RFE and Chi-Square Based Feature Selection Approach for Detection of Diabetic Retinopathy

Alifah[1], Titin Siswantining[1], Devvi Sarwinda[1*], Alhadi Bustamam[1]

[1]*Department of Mathematics, Faculty of Mathematics and Natural Sciences, Universitas Indonesia, Depok, Indonesia*
[*]*Corresponding author. Email:* *devvi@sci.ui.ac.id

ABSTRACT

Diabetic retinopathy, which is one of the complications in diabetes, is an eye disease that can lead to blindness. The damage happens in retina as result of a long period of diabetic mellitus. People usually do research using image data in diabetic patients. This paper presents the idea of using feature selection in diabetic retinopathy. In this study, we use the data of diabetic patients that will be extracted with feature selection method. The feature selection used in this study is Recursive Feature Elimination (RFE) and Chi-Square. Then, the classification of diabetic retinopathy is done by using Support Vector Machine (SVM). Based on the experiment's result with various hyperparameters tunning, the classification model obtains the accuracy of 97%-100% for both methods.

Keywords: *programming assessment tool, programming, clustering, K-Means*

1. INTRODUCTION

Nowadays, diabetes is one of the fastest growing disease in the world. According to World Health Organization (WHO) diabetes mellitus is in the top 10 causes of death in the world, which was issued in 2016 and was in the top of 6 non-communicable disease or chronic disease [1]. According to IDF (International Diabetes Federation), there were more than 382 million people suffering from diabetes. The number of people with diabetes increases every year. In 2014, the number of people with diabetes was 422 million patients, and diabetes became the 7th prominent reason of death worldwide in the future. Diabetes is one of the chronic diseases and one of the metabolism disorders that has characteristic of hyperglycemia, which occurs when the pancreas does not have enough insulin (it is usually called diabetes type 1 or insulin-dependent) or when the body cannot effectively use the insulin (it is usually called diabetes type 2 or non-insulin-dependent) [2]. Insulin is a hormone that regulates blood sugar.

If the diabetes is not treated properly for a long time, it can lead to a serious damage to most systems in our body, especially the nerves and blood vessels. This will cause complication. One of the complications that might happen from diabetic retinophaty is the patient might show blood vessels abnormality in retina, that later could also lead to visual impairment [3]. Diabetic retinopathy is the major cause of blindness in adult age, ranging from 20 years old to 74 years old [4]. A study by DiabCareAsia in 2008 showed that 42% of people suffer from diabetic retinopathy in Indonesia, and another study by Health Ministry of Indonesia in 2011 shows that diabetic retinopathy was the second major complication with a percentage of 33,40% after neuropathy, and was ranked in the 4th place globally for the cause of blindness [5]. Diabetic retinopathy is a long-term microvascular disease caused by diabetes mellitus [6].

It takes a lot of work to diagnose whether the patients suffer retinopathy or not. In diabetes mellitus, medical data record usually contains numerical data or Citra. Thus, the data can be processed with machine learning. Previously, there had been some experiments using machine learning to predict retinopathy diabetic. Bhatia, Karan. et. al. (2016) diagnostic of diabetic retinopathy using Amplitude Modulation Frequency

Modulation (AM/FM) method [7]. Abdillah, Bariqi et. al. (2017) classification of diabetic retinopathy with data image using Local Binary Pattern method [8]. Hariany. S. F. (2018) predicted diabetic retinopathy with data numerical and categorical using Classification and Regression Tree (CART) [9]. Sanjay, Amrita et. al. (2018) predicted breast cancer using improved feature selection method [10]. This research used feature selection (RFE and Chi-Square) and the accuracy of each methods is quite satisfactory, which scored more than 80%.

In this study, we evaluated two feature selection methods (RFE and Chi-Square) to classify Diabetic Retinopathy. The goal of this study was to classify diabetic retinopathy and gain knowledge about the best features to diagnose diabetic retinopathy from medical data. An explanation of the material and methods of used in this study (e.g., the data, preprocessing method, feature extraction method, feature selection method, and classification method) is given in Section II. In Section III, we described the results of the evaluation of our proposed method and concluded our paper in Section IV.

2. METHOD

In this section, we will describe the flow of the research, data, feature selection method, and the classifier method.

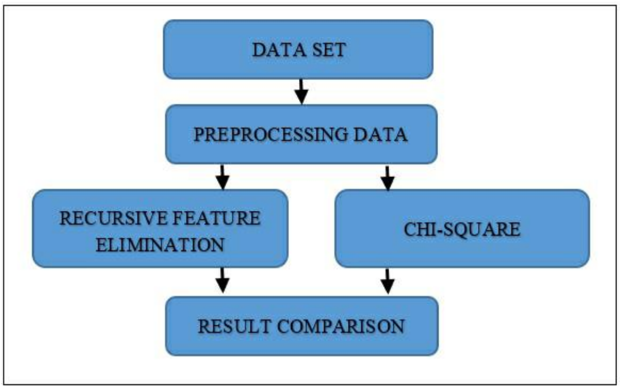

Figure 1. Overall System Flowchart

2.1 Data Acquisition

This research used medical record data of patient with diabetes type 2 in Cipto Mangunkusumo Hospital, which is located in Jakarta, Indonesia. The criteria of the sample for this research was patients with age from 18 years old. There were 174 patients with diabetes type 2 selected in this research. The data that we used in this research contained 18 features for diabetic patients also experienced retinopathy. However, the data was not complete, there were some missing values which could lead to incorrect statistical inference.

2.2 Data Pre-processing

Data preprocessing [11,12] is a major phase within the knowledge discovery process, but it is less known than other steps such as data mining. Data preprocessing actually needs more effort and time to be used [13]. Raw data usually comes with many problems such as noises, inconsistencies, redundancies, and missing values. Using raw data in the algorithm with low quality and many imperfections may cause the subsequent can be undetermined. Thus, with a proper preprocessing data step, we were able to significantly influence the quality of subsequent discoveries and decisions for the next step.

The data obtained in the beginning was in a form of raw data that could not be used. In the data set of the diabetic patients we found the missing value. The missing value of the data might be caused by many circumstances, such as failure of inputting information or incomplete extraction. For handling this problem, there were many methods could be used, for instance, removing the feature of the missing value, predicting the missing value by using statistic method and filling up the missing value by the related feature averages. In this research, preprocessing data method used by the writer was filling the empty value by adding mean value from the features.

2.3 Feature Selection Using Recursive Feature Elimination (RFE)

Our aim in using Recursive Feature Elimination (RFE) was to get the most prominent features from the data set. This method removed the lowest ranking of the features recursively by using logistic regression. First, Logistic Regression was applied to the data set and then the coefficient value for each feature was determined by the model regression. The feature with the least coefficient value was ranked as the lowest, it would be removed and after that regression was applied for the second time on the rest of the features. It can be done until all the features are been removed from the model regression. next we obtained the prominent features from the data set. this method used Logistic Regression and the formula is given by:

$$ln\left\{\frac{E(y)}{1-E(y)}\right\} = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki} \quad (1)$$

Where $E(y)$ is probability when the score of y is 1, and $1 - E(y)$ is probability when the score of y is 0.

Tabel 1. Description Of Features

Features	Description
Gender	Patient gender
Age	Patient age
Systolic Blood Pressure	Blood pressure when heart beated
Diastolic Blood Pressure	Blood pressure when heart rest between beats
Family Background	History of diabetes millitus
Microalbuminuria	Urinary albumin excretion
Blood Creatinine	Measures kidney function
e GFR	Flow of plasma from glomerulus into bowman's space
Triglyceride	Type of fat in blood
Cholesterol Total	Total amount of cholesterol in blood
Cholesterol HDL	High density lipoprotein in blood
Cholesterol LDL	Low density lipoprotein in blood
Fasting Glucose	Blood sample taken after fasting
Random Glucose	Blood sample taken from a non-fasting
Glucose 2 Hours After Eating	Blood sample taken 2 hours after aeating
Glico Hemoglobin	Average blood glucose levels for the last three months
Duration of Diabetes	How long suffer from diabetes
Body Mass Index	Measure body fat based on height and weight

2.4 Feature Selection Using Chi-Square

If we used Chi-Square, K-Best method would be selected to decide the value for K, it meant K number of the features was extracted from the data set. However, before we selected the amount of the K feature we ranked first all the features using a mathematical function called Chi-Square, a method of a statistical test to find prominent features from data set. the Chi-Square formula is given by:

$$X^2 = \sum_{i=1}^{r} \quad \frac{(n_{i1}-\mu_{i1})^2}{\mu_{i1}} + \frac{(n_{io}-\mu_{io})^2}{\mu_{io}} \quad (2)$$

Where n_{ij} is the number of instances that have a value of C_i (feature label) with $i = 1, 2, \ldots, r$ and class j, $n_{*j} = \sum_{i=1}^{k} \quad n_{ij}$ as a number of sample in class j, $n_{i*} = \sum_{j=1}^{m} \quad n_{ij}$ is the number of sample in feature C_i, expected value $\mu_{ij} = \frac{n_{*j}n_{i*}}{n}$, and n is a number of sample. We calculated Chi-Square between each feature and the target, so we could get the Chi-Square score for each feature. After that, all of the features were ranked and K-Best method was applied to get K features.

2.5 Support Vector Machine (SVM)

Support Vector Machine (SVM) is one of a supervised learning that uses a statistical approach and one of a classifier algorithm that separates into two classes. Support Vector Machine (SVM) is usually used to solve the regression and classification problem. For the first step, Support Vector Machine (SVM) was created with the purpose to solve binary classification [14]. The basic idea from Support Vector Machine (SVM) was to implicitly map the training data set into a high dimensional feature space. One kernel function that is frequently used is Radial Basis Function (RBF)[15].

The concept of the SVM method is to create an optimal hyperplane which separates data into two classes. The optimal hyperplane is an area which splits the data into classes and it is located perpendicular to the closest pattern. Patterns are dots that describe a data set. To get the optimal hyperplane, we have to find the maximum margin. Margin is a range between the hyperplane with the closest pattern for each class, while support vector is the nearest pattern to the optimal hyperplane [16].

For example, there are N sample, x_i, y_i with i = 1, 2,…, N and $y_i \epsilon\ 0{,}1$ as a class label from the data set (retinopathy diabetic and normal). Thus, the hyperplane is defined with [17]:

$$y(x) = w^T x + b \quad (3)$$

Where w is a vector of the weight parameter values, and b as a bias that has scalar value. The hyperplane that is formed will separate the data into two classes from the data set. The process will be defined with the equations below:

$$w^T x + b \geq 1, y_i \ = 1 \quad (4)$$

$$w^T x + b \leq 1, y_i \ = 0 \quad (5)$$

The equation (4) and (5) can generally be expressed in the equation:

$$y_i(w^T x + b) \geq 1 \quad (6)$$

Then the margin between the planes is $\frac{2}{|w|}$ and its maximization leads to the constrained optimization problem under the inequality constraints of (6).

$$min \left\{\frac{1}{2}\|w\|\right\} \quad (7)$$

for nonlinear mapping problem, there are some slack variables, $\xi_i \geq 0, \ i \ = 1, \ldots, N$. Thus, the equation (6) becomes:

$$y_i(w^T x + b) \geq 1 - \xi_i \quad (8)$$

So, the optimization from (7) becomes

$$min \left[\frac{1}{2}\|w\|^2 + C \sum_{i=1}^{N} \quad \xi_i\right] \quad (9)$$

Under the constrain of (8).

If we use a linear hyperplane for nonlinear problem, it may lead to problem where the spreading of the data distribution is affected. To handle this problem, the training data must be projected into a high dimensional space. To solve high dimensional problem, we used a function φthat transforms data into a high dimensional data. The data point x can be represented into $\varphi(x)$ to be a high dimensional function. This function called positive definite kernel with equation

$$k(x, x_i) = (\varphi(x).\varphi(x_i)) \quad (10)$$

That builds the decision function of the form

$$f(x) = sgn\left(\sum_{i-1}^{N} \quad \alpha_i y_i k(x, x_i) + b\right) \quad (11)$$

Where α_i is a Lagrange Multiplier.

2.6 Kernel Function

We used the Radial Basis Function (RBF) for kernel function. With the equation below [18]:

$$K(x_i, x_j') = exp(-\gamma\|x_i - x_j\|^2), \gamma \geq 0 \quad (12)$$

2.7 Experimental Design

We used Recursive Feature Elimination (RFE) and Chi-Square method to rank the features. The features selected consisted of 1, 5, 9, 13, and 18 features for each method. The features were classified to obtain accuracy and running time using Support Vector Machine (SVM). In classification process, we splitted the data into 20%, 25%, and 40% for testing data. Some parameters were optimized with grid search method. The grid search method was used to find one by one combination of the parameters that would produce the optimum model. We tried five different numbers, 0.1, 1, 5, 10, and 100 for parameter C while 0.1, 0.001, 0.005, 0.0001, and 0.0005 for parameter gamma (γ). The grid search chose the best parameters for the model, thus best output from that parameters could be obtained.

2.8 Performance Evaluation

In this research, the analysis of the results was shown by the accuracy score and the running time. The result evaluation formula is given below:

$$Accuracy \ = \frac{TP+TN}{N} \quad (13)$$

Where TP: True Positive; TN: True Negative; and N: the total of sample [19], while the running time is how long the algorithm run the process until the result is obtained from the program

3. RESULT AND DISCUSSION

In this research, Recursive Feature Elimination (RFE) and Chi-Square method were used as the feature selection method for the data set. However, before feature selection method was applied to the data set, data preprocessing needed to be done by inputting the missing value using the mean score for each feature. After that, feature ranking for diabetes mellitus was obtained from the future selection. Figure 2 explained about feature ranking using RFE.

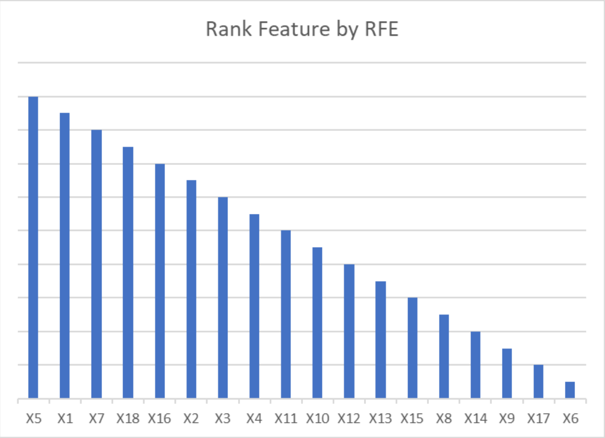

Figure 2. Feature Ranking Using Recursive Feature Elimination method

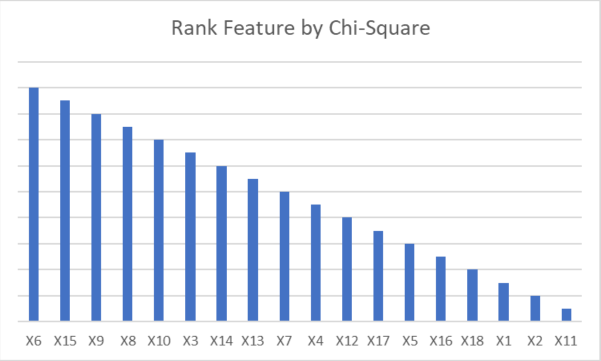

Figure 3. Feature Ranking Using Chi-Square Method

Figure 3 shows the ranking of the feature from the data set using Chi-Square method. The description of variable in Figure 2 and Figure 3 is described in the Table II. Then, the next step was to calculate the accuracy and running time using Support Vector Machine (SVM) with the parameters already determined.

Table 2. List Of Feature

Feature	Information of The Feature
X1	Gender
X2	Age
X3	Systolic Blood Pressure
X4	Diastolic Blood Pressure
X5	Family Background
X6	Microalbuminuria
X7	Blood Creatinine
X8	e GFR
X9	Triglyceride
X10	Cholesterol Total
X11	Cholesterol HDL
X12	Chholesterol LDL
X13	Fasting Glucose
X14	Random Glucose
X15	Glucose 2 Hours After Eating
X16	Glico Hemoglobina
X17	Duration of Diabetes
X18	Body Mass Index

In Table III, the accuracy from each feature selection method with 25% data test were recorded from the data set. The result showed that there was no different between Recursive Feature Elimination (RFE) and Chi-Square method. In Table IV, the accuracy from each feature selection method with 20% data test was recorded from the data set. The result showed that there wasno different between Recursive Feature Elimination (RFE) and Chi-Square method. However, the accuracy score gave a quiet satisfactory result.

Table 3. Comparison Of Accuracy Result For 25% Data Test

Number of	Recursive Feature	Chi-Square

Feature	Elimination	
1	0.9772727272727273	0.9772727272727273
5	0.9772727272727273	0.9772727272727273
9	0.9772727272727273	0.9772727272727273
13	0.9772727272727273	0.9772727272727273
18	0.9772727272727273	0.9772727272727273

Table 4. Comparison Of Accuracy Result For 20% Data Set

Number of Feature	Recursive Feature Elimination	Chi-Square
1	1.0	1.0
5	1.0	1.0
9	1.0	1.0
13	1.0	1.0
18	1.0	1.0

Table 5 Comparison Of Running Time Result For 25% Data Set

Number of Feature	Recursive Feature Elimination	Chi-Square
1	0.2117622803960444	0.5825519911777519
5	0.2388392388981515	0.6583141037845053
9	0.24783514189999778	0.994546855763474
13	0.23161092217486612	0.7692021407383436
18	0.28355189791284374	0.798209565848083

In Table V, the running time from each feature selection method with 25% data test was recorded from the data set. The result showed that the Recursive Feature Elimination (RFE) was better and more efficient than Chi-Square. In Table VI, the running time from each feature selection method with 20% data test was recorded from the data set. The result showed that Recursive Feature Elimination (RFE) was better and more efficient t than Chi-Square.

Table 6. Comparison Of Running Time Result For 20% Data Set

Number of Feature	Recursive Feature Elimination (in second)	Chi-Square (in second)
1	0.19039996715619623	0.6190005930620828
5	0.25024353956359846	0.6911498860499705
9	0.26993911845596585	0.7694805991668545
13	0.2697945342715684	0.7658262785334955
18	0.28555153288243673	0.8030616147298133

4. CONCLUSION

In this paper, we presented feature selection method with medical data diabetes mellitus and test the result to compare we calculation the accuracy from the feature selection using machine learning method with Support Vector Machine (SVM). From our experimental result, we analyzed that classification for data diabetes mellitus achive above 95% accuracy value. The execution time obtained in Recursive Feature Elimination (RFE) method was less than 0,3 ms and the Chi-Square Method less than 1,0 ms. The best accuracy was obtained when the data test size 0,20% in each method we get the 1.0 for each method. Meanwhile, if we use the 0,25% data test size, then the accuracy score is 0.9772727272727273 for every experiment with different feature or different method.

However, from the result of feature selection method using Recursive Feature Elimination (RFE) and Chi-Square, we could see the feature ranking is very different between each method. For example, with Recursive Feature Elimination (RFE) method in feature X6 gets the lowest ranking, but in the Chi-Square method, the feature X6 had the highest ranking. All of this result was affected by the process and the content of the data set .

ACKNOWLEDGMENT

This research is financially supported by PITTA B 2019 Research Grant from Universitas Indonesia (ID number NKB-0678/UN2.R3.1/HKP.05.00/2019).

REFERENCES

[1] https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death

[2] American Diabetes Association, Diagnosis and Classification of Diabetes Mellitus, Diabetes Care, 2009.

[3] Mathebula. Solani D., Biochemical Changes in Diabetic Retinopathy Triggered by Hyperglycaemia: A Review, Aveh Journal, 2017.

[4] Lee. Ryan et. al. Epidemiology of Diabetic Retinopaathy, Diabetic Macular Edema and Related Vision Loss, NCBI, 2015.

[5] Kementrian Kesehatan Republik Indonesia, Info Datin: Situasi dan Analisis Diabetes (Pusat Data dan Informasi Kementrian Kesehatan Republik Indonesia, Jakarta, 2014). Available at http://www.depkes.go.id/resources/download/pusdatin/infodatin/infodatin-diabetes.pdf

[6] Chawla. Aastha, Chawla. Rajeev, and Jaggi. Shalini, Microvascular and Macrovascular Complications in Diabetes Mellitus: Distinct or Continu?, NCBI, 2016.

[7] Bhatia. Karan, Arora. Shikhar, Tomar. Ravi, Diagnosis of Diabetic Retinopathy Using Machine Learning Classification Algorithm, 2nd International Conference on Next Generation Computing Technologies (NGCT), 2016.

[8] Abdillah. Bariqi, Bustamam. Alhadi, Sarwinda. Devvi, Classification of Diabetic Retinopathy Through Texture Features Analysis, International Conference on Advanced Computer Science and Information Systems (ICACSIS), 2017.

[9] Hariany. S. f., Siswantining. T., Bustamam. A., Budiman. B., Result Comparison Between Categorical and Numerical Predictor Variables on CART Method in Predicting Factors Related to Diabetic Retinopathy in Patients With Type 2 Diabetes Mellitus, 2018.

[10] Sanjay. Amrita, Nair. H. Vinayak, Murali. Sruthy, A Data Mining To Predict Breast Cancer Using Improved Feature Selection Method On Real Time Data, International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2018.

[11] S. García, J. Luengo, F. Herrera, Data Preprocessing in Data Mining, Springer, 2015

[12] S. Garcia, J. Luengo, Ferrera. Tutorial on Practical Tips of The Most Influential Data Preprocessing Algorithms in Data Mining. Knowledge-Based Systems, 1-29, 2016.

[13] D. Pyle. Data Preparation for Data Mining, Morgan Kaufmann Publishers Inc., 1999.

[14] P V. Arivoli and T. Chakravarthy, Document Classification Using Machine Learning Algorithms-A Review, International Journal of Scientific Engineering and Research (USER), vol. 5, iss. 2, 48-54, 2017

[15] M. H. Afif and A. R. Hedar, Data Classification using Support Vector Machine Integrated with Scatter Search Method, Japan-Egypt Conference on Electronics, Communications, and Computers, 2012.

[16] Srivastava, D.K and Bhambhu, L, Data Classification Using Support Vector Machine, Journal of Theoritical and Applied Information Technology, 2005.

[17] Rustam, Z., Yaurita, F, Insolvency Prediction in Insurance Companies Using Support Vector Machines and Fuzzy Kernel C-Means, Journal of Physics: Conference Series, Volume 1028, conference 1, 2018.

[18] Patle. Arti, Chouhan. Deepak Singh, SVM Kernel Function for Classification, International Conference on Advances in Technology and Engineering (ICATE), 2013.

[19] Sarwinda. Devvi, Siswantining. Titin, Bustamam. Alhadi, Classification of Diabetic Retinopathy Stages Using Histogram of Oriented Gradients and Shallow Learning, International Conference on Computer, Control, Informatics and its Applications, 2018.