Malaysian Journal of Computing, 5 (2): 572-586, 2020
Copyright © UiTM Press
eISSN: 2600-8238 online

PREDICTION OF DIABETIC RETINOPATHY AMONG TYPE II DIABETIC PATIENTS USING DATA MINING TECHNIQUES

Zuraida Khairudin[1], Nurfatin Adila Abdul Razak[2], Hezlin Aryani Abd Rahman[3], Norbaizura Kamarudin[4] and Nor Azimah Binti Abd Aziz[5]

[1,2,3,4]*Faculty of Computer and Mathematical Sciences, UiTM Shah Alam, Selangor, Malaysia*
[5]*Faculty of Medicine, UiTM Sg Buloh, Selangor, Malaysia*
[1]zuraida_k@tmsk.uitm.edu.my, [2]nurfatinrazak9570@gmail.com, [3]hezlin@tmsk.uitm.edu.my,
[4]norbaizura@tmsk.uitm.edu.my, [5]azimah80@uitm.edu.my

ABSTRACT

Diabetic retinopathy is one of the leading causes of visual disability and blindness worldwide. It is estimated that 4.8% out of 37 million cases of blindness were due to diabetic retinopathy, globally. It affects patients suffering from prolonged diabetes, which probably results in permanent blindness. The earliest symptoms surfaced when the patients have vision problems. Therefore, regular eyes examination and early intervention normally controls this disease. Many studies for early intervention and prevention of diabetic retinopathy uses various predictive models. The booming of database and digital storage technology creates an abundance of health records. Thus, data mining techniques helps uncover meaningful patterns while attending to sensitivity health record issues. Hence, this study took the data mining approach in predicting the presence of diabetic retinopathy narrowing to only Type II diabetic patients as well as to determine the risk factors that contribute to the presence of diabetic retinopathy. The data mining models selected for this study is the Logistic Regression, Decision Tree and Artificial Neural Network. The dataset of 361 Type II diabetic patients from Ophthalmology Clinic, UiTM Medical Specialist Centre were selected between January 2014 to December 2018, consists of 17 variables. The result shows that the Logistic Regression using Forward selection method model is the best model since it had the highest sensitivity (Sen=50.0%), specificity (Spe=79.03%) and accuracy rate (Acc=66.36%) on the validation dataset compared to other Logistic Regression selection options. Meanwhile among the Decision Tree models, DT using Gini is the best model. Logistic Regression (Forward) and Decision Tree (Gini) were then compared with Artificial Neural Network model (Sen=56.25%, Spe=70.97%, Acc=64.55%). The results demonstrated that Logistic Regression using Forward selection method was the best model to predict the presence of diabetic retinopathy among the Type II diabetic patients compared to other models. The significant risk factors associated with the presence of the diabetic retinopathy obtained are duration of diabetes, HbA1C level, diabetic foot ulcer, nephropathy, and neuropathy.

Keywords*: Data Mining. Diabetic Retinopathy, Predictive Modelling, Type II Diabetes Mellitus.*

Received for review: 30-06-2020; Accepted: 18-09-2020; Published: 29-09-2020

1. Introduction

Diabetic retinopathy (DR) occurs when the high glucose level causes damage to the tiny blood vessels inside the retina. DR is a complication of diabetes that is characterized by signs of microaneurysms, haemorrhages, cotton wool spots, intraretinal microvascular abnormalities, neovascularisation and signs of increased retinal vascular permeability (American Academy of Ophthalmology (2016)). These complications of diabetes are classified into microvascular and macrovascular complications. Microvascular complications of diabetes affect small blood vessels while macrovascular complications affect larger blood vessels. DR is the example of microvascular complication including nephropathy and neuropathy. Macrovascular complications can include ischemic heart disease, chronic heart failure and peripheral arterial disease (Lee, 2013). Visual symptoms of this disease are not apparent during the early stage and if it is not treated in time, DR can cause blindness which has affected millions of people worldwide (Gadekallu *et al*., 2020).

In Malaysia, the common cause of visual loss among the working age persons is related to diabetic eye disease. Although it is an alarming fact, the National Health Malaysian Survey (2006) reported that approximately only 45% of the diabetic patients had undergone an eye examination at least once after their diagnosis of diabetes (Letchuman *et al*., 2010). The duration of having DM is closely linked to the prevalence of DR. There were less than 5% had retinopathy at diagnosis, while the prevalence increased up to 40 - 50% after 10 years. More than 60% of Type II Diabetes Melitus (T2DM) patients were expected to have diabetic retinopathy after 20 years (Clinical Practice Guidelines, 2011). Therefore, it is advisable for diabetic patients to be screened at least every two years.

Numerous studies have been conducted to investigate the risk factors contributing to the presence of diabetic retinopathy by researchers such as Shriwas *et al.* (1996), Abougalambou & Abougalambou (2015) and Harb *et al*. (2018). Studies employing data mining techniques to identify the causes that associated with diabetic retinopathy among T2DM patients among others are studied by Mohd Ali *et al.* (2016), Senthilvel & Radhakrishnan (2012), Abougalambou & Abougalambou (2015), Mallika *et al*. (2011) and Senthilvel & Radhakrishnan (2012). Other studies applied univariate and multivariate survival analysis to determine the relationship between early markers of diabetes control and diabetic retinopathy (Diallo *et al*., 2018). Comparison of models were proposed by Ananthapadmanabhan & Parthiban (2014) who applied Naive Bayes and Support Vector Machine (SVM) on T2DM patients. Naive Bayes was more effective than SVM as it has the highest percentage of correct predictions by 83.37%. Mahendran & Dhanasekaran (2015) also used predictive model to compare the best model to assess the severity of DR by using SVM and Probabilistic Neural Network (PNN). The results revealed that SVM classifiers were able to identify all the normal classes compared to PNN. The average classification accuracy for SVM and PNN were 97.89% and 94.76% respectively. Tsao et al. (2018) predicted DR on T2DM patients using SVM, decision trees, ANN and logistic regressions. In this study, the prediction performance of SVM was better compared to other machine learning algorithms since it had the highest accuracy rate (0.817) and area under the receiver operating characteristics curve AUC (0.744). Reddy *et al.* (2020) used deep neural network for DR classification. The analysis revealed that this generated classifications results with enhanced accuracy.

Though data mining technique has been applied widely in DR patients, the focus of this study is to predict the presence of diabetic retinopathy, narrowing to only among the Type II diabetic (T2DM) patients. As technology and digitization grows, there is a large database and digital storage of health records made available. Data mining technique has an important role in uncovering patterns existing in these health records to provide more interesting insights in medical diagnosis and prognosis for various ailments. Due to the sensitivity of the health records, the data mining techniques are suitable to predict the diseases accurately. Therefore, this study compares the performance of data mining models in predicting DR patients specifically in Malaysia, which thus far, is still lacking. Hence, the best predictive model using data mining techniques (Logistic Regression, Decision Tree and Artificial Neural Network) in predicting DR status patients, risk factors associated to DR among T2DM patients can be identified.

2. Materials and Methods

In this section we will further discuss the scope of study and statistical analyses used in predicting DR status patients, risk factors associated to DR among T2DM patients can also be identified.

2.1 Scope of study

This study focused on Type II diabetic patients (T2DM) above 18 years old who had undergone diabetic retinopathy (DR) screening at Ophthalmology Clinic, UiTM Medical Specialist Centre during first visit. The sample data were from January 2014 to December 2018. Socio-demographic data and clinical and metabolic profiles of 361 T2DM patients with 17 variables (as described in Table 1) were obtained from the manual doctors endorsed medical records which was transferred to digital medical records. The eye with more severe retinopathy was used to represent the overall retinopathy severity for particular patients.

This study compared the performance of three predictive models (Logistic Regression, Decision Tree and Artificial Neural Network) in predicting the presence of diabetic retinopathy. Based on the best predictive model, significant factors that contribute to the presence of diabetic retinopathy were determined.

2.2 Statistical Analysis

In this section we will elaborate on description of variables, missing values treatment and the data mining techniques used in predicting DR status patients, risk factors associated to DR among T2DM patients.

2.2.1 Description of Variables

Table 1 shows the independent variables of this study are demographic profile, diabetic history and blood result. The presence of Diabetic Retinopathy is the dependent variable of this study.

Table 1. Description and Distribution of Variables

Role	Variable Name	Description	Frequency	Percentage
Dependent variable	Presence of DR	0: No DR	204	56.51%
		1: DR	157	43.49%
Independent variable	Gender	1: Male	217	60.11%
		2: Female	144	39.89%
	Ethnicity	1: Malay	280	77.56%
		2: Chinese	40	11.08%
		3: Indian	38	10.53%
		4: Others	3	0.83%
	Hypertension	0: No	34	9.20%
		1: Yes	327	90.58%
	Nephropathy	0: No	276	76.45%

		1: Yes	85	23.55%
	Neuropathy	0: No	297	82.27%
		1: Yes	64	17.73%
	Stroke	0: No	345	95.57%
		1: Yes	16	4.43%
	Ischaemic Heart Disease (IHD)	0: No	228	63.16%
		1: Yes	133	36.84%
	Diabetic Foot	0: No	336	93.07%
		1: Yes	25	6.93%
	Dyslipidemia	0: No	47	13.02%
		1: Yes	314	86.98%
	Asthma	0: No	341	94.46%
		1: Yes	20	5.54%

2.2.2 Missing Value Treatment

Data cleaning or screening process was the first stage in checking the existence of missing values in the DR dataset. If the missing values are not handled properly, the researcher may draw an inaccurate inference about the data. Other than that, it can lead to poor model performance at the model development stage. In this study, variables with over 15% missing information will be excluded from the model as suggested by Masconi *et al.* (2015). Data imputation is commonly used to handle the missing values where this process assigns new values to the missing items. For interval inputs, the missing values were replaced with the mean of the non-missing values. Meanwhile, for categorical inputs, the missing values were replaced with the most frequent category (SAS Institute Inc., 2010).

2.2.3 Data Mining Techniques

For Logistic Regression, Decision Tree and Artificial Neural Network analysis, 70% of the data was used as the training set and the remaining 30% for the validation set.

2.2.3.1 Logistic Regression

Logistic Regression model is one type of regression model that is suitable to use when the response variable has two possible outcomes or dichotomous responses (Park, 2013). The objective of conducting the logistic regression is to estimate the probability of occurrence of an event such as patients having diabetic retinopathy, (Y=1). The general logistic regression model can be written as (Hosmer & Lemeshow, 2000):

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 ..... + \beta_i X_i \qquad (1)$$

where *p* is the probability of an event (Y=1), β_i is the estimated coefficient and X_i is the explanatory variables ($i=1,2,\ldots,k$). According to Haque *et al.* (2018), there are four types of model selection in logistic regression which are known as Enter, Stepwise, Forward and Backward.

2.2.3.2 Decision Tree

According to Wu *et al*. (2008), a decision tree is a flowchart like tree structure where each internal node denotes a test on an attribute, each branch represents outcome of the test and leaf nodes represents the classes (having DR or not having DR group). The potential split is measured based on the one that increases the purity of the target variable in the children by the greatest amount (Matignon, 2007). Three types of model used in this study: Gini index, Entropy and Chi-square test.

2.2.3.3 Artificial Neural Network

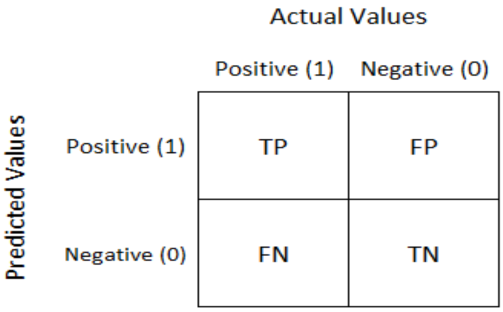
Recently, many studies applied ANN model to solve estimation, classification and prediction problems whereby this model is a computational based information system that mimic the structures of a biological nervous system such as the human brain. ANNs can discover and learn interrelated patterns between input data sets and corresponding target values (Lek & Gue, 1999). The core element in ANN model is the neural processing unit located in hidden layer. Each neuron determine the optimal connection weight, $w=(w_1,w_2,\ldots,w_n)$ of each input through the learning algorithm set in the network. The neuron aggregated the weighted value form each input into a single value using summation calculation. Then, the result is transformed into the output through a nonlinear transfer function. Every neuron in a layer is connected to every neuron in the adjacent layers. Each interconnection has associated with a scalar weight which acts to modify the strength of the signal passing through it (Balaji & Baskaran, 2013).

2.2.4 Evaluation of Model Performances

The performance of the predictive models by using validation data set were compared based on accuracy, sensitivity, specificity, misclassification rate and also receiver operating curves (ROC). The main goal of the matrix is to determine which model is the most accurate and have highest performances classification for diabetic retinopathy.

2.2.4.1 Confusion Matrix

The confusion matrix displays the number of correct and incorrect predictions made by the model compared with the actual classifications in the validation data (Witten & Frank, 2005). The matrix contains True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN), as illustrated in Figure 1. The accuracy is the proportion of the total number of predictions that were correct and calculated as the ratio between the number of cases correctly classified and the total number of cases (Equation 2). Sensitivity is the proportion of actual positive cases which are correctly predicted (Equation 3), and specificity is the proportion of actual negative cases which are correctly predicted (Equation 4). The misclassification rate is error that indicates the probability of cases was classified incorrectly (Equation 5). Model with the highest accuracy, sensitivity and specificity, and lowest misclassification rate were selected as the best model.



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

Figure 1. Confusion Matrix

$$Accuracy = \frac{TP + TN}{TP + TN + FN + FP} \tag{2}$$

$$Sensitivity = \frac{TP}{TP + FN} \tag{3}$$

$$Specificity = \frac{TN}{TP + FP} \tag{4}$$

$$Misclassification\ \ Rate = \frac{FP + FN}{TP + TN + FN + FP} \tag{5}$$

2.2.4.2 Receiver Operating Curves (ROC)

ROC is a graph of true positive rate (sensitivity) against false positive rate (1-specificity) at various cut-off points. A test is more accurate when the curve is closer to the left-hand border and the top border of the ROC space. On the other hand, the test is less accurate when the curve comes closer to 45-degree diagonal of the ROC curve. The accuracy is measured based on area under the ROC curve (AUC) which range from 0.5 to 1.0 where perfect classifiers give AUC equal to 1.0. In general, higher AUC values indicate better prediction performance (Witten & Frank, 2005).

3.0 Results

In this section we will discuss the results obtained from the analysis performed in this study to predict the DR status patients, risk factors associated to DR among T2DM patients.

3.1 Descriptive Statistics

Table 1 presents the detailed results for categorical variables in this study. For the demographic profile, majority of the patients are male which is about 217 (60.11%) out of 361 patients while 144 (39.89%) are female patients. In this study, Malay patients were the majority. Furthermore, most of the diabetic patients have hypertension, consisting about 327 (90.58%) out of 361 while patients. About 86.98% of the patients suffered from dyslipidemia (a total of 314 out of 361). However, about 276 (76.45%) of them have no presence of nephropathy while only 85 (23.55%) patients are with nephropathy. Other than that, 345 (95.57%) of the patients never had stroke and only 16 (4.43%) of them had stroke. Majority of the patients do not have diabetic foot ulcer and asthma in which only 25 (6.93%) and 20 (5.54%) out of 361 are with diabetic foot ulcer and asthma, respectively. There are 133 (36.84%) patients with IHD and majority 228 (63.16%) of them do not suffer from IHD.

3.2 Model Comparison

Model comparison results are thoroughly discussed in this section which details for logistic regression, decision tree and artificial neural network. The results of training and validation performance (sensitivity, specificity, accuracy and ROC index) are compiled and discussed. The training dataset is compared to the validation dataset to identify overfitting and underfitting of the model results. Training dataset is used to obtain the model and validation dataset is to test the validity of the model obtained.

3.2.1 Model Comparison Logistic Regression

The summary of model performances and significant variables is shown in Table 2. Based on the results, Logistic Regression (Forward) is selected as the best model since it has the highest sensitivity (50.0%), specificity (79.03%) and accuracy rate (66.36%) on the validation dataset compared to other selection methods. The variable is statistically significant when p-value is less than 0.05.

Table 2. Comparison of Model Performance (Logistic Regression)

Model		Sensitivity	Specificity	Accuracy Rate	ROC Index	Significant Variables
Enter	Training	70.64	80.99	76.49	0.833	Duration of DM, Nephropathy, Neuropathy and HbA1C
	Validation	**50**	74.19	63.64	0.644	
Stepwise	Training	66.97	81.69	75.3	0.819	Duration of DM, Diastolic BP, Nephropathy, Neuropathy, Diabetic Foot and HbA1C
	Validation	47.92	74.19	62.73	0.693	
Forward	Training	65.14	80.28	73.71	0.816	Duration of DM, Nephropathy, Neuropathy, Diabetic Foot and HbA1C
	Validation	**50**	**79.03**	**66.36**	0.709	
Backward	Training	61.47	80.28	72.11	0.812	Duration of DM, Nephropathy, Neuropathy, Diabetic Foot and HbA1C
	Validation	47.92	77.42	64.55	**0.717**	

The estimated model for Logistic Regression (Forward) is as follows:

$$P(Y=1)=\frac{1}{1+e^{-z}} \tag{6}$$

where z = – 0.3974 – 0.0292 (Age) – 0.7343 (Diabetic Foot = No) + 0.1554 (Duration of DM) + 0.1853 (HbA_{1C}) – 0.4462 (Nephropathy = 0) – 0.5322 (Neuropathy = 0)

Based on Equation 6, the significant factors can be discussed as follows:

Table 3. Odds Ratio Interpretation for Logistic Regression Model.

Variable	**Odds ratio**	**Interpretation**
Age	$e^{-0.0292} = 0.9712$	For every one-year increase in age, the odds of having DR decreased by 2.88%
Diabetic Foot	$e^{-0.7343} = 0.480$	Patients without diabetic foot ulcer (Diabetic Foot=No) are less likely to have diabetic retinopathy compared to patients who suffered from diabetic foot ulcer (Diabetic Foot=Yes).
Duration of DM	$e^{0.1554} = 1.168$	For every one-year increase in duration of DM, the odds of having diabetic retinopathy increased by 16.8%.
HbA_{1C}	$e^{0.1853} = 1.204$	For everyone unit increase in HbA_{1C} level, the odds of having diabetic retinopathy increased by 20.4%.
Nephropathy	$e^{-0.4462} = 0.640$	The odds of patients without nephropathy (Nephropathy=No) are less likely to have diabetic retinopathy compared to patients who suffered from nephropathy (Nephropathy=Yes).
Neuropathy	$e^{-0.5322} = 0.587$	The odds of patients without neuropathy (Neuropathy = No) are less likely to have diabetic retinopathy compared to patients who suffered from neuropathy (Neuropathy=Yes).

3.2.2 Model Comparison Decision Tree

The summary of Decision Tree Analysis is compiled in Table 4. Gini Index is chosen as the best model as it has the highest specificity (70.97%) and accuracy (64.55%) rate and the lowest misclassification rate (0.3546) compared to Entropy and Chi-Square as shown in Table 4. The variable is statistically significant when p-value is less than 0.05.

Table 4. Comparison of Model Performance (Decision Tree)

Model		**Sensitivity**	**Specificity**	**Accuracy Rate**	**ROC Index**	**Significant Variables**
Gini	Training	70.64	83.8	78.09	0.832	Age, BMI, Duration of DM, Systolic of BP, Nephropathy and HbA1C
	Validation	56.25	70.97	64.55	0.635	
Entropy	Training	66.06	85.21	76.89	0.814	Age, Duration of DM, Systolic of BP, Nephropathy and HbA1C
	Validation	54.17	70.97	63.64	0.671	
Chi-Square	Training	88.07	54.23	68.92	0.711	Duration of DM
	Validation	72.92	51.61	60.91	0.623	

3.2.3 Artificial Neural Network (ANN)

The result of ANN analysis is shown in Table 5. The performance of ANN model is better in training dataset compared to the validation dataset where the sensitivity rate is 86.24%, specificity is 87.32% and accuracy rate is 86.85%. From prediction performance analysis for ANN model, 56.25% of the patients having diabetic retinopathy samples were correctly classified as having diabetic retinopathy while 70.97% of patients without diabetic retinopathy samples were correctly classified as not having diabetic retinopathy. The variable is statistically significant when p-value is less than 0.05.

Table 5. Artificial Neural Network Result

Model		**Sensitivity**	**Specificity**	**Accuracy Rate**	**ROC Index**	**Significant Variables**
Artificial Neural Network (ANN)	Training	86.24	87.32	86.85	0.919	Duration of DM, Nephropathy, Neuropathy, Diabetic Foot and HbA1C
	Validation	56.25	70.97	64.55	0.667	

3.3 Performance Indicator

The best models from Logistic Regression (Forward) and Decision Tree using Gini Index (DT-Gini) were then compared with Artificial Neural Network (ANN) model. The performances of these models were compared based on the sensitivity, specificity, accuracy rate and ROC index. Table 6 summarized the result of prediction performances for these three predictive models. Based on the results, DT-Gini and ANN had the highest sensitivity rate (56.25%) in the validation dataset. Other than that, ANN performed the best in the training dataset compared to other models with the highest specificity (86.24%), specificity (87.32%), accuracy rate (86.85%) and also ROC index (0.919). However, among the three data mining techniques, Logistic Regression (Forward) performed the best on the validation since it achieved the best prediction performance with 79.03% and 66.36% in specificity and accuracy rate, respectively. This indicates that Logistic Regression (Forward) model was able to correctly predict the presence of diabetic retinopathy as much as 66.36%. Besides, from the classification performance analysis for Logistic Regression (Forward), 50.0% of the diabetic retinopathy patients were correctly classified as having diabetic retinopathy, while 79.03% of the non-diabetic retinopathy patients were correctly classified as not having diabetic retinopathy.

Table 6. Comparison of best model results

Model		**Performance Measures**			
		Sensitivity	**Specificity**	**Accuracy Rate**	**ROC Index**
Forward	Training	65.14	80.28	73.71	0.816
	Validation	50.00	**79.03**	**66.36**	0.709
Gini	Training	70.64	83.8	78.09	0.832
	Validation	**56.25**	70.97	64.55	0.635
ANN	Training	86.24	87.32	86.85	0.919
	Validation	56.25	70.97	64.55	0.667

The ROC charts for Logistic Regression (Forward), Decision Tree (Gini) and ANN models are shown in Figure 2. In the validation dataset, the curve of Logistic Regression using Forward selection method is slightly highest than Gini and ANN. Furthermore, Logistic Regression (Forward) had the highest ROC index, which is 0.709, followed by ANN (0.667) and Gini (0.635). This indicates that Forward model had better classification performance than Gini and ANN. Thus, Logistic Regression using Forward selection method was chosen as the best model to predict the presence of DR among the Type II diabetic patients.

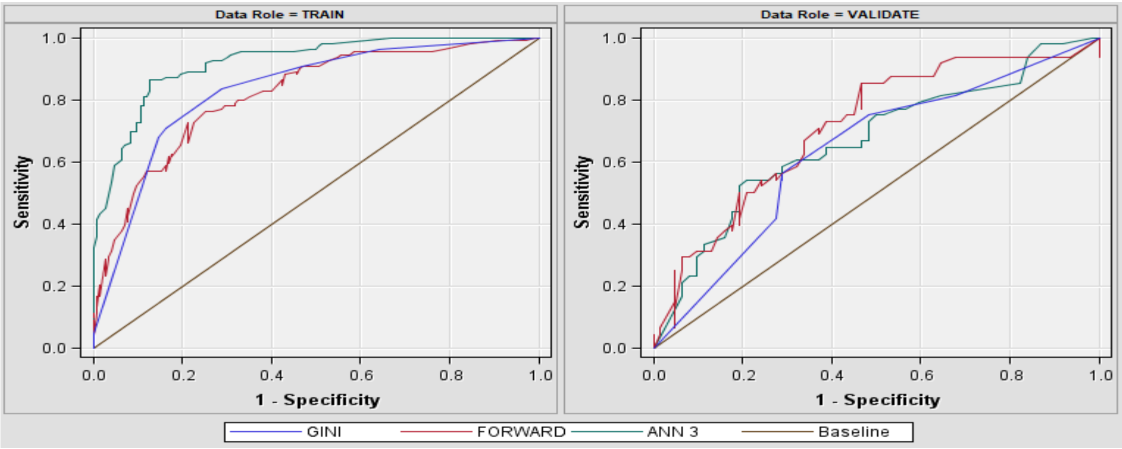


Figure 2. ROC Chart for Logistic Regression (Forward), Decision Tree (Gini) and ANN

4.0 Conclusion

The predictive models were compared based on sensitivity, specificity, accuracy rate and ROC Index. The best selected model for prediction of diabetic retinopathy was Logistic Regression using Forward selection method since it had the highest sensitivity, specificity and accuracy rate than other predictive models. The performance of Decision Tree using Gini and ANN model ranked the second and third respectively.

By using Logistic Regression (Forward selection) method, this study found that there were five significant variables that contribute to the presence of diabetic retinopathy which are duration of DM, Nephropathy, Neuropathy, Diabetic Foot and HbA_{1C} level. This study revealed that the presence of DR correlated with the length of time the patients had been diagnosed with DM. This means that the longer a patient suffers with DM, the more likely the patient are to have diabetic retinopathy. From the Logistic Regression equation, the odds of having diabetic retinopathy increased by 16.8% every year. This finding is consistent with the result of local study in which the patients with longer duration of DM had a higher risk of developing DR (Abougalambou & Abougalambou, 2015; Mallika *et al*., 2011; Mohd Ali *et al*., 2016; Naveen *et al.,* 2014). The duration of diabetes seemed to be the strongest factor associated with the occurrence of retinopathy, as supported by several studies (Cardoso, Leite, Dib, & Salles, 2017; Casanova *et al*., 2014; Karadeniz & Yilmaz, 2007; Kim *et al.,* 2018; Maria *et al*., 2003; Tsao *et al*., 2018).

This study also revealed that nephropathy was significantly associated with the presence of DR in which patients who suffered from nephropathy were twice more likely to have diabetic retinopathy compared to patients without nephropathy. This findings were confirmed by previous studies by (Diallo *et al*., 2018; Maria *et al*., 2003; Mohd Ali *et al*., 2016). In contrast with Abougalambou & Abougalambou (2015), no relationship was found between the development of retinopathy and the presence of nephropathy in their study. Zhang *et al*. (2018) also reported that nephropathy and diabetic retinopathy are the two major chronic microvascular complications of T2DM which increased in parallel with diabetes. This study provides important information about risk factors that associated with the presence of diabetic retinopathy among Type II diabetic patients. In addition, the knowledge from this study is hoped to create awareness to seek early treatment and undergo regular eye examinations for diabetic patients. Furthermore, it benefits not only the ophthalmology department but also other multidisciplinary areas that involves in the management of DR patients.

This is a retrospective study that relied on medical records. Thus, some of the patients' details were not available or recorded in the medical records. In addition, the majority of diabetic patients in this study were Malays and therefore, other ethnic groups were not equally covered. Thus, this study may not be representative to a larger community. Therefore, for future studies, it is suggested to have more samples and risk factors such as family history, lipid profile and fasting blood sugar so that it can provide a better overview of the disease.

Acknowledgement

The authors would like to thank Faculty of Medicine, UiTM Sungai Buloh for providing dataset in this study.

Reference

Abougalambou, S. S. I., & Abougalambou, A. S. (2015). Risk Factors Associated With Diabetic Retinopathy among Type 2 Diabetes Patients at Teaching Hospital in Malaysia. Diabetes and Metabolic Syndrome: Clinical Research and Reviews, 9(2), 98–103. https://doi.org/10.1016/j.dsx.2014.04.019.

American Academy of Ophthalmology. (2016). AAO Guidelines for Diabetic Retinopathy. https://doi.org/10.1016/S0140-6736(09)62124-3.

Ananthapadmanabhan, K. R., & Parthiban, G. (2014). Prediction of Chances-Diabetic Retinopathy Using Data Mining Classification Techniques. *Indian Journal of Science and Technology*, *7*(10), 1498–1503. https://doi.org/10.17485/IJST/2014/V7I10/51098

Balaji, S. A., & Baskaran, K. (2013). Design and Development of Artificial Neural Networking (ANN) System Using Sigmoid Activation Function to Predict Annual Rice Production in Tamilnadu. *International Journal of Computer Science, Engineering and Information Technology (IJCSEIT)*, *3*(1), 13–31. https://doi.org/10.5121/ijcseit.2013.3102.

Cardoso, C. R. L., Leite, N. C., Dib, E., & Salles, G. F. (2017). Predictors of Development and Progression of Retinopathy in Patients with Type 2 Diabetes : Importance of Blood Pressure Parameters. *Scientific Reports*, *7*(4867), 1–10. https://doi.org/10.1038/s41598-017-05159-6.

Casanova, R., Saldana, S., Chew, E. Y., Danis, R. P., Greven, C. M., & Ambrosius, W. T. (2014). Application of Random Forests Methods to Diabetic Retinopathy Classification Analyses. *PLoS ONE*, *9*(6), 1–8. https://doi.org/10.1371/journal.pone.0098587.

Clinical Practice Guidelines. (2011). Screening of Diabetic Retinopathy. *Ministry of Health Malaysia*. https://doi.org/10.1002/pdi.299.

Diallo, A. M., Novella, J. L., Lukas, C., Souchon, P. F., Dramé, M., François, M., … Delemer, B. (2018). Early predictors of diabetic retinopathy in type 1 diabetes: The Retinopathy Champagne Ardenne Diabète (ReCAD) study. *Journal of Diabetes and Its Complications*, *32*(8), 753–758. https://doi.org/10.1016/j.jdiacomp.2018.05.011.

Gadekallu, T.R.; Khare, N.; Bhattacharya, S.; Singh, S.; Maddikunta, P.K.R.; Ra, I.-H.; Alazab, M.(2020. Early Detection of Diabetic Retinopathy Using PCA-Firefly Based Deep Learning Model. Electronics. 2020; 9(2):274.

Haque, M., Rahman, A., Hagare, D., & Chowdhury, R. K. (2018). A Comparative Assessment of Variable Selection Methods in Urban Water Demand Forecasting. *Water*, *10*(419), 1–15. https://doi.org/10.3390/w10040419.

Harb, W., Harb, G., Chamoun, N., Kanbar, A., Harb, M., & Chanbour, W. (2018). Severity of diabetic retinopathy at the first ophthalmological examination in the Lebanese population. *Therapeutic Advances in Ophthalmology*, *10*, 2515841418791950. https://doi.org/10.1177/2515841418791950.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied Logistic Regression* (Second Edi). Canada: John Wiley & Sons, Inc. Retrieved from http://resource.heartonline.cn/20150528/1_3kOQSTg.pdf.

Karadeniz, Z. S., & Yilmaz, M. T. (2007). Duration of diabetes and prevalence of diabetic retinopathy : Istanbul Diabetic Retinopathy Study – IDRS results 1. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, *1*, 43–48. https://doi.org/10.1016/j.dsx.2006.11.008

Kim, K., Kim, E. S., & Yu, S. (2018). Longitudinal relationship between retinal diabetic neurodegeneration and progression of diabetic retinopathy in patients with type 2 diabetes. *American Journal of Ophthalmology*. https://doi.org/10.1016/j.ajo.2018.08.053.

Lee, J. (2013). Macrovascular and Microvascular Complications of Diabetes Mellitus. *Drug Topics*, 44–53. https://doi.org/10.5005/jp/books/10202.

Lek, S., & Gue, J. F. (1999). Artificial neural networks as a tool in ecological modelling , an introduction. *Ecological Modelling*, *120*(2–3), 65–73. https://doi.org/10.1016/S0304-3800(99)00092-7.

Letchuman, G. R., Nazaimoon, W. M. W., Mohamad, W. B. W., Chandran, L. R., Tee, G. H., Jamaiyah, H., … Faudzi, Y. A. (2010). Prevalence of Diabetes in the Malaysian National Health Morbidity Survey III 2006 The Malaysian National Health Morbidity Survey III (NHMS. *Medical Journal of Malaysia*, *65*(3), 173–179. Retrieved from http://www.e-mjm.org/2010/v65n3/Diabetes.pdf.

Mahendran, G., & Dhanasekaran, R. (2015). Investigation of the severity level of diabetic retinopathy using supervised classifier algorithms. *Computers and Electrical Engineering*, *45*, 312–323. https://doi.org/10.1016/j.compeleceng.2015.01.013.

Mallika, P., Lee, P., Cheah, W., Wong, J., Syed Alwi, S., Nor Hayati, H., & Tan, A. (2011). Risk Factors For Diabetic Retinopathy In Diabetics Screened Using Fundus Photography At A Primary Health Care Setting In East Malaysia. *Malaysian Family Physician : The Official Journal of the Academy of Family Physicians of Malaysia*, *6*(2–3), 60–65. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/25606225%5Cnhttp://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=PMC4170425.

Maria, Z., Corrêa, S., & Freitas, A. M. (2003). Risk factors related to the severity of diabetic retinopathy. *Arq Bras Oftalmol*, *66*(5), 739–743. https://doi.org/10.1590/S0004-27492003000700002.

Masconi, K. L., Matsha, T. E., Echouffo-tcheugui, J. B., Erasmus, R. T., & Kengne, A. P. (2015). Reporting and handling of missing data in predictive research for prevalent undiagnosed type 2 diabetes mellitus : a systematic review. *The EPMA Journal*, *6*(7), 1–11. https://doi.org/10.1186/s13167-015-0028-0.

Matignon, R. (2007). *Data Mining Using SAS® Enterprise Miner TM*. South San Francisco, CA: A John Wiley & Sons, Inc. Retrieved from https://doc.lagout.org/Others/Data Mining/Data Mining using SAS Enterprise Miner %5BMatignon 2007-08-03%5D.pdf.

Mohd Ali, M. H., Draman, N., Mohamed, W. M. I. W., 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

Naveen Kumar, T., Nagi Reddy, T., & Radha Kishan, T. (2014). Diabetic eye screening in multi ethnic population of Malaysia : epidemiological risk factors for development of diabetic retinopathy. *International Journal of Research in Medical Sciences*, *2*(3), 1045–1049. https://doi.org/10.5455/2320-6012.ijrms201408069.

Park, H.-A. (2013). An Introduction to Logistic Regression : From Basic Concepts to Interpretation with Particular Attention to Nursing Domain. *Korean Society of Nursing Science*, *43*(2), 154–164. http://dx.doi.org/10.4040/jkan.2013.43.2.154.

Reddy, G. T., Bhattacharya, S., Ramakrishnan, S. S., Chowdhary, C., Hakak, S., Kaluri, R., *et al*. (2020). An ensemble based machine learning model for diabetic retinopathy classification. International Conference on Emerging Trends on Information Technology and Engineering,2020, 1–6.

SAS Institute Inc. (2010). *Applied Analytics Using SAS® Enterprise Miner^{TM}*. Retrieved from https://is.muni.cz/el/1431/podzim2012/M9DM2/um/LWAAEM61_002_2010.pdf.

Senthilvel, V., & Radhakrishnan, R Sathiyamoorthi, R. (2012). Prediction of Diabetic Rretinopathy Among Diabetics using Binary Logistic Regression Approach. *Indian Journal of Medical Specialities*, *3*(1), 18–20. https://doi.org/10.7713/ijms.2012.0005.

Shriwas, S. R., Isa, A. B., Reddy, S. C., & Mohammad, M. (1996). Risk Factors for Retinopathy in Diabetes Mellitus in Kelantan Malaysia. *Med J Malaysia*, *51*(4), 447–452. Retrieved from https://e-mjm.org/1996/v51n4/Diabetes_Mellitus.pdf.

Tsao, H., Chan, P., & Su, E. C. (2018). Predicting diabetic retinopathy and identifying interpretable biomedical features using machine learning algorithms. *BMC Bioinformatics*, *19*(283), 195–205. https://doi.org/https://doi.org/10.1186/s12859-018-2277-0.

Witten, I. H., & Frank, E. (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. (J. Gray, Ed.) (Second Edi). Morgan Kaufmann Publishers. Retrieved from ftp://ftp.ingv.it/pub/manuela.sbarra/Data Mining Practical Machine Learning Tools and Techniques - WEKA.pdf.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., … Dan, J. H. (2008). Top 10 algorithms in data mining. *Knowledge and Information Systems*, *14*(1), 1–37. https://doi.org/10.1007/s10115-007-0114-2.

Zhang, J., Wang, Y., Li, L., Zhang, R., Guo, R., & Li, H. (2018). Diabetic retinopathy may predict the renal outcomes of patients with diabetic nephropathy, *6049*. https://doi.org/10.1080/0886022X.2018.1456453.