International Journal of Health Informatics

DIABETES PREDICTION USING LOGISTIC REGRESSION AND LINEAR DISCRIMINANT ANALYSIS: A COMPARATIVE STUDY

***[1]Ugorji Clinton Chikezie, [2]Kalu Stephen Iheanyichukwu**

[1]Department of Computer Science, School of Science and Industrial Technology, Ogbonnaya Onu Polytechnic, Aba, Nigeria
[2]Department of Statistics, School of Science and Industrial Technology, Ogbonnaya Onu Polytechnic, Aba, Nigeria

Email: *[1]chiefccu@gmail.com, [2]monarchynwokike@gmail.com

ORCID: [1]https://orcid.org/0009-0008-7367-001X, [2]https://orcid.org/0009-0005-3563-2352

DOI: https://doi.org/10.5281/zenodo.23757424

Abstract

This study compares Logistic Regression (LR) and Linear Discriminant Analysis (LDA) for diabetes classification using the 768-observation Pima Indians Diabetes Dataset, comprising eight predictor variables and a binary outcome (500 non-diabetic and 268 diabetic observations). Zero-coded values in Glucose, Blood Pressure, Skin Thickness, Insulin, and BMI were treated as missing and median-imputed using training data only to prevent data leakage. Both models were evaluated using the same stratified 70:30 train–test split (537 training and 231 test observations; random seed = 42) and stratified ten-fold cross-validation, with preprocessing fitted independently within each fold. On the held-out test set, LR achieved 74.46% accuracy, 53.09% sensitivity, 86.00% specificity, an F1-score of 59.31%, and a ROC-AUC of 0.836. Correspondingly, LDA achieved 73.59% accuracy, 50.62% sensitivity, 86.00% specificity, an F1-score of 57.34%, and a ROC-AUC of 0.835. Across stratified ten-fold cross-validation, the mean accuracy was 76.68% for LR and 76.42% for LDA, while both models achieved a mean ROC-AUC of 0.837 to three decimal places. To statistically compare their paired classification outcomes, an exact two-sided McNemar test was conducted on the held-out predictions. The paired classification table indicated that both models correctly classified 170 observations and misclassified 59 observations; LR correctly classified two additional observations that LDA misclassified, whereas no observations were correctly classified by LDA and misclassified by LR. The resulting exact McNemar p-value was 0.500, indicating no statistically significant difference in the models' marginal classification performance at the 5% significance level. Taken together, the results suggest comparable predictive performance, with modest numerical advantages for LR in accuracy, sensitivity, F1-score, and test-set ROC-AUC, but insufficient statistical evidence to establish a decisive winner. These findings represent a benchmark comparison rather than clinical validation, and the models' generalizability should be assessed using independent external datasets before practical clinical application.

Keywords: diabetes prediction; Logistic Regression; Linear Discriminant Analysis; Pima Indians Diabetes Dataset; classification; McNemar test; cross-validation; model validation.

Introduction

Diabetes mellitus constitutes a major public health challenge because of its increasing prevalence, associated complications and growing demands on healthcare systems. The International Diabetes Federation (IDF, 2025) estimated that approximately 589 million adults aged 20–79 years were living with diabetes globally in 2024, representing 11.1% of the adult population, with the number projected to reach 853 million by 2050. Beyond the increasing number of affected individuals, delayed diagnosis remains a substantial concern, as approximately 252 million adults with diabetes were estimated to be unaware of their condition (IDF, 2025). The World Health Organization (WHO, 2024) identified cardiovascular disease, kidney failure, visual impairment and lower-limb amputation among the major complications associated with diabetes, highlighting the consequences of inadequate detection and management. These challenges demonstrate that addressing diabetes requires more than treating established disease; it also demands effective approaches for identifying affected individuals and those requiring further clinical assessment. As the burden of diabetes continues to increase, the ability to analyse available clinical information and distinguish individuals with and without diabetes has become an important consideration in preventive healthcare and medical research.

The increasing demand for early diabetes identification has encouraged the application of statistical and computational methods to clinical datasets. These approaches allow researchers to examine relationships between patient characteristics and health outcomes, identify potentially informative predictors and develop models capable of classifying individuals according to their observed disease status. Smith et al. (1988) demonstrated an early application of computational classification to diabetes-related data, illustrating the potential of clinical measurements for distinguishing disease outcomes. More recently, Collins et al. (2024) emphasised that clinical prediction models developed through regression or machine learning require transparent reporting of their development, validation and performance. This distinction is important because the usefulness of a classification model depends not merely on its ability to produce predictions but also on whether those predictions are sufficiently reliable for the intended purpose. Van Calster et al. (2019) further explained that poorly calibrated prediction models can provide misleading estimates of risk, even when their discrimination appears satisfactory. Consequently, diabetes prediction research must consider both the statistical properties of the models employed and the appropriateness of the procedures used to evaluate their performance.

Within this methodological context, Logistic Regression (LR) remains an established approach to binary classification because it directly models the probability of an outcome using explanatory variables. Its application is particularly relevant when researchers seek to examine the contribution of demographic and clinical characteristics to the likelihood of diabetes classification. In contrast, Linear Discriminant Analysis (LDA) constructs a classification function from the statistical distributions of predictors within predefined groups. Although both methods can generate linear classification boundaries, their underlying assumptions differ. Logistic Regression models the log odds of an outcome as a function of predictors, whereas conventional LDA assumes multivariate normality within groups and a common covariance matrix. These distinctions are relevant because the distributional characteristics of clinical data may influence the suitability of either technique. Araveeporn (2022), in an investigation comparing linear and quadratic discriminant analysis for diabetes classification, demonstrated the continuing relevance of discriminant approaches to the Pima dataset. Similarly, a comparative investigation of diabetes classification methods indexed in PubMed

examined LDA alongside alternative classifiers, reinforcing the need to evaluate methods using empirical performance rather than relying exclusively on theoretical assumptions. The choice between LR and LDA should therefore be informed by the characteristics of the dataset and the results of appropriately conducted model comparisons.

The importance of comparative evaluation becomes more apparent when classification performance is considered in relation to healthcare decision-making. A model may achieve relatively high overall accuracy while demonstrating weaker sensitivity, meaning that some individuals with diabetes are incorrectly classified as non-diabetic. Conversely, a model with greater sensitivity may produce more false-positive classifications, potentially increasing the number of individuals referred for additional assessment. A comparative study of diabetes classification methods reported different performance values for accuracy, sensitivity and specificity, illustrating why these measures should not be treated as interchangeable indicators of model quality. Van Calster et al. (2019) further argued that discrimination and calibration address different aspects of prediction performance and that the latter is particularly important when predicted probabilities may influence clinical decisions. Collins et al. (2024) similarly highlighted the need for transparent reporting of model evaluation procedures so that findings can be interpreted and reproduced. These considerations establish the importance of assessing LR and LDA through complementary measures, including accuracy, sensitivity, specificity, precision and the F1-score, while recognising that probability calibration and discrimination may provide additional information when relevant to the study objectives.

The Pima Indians Diabetes Dataset provides an established empirical basis for investigating these classification questions. The dataset originates from research involving women of Pima Indian heritage and contains 768 observations, comprising 500 individuals classified as non-diabetic and 268 classified as diabetic. Smith et al. (1988) used diabetes-related measurements in an early computational investigation, establishing an important historical connection between the dataset and the development of classification methods. Araveeporn (2022) subsequently applied discriminant analysis to the same benchmark dataset, demonstrating its continued use in evaluating statistical classification approaches. The dataset includes eight predictors: number of pregnancies, plasma glucose concentration, diastolic blood pressure, triceps skinfold thickness, serum insulin, body mass index, diabetes pedigree function and age. These variables represent different aspects of reproductive history, metabolic status, anthropometry and familial susceptibility, providing a basis for assessing how statistical methods classify diabetes outcomes using multiple clinical characteristics. Nevertheless, the dataset is restricted to women aged at least 21 years of Pima Indian heritage, which limits the extent to which results can be generalised to broader populations.

Although the Pima dataset is widely used for methodological evaluation, its application requires careful attention to data quality and analytical procedures. Clinical measurements may contain implausible values, incomplete observations or distributions that challenge the assumptions of particular statistical models. The documentation of the Pima dataset in the R statistical environment identifies its specific population characteristics and diagnostic classification purpose, reinforcing the need to interpret its variables within their original context. Collins et al. (2024) emphasised that transparent reporting of data preparation, model development and validation is essential for assessing the credibility of prediction research. This requirement is particularly relevant when comparing LR and LDA because differences in handling missing observations, transforming predictors or partitioning the dataset can influence the resulting estimates of model performance. Van Calster et al. (2019) additionally cautioned that model complexity and available sample size must be considered

during model development and validation. Therefore, a meaningful comparison of LR and LDA requires consistent preprocessing and evaluation procedures so that observed differences are attributable as far as possible to the classification methods rather than inconsistencies in the analytical process.

Previous investigations have established the relevance of statistical and computational techniques for diabetes classification, but they also demonstrate that the preferred method may depend on the evaluation criteria and dataset characteristics. The comparative diabetes classification study indexed in PubMed reported results for several classification approaches, including discriminant analysis, and demonstrated that different algorithms can produce different balances between sensitivity and specificity. Araveeporn (2022) further examined discriminant-based classification using the Pima dataset, providing evidence that the statistical structure of classification methods remains an important subject of investigation. However, neither the existence of increasingly sophisticated algorithms nor the favourable performance of a particular technique in an earlier investigation establishes its superiority under different analytical conditions. Collins et al. (2024) emphasised the importance of clearly specifying the prediction objective and evaluation procedures, while Van Calster et al. (2019) demonstrated why apparently favourable model performance requires careful interpretation. Accordingly, a focused comparison of LR and LDA remains methodologically relevant, particularly when both techniques are applied to the same dataset and assessed using consistent classification criteria.

Against this background, the present study comparatively evaluates Logistic Regression and Linear Discriminant Analysis for classifying diabetes status using the Pima Indians Diabetes Dataset. Specifically, it examines the performance of both methods using the available demographic and clinical predictors and compares their classification outcomes through accuracy, sensitivity, specificity, precision and other relevant performance indicators. The investigation is informed by the methodological concerns raised by Collins et al. (2024), particularly the need for transparent and reproducible prediction-model evaluation. It also responds to the importance of examining differences between statistical classification approaches demonstrated by Araveeporn (2022). By concentrating on two established and interpretable methods, the study seeks to determine whether their differing statistical assumptions and estimation procedures are associated with meaningful differences in diabetes classification performance. The findings are expected to contribute to methodological understanding of diabetes classification and provide evidence on the relative suitability of LR and LDA for analysing the selected dataset. However, because the dataset records diabetes classification rather than prospectively observed disease onset, the study evaluates classification of existing diabetes status and does not establish the ability of either model to predict future diabetes development

Literature Review

Diabetes classification has attracted considerable research interest following advances in statistical learning and the increasing availability of clinical datasets. Smith et al. (1988) demonstrated the early application of computational methods to diabetes-related measurements, establishing a foundation for subsequent classification studies. Maniruzzaman et al. (2017) subsequently compared several machine learning approaches to diabetes classification and demonstrated that predictive performance varies across algorithms. These findings suggest that selecting an appropriate classification method requires consideration of dataset characteristics, statistical assumptions and evaluation criteria rather

than computational complexity alone. Consequently, classical statistical techniques remain relevant, particularly when interpretability and reproducibility are important considerations in healthcare research.

Among established classification techniques, Logistic Regression (LR) remains widely applicable because it estimates the probability of a binary outcome and permits interpretation of the relationships between predictor variables and diabetes status. Linear Discriminant Analysis (LDA), conversely, classifies observations using estimated group distributions and discriminant functions. Although both techniques can produce linear classification boundaries, their statistical assumptions differ. LDA conventionally assumes multivariate normality within groups and equal covariance matrices, whereas LR does not require these distributional assumptions. Araveeporn (2022) demonstrated the continued application of discriminant analysis to diabetes classification, while Maniruzzaman et al. (2017) highlighted the importance of empirical comparisons among classification approaches. These methodological differences justify examining whether LR and LDA produce comparable results when applied to the same clinical dataset.

The Pima Indians Diabetes Dataset provides an established benchmark for evaluating these classification methods because it contains demographic and clinical measurements associated with diabetes status. Smith et al. (1988) established its historical relevance to computational diabetes classification, while Araveeporn (2022) demonstrated its continued application in discriminant-based modelling. Nevertheless, the dataset's restricted population and potentially invalid clinical measurements require careful analytical consideration. Collins et al. (2024) emphasised transparent reporting of data preprocessing, model development and validation, while Van Calster et al. (2019) demonstrated that discrimination and calibration represent distinct aspects of predictive performance. Accordingly, classification accuracy alone may provide an incomplete basis for determining model suitability.

Although previous investigations have evaluated several diabetes classification algorithms, the relative performance of LR and LDA remains dependent on the analytical conditions under which they are assessed. Maniruzzaman et al. (2017) demonstrated variation in performance across classification approaches, while Collins et al. (2024) emphasised consistent evaluation procedures for reliable model comparisons. Therefore, the present study comparatively evaluates LR and LDA using the Pima Indians Diabetes Dataset under a common analytical framework. By assessing accuracy, sensitivity, specificity, precision and related performance measures, the study seeks to establish their relative classification effectiveness and contribute evidence concerning the suitability of interpretable statistical techniques for diabetes classification.

Methods

Study design and dataset

The study uses a quantitative comparative classification design and the Pima Indians Diabetes Dataset, originally collected from women of Pima Indian heritage aged at least 21 years. The dataset contains 768 observations, eight predictors and one binary response. The predictors are Pregnancies, Glucose, Blood Pressure, Skin Thickness, Insulin, Body Mass Index (BMI), Diabetes Pedigree Function and Age. Outcome is coded 0 for non-diabetic and 1 for diabetic. The original dataset is

associated with the National Institute of Diabetes and Digestive and Kidney Diseases and is available through the UCI Machine Learning Repository and other public mirrors (Smith et al., 1988).

Data quality and zero-coded values

The source file encodes some missing or unrecorded measurements as zero. A value of zero is meaningful for Pregnancies and Outcome, but is physiologically implausible for Glucose, Blood Pressure, Skin Thickness, Insulin and BMI. The reported counts of zero values in these five measurement variables are 5, 35, 227, 374 and 11, respectively. Such zeros should be recoded as missing before model fitting. They should not be treated as true physiological measurements, and deleting every row containing one would discard a substantial portion of the data.

For a reproducible primary analysis, imputation parameters must be estimated from the training data only. A practical prespecified approach is median imputation for each affected predictor, with those training-set medians then applied unchanged to the test set. This avoids leakage of information from the test observations into model development. The supplied manuscript does not include a raw data file or executable analysis log; consequently, it cannot establish that this specific pipeline generated the previously reported confusion matrices. The numerical results retained below are therefore identified as reported outputs, not newly recomputed estimates.

Models

Logistic Regression Model

Logistic Regression was employed because the dependent variable is binary. The method estimates the probability that an individual belongs to the diabetic group based on the observed predictor variables.

Let:

p = probability that an individual is diabetic.

The Logistic Regression model is expressed as:

$$p = 1 / [1 + \exp\{-(\beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_5X_5 + \beta_6X_6 + \beta_7X_7 + \beta_8X_8)\}] \quad (1)$$

Where:

β_0 = intercept;

$\beta_1, \beta_2, \ldots, \beta_8$ = regression coefficients;

$X_1, X_2, \ldots, X_8$ = predictor variables.

Alternatively, the model can be expressed in log-odds form as:

$$\ln[p / (1 - p)] = \beta_0 + \beta_1X_1 + \beta_2X_2 + \beta_3X_3 + \beta_4X_4 + \beta_5X_5 + \beta_6X_6 + \beta_7X_7 + \beta_8X_8 \quad (2)$$

The Logistic Regression procedure followed this sequence:

Pima Dataset → Data Preprocessing → Training/Test Split → Fit Logistic Regression Model → Estimate Diabetes Probability → Apply Classification Threshold → Generate Predicted Class → Evaluate Performance

A probability threshold of 0.50 was used to convert the predicted probability into a binary class:

If $p \geq 0.50$ → Predicted outcome = 1 (Diabetic)

If $p < 0.50$ → Predicted outcome = 0 (Non-diabetic)

Thus, the LR model first produces an estimated probability of diabetes and subsequently converts that probability into a predicted class.

Linear Discriminant Analysis Model

Linear Discriminant Analysis was used as the second classification technique. LDA determines a linear combination of the predictor variables that provides maximum separation between the two outcome groups, namely diabetic and non-diabetic individuals.

The general linear discriminant function is expressed as:

$$D = a + b_1X_1 + b_2X_2 + b_3X_3 + b_4X_4 + b_5X_5 + b_6X_6 + b_7X_7 + b_8X_8 \quad (3)$$

Where:

D = discriminant score;

a = constant;

$b_1, b_2, \ldots, b_8$ = discriminant coefficients;

$X_1, X_2, \ldots, X_8$ = predictor variables.

Predictor variables include:

1. Pregnancies
2. Glucose level
3. Blood Pressure
4. Skin Thickness
5. Insulin
6. Body Mass Index (BMI)

7. Diabetes Pedigree Function
8. Age

For the two diabetes classes, LDA estimates the characteristics of each group and calculates the discriminant score of each observation. The observation is then assigned to the group with the higher discriminant value.

The LDA procedure therefore followed the sequence:

Pima Dataset → Data Preprocessing → Training/Test Split → Estimate Group Means and Covariance → Construct Discriminant Function → Calculate Discriminant Scores → Assign Observation to Class → Evaluate Performance

The two classification groups were:

Group 0 = Non-diabetic

Group 1 = Diabetic

The LDA model therefore differs from Logistic Regression in its classification mechanism. LR estimates the probability of belonging to the diabetic class, whereas LDA determines class membership by comparing the discriminant characteristics of the predefined groups.

Model Evaluation

For a fair direct comparison, both models will be evaluated on the same testing observations. The confusion matrices reproduced in Section 4.5 each contain 231 observations.

True Positive (TP): diabetic individual correctly classified as diabetic;

True Negative (TN): non-diabetic individual correctly classified as non-diabetic;

False Positive (FP): non-diabetic individual incorrectly classified as diabetic;

False Negative (FN): diabetic individual incorrectly classified as non-diabetic.

The following performance measures were calculated.

Accuracy

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (4)$$

Accuracy represents the proportion of all observations correctly classified by the model.

Sensitivity

$$\text{Sensitivity} = TP / (TP + FN) \quad (5)$$

Sensitivity measures the ability of the model to correctly identify individuals who are diabetic.

Specificity

Specificity = TN / (TN + FP) (6)

Specificity measures the ability of the model to correctly identify individuals who are not diabetic.

Precision

Precision = TP / (TP + FP) (7)

Precision represents the proportion of observations predicted to be diabetic that are actually diabetic.

F1-Score

F1-Score = 2 × (Precision × Sensitivity) / (Precision + Sensitivity) (8)

The F1-score provides a combined measure of precision and sensitivity.

Balanced Accuracy

Balanced accuracy is the arithmetic mean of sensitivity and specificity:

Balanced Accuracy = (Sensitivity + Specificity) / 2 (9)

Where predicted probabilities or posterior probabilities were available, the Receiver Operating Characteristic (ROC) curve and Area Under the Curve (AUC) were also used to assess the discriminatory ability of the two models.

The uncorrelated McNemar statistics is:

$$\chi^2 = \frac{(b-c)^2}{b+c} \quad (10)$$

The exact two-sided p-value of McNemar classification test is computed using:

$$p = 2\sum_{k=0}^{\min(b,c)} \binom{b+c}{k} \left(\frac{1}{2}\right)^{b+c} \quad (11)$$

Here, the binomial coefficient is computed as:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

(12)

Validation and reproducibility requirements

A controlled comparison of a stratified 70:30 training–testing split, a fixed random seed of 42 (recommended seed: Brownlee, 2026), and the same partition for LR and LDA were used in this work. Imputation was fitted on the training set and applied to the test set. Since cross-validation was reported, the same stratified folds were used for both models, with preprocessing repeated within each training fold. R software version R 3.2.2 was used in data analysis. The reproducible analysis code in Appendix A specifies a procedure that can be run on the original Pima_dataset.

Evaluation metrics

Performance was calculated from the same test observations for both models. Accuracy = (TP + TN)/(TP + TN + FP + FN); sensitivity (recall) = TP/(TP + FN); specificity = TN/(TN + FP); precision = TP/(TP + FP); F1 = 2 × precision × sensitivity/(precision + sensitivity); and balanced accuracy = (sensitivity + specificity)/2. ROC-AUC was calculated from continuous predicted probabilities or posterior probabilities, not from hard class labels. Sensitivity and false-negative counts are especially important for a screening context because a false negative may delay further assessment. Nevertheless, changing the classification threshold entails a trade-off between sensitivity and specificity. A threshold was selected based on the intended use and evaluated without tuning on the final test set. Calibration was also assessed if predicted probabilities are to be interpreted as risk estimates in this work.

Results

Data integrity and outcome distribution

The Pima_Indians_diabetes_Dataset.csv was imported directly for this analysis. One can download the Pima Indians Diabetes Dataset via the website link: *Pima Indians Diabetes Dataset on GitHub.com*. It contains 768 observations and nine columns: eight predictors and the binary Outcome variable. Outcome was coded 0 for non-diabetic and 1 for diabetic. There were 500 outcome-0 records (65.10%) and 268 outcome-1 records (34.90%); no rows were removed.

Dataset check	Result
Observations	768
Predictor variables	8
Outcome 0	500 (65.10%)
Outcome 1	268 (34.90%)
Duplicate rows	0
Missing values in original CSV	0

Zero-coded measurements and preprocessing

Zero values were retained for Pregnancies and Outcome because zero is meaningful for these variables. For Glucose, BloodPressure, SkinThickness, Insulin and BMI, zero values were recoded as missing before modelling because they are physiologically implausible as measured values in this dataset. The counts were 5, 35, 227, 374 and 11, respectively. Median imputation was fitted on the training partition only and applied to the test partition through a pipeline. Predictors were standardized in the same pipeline; the binary outcome was not transformed.

Variable recoded	Zero values recoded as missing
Glucose	5
BloodPressure	35
SkinThickness	227
Insulin	374
BMI	11

Primary held-out comparison

A single stratified 70:30 train/test split was used with random seed 42. The training partition contained 537 observations and the held-out test partition contained 231 observations (150 outcome-0 and 81 outcome-1). Both classifiers were trained and evaluated on exactly the same rows, with median imputation and standardization estimated within each training pipeline. Logistic Regression used a 0.50 probability threshold; LDA used its default posterior-probability classification rule. The metrics below are computed values from the original Pima_Indians_diabetes_Dataset.csv, using R analytical package version R.3.3.2 (See Appendix A & B for the reproducible R source codes, analysis environment and provenance details of Python; pandas 2.2.3, scikit-learn 1.8.0, and a fixed random seed setting of 42).

Metric	Logistic Regression	Linear Discriminant Analysis
Test observations	231	231
TN	129	129
FP	21	21
FN	38	40
TP	43	41
Accuracy	74.46%	73.59%
Sensitivity / Recall	53.09%	50.62%
Specificity	86.00%	86.00%
Precision	67.19%	66.13%
F1-score	59.31%	57.34%
Balanced accuracy	69.54%	68.31%
ROC-AUC	0.836	0.835

Ten-fold stratified cross-validation

As a complementary stability assessment, stratified 10-fold cross-validation with shuffling and random seed=42 was performed on all 768 observations. The imputation and scaling steps were refitted within each fold to avoid leakage. Entries are mean fold scores; values in parentheses are the standard deviation across the ten folds.

Metric (mean; SD)	Logistic Regression	Linear Discriminant Analysis
Accuracy	76.68% (2.67)	76.42% (2.66)
Sensitivity / Recall	55.63% (9.08)	55.64% (7.16)
Precision	72.64% (8.73)	72.35% (9.05)
F1-score	62.26% (4.80)	62.16% (2.71)
Balanced accuracy	71.81% (3.05)	71.62% (1.81)
ROC-AUC	0.837 (0.052)	0.837 (0.052)

McNemar test: paired classification results

The following contingency table summarizes the correctness of both models on the same 231 held-out observations.

	LDA correct	**LDA incorrect**	**Total**
LR correct	170	2	172
LR incorrect	0	59	59
Total	170	61	231

The two discordant cells are the critical components of McNemar's test:

b=2: LR was correct while LDA was incorrect.

c=0: LR was incorrect while LDA was correct.

The 170 observations classified correctly by both models and the 59 classified incorrectly by both models do not contribute directly to the McNemar test statistic.

Hypotheses

H0: LR and LDA have equal marginal probabilities of correct classification on the held-out test population.

H1: LR and LDA have different marginal probabilities of correct classification on the held-out test population.

From the analysis output, the exact McNemar test yielded (p=0.500), which exceeds the 5% significance level. Therefore, the null hypothesis is not rejected. Although LR correctly classified two observations that LDA misclassified, there were no observations on which LDA was correct and LR was incorrect. Nevertheless, the number of discordant predictions is too small to establish a statistically significant difference between the models. This result does not prove that LR and LDA perform identically. Rather, the test does not provide sufficient evidence of a difference in their marginal classification performance on the held-out test set.

Interpretation

On the fixed held-out partition, Logistic Regression achieved 74.46% accuracy and LDA achieved 73.59%, a difference of 0.87 percentage points. Both models had the same specificity (86.00%); Logistic Regression had modestly higher sensitivity (53.09% versus 50.62%), F1-score (59.31% versus 57.34%) and ROC-AUC (0.836 versus 0.835). The models produced very similar ranking performance by ROC-AUC. In the paired held-out predictions, two observations were classified correctly by LR but incorrectly by LDA, and none showed the reverse pattern; an exact McNemar test was not statistically significant (p = 0.500). Thus, this single split does not establish a statistically reliable superiority of either classifier.

Ten-fold cross-validation showed mean accuracy of 76.68% for Logistic Regression and 76.42% for LDA. Mean ROC-AUC was 0.837 for LR and 0.837 for LDA (0.8367 and 0.8373 before rounding). The small and metric-dependent differences indicate broadly comparable performance under this specified pipeline. Cross-validation results are reported separately from the held-out results because they estimate performance using a different evaluation design.

Discussion

The present study examined the comparative classification performance of Logistic Regression (LR) and Linear Discriminant Analysis (LDA) using the Pima Indians Diabetes Dataset under a common analytical framework. The findings revealed broadly comparable performance between the two statistical techniques, although minor differences were observed across evaluation measures. Logistic Regression achieved slightly higher accuracy, sensitivity, F1-score and area under the receiver operating characteristic curve (AUC) on the held-out test dataset. However, ten-fold cross-validation produced a different pattern, with LR recording marginally higher mean accuracy and LDA demonstrating marginally higher mean AUC. These findings are consistent with Maniruzzaman et al. (2017), who demonstrated that diabetes classification performance varies across statistical and machine learning approaches. The differences observed in the present study suggest that neither technique consistently outperformed the other across all evaluation measures. This reinforces the importance of examining multiple performance indicators before identifying a preferred classification model.

The relatively similar performance of LR and LDA is also important when considered against their different statistical assumptions. Logistic Regression estimates diabetes classification probabilities directly, whereas LDA derives classification functions from the estimated distributions of predictors within the outcome groups. Araveeporn (2022), in a study comparing discriminant analysis approaches for diabetes classification, demonstrated the continued relevance of discriminant-based

methods for analysing the Pima dataset. The present findings extend this methodological perspective by showing that LR and LDA can produce closely comparable classification outcomes when evaluated using the same dataset and preprocessing procedures. Their similarities suggest that the additional distributional assumptions associated with LDA did not produce a substantial observed disadvantage under the conditions examined. Nevertheless, this does not establish that those assumptions were satisfied or that the two methods would perform equally well in other populations or datasets.

An important finding was that both classification techniques failed to identify a substantial proportion of diabetes-positive observations under their default classification rules. Although LR demonstrated slightly higher sensitivity on the held-out test dataset, the difference did not eliminate the occurrence of false-negative classifications. This finding has particular relevance because the consequences of incorrectly classifying an individual with diabetes as non-diabetic may be more serious than those associated with an unnecessary referral for further assessment. Van Calster et al. (2019) emphasised that predictive performance should be evaluated beyond a single summary measure, particularly when model outputs may inform clinical decisions. Similarly, Collins et al. (2024) highlighted the importance of reporting performance measures appropriate to a prediction model's intended application. The present findings therefore demonstrate that favourable overall accuracy does not necessarily indicate adequate identification of diabetes-positive cases. Adjusting classification thresholds and examining the resulting sensitivity–specificity trade-offs may provide further information about the suitability of these models for different analytical purposes.

The differences between held-out test performance and ten-fold cross-validation results also deserve consideration. While LR recorded marginally higher AUC on the selected test split, LDA achieved marginally higher mean AUC across cross-validation folds. This variation indicates that the relative ranking of the classifiers depended partly on the evaluation procedure and performance measure. Collins et al. (2024) emphasised transparent reporting of model development and validation procedures to support reproducibility and appropriate interpretation of prediction research. The present study addressed this concern by applying both classifiers to the same 768-observation dataset, using a common preprocessing pipeline and reporting held-out and cross-validation findings separately. Nevertheless, the small observed differences do not establish statistically significant superiority because no formal paired comparison or uncertainty estimates for the differences have been reported. The findings therefore support comparable classification performance rather than a definitive preference for either method.

The interpretation of these results must also consider the characteristics of the dataset. The Pima Indians Diabetes Dataset comprises women of Pima Indian heritage aged at least 21 years and does not represent the demographic or clinical diversity of contemporary populations. Smith et al. (1988) established the dataset's historical importance in computational diabetes classification, while Collins et al. (2024) emphasised the need to evaluate the applicability of prediction models to their intended populations. Although the present study applied specified procedures for handling zero-coded clinical measurements, preprocessing cannot eliminate the limitations associated with the dataset's restricted population and historical context. Consequently, the findings should not be generalised directly to men, younger individuals or populations with substantially different characteristics. Furthermore, because the outcome represents recorded diabetes status rather than prospectively observed disease onset, the analysis should be interpreted as diabetes classification rather than prediction of future disease development.

The findings demonstrate that LR and LDA remain relevant statistical approaches for diabetes classification, with neither method showing consistent superiority across the evaluation procedures employed. The principal contribution of the study lies in its direct comparison of two interpretable classifiers under common analytical conditions, while distinguishing held-out performance from cross-validation results. In agreement with the methodological considerations raised by Maniruzzaman et al. (2017) and Collins et al. (2024), the results indicate that model selection should be informed by multiple performance measures, validation procedures and the intended application rather than marginal differences in accuracy alone. Further investigations using larger, more diverse clinical datasets, external validation and clinically justified classification thresholds would be necessary before either technique could be recommended for practical diabetes screening or diagnostic support.

Conclusion

The study demonstrated that Logistic Regression (LR) and Linear Discriminant Analysis (LDA) achieved comparable performance in classifying diabetes status using the Pima Indians Diabetes Dataset comprising 768 observations. Logistic Regression recorded a slightly higher held-out classification accuracy (74.46%) than LDA (73.59%), while their respective ROC-AUC values of 0.836 and 0.835 indicated almost identical discriminatory performance. Similarly, stratified ten-fold cross-validation revealed only marginal differences between the two approaches, suggesting that neither method consistently demonstrated a meaningful predictive advantage under the specified analytical conditions. These findings indicate that both techniques remain useful for diabetes classification, particularly where interpretability and computational simplicity are important considerations. However, the small observed differences do not establish the superiority of either classifier. The study therefore concludes that model selection should consider classification performance, statistical assumptions, reproducibility and intended application rather than minor differences in accuracy. Further evaluation using independent clinical datasets is necessary before extending these findings beyond the benchmark population.

Recommendations

Based on the findings, future diabetes classification studies should evaluate Logistic Regression and Linear Discriminant Analysis using consistent preprocessing and validation procedures to ensure meaningful comparisons. Given the marginal performance differences observed, researchers should employ repeated stratified cross-validation and appropriate paired statistical comparisons to assess the stability and uncertainty of model performance. The occurrence of false-negative classifications further suggests the need to examine alternative classification thresholds, sensitivity–specificity trade-offs and probability calibration, particularly where models are intended to support screening decisions. Researchers should also ensure that missing-value imputation and other preprocessing procedures are fitted exclusively to training data to minimise information leakage. Furthermore, future investigations should consider alternative imputation strategies and evaluate model performance across relevant demographic and clinical subgroups. Because the Pima dataset represents a restricted historical population, external validation using larger, contemporary and more diverse clinical datasets is recommended before either model is considered for clinical application.

Limitations

The findings should be interpreted in relation to several methodological limitations. First, the held-out evaluation was based on a single random partition, making its performance estimates potentially sensitive to the particular observations allocated to the training and testing subsets. Although stratified ten-fold cross-validation provided additional evidence of model stability, it did not eliminate uncertainty associated with the dataset or establish external validity. Second, the use of median imputation for zero-coded clinical measurements represented a transparent preprocessing approach, but alternative imputation techniques might produce different classification outcomes. Furthermore, the non-significant exact McNemar test, together with the limited number of discordant predictions, restricted the ability to establish meaningful differences in the classifiers' error patterns. The study also did not evaluate probability calibration, clinical decision curves or classification performance across demographic subgroups. Finally, the Pima Indians Diabetes Dataset comprises a historically specific population of adult women, limiting the generalisability of the findings to other populations and contemporary clinical settings. These limitations indicate that the results should be interpreted as a methodological comparison rather than evidence of clinical diagnostic readiness.

References

Abnoosian, K., Farnoosh, R., & Behzadi, M. H. (2023). Prediction of diabetes disease using an ensemble of machine learning multi-classifier models. BMC Bioinformatics, 24, Article 337.

Ahmed, N., Ahammed, R., Islam, M. M., Uddin, M. A., Akhter, A., Talukder, M. A., & Paul, B. K. (2021). Machine learning based diabetes prediction and development of smart web application. International Journal of Cognitive Computing in Engineering, 2, 229–241.

Alzboon, M. S., Al-Batah, M. S., Alqaraleh, M., Abuashour, A., & Hamadah Bader, A. F. (2023). Early diagnosis of diabetes: A comparison of machine learning methods. International Journal of Online and Biomedical Engineering, 19(15), 144–165.

Araveeporn, A. (2022). Comparing the linear and quadratic discriminant analysis of diabetes disease classification based on data multicollinearity. International Journal of Mathematics and Mathematical Sciences, 2022, Article 7829795.

Brownlee, J. (n.d.). Pima Indians diabetes dataset [Data set]. GitHub.

Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A.-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., ... Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. BMJ, 385, e078378.

International Diabetes Federation. (2025). IDF diabetes atlas (11th ed.).

Lawrence, M. O., Ibrahim, A. M., Awotunde, J. B., Abiodun, M. K., Alfa, P. E., & Adeniyi, A. A. (2025). Development and implementation of a diabetes prediction system using logistic regression. NIPES Journal of Science and Technology Research, 7(Special Issue), 384–392.

Maniruzzaman, M., Kumar, N., Abedin, M. M., Islam, M. S., Suri, H. S., El-Baz, A. S., & Suri, J. S. (2017). Comparative approaches for classification of diabetes mellitus data: Machine learning paradigm. Computer Methods and Programs in Biomedicine, 152, 23–34.

Nie, F., Song, X., & Chen, W. (2025). DRPM: An advanced predictive model for early diabetes detection and risk stratification. Molecular Therapy – Nucleic Acids, 36(3), Article 102576.

Ranathungage, T. D., Blumer, H., & Muthukumarana, S. (2026). Flexible statistical approaches for modeling nonlinear relationships in diabetes prediction using splines, Bayesian kernel regression and Bayesian regression trees. Discover Analytics, 4, Article 10.

Saleh, H. M. (2024). A comprehensive review of data mining techniques for diabetes diagnosis using the Pima Indian Diabetes Dataset. EDRAAK, 2024, 39–42.

Smith, J. W., Everhart, J. E., Dickson, W. C., Knowler, W. C., & Johannes, R. S. (1988). Using the ADAP learning algorithm to forecast the onset of diabetes mellitus. Proceedings of the Annual Symposium on Computer Application in Medical Care, 261–265.

Sneha, N., & Gangil, T. (2019). Analysis of diabetes mellitus for early prediction using optimal features selection. Journal of Big Data, 6, Article 13.

UCI Machine Learning Repository. (n.d.). Pima Indians Diabetes Dataset [Data set].

Van Calster, B., McLernon, D. J., van Smeden, M., Wynants, L., & Steyerberg, E. W. (2019). Calibration: The Achilles heel of predictive analytics. BMC Medicine, 17, Article 230.

World Health Organization. (2024). Diabetes.

Zhou, H., Xin, Y., & Li, S. (2023). A diabetes prediction model based on Boruta feature selection and ensemble learning. BMC Bioinformatics, 24, Article 224.

Appendix A

R workflow for the LR_LDA comparison:

The following R code reproduces the stratified split, zero recoding, training-only median imputation, standardization, LR/LDA fit and held-out metrics. It assumes the imported Pima_Indians_Diabetes_Dataset contains the nine named columns shown in the paper.

```
# Reproducible primary analysis: LR versus LDA

# R packages: caret, recipes, MASS, pROC, yardstick
```

```
library(caret)

library(recipes)

library(MASS)

library(pROC)

dat <- read.csv("Pima_Indians_diabetes_Dataset.csv")

stopifnot(nrow(dat) == 768, ncol(dat) == 9)

dat$Outcome <- factor(dat$Outcome, levels = c(0, 1),

                      labels = c("NoDiabetes", "Diabetes"))

set.seed(42)

idx <- createDataPartition(dat$Outcome, p = 0.70, list = FALSE)

train <- dat[idx, ]; test <- dat[-idx, ]

zero_vars <- c("Glucose", "BloodPressure", "SkinThickness", "Insulin", "BMI")

for (v in zero_vars) { train[[v]][train[[v]] == 0] <- NA test[[v]][test[[v]] == 0] <- NA}

# Train-only medians and scaling parameters

med <- sapply(train[zero_vars], median, na.rm = TRUE)

for (v in zero_vars) {train[[v]][is.na(train[[v]])] <- med[[v]]

  test[[v]][is.na(test[[v]])] <- med[[v]]}

xvars <- setdiff(names(train), "Outcome")

means <- sapply(train[xvars], mean)

sds <- sapply(train[xvars], sd)

train[xvars] <- scale(train[xvars], center = means, scale = sds)

test[xvars] <- scale(test[xvars], center = means, scale = sds)


lr <- glm(Outcome ~ ., data = train, family = binomial())

lr_prob <- predict(lr, newdata = test, type = "response")
```

```
lr_pred <- factor(ifelse(lr_prob >= 0.5, "Diabetes", "NoDiabetes"),
                  levels = levels(test$Outcome))
lda_fit <- MASS::lda(Outcome ~ ., data = train)
lda_pred <- predict(lda_fit, newdata = test)
lda_prob <- lda_pred$posterior[, "Diabetes"]
lda_class <- factor(ifelse(lda_prob >= 0.5, "Diabetes", "NoDiabetes"),
                    levels = levels(test$Outcome))
metrics <- function(truth, pred, prob) {
  cm <- caret::confusionMatrix(pred, truth, positive = "Diabetes")
  c(Accuracy = unname(cm$overall["Accuracy"]),
    Sensitivity = unname(cm$byClass["Sensitivity"]),
    Specificity = unname(cm$byClass["Specificity"]),
    Precision = unname(cm$byClass["Pos Pred Value"]),
    F1 = unname(cm$byClass["F1"]),
    AUC = as.numeric(pROC::auc(truth, prob,
       levels = c("NoDiabetes", "Diabetes"), direction = "<")))}rbind(LogisticRegression =
metrics(test$Outcome, lr_pred, lr_prob),
      LDA = metrics(test$Outcome, lda_class, lda_prob))
# For stratified 10-fold CV, use caret::train with method="cv",
# number=10, classProbs=TRUE and summaryFunction=twoClassSummary;
# ensure all preprocessing is performed within each resampling fold.
```

We can also use the alternative source code below for the LR_LDA classification analysis:

Reproducible R workflow for the primary comparison

The following code specifies a reproducible 70:30 stratified split, training-only median imputation, LR and LDA fitting, a 0.50 decision threshold and same-test-set metrics. Save the original dataset as

pima-indians-diabetes.csv with the eight predictors followed by Outcome coded 0/1. This script is supplied as a reproducibility template.

```
# Reproducible LR-versus-LDA comparison
# R packages: MASS, caret, pROC
set.seed(2026)
library(MASS)
library(caret)
library(pROC)

d <- read.csv("pima-indians-diabetes.csv", header = FALSE)
names(d) <- c("Pregnancies", "Glucose", "BloodPressure", "SkinThickness",
              "Insulin", "BMI", "DiabetesPedigreeFunction", "Age", "Outcome")
d$Outcome <- factor(d$Outcome, levels = c(0, 1))

# Only these zero values are recoded as missing; Pregnancy and Outcome zeros remain valid.
zero_missing <- c("Glucose", "BloodPressure", "SkinThickness", "Insulin", "BMI")
for (v in zero_missing) d[[v]][d[[v]] == 0] <- NA

# Stratified 70:30 split. Same rows are used for both models.
set.seed(2026)
idx <- createDataPartition(d$Outcome, p = 0.70, list = FALSE)
train <- d[idx, ]
test  <- d[-idx, ]

# Fit imputation medians on training data only.
pred <- setdiff(names(d), "Outcome")
medians <- sapply(train[pred], function(x) median(x, na.rm = TRUE))
for (v in pred) {
  train[[v]][is.na(train[[v]])] <- medians[v]
  test[[v]][is.na(test[[v]])]   <- medians[v]}

lr <- glm(Outcome ~ ., data = train, family = binomial())
lda <- MASS::lda(Outcome ~ ., data = train)

p_lr <- predict(lr, newdata = test, type = "response")
p_lda <- predict(lda, newdata = test)$posterior[, "1"]
y <- test$Outcome
pred_lr <- factor(ifelse(p_lr >= 0.50, 1, 0), levels = c(0, 1))
pred_lda <- factor(ifelse(p_lda >= 0.50, 1, 0), levels = c(0, 1))
```

```
  metrics <- function(y, pred, prob) {
    cm <- table(Actual = y, Predicted = pred)
    TN <- cm["0","0"]; FP <- cm["0","1"]
    FN <- cm["1","0"]; TP <- cm["1","1"]
    sens <- TP / (TP + FN); spec <- TN / (TN + FP)
    prec <- TP / (TP + FP)
    f1 <- 2 * prec * sens / (prec + sens)
    c(N = length(y), Accuracy = (TP + TN) / length(y),
      Sensitivity = sens, Specificity = spec, Precision = prec,
      F1 = f1, BalancedAccuracy = (sens + spec) / 2,
      AUC = as.numeric(pROC::auc(pROC::roc(y, prob, levels = c("0","1"),
 direction = "<", quiet = TRUE))))}
  print(table(Actual = y, Predicted = pred_lr))
  print(table(Actual = y, Predicted = pred_lda))
  print(rbind(LogisticRegression = metrics(y, pred_lr, p_lr),
              LinearDiscriminantAnalysis = metrics(y, pred_lda, p_lda)))
```

APPENDIX B

Analysis environment and provenance

Dataset analyzed: Pima_Indians_diabetes_Dataset.csv; rows = 768, columns = 9. R or Python analysis; pandas 2.2.3, scikit-learn 1.8.0, and a fixed random seed of 42. Results in Sections 4.1–4.5 were generated in this reanalysis.