IRAQI Academic Scientific Journals

KHAZAYIN OF ECONOMIC AND ADMINISTRATIVE SCIENCES (2025) (02) (04): P(24-37)

خزائن للعلوم الاقتصادية والإدارية
KHAZAYIN OF ECONOMIC AND ADMINISTRATIVE SCIENCES
ISSN: 2960-1363 (Print)
ISSN: 3007-9020 (Online)

Comparison of linear discriminant analysis and binary logistic regression analysis in predicting Alzheimer's disease

Hadeel Imad Naser
College of Administration and Economics-Kirkuk University, Kirkuk, Iraq
hadeel.imad@uokirkuk.edu.iq

Abstract. Alzheimer's disease is a neurological disease caused by several factors affecting brain functions leading to memory loss, which in turn impacts the lives of those affected and those around them, often leading to the death of many elderly people. This study aimed to diagnose this disease and predict whether a person is affected by it. For this purpose, binary logistic regression and linear discriminant analysis were used and compared to each other to arrive at a result that clarifies which of the methods was the best and most powerful in classifying and diagnosing the disease. These methods were applied to data on reached (2149) Alzheimer's patients that were obtained from kaggle. When analyzed using the statistical program SPSS v.25, the results showed the efficiency of both methods in classification and in predicting the affiliation of group elements, as the binary logistic regression analysis method was able to correctly classify (84.8%) while it was (83.2%) for the linear discriminant analysis. As for the comparison between the two models, the results showed that the sensitivity, specificity and AUC of the linear discriminant analysis reached (83.4, 83% and 83.2%) and (74.6%, 90.4% and 82.5%) for the binary logistic regression analysis, which making the LDA model the best for reducing misdiagnosis in patients and the BLR model the best for avoiding misdiagnosis in uninfected people.

Keywords: Linear discriminant analysis, Binary logistic regression and Alzheimer's disease.

DOI: 10.69938/Keas.2502043

مقارنة بين التحليل التمييزي الخطي وتحليل الانحدار اللوجستي الثنائي في التنبؤ بمرض الزهايمر

م.م. هديل عماد ناصر
كلية الادارة والاقتصاد-جامعة كركوك، كركوك، العراق

المستخلص. مرض الزهايمر مرض عصبي ناجم عن عدة عوامل التي تؤثر على وظائف الدماغ مما يؤدي إلى فقدان الذاكرة، وهذا بدوره يؤثر على حياة المصابين به ومن حولهم، والذي غالباً ما يؤدي إلى وفاة العديد من كبار السن. هدفت هذه الدراسة إلى تشخيص هذا المرض والتنبؤ بإصابة الشخص به. ولهذا الغرض، استُخدم أسلوبي الانحدار اللوجستي الثنائي والتحليل التمييزي الخطي ومقارنتهما معاً للوصول إلى نتيجة توضح أيهما الأفضل والأكثر فعالية في تصنيف وتشخيص المرض. طُبقت هذه الأساليب على بيانات بلغت (2149) مريضاً بالزهايمر تم الحصول عليها من منصة Kaggle. عند التحليل باستعمال البرنامج الإحصائي SPSS v.25، أظهرت النتائج كفاءة كِلا الاسلوبين

في التصنيف والتنبؤ بإنتماء عناصر المجموعة، حيث استطاع اسلوب تحليل الانحدار اللوجستي الثنائي تصنيف المرض بشكل صحيح بنسبة (84.8%) بينما بلغت نسبة دقة التحليل التمييزي الخطي (83.2%). أما بالنسبة للمقارنة بين النماذج فقد أظهرت النتائج أن حساسية وخصوصية و AUC للتحليل التمييزي الخطي بلغت (83.4%، 83% و83.2%) و(74.6%، 90.4% و82.5%) لتحليل الانحدار اللوجستي الثنائي، وهذا ما يجعل نموذج LDA هو الأفضل للحد من التشخيص الخاطئ لدى المرضى ونموذج BLR هو الأفضل لتجنب التشخيص الخاطئ لدى الأشخاص غير المصابين.

الكلمات المفتاحية: التحليل التمييزي الخطي، الانحدار اللوجستي الثنائي ومرض الزهايمر.

Corresponding Author: E-mail: hadeel.imad@uokirkuk.edu.iq

1.Introduction:

Alzheimer's disease is one of the diseases of our time that mostly affects the elderly. It is a neurological disease that occurs when there is atrophy or loss of tissue throughout the brain, in addition to other causes, which leads to a loss of cognitive ability and memory in those afflicted with the disease. As a result of this disease, those afflicted and the people who live with them are affected. The incidence of the disease is increasing rapidly, as statistics show that there are (47) million Alzheimer's patients in (2020), which is expected to approach (75) million people in (2030). This increase poses a challenge to health and social care systems. Therefore, it is necessary to conduct an examination for early diagnosis of the disease, as early diagnosis and prevention are important, and treatment reduces the burden on society (Salunkhe et al., 2021).

There were Several studies that dealt with this disease, which several different methods and techniques have been used to diagnose and predict the disease occurrence.

In (2017), (Benyoussef et al.) propose a model composed by logistic regression, discriminant analysis and decision tree to classify and prediction Alzheimer's disease patients. They found that the accuracy of the decision tree, logistic regression and discriminant analysis models in performance to classify and prediction the disease was (60%, 70%, and 66%) for OASIS dataset (Benyoussef et al., 2017).

In (2018), (Yu et al.) applied logistic regression and random forest analysis to develop a suitable algorithm for detecting Alzheimer's disease in a sample of (156) control patients. They found that the random forest method generated a more robust predictive model than the logistic regression method and the eight-protein-based algorithm was the strongest where the sensitivity measure reached (97.7%), the specificity (88.6%) and the area under the curve (99%) (Yu et al., 2018).

In 2020, (Vidushi et al.) used a different machine learning algorithm on oasis_longitudinal MRI data, trained different machine learning models in addition used (binary logistic regression classifier, hierarchical decision tree, support vector machine, and ensemble random forest and boosting adaboost) to early detection and effective diagnosis the disease. They finally concluded that random forests and (adaboost) achieved the highest accuracy in diagnosing the disease (Vidushi et al., 2020).

In the same year, (Baglat et al.) used the techniques of logistic regression, decision tree, random forest classifier, adaboost and (SVM) for early diagnosis and classification of disease (using the open access series of imaging studies (OASIS) dataset). They found that the performance of random forest and adaptive boosting classifier was the strongest with accuracy reaching (86%) for disease diagnosis, while (78%) from logistic regression and (81%) from SVM, decision tree (Baglat et al., 2020).

(Alroobaea et al.) applied some supervised machine learning techniques in (2021) on the data obtained "from the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the Open Access Series of Imaging Studies (OASIS) from Brain Datasets" to detect Alzheimer's disease, the researchers found that the logistic regression classifiers and the support vector machine gave the highest accuracy of (99.43%) and (99.10%) for the "ADNI data set", accuracy rates of (84.33%) and (83.92%) were obtained by the logistic regression classifier and random forest for the "OASIS dataset" (Alroobaea & Al, 2021).

In the same year ,(Said et al.) used some of machine learning techniques (LR, LDA, KNN, CART, GNB, SVM) on s- parameter data obtained from 6 antennas that were placed around the patient's head to noninvasively capture changes in the brain. The data was processed by machine learning algorithms, they created the prediction score and accuracy of each algorithm and compared them to

determine which algorithm that can be used to classify the stages of the disease, they concluded that the LR model obtained the best accuracy of (98.97%) and efficiency in distinguishing between 4 different stages of the disease (Saied et al., 2021).
In their study (Mabrouk et al.) in (2023), sought to develop a four-way approach for multi-class disease diagnosis using linear discriminant analysis on data (available from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database). They also proposed a novel approach that (optimal LDA subspace was combined with the Pearson's correlation coefficient (PCC) method to overcome the singularity problem). They concluded that the proposed method offers promising performance compared to traditional methods (Mabrouk et al., 2023).
In 2024, (Ahmadi et al.) used a hybrid of 12 machine learning methods to diagnose the severity disease. 6 traditional machine learning methods were applied, namely (linear discriminant analysis, decision tree, naive bayes, k-nearest neighbor, support vector machine and ensemble learning methods), After optimization and training the CNN model using the machine learning algorithm to identify specific patterns, they concluded that the accuracy of the algorithms was, respectively (Naïve Bayes, Support Vector Machines, K-nearest neighbor, Linear discrimination classifier, Decision tree, Ensembled learning, and presented CNN architecture) are (67.5%, 72.3%, 74.5%, 65.6%, 62.4%, 73.8% and, 95.3%),so CNN the best to diagnose Alzheimer severity (Ahmadi et al., 2024).
In the same year, (Liau et al.) used ordinal logistic regression to diagnose and compare longitudinal changes in the use prevalence of symptomatic and preventive medications between those with and without the disease for a sample (58496) over the five years before and after diagnosis. It showed differences in symptoms ($p < 0.001$) and preventive medications ($p = 0.006$) at and after diagnosis (Liau et al., 2024).
In 2025, (Demir & Selvitopi) used several machine learning models were used on the" Open Access Imaging Studies Series (OASIS) data" to predict Alzheimer's disease. The models were evaluated using (precision, F1 score, and recall). The researchers found that the models demonstrated their efficiency, achieving an average accuracy of (85%). The highest classification accuracy was achieved by the RF model (0.8684). The ANN model outperformed the model with a recall rate of (0.9166) in identifying positive cases of the disease, while the DT model achieved a low recall rate. The SVM model provided balanced precision and recall while the LR model performed slightly in precision and good in recall (Demir & Selvitopi, 2025).
To determine the factors that affect Alzheimer's disease and to obtain the best diagnosis and prediction of the disease in the future, this study used two machine learning classifiers, (the linear discriminant analysis and the binary logistic regression analysis), and compared them to obtain the most powerful model for classifying and predicting the disease. This study is divided into four sections, the first includes an introduction and literature review in the area of Alzheimer's disease, the second section includes the used methodology, the third presents data collection and Result and discussion, and finally, section four include a conclusion.

2. Materials and Methods

2.1 Linear Discriminant Analysis (LDA):

It is considered one of the methods used in analyzing multivariate data. It is called discriminant analysis because of its ability to distinguish or separate two or more groups of observations, as well as its ability to classify new elements according to previously defined groups. This is done by constructing discriminant functions whose purpose is to distinguish between categories of the dependent variable, which is a descriptive variable. It is also distinguished by its ability to predict group membership based on a linear combination of quantitative variables. The functions maximize the differences between groups and minimize the variance within each group (Elgohari, 2017),(PREMPEH, 2009). Discriminant analysis has several assumptions, the most important of which are the distribution of the independent variables is normal, there is no strong correlation

between them and that the variance matrix is homogeneous (Polat, 2018),(Mahmood & Khider, 2023).

The discriminant analysis function can be written as:

$$\hat{L} = b_1 x_1 + b_2 x_2 + \cdots + b_k x_k \qquad (1)$$

Where $\hat{L}$ is the discriminatory degree, b_k are the coefficients of the discriminative model that are used in the classification process, k is the number of variables and x is the vector of variables.

$b = v^{-1}d$ where ,v^{-1} is inverse of the variance and covariance matrix, d is the distance matrix between the average of the variables in each of the two groups, b is calculated as:

$$\begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_k \end{bmatrix} = \begin{bmatrix} v_{11} & v_{12} & \dots & v_{1k} \\ v_{21} & v_{22} & & v_{2k} \\ \vdots & \vdots & & \vdots \\ v_{k1} & v_{k2} & \dots & v_{kk} \end{bmatrix} \begin{bmatrix} d_1 \\ d_2 \\ \vdots \\ d_k \end{bmatrix} \qquad (2)$$

(Nwanamidwa, 2018),(Matrood & Al-quraishi, 2023).

The relative importance of the factors influencing the process of discrimination and separation between groups is calculated after the stage of constructing the discriminant function, where the standard coefficients with the highest value are considered the most important. The percentage contribution of the variable to the discrimination process is calculated by the canonical correlation coefficient. The function's ability to discriminate is tested through several tests, including Wilk's lambda statistic which is calculated using the formula : $W = \frac{|A|}{|A+B|}$ where "A: the variance and covariance matrix within the groups, B: the variance and covariance matrix between groups". The value of Wilk's lambda lies between (0,1), if the value is equal to 1 this means that the averages for the two groups are equal, therefore the function is unable to distinguish. If its value is equal to 0, it means that the averages for the two groups are not equal, thus the function is able to distinguish (AlKhayyat & Alhaqbani, 2021),(PREMPEH, 2009),(Matrood & Al-quraishi, 2023).

2.2 Binary Logistic Regression (BLR):

It is concerned with studying the relationship of the influence of several independent variables, which may be quantitative or descriptive variables, on the dependent variable, which is a descriptive variable and may have only two values. In this case, the model is called binary logistic regression. If it has more than one value, it is called a multiple logistic regression model. The logistic regression model aims to classify and predict the probability of a positive outcome. It is distinguished by the fact that it does not require many assumptions, especially since it does not require the data to be normally distributed , It requires that the variables be independent of each other and that the relationship between them be linear. The dependent variable in binary logistic regression follows the Bernoulli distribution. with probability p (probability of success = 1) and q (probability of failure = 0). The logistic regression equation can be written as:

$$p(x) = \frac{1}{1 + e^{-(b_0 + \sum b_j x_{ij})}} \qquad (3)$$

Where $b_0, , b_j$ are the estimated model parameters, e is the natural logarithm, x_{ij} is the independent variables and p(x) is the dependent variable (Hilbe, 2015),(PENG et al., 2002).

By using the maximum likelihood method, the model parameters can be estimated, the method depends on repeating the calculations in order to reach the best estimate (Kleinbaum & Klein, 2010). To find out whether the independent variables in the model are significantly related to the dependent variable (i.e., verifying the model's suitability), there are several measures that can be used including the pseudo test and R^2 statistic which is calculated:

$$R_L^2 = -2\ \log(L_0) - \frac{[-2\ \log(L_m)]}{-2\log(L_0)} \qquad (4)$$

Where L_0 represents the log likelihood when including the fixed term in the model (i.e. the model does not contain independent variables), L_m represents the log likelihood of the estimated model (i.e. including independent variables in the model) (Domínguez-Almendros et al., 2011),(C.Pampel, 2021). We can determine the extent to which the data matches the model (i.e. The extent to which the expected values match the variable dependent on the observed values) by the Hosmer and Lemeshow test which depends on grouping the sample cases based on the percentiles of the expected probabilities, the cases are distributed into n after arranging them in ascending order according to the expected values of the probabilities, the observed and expected values of the cases are grouped according to the two values of y ,then the test value is calculated according to the Pearson x^2statistic from Table 2xg. The Hosmer and Lemeshow statistic calculated:

$$H_L = \sum_{k=1}^{g} [\frac{O_k - n_k \bar{P}_k}{n_k \bar{P}_k (1 - \bar{P}_k)}] \qquad (5)$$

O_k represents the observed frequency of the number of successes within group, n_k is the total number of cases within the group and $\bar{P}_k$ is the average expected cases for group k where $\bar{P}_k = \sum_{i=1}^{n_k} \frac{p_i}{n_k}$. Another test called the Wald test the way to evaluate the individual contributions of the independent variables to the model is to calculate the square of the regression coefficient to the square of the standard error of the coefficient, which is close to the x^2 distribution and can be expressed as:

$$W_j = \frac{B_j^2}{S\ EB_j^2} \qquad (6)$$

(Park, 2013),(Kleinbaum & Klein, 2010).

2.3 Evaluation of classification models:

To evaluate the performance of methods, we can use the confusion matrix as in table 1. It is a binary matrix which divided into four categories, Where true values refer to their expected values according to the classification.

Table 1: Confusion matrix

Observed	Predicted	
	Positive	Negative

Positive	TP	FN
Negative	FP	TN

Whereas :

(TP True Positive) represents positive cases and their cases (predictive values) are classified as positive.

(TN True Negative) represents negative cases and their cases (predictive values) are classified as negative.

(FP False Positive) Represents negative cases and their cases (predictive values) are classified as positive.

(FN False Negative) represents positive cases and their cases (predictive values) are classified as negative.

The sensitivity measure, expressed as the True Positivity Rate (TPR), is the probability that positive cases are classified as positive, is calculated as:

$$Se = \frac{TP}{TP + FN}$$

The specificity measure, expressed as the True Negative Rate (TNR), is the probability that cases are negative and are classified as negative, is calculated as:

$$Sp = \frac{TN}{TN + FP}$$

The Receiver Operating Characteristic (ROC) technique is a graph used to evaluate the quality of models, compare and select classifiers based on their performance. It is a graph in which the (TPR) is plotted on the y-axis and the (FPR) is plotted on the x-axis, when (TPR=1, FPR=0) then the curve is perfect. The distance under the curve (ROC) is defined as the unit squared ratio under the curve (Area Under the Curve (AUC)), it's values between (0,1), when (AUC > 0.8) It means that the model is good, It means the probability of a positive sample having a higher classification score than a negative sample (Rainio et al., 2024),(Mohammed et al., 2019),(Mohammed et al., 2018).

3.Results and Discussion

Data description:

The data were obtained through url={https://www.kaggle.com/dsv/8738477}, DOI={10.34740/KAGGLE/DSV/8738477}, This data contains records of patients diagnosed with Alzheimer's disease, which were created from real or simulated medical records. The data includes a set of (32) different clinical, demographic, and diagnostic features for a sample size of (2149). It does not contain missing values. The dependent variable is the diagnosis (Alzheimer's disease) which was assigned a score of (1= infected) for (760) patients (35.4%) and (0= uninfected) for (1389) patients (64.6%),Since the data are not balanced, It were rebalanced by assigning weights to the cases and then selecting equal samples (random selection) . The SPSS v.25 statistical program was used to

analyze the data, apply, and compare the methods used in the study. The results of the analysis were as follows:

Linear discriminant analysis results:

To test the contribution of variables to the discrimination or separation process, this is done by calculating the canonical correlation as shown in Table 2 below:

Table 2: canonical correlation coefficient values for variables

Function	Eigenvalue	% of Variance	Cumulative %	Canonical Correlation
1	0.774	100.0	100.0	0.661

As shown in Table 2 above, the value of the canonical correlation coefficient is (0.661), which indicates the presence of a fairly strong correlation (the correlation of the variables with the discriminant function), where the closer its value is to 1, the stronger the discrimination between the groups is with respect to the discriminant function.

To test the discriminant function's ability to discriminate and classify, this is done by calculating the value of the Wilks' lambda statistic:

Table 3: Calculation of value wilks' lambda

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	0.564	1222.022	32	0.000

Table 3 above shows that the Wilks' Lambda statistic value is (0.564) and the sig value is (0.000), which is less than the significance level of (0.001). This leads us to say that the discriminant function is capable of classifying the group members (infected and uninfected).

To determine the relative importance of each variable in the discrimination process, this is done by calculating the coefficients of the standardized canonical discriminant function, where the larger the absolute value of the coefficient indicates its greater ability to discriminate.

Table 4: Standardized Canonical Discriminant Function Coefficients

Function 1					
Age	0.058	Cardiovascular Disease	-0.036	MMSE (Mini-Mental State Examination)	0.459
Gender	0.007	Diabetes	0.001	Functional Assessment	0.631
Ethnicity	0.017	Depression	-0.017	Memory Complaints	-0.573
Education Level	0.046	Head Injury	0.040	Behavioral Problems	-0.469
BMI	0.012	Hypertension	-0.047	ADL (Activities of Daily Living)	0.597
Smoking	0.053	Systolic BP	0.005	Confusion	0.027
Alcohol Consumption	0.028	Diastolic BP	-0.013	Disorientation	0.031
Physical Activity	0.002	Cholesterol Total	0.002	Personality Changes	0.001
Diet Quality	-0.016	Cholesterol LDL	0.065	Difficulty Completing Tasks	-0.005
Sleep Quality	0.058	Cholesterol HDL	-0.051	Forgetfulness	-0.007

Family History Alzheimer's	0.026	Cholesterol Triglycerides	-0.052		

Table 4 above shows that the functional assessment factor had the highest discrimination ability, with a value of (0.631), followed by the normal daily activities factor with a value of (0.597), then the memory complaints factor with a value of (-0.573) and the MMSE factor with a value of (0.459). The least discriminating factors were personality changes and diabetes, with a value of (0.001) and physical activity and total cholesterol with a value of (0.002).

Table 5 below shows the classification function coefficients for both groups, through which the two groups can be compared, as the cases are assigned to the group whose function obtained the highest score.

Table 5: Classification Function Coefficients

	Diagnosis			Diagnosis			Diagnosis	
	0	1		0	1		0	1
Age	0.960	0.948	Cardiovascular Disease	1.781	1.968	MMSE (Mini-Mental State Examination)	0.318	0.217
Gender	1.521	1.497	Diabetes	0.304	0.300	Functional Assessment	0.996	0.565
Ethnicity	0.532	0.500	Depression	1.864	1.940	Memory Complaints	-2.094	0.635
Education Level	2.446	2.352	Head Injury	2.455	2.199	Behavioral Problems	-3.199	-0.767
BMI	0.565	0.562	Hypertension	1.809	2.051	ADL (Activities of Daily Living)	1.106	0.712
Smoking	2.205	1.992	Systolic BP	0.212	0.211	Confusion	1.151	1.028
Alcohol Consumption	0.388	0.379	Diastolic BP	0.296	0.297	Disorientation	-0.078	-0.232
Physical Activity	0.590	0.589	Cholesterol Total	0.125	0.125	Personality Changes	2.111	2.106
Diet Quality	0.521	0.532	Cholesterol LDL	0.067	0.065	Difficulty Completing Tasks	2.095	2.118
Sleep Quality	1.946	1.886	Cholesterol HDL	0.100	0.104	Forgetfulness	1.519	1.545
Family History Alzheimer's	1.368	1.257	Cholesterol Triglycerides	0.023	0.024	(Constant)	-119.682	-113.963

Binary logistic regression analysis results :

When the data was analyzed using binary logistic regression, the results were as follows:

Table 6: Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	1210.173	32	0.000
	Block	1210.173	32	0.000
	Model	1210.173	32	0.000

Table 6 above shows the effect of the independent variables on the dependent variable in the model. The results showed that the chi-square value is (1210.173) with a significance level of (sig=0.000), which is less than (0.001), this leads us to say that the independent variables together affect the dependent variable meaning that there is a relationship between the dependent variable and the independent variables.

Table 7: pseudo R square

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	1582.146	0.431	0.592

Table 7 above shows the value of the pseudo R square statistic which was calculated using Coxs and Snells R squares and Nagelkerkes R squares. According to the results shown in the table, the value of Coxs and Snells R square is (0.431) which means that (43.1%) of the variance is explained by the independent variables. The value of Nagelkerke's R square is (0.592) which means that (59.2%) of the variance is explained by the independent variables.

Table 8: Variables in the Equation

	B	S.E.	Wald	df	Sig.	Exp(B)
Age	-0.010	0.007	2.147	1	0.143	0.990
Gender	-0.048	0.128	0.141	1	0.707	0.953
Ethnicity	-0.037	0.065	0.331	1	0.565	0.963
Education Level	-0.088	0.071	1.542	1	0.214	0.916
BMI	-0.004	0.009	0.243	1	0.622	0.996
Smoking	-0.214	0.142	2.274	1	0.132	0.807
Alcohol Consumption	-0.009	0.011	0.627	1	0.428	0.991
Physical Activity	-0.007	0.022	0.110	1	0.740	0.993
Diet Quality	0.010	0.022	0.204	1	0.651	1.010
Sleep Quality	-0.055	0.036	2.328	1	0.127	0.946
Family History Alzheimer’s	-0.111	0.149	0.555	1	0.456	0.895
Cardiovascular Disease	0.168	0.176	0.906	1	0.341	1.182
Diabetes	0.015	0.183	0.007	1	0.933	1.015
Depression	0.066	0.156	0.179	1	0.672	1.068
Head Injury	-0.327	0.222	2.166	1	0.141	0.721
Hypertension	0.198	0.178	1.243	1	0.265	1.219
Systolic BP	-0.001	0.002	0.084	1	0.773	0.999
Diastolic BP	0.002	0.004	0.268	1	0.605	1.002
Cholesterol Total	0.000	0.001	0.025	1	0.874	1.000
Cholesterol LDL	-0.003	0.001	3.836	1	0.050	0.997
Cholesterol HDL	0.005	0.003	2.954	1	0.086	1.005
Cholesterol Triglycerides	0.001	0.001	1.380	1	0.240	1.001

MMSE (Mini-Mental State Examination)	-0.107	0.008	170.916	1	0.000	0.898
Functional Assessment	-0.448	0.026	286.149	1	0.000	0.639
Memory Complaints	2.597	0.167	242.548	1	0.000	13.420
Behavioral Problems	2.505	0.185	183.335	1	0.000	12.248
ADL (Activities of Daily Living)	-0.422	0.026	263.111	1	0.000	0.656
Confusion	-0.154	0.160	0.934	1	0.334	0.857
Disorientation	-0.121	0.176	0.472	1	0.492	0.886
Personality Changes	-0.071	0.182	0.152	1	0.697	0.931
Difficulty Completing Tasks	0.101	0.174	0.335	1	0.562	1.106
Forgetfulness	0.004	0.139	0.001	1	0.980	1.004
Constant	5.298	0.969	29.889	1	0.000	199.850

Table 8 above shows the value of the B coefficients, the odds ratio (which is the value of the exponential function of the regression coefficient which expresses the multiplier by which the odds ratio changes (i.e. the probability of an event occurring to the probability of its non-occurrence), as well as the Wald test, through which it is possible to know which of the independent variables are statistically significant and have an impact on the dependent variable. According to the results shown in Table (8), the following can be interpreted the functional assessment variable ranked first in its impact on the disease diagnosis variable, as the value of the Wald test reached (286.149) at a significance level of (sig = 0.000) and its regression coefficient reached (-0.448), meaning that the change in disease diagnosis will decrease by a probability of (0.448) times in the logarithm of preference for the dependent variable, holding all other variables constant followed by the ADL variable as the Wald test value reached (263.111) At a significance level of (sig = 0.000), its regression coefficient was (-0.422) meaning that the change in the diagnosis of the disease will decrease the disease by a probability of (0.422) times in the logarithm of preference of the dependent variable with the rest of the variables constant. followed by the memory complaints variable, where the value of the Wald test reached (242.548) at a significance level of (sig = 0.000) and its regression coefficient reached (2.597) meaning that the change in the diagnosis of the disease will increase by a probability of (2.597) times in the logarithm of preference of the dependent variable with the rest of the variables constant. followed by the behavioral problems variable, where the value of the Wald test reached (183.335) at a significance level of (sig = 0.000) and its regression coefficient reached (2.505) meaning that the change in the diagnosis of the disease will increase by a probability of (2.505) times in the logarithm of preference of the dependent variable, with the rest of the variables constant. followed by the MMSE variable, where the value of the Wald test reached (170.916) at a significance level of (sig = 0.000) and its regression coefficient is (-0.107) which means that the change in diagnosis will decrease by a probability of (0.107) times the logarithm of preference for the dependent variable, holding other variables constant.

The binary logistic regression function can also be written as:

$$p(x) = \frac{1}{1 + e^{-(5.298 - 0.448(FunctionalAssessment) - 0.422(ADL) + 2.597\,(MemoryComplaints) + 2.505(BehavioralProblems) - 0.107\,(MMSE)}}$$

Comparison of models performance:

To compare and evaluate the performance of linear discriminant analysis and binary logistic regression models in correctly classifying the membership of the dependent variable (infected = 1, uninfected = 0), the following is calculated:

- The classification matrix (confusion matrix), which is the actual membership versus the predicted membership of the group:

Table 9 above shows that the linear discriminant analysis correctly classified group membership (83.2%) of the time, classifying the uninfected individuals as uninfected (1153) and infected individuals as infected (634), but incorrectly classifying the uninfected individuals as infected (236) and infected individuals as uninfected (126).

Table 9: Classification Results (LDA)

	Diagnosis	Predicted Group Membership		Total
		0	1	
Count	0	1153	236	1389
	1	126	634	760
%	0	83.0	17.0	100.0
	1	16.6	83.4	100.0

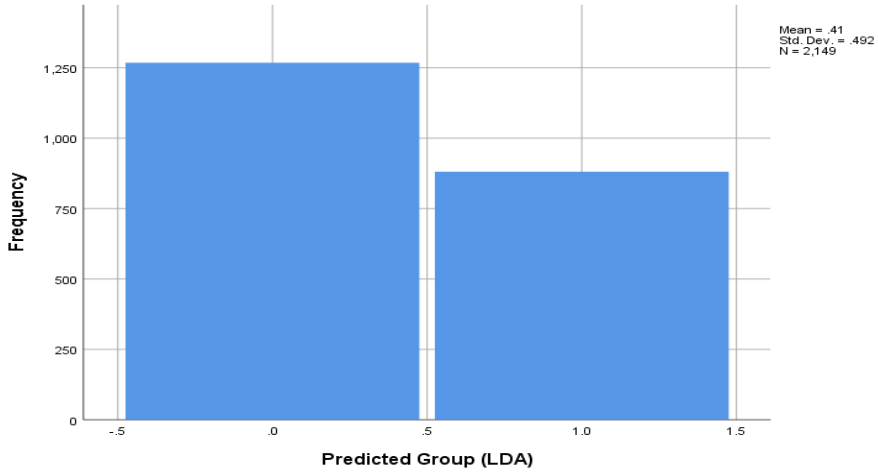

Figure 1: predicted group (LDA)

Table 10 shows that the binary logistic regression model was able to correctly classify disease diagnoses (84.8%) of the time, classifying uninfected individuals as uninfected in (1255) cases and infected individuals as infected in (567) cases. However, it incorrectly classifying the uninfected individuals as infected in (134) cases and infected individuals as uninfected in (193) cases.

Table 10: Classification Results (BLR)

Observed		Predicted		
		Diagnosis		Percentage Correct
		0	1	
Diagnosis	0	1255	134	90.4
	1	193	567	74.6
Overall Percentage				84.8

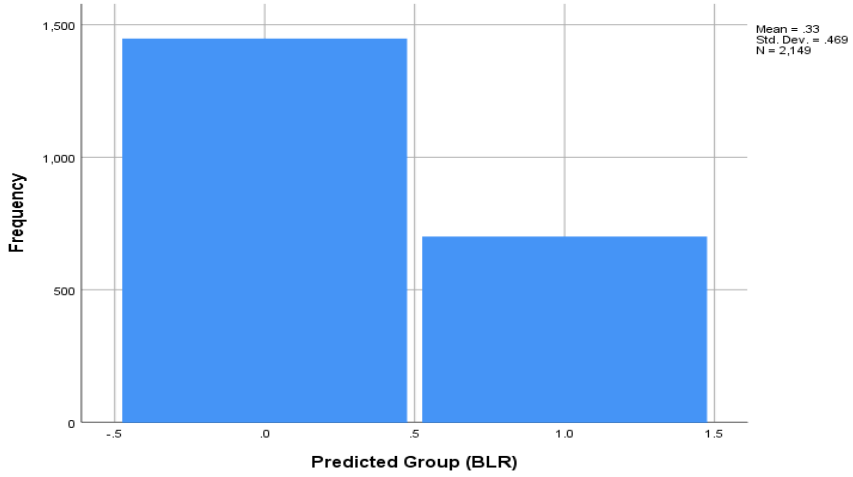

Figure 2: predicted group (BLR)

- Calculating Sensitivity, Specificity, and Area Under the Curve (AUC):

The performance of the models can also be compared by calculating the sensitivity, specificity, and Area Under the Curve (AUC) for both models, The results of the analysis were as follows:

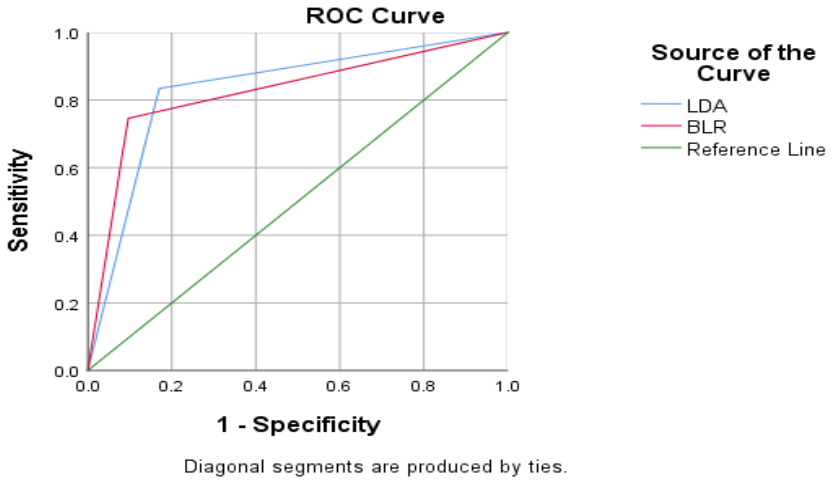

Figure 3: shows the comparison between the two models by ROC

The Table 11 above shows the classification rate of both models was good but had differences in Sensitivity, Specificity and (AUC). The table above shows that the LDA model outperforms the BLR model in sensitivity (83.4% vs 74.6%), making it the best choice for reducing misdiagnosis in patients, while the BLR model outperforms the LDA model in specificity (90.4% vs 83%), making it the best choice for avoiding misdiagnosis in uninfected people.

Table 11: Calculate Sensitivity, Specificity and (AUC)			
Test Result Variable(s)	% (AUC)	%Sensitivity	%Specificity
LDA	83.2	83.4	83
BLR	82.5	74.6	90.4

4.Conclusions:

The study included the diagnosis and prediction of Alzheimer's disease (infected and uninfected) using linear discriminant analysis and binary logistic regression analysis on a sample of (2149) consisting of (32) independent variables and a diagnostic (dependent) variable. using SPSS v.25, the results showed that the linear discriminant analysis performed well in the discrimination and prediction of group membership. The results also demonstrated the discriminant function's ability to discriminate by calculating the Wilks' Lambda statistic value, which reached (0.564) at a significance level of (sig = 0.000), where the functional assessment factor had the highest discriminatory ability, while the least discriminatory factors were personality changes and diabetes. The binary regression analysis also demonstrated the ability to correctly and accurately classify, as the chi-square value showed a significant effect of the independent variables on the dependent variable, as the test value reached (1210.173) at a significance level of (sig = 0.000), The functional assessment factor was ranked first in its impact on the disease diagnosis variable, while the Forgetfulness factor was ranked last . According to comparing the performance of the two models in classifying and predicting group membership, the classification results showed that the linear discriminant analysis model correctly classified group membership with a percentage of (83.2%), a sensitivity of (83.4%), a specificity of (83%), and an area under the curve of (83.2%). As for the binary logistic regression analysis, the classification results showed that the model correctly classified group membership with a percentage of (84.8%), a sensitivity of (74.6%), a specificity of (90.4%), and an area under the curve of (82.5%), Which makes the LDA model the best for reducing misdiagnosis in patients and the BLR model the best for avoiding misdiagnosis in uninfected people. Hence, it can be said that it is preferable to use the BLR model in medical applications because it gives the probability of contracting the disease and is easy to interpret, while it is preferable to use LDA if the goal is to classify patients.

5.References:

1. Ahmadi, M., Javaheri, D., Khajavi, M., Danesh, K., & Hur, J. (2024). A deeply supervised adaptable neural network for diagnosis and classification of Alzheimer's severity using multitask feature extraction. *PLoS ONE*, *19*(3), 1–20. https://doi.org/10.1371/journal.pone.0297996
2. AlKhayyat, S. L., & Alhaqbani, A. M. (2021). A comparison between binary logistic regression and discriminant analysis in determining the most important factors affecting the instability of marital life in the Kingdom of Saudi Arabia. *Applied Mathematical Sciences*, *15*(15), 725–736. https://doi.org/10.12988/ams.2021.916585
3. Alroobaea, R., & Al, E. (2021). *Alzheimer ' s Disease Early Detection Using Machine Learning Techniques*. 1–16.
4. Baglat, P., Salehi, A. W., Gupta, A., & Gupta, G. (2020). Multiple Machine Learning Models for Detection of Alzheimer ' s Disease Using OASIS Dataset. *International Working Conference on Transfer and Diffusion of IT (TDIT)*, 614–622.
5. Benyoussef, E. M., Elbyed, A., & El Hadiri, H. (2017). Data mining approaches for Alzheimer's disease diagnosis. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10542 LNCS*, 619–631. https://doi.org/10.1007/978-3-319-68179-5_54
6. C.Pampel, F. (2021). *Logistic Regression A Primer SECOND EDITION*.
7. Demir, S., & Selvitopi, H. (2025). Early Diagnosis of Alzheimer's Disease Using Machine Learning

Methods. *Procedia Computer Science*, *258*, 107–117. https://doi.org/10.1016/j.procs.2025.04.201
8. Domínguez-Almendros, S., Benítez-Parejo, N., & Gonzalez-Ramirez, A. R. (2011). Logistic regression models. *Allergologia et Immunopathologia*, *39*(5), 295–305. https://doi.org/10.1016/j.aller.2011.05.002
9. Elgohari, H. (2017). Efficiency of Discriminant Analysis and Multivariate Logistic Regression for the Detection of Anemic Children with Chronic Kidney Disease. *International Journal of Statistics and Applications*, *7*(2), 131–136. https://doi.org/10.5923/j.statistics.20170702.09
10. Hilbe, J. M. (2015). *Practical Guide to Logistic Regression*.
11. Kleinbaum, D. G., & Klein, M. (2010). *Logistic Regression A Self-Learning Text*.
12. Liau, S. J., Bell, J. S., Lalic, S., Tolppanen, A. M., & Hartikainen, S. (2024). Symptomatic and Preventive Medication Use before and after Alzheimer's Disease Diagnosis: A 10-Year Matched Cohort Study. *JAMDA*, *25*. https://doi.org/10.1016/j.jamda.2024.04.001
13. Mabrouk, B., Hamida, A. Ben, Mabrouki, N., Bouzidi, N., & Mhiri, C. (2023). A novel approach to perform linear discriminant analyses for a 4-way alzheimer's disease diagnosis based on an integration of pearson's correlation coefficients and empirical cumulative distribution function. *Multimedia Tools and Applications*, 0–18. https://doi.org/10.1007/s11042-024-18532-1
14. Mahmood, S. H., & Khider, H. H. (2023). Comparison of Linear Discriminant Analysis and Artificial Neural Networks for Stroke patients Classification. *University of Kirkuk Journal For Administrative and Economic Science*, *13*(2), 208–220.
15. Matrood, D. O., & Al-quraishi, O. A. K. (2023). The use of linear discriminant analysis to determine the variables affecting cancer diseases. *University of Kirkuk Journal For Administrative and Economic Science*, *13*(1), 85–102.
16. Mohammed, S. N., Guzel, M. S., & Bostanci, G. E. (2019). Classification and Success Investigation of Biomedical Data Sets Using Supervised Machine Learning Models. *3rd International Symposium on Multidisciplinary Studies and Innovative Technologies, ISMSIT*. https://doi.org/10.1109/ISMSIT.2019.8932734
17. Mohammed, T. A., Alhayali, S., Bayat, O., & Uçan, O. N. (2018). Feature reduction based on hybrid efficient weighted gene genetic algorithms with artificial neural network for machine learning problems in the big data. *Scientific Programming*, *2018*, 1–10. https://doi.org/10.1155/2018/2691759
18. Nwanamidwa, T. S. (2018). *A comparative study of multiple discriminant analysis and multinomial logistic regression applied to students' performance*.
19. Park, H. A. (2013). An introduction to logistic regression: From basic concepts to interpretation with particular attention to nursing domain. *Journal of Korean Academy of Nursing*, *43*(2), 154–164. https://doi.org/10.4040/jkan.2013.43.2.154
20. PENG, C.-Y. J., LEE, K. L., & INGERSOLL, G. M. I. (2002). An Introduction to Logistic Regression Analysis and Reporting. *The Journal of Educational Research*, *96*(1).
21. Polat, C. (2018). Performance Evaluation of Logistic Regression, Linear Discriminant Analysis, and Classification and Regression Trees under Controlled Conditions. *Electronic Theses and Dissertations. 1503.*, 174.
22. PREMPEH, E. A. (2009). *COMPARATIVE STUDY OF THE LOGISTIC REGRESSION ANALYSIS AND THE DISCRIMINANT ANALYSIS*.
23. Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, *14*(1), 1–14. https://doi.org/10.1038/s41598-024-56706-x
24. Saied, I. M., Arslan, T., Member, S., & Chandran, S. (2021). Classification of Alzheimer's Disease Using RF Signals and Machine Learning. *IEEE Journal of Electromagnetics, RF and Microwaves in Medicine and Biology*. https://doi.org/org/10.1109/JERM
25. Salunkhe, S., Bachute, M., Gite, S., Vyas, N., Khanna, S., Modi, K., Katpatal, C., & Kotecha, K. (2021). Classification of alzheimer's disease patients using texture analysis and machine learning. *Applied System Innovation*, *4*(49). https://doi.org/10.3390/asi4030049
26. Vidushi, Rajak, A., & Shrivastava, A. K. (2020). Diagnosis of Alzheimer Disease using Machine Learning Approaches. *International Journal of Advanced Science and Technology*, *29*(04), 7062–7073.
27. Yu, S., Liu, Y. P., Liu, H. L., Li, J., Xiang, Y., Liu, Y. H., Jiao, S. S., Liu, L., Wang, Y., & Fu, W. (2018). Serum Protein-Based Profiles as Novel Biomarkers for the Diagnosis of Alzheimer's Disease. *Molecular Neurobiology*, *55*(5). https://doi.org/10.1007/s12035-017-0609-0