ENHANCED KNOWLEDGE IN SCIENCES AND TECHNOLOGY
e-ISSN: 2773-6385

EKST

Vol. 6 No. 1 (2026) 581-591
https://publisher.uthm.edu.my/periodicals/index.php/ekst

Multinomial Logistic Regression for Diabetes Prediction using LASSO and Elastic Net Regularization

Tai Wei Hong[1], Khuneswari P Gopal Pillay[1]*

[1] *Department of Mathematics and Statistics, Faculty of Applied Sciences and Technology, UTHM Kampus Cawangan Pagoh, Hab Pendidikan Tinggi Pagoh, KM 1, Jalan Panchor, 84600 Pagoh, Muar, Johor, MALAYSIA.*

*Corresponding Author: khuneswari@uthm.edu.my
DOI: https://doi.org/10.30880/ekst.2026.06.01.064

Article Info

Received: 4 January 2026
Accepted: 6 June 2026
Available online: 9 July 2026

Keywords

Diabetes Prediction, Multinomial Logistics Regression, Classification, Diabetic, Prediabetic, Non-Diabetic, LASSO, Elastic Net, Body Mass Index, Age, HbA1c, Triglycerides and Cholesterol

Abstract

Diabetes is a chronic metabolic disorder characterized by persistently elevated blood glucose levels and is associated with severe long-term health complications if not detected early. Accurate classification of individuals into Non-Diabetic, Prediabetic, and Diabetic categories are therefore essential for effective prevention and clinical intervention. This study is a multiclass diabetes prediction model by using Multinomial Logistic Regression integrated with LASSO and Elastic Net regularization techniques. A dataset of 1,000 patient records containing demographic and biochemical variables was analysed following data cleaning, normalization, and an 80:20 train–test split. Model training and hyperparameter tuning were conducted using 10-fold cross-validation. Performance was evaluated by using accuracy, Cohen's Kappa, F1-score, Brier score, macro-AUC, and information criteria. Elastic Net model achieves the highest accuracy of 0.9141, the highest Cohen's Kappa of 0.6340, the highest macro sensitivity of 0.6607 and the highest macro balanced accuracy of 0.7303 among the models. In addition, it consists of fewest variables and provided the most balanced and stable result of AIC, BIC, Pseudo-R2, F1-Macro, F1-Weighted, Brier score and Macro specificity which is 387.8591. 401.9204, 0.5556, 0.9559, 0.5737, 0.8909 and 0.8665 respectively. The results indicate that Elastic Net regularization provides the most stable and balanced classification performance while effectively handling correlated variables. Key variables that influenced diabetes status include body mass index, age, HbA1c, triglycerides, and cholesterol. The proposed approach offers an interpretable and reliable framework for multiclass diabetes classification and early risk stratification.

1. Introduction

Diabetes mellitus is a chronic metabolic disorder characterised by persistent hyperglycaemia caused by insufficient insulin secretion or impaired insulin action [1]. The rapid increase in diabetes prevalence worldwide has made it a major public health concern, as delayed diagnosis often leads to severe complications such as cardiovascular disease, kidney failure, and neuropathy [2]. These complications significantly increase healthcare costs and mortality rates, highlighting the importance of early and accurate detection [3]. Most existing diagnostic and predictive approaches classify individuals into binary outcomes, namely diabetic and non-diabetic. However, this approach neglects the prediabetic stage, during which early intervention can substantially reduce the risk of progression to type 2 diabetes [4]. With the growing availability of clinical data, statistical and machine learning methods have been increasingly applied to improve disease prediction [5]. Among these methods, multinomial

logistic regression is well suited for multiclass classification problems and offers interpretable probabilistic predictions, making it particularly valuable in medical applications [6]. However, the presence of multicollinearity and high-dimensional variables can reduce model reliability and prediction accuracy. Regularization methods such as LASSO and Elastic Net have been shown to effectively address these challenges by shrinking regression coefficients, performing feature selection, and improving model generalisation [7]. Lasso encourages sparse solutions by eliminating less relevant variables, while Elastic Net combines L1 and L2 penalties to better handle correlated variables [8]. Integrating these regularization techniques with multinomial logistic regression provides a robust and interpretable framework for multiclass diabetes prediction, supporting early risk stratification and informed clinical decision-making [9].

Despite continuous advancements in diabetes diagnosis and management, the disease remains a major global health challenge due to its increasing prevalence and severe long-term complications. Most existing predictive models rely on binary classification approaches that categorise individuals as either diabetic or non-diabetic, thereby overlooking the prediabetic stage, where early intervention can significantly reduce disease progression. Furthermore, diabetes-related clinical datasets often contain multiple correlated variables, which may lead to multicollinearity, overfitting, and reduced interpretability in conventional multinomial logistic regression models. Consequently, there is a need for a robust and interpretable multiclass prediction framework that can accurately classify diabetic, non-diabetic, and prediabetic individuals while effectively addressing model complexity and feature redundancy through appropriate regularization techniques. The objectives of the study are to determine the relationship between diabetes and key patient attributes such as blood sugar level, BMI, age, cholesterol levels, HbA1c and other clinical indicators, constructing a multinomial logistic regression model for predicting diabetes status using Lasso and Elastic Net Regularization techniques and identifying the most significant patient variables that contribute to accurate classification of diabetes status based on the best models.

Previous studies have shown that diabetes prediction commonly relies on clinical and demographic variables such as age, body mass index (BMI), HbA1c, and lipid profiles, which are strongly associated with glycaemic control and insulin resistance [10]. Early identification of individuals in the prediabetic stage is particularly important, as timely intervention can significantly reduce progression to type 2 diabetes [11]. Statistical and machine learning methods have been widely applied in diabetes prediction, including logistic regression and other classification techniques [12]. However, most existing models focus on binary outcomes, classifying individuals as diabetic or non-diabetic, which limits early risk stratification and neglects the clinically important prediabetic group [16]. Multinomial logistic regression is well suited for multiclass medical classification problems and offers interpretable probabilistic predictions [13]. Nevertheless, correlated and redundant variables in medical datasets may lead to overfitting and unstable coefficient estimates. Regularization techniques such as LASSO and Elastic Net address these issues by performing feature selection and improving model generalisation [14]. Both LASSO and Elastic Net Regularization can select important features and manage complex data to avoid overfitting occur [15].

2. Methodology

In this study, the dataset of patient about the diabetes information are obtained from the Iraqi society where it came from a laboratory of Medical City Hospital.

2.1 Data Description

This study utilises a structured dataset comprising medical and demographic information collected from patients, aimed at facilitating the prediction and classification of diabetes status. The dataset consists of 1000 observations of patients and includes a total of 14 attributes which include both clinical indicators and categorical variables. The variables recorded are patient identification number, patient reference number, age, gender, creatinine ratio (Cr), body mass index (BMI), urea, total cholesterol (Chol), fasting lipid profile components such as low-density lipoprotein (LDL), very-low-density lipoprotein (VLDL), triglycerides (TG) and high-density lipoprotein (HDL), along with hemoglobin A1c (HbA1c) levels. The dataset is obtained from an open-access repository Mendeley Data and the data were collected from patients in Iraq through medical records and laboratory results from Medical City Hospital and the Specialised Centre for Endocrinology and Diabetes in Al-Kindy Teaching Hospital. List of variables in the diabetes dataset including their descriptions and the types is explained in Table 1.

Table 1 Variables in the Dataset

Variables	Description	Data Type
ID	Unique identifier for each patient entry	Numerical
No_Pation	Patient identification number	Numerical
Gender	Gender of the patient (M= Male, F = Female)	Categorical

AGE	Age of the patient	Numerical
Urea	Urea level in the blood	Numerical
Cr	Creatinine ratio	Numerical
HbA1c	Glycated haemoglobin	Numerical
Chol	Total cholesterol level	Numerical
TG	Triglyceride level in blood	Numerical
HDL	High-Density Lipoprotein	Numerical
LDL	Low-Density Lipoprotein	Numerical
VLDL	Very-Low-Density Lipoprotein	Numerical
BMI	Body Mass Index	Numerical
CLASS	Diabetes status classification (Y=Diabetic, N= Non-diabetic and P =Prediabetic)	Categorical

(Source: Mendeley Data, 2020)

2.2 Descriptive Analysis

To develop the predictive model, descriptive analysis was conducted to get the general idea about the data. Cross-tabulation and chi-square tests were applied to evaluate associations between categorical variables and diabetes outcome [16]. For continuous variables, the Kruskal–Wallis test was employed to assess statistically significant differences across the diabetic, prediabetic, and non-diabetic groups [17]. These analyses provide preliminary justification for variable inclusion in the predictive models.

2.3 Model Development

In this study, multinomial logistic regression (MLR) was employed to model the multiclass diabetes outcome consisting of diabetic, prediabetic, and non-diabetic categories. MLR estimates the probability of each outcome category as a function of multiple predictor variables, making it suitable for multiclass medical classification problems. The non-diabetic group was selected as the reference category. To improve model stability and reduce overfitting caused by multicollinearity among variables, regularization techniques were incorporated into the MLR framework. LASSO regularization applies an L1 penalty to the regression coefficients, encouraging sparse solutions by shrinking less relevant coefficients to zero. Elastic Net regularization combines L1 and L2 penalties, enabling effective handling of correlated variables while maintaining model interpretability. The dataset was divided into training (80%) and testing (20%) subsets using stratified sampling to preserve class proportions. Regularization parameters were optimised using 10-fold cross-validation on the training dataset. Model estimation and tuning were conducted using the training data, while the testing data were reserved exclusively for performance evaluation.

2.3.1 Multinomial Logistic Regression

Multinomial Logistic Regression (MLR) is a statistical method for predicting outcomes when the dependent variable has more than two nominal categories [18]. It extends binary logistic regression to handle multi-class classification problems. The general equation of MLR is shown as in Equation (1) [19].

$$\boldsymbol{P(Y = k|x)} = \frac{\boldsymbol{e^{\beta_0+\beta_1 X_1+\cdots+\beta_k X_k}}}{\boldsymbol{\sum_{j=1}^{K} e^{\beta_0+\beta_1 X_1+\cdots+\beta_k X_k}}} \quad (1)$$

Where for k = 1, 2, ..., K, $P\ (Y = k|x)$ = probability of variables k given input x, β_0 = intercept term, β_k = coefficient of variables k for k=1, 2,..., K,x = vector of independent variables (variables).

2.3.2 LASSO Regression

In this study, LASSO is applied with multinomial logistic regression to identify the most influential patient features for diabetes diagnosis [20]. The mathematical formula of LASSO is shown as in Equation (2) [21].

$$(\hat{\alpha}, \hat{\beta}) = \arg min \left\{ \sum_{i=1}^{N} \left(y_i - \alpha - \sum_j \beta_j x_{ij} \right)^2 \right\} \quad (2)$$

Where $\hat{\alpha}$ = estimated intercept, $\hat{\beta}$ = estimated regression coefficients, N = total number of observations, y_i = response variable, x_{ij} = predictor value for observation i and predictor j, α = intercept parameter, β_j = coefficient for predictor j, $\sum_{i=1}^{N}$ = summation over all observations, $\sum_j$ = summation over all variables.

2.3.3 Elastic Net Regression

Elastic Net regression is applied with multinomial logistic regression to classify diabetes into Diabetic, Non-Diabetic, and Prediabetic. The mathematical formulation of Elastic Net is shown as in Equation (3) [22].

$$\left(\widehat{\beta_0}, \hat{\beta}\right) = \arg min_{(\beta_0,\beta)\in\mathbb{R}^{p+1}} \left[\frac{1}{2n}\sum_{i=1}^{n}(y_i - \beta_0 - x_i^T\beta)^2 + \lambda(\frac{1-\alpha}{2}\left|\left|\beta\right|\right|_2^2 + \alpha||\beta||_1\right] \quad (3)$$

Where $\widehat{\beta_0}$ = estimated intercept, $\hat{\beta}$ = estimated coefficient vector, β_0 = intercerpt, β = coefficient vector, n = number of observations, y_i = response variable, x_i = predictor variable vector for observation, $x_i^T\beta$ = dot product of predictor variables and coefficients, λ = regularization parameter, α = mixing parameter, $||\beta||_1$ = $L1$ norm, $\left|\left|\beta\right|\right|_2^2$ = Squared $L2$ norm.

2.4 Model Evaluation and Selection

Model selection is conducted to identify the most suitable model for classifying diabetes status into Diabetic, Non-Diabetic, and Prediabetic. Model performance is assessed using multi-class evaluation metrics including accuracy, specificity, sensitivity, F1-score, and macro-averaged AUC-ROC. To reduce overfitting, k-fold cross-validation is applied, and performance measures are averaged across folds to obtain stable estimates. In addition to predictive performance, model simplicity and interpretability are considered. The final model is selected based on a trade-off between accuracy, generalizability, and interpretability to ensure clinical relevance. Besides, the Corrected Akaike Information Criterion (AICc) and Bayesian Information Criterion (BIC) will also be considered to assess model fit and complexity. Kappa , F1-Macro and F1-Weighted and brier score are also important as the key metrics to evaluate the models. The Brier Score was used to assess the accuracy and calibration of the class probability predicted by the model where the lower the score, the better the probabilistic predictions [23]. Cohen's Kappa was adopted to quantify the extent of agreement between the predicted and actual class labels which also takes into consideration the chance of agreement, especially when class imbalance exists [24]. Furthermore, both F1-Macro and F1-Weighted scores were used for evaluating balance between precision and recall for the three diabetes outcome classes with F1-Macro giving equal weight to each class and F1-Weighted taking class distribution into consideration [25].

A confusion matrix is a crucial performance measurement tool for classification models by comparing actual and predicted class labels. True positive (TP), true negative (TN), false positive (FP) and false negative (FN) are the main components of the confusion matrix to show the performance of the model. Table 2 shows the structure of confusion matrix with multiclass [26]:

Table 2 Structure of Confusion Matix with Multiclass

		Actual		
	Class	Y	N	P
Predicted	Y	TP	FN	FN
	N	FP	TP	FN
	P	FP	FP	TP

2.5 Goodness-of-Fit Test

Goodness of fit test is a very crucial element in the analysis of a statistical model assessment Goodness of fit test is very crucial element. A flowchart is important in study research as it summarises the systematic steps used to predict diabetes status using regularized multinomial logistic regression models [27]. The process begins with data collection and preprocessing, including missing value treatment, categorical variable encoding, and normalization of continuous variables such as blood sugar, BMI, cholesterol, and HbA1c. Next, multinomial logistic regression models with LASSO and Elastic Net regularization are fitted to improve predictive performance and reduce overfitting by penalizing less important variables. Cross-validation is applied to optimize key hyperparameters, including the penalty parameter λ and the mixing parameter α for Elastic Net. Finally, model performance is evaluated using accuracy, Kappa, Brier score, F1-score, and AUC-ROC and the best-performing model is selected based on predictive performance and interpretability.

3. Results and Discussions

This section is included cross-tabulation analysis, chi-squared test, Kruskal-Wallis test, model building, model selection and goodness-of-fit test.

Enhanced Knowledge in Sciences and Technology Vol. 6 No. 1 (2026) 581-591

3.1 Cross-tabulation Analysis of Categorical Variables

Cross-tabulation analysis was performed to analyse the relationship between categorical variables and diabetes classification. Table 3 shows the cross-tabulation analysis of gender by diabetes classification.

Table 3 Cross-Tabulation of Gender by Diabetes Classification

	Diabetic	Non-Diabetic	Prediabetic	Sum
Male	490	39	36	565
Female	354	64	17	435
Sum	844	103	53	1000

The results show that among male respondents, 490 individuals were classified as Diabetic, 39 as non-diabetic and 36 as Prediabetic, giving a total of 565 males. Among female respondents, 354 were Diabetic, 64 were non-diabetic, and 17 were Prediabetic, with a total of 435 females. Overall, diabetic cases dominate both genders, with a higher number of diabetic males than females. Table 4 shows the row-wise proportions of gender by diabetes classification.

Table 4 Row-Wise Proportions of Gender by Diabetes Classification

	Diabetic (%)	Non-Diabetic (%)	Prediabetic (%)	Sum (%)
Male	86.73	6.90	6.37	100
Female	81.38	14.71	3.91	100

Among males, 86.73% were Diabetic, 6.90% were non-diabetic, and 6.37% were Prediabetic. Among females, 81.38% were Diabetic, 14.71% were non-diabetic and 3.91% were Prediabetic. The results indicate that while diabetes prevalence is high for both genders, males have a slightly higher proportion of diabetic cases compared to females.

3.2 Chi-Squared Test for Categorical Variables

A chi-squared test of independence was conducted to determine the association between gender and diabetes status. Table 5 shows the results of the chi-square test of the association between gender and diabetes status.

Table 5 Chi-Square Test of Association Between Gender and Diabetes Status

Variable	Chi-Square Value	df	*p*-value
Gender	18.2020	2	< 0.001

The chi-square statistic of 18.2020 with 2 degrees of freedom and a p-value < 0.001 shows that there is statistical significance association between Gender and diabetes status (Diabetic, Non-Diabetic and Prediabetic) at the 5% of significance level since p-value is less than 0.05. This means that the distribution of Diabetic, Non-Diabetic and Prediabetic cases differ significantly between males and females.

3.3 Kruskal-Wallis Test in Numerical Variables

A Kruskal-Wallis test was conducted to determine whether the distributions of numerical variables differed significantly across the three diabetes groups. This non-parametric test was selected because several variables did not meet the normality assumption required for ANOVA. The test compares the median ranks of each variable across the Diabetic, Non-diabetic and Prediabetic groups to identify meaningful group differences. Table 6 presents the results of the Kruskal-Wallis test for numerical variables by diabetes classification.

Table 6 Kruskal–Wallis Test Results for Numerical Variables by Diabetes Classification

Variable	Chi-Square Value	df	*p*-value
Age	230.8800	2	< 0.05
Urea	8.2451	2	0.0162
Cr	2.3185	2	0.3137

HbA1c	334.0500	2	< 0.05
Chol	35.5720	2	< 0.05
TG	62.0410	2	< 0.05
HDL	1.2112	2	0.5457
LDL	0.3463	2	0.8410
VLDL	54.8280	2	< 0.05
BMI	324.8500	2	< 0.05

The results show that several variables including age, Urea, HbA1c, Chol, TG, VLDL and BMI, exhibited statistically significant differences between the diabetes categories which the *p*-values are less than 0.05. This suggests that these clinical and demographic factors vary significantly across diabetes classifications and may play a crucial role in distinguishing diabetic from non-diabetic individuals. On the other hand, variables such as Cr, HDL and LDL did not show significant differences since their *p*-values are not less than 0.05 which means that their distributions are similar across all diabetes groups.

3.4 Model Building

There are three model are built which are MLR, LASSO and Elastic Net Regression. The MLR model was fitted using the training dataset with all variables included to evaluate the association with diabetes classification. Table 7 shows the coefficient estimates for three model:

Table 7 Coefficient Estimates for the MLR Model

Term	Estimate	Std. Error	*z* value	Pr(>|z|)
(Intercept)	6.7254	32.4565	0.207	> 0.05
Gender (Female)	2.8247	32.4329	0.087	> 0.05
Gender (Male)	4.1409	32.4361	0.128	> 0.05
Age	-0.6083	0.3936	-1.545	> 0.05
Urea	-0.1086	0.3379	-0.322	> 0.05
Cr	0.0031	0.2892	0.011	> 0.05
HbA1c	4.5771	0.9599	4.768	< 0.05
Chol	0.9632	0.3885	2.480	< 0.05
TG	0.9947	0.3738	2.661	< 0.05
HDL	-0.2037	0.3181	-0.640	> 0.05
VLDL	0.0703	0.4714	0.149	> 0.05
BMI	3.9785	0.9056	4.393	< 0.05

HbA1c, Chol, TG and BMI were statistically significant ($p < 0.05$), indicating that higher levels of these clinical parameters are strongly associated with shifts in diabetes classification. Among them, HbA1c and BMI displayed the largest coefficient magnitudes, suggesting that they are particularly influential variables in differentiating between individuals with diabetes, prediabetes and those without diabetes. Other variables, including Gender, AGE, Urea, Cr, HDL, LDL and VLDL were not statistically significant ($p > 0.05$), implying that they did not meaningfully contribute to the classification after adjusting for the effects of other variables.

The non-significance of gender variables (Female and Male) suggests that the observed differences between males and females in earlier descriptive analysis were not strong enough to hold when controlling for clinical measurements. After evaluating the significance of the model coefficients, the performance of the MLR model was assessed using the testing dataset. Table 8 shows the coefficient estimates for three models of MLR, LASSO and Elastic Net:

Enhanced Knowledge in Sciences and Technology Vol. 6 No. 1 (2026) 581-591

Table 8 Coefficient Estimates between the MLR, LASSO and Elastic Net models

Variable	Coefficient		
	MLR	LASSO	Elastic Net
Gender (Female)	2.8247	-0.0574	0
Gender (Male)	4.1409	0	0
Age	-0.6083	-0.0362	0.1241
Urea	-0.1086	-0.0133	0
Cr	0.0031	0	0
HbA1c	4.5771	-0.5567	0.0541
Chol	0.9632	0.0015	0.0806
TG	0.9947	-0.2369	-0.0581
HDL	-0.2037	-0.0376	0
LDL	-0.4863	-0.0398	0
VLDL	0.0703	0.0814	0
BMI	3.9785	-0.1944	0.2450

For MLR model, it includes all variables and indicates that HbA1c, BMI, TG, and Chol have relatively large coefficient magnitudes which show that they have strong associations with diabetes classification. For LASSO model, the results show that the variables of HbA1c, TG and BMI play the most important or influential roles in distinguishing between diabetic, prediabetic and non-diabetic individuals. Variables like VLDL, LDL, HDL, AGE, Urea and cholesterol also showed non-zero coefficients with coefficients of 0.0814, -0.0398, -0.0376, -0.0362, -0.0133, and 0.0015 respectively. It means that these variables make a significant but weaker contribution to the model compared to the main variables which are HbA1c, TG and BMI. GenderF, which is female, shows a negative coefficient with the absolute coefficients of 0.0574, while GenderM, which is male, shows an absolute coefficient of 0.000013 in the model. For both genders, it shows female contributes limited predictive influence in this model, while male have a negligible effect on the model since the absolute coefficient is extremely closed to zero, so it is removed from the model. There is another variable which is Cr, that is reduced to a coefficient of zero by the LASSO penalty, as it does not have any effect or influence on this model, and it is removed from the final model. Therefore, variables for Cr and GenderM are eliminated and excluded from the final model. For Elastic Net model, the results show that BMI, AGE, Chol, TG and HbA1c play the most influential roles in distinguishing between individuals with diabetes, prediabetes, and those without diabetes. There are six variables including the categorical data which are both genders, GenderF (female), GenderM (male), Urea, Cr, HDL, LDL and VLDL, that are completely shrunk to zero by elastic net. It means that these variables do not contribute any predictive information and show no influence on this model, therefore variables GenderF, GenderM, Urea, Cr, HDL, LDL and VLDL are eliminated and excluded from this model.

3.5 Model Comparison and Model Selection

There is some evaluation metrics used to compare the difference between these three models. The evaluation included the accuracy, Cohen's Kappa, information criteria, which are AIC and BIC, pseudo-R^2 and macro-averaged AUC. Table 9 shows the overall comparison of the three models based on these performance metrics.

Table 9 Overall Predictive Performance Metrics

Model	Accuracy	Kappa	AIC	BIC	Pseudo-R^2	MacroAUC
MLR	0.9091	0.6322	307.7669	429.6317	0.7023	0.9696
LASSO	0.8990	0.5846	271.1341	285.1954	0.6914	0.9675
Elastic Net	**0.9141**	**0.6340**	**387.8591**	**401.9204**	**0.5556**	**0.9559**

The Elastic Net model achieves the highest accuracy (0.9141) and shows that it correctly classifies the largest proportion of test observations followed closely by MLR (0.9091) and LASSO (0.8990). For Cohen's Kappa, where Elastic Net achieves highest value, which is 0.6340, as compared to model MLR and LASSO, which are 0.6322 and 0.5846, respectively. In terms of information criteria, LASSO records the lowest AIC and BIC values, which are 271.13 and 285.20, meaning that it offers a balance between model fit and complexity among the models. However, Elastic Net retains fewer variables, so it is the most parsimonious in terms of variable reduction although having

higher values of AIC and BIC. For explanatory power, MLR shows the strongest Pseudo-R^2, which is 0.7023, among three models. In terms of discrimination ability, all models achieve high values of macro-AUC, while MLR shows the best, which is 0.9696 among the three models. Overall, Elastic Net shows the highest accuracy, LASSO offers the best AIC/BIC efficiency and MLR shows the strongest explanatory capability.

In addition, several class-based metrics were compared using the F1-Macro score, the F1-Weighted score, and the Brier Score. These measures can show a more detailed assessment of each model. Table 10 shows the comparative F1 and Brier Score performance for the MLR, LASSO and Elastic Net models.

Table 10 Classification Performance Metrics for MLR, LASSO and Elastic Net

Model	F1-Macro	F1-Weighted	Brier Score
MLR	0.6151	0.8986	1.8065
LASSO	0.5530	0.8845	0.1121
Elastic Net	**0.5737**	**0.8909**	**0.1233**

The MLR model achieves the highest F1-Macro score, which is 0.6151, meaning that it has the most balanced performance across all classes. Its F1-Weighted score (0.8986) is also the highest, reflecting strong overall predictive performance when class prevalence is considered. However, the MLR model has the largest Brier Score, which is 1.8065, followed by the Elastic Net of 0.1233 and the LASSO of 0.1121. It shows that although its classifications are accurate, its probability estimates are poorly calibrated. In contrast, LASSO and Elastic Net achieve lower F1-Macro values, which are 0.5530 and 0.5737, respectively, and show slightly weaker balance across classes, indicating difficulties in predicting the Prediabetic category. Overall, MLR provides a better class-level predictive balance, while the performance of LASSO and Elastic Net is similar, with only differences noted in the F1-score, F1-Weighted, and Brier Score metrics.

To further evaluate the classification performance of the three models, their class-level diagnostic ability is examined using macro-averaged sensitivity, specificity and balanced accuracy. These metrics provide an overall view of how well each model performs across all three classes by treating each category equally, regardless of its prevalence. Table 11 presents the macro-averaged sensitivity, specificity, and balanced accuracy for the MLR, LASSO, and Elastic Net models.

Table 11 Macro-Averaged Class-Based Metrics for MLR, LASSO and Elastic Net

Model	Macro Sensitivity	Macro Specificity	Macro Balanced Accuracy
MLR	0.6087	0.8835	0.7262
LASSO	0.5940	0.8706	0.7147
Elastic Net	**0.6607**	**0.8665**	**0.7303**

The Elastic Net achieves the highest macro-sensitivity, with a value of 0.6607 among the three models, indicating that it is comparatively better at correctly identifying cases across classes. In terms of macro-specificity, all models perform well, with values above 0.86, although MLR records the highest score of 0.8835 which reflects its stronger ability to correctly recognise non-target classes. For macro- balanced accuracy, which accounts for both true positive and true negative rates, Elastic Net performs the best (0.7303), followed by MLR of (0.7262) and LASSO of (0.7147). Overall, these results demonstrate that Elastic Net achieves the most balanced performance across all classes, with LASSO performing slightly weaker and MLR excelling mainly in specificity, albeit not as strongly in sensitivity.

A direct comparison of the most influential variables from LASSO and Elastic Net is important to determine whether the selected model maintains clinical interpretability while avoiding unnecessary complexity. Fig. 1 shows the comparison of the absolute coefficient magnitudes of selected variables between LASSO and Elastic Net models.

Enhanced Knowledge in Sciences and Technology Vol. 6 No. 1 (2026) 581-591

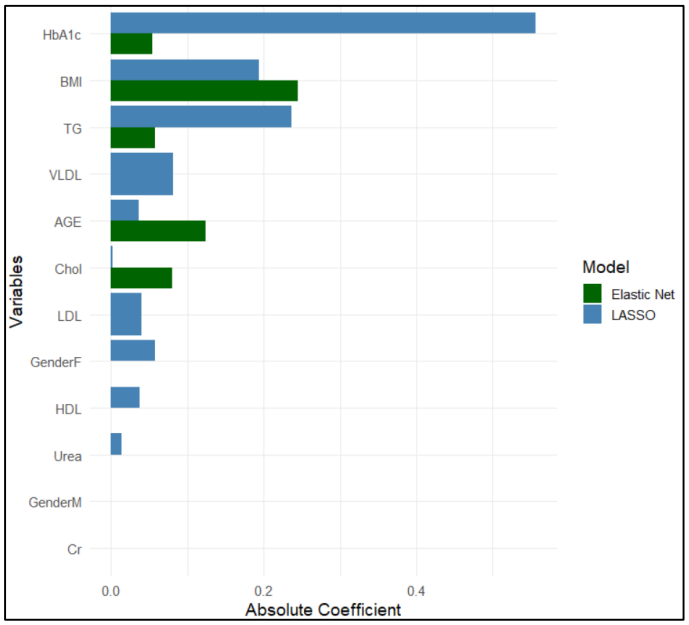

Fig. 1 Comparison of the Absolute Coefficient Magnitudes of Selected Variables between LASSO and Elastic Net Models

Both models identify HbA1c, BMI, and TG as the most influential variables in classifying diabetes status, showing strong consistency in variable importance across methods. Elastic Net retains a smaller and more stable variable, with 5 variables as compared to LASSO, which retains 10 variables. Therefore, Elastic Net is chosen as the best model with 5 variables of BMI, AGE, Chol, TG and HbA1c since it offers the highest accuracy, better balance between bias and variance, superior macro-level sensitivity, and a more parsimonious variable structure while maintaining strong classification performance across major diabetes categories

3.6 Goodness-of-Fit Test

This section will evaluate and check the selected model of Elastic Net fitting on the observed data, which is the training data. Fig. 2 illustrates the trends in deviance residuals for each fitted probability in each class as part of the Elastic Net model.

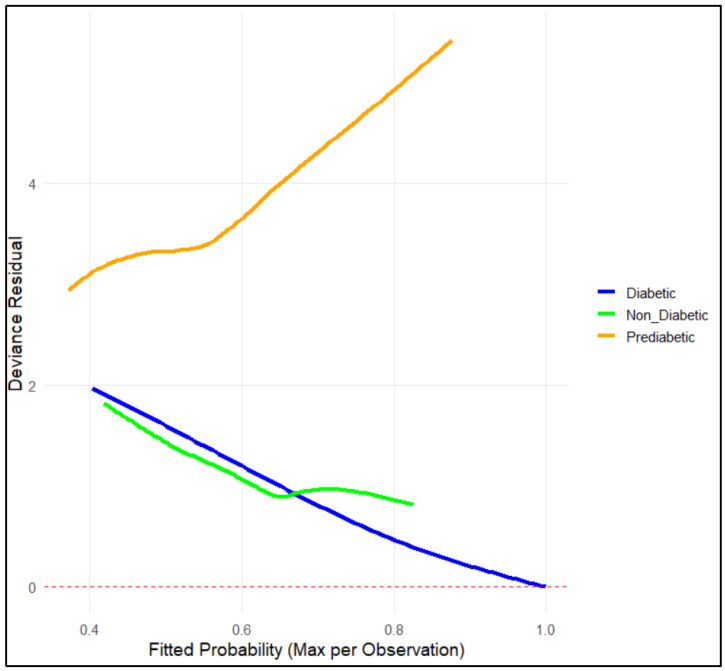

Fig. 2 Relationship between Deviance Residuals and Fitted Probabilities Across Diabetes Classes in Elastic Net Model

Overall, the goodness-of-fit tests indicate that the model Elastic Net represents an acceptable and stable fit for the data. The residual trend plot further supports this by both the Diabetic and Non-Diabetic classes having decreasing residuals with increasing fitted probabilities which is an expected pattern with a well-behaved model. Although the Prediabetic class has higher and more varied residuals, it is consistent with previous results of less predictive success with this minority class rather than a model failure. These results suggest that the Elastic Net model fits most of the observations well and does not show signs of substantial lack-of-fit.

4. Conclusion

Primary objectives in this study are achieved well through the analysis which are investigating the relationship between diabetes status and key patient attributes through descriptive analysis including cross-tabulations, Chi-squared test and Kruskal-Wallis tests. It shows that variables like Age, Urea, Hb1Ac, Cholesterol, TG, VLDL and BMI are influential to the diabetes individual classes. Besides, MLR model using LASSO and Elastic Net is successful conducted and all the models were trained by 10-fold cross-validation technique. Final objective was to identify the most important variables that contribute to the accurate classification of diabetes. Through best model of Elastic Net, five variables which are HBA1c, BMI, TG, Age and Cholesterol were the consistently most influential variables to the classification of diabetes where the result is also fulfilled as the result shown in Kruskal-Wallis test. To put it into a nutshell, the study achieved success for all research goals while statistically meaningful relationships were established between the variables and diabetes classification, well-constructed regularised MLR models were built, and the most important variables were found through a well-validated model selection process.

Although the study has achieved its objectives, there are some limitations throughout the whole research. Firstly, the database was collected from a single source and it may not be used as a representative of wider population differences in demographic and biochemical profiles. Previous studies have proved that diabetes risk factors and disease patterns may vary across regions, healthcare systems and population groups, thereby limiting the generalizability of findings derived that only from one single source [28]. Secondly, there was an obvious class imbalance in the dataset where the Diabetic class was found to be the majority class and the Prediabetic class was relatively small. Similar challenges have been reported in diabetes prediction studies, where minority classes are often misclassified due to overlapping clinical characteristics and insufficient representation [29]. To further improve generalisability, much larger and more diverse datasets from multiple populations of hospitals should be used. Advanced data balancing techniques to overcome class imbalance such as oversampling, under sampling or synthetic data generation methods like SMOTE are also encouraged for the improvement of model detection of minority classes, especially for Prediabetic classes [30]. The model performance would improve with a more holistic understanding of diabetes risk factors would be achieved by expanding the range of variables to behavioural, lifestyle and clinical history variables. For future studies, some advanced forms of machines learning techniques like Random Forest, XGBoost, Support Vector Machines or Neural Networks, which may capture complex interactions and non-linear relationships in more ways [31]. Finally, the validation of the chosen model of the Elastic Net on an independent dataset is recommended to further assess its robustness and applicability in real-world clinical scenarios [32].

Acknowledgement

The authors would like to thank the Faculty of Applied Sciences and Technology at Universiti Tun Hussein Onn Malaysia for its support.

Conflict of Interest

Authors declare that there is no conflict of interests regarding the publication of the paper.

Author Contribution

The authors confirm contribution to the paper as follows: ***study conception and design:*** *Tai Wei Hong, Khuneswari P Gopal Pillay;* ***data collection:*** *Tai Wei Hong;* ***analysis and interpretation of results:*** *Tai Wei Hong, Khuneswari P Gopal Pillay;* ***draft manuscript preparation:*** *Tai Wei Hong, Khuneswari P Gopal Pillay. All authors reviewed the results and approved the final version of the manuscript.*

References

[1] Herman, W. H., Teutsch, S. M., & Geiss, L. S. (2001). Diabetes mellitus. *Nephrology Dialysis Transplantation, 16*(6).

[2] World Health Organization: WHO & World Health Organization: WHO. (2024). Diabetes. https://www.who.int/news-room/fact-sheets/detail/diabetes

[3] Reshmi, S., Biswas, S. K., Boruah, A. N., Thounaojam, D. M., & Purkayastha, B. (2022). Diabetes prediction using machine learning analytics. In *2022 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COM-IT-CON)* (Vol. 1, pp. 108–112). IEEE.
[4] Centers for Disease Control and Prevention. (2024). Prediabetes—United States. https://www.cdc.gov/diabetes/communication-resources/prediabetes-statistics.html
[5] Nilashi, M., bin Ibrahim, O., Ahmadi, H., & Shahmoradi, L. (2017). An analytical method for diseases prediction using machine learning techniques. *Computers & Chemical Engineering, 106*, 212-223.
[6] Kwak, C., & Clayton-Matthews, A. (2002). Multinomial logistic regression. *Nursing research*, 51(6), 404-410.
[7] Fan, L., Chen, S., Li, Q., & Zhu, Z. (2015). Variable selection and model prediction based on lasso, adaptive lasso and elastic net. In *2015 4th International Conference on Computer Science and Network Technology (ICCSNT)* (Vol. 1, pp. 579-583). IEEE.
[8] Waldmann, P., Mészáros, G., Gredler, B., Fuerst, C., & Sölkner, J. (2013). Evaluation of the lasso and the elastic net in genome-wide association studies. *Frontiers in genetics*, 4, 270.
[9] Obi, J. C., & Jecinta, I. C. (2023). A review of techniques for regularization. *International Journal of Research in Engineering and Science, 11*(1), 360-367.
[10] Banday, M. Z., Sameer, A. S., & Nissar, S. (2020). Pathophysiology of diabetes: An overview. *Avicenna journal of medicine,* 10(04), 174-188.
[11] Remuzzi, G., Benigni, A., & Remuzzi, A. (2006). Mechanisms of progression and regression of renal lesions of chronic nephropathies and diabetes. *The Journal of clinical investigation*, *116*(2), 288-296.
[12] Zhang, L., Wang, Y., Niu, M., Wang, C., & Wang, Z (2025). *Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: the Henan rural cohort study. Sci Rep. 2020; 10* (1): 4406.
[13] Deshmukh, S. (2020). *A machine learning approach to predict retention of computer science students at University of Nevada, Las Vegas* (Master's thesis, University of Nevada, Las Vegas).
[14] Abqorunnisa, F., Erfiani, E., & Djuraidah, A. (2023). Performance of LASSO and elastic-net methods on non-invasive blood glucose measurement calibration modeling. *BAREKENG: Jurnal Ilmu Matematika dan Terapan, 17*(1), 0037–0042.
[15] Liu, L., Gao, J., Beasley, G., & Jung, S.-H. (2023). LASSO and Elastic Net tend to over-select features. *Mathematics, 11*(17), 3738. https://doi.org/10.3390/math11173738
[16] Tallarida, R. J., & Murray, R. B. (1987). Chi-square test. In *Manual of pharmacologic calculations: with computer programs* (pp. 140-142). New York, NY: Springer New York.
[17] McKight, P. E., & Najab, J. (2010). Kruskal-wallis test. *The corsini encyclopedia of psychology*, 1-1.
[18] Böhning, D. (1992). Multinomial logistic regression algorithm. *Annals of the institute of Statistical Mathematics*, *44*(1), 197-200
[19] El-Habil, A. M. (2012). An application on multinomial logistic regression model. *Pakistan journal of statistics and operation research*, 271-291.
[20] Fonti, V., & Belitser, E. (2017). Feature selection using lasso. *VU Amsterdam research paper in business analytics*, *30*, 1-25.
[21] Roth, V. (2004). The generalized LASSO. *IEEE transactions on neural networks, 15*(1), 16-28.
[22] De Mol, C., De Vito, E., & Rosasco, L. (2009). Elastic-net regularization in learning theory. *Journal of Complexity, 25*(2), 201-230.
[23] Roulston, M. S. (2007). Performance targets and the Brier score. *Meteorological Applications: A journal of forecasting, practical applications, training techniques and modelling*, *14*(2), 185-194.
[24] Douma, J. C., Cornwell, W. K., & van Bodegom, P. M. (2017). A new metric to assess the predictive accuracy of multinomial land cover models. *Journal of Biogeography*, *44*(6), 1212-1224.
[25] Opitz, J. (2024). A closer look at classification evaluation metrics and a critical reflection of common evaluation practice. *transactions of the association for computational linguistics*, *12*, 820-836.
[26] Tharwat, A. (2021). Classification assessment methods. *Applied computing and informatics*, *17*(1), 168-192.
[27] Kuchiki, A. (2005). A flowchart approach. In *Industrial clusters in Asia: Analyses of their competition and cooperation* (pp. 169-199). London: Palgrave Macmillan UK.
[28] Joshi, R. D., & Dhakal, C. K. (2021). Predicting type 2 diabetes using logistic regression and machine learning approaches. *International journal of environmental research and public health*, *18*(14), 7346.
[29] Wang, S., Chen, Y., Cui, Z., Lin, L., & Zong, Y. (2024). Diabetes risk analysis based on machine learning LASSO regression model. *Journal of Theory and Practice of Engineering Science*, *4*(01), 58-64.
[30] Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.
[31] Chen, T. (2016). XGBoost: A Scalable Tree Boosting System. *Cornell University*.
[32] Steyerberg, E. W., & Harrell Jr, F. E. (2015). Prediction models need appropriate internal, internal-external, and external validation. *Journal of clinical epidemiology*, *69*, 245.