International Journal of Engineering & Technology, 7 (4.31) (2018) 521-527

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

A Study of Diabetic Retinopathy Classification Using Support Vector Machine

Nur Izzati Ab Kader[1], Umi Kalsom Yusof [2*], Syibrah Naim[3]

[1]A School of Computer Sciences, Universiti Sains Malaysia,11800, Georgetown, Pulau Pinang, Malaysia
**Corresponding author E-mail: umiyusof@usm.my*

Abstract

Diabetic Retinopathy (DR) is a diabetic complication which can cause blindness. As DR cases keep increasing, ophthalmologists are forced to diagnose a large number of retinal images daily. Generally, the diabetic eye screening is done manually using qualitative scale to detect abnormalities on the retina. Although this approach is useful, the detection is not accurate; and create a need for a tool that can help the experts to classify the severity of DR to establish adequate therapy. Previous researchers have studied machine learning to propose an automatic DR classification. However, it needs to be improvised especially in terms of accuracy. Hence, this paper aimed to find classifier with optimal performance in the study of DR classification. This study considered three classes of diabetic patients which were patients who do not have DR (NODR), patients with non-proliferative DR (NPDR) and patients with proliferative DR (PDR), instead of focusing only on two classes (NO DR, DR). Support Vector Machine was used in this research due to the success of many classification problems that had been proposed which produced good result. The results obtained showed that SVM gave the best accuracy, 76.62% with average sensitivity of 0.8081 and average specificity of 0.8376 respectively.

Keywords: *Diabetic Retinopathy; Machine Learning; Classification; Support Vector Machine; Kernel.*

1. Introduction

Diabetic Retinopathy (DR) is a part of complication of Diabetes Mellitus (DM) and it affects 1 in 3 persons with DM. DR is caused by damage to the blood vessels of the retina and the light sensitive tissue at the back of the eye. The number of DR prevalence is increasing year on year. According to WHO Global report, the number of adults living with diabetes has almost quadrupled since 108 million in 1980 to 422 million adults in 2016. This dramatic rise is largely due to the rise in type 2 diabetes and factors driving it include overweight and obesity [1]. With the increasing number of cases nowadays, abnormal retinal classification become a challenging task for ophthalmologist as they need to deal with a large number of retinal images to be diagnose every day. Screening and early detection of DR are playing an important role to help reduce the incidence of visual morbidity and vision loss. The screening tasks are done manually in most country [2]. Issue of variability in grading arise from this manual grading as the boundaries between the grades may differ between observers and also prone to error [3].

The process is carried out through naked eyes inspection. This inspection is carried out using an ophthalmoscope to directly inspect the fundus of the eye. The pupil will be dilated before it is examined. Usually, the experts identify relative characteristics such as to differentiate between normal and abnormal retina based on their experience. The retinal is mostly evaluated using qualitative scale such as mild, moderate, severe and extreme. Occasionally, it is useful but not very effective. Issue of variability in grading arise from this manual grading as the boundaries between the grades may differ between observers and also prone to error [3]. The prevalence of the disease is drawing an attention for all the parties to play their parts towards the prevention and treatment of the disease. Collaboration between experts from different areas can be achieved with the sophisticated technology nowadays. Currently, the application of computational technique has made a huge impact in health sector. Computational technique such as supervised machine learning is popularly used to predict the presence and absence of the disease. These methods play vital roles in improving the way for detection, diagnosis and treatment of the disease.

Among the solution that have been proposed by previous researchers is to come out with DR classification that can help ophthalmologist for grading process. There are various methods have been applied for DR classification. Some of them are classifying using retinal imaging which is a classification technique performed based on the abnormalities found on retinal fundus image such as exudates, micro aneurysm, hemorrhages and also blood vessels. Although the retinal imaging technique facilitate early detection of DR, they required additional equipment which is quite cost-prohibitive or sometimes unavailable especially in rural areas. On the other hand, several DR classifiers have been developed using clinical variables as an alternative to retinal imaging. However, there is still some space for improvement especially in the accuracy of the classifiers.

Therefore, this study is proposed to classify DR with the objective to find DR classifier with optimal or near-optimal performance matrices using Support Vector Machine. Support Vector Machine is chosen to be adapted in this study as it can helps to improves sensitivity and/or specificity of disease detection and diagnosis. This classifier is built based on the clinical variables data. There are several advantages of this study. First and foremost, the clinical variables used in this study are selected by doctors, thus the validity of the features used are unquestionable. Equally important, this dataset encompasses of three classes of diabetic patients which are patients that do not have DR (NODR), patients

with non-proliferative DR (NPDR) and patients with proliferative DR (PDR). Usually, DR classification focus only on two classes which are to classify whether a person being diagnosed with DR or not. This classification can assist the doctors to perform an optimum decision-making regarding the type and medication to be prescribed. In addition, unnecessary testing and check ups can be prevented (Figure 1).

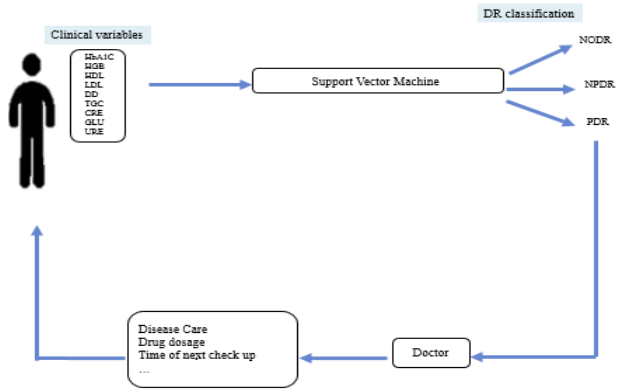

Fig. 1: The classification of diabetic retinopathy among diabetic patients

2. Related works

In the area of DR classification, many researchers have studied DR with different intelligent methods and aims. Most of the existing DR classification or detection have mainly focused on the computational analysis of the eye fundus using image processing classifiers. Currently, research in image processing are studying on how to extract signs of DR from the fundus image [4]. Usually, computer vision techniques is used to build models for the detection of the signs. These classifiers facilitate early detection of DR, thus retinal image is required. Therefore, they are unable to address the evident barrier of patients' access to the specialist even though they might ease their burden to assess the image. Besides, there are also studies performed to build clinical decision support system (CDSS) that matches with lenses or an ophthalmoscope that can be used on smartphone. A smartphone-based classifier integrated with microscopic lenses was proposed by [5] to capture retinal images. A neural network model has been used in their study to analyze images and provide the results. In the other study, a portable smartphone-based using image analysis and machine learning was proposed. This portable smartphone can be used for initial screening by attaching an ophthalmoscope to capture fundus image. The classifier that was install in the smartphone will play role to process the captured image. Despite all the sophisticated features and benefits offered by the classifiers presented in these studies, they are cost-prohibitive as additional equipment for retinal imaging is required.

Previous researchers have studied the association between DR and clinical variables of the patient. Although the potential of developing classification model using clinical variables has been proved, not much attention given to this approach. A few studies have been conducted to develop DR classification using clinical variables with adopting supervised machine learning. In supervised machine learning, the model has to learn a function named target function, which is an expression of a model describing the data. A few studies have been conducted using supervised machine learning techniques. Previously, [6] have developed a clinical decision support system (CDSS) for DR using logistic regression, random forest, decision tree and ensemble models. The CDSS was built from demographic and lab data in order to detect patient's susceptibility to retinopathy. Their work reached an accuracy of 92.76%.

In another work, a study had been conducted to explore the use of two kinds of ensemble classifiers to determine whether a patient is in risk of developing DR: fuzzy random forest and dominance-based rough set balanced rule ensemble. This study employed the clinical attributes which represent main risk factors to perform the prediction. The performance obtained in the study are over 80% for specificity and sensitivity. Besides, [7] built DR classifier to predict the risk of DR using Naive Bayes and Decision Tree. From the study, it was found that Decision Tree method has 90% of accuracy. [8] have conducted a study on prediction and diagnosis of DR using Naive Bayes. The study employed clinical variables to perform the prediction. From the result, Naive Bayes able to obtained 89.11 % of accuracy. [9] built a DR predictive system using 140 diabetic patients' data to predict prevalence in Malaysia. They adopted a voting mechanism to select the final results of Decision Tree and Case Base Reasoning. [10] develop DR classifier to predict the risk of DR using data from 55 type 1 diabetes patients. They applied Classification and Regression Tree (CART), Neural network, classification-based Rule Induction with C5.0, Hybrid Wavelet Neural Network (HWNN) and merged their result using voting mechanism.

While the machine learning for DR classification has been adopted in some form, it is limited in several ways. First, the current accuracy of DR is still low and need to be improvised. To the best of our knowledge, the work by [6] yield the highest accuracy, but with an overall accuracy 92.76%, it leaves room for improvement for DR classification. Besides, a few studies [9],[10] considered small number of instances in the dataset which are not enough to build a good classifier. In addition, very little work have been done with regards to DR classification focusing on three classes of DR (NO DR, NPDR, PDR). Most of the studies done previously focused only on two of the classes (NO DR, DR). Therefore, in the present research, the effort is to develop Support Vector Machine classifier that address the limitations of the extant literature, with the motivation to improve their results.

3. Data and Method

This section explains how the machine learning techniques mentioned in the previous section was used for DR classification. The dataset used in this study was provided by Eye Clinic of the Sakarya University Educational and Research Hospital, a hospital located in the city of Adapazari, the capital of the Turkish province of Sakarya.

3.1. The Data from The Electronic Health Record

The dataset contained the information of 385 diabetic patients, who were already labelled according to the DR reference: 79 patients were not suffering from DR, 161 patients presented NPDR and 145 patients presented PDR. Therefore, there were two types of attributes: numerical (Glycated Hemoglobin (HbA1C), Hemoglobin (HGB), High-Density Lipoprotein (HDL), Low Density Lipoprotein (LDL), Diabetes Duration, Triglyceride, Creatine, Glucose and URE) and categorical (NODR, NPDR, PDR). Details of the dataset are shown in Table 1, which had previously been used in [8], that investigated the DR prediction using Naive Bayes. It is always said that in any comprehensive machine learning study, data understanding is required - prior to the development of the classifier. Most of the researchers in novel machine learning field pay very serious attention to data as the quality of the data affects the machine learning results. Hence, in this study, data understanding was performed prior to model development to have a general insight about the data quality, representativeness, in formativeness, and presence or absence of outliers and missing values. At the end of the data understanding phase, the data was found to be in a good quality and no outliers and missing values were found. The dataset was then utilised in the development of the models.

3.2. Modeling Procedure

To develop the classification model, Support Vector Machine (SVM) is applied in this study. Building of this model involved several stages. In this study, the modeling procedure encompasses of five stages, as shown in Figure 2. The target variable, NODR denoted no diabetic retinopathy diagnosis and NPDR denoted non-proliferative diabetic retinopathy and PDR denoted proliferative diabetic retinopathy. Hence, the aim of performing classification is to predict categorical class label for unknown data based on the classification model built by training data. It played its role of mapping an input attributes set x into its class label y.

Table 1: Description of features in dataset

Features	Description
HbA1c	Shows average level of sugar over the past 2 to 3 months
Haemoglobin	Substance inside red blood cells that transports oxygen to the cells of the body
HDL	Carries LDL cholesterol away from arteries and back to the liver
LDL	A bad cholesterol that contributes to fatty build up in arteries
Diabetes Duration	Length of time they have diabetes
Triglyceride	Fat in blood that body uses for energy
Glucose	Indicates concentration of blood sugar at a single point in time
URE	Indicates blood urea concentration
Creatine	Facilitate recycling of energy

Classifiers should also be able to generalise previously unseen data. If not, it would result in the poor generalisation that can be characterised by over-training. The over-training model just memorised the training examples and was unable to give correct prediction output for the sample that were not in the training set (test data). These two crucial demands (good prediction on test data and good generalization) are conflicting and known as the Bias and Variance dilemma. One of the techniques to balance between minimal Bias and minimal Variance of the model is cross-validation (CV). CV technique helps to solve improper data splitting with the sophisticated sampling method. Improper split of the dataset can lead especially to an excessively high Variance of the model performance. The basic idea of CV is based on data splitting, part of the data is used for fitting each competing model and other than that is used to measure the predictive performance. The main goal was to achieve a stable and confident estimate of model performance. There are two type of CV that are commonly used which are hold-out CV and k-fold CV. Hold-out method is popular for its efficiency and easiness while k-fold gain advantage on its ability s to gain a stable estimate of the model error using s a combination of more tests. It is useful if not enough data for the hold-out cross-validation is available. In the present study, 10-fold cross-validations were applied.

3.3. Support Vector Machine

Support Vector Machine (SVM) is a classifier introduced by Corinna Cortes and Vladimir Vapik. It is a classifier with a learning routine used for classification of input data received by a computing system and also for regression task. It is categorized as supervised machine learning method with objective to classify data points by maximizing the margin between classes in a high-dimensional space. SVM work by generates a hyperplane to discriminate between each class after the input data have been transformed into high-dimensional space. The specialty of SVM is that can efficiently perform non-linear classification using kernel trick.

Fig. 2: Process flow diagram for development of DR classification model. The diagram represents the general modelling procedure of SVM

The kernel-trick function is to allows constructing the classifier in non-linear data by increasing the dimension of SVM [11]. In the earlier version, SVM was originally designed for binary classification. However, it has been effectively extend for multiclass classification problem by a few researchers by introducing method of one against one, one against all and directed acyclic graph SVM [12], [13],[14]. The basis of their method are either to decompose the multi class problem into several binary classification problem and build a standard SVM for each or directly considering all data at once.

One-against-one: In this study, SVM deal with three classes which are NODR, NPDR and PDR. The method of one against one (or also known as pairwise coupling or round robin) is used to ensure that SVM can handle this multiclass classification. In the earlier, this method was introduced by [15] and the first used of this method on SVM was by [16]. In this study, the SVM is build for each one pair of classes to distinguish the samples of one class from the samples of other classes. This method works by construct $k(k\text{-}1)/2$ classifiers where each one is trained using the data from two classes. The value of k the number denoted the number of classes.

Thus, in this case, given l training data $(x1,y1)$, where $x1$, i=1,...,l and y=1,...,k with k is the class of x. The training data $x1$ are mapped to n- th dimensional space by the function ϕ and C is the penalty parameter. One-against-one method solve the following classification problem:

$$\min_{w^{ij},b^{ij},\xi^{ij}} \frac{1}{2}\left((w^{ij})^T\right)(w^{ij})+C\,\Sigma_t\xi_t^{ij}(w^{ij})^T,$$

$$(w^{ij})^T\phi(x_t)+b^{ij}\geq 1-\xi_t^{ij},\ y_t=i,\ (w^{ij})^T\phi(x_t)+b^{ij}\leq -1+\xi_t^{ij},\ y_t=j,$$

$$\xi_t^{ij}\geq 0. \qquad (1)$$

Based on the Equation 1, by minimizing 1/2 $((w^{ij})^T)(w^{ij})$, the algorithm would be able to maximize the margin between two classes, $2/\|wi\|$. The penalty is defined by C $\sum t\ \xi^{ij}(w^{ij})^T$ and it would reduce the number of training errors. It would be better for the algorithms performance if it is able to find a balance between 1/2 $((w^{ij})^T)(w^{ij})$, and the training errors. After $k(k\text{-}1)/2$ have been constructed, the future testing have to be done. Voting strategy is used in this study. Based the voting strategy, if sign determine that x is in the i class, then the vote for i is added one. Otherwise, the j is increased by 1. The label of x is based on the largest vote.

Parameter initialization: SVM has two significant parameters which are, cost and gamma. The function of cost is to control on how the decision boundary is crafted around while the gamma controls on how many vectors and curves are allowed for the im-

plementation of SVM. The best value for cost and gamma depend on the data, it may vary between low values such as 0.000001 and quite high values such as 1000 or higher. In this study, the value of cost is assigned to 1, and gamma is 0.11. The kernel used for this study is radial basis function (RBF) kernel.

Finding the boundary: After the parameters have been initialized, the next step is to find the boundary/hyperplane. The boundary can be found by connecting every point in one class to each other. From the connection, the outline will be emerges, and it defines the boundary of the class. Usually the outline emerge can be more than one. The classes that are linearly separable will not intersect each other but in non-linearly separable data, the boundaries are intersect to each other. Example of hyperplane is shown in Figure 3. The hyperplane is usually calculated based on the following equation:

$$H = \text{bias} + \text{weight} * x = 0 \quad (2)$$

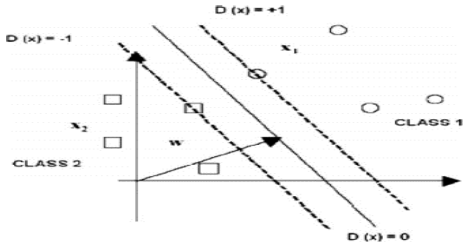

Fig. 3: Example of hyperplane in SVM. The hyperplane is used to discriminate between the two classes of data [17].

Kernel Function: Among the specialty that own by SVM is the kernel function K*xn,Xi* which can help to separate the data easily especially in non linear separable data. The data used in this study is not linearly separable. The kernel transforms the non-linear separable data to linear generally by taking the square root of *x*. In this case which using RBF kernel, the function can be represented in terms of this following equation:

$$K(x1, x\,j) = C^{-\gamma|xi - xj|2} \quad (3)$$

Termination Condition: The stopping criterion in this study is de- rived from the method used by Crammer and Singer, which following the equation of:

$$\max\left(\max_{a^i<C, y_i=1} - \nabla f(\alpha)i \max_{a^i<C, y_i=-1} \nabla f(\alpha)i\right),$$
$$\leq \min\left(\min_{a^i<C, y_i=-1} \nabla f(\alpha)i \min_{a^i<C, y_i=1} - \nabla f(\alpha)i\right) + 10^{-3} \quad (4)$$

The process of SVM will be terminated once all the test data has been successfully classified.

3.4. Performance Evaluation

Performance evaluation is beneficial for comparing the quality of classification across systems. The accuracy of the classifier was indicated by the percentage of the test dataset that was correctly classified by the classifier. It was calculated using the value of true positive (TP), true negative (TN), false positive (FP) and false negative (FN).
Besides, confusion matrix was also used to measure the general performance of classifier using confusion matrix. It determines the ability of the classifier to produce accurate diagnosis for DR. Confusion matrix is a table that consist of performance of classification model which are true values are known. It contains information regarding actual and predicted classification done by a classification system. It is evaluated by calculating the correctly predicted TP and TN classifications based on the formulas as in Equation 5, 6 and 7. Besides, sensitivity and specificity also can be measured from the confusion matrix in order to get more specific information on performance of classifier. Sensitivity measures selected relevant instances while specificity measures exactness of classifier.

The words sensitivity and specificity have their origins in screening tests for diseases. Sensitivity refers to the probability for the test to determine a person has the disease when in fact they really do have the disease. In other words, it measure how likely it is for a classifier to pick the presence of a disease in a person who has it.[18] suggest that the index, sensitivity is the first priority to be considered. This is because in the reality of medical-care case, if a patient is true positive but he/she is no further cured, he/she will suffer an irreparable damage or in DR permanent blindness.
On the other hand, specificity refers the probability for the classifier to determine that a person does not have the disease when they are in fact disease free. It was also an important measure to be considered. An ideal classifier should have high sensitivity and high specificity value. Thus, in this paper we are looking for high sensitivity and specificity classifier with an emphasis on sensitivity value as one of the motivations for this research was to minimise cases of visual loss.

$$AC = (TP + TN)/TP + FP + TN + FN \quad (5)$$

$$PN = TP/(TP+FP) \quad (6)$$

$$RC = TP/TP+FN \quad (7)$$

The confusion matrix can be understand as follows:

- Accuracy (AC): overall performance of classifier
- True Positive (TP): correct positive prediction
- True Negative (TN): correct negative prediction
- False Positive (FP): incorrect positive prediction
- False Negative (FN): incorrect negative prediction

Besides, in order to determine which classifier has the best performance, it is good to evaluate the classifiers with additional evaluation metrics. In this paper F-measure, kappa statistic and confidence interval (CI) are used as additional evaluation metrics. F-measure is a harmonic mean of precision (positive predictive value) and recall (exactness of classifier). According to Van Rijsbergen (1979), F-measure is defined as a combination of recall (R) and precision (P) with an equal weight in the following formula [19]:

$$F = \frac{2PR}{P+R} \quad (8)$$

Precision can be defined as the probability that a randomly chosen predicted instance (positive) will be relevant while Recall is how close we are to a specific target on average. Kappa statistic, which introduced by Jacob Cohen in 1960, was used in this paper in order to test interrater reliability (consistency of measurement obtained). There are various methods to measure interrater reliability. It is used to account the possibility that raters actually guess on at least some variables due to uncertainty with the range from -1 to +1, similar to correlation coefficient. It is calculated based on formula:

$$K = \frac{Pr(a)-Pr(e)}{1-Pr(e)} \quad (9)$$

In the equation, Pr(a) represent the actual observed agreement and Pr(e) represents chance agreement. The interpretation for kappa statistic is based on it standard range [20]:

- less than 0.20: Poor
- 0.21-0.40: Fair
- 0.41-0.60: Moderate
- 0.61-0.80: Good
- 0.81-100: Very Good

Once the kappa has been calculated, CI is calculated to evaluate the meaning of the obtained kappa. In this paper 95 % CI is used as it is the most frequent value desired. It would indicate the range that 95% of those true means value would fall into. CI is represented by subtracting from kappa from the value of the desired CI level times the standard error of kappa *SEk*. The formula used 1.96 as the constant by which the standard error of kappa is multiplied. Thus, the formula for a CI is calculated as:

$$(k - 1.96 \times SE_k) \text{to} (k + 1.96 \times SE_k) \quad (10)$$

4. Computational experiment and result analysis

To evaluate the performance of the proposed classifiers, computational experiments were conducted using benchmark dataset. The proposed SVM was implemented using R software version 2016 run on an Intel Core i3-3110M CPU 2.4 GHz, on a 64-bit windows 8 operating system.

In the beginning, the dataset is divided into training and testing data; the proportion of elements in each part was 70 and 30 % respectively. A standard procedure, k-fold cross validation is used for evaluating the performance of classifier. The instances in the dataset *D* is divided into k subsets (the folds: *D*1, *D*2,*D*3...*Dk*) of the same size approximately. However, a concern has arisen from this technique is on how to choose k. In most applications k = 10 is chosen. Thus, in this study, k=10 was used. There are two subprocesses in CV which are training set and testing set. One part of k-fold forms the testing set *Tv*, while the remaining k-fold form the training set *Tt* . The classifier is built on the training set while the evaluation of performance is made in the testing set. During SVM training, the RBF kernel plays a very significant role as it help SVM to classify the non-linear separable data in this study into linear separable form. Figure 4 illustrates on how the kernel trick function.

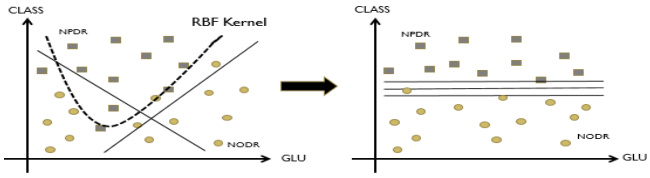

Fig. 4: The kernel trick enables SVM to transform the data into linear form

Model building and evaluation was executed for 10 iterations and the result taken is from the average of the 10 sets result in order to get the best results. The result was measured using the performance metrics. The objective is on the maximization of the classifiers performance in each matrix. The main performance metric is the confusion matrix that represent the accuracy, sensitivity and the specificity while others are considered as additional matrices that are used to examine the performance in detail. The metrics that were used to measure the performance are based on the Equation 5 until Equation 10 stated in Section 3. The overall process flow for the computational experiment is shown in the Figure 5.

Firstly, the performance measure observed is accuracy of the classifier. The accuracy of the classifier is indicated by the percentage of the test dataset that are correctly classified by the classifier. In SVM, two parameters that being the key for performance are the value of cost and gamma.

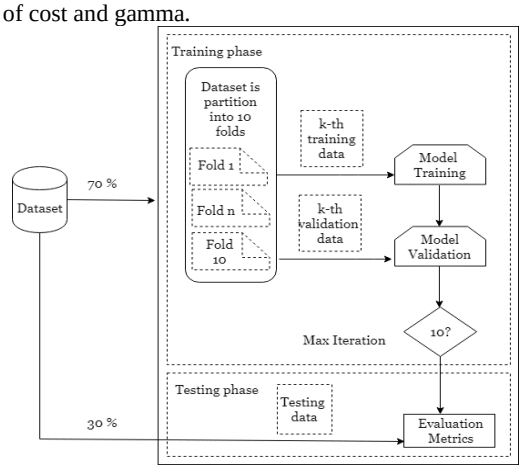

Fig. 5: Process flow diagram for computational experiment. This diagram shows detail process of the modelling SVM

The accuracy of SVM is the high which is 76.62%. Besides, in order to get a good understanding on the performance of algorithms, sensitivity and specificity also investigated for each algorithm. High sensitivity values mean that the algorithm has low False Negative (incorrect negative prediction). High specificity indicates that the samples for each class are classified with high precision.

F-measure is then calculated. F-measure intended to combine the precision and also recall into a single measure. It is a good performance evaluation other than accuracy, as it solves the bias issue that can affects both precision and recall. In the F-measure, the objective of the experiment is to find the algorithm with the highest value, or in other words, we want it to be 1. In contrast, the measure is 0 when either the precision or the sensitivity is 0. That bring a meaning that, low precision or the sensitivity will affect the F-measure. Based on the table, SVM also shows high F-measure, 0.8758 with high precision and recall values. Thus, it can be concluded that SVM conveys the balance between the precision and recall.

Besides, kappa statistics recently has been applied in machine learning to rate algorithms. Performance evaluation for kappa statistic is based on its standard range. Generally, the evaluation is understood as (higher = better algorithm. The algorithm is categorized using their value that is calculated using the formula (refer Equation 10).

Finally, the last performance metric that is used to evaluate the algorithms is confidence interval. Unlike the other measures that use only a single measure, the confidence interval provides the result using the range value. Basically, the result for the confidence interval will reflect the accuracy value. Therefore, the result from SVM demonstrated a highly confidence level with a range between 72.07% to 90.76%, as it also gains the highest accuracy value which is 76.62%. This means that a range of possible accuracy for SVM is from 72.07 to 80.76 with confidence of 95%. The overall result of SVM can be found in Table 2.

Table 2: Result for SVM classification based on each class of DR

Class	Accuracy	Sensitivity	Specificity	Precision	Recall	F-Measure	Kappa	CI
NODR		0.8481	0.9771	0.9054	0.8481	0.8758		
NPDR	76.62	0.7950	0.7545	0.6995	0.7950	0.7442	63.28	72.07,80.76
PDR		0.7813	0.7812	0.6897	0.7813	0.7327		

Table 3: Result for classification based on each class of DR

Class	Techniques	Accuracy (%)	Sensitivity	Specificity	Precision	Recall	F-Measure
Class NODR	**SVM**	**76.62**	0.8481	0.9771	0.9054	**0.8481**	**0.8758**
	ANN	72.47	**0.9241**	0.9608	**0.9825**	0.8241	0.8235
	KNN	67.01	0.7215	**0.9850**	0.6706	0.7215	0.6951
Class NPDR	**SVM**	**76.62**	**0.7950**	0.7545	**0.6995**	**0.7950**	**0.7442**

	ANN	72.47	0.6521	**0.7946**	0.5585	0.6521	0.6017
	KNN	67.01	0.6335	0.7857	0.6800	0.6335	0.6559
Class PDR	**SVM**	**76.62**	**0.7813**	0.7812	**0.6897**	**0.7813**	**0.7327**
	ANN	72.47	0.5793	**0.8000**	0.6000	0.5793	0.5895
	KNN	67.01	0.6828	0.7875	0.6600	0.6828	0.6719

4.1. Comparison of Algorithms

There are many classifiers in the area of machine learning that have the potential to solve the problem of DR classification. However, there were two classifiers that had been decided as comparison classifiers for this study. The comparison classifiers were selected based on a few criteria. First, the classifier must be a supervised machine learning as the data used for this study was labelled data. Next, the classifier must have the ability to handle numerical and categorical variables as the variables in the dataset were numerical and categorical variables.

Besides, the classifier is known to have solved many classification problems with good result. Based on the highlighted criteria, KNN and ANN satisfied all the criteria and hence chosen to be used in this study as the comparison classifiers. Based on previous literatures, these two classifiers have shown high capability in solving classification problems. The parameter of KNN was set as k=5 and for ANN, the hidden nodes = 70, maximum number of iterations = 100. Performance results of SVM were compared to the performance result of KNN and ANN. They were compared based on the same metrics as mentioned in Section 3. Based on the result, it was clear that the proposed SVM outperformed KNN and ANN. When compared to ANN, the accuracy of SVM and ANN were not much different as they were differed only by 4.15 %. However, SVM showed much greater performance when compared to KNN with 9.61 % differences in accuracy. These findings showed that SVM was more capable than ANN and KNN to handling multiclass DR Classification.

For the class of NODR, the performance of SVM was high which was more than 0.8481 % of each performance metrics. However, it was lower than the sensitivity of ANN. The lower performance of SVM was due to the fact that SVM was wrongly classified 6 data to NPDR, which means it had wrongly recognised the clinical variables of a person that did not have DR as a person with NPDR, while ANN only misclassified three persons to NPDR. Compared to KNN, SVM showed a better performance compared to KNN, as it only had metrics reading starting from 0.6706 to 0.9850 %. The largest difference was on the precision of the algorithms where SVM able to show 0.9825 while KNN only had 0.6706 of sensitivity. However, the specificity value of SVM was lower than KNN, which means SVM had lower capability to recognized patients that do not have DR compared to KNN as the purpose of the sensitivity is to test how good the algorithm is to recognise the positives.

In the class of NPDR, SVM also showed a better performance com- pared to ANN and KNN. However, the performance is started to lower due to the fact that more data were not successfully classified.

From the Table 3, the F-measure showed a balance between precision and recall of this class which was lower than in NODR, with difference of 0.1316 %. However, SVM was better than ANN and KNN which have 0.6017 and 0.6559 of F-measure respectively.

In the classes of PDR, most of the data that were not correctly classified are fell into the category of NPDR. It brings a meaning that the SVM as well as KNN and ANN had a tendency to classify a person with PDR as having NPDR. There was only a case which involved misclassification of PDR persons into NODR. The reason that the algorithms had more tendency to misclassify into NPDR was because the characteristic of clinical variables for PDR person is quite similar to the NPDR's. In references to the stages of DR severity, PDR patients are ones from the NPDR stage. Thus, the algorithms should have a good ability to recognize if a patient is still in NPDR stage or starting to move into the PDR stage based on the clinical variables.

The CI and kappa statistic showed in the Table 4 were used to reverify again the performance mention in the previous paragraph. From the table, it shows that the pattern of SVM performance was the same as mentioned before. It had the highest performance compared to ANN and KNN.

Table 4: Confidence Interval and Kappa Statistic

Algorithm/Result	SVM	ANN	KNN
Lower Limit	**72.07**	67.71	63.67
Upper Limit	**80.76**	76.87	73.18
Accuracy	**76.62**	72.47	67.01
Kappa Statistic	**67.55**	57.30	51.38

5. Analysis and Discussion

SVM is known to have a good performance in the previous research. It has demonstrated high ability in solving classification problems in many fields such as in biomedical and bioinformatics. So, it is not uncommon when it also showed a high performance in this study. The factors contributed to a high performance of SVM were analysed. It was found that its performance was, indeed, depended on the choice of parameters, cost and gamma. A good combination of cost and gamma values helps SVM to get an optimal separating hyperplane in the feature space. The hyperplane helps to discriminate between two classes after the input data have been transformed into high-dimensional space. However, in this case, it worked with multi- class scenarios to discriminate between the three classes.

The performances of SVM for each class were also analysed and they were compared to other classes of DR in KNN and ANN. While comparing the performances of SVM based on each class, it was obvious that SVM also had a better performance in each of the classes. For the selected algorithms that were used for comparison classifiers in this study which were KNN and ANN, a few reasons could explain the inability of both algorithms to outperform SVM. First, the parameters used during the development of the algorithms might still not be the optimal parameters. With parameter (k=5) for KNN, it could not effectively classify the data to the correct group of neighbours.

While for ANN, with a small number of hidden nodes (h=70) and (Imax=100), it did not help the ANN to learn the data. Therefore, they were unable to classify the data at the highest capacity.

In this study, the result performance that demanded a high attention was the sensitivity value of an algorithm. It has already been mentioned by [18] that suggested the index, sensitivity has to be the first to be considered. This is because in the reality of medical care cases, if a patient is true positive but he/she has not cured further, blindness. Based on the analysis of the performance, SVM had a high sensitivity value for each class. In the other algorithms that were compared to SVM, the sensitivity values were quite low, such as in KNN the lowest value was 0.6335 % and it has to be improved.

6. Conclusion

In this paper, a study on the most well-known classifier, SVM has been done in detail. The aim was to find classifier with an optimal or near-optimal accuracy value in DR classification. From the study, re- sults obtained show that SVM gives the best performance compared to KNN and ANN with 76.62 % of accuracy. The implementation of the proposed DR classification with a good performance will be able to serve as an aid in assisting experts in

the diagnosis of DR. It can help experts in improving decision making and become a standard guideline for the diagnosis.
In addition, it is highly important to classify and categorise the stage of severity of DR in order to establish adequate therapy. With proper management, cases of visual loss can be prevented. With the healthcare industry continually looking to improve the efficiency and throughput, this study seems to be a satisfactory solution that can provide quick result and timely manage eye screening.
In conclusion, this research is expected to give significant impacts to community and would become one of the key for optimizing the health sector service in the future. Further studies should be con- ducted to improve performance of these classification techniques by using larger dataset. The other supervised machine learning tech- niques such as hyperparameter optimization and hybrid supervised machine learning can also be incorporated. The other performance measure such as time complexity can also be included.

Acknowledgement

The authors wish to thank Universiti Sains Malaysia for the support it has extended in the completion of the present research through the support of USM Fellowship. The authors also wish to thank Mr Evirgen Menduh of Eye Clinic of the Sakarya University Educational and Research Hospital for dataset sharing.

References

[1] "Global reports on diabetes", *World Health Organization*, Tech. Rep.,2016.
[2] W. M. D. W. Zaki, M. A. Zulkifley, A. Hussain, W. H. W. Halim,N. B. A. Mustafa, and L. S. Ting, "Diabetic retinopathy assessment:Towards an automated system", *Biomedical Signal Processing and Control*, vol. 24,2016, pp. 72–82.
[3] B. Wu, W. Zhu, F. Shi, S. Zhu, and X. Chen, "Automatic detection of microaneurysms in retinal fundus images", *Computerized Medical Imaging and Graphics*, vol. 55, 2017, pp. 106–112.
[4] P. J. Navarro, D. Alonso, and K. Stathis, "Automatic detection of microaneurysms in diabetic retinopathy fundus images using the l* a*b color space", *JOSA A*, vol. 33, no. 1, 2016, pp. 74–83.
[5] A. Bourouis, M. Feham, M. A. Hossain, and L. Zhang, "An intelligent mobile based decision support system for retinal disease diagnosis", *Decision Support Systems*, vol. 59, 2014, pp. 341–350.
[6] S. Piri, D. Delen, T. Liu, and H. M. Zolbanin, "A data analytics approach to building a clinical decision support system for diabetic retinopathy: Developing and deploying a model ensemble", *Decision Support Systems*, 2017.
[7] G. Parthiban and S. Srivatsa, "Comparing naive bayes and decision-tree techniques for predicting the risk of diabetic retinopathy", *Digital Signal Processing*, vol. 7, no. 5, 2015, pp. 141–145
[8] H. Evirgen and M.C erkezi, "Prediction and diagnosis of diabetic retinopathy using data mining technique", *Turkish Online Journal of Science and Technology*, vol. 4, no. 3, 2014.
[9] V. Balakrishnan, M. R. Shakouri, H. Hoodehet al., "Predictions using data mining and case-based reasoning: a case study for retinopathy," *World Academy of Science, Engineering and Technology, International Journal of Medical, Health, Biomedical, Bioengineering and Pharma- ceutical Engineering*, vol. 6, no. 3, 2012, pp. 55–58.
[10] M. Skevofilakas, K. Zarkogianni, B. G. Karamanos, and K. S. Nikita,"A hybrid decision support system for the risk assessment of retinopathy development as a long term complication of type 1 diabetes mellitus", *IEEE*, 2010, pp. 6713–6716.
[11] H.-Y. Huang and C.-J. Lin, "Linear and kernel classification: When touse which?", *inProceedings of the 2016 SIAM International Confer- ence on Data Mining*, 2016, pp. 216–224.
[12] J. Weston and C. Watkins, "Multi-class support vector machines",*Cite-seer*, 1998.
[13] V. Franc and V. Hlav ac, "Multi-class support vector machine,", *IEEE*, vol. 2, 2002, pp. 236–239.
[14] Z. Wang and X. Xue, "Multi-class support vector machine", *Springer*, 2014, pp. 23–48.
[15] S. Knerr, L. Personnaz, and G. Dreyfus, "Single-layer learning revis- ited: a stepwise procedure for building and training a neural network", *in Neurocomputing. Springer*, 1990, pp. 41–50
[16] J. Friedman, "Another approach to polychotomous classification", *Technical report, Department of Statistics, Stanford University*, Tech. Rep.,1996.
[17] K. Polat, S. G unes , and A. Arslan, "A cascade learning system for clas- sification of diabetes disease: Generalized discriminant analysis and least square support vector machine", *Expert systems with applications*, vol. 34, no. 1, 2008, pp. 482–487.
[18] C.-H. Weng, T. C.-K. Huang, and R.-P. Han, "Disease prediction-with different types of neural network classifiers", *Telematics andInformat- ics*, vol. 33, no. 2, 2016, pp. 277–292
[19] L. M.Manevitz and M. Yousef, "One-class svms for documet classifi- cation", *Journal of Machine Learning Research*, pp. 139–154, 2001.
[20] M.L. McHugh, "Interrater reliability: the kappa statistic", *Biochemi- amedical*, vol. 22, no. 3, 2012, pp. 276–282.