Malaysian Journal of Computing, 9 (2): 1916-1929, 2024
Copyright © UiTM Press
eISSN: 2600-8238

PREDICTION OF DIABETIC RETINOPATHY AMONG DIABETIC NEUROPATHY IN T2DM PATIENTS USING DATA MINING ALGORITHM

Nur Balqis Oon[1], Zuraida Khairudin[2*], Hezlin Aryani Abd Rahman[3], Norbaizura Kamarudin[4], Nur Syamimi Haji Abu Bakar[5] and Nor Azimah Abd Aziz[6]

[1,2*,3,4,5] *College of Computing, Informatics and Mathematics, Universiti Teknologi Mara, Shah Alam, Selangor, Malaysia*

[6]*Faculty of Medicine, Universiti Teknologi Mara, Sg Buloh, Selangor, Malaysia*

[1] alqisoon42@gmail.com, [2*]zuraida_k@fskm.uitm.edu.my [3], hezlin@fskm.uitm.edu.my[4], norbaizura404@uitm.edu.my[5,], syamimi@uitm.edu.my[6], azimah80@uitm.edu.my

ABSTRACT

Diabetic retinopathy (DR) and diabetic neuropathy (DN) are the most common complications among diabetes mellitus (DM) patients. Despite the widespread awareness, the implications of these serious diabetes complications remain poorly understood. Hence, this study aims to determine the association between DR and DN, predict DR and identify the significant risk factors associated with DR among DN patients based on the best predictive model obtained. Three models are employed in this study; Logistic Regression (LR) (Forward, Backward, Enter and Optimize), Decision Tree (Information Gain, Gini Index and Gain Ratio) and Artificial Neural Network with a splitting of 70-30. *This study involved 361 T2DM patients who had undergone DM screening at the Ophthalmology Clinic, UiTM Medical Specialist Centre. Results of this study show that the prevalence of DR in individuals with DN was 1.75 times more than in individuals without DN. The LR (Optimize Evolutionary) is the best model for LR with accuracy=68.42% and AUC =0.423, compared to the other models; LR Forward (Accuracy=68.42%, AUC = 0.731), LR Backward ((Accuracy=57.89%, AUC=0.487) and LR Enter (Accuracy=57.89%, AUC =0.487). The DT Information Gain is the best model for the Decision Tree model (Accuracy=92.31%, AUC=0.667) compared to the DT Gini Index (Accuracy=92.31%, AUC=0.333) and DT Gain Ratio (Accuracy=84.62%, AUC=0.50). The ANN model gives an accuracy of 68.42% and ROC=0.50. Thus, the DT Information Gain is the best model to predict the presence of DR in T2DM patients with significance factors; duration of DM, Age, diastolic BP and BMI. The significance of this study can be applied globally to promote better health understanding in predicting the presence of DR among T2DM with DN patients and future prevention.*

Keywords: *Data Mining, Diabetes Complications, Diabetic Neuropathy, Diabetic Retinopathy, Risk Factors.*

Received for review: 09-09-2024; Accepted: 27-09-2024; Published: 01-10-2024
DOI: 10.24191/mjoc.v9i2.26720

1. Introduction

Diabetes mellitus (DM) is a condition in which the body's blood glucose, also known as blood sugar, is too high (National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK), 2016; Wu *et al*., 2018). The Institute for Public Health (2020) reports that one in five individuals in

Malaysia, aged 18 years and above, have DM. Furthermore, the prevalence rate of individuals diagnosed with DM and those who did not know they had it increased from 13.4% in 2015 to 18.3% in 2019.

In Malaysia, DM is considered a serious public health problem with Type II diabetes mellitus (T2DM) now accounting for 20.8% of those over the age of 30, impacting 2.8 million people with 1.5 million (48%) of fatalities reported in 2019 (Hussein *et al*., 2015). There are two types of complications in DM; microvascular and macrovascular (Chawla et al., 2016). The most common complications for macrovascular are coronary arteries, peripheral arteries, and cerebrovascular. Meanwhile, long-term effects of DM that impact small blood vessels are known as microvascular where the most common complications are known as diabetic neuropathy, diabetic retinopathy, and diabetic nephropathy.

Diabetic retinopathy (DR) is one of the most common complications of diabetes mellitus (DM), which causes progressive retinal damage because of a microvascular complication of diabetes mellitus (Ministry of Health, 2017). DR is an eye condition that affects people who have DM when high blood sugar levels damage the blood vessels in the retina. It is reported that one out of ten patients with T2DM will get DR, a condition that might potentially result in blindness and pose a threat to eyesight (Yau *et al*., 2012). According to Ali *et al*. (2016), DR affects 12.3% of T1DM while it affects 22.3% of T2DM. Another microvascular complication is diabetic neuropathy (DN), known as a kind of nerve damage that can develop into diabetic nerve injury where high blood sugar levels can harm nerves all over the body mostly in the legs and feet (Saxena *et al*., 2021). DN is the most common serious DM complication where about half of all individuals with DM may have DN (Anandhanarayanan *et al*., 2022).

In this study, we focus on diabetic retinopathy and neuropathy. Interestingly, according to Boulton and Malik (1998), the pathogenesis of both complications is the same and as a result, the occurrence of a particular complication may indicate the presence of a disease in a person or even its severity. Both complications may be somewhat reversible if caught early and appropriately treated. According on Eyecare Medical Group (2017), the doctors believed that patients with retinopathy will have blood vessel damage and then develop neuropathy. Therefore, the doctors were concentrating on early identification and treatment to avoid blindness, which the doctors believed would then avoid the damage caused by neuropathy. Besides, patients with diabetic neuropathy also had higher rates of retinopathy, and it progresses in a way that slows down nerve conduction velocity (Kalpana *et al*., 2020). Moreover, the prevalence of retinopathy in individuals with and without neuropathy has been studied and it was discovered that those with neuropathy had a higher prevalence than those without neuropathy (Ashok *et al*., 2002; Gokhale *et al*., 2015). Therefore, this study also studies the correlation between the prevalence of DR in those with and without DPN.

There are many significant risk factors such as Age, Gender, and Body Mass Index (BMI) in DM that have been discussed in various forms. However, most people are unaware of the factors that lead to DM. According to Fiarni *et al*. (2019), factors associated with DM are Age, Gender, Body Mass Index (BMI), Blood Pressure and Duration of Diabetes Mellitus, Neuropathy, and Retinopathy. Other than that, risk factors such as Male, Hypertension, Neuropathy, Retinopathy, Nephropathy, Duration of DM, and Dyslipidaemia also contribute to DM (Bui *et al*., 2019). Data mining algorithms were also discovered to be highly helpful in prediction related to diabetes and classification models (Marinov *et al*., 2011; Khairudin *et al*., 2020; Pala, *et al*., 2014; Sunardi *et al*., 2023; Hashad *et al*., 2024, Noh *et al*., 2023). DR and DN can affect everyone with diabetes mellitus. Early diagnosis increases the chances of receiving the best cure to prevent complications. Besides, not many people are aware that it is possible to have two complications at one time. Hence, it is important to maintain a healthy lifestyle by eating healthy foods, doing more physical activity and most importantly going for check-ups regularly. Since predicting diabetic retinopathy with the presence of diabetic neuropathy is important, the data mining algorithm used in this study helps to gain a better understanding of the problem and make better decisions for the screening and treatment of the complications. Therefore, in this study, diabetic retinopathy with the presence of diabetic neuropathy will be focused on and reviewed since the association between the complications is higher. Besides, the following factors also increase the chances of developing those complications, such as blood sugar control, kidney disease, diabetes history, body mass index (BMI) and smoking. Thus, factors that contributed to DR among DN in T2DM patients were determined and predicted.

2. Materials and Methods

This section provides a brief description of the study's scope and the statistical methods used for the Chi-Square Test for Association, Logistic Regression (LR), Decision Tree (DT), and Artificial Neural Network (ANN).

2.1 Scope of study

Data used in this study is secondary data consisting of 361 T2DM patients who had undergone DM screening at the Ophthalmology Clinic, UiTM Medical Specialist Centre, between the years 2013 and 2018. The data consisted of three parts demographic profile (age, gender, ethnicity, and BMI), diabetes history on the patients (duration of DM, stroke, hypertension, and IHD) and clinical factors (systolic BP, diastolic BP and HbA1c). This study focused on predicting DR among DN patients which involves eyes and nerves and not focusing on other areas affected by diabetes mellitus such as the kidney, heart, and eyes. There are three predictive models involved in this study which are Logistic Regression, Decision Tree, and Artificial Neural Network.

2.2 Statistical Analysis

This section elaborates on the description of variables, and data mining techniques used in predicting the DR status of patients and the risk factors associated with DR among T2DM patients described. Discussion on the chi-square method and data mining models used in this study is discussed and evaluation criteria are also explained.

2.2.1 Description of Variables

Table 1 describes the variables used in this study. There are 361 diabetic patient records with 12 categorical and continuous variables

Table 1. Descriptions of Variables

Role	Variable Name	Measurement Level	Description	Frequency	Percentage (%)
Dependent variable	Presence of DR	Binary	0: No DR	204	57
			1: DR	157	43
Independent variable (Categorical)	Gender (x_1)	Binary	1: Male	217	60.1
			2: Female	144	39.9
	Ethnicity (x_2)	Nominal	1: Malay	280	77.6
			2: Chinese	40	11.1
			3: Indian	38	10.5
			4: Others	3	0.8
	Hypertension (x_3)	Binary	0: No	34	9.4
			1: Yes	327	90.6
	Stroke (x_4)	Binary	0: No	345	95.6
			1: Yes	16	4.4
	Ischaemic Heart Disease (IHD) (x_5)	Binary	0: No	228	63.2
			1: Yes	133	36.8
Independent variable (Continuous)	**Variable Name**	**Measurement Level**	**Description**	**Mean**	**Standard Deviation**
	BMI (x_6)	Continuous	Body Mass Index (BMI) (kg/m^2)	28.8	4.953
	Age (x_7)	Continuous	Patient's age during first visit (years)	58.36	9.695
	Duration of DM (x_8)	Continuous	Duration of diabetes mellitus (years)	9.72	7.792
	Systolic Blood Pressure (BP) (x_9)	Continuous	Systolic blood pressure (mmHg)	138.37	17.665
	HbA1c (x_{10})	Continuous	Glycosylated Haemoglobin Level (%)	8.67	2.037
	Diastolic Blood Pressure (BP) (x_{11})	Continuous	Diastolic blood pressure (mmHg)	77.67	11.832

2.2.2 Chi-Square Test for Association

The chi-squared test in Equation 1 is used to determine if there's a significant association between categorical variables. It compares the observed frequencies in the data with the frequencies that would be expected if the two variables had no relationship. In this study, the objective is to determine the association between diabetic retinopathy (DR) and diabetic neuropathy (DN) among T2DM patients. According to Moore *et al.* (2017), percentages provide more information than counts do by dividing the sum of each row by the sum of the table, then converting it to a percentage. Marginal distributions are the distribution of values for a categorical variable (DR and DN) among all the T2DM patients while conditional distributions are the distribution of values for no DN and no DR patients among no DN patients.

$$X^2 = \sum \frac{(Observed - Expected)^2}{Expected} \tag{1}$$

2.2.3 Data Mining Techniques

Analyses in this study were conducted using LR (Enter, Forward Backward, Optimize Evolutionary), DT (Information Gain, Gini Index, Gain Ratio) and ANN. The distribution of training and validation datasets was set for 70:30, based on previous medical studies by Montagna *et al*. (2022) and Khairudin *et al*. (2020).

2.2.3.1 Logistic Regression

Logistic regression (LR) is a type of predictive analysis that describes the relationship between a collection of independent variables as described in Table 1 and the dependent variable is the presence of DR where Y=1 (with DR) and Y=0 (without DR). According to Hosmer *et al.* (2013), the general LR model can be written as shown in Equation 2:

$$\hat{Y} = In\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \ldots\ldots + \beta_i X_i \tag{2}$$

where p = probability of an occurrence (Y=1),
β_i = estimated coefficient,
X_i =explanatory variables ($i=1,2,\ldots,11$).

2.2.3.2 Decision Tree

Decision Tree (DT) models are a machine learning technique that consists of principles for splitting a large diverse population into smaller, more homogenous groups in relation to a certain target. In this study, there are three splitting criteria chosen which are the Gini index, entropy, and gain information (Rokach & Maimon, 2006).

The Gini Index in Equation 3 is a metric employed in decision trees to quantify the level of impurity or uncertainty within a dataset. It is the sum of squares of the proportion of the classes in the node. A perfectly pure node has a Gini score of 1. A node that is evenly balanced has a Gini score of 0.5. Information gain in Equation 4 is the total the entropy of all terminal nodes in the decision tree. The gain ratio in Equation 5 is an advanced version of Information gain.

$$GINI = 1 - \sum_{i=1}^{C} (p_i)^2 \tag{3}$$

$$\text{Information gain} = -\sum p_i \, log_2 \, (p_i) \tag{4}$$

$$Gain\ Ratio = \frac{Information\ Gain\ (xi, S)}{Entropy\ (xi, S)} \tag{5}$$

p_i = probability of the ith category of the target variable occurring the in a particular node

C= total number of classes

S = original dataset

x_i= independent variables (*i=1,2,…,11*)

2.2.3.3 Artificial Neural Network

Artificial Neural Network (ANN) is one of the most well-known and extensively researched to solve estimation, classification, and prediction problems. Neural Network was illustrated on the structure of neurons in the human brain. The neural processing unit, which is hidden layer based, is the central component of the ANN model. Through the learning algorithm defined in the network, each neurone determines the ideal connection weight, W_i of each input. Using a summation computation, the neurone combined the weighted values from each input into a single value. After that, a nonlinear transfer function converts the result into the output. Every neurone in a layer is linked to every other layer's neurone as well. Every link has a scalar weight attached to it that modifies the strength of the signal that flows through it (Kukreja *et al*., 2016; Kuldeep & Anitha, 2015) as illustrated in Figure 1 and Equation 6.

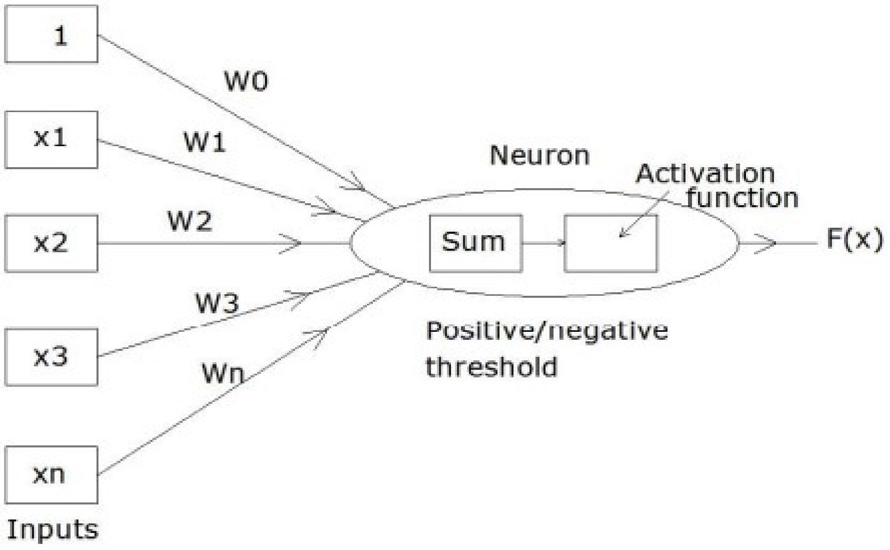


Figure 1. Model of an Artificial Neural Network

$$F\,(x) = \frac{1}{1 + e^{-\sum_{i=0}^{n} x_i W_i}} \tag{6}$$

where x_i is the independent variables (*i=1,2,…,11*) and W_i is the calculated weights.

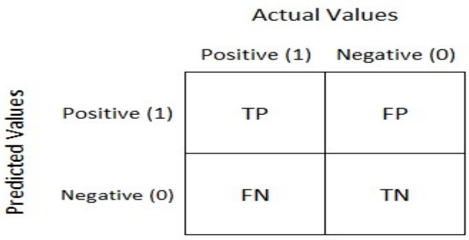
2.2.4 Evaluation of Model Performance

The prediction models were assessed on the training and validation dataset using confusion metrics to calculate accuracy (ACC) and receiver operating characteristic (ROC) index (AUC). The primary goal of this evaluation is to determine the best-performing model in predicting the presence of DR in T2DM patients.

2.2.4.1 Confusion Matrix

The performance of the model can be further investigated by using the confusion matrix. The number of correct and incorrect predictions made by the model compared with the actual classifications in the validation data is displayed in the confusion matrix illustrated by Figure 2. The matrix contains True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). Accuracy is the probability of cases that are correctly classified which is calculated as the ratio between the number of cases correctly classified and the total number of cases defined in Equation 7. The best-performing model is determined to be the one with the highest percentage accuracy (Karimi, 2021).

$$Accuracy = \frac{True\ Positive\ (TP) + True\ Negative\ (TN)}{\mathrm{TP + TN + FN + FP}} \tag{7}$$



		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 2. Confusion Matrix

2.2.4.2 Receiver Operating Curves Index (AUC)

According to Khairudin *et al.* (2020), a graph of true positive rate (sensitivity) vs. false positive rate (1-specificity) at various cut-off points is called a receiver operating characteristic (ROC). When the curve is closer to the left-hand boundary and the top border of the ROC space, the test is more accurate. The test, on the other hand, becomes less accurate when the curve approaches the ROC curve's 45-degree diagonal. The accuracy of a classifier is determined by the area under the ROC curve index (AUC), which ranges from 0.5 to 1.0, with ideal classifiers having an AUC of 1.0. Higher AUC values, in general, reflect greater prediction performance. The discrete points from the ROC curve can be calculated as shown in Equation 8.

$$\mathrm{AUC} = \sum_{i=1}^{n-1} \left[\left(\frac{\mathrm{FP}}{\mathrm{FP+TN}} \right)_{i+1} - \left(\frac{\mathrm{FP}}{\mathrm{FP+TN}} \right)_{i} \right] \times \left[\frac{\left(\frac{\mathrm{TP}}{\mathrm{TP+TN}} \right)_{i+1} - \left(\frac{\mathrm{TP}}{\mathrm{TP+TN}} \right)_{i}}{2} \right] \tag{8}$$

3. Results

In this section, descriptive analysis as well as the result and findings for the predictive model are provided. The performance of the data mining models is further discussed and compared.

3.1 Descriptive Statistics

Table 1 includes the descriptive statistics of the variables used in this study. In terms of demographic profile, the majority of the T2DM had no DR and no DN. Besides, most of the T2DM patients are male with about 217 (60.1%) while the remaining 144 (39.9%) patients are female. More than that, in this study majority of the T2DM patients, are Malay followed by Chinese, Indian and others with about 280 (77.6%), 40 (11.1%), 38 (10.5%) and 3 (0.8%) respectively. Meanwhile, for diabetic history, there are three categorical variables of independent variables which are IHD, hypertension, and stroke. Patients with T2DM have IHD of about 133 (36.8%) while those without IHD are about 228 (63.2%) patients. Interestingly, the majority of the patients with T2DM suffer from hypertension which is about 327 (90.6%) but only 34 (9.4%) are not suffering from hypertension.16 (4.4%) patients have a stroke and the remaining 345 (95.6%) did not have a stroke.

3.2 Association between Diabetic Retinopathy (DR) and Diabetic Neuropathy (DN)

The prevalence of DR in T2DM patients with DN was investigated to determine the association between DR and DN. Table 2 shows the Chi-Square test crosstabulation of DR with DN and the prevalence of DR in patients with and without DN. At a 5% significance level, the presence of DR was found to be significantly associated with the presence of DN (p=0.000 < 0.05). DR was found in 67% of patients with DN, which is 1.75 times higher than those without DN. Therefore, there is an association between the presence of DR among DN in T2DM patients, since the prevalence rate of retinopathy in neuropathy is higher.

Table 2. Crosstabulations of Diabetic Retinopathy with Diabetic Neuropathy

		DR		**Pearson**	**p-value**
		0: No DR	**1: DR**	**Chi - Square**	
DN	**0: No DN**	183 (61.6%)	114 (38.4%)	17.775	0.000
	1: DN	21 (32.8)	43 (67.2%)		

3.2 Model Comparison of Machine Learning Approaches

A comparison of model performances for each method applied in this study and significant variables were shown in Table 3. For LR models, LR Enter and LR Backward were found to have an overfitting issue, where the training ACC and AUC for both were too different from the validation results. In addition, no significant factors contributed to the presence of DR in the LR Forward. Based on the results, LR Optimize Evolutionary is selected as the best model ACC = 68.42% and AUC = 0.423 for validation dataset. Hence, the estimated model for the LR Optimize Evolutionary model is given in Equation 9 as follows:

$$P\ (Y = 1) = \frac{1}{1 + e^{-z}} \tag{9}$$

where z = 17.918 – 3.997 *(Ethnicity = Chinese)* + 0.289 *(Duration of DM)* – 0.063 *(Systolic BP)*

Based on Table 3, all the Decision Tree (DT), either Information Gain, Gini Index, and Gain Ratio show good results in ACC and AUC. DT Information Gain has the highest ACC (92.31%) and AUC (0.667) for the validation dataset. Meanwhile, the DT Gini Index has the same ACC (92.31%) as DT Information Gain but the lowest AUC (0.333) among all three models. Therefore, DT Information Gain is selected as the best model.

Overall comparison, DT Information Gain showed the highest ACC and AUC compared to the LR Optimize Evolutionary and ANN models for both the training and validation datasets, which indicated that the DT Information Gain model is the best predictive model and was able to correctly predict the presence of diabetic retinopathy among diabetic neuropathy patients as much as 92.31%.

Table 3. Comparison of Model Performances

Machine Learning	**Model**		**Accuracy (%)**	**ROC Index (AUC)**	**Significant Variables**
Logistic Regression (LR)	Enter	Training	84.44	0.896	Duration of DM, and Systolic BP
		Validation	57.89	0.487	
	Forward Selection	Training	66.67	0.602	None
		Validation	68.42	0.731	
	Backward Elimination	Training	84.44	0.898	Duration of DM, and Systolic BP
		Validation	57.89	0.487	
	Optimize Evolutionary	Training	77.78	0.871	Duration of DM, Ethnicity, and Systolic BP
		Validation	**68.42**	**0.423**	
Decision Tree (DT)	**Information Gain**	Training	96.67	0.999	Duration of DM, Age, Diastolic BP, and BMI
		Validation	**92.31**	**0.667**	
	Gini Index	Training	93.33	0.994	Duration of DM, Ethnicity, Diastolic BP, Gender, HbA1c, BMI, and IHD
		Validation	92.31	0.333	
Decision Tree (DT)	Gain Ratio	Training	96.67	0.998	Duration of DM, Systolic BP, Diastolic BP, HbA1c, BMI, and IHD
		Validation	84.62	0.500	
Artificial Neural Network	ANN	Training	82.22	0.920	Duration of DM
		Validation	68.42	0.500	

3.4 Odds Ratio Interpretation for Logistic Regression Model

Based on Table 3, LR Optimize Evolutionary is selected as the best model among the LR models. Thus, the odds ratios in Table 4, stated that Chinese patients with DN will have higher chances of getting diabetic retinopathy compared to the other ethnicity. The longer diabetic neuropathy patients have diabetes, the higher their chances of developing diabetic retinopathy. Also, the higher systolic blood pressure of diabetic neuropathy patients is more likely to have diabetic retinopathy.

Table 4. Odds Ratio Interpretation for Logistic Regression Model

Variable Name	Odds Ratio	Interpretation
Ethnicity Chinese	$e^{-3.997} = 0.0184$	The odds ratio of 0.0184 for Chinese patients indicates that if Chinese patients with diabetic neuropathy have diabetes, the greater the chances of having diabetic retinopathy. For every one unit increase in ethnicity = Chinese, the odds of having diabetic retinopathy increased by 1.84%.
Duration of DM	$e^{0.289} = 1.335$	The odds ratio of 1.335 for duration of DM indicates that the longer the diabetic neuropathy patients have diabetes, the greater the chances of having diabetic retinopathy. For every one year increase in duration of DM, the odds of having diabetic retinopathy increased by 33.5%.
Systolic BP	$e^{-0.063} = 0.939$	The odds ratio of 0.749 for systolic blood pressure indicates that the higher systolic blood pressure of diabetic neuropathy patients is more likely to have diabetic retinopathy. For every one unit increase in systolic blood pressure, the odds of having diabetic retinopathy increased by 93.9%.

3.5 Significant Factors to Predict DR among DN in T2DM Patients

Based on Table 3, the DT Information Gain is the best model. The DT Information Gain flowchart (Figure 3) revealed that if the duration of DM is more than 5.5 years and age is less than or equal to 56.5 years, the presence of DR is true (Y=1), if the duration of DM is more than 5.5 years, age is more than 56.5 years and diastolic blood pressure is more than 90mmHg, the presence of DR is true (Y=1), if the duration of DM is more than 5.5 years, age is more than 56.5 years diastolic blood pressure is between 73.5 and 90mmHg, duration of DM less than or equals to 10.5 years, the presence of DR is true (Y=1), if the duration of DM is more than 5.5 years, age is more than 56.5 years diastolic blood pressure is between 73.5 and 90mmHg, BMI less than or equals to 32 kg/m^2, the presence of DR is true (Y=1) and if the duration of DM is less than or equals to 5.5 years, diastolic blood pressure is less than or equals to 81mmHg and duration of DM more than 4.5 days, the presence of DR is true (Y=1).

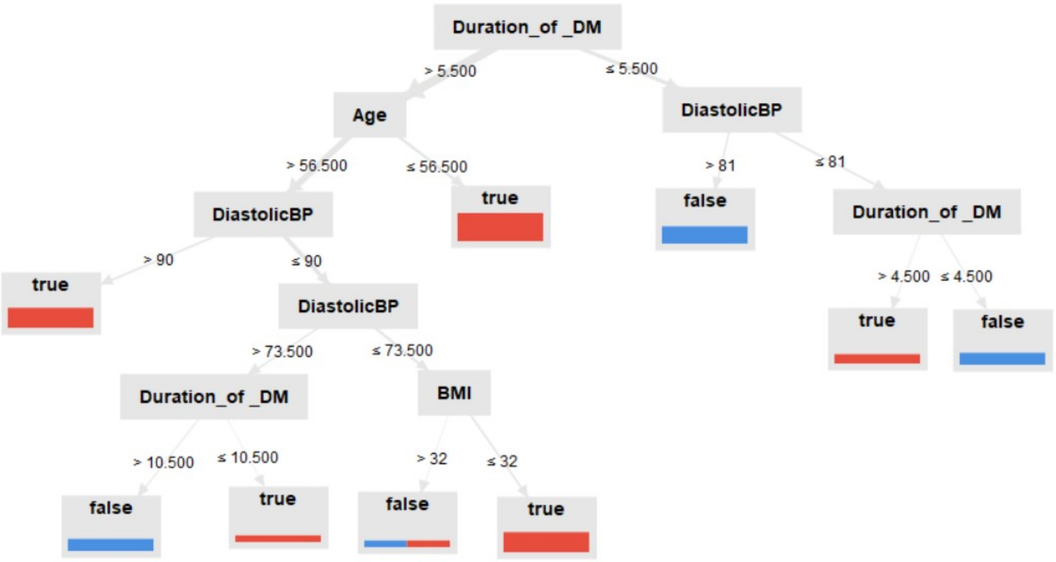


Figure 3. Decision Tree (Information Gain)

4. Conclusion

This study was conducted to predict the presence of diabetic retinopathy (DR) among diabetic neuropathy (DN) patients in T2DM patients. The Chi-Square Test of association concluded that DR was found in 67% of patients with DN, which is 1.75 times higher than in patients without DN. This is supported by Boulton and Malik (1998) and Eyecare Medical Group (2017). Model comparison based on ACC and AUC concluded that DT Information Gain was the best model in predicting the presence of DR. This result is similar to Fiarni *et al.* (2019). Hence, the significant factors obtained are duration of DM, Age, Diastolic BP, and BMI. This is also supported by Abougalambou and Abougalambou (2015), Ali *et al.* (2016) and Mimi *et al.* (2003).

A sample size of just above 350 is considered small in data mining. However, in medical studies, this is quite common, and although with these constraints, this study achieved all the objectives aligned. Thus, it is proposed further research to increase the sample size to allow the generalization of the findings and produce more meaningful results. In addition, sensitivity and specificity can be added to measure the performance of the model. Besides, other risk factors could be included, i.e. family history, cholesterol, and smoking. DR levels can also be classified into different severity levels (mild, moderate, and severe) (Slakter *et al.*, 2015), and types of DN specified (peripheral neuropathy, autonomic neuropathy, proximal neuropathy, and focal neuropathy) (Bui *et al.*, 2019). The results might be richer with new information. It is also proposed to apply other data mining techniques i.e., Naïve Bayes, SVM, and Random Forest, to develop significantly better prediction models.

Acknowledgment

The authors would like to thank the Ophthalmology Clinic, UiTM Medical Specialist Centre for providing the dataset in this study.

Author Contribution

Nur Balqis Oon collected and analyzed the data for the study. Zuraida Khairudin is drafting the manuscript. Hezlin Aryani Abd Rahman, Norbaizura Kamarudin and Nur Syamimi Haji Abu Bakar have participated in revision of the intellectual content. Dr Azimah from the Faculty of Medicine, UiTM providing her expert opinion.

Conflict of Interest

The authors have no conflicts of interest to declare.

References

Abougalambou, S. S. I., & Abougalambou, A. S. (2015). Explorative study on diabetes neuropathy among type II diabetic patients in Universiti Sains Malaysia Hospital. Diabetes & Metabolic Syndrome Clinical Research & Reviews, 6(3), 167–172. https://doi.org/10.1016/j.dsx.2012.09.002.

Ali, M. H. M., Draman, N., Mohamed, W. M., Yaakub, A., & Embong, Z. (2016). Predictors of proliferative diabetic retinopathy among patients with type 2 diabetes mellitus in Malaysia as detected by fundus photography. Journal of Taibah University Medical Sciences, 11(4), 353–358. https://doi.org/10.1016/j.jtumed.2016.03.002.

Anandhanarayanan, A., Teh, K., Goonoo, M., Tesfaye, S., & Selvarajah, D. (2022, March 15). Diabetic neuropathies. Endotext - NCBI Bookshelf. https://www.ncbi.nlm.nih.gov/books/NBK279175/?report=classic.

Ashok, S., Ramu, M., Deepa, R., & Mohan, V. (2002). Prevalence of neuropathy in type 2 diabetic patients attending a diabetes centre in South India. ResearchGate. https://www.researchgate.net/publication/11223182.

Bui, H. D. T., Jing, X., Lu, R., Chen, J., Ngo, V., Cui, Z., Liu, Y., Li, C., & Ma, J. (2019). Prevalence of and factors related to microvascular complications in patients with type 2 diabetes mellitus in Tianjin, China: a cross-sectional study. Annals of Translational Medicine, 7(14), 325. https://doi.org/10.21037/atm.2019.06.08.

Boulton, A. J., & Malik, R. A. (1998). DIABETIC NEUROPATHY. Medical Clinics of North America, 82(4), 909–929. https://doi.org/10.1016/s0025-7125(05)70029-8.

Chawla, A., Chawla, R., & Jaggi, S. (2016). Microvasular and macrovascular complications in diabetes mellitus: Distinct or continuum? Indian Journal of Endocrinology and Metabolism, 20(4), 546. https://doi.org/10.4103/2230-8210.183480.

Eyecare Medical Group. (2017). Diabetic retinopathy and neuropathy. Eyecare Medical Group. https://www.eyecaremed.com/news/diabetic-retinopathy-and-neuropathy/

Fiarni, C., Sipayung, E. M., & Maemunah, S. (2019). Analysis and Prediction of Diabetes Complication Disease using Data Mining Algorithm. Procedia Computer Science, 161, 449–457. https://doi.org/10.1016/j.procs.2019.11.144.

Gokhale, V. S., Chaudhari, N. C., Kakrani, A. L., & Shah, B. P. (2015). High incidence of retinopathy in neuropathy proven diabetic patients: A cohort study. International Journal of Medicine and Public Health, 5(4), 289. https://doi.org/10.4103/2230-8598.165952.

Hashad, A. A., Wah, K. K., Alnoor, A., & Chew, X. (2024). Exploratory Analysis With Association Rule Mining Algorithms In The Retail Industry. Malaysian Journal of Computing (MJoC), 9(1), 1746–1758. https://doi.org/10.24191/mjoc.v9i1.21433.

Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied Logistic Regression. In Wiley series in probability and statistics. https://doi.org/10.1002/9781118548387.

Hussein, Z., Taher, S. W., Singh, H. K. G., & Swee, W. C. S. (2015). Diabetes care in Malaysia: problems, new models, and solutions. Annals of Global Health, 81(6), 851. https://doi.org/10.1016/j.aogh.2015.12.016.

Institute for Public Health. (2020). National health and morbidity survey (NHMS) 2019: Non-communicable diseases, healthcare demand, and health literacy: Key findings.

Kalpana, R., Kanmani, K., & Anuradha, P. (2020). Association of peripheral neuropathy with retinopathy in diabetic patients. Indian Journal of Clinical and Experimental Ophthalmology, 4(1), 115–119. https://doi.org/10.18231/2395-1451.2018.0026.

Karimi, Z. (2021). Confusion Matrix. ResearchGate. https://www.researchgate.net/publication/355096788.

Khairudin, Z., Razak, N. a. A., Rahman, H. a. A., Kamaruddin, N., & Aziz, N. a. A. (2020). Prediction Of Diabetic Retinopathy Among Type II Diabetic Patients Using Data Mining Techniques. Malaysian Journal of Computing, 5(2), 572. https://doi.org/10.24191/mjoc.v5i2.10554.

Kukreja, H. N., Bharath, N., Siddesh, C. S., & Kuldeep, S. (2016). An introduction to artificial neural network. International Journal of Engineering and Techniques, 1(5).

Kuldeep, S., & Anitha, G. S. (2015). Neural network approach for processing substation alarms. International Journal of Power Electronics Controllers and Converters. Retrieved from www.journalspub.com.

Marinov, M., Mosa, A. S., Yoo, I., & Boren, S. A. (2011). Data-mining technologies for diabetes: a systematic review. Journal of diabetes science and technology, 5(6), 1549–1556. https://doi.org/10.1177/193229681100500631.

Mimi, O., Teng, C. L., & Chia, Y. C. (2003). The prevalence of diabetic peripheral neuropathy in an outpatient setting. The Medical journal of Malaysia, 58(4), 533–538.

Ministry of Health. (2017). Diabetic retinopathy screening module (2nd ed.).

Montagna, S., Pengo, M. F., Ferretti, S., Borghi, C., Ferri, C., Grassi, G., Muiesan, M. L., & Parati, G. (2022). Machine Learning in Hypertension Detection: A study on World Hypertension Day data. Journal of Medical Systems, 47(1). https://doi.org/10.1007/s10916-022-01900-5.

Moore, D. S., Notz, W. I., & Fligner, M. A. (2017). The basic practice of statistics (6th ed.). W. H. Freeman and Company.

National Institute of Diabetes and Digestive and Kidney Diseases (NIDDK). (2016, December).Diabetes.https://www.niddk.nih.gov/healthinformation/diabetes/overview/what-is-diabetes.

Noh, S. S. M., Ibrahim, N., Mansor, M. M., & Yusoff, M. (2023). Hybrid filtering methods for feature selection in high-dimensional cancer data. International Journal of Power Electronics and Drive Systems/International Journal of Electrical and Computer Engineering, 13(6), 6862. https://doi.org/10.11591/ijece.v13i6.pp6862-6871.

Pala, Tuba and Camurcu, Ali Yilmaz (2014) *Evaluation of data mining classification and clustering techniques for diabetes / Tuba Pala and Ali Yilmaz Camurcu.* Malaysian Journal of Computing (MJoC), 2 (1): 4. pp. 37-45. ISSN 2231-7473.

Rokach, L., & Maimon, O. (2006). Decision Trees. In Springer eBooks (pp. 165–192). https://doi.org/10.1007/0-387-25465-x_9.

Saxena, A., Jain, M., & Shrivastava, P. (2021). Data mining techniques based diabetes prediction. Indian Journal of Artificial Intelligence and Neural Networking, 1(2), 29–35. https://doi.org/10.35940/ijainn.b1012.041221.

Slakter, J. S., Schneebaum, J. W., & Shah, S. A. (2015). Digital Algorithmic Diabetic Retinopathy Severity Scoring System (An American Ophthalmological Society Thesis). Transactions of the American Ophthalmological Society, 113, T9.

Sunardi, Fadlil, A., & Perdana Kusuma, N. M. (2023). Cyber fraud profiling with routine activity theory using data mining techniques. Malaysian Journal of Computing (MJoC), 8(2), 1517–1533. ISSN 2600-8238.

Wu, H., Yang, S., Huang, Z., He, J., & Wang, X. (2018). Type 2 diabetes mellitus prediction model based on data mining. Informatics in Medicine Unlocked, 10, 100–107. https://doi.org/10.1016/j.imu.2017.12.006.

Yau, J. W., Rogers, S. L., Kawasaki, R., Lamoureux, E. L., Kowalski, J. W., Bek, T., Chen, S., Dekker, J. M., Fletcher, A., Grauslund, J., Haffner, S., Hamman, R. F., Ikram, M. K., Kayama, T., Klein, B. E., Klein, R., Krishnaiah, S., Mayurasakorn, K., O'Hare, J. P., Wong, T. Y. (2012). Global prevalence and major risk factors of diabetic retinopathy. Diabetes Care, 35(3), 556–564. https://doi.org/10.2337/dc11-1909.