline{X}}^{(1)} - \bar{\underline{X}}^{(2)})D^2$$

A measure of the distances between the centers of groups, called the Mahalonobis distance, was named after the Indian scientist (1930). The larger its value, the greater the difference between the two groups.

$$D^2 = \left((\bar{\underline{X}}^{(1)} - \bar{\underline{X}}^{(2)})'\underline{\beta}\right) \qquad (12)$$

$$\therefore (\bar{\underline{X}}^{(1)} - \bar{\underline{X}}^{(2)})D^2 = a\underline{S}\underline{\beta}$$

$$\hat{\underline{\beta}} = \underline{S}^{-1}(\bar{\underline{X}}^{(1)} - \bar{\underline{X}}^{(2)}) = \underline{S}^{-1}d \qquad (13)$$

($\bar{\underline{X}}^{(1)}$ - $\bar{\underline{X}}^{(2)}$) : the difference vector between the means of the first and second groups.

$\underline{S}^{(-1)}$ : Inverse of the estimated variance–covariance matrix. Thus, the estimated linear LDF is given as the equation [13]:

$$\hat{Y}_i = \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + \ldots + \hat{\beta}_m X_{mi} \qquad ; i = 1,2,\ldots,n \qquad (14)$$

## 3.3. Test the significance of a LDF (the ability of a function to discriminate)

The Hotelling test assesses significant differences between group means, whereas the F-test evaluates parameter significance to distinguish between groups and construct statistically valid **DF**s.

### 3.3.1. Test the significance of differences among means by test the hypothesis

$H_0$: A function does not has the ability to distinguish versus

$H_1$:A function has the ability to distinguish.

$$H_0: \underline{\mu}_1 = \underline{\mu}_2 = \cdots = \underline{\mu}_m = 0 \quad VS$$

$$H_1: \underline{\mu}_1 \neq \underline{\mu}_2 \neq \ldots \neq \underline{\mu}_m \neq 0 \qquad (15)$$

There are three types of tests

**1.** Hotelling $(T^2)$ Statistics

In 1931, Hotelling introduced the statistical index $(T^2)$ to test whether two multivariate normal groups with equal covariance matrices are significantly different [14]. This index $(T^2)$ is used prior to applying discriminant analysis to assess the hypothesis (15) of group differences:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} D^2 \qquad (16)$$

The Hotelling $(T^2)$ statistic is proportional to and dependent on the Mahalanobis $(D^2)$ statistic. This corresponds to the F-value from ANOVA and can be approximated using the F-distribution, as follows:

$$F_{cal} = \frac{MSB}{MSE} = \frac{n_1 + n_2 - m - 1}{(n_1 + n_2 - 2)m} * (T^2)$$

Table value $F_{tab(\alpha, m-1, n_1 = n_2 - m - 1)}$

If $(F_{cal} > F_{tab})$, $(H_0)$ is rejected, that is, the function has the ability to distinguish.

**2. ANOVA to test the significance of a LDF:**

For test the hypothesis [15]:

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_m = 0 \quad vs$$

$$H_1: \beta_1 \neq \beta_2 \neq \cdots \neq \beta_m \qquad (17)$$

Using ANOVA table, differences between and within groups are measured:

ANOVA table for discriminant analysis

| Sours | S.S | d.f | M.S | F |
|---|---|---|---|---|
| Between X,s | $SSB = \frac{(n_1 n_2)(D^2)^2}{(n_1 + n_2)(n_1 + n_2 - 2)}$ | $m-1$ | MSB | $F_{cal} = \frac{MSB}{MSE}$ $F_{tab}(\alpha, m-1, n_1 + n_2 - m)$ |
| Within X,s Error | $SSE = D^2 = \hat{\beta}_1 d_1 + \hat{\beta}_2 d_2 + \ldots + \hat{\beta}_m d_m$ | $n_1 + n_2 - m$ | MSE | |
| Total | $SST = SSB + SSE$ | $n_1 + n_2 - 1$ | | |

If $(F_{cal} > F_{tab})$, $(H_0)$ is rejected, that is, the LDF has the ability to distinguish.

**3.** Wilk's Lambda test $(\Lambda)$

It is used to test the hypothesis in (15), where the calculated value is

$$\Lambda = \prod_{i=1}^{m} \frac{1}{1 + \lambda_i} \qquad (18)$$

Where: $\lambda_i$ is the eigenvalue of all variables, m is the number of variables, Wilks' Lambda $0 \leq \Lambda \leq 1$ measures the discriminative power of the function, and values near zero indicate strong discrimination, whereas values near one suggest weak discrimination. Variables with the lowest Lambda and highest **F** values were retained in the model.

### 3.3.2. Test the significance of each variable within the DA

Many explanatory variables may be insignificant and excluded to enhance the discrimination. Their significance

©2026 The Author(s). This article is licensed under CC By 4.0

was tested using the (t-test) statistic to compare group means.

$$H_0: \underline{\mu}_1 = \underline{\mu}_2 \quad vs \quad H_0: \underline{\mu}_1 \neq \underline{\mu}_2 \tag{19}$$

The test function is:

$$T = \frac{\bar{X}^{(1)} - \bar{X}^{(2)}}{S_P\sqrt{\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \tag{20}$$

### 3.4. Relative importance of independent variables

Relative importance identifies the key independent variables that contribute to group classification and distinction using the following equation:

$$\beta_i' = \hat{\beta}_i\sqrt{S_{ii}} \tag{21}$$

It reflects the variance extracted from the variance matrix, where larger absolute coefficients indicate more important variables in the discrimination process, ranked from the most to least significant [2].

### 3.5. Cut Point=CP

The cut point between the two groups is written as follows:

$$CP = \frac{1}{2}\left(\vec{Y}^{(1)} + \vec{Y}^{(2)}\right) \tag{22}$$

To obtain the classification function $Y^*$, the cut point can be combined with the discriminant function as follows:

$$Y^* = CP + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m \tag{23}$$

### 3.6. Classification Role

Based on the classification function $Y^*$, the observation is classified for groups according to the following relationship:

$$Y^* \begin{cases} > 0 Group I \\ < 0 Group II \\ 0 Cannot classify \end{cases} \tag{24}$$

### 3.7. Probability of classification error

The use of the DF and a cut-off point may lead to misclassification, where an observation is assigned to the wrong group. This results in two types of classification errors.

**1. Apparent error rate**

Misclassifying an observation from group one into group two is denoted as $n_{21}$, and misclassifying an observation from group two into group one is denoted as $n_{12}$. The respective misclassification percentages were calculated as follows:

$$L_{12} = \frac{n_{12}}{n_1} \quad , L_{12} = \frac{n_{12}}{n_1} \tag{25}$$

**2. The real error**

It is calculated from the following probability:

$$L_{12} = L_{21} = \Phi\left(-\frac{\sqrt{D^2}}{2}\right) \tag{26}$$

where $D^2$ denotes the Mahalonobis statistics and $\Phi$ denotes the normal distribution function [13, 16].

### 3.8. Stepwise Regression Procedure for selecting variables

Stepwise methods were used to select the most significant independent variables by minimizing Wilks' Lambda and maximizing the F-value. Originally developed for regression, this approach has also been applied in discriminant analysis with a binary qualitative dependent variable [15, 16].

## 4. THE PRACTICAL ASPECT

Linear discriminant analysis steps were applied to identify the significant factors used to construct a highly accurate discriminant and classification function that distinguishes diabetes from non-diabetes mellitus. The results of the analysis are presented below.

The independent variables (influencing factors) were nominated for study by doctors and through previous studies, as shown in Table (1).

©2026 The Author(s). This article is licensed under CC By 4.0

Table (1) Symbol and names of the independent variables nominated for the study

| Variable N. | Variable name | Variable N. | Variable name |
|---|---|---|---|
| $x_1$ | Gender | $x_6$ | Gout |
| $x_2$ | Age | $x_7$ | Smoking |
| $x_3$ | Marital status | $x_8$ | Sports |
| $x_4$ | Immunohistochemistry | $x_9$ | Blood pressure |
| $x_5$ | Boody mass index | $x_{10}$ | Cholesterol |

Table (2) Distribution of sample observations according to Diabetes mellitus (DM) incidence

| Groups | Names of the Groups | N. Cases | Percentage % |
|---|---|---|---|
| 1 | Diabetic mellitus | 225 | 61.1 |
| 2 | Non- Diabetic | 143 | 38.9 |
| Total | | 368 | 100 |
| Source: Researcher's analysis | | | |

Table (3) The means and S.D. for each independent variable

| Variable symbol | Diabetes mellitus DM | | Non- Diabetes mellitus DM | |
|---|---|---|---|---|
| | Means | Std. Devotion | Means | Std. Devotion |
| $x_1$ | 0.5803 | 0.4750 | 0.5385 | 0.5003 |
| $x_2$ | 52.1383 | 9.4617 | 45.990 | 9.647 |
| $x_3$ | 0.8169 | 0.38720 | 0.79720 | 0.4035 |
| $x_4$ | 0.22786 | 0.42067 | 0.32867 | 0.4714 |
| $x_5$ | 31.9016 | 6.12465 | 26.514 | 5.6667 |
| $x_6$ | 0.4576 | 0.499490 | 0.0979 | 0.29822 |
| $x_7$ | 0.43394 | 0.49960 | 0.2867 | 0.4532 |
| $x_8$ | 0.41589 | 0.498052 | 0.41958 | 0.4953 |
| $x_9$ | 0.37538 | 0.48563 | 0.7623 | 0.42742 |
| $x_{10}$ | 0.65635 | 0.47412 | 0.22378 | 0.418239 |

Source: Researcher's analysis

Table (3) presents the descriptive statistics for the two categories of the dependent variable—DM and non-DM individuals–showing the means and standard deviations for each independent variable. Notable differences can be observed among the means of most independent variables for the two groups.

## 4.1. Results

### *4.1.1. Checking the conditions for applying DA*

**First: Testing the normal distribution of the sample**

Since the sample size is 368, which exceeds 30, the data can be assumed to follow a normal distribution, according to the Central Limit Theorem.

**Second: Equal variance test**

Using M Box's covariance matrix equality test, to test the following hypothesis

$$H_0: \Sigma_0 = \Sigma_1 \quad VS \quad H_1: \Sigma_0 \neq \Sigma_1$$

Where: $\sum_0, \sum_1$ represents the variance of variables for DM and non- DM, respectively.

Table (4) Results of Box's M test for the homogeneity of variance test

| | F | | | |
|---|---|---|---|---|
| Box's M | Approx | $df_1$ | $df_2$ | Sig |
| 90.401 | 5.932 | 15 | 367708.21 | 0.100 |
| Source: Researcher's analysis | | | | |

As presented in Table (4), the assumption of homogeneity of variance between the two groups is satisfied (Sig= 0.100).

**Third: Testing the equality of means between the two groups**

Hypothesis (15) tests the equality of means for DM and non-DM.

$$H_0: \mu_0 = \mu_1 \, VS \quad H_1: \mu_0 \neq \mu_1$$

where $\mu_1, \mu_0$ represent the variances of the variables for DM and non-DM, respectively.

Table (5) Significant differences among the means the groups

| Variable symbol | Free dome | | Statistics | | Sig |
|---|---|---|---|---|---|
| | Wilkes' Lambda | F | df 1 | df2 | |
| $x_1$ | 0.998 | 0.621 | 1 | 363 | 0.019 |
| $x_2$ | 0.911 | 35.62 | 1 | 363 | < 0.001 |
| $x_3$ | 0.999 | 0.220 | 1 | 363 | 0.541 |
| $x_4$ | 0.988 | 4.580 | 1 | 363 | 0.033 |
| $x_5$ | 0.811 | 85.00 | 1 | 363 | < 0.001 |
| $x_6$ | 0.854 | 62.57 | 1 | 363 | < 0.001 |
| $x_7$ | 0.978 | 8.097 | 1 | 363 | < 0.001 |
| $x_8$ | 1.000 | 0.007 | 1 | 363 | 0.930 |
| $x_9$ | 0.857 | 60.92 | 1 | 363 | < 0.001 |
| $x_{10}$ | 0.822 | 79.07 | 1 | 363 | < 0.001 |

Source: Researcher's analysis

The Wilks' lambda values and significance levels (Sig < 0.001) reported in Table (5) indicate the presence of statistically significant differences in the mean values between the DM and non-DM groups for the variables

©2026 The Author(s). This article is licensed under CC By 4.0

$x_2, x_4, x_5, x_6, x_7, x_9, x_{10}$, therefore, hypothesis $(H_1)$ is supported, as these variables have a significant effect on DM, in contrast to the remaining variables $x_1, x_3, x_8$.

## 4.2. Building a DA model (function)

### 4.2.1. Identify statistically significant variables

Table (6) Identify the independent variables that contribute to constructing the DF

| Variable symbol | Variable name | Wilkes' s Lambda statistics | Statistics F | Degrees of freedom | | Sig |
|---|---|---|---|---|---|---|
| | | | | $df_1$ | $df_2$ | |
| $x_5$ | BMI | 0.811 | 85.03 | 1 | 363.0 | < 0.00 |
| $x_9$ | Blood pressure | 0.857 | 60.920 | 1 | 363.0 | < 0.00 |
| $x_{10}$ | Cholesterol | 0.822 | 79.055 | 1 | 363.0 | < 0.00 |
| $x_6$ | Gout | 0.854 | 62.573 | 1 | 363.0 | < 0.00 |
| $x_2$ | Age | 0.9115 | 35.621 | 1 | 363.0 | < 0.00 |
| Source: Researcher's analysis | | | | | | |

In table (6), smaller values of Wilks' lambda indicate greater discriminatory ability of the function, the five more important independent variables that contribute significantly to the prediction by the discriminant function as indicated by the smaller values of the Wilks's lambda and based on the stepwise selection are $(x_{5,}, x_9, x_{10}, x_6, x_2)$.

### 4.2.2. Finding the estimated values of the parameters of a LDF

**1.** The distance $(d_i)$ between the two groups was determined using the Eq: (3):

$$d = \begin{pmatrix} 5.388 \\ -0.388 \\ 0.4326 \\ 0.3588 \\ 9.148 \end{pmatrix}$$

**2.** Find the variance and covariance matrix of the two groups as in the Eq. (6):

$$S = \begin{array}{c} \\ x_5 \\ x_9 \\ x_{10} \\ x_6 \\ x_2 \end{array} \begin{array}{c} \begin{array}{ccccc} x_5 & x_9 & x_{10} & x_6 & x_2 \end{array} \\ \begin{pmatrix} 36.62 & 0.788 & -0.620 & 0.574 & 13.430 \\ 0.788 & 0.251 & 0.002 & 0.050 & 0.680 \\ -0.620 & 0.002 & 0.250 & -0.055 & -0.995 \\ 0.574 & 0.050 & -0.055 & 0.219 & 0.791 \\ 13.430 & 0.680 & -0.995 & 0.791 & 101.27 \end{pmatrix} \end{array}$$

**3.** Find the estimated values of the parameters of the linear discriminant function using Eq. (13):

$$\hat{\beta} = S^{-1}d \quad ; d = (\underline{\bar{X}}^{(1)} - \underline{\bar{X}}^{(2)})$$

$$\hat{\beta} = \begin{pmatrix} 0.545 \\ 0.526 \\ 0.461 \\ -0.468 \\ 0.353 \end{pmatrix}$$

The DF based on standardized coefficients is:

$$Y^* = 0.545X_5 + o.526X_{10} - 0.461X_9 + 0.468X_6 + 0.353X_2$$

Table (7) Coefficients of the variables that contribute most to discrimination

| Variable symbol | Variable name | DF coefficients $\hat{\beta}$ | Relative importance % $\beta_i'$ |
|---|---|---|---|
| $x_5$ | BMI | 0.078 | 0.545 |
| $x_9$ | Blood pressure | 1.233 | −0.461 |
| $x_{10}$ | Cholesterol | −1.066 | 0.526 |
| $x_6$ | Gout | 0.801 | 0.468 |
| $x_2$ | Age | 0.024 | 0.353 |
| Source: Researcher's analysis | | | |

Table (7) illustrates the relative importance of the influencing factors and their contribution to discrimination and prediction in the discriminant model. The most influential variables, in descending order of importance, are $(x_5)$ Body Mass index (BMI), $(x_{10})$ cholesterol, $(x_6)$ gout, $(x_9)$ blood pressure, and $(x_2)$ Age, This ranking reflects the extent to which each variable contributes to the discrimination between the studied groups.

### 4.2.3. Testing the significance of the LDF model

There are several tests:

**1. Wilkes"s Lambda test**:

Table (8) Wilks's Lambda test for DF significance

| Test of Function | Wilkes' Lambda statistic | Chi square | df | (Sig) |
|---|---|---|---|---|
| 1 | 0.401 | 331.250 | 5 | < 0.001 |

Table (8) presents the results of Wilks' Lambda and Chi-square $(x^2)$ test, which are statistically significant at (Sig < 0.001). This indicates that the DF significantly differentiates between the groups, confirming the model's effectiveness in classification.

**2. Canonical correlation**

Table (9) Eigenvalue and Canonical Correlation

| Function | Eigenvalue | Variance of % | Cumulative% | Canonical Correlation | Eta square |
|---|---|---|---|---|---|
| 1 | 2.15 | 100 | 100 | 0.826 | 0.682 |

As shown in Table (9), the eigenvalue of 2.15 indicates the proportion of variance explained between the two groups— DM and non- DM which is attributed to the differences captured by the single DF. Meanwhile, the overall canonical correlation reached 0.826, indicating a strong association between the independent variables and the dependent variable in forming this group. Moreover, the coefficient of determination reached 0.682, in-

©2026 The Author(s). This article is licensed under CC By 4.0

dicating that 68.2% of the variance is attributable to the difference between the two groups, as explained by the DF.

### 3. Hotelling$(T^2)$ Statistics

To test Hypothesis (16), the statistic defined in Equation (17) is used as follows

$$D^2 = (\bar{X}_1 - \bar{X}_2)' S^{-1} (\bar{X}_1 - \bar{X}_2) = d' S^{-1} d = 5.994$$

$$\therefore \quad T^2 = \frac{n_1 n_2}{n_1 + n_2} D^2 = 524.068$$

$$F_{cal} = \frac{n_1 + n_2 - m - 1}{(n_1 + n_2 - 2)m} * (T^2) =$$

$$\frac{143 + 225 - 5 - 1}{(143 + 225 - 2)(5)} * (524.068) = 103.67$$

Using the significance of ($\alpha$=0.05), the tabular value of the distribution (F) is:

$$F_{tab}(\alpha, m-1, n_1 + n_2 - m - 1) = F_{tab}(0.05,4,362) = 2.21$$

Because $F_{cal} > F_{tab}$, the alternative hypothesis is accepted, indicating that the LDF is effective in distinguishing between the two groups, with statistically significant differences observed.

### 4. ANOVA for DF Significance

ANOVA was used to test Hypothesis (20):

$$\because \quad SSB = \frac{(n_1 n_2)(D^2)^2}{(n_1+n_2)(n_1+n_2-2)} = 9.73, \quad and \quad SSE = D^2 = 5.994$$

Table (10) Analysis of variance for discriminant analysis

| Source | S.S | d.f. | M.S | F |
|---|---|---|---|---|
| SSB | 9.73 | 4 | 2.431 | $F_{cal} = \frac{MSB}{MSE} = 146.82$ <br> $F_{tab}(0.05,4,362) = 2.21$ |
| *SSE* | 5.994 | 362 | 0.0166 | |
| Total | 15.724 | 366 | | |

Source: Researcher's analysis

In table (10) because $F_{cal} > F_{tab}$, the alternative hypothesis is accepted, indicating that the DF can distinguish between groups and perform accurate classification.

#### 4.2.4. Discriminant and Classification Function Model

The actual prediction equation based on the unstandardized coefficients can be used as a predictive model for the classification of new patients. These predictor variables provided the best discrimination between groups; the fitted linear discriminant model $Y^*$ is

$$\hat{Y} = -3.840 + 0.078X_5 + 1.233X_9 - 1.066X_{10} + 0.801X_6 + 0.024X_2$$

#### 4.2.5. Probability of classification error

Using Eq. (26) The apparent classification errors presented in Table (11) were obtained.

Table (11) Classification of the dependent variable by DF

| Dependent variable (premature baby status) | | Diabetes | Non–Diabetes | Apparent classification error |
|---|---|---|---|---|
| Count | Diabetes | 197 | 28 | 225 |
| | Non– Diabetes | 9 | 134 | 143 |
| % | Diabetes | 87.6 | 12.4 | 100 |
| | Non– Diabetes | 6.3 | 93.7 | 100.0 |

Source: Researcher's analysis

Table (11) illustrates the prediction accuracy, showing that 87.6% of diabetes cases were correctly classified with an error rate of 12.4%, while 93.7% of non- diabetes were correctly classified, with an error rate of 6.3%. In general, the LDF correctly classified 89.9% of the data.

## 5. DISCUSSION

The current study found that the variables influencing between the diabetes and non – diabetes age, smoking, as the classification rate reached 89.9%, which is a high rate, and the classification error rate is small, 10.1%, and this indicates high efficiency in the classification model. These results align with [17] which concluded that the most influential and important factors in developing diabetes are weight, then blood pressure, then smoking, as the classification rate reached 90.6%. Also [18] reported that DFA models were developed using variables such as age, fasting blood glucose, BMI, waist girth, blood pressure, HDL, triglycerides, and total cholesterol and these results showed a high classification specificity of approximately 97% in identifying non-T2DM individuals. These findings indicate that DFA can effectively discriminate between diabetic and non-diabetic individuals. Another study [19] found significant differences between diabetic patients with and without kidney failure, with urea and creatinine being the most influential variables, where the discriminant model achieved a high classification accuracy of 91%. [20] the Fisher's linear Discriminant function FLDF was used, where patient's age and gender were found to be the two most important contributing variables in classifying a patient between the two groups diabetic or non-diabetic patient, Up to 65.4% correct classification was achieved. [21] Found that age, male sex, positive family history of type 2 diabetes, high BMI, unhealthy lifestyle, anxiety, depression, high blood pressure, high triglycerides, and a high fatty liver index are risk factors for the progression from prediabetes to type 2 diabetes and should be given sufficient attention. Our study also corresponded with [22] Confirmed the efficiency of the probabilistic model in classifying heart patients, identifying smoking, blood pressure, weight, and age as key factors, with a low misclassification rate of 10%. [23] which developed a discriminant model to predict diabetic foot problems using variables such as duration of infection, blood glucose, age, weight, gender, and blood pressure. Results showed that all variables

©2026 The Author(s). This article is licensed under CC By 4.0

had a significant effect, with the duration of infection being the most influential, followed by blood pressure, while weight had the least impact. [24] Indicted that age > 44 years, divorced/widowed, overweight, central obesity, hypertension, dyslipidemia, a sedentary lifestyle and a family history of diabetes were risk factors for DM. [25] The results highlighted DFA's potential in OA diagnosis, suggesting its utility in managing complex data and aiding personalized treatment strategies. The study underscores the need for larger sample sizes and additional biomarkers to enhance diagnostic robustness and provides a foundation for integrating DFA into clinical practice for early OA detection. [26] Another study demonstrated the effectiveness of the Adaptive Behavior Scale in distinguishing between normal and mentally retarded students, with statistically significant differences favoring normal students. The overall classification accuracy reached 100%, while gender-based classification accuracy was 60.2% for normal students and 60.6% for mentally retarded students.

## 6. CONCLUSIONS

This study highlighted the potential of the discriminant function in distinguishing and classifying individuals based on diabetes disease mellitus, with the aim of developing a predictive model for accurate identification of diabetes disease and non- diabetes. The study found that the variables included in the discriminant equation—$(x_5)$ Body Mass index (BMI), $(x_{10})$ cholesterol, $(x_6)$ gout, $(x_9)$ blood pressure, and $(x_2)$ age —were statistically significant factors. These variables demonstrated discriminative and classificatory power in differentiating patients with diabetes from those without, achieving a correct classification accuracy of 89.9%.

## 7. RECOMMENDATIONS

Therefore, it is recommended to employ DA analysis and the DF model to identify the factors affecting the incidence of diabetes and facilitate early diagnosis.

Further statistical studies are recommended to investigate the impact of additional influencing factors, such as psychological status, sudden shocks or stress, genetic factors, and other related variables

Moreover, it is recommended to continue conducting statistical research on diabetes, given its high prevalence, to develop effective strategies for its prevention and treatment.

## REFERENCES


[1] D. J. Magliano and E. J. Boyko, *IDF Diabetes Atlas*, 10th. Brussels: International Diabetes Federation, 2021.

[2] H. Sun, P. Saeedi, et al., "Erratum to "idf diabetes atlas: Global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045"," *Diabetes Res. Clin. Pract.*, vol. 183, 2023, j.diabres.2023.110945.

[3] Y. Shang, A. Marseglia, et al., "Natural history of prediabetes in older adults from a population-based longitudinal study," *J. Intern. Med.*, vol. 286, no. 3, pp. 326–340, Sep. 2019.

[4] A. Agresti, *Categorical Data Analysis*, 2nd ed. Hoboken, NJ: Wiley, 2002.

[5] C. J. Huberty and S. Olejnik, *Applied MANOVA and Discriminant Analysis*. New York, NY: John Wiley & Sons, 2006.

[6] D. Huang, Y. Quan, M. He, B. Zhou, et al., "Comparison of linear discriminant analysis methods for the classification of cancer based on gene expression data," *J. Exp. & Clin. Cancer Res.*, vol. 28, no. 1, p. 149, 2009. DOI: 10.1186/1756-9966-28-149.

[7] H. N. Al-Rifaim, *Data Analysis and Modeling Using Computers: A Comprehensive Application of the SPSS Package*. Jordan: Al-Ahlia Publishing, 2006.

[8] G. Mahfut, *Statistical Analysis Using SPSS*, 1st ed. Amman, Jordan: Wael Printing House, 2008.

[9] A. A. Afifi and V. Clark, *Computer-Aided Multivariate Analysis*. Belmont, California, USA: Lifetime Learning Publications, 1984.

[10] A. A. Kenan, "The effectiveness of using cluster analysis and discriminant analysis in verifying the discriminant significance of intelligence and personality tests," Ph.D. dissertation, University of Damascus, Syria, 2015.

[11] T. Raykov and G. A. Marcoulides, *An Introduction to Applied Multivariate Analysis*. New York: Taylor & Francis Group, 2008.

[12] M. A. Suleiman, "Comparison between discriminant analysis, the binary logistic model, and neural network models in classifying observations," Ph.D. dissertation, Sudan University of Science and Technology, Sudan, 2015.

[13] R. D. Bock, *Multivariate Statistical Methods in Behavioral Research*. USA: McGraw-Hill, 1975.

[14] D. F. Morrison, *Multivariate Statistical Methods*, 3rd ed. New York: McGraw-Hill Book Company, 1976.

[15] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis*, 2nd ed. New York: John Wiley & Sons, 1984.

[16] A. M. Muhammad, "Using the linear discriminant function to distinguish children with and without diabetes," Ph.D. dissertation, University of Gezira, Wad Madani, Sudan, 2023.

[17] A. Bassiouni, "Using discriminant analysis in classification and prediction," *J. Commer. Finance, Tanta Univ.*, no. 3, pp. 299–325, 2021.

[18] E. N. Liberda et al., "Fisher's linear discriminant function analysis and its potential utility as a tool for the assessment of health-and-wellness programs in indigenous communities," *Int. J. Environ. Res. Public Health*, vol. 17, no. 21, p. 7894, 2020. DOI: 10.3390/ijerph17217894.

[19] F. Al-Nuwairi, "Using the linear discriminant function to distinguish diabetic patients with and without kidney failure," Ph.D. dissertation, University of Sudan, 2013.

[20] N. P. Dibal and C. A. Abraham, "On the application of linear discriminant function to evaluate data on diabetic patients at the university of port harcourt teaching hospital, rivers, nigeria," *Am. J. Theor. Appl. Stat.*, vol. 9, no. 3, pp. 53–56, 2020. DOI: 10.11648/j.ajtas.20200903.1.

[21] Y. Liu et al., "Risk factors for progression to type 2 diabetes in prediabetes: A systematic review and meta-analysis," *BMC Public Health*, vol. 25, no. 1, 2025. DOI: 10.1186/s12889-025-21404-4.

[22] A. Al-Marsomy and S. Al-Hashmi, "Diagnosis of factors affecting heart disease using discriminatory analysis," *J. Adm. Econ.*, vol. 42, pp. 367–377, 2019.


©2026 The Author(s). This article is licensed under CC By 4.0

[23] H. H. Fadl, "Used the discriminant function and the criteria for distinguishing foot problems in diabetic patients," Ph.D. dissertation, University of Sudan, Sudan, 2007.

[24] S. B. Liu, "Study on the prevalence and influencing factors of prediabetes and diabetes in cohort population in 10 provinces and cities," *Chin. Cent. for Dis. Control. Prev.*, pp. 48–52, 2020.

[25] L. J. Coleman, J. L. Byrne, S. Edwards, and R. O'Hara, "Utilizing discriminant function analysis (dfa) for classifying osteoarthritis (oa) patients and volunteers based on biomarker concentration," *Diagnostics*, vol. 14, no. 15, p. 1660, 2024. DOI: 10.3390/diagnostics14151660.

[26] A. Abdi, "Discriminatory analysis of the responses of a sample of normal and mentally retarded students on the american association for mental retardation adaptive behavior scale, part one," Unpublished research, Journal of the Association of Arab Universities, Damascus, Syria, Damascus, Syria, 2013.

©2026 The Author(s). This article is licensed under CC By 4.0