A Preliminary Study on Identifying Probable Biomarker of Type 2 Diabetes using Recursive Feature Extraction

1st Nur Nilamyani
Department of Computer Science
Hasanuddin University
Makassar Indonesia
nilamyanin14@student.unhas.ac.id

2nd Armin Lawi
Department of Computer Science
Hasanuddin University
Makassar Indonesia
armin@unhas.ac.id

3rd Sri Astuti Thamrin
Department of Statistics
Hasanuddin University
Makassar, Indonesia
tuti@unhas.ac.id

***Abstract*—Microarray technology has the ability to measure the level expression of thousand genes by single experiment and it can be used by biologist to study about the effect of treatments, disease and developmental stages on their expressions. Microarray based on gene expression profiling can be used to observing the response of expression genes to pathogens and identify which expressions genes are changed by comparing the expression in infected to that uninfected cells or tissue. Type 2 diabetes mellitus is a metabolic disorder that causes an increase in blood sugar due to decreased insulin secretion by pancreatic beta cells and insulin disorder (insulin resistance). The number of incidences of diabetes mellitus in Indonesia reached 10 million and 53% from the patients do not realized that they are infected and 90% case of diabetes from whole world is type 2 of diabetes. Therefore, in this paper, we identify probable biomarker of type 2 diabetes using microarray based on gene expression data. But the risk of using microarray data is the large dimension of data so have to find a way how to solve that problem to get a good prediction result. In this paper will use recursive feature extraction for predicting biomarkers of diabetes mellitus type 2 from microarray gene expression data.**

Keywords—biomarker, microarray, feature ranking, feature selection, type 2 diabetes

I. Introduction

Type 2 diabetes mellitus is a metabolic disorder disease that causes an increase in blood sugar due to decreased insulin secretion by pancreatic beta cells or insulin resistance and it makes the body is unable to respond to insulin as a normally. This type of diabetes does not show symptoms that can be seen directly so it is difficult to detect the disease. The number of Type 2 Diabetes Mellitus sufferers continues to increase. According to the results of research from International Diabetes Federation (IDF) in 2015, there were 415 million people in the whole world suffering from diabetes and 95% of it are people who suffering in type 2 of diabetes and about 10 million in Indonesia. More than 60 percent of people with diabetes are not aware of diabetes, so often Type 2 DM sufferers are diagnosed after complications occur. Late handling of Type 2 Diabetes Mellitus patients can have serious consequences.

Until now there is no medical treatment known to cure type 2 diabetes permanently. Treatment measures in patients only function to keep blood glucose levels as normal as possible and control the symptoms that appear later on not to cure the disease permanently. Therefore, research on appropriate treatment for type 2 diabetes's patients continues to be done, one of which is identifying biomarkers from Type 2 Diabetes Mellitus. Biomarkers can help early detection of a disease so that proper treatment of the disease can be done. Biomarker identification of Type 2 Mellitus diabetes continues to develop, one of which is using DNA microarray technology.

DNA microarray technology is used to determine the expression level of thousands of genes carried out in one experiment, and simultaneously monitor ongoing biological processes [1]. Identification of biomarkers using gene expression can provide very important results because it can help the development of personalized medicine and also help researchers to find appropriate treatment for serious diseases, including type 2 diabetes mellitus [2]. Several studies using microarray data have also been conducted to find biomarkers in several types of complex diseases such as research on biomarkers of cancer using micro array data using Feature Selection and Semi supervised learning [3] and also conducted a similar study which identified biomarkers of cancer but used the Network-Constrained Support Vector Machine Method [4]. In addition to cancer, another complex disease that is also of concern is Type 2 Diabetes Mellitus. Several studies have been conducted to find biomarkers of type 2 DM using the PreDx DRS algorithm [5]. Similar research was also identifying type 2 DM biomarkers using Discriminative Area of Functional Activity [2].

The main problem of microarray data itself is the large of number dimensions because the number of genes is bigger than the number of samples, we are using feature selection method is needed to select informative genes. In this paper, we will use recursive feature extraction to identify biomarker of type 2 diabetes.

II. Methodology

A. Biomarker

Biomarker or "biological markers", is an objective indication of medical conditions observed from outside the patient which can be measured accurately and reproductively. Biomarkers or biological markers are marker molecules that are typical for cells, which can be used to diagnose a disease and therapeutic target molecules that cause certain diseases [6]. Another opinion says that biomarkers function as an early warning system for disturbances that can be the potential for the emergence of a disease due to pressure experienced by organisms [7].

B. Type 2 of Diabetes

Diabetes mellitus is a disease which marked by the occurrence of hyperglycemia and impaired metabolism of carbohydrates, fats, and proteins associated with insulin

978-1-5386-8050-6/18/$31.00 ©2018 IEEE

secretion. According to the etiological classification DM according to the American Diabetes Association is divide into 4 types, namely Type 1 Diabetes Mellitus or Insulin Dependent Diabetes Mellitus, Type 2 Diabetes Mellitus or Non-dependent Insulin Diabetes Mellitus. Other type of Diabetes Mellitus, Gestational Diabetes Mellitus [8]

Patients with Type 2 Diabetes Mellitus there is hyperinsulinemia but insulin cannot bring glucose into the tissue because of insulin resistance which is a decrease the insulin's ability to stimulate glucose uptake by peripheral tissues and to inhibit glucose production by the liver. Because the occurrence of insulin resistance (insulin receptors are inactive because they are considered high levels in the blood) will result in relative insulin deficiency. This can lead to reduced insulin secretion in the presence of glucose with other insulin secretion ingredients so that the pancreatic beta cells will experience desensitization in the presence of glucose. This type of DM onset occurs slowly because the symptoms are asymptomatic. The existence of resistance that occurs slowly will result in reduced receptor sensitivity for glucose. This type of DM is often diagnosed after complications [9]

C. DNA Microarray

Microarray (DNA microarray) is a chip which measures the level of gene expression. There are several different types of microarray technologies such as Affymetrix, Agilent etc. This technology helps researchers in studying various diseases, especially cancer. Therefore, microarray technology can help diagnose, monitor and predict a disease. This method uses a tool in the form of a slide made of glass and consists of thousands and even tens of thousands of blocks [7], Fig. 1.

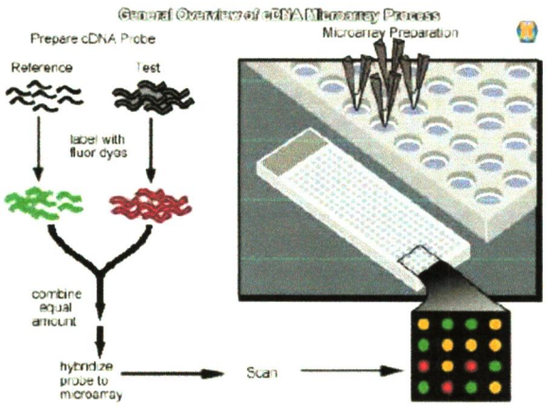

Fig. 1. DNA Microarray Technology

Principle working of microarray technology is to measure the amount of hybridization of mRNA in the cDNA in the chip. In general, the microarray analysis uses two different samples such as normal skin cells with diseased skin cells. Both of these samples were isolated from the mRNA and then placed in a microarray chip. Then, the chip is given a radioactive marker to produce color fluorescent after the scanner is connected. Finally, the existing color patterns will analyze using the two samples [7]

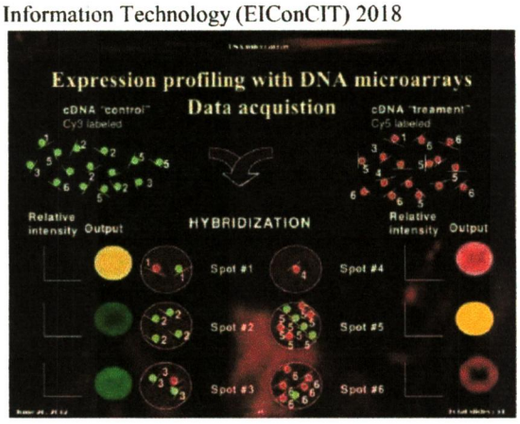

Fig. 2. Interpretation result of DNA microarray technology

According to Fig. 2, the interpretation results of DNA microarray technology. if the particular gene expression is higher, then 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, it will appear yellow. A black dot indicates that no cDNA is bound to the DNA in the gene located in that place. This shows that the gene is inactive (all genes are active) [7]

Data generated from Microarray is the type of data used in bioinformatics. The characteristics of data microarray are the small amount of data and the large number of features. This data contains gene information because the number of features is very large. Although these types of data are difficult to screen but, the results obtained will be very useful such as the discovery of new drugs (drug discovery) and the determination of the type of treatment for patient handling.

D. Dimensional Reduction

Dimensionality is the main problem of microarray data. The dimensionality problem in microarray data is very influential on computation time and the level of accuracy in the classification process. Dimensional problems in micro array data can cause curse of dimensionality which greatly affects the level of accuracy in the classification process. This is due to the number of features is bigger than the number of samples available. Therefore, a dimensional reduction process is needed on micro array data in order to save computational time and also increase accuracy. Kind of dimension reduction technique are feature ranking and feature selection.

E. Normalization

Normalization is a preprocessing stage that aims to correct systematic differences between genes. The experimental microarray data consists of several samples, having two or more sample groups (the group here refers to conditions or times). These samples are placed in different arrays, sometimes there is an imbalance between the samples. So, the difference between arrays can only be done if the array is normalized first. The process of normalizing microarray data depends heavily on technology. In this study, the RMA (Robust Multi-Array Average) method will be used to normalize the microarray data used [8].

F. Feature Ranking

There is a considerable difference between the number of samples and features (genes) in microarray data. Therefore, the use of the feature selection approach for different features as much as it sometimes does not give good results. Therefore, before the feature selection process, the ranking feature will be processed by giving a score to each feature that aims to eliminate irrelevant features using several ranking feature methods, namely:

1) Chi-Squared: to find the weights of attributes we will using chi-squared tes to calculate it

$$\chi^2 = \sum_{i=1}^{b} \sum_{j=1}^{m} \frac{(o_{ij}-e_{ij})^2}{e_{ij}} \quad (1)$$

Where b is row, m is column, o_{ij} is the observed value, e_{ij} is the expected value.

2) Information Gain will used to find the weights of discrete attributes based on their correlation with continuous class attribute[9].

$$H(Class) + H(Attribute) - H(Class, Attribute) \quad (2)$$

3) Random Forest Importance: The algorithm finds weights of attributes using Random Forest algorithm and for the classifier we are using OneR classifier. OneR algorithm generates one rule for each predictor in the data, then selects the rule with the smallest total error as it's "one rule".

OneR Algorithm
For each predictor
  For each value of that predictor, make a rule as follows:
    Count how often each value of target (class) appear
    Find the most frequent class
    Make the rule assign that class to this value of the predictor
  Calculate the total error of the rules of each predictor
Choose the predictor with the smallest total error

Fig. 3. Classifier Algorithm

G. Feature Selection

We are using feature selection to eliminate the non-relevant features and extract the features which best distinguish between the given classes (conditions) by feature ranking. We are applying feature selection for retaining features (genes). We are using this following feature selection approaches:

1) Linear Discriminant Analysis is a method for dimensionality reduction processes used in statistics, pattern recognition, machine learning, and bioinformatics to look for linear combinations of features that characterize or separate two or several objects and minimize distances within the same object class. LDA is developed to transform features into lower dimensional space by maximizing the ratio of variance between classes to variance in the class by ensuring maximum class separator [10].

There are three steps in LDA:

- Calculates separators between different classes called between-class variance (S_B) or between-class matrix.
- Calculate the distance between the average and the sample for each class, called within class variance (S_W) or within-class matrix.

Construct low dimensional space by maximizing between-class variance (S_B) and minimizing within class variance (S_W).

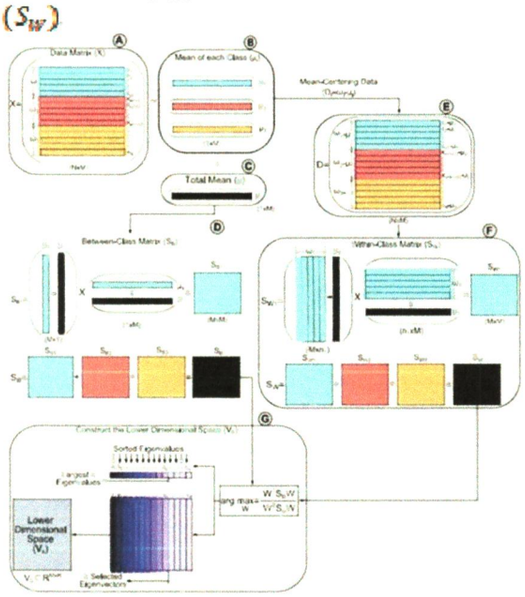

Fig. 4. LDA Visualization

2) Random Forest Recursive Feature Elimination: Random forest algorithm is one method of learning that can solve the problem of classification or regression. The main principle of random forest is combining several decision trees that are built using several bootstrap samples taken from the main sample L and randomly selecting each node which is a subset of the explanatory variable X. This algorithm can provide the dependent variable in a number of class trees that are formed. Then combined with the Recursive Feature Elimination algorithm to select the best features of the microarray data.

3) Support Vector Machine is kind of supervised learning methods, because the training data set is in the form of input vectors that are given targets as output. The purpose of this learning is to build a model that can produce the correct output if given a new input.

III. Results and Discussion

A. Data Description

The experiment using gene expression dataset (GSE 18732) for T2DM from Gene Expression Omnibus (GEO) [11] which downloaded from public data portals. The gene expression dataset consists of mRNA extracted from skeletal muscle of 47 normal samples, 26 glucose intolerant samples and 45 type 2 diabetes samples and only normal samples and type 2 diabetes samples are using in this work.

B. Methods

1) RMA algorithm was using to normalize the expression data

2) Feature Ranking We first calculate weights of features using (1) and select the best 500 features for the next processing stage. We then select the best 400 genes by calculate the information gain for the retained 500 genes (2) and select the best 300 genes for further analysis by calculate the random forest importance of these 400 genes.

3) Feature selection First we run LDA algorithm we are using 10-fold cross validation approach and to find the model with highest accuracy corresponding to the number of features retained then we run the random forest recursive feature elimination method and as the final feature selection procces we run the Support Vector Machine based Random Forest RFE result and to train the SVM model we use the Radial kernel SVM for analysis and follow a bootstrapping approach.

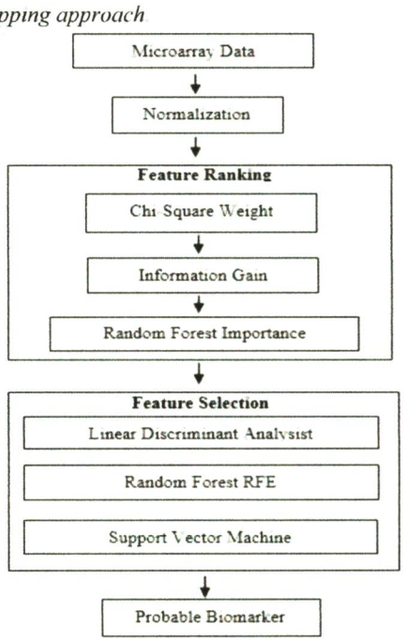

Fig. 5. Proposed Framework

IV. CONCLUSION

This paper has explained how to identify the probable biomarker of type 2 diabetes by using Recursive Feature Extraction. Biomarkers can be used to diagnose a disease and therapeutic target molecules caused certain diseases. Therefore, the result of this research can be used to find the best medical treatment to totally cure the type 2 diabetic. For future works, we will develop a classification model that is better for detecting biomarkers. This model is not only implemented in type 2 diabetes mellitus but can also be used in other diseases.

REFERENCES

[1] Siang, Tan Ching, et al. "A review of cancer classification software for gene expression data." International Journal of Bio-Science and Bio-Technology 7.4 (2015): 89-108. Xindong Zhang, Lin Gao, Zhi-Ping Liu and Luonan Chen. "Identifying module biomarker in type 2 diabetes mellitus by discriminative area of functional activity." 2015.

[2] Zhang, Xindong, et al. "Identifying module biomarker in type 2 diabetes mellitus by discriminative area of functional activity." BMC bioinformatics 16.1 (2015): 92.

[3] Chakraborty, Debasis, and Ujjwal Maulik. "Identifying cancer biomarkers from microarray data using feature selection and semisupervised learning." IEEE journal of translational engineering in health and medicine 2 (2014): 1-11.

[4] Chen, Li, et al. "Identifying cancer biomarkers by network-constrained support vector machines." BMC systems biology 5.1 (2011): 161.

[5] Kolberg, Janice A., et al. "Biomarkers in Type 2 diabetes: improving risk stratification with the PreDx® Diabetes Risk Score." Expert review of molecular diagnostics 11.8 (2011): 775-792.

[6] Mishra, Arpit, et al. "Probable Biomarker Identification Using Recursive Feature Extraction and Network Analysis." Data Mining Workshops (ICDMW), 2017 IEEE International Conference on. IEEE, 2017.

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

[8] Irizarry, Rafael A., et al. "Exploration, normalization, and summaries of high density oligonucleotide array probe level data." Biostatistics 4.2 (2003): 249-264.

[9] SHANNON, Claude. "A Mathematical Theory of Communication-The Bell System Technical Journal, vol. 27, pag. 379-423, 623-656." (1948).

[10] Tharwat, Alaa, et al. "Linear discriminant analysis: A detailed tutorial." AI communications 30.2 (2017): 169-190.

[11] Edgar, Ron, Michael Domrachev, and Alex E. Lash. "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository." Nucleic acids research 30.1 (2002): 207-210.