Selection of Informative Genes to Classify Type 2 Diabetes Mellitus using Support Vector Machine

Firda Nurul Auliah
Department of Computer Science
Hasanuddin University
Makassar, Indonesia
auliahfn14h@student.unhas.ac.id

Armin Lawi
Department of Computer Science
Hasanuddin University
Makassar, Indonesia
armin@unhas.ac.id
http://orcid.org/0000-0003-1023-6925

Sri Astuti Thamrin
Department of Statistics
University of Hasanuddin
Makassar, Indonesia
tuti@unhas.ac.id
http://orcid.org/0000-0002-2512-0266

Edy Budiman
Department of Informatics
Mulawarman University
Samarinda, Indonesia
edybudiman.unmul@gmail.com
http://orcid.org/0000-0002-3164-5157

***Abstract*—Type 2 Diabetes Mellitus is a group of disorders with characteristics such as insulin resistance, impaired insulin secretion, and increased glucose production, which has patients reaching 90% - 95% of the overall population of people with diabetes mellitus. There at most 70% of Indonesians are unaware attacked by diabetes. Therefore, an early detection has an important role and the utilization of microarray technology can cope this problem. One of the challenges for microarray applications is to select an appropriate number of the most significant genes for data analysis. Besides that, it is hard to accomplish a satisfactory classification results by Support Vector Machine (SVM) due to the dimensionality and the over-fitting problems. For this reason, it is desirable to select informative genes firstly in order to improve classification accuracy of SVM classifier. In this study, we use the Information Gain in order to determine informative features of data to get better classification performances and then the SVM is applied to the selected features. Based on the result, the informative genes were selected from 25,770 to 309 genes. SVM can predict sample with accuracy 100% and area under curve 100%. There is a probe which is a gene that has various functions, including regulating neurotransmitter releases, heart rate, insulin secretion, neuronal excitability, epithelial electrolyte transport, smooth muscle contraction, and cell volume.**

Keywords*—*Type 2 diabetes mellitus, microarray, information gain, support vector machine

I. INTRODUCTION

The Type 2 Diabetes Mellitus (T2DM) is a group of disorders with characteristics such as insulin resistance, impaired insulin secretion, and increased glucose production, which has patients reaching 90-95% of the overall population of people with diabetes mellitus [1]. The World Health Organization (WHO) predicts that the number of DM patients in Indonesia will increase from 8.4 million in 2000 to around 21.3 million in 2030 [2] and around 70% of Indonesian people are not aware affected by diabetes. Therefore, early detection has an important role and microarray technology has a role in this case.

The official website of the National Center for Biotechnology Information (NCBI) define that microarray is a hybridization of nucleic acid samples (targets) to a very large set of oligonucleotide probes, which are attached to a solid support, to determine the sequence or to detect variations in a gene sequence or expression or for gene mapping. The implementation of microarray is stored as gene data of patients who affected by T2DM.

Recent years, application of microarray have a major influence in determining the informative gene that causes disease. Micorarray can determine the expression of thousands genes and simultaneously monitor the ongoing biological processes [3], [4], [5]. One of the challenges for microarray applications is to choose the number of genes that are most significant for data analysis. Moreover, it is difficult getting satisfactory classification results due to high dimensional and over-fitting problems. Therefore, selecting informative genes can be used to improve classification accuracy [4], [5], [6], [7].

In this paper, we use Information Gain (IG) in order to determine informative features of the data to get a better classification, and the SVM classification method is applied to classify T2DM from the selected features. The paper is organised as follows. We explained introduction in the first section, followed by material and methods. In this second section, we describe IG and SVM. In the next section, we present the result. The conclusions and the future work are presented further in the last section.

II. METODOLOGY

A. Type 2 of Diabetes Mellitus

Diabetes mellitus (DM) is a metabolic disorder characterized by hyperglycemia that is related to the problem of carbohydrate, fat and protein metabolism [8]. T2DM is a group of abnormalities with characteristics such as insulin resistance, impaired insulin secretion, and increased glucose production. T2DM begins with an abnormal glucose homeostasis period, which is impaired fasting glucose (IFG) or impaired glucose tolerance (IGT) [9].

T2DM is a more common type of diabetes that has more patients than Type 1 DM. T2DM sufferers reach 90-95% of the total population of DM patients, generally aged over 45 years old, but recently people with T2DM among teenagers and children have increases the population. The etiology of T2DM is multifactorial which has not been fully revealed clearly. Genetic factors and environmental influences play an important role in causing T2DM, including obesity, high-fat and low-fiber diets, and lack of exercise [10].

B. Microarray

Microarray has been widely used in biomedical research. Microarray is a tool designed to measure expression levels of thousands of genes in a disease or cell type. Microarray technology capabilities can be used to find out and examine differences in diseases that are often found in molecular biology and medical biology. This method uses a tool in the form of a slide made of glass and consists of thousands or even tens of thousands of blocks [11].

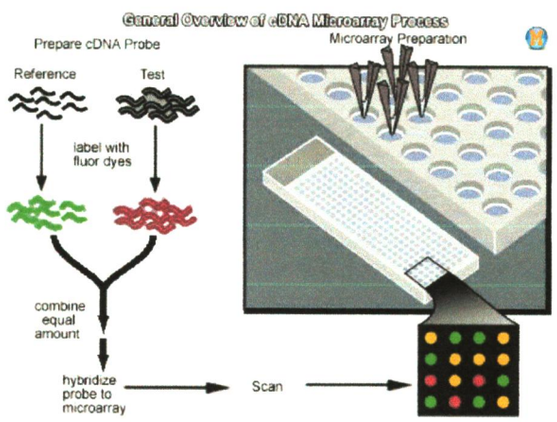

Fig. 1. General Overview Microarray

As depicted on Fig. 1, after taking a sample in the form of blood, then the mRNA is isolated. The mRNA is transformed into cDNA, and then the cDNA is labeled with two different dyes. The first DNA is labeled with fluorescent green Cy3 and the second DNA is labeled fluorescent red Cy5, then cDNA was mixed and put into a microarray slide that had been previously probed. After that, the hybridization process, which is incubating the microarray chip for one night at 60°. The next step is scanning laser light, the aim is to detect the fluorescent intensity visualized on the computer. The final step is to do data analysis that displays different gene expressions [12].

The results of the interpretation of DNA microarray technology shows in Fig. 2, if the expression of genes is higher it will appear red. Conversely, if the expression in the experimental sample is lower, it will appear green. Finally, if there are two similar expressions in a sample, yellow will appear. A black dot indicates that there is no cDNA bound to the DNA in the gene located in that place. This shows that the gene is inactive (all genes in the experiment are active) [11].

Data collected through microarrays can be used to create gene expression profiles. The profile shows simultaneous changes in the expression of genes to response a particular condition or treatment.

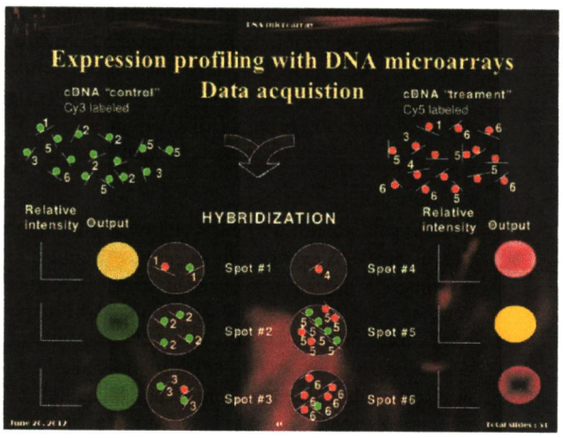

Fig. 2. Interpretation Result of DNA Microarray Technology

C. Preprocessing Data

The first step that must to do on data is preprocessing. Preprocessing is a process or step that converts initial data into better quality data. Preprocessing is useful for improving data quality and reducing data noise [13]. Preprocessing is done by data transformation and data normalization. In this study, preprocessing data is done by data transformation and data normalization.

D. Information Gain

Information Gain (IG) is the simplest feature selection method by ranking attributes and widely used in text categorization applications, microarray data analysis and image data analysis [14]. IG can reduce noise caused by irrelevant features. Information Gain detect the features detect the features that have the most information based on a particular class. Determining the best attributes performed by calculating the entropy value first. *Entropy* is a measure of class uncertainty by using probability events or attributes [15]. The Entropy can be evaluated using equation (1) and then the Information Gain using equation (2) [16].

$$Entropy(S) = \sum_{i=1}^{c} (-p_i) \log_2 p_i \tag{1}$$

Here c is amount of value in the classification class and P_i is the proportion of S that belongs to class i.

$$Gain(S, A) = Entropy(S) - \sum_{Values(A)} \frac{|S_v|}{|S|} Entropy\ (S) \tag{2}$$

With A being an attribute, v is a possible value for attribute A, Values (A) is a set of possible values for A, | Sv | is the number of samples for the value of v, | S | is the sum of all data samples and Entropy (Sv) is entropy for samples that have a value of v.

E. Support Vector Machine

Support Vector Machine (SVM) developed by Boser, Guyon, and Vapnik, was first introduced in 1992 at the Annual Workshop on Computational Learning Theory. The basic concept of the SVM method is actually a combination or combination of computational theories that have existed in the previous year, such as hyperplane margins. SVM method uses dot product function. SVM is an attempt to find the best hyperplane that functions as a separator of two classes in input space [17]. The SVM method have been used for classification problem in Bioinformatics [18], especially using microarray gene expression data [4], [5], and also for quality control of DNA sequencing [12].

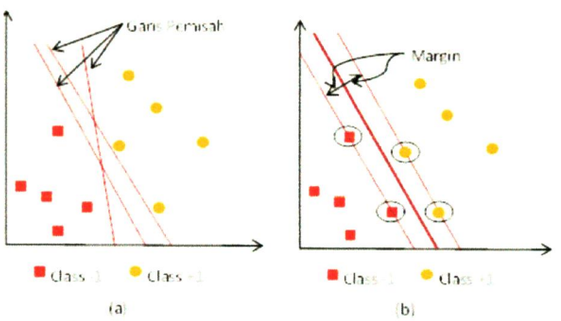

Fig. 3. Determine the Best Hyperplane

Fig. 3 (a) shows several patterns that are members of two classes: +1 and -1. The pattern incorporated in class -1 is symbolized in red (box), while the pattern in class +1 is symbolized in yellow (circle). Classification problems can be translated by finding a line (hyperplane) that separates the two groups. Various discrimination boundaries are shown in Fig. 3 (a).

The best hyperplane separator between the two classes can be found by measuring the hyperplane margin and finding the optimum point of the hyperplane. Margin is the distance between the hyperplane and the closest pattern of each class. The closest pattern is called support vector. The solid line in Fig. 3 (b) shows the best hyperplane, which is located in the middle of the two classes, while the red and yellow points in the black circle are support vector. Attempts to locate this hyperplane is the point of the learning process on SVM.

For example the available data that notated as $\vec{x_i} \in \Re^d$ while each label notated $y_i \in \{-1, +1\}$ for $i = 1, 2, \ldots, l$, where l is the amount of data. It is assumed that both class -1 and +1 can be completely separated by a hyperplane dimension d, which is defined:

$$\vec{w} \cdot \vec{x} + b = 0 \tag{3}$$

Pattern $\vec{x}$ which include to class -1 (negative samples) can be formulated as pattern that fulfill inequality that include *class* -1 (negative samples) can be formulated as *pattern* that meets inequality:

$$\vec{w} \cdot \vec{x} + b \leq -1 \tag{4}$$

while *pattern* $\vec{x}$ that include *class* +1 (positif samples):

$$\vec{w} \cdot \vec{x} + b \geq +1 \tag{5}$$

The biggest margin can be found by optimizing the distance value between the hyperplane and its closest point, that is $\frac{1}{\|\vec{w}\|}$. This can be formulated as *Quadratic Proraming (QP) problem*, that is, looking for the minimum point of the equation (6), with *constraint* equation (7):

$$\min_{\vec{w}} \tau(w) = \frac{1}{2}\|\vec{w}\|^2 \tag{6}$$

$$y_i(\vec{w} \cdot \vec{x}_i + b) - 1 \geq 0, \quad \forall_i \tag{7}$$

Input data that notated x_i, is output from data x_i, $\vec{w}, b$ is the parameters that are searched for value. The formulation above, wants to minimize the objective function (*obyective function*) $\frac{1}{2}\|\vec{w}\|^2$ or maximize quantity $\|\vec{w}\|^2$ by paying attention to the delimiter as the equation (4) and (5). If the output data $y_i = +1$, so the limiter become $\vec{w} \cdot \vec{x} + b \geq +1$. Conversely if $y_i = -1$, the limiter become $\vec{w} \cdot \vec{x} + b \leq -1$.

These problems can be solved by various computational techniques, like *Lagrange Multiplier*:

$$L(w, b, \alpha) = \frac{1}{2}\|\vec{w}\|^2 - \sum_{i=1}^{l} \alpha_i (y_i((\vec{x_i} \cdot \vec{w} + b) - 1)), \tag{8}$$

where $i = 1, 2, \ldots, l$ and α_i is *Lagrange Multiplier* with $\alpha_i \geq 0$. Optimal value of the equation (8) can be determined by minimizing L against $\vec{w}$ and b, and maximize L against α_i. If we paying attention to the quality that at the optimal point of *gradient* $L = 0$ in equation (8) can be modified as a problem maximization that only contains α_i, as shown in the equation (9):

$$\max \sum_{i=1}^{l} \alpha_i - \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} \alpha_i \alpha_j y_i y_j \vec{x_i} \cdot \vec{x_j} \tag{9}$$

Subject to:

$$\sum_{i=1}^{l} \alpha_i y_i = 0, \ \alpha_i \geq 0 \ (i = 1, 2, \ldots, l) \tag{10}$$

According to the equation (9) and (10), we will get α_i which is mostly positive referred to *support vector*.

SVM is a two-class classifier inherently. The natural way to classify into multiclass using SVM is to use the methods of either one-versus-all (OVA) or one-to-one (OVO). These methods are using the SVM two-class classifier repeatedly [19], [20], [21].

III. DISCUSSION

A. Data Description

We applied the method using microarray dataset T2DM in a public functional genomics data repository of the Gene Expression Omnibus (GEO) by the National Center for Biotechnology Information (NCBI), which is available at [22]. The dataset is mRNA expression data from skeletal muscle of type 2 diabetes that consisted of 118 samples which contains 25.770 gene expression. 26 samples of Glucoseintolerant, 47 samples of normal and 45 samples of T2DM who had been free of hypoglycemic medication for one week. RNA was hybridized to Affymetric

HGU133plus2 platform following manufacturer's directions [23].

In this study, the data is divided by the proportion of 80% training data and 20% test data. Each data from the training data and test data was taken randomly with emphasis on all classifications included in the two data. Known T2DM data from GSE18732 there are 2 classes, namely normal (dmverified: 0) and T2DM (dmverified: 1). GSE18732 data consists of 118 samples used, so 94 (80%) samples taken must have 2 classes and 24 (20%) samples of test data must also consist of 2 classes of T2DM classification.

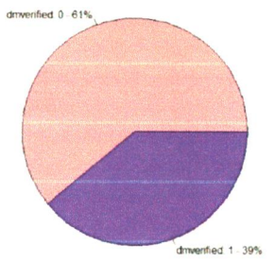

Fig. 4. Description Normal Sample and T2DM Sample

B. Methods

1) Preprocessing: On preprocessing, we transform a dataset using *logarithmic transformation* and *quantile normalization*.

2) Information Gain: we first calculate the entropy of features using eq. (1) and then calculate the information gain using eq. (2). We use InfoGainAttributeEval and Ranker evaluation tools of WEKA, where the results show that there are 390 genes selected from 25770 genes. The best genes are taken by ranking the 50, 40, 30, 20, and 10 best features of the available data, to find out the most appropriate ranking used in this data. The the results show that the top 10 is most suitable for the data.

TABLE 1. GENE INFORMATIVE

Features/Genes	Rank Value
X8292	0.2198
X24275	0,2198
X10089	0.2111
X6924	0.1656
X3269	0.1602
X13319	0.1492
X6348	0.1442
X6349	0.1442
X21348	0.1395
X24981	0.1395

3) Support Vector Machine: The next step is using SVM to classify the dataset after getting informative genes. The classification result were tested using top 10, 20, 30, 40, and 50 informative genes with confusion matrix. It shows that the highest accuracy is using top 10 of informative genes and the results classification of the test data were 100% with 1000 times iteration. In addition, Area Under Curve (AUC) is also used to further review the classification results. The result showed that testing with AUC was 100%.

4) Informative Genes: After obtaining the results of 10 genes, then look for the symbol of the probe that can be used to find more information about the probe that has been obtained. The results show that there are 8 probes are known which can be seen in the Table 2.

TABLE 2. GENE SYMBOL

Feature/Gene	*Symbol*
ENST00000262662	CDKN2C
ENST00000298694	ARHGEF40
ENST00000303372	TCTN2
ENST00000313860	LIMCH1
ENST00000328032	KCNH7
ENST00000348956	CKB
ENST00000381753	LIMCH1
ENST00000383362	C2

IV. CONCLUSION

In this paper, we have classified the T2DM patients with the following three steps. The first step is to preprocess data, we transform and normalize the dataset using *logarithmic transformation* and *quantile normalization*. The next step is to select informative gene using IG from 25,770 to 390 genes. The obtained results that there is 1 gene/probe has informative that related with T2DM, namely KNCH7. According to NCBI, the KNCH7 has various functions, including regulating the release of neurotransmitters, heart rate, insulin secretion, smooth muscle contraction, and cell volume. The last step is to classify the informative gene using SVM. The results classification of the test data was 100% with 1,000 iterations and 100% AUC.

REFERENCES

[1] American Diabetes Association, et al., "Diagnosis and classification of diabetes mellitus," *Diabetes care*, 37 Supplement 1, S81-S90, 2014.

[2] Soewondo, P., "Pemantauan Kendali Diabetes Melitus," In: Soegondo, S., Soewondo, P., Subekti, I., Eds. Penatalaksanaan Diabetes Melitus Terpadu bagi dokter maupun edukator diabetes". Jakarta: Fakultas Kedokteran Universitas Indonesia, pp. 111-133, 2011.

[3] Diaz-Uriarte, Ramon, and Sara Alvarez De Andres. "Gene selection and classification of microarray data using random forest." *BMC bioinformatics*, vol. 7, no. 1, p. 3, 2006.

[4] M.P.S. Brown, W.N. Grundy, D. Lin, N. Cristianini, C.W. Sugnet, T.S. Furey, Jr.M. Ares, and D. Haussler, "Knowledge-based analysis of microarray gene expression data by using support vector machines," PNAS, 97(1), pp. 262–267, 2000.

[5] Mukherjee, Sayan, P. A. S. D. Tamayo, D. Slonim, A. Verri, T. Golub, J. Mesirov, and T. Poggio. "Support vector machine classification of microarray data". http://www.ai.mit.edu/research/abstracts/abstracts2001/machine-learning/11mukherjee2.pdf, 2001. (Accessed on March 10, 2019)

[6] Nguyen, Danh V., and David M. Rocke. "Tumor classification by partial least squares using microarray gene expression data." *Bioinformatics*, vol. 18, no. 1, pp. 39-50, 2002.

[7] Vural, Halit, and A. Subasi. "Data-mining techniques to classify microarray gene expression data using gene selection by SVD and information gain." *Model Artificial Intel*, vol. 6, pp. 171-182, 2015.

[8] Kang-Birken, S. I., dan Dipiro, J. T."Sepsis and septic shock". Pharmacotherapy: A Pathophysiologic Approach, Seventh Edition. *McGraw-Hill Companies*, United States of America, pp.1943-1944, 2008.

[9] Powers A. "Diabetes Mellitus In: Gibson R.J. ed. The 16th Edition Of Harrison's Principles Of Internal Medicine". USA. *The McGraw-Hill Companies*, vol 16, pp. 3779-829, 2005.

[10] Isselbacher, K. J., Braunwald, E., Wilson, J. D., Martin, J. B., Fauci, A. S., and Kasper, D. L. "Harrison's Principles of Internal Medicine. Boston". *McGraw-Hill Inc. Shock*, vol. 5, no. 1, pp. 78, 1996.

[11] Razavi, Amirnader Emami. "DNA Microarray". *Isfahan University of Medical Science. School of Pharmacy. Department of Clinical Biochemistry*, 2012.

[12] Öz, Ersoy, and Hüseyin Kaya. "Support vector machines for quality control of DNA sequencing." *Journal of Inequalities and Applications*, no. 1, p. 85, 2013.

[13] Setyohadi, D. B., Kristiawan, F. A., dan Ernawati, E., "Perbaikan Performansi Klasifikasi Dengan Preprocessing Iterative Partitioning Filter Algorithm". *Telematika vol 14, no. 01, 2017*.

[14] Chormunge, Jena, S. "Efficient Feature Subset Selection Algorithm for High Dimensional Data". *International Journal of Electrical and Computer Engineering (IJECE) vol*, 6, no 4, pp. 1880-1888, 2016.

[15] Shaltout, N. A., El-Hefnawi, M., Rafea, A., and Moustafa, A. "Information Gain as a Feature Selection Method for the Efficient Classification of Influenza Based on Viral Hosts". *Proceedings of the World Congress on Engineering*. London, U.K, WCE, vol. 1, pp. 3-7, 2016.

[16] Mitchell, T. M. "Machine learning". *Burr Ridge, IL: McGraw Hill*, vol. 45, no. 37, pp. 870-877, 1997.

[17] Lawi, Armin, Ali Akbar Velayaty, and Zahir Zainuddin. "On identifying potential direct marketing consumers using adaptive boosted support vector machine." In *2017 4th International Conference on Computer Applications and Information Processing Technology (CAIPT)*, pp. 1-4. IEEE, 2017.

[18] Byvatov, Evgeny, and Gisbert Schneider. "Support vector machine applications in bioinformatics." *Applied bioinformatics*, vol. 2, no. 2, pp. 67-77, 2003.

[19] Jafar, Nurkamila, Sri Astuti Thamrin, and Armin Lawi. "Multiclass classification using Least Squares Support Vector Machine." In *2016 International Conference on Computational Intelligence and Cybernetics*, pp. 7-9. IEEE, 2016.

[20] Lawi, Armin, and Yudhi Adhitya. "Classifying Physical Morphology of Cocoa Beans Digital Images using Multiclass Ensemble Least-Squares Support Vector Machine." In *Journal of Physics: Conference Series*, vol. 979, no. 1, p. 012029. IOP Publishing, 2018.

[21] Lawi, Armin, and M. Sya'Rani Machrizzandi. "Facial Expression Recognition using Multiclass Ensemble Least-Square Support Vector Machine." In *Journal of Physics: Conference Series*, vol. 979, no. 1, p. 012032. IOP Publishing, 2018.

[22] NCBI, "mRNA expression data from skeletal muscle of type 2 diabetes". *Gene Expression Omnibus*. https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE18732 (online) Jan 19, 2010. (Accessed on March 10, 2019).

[23] Gallagher, Iain J., Camilla Scheele, Pernille Keller, Anders R. Nielsen, Judit Remenyi, Christian P. Fischer, Karim Roder, et al. "Integration of microRNA changes in vivo identifies novel molecular features of muscle insulin resistance in type 2 diabetes". *Genome medicine*, vol. 2, no. 2, p. 9, 2010.