JURNAL INFOTEL
Informatics - Telecommunication - Electronics

Website Jurnal : http://ejournal.st3telkom.ac.id/index.php/infotel
ISSN : 2085-3688; e-ISSN : 2460-0997

# On the Feature Selection of Microarray Data for Cancer Detection based on Random Forest Classifier

Tita Nurul Nuklianggraita[1*], Adiwijaya[2], Annisa Aditsania[3]
[1,2,3] School of Computing, Telkom University
[1,2,3]Telekomunikasi Street, Terusan Buah Batu, Bandung 40257, West Java, Indonesia
*Corresponding email: titanggraita@gmail.com

Received 27 May 2020, Revised -, Accepted 12 June 2020

Abstract — Cancer is a disease that can affect all organs of humans. Based on data from the World Health Organization (WHO) fact sheet in 2018, cancer deaths have reached 9.6 million. One known way to detect cancer that is with Microarray Technique, but the microarray data have large dimensions due to the number of features that are very much compared to the number of samples. Therefore, dimension reduction should be made to produce optimum accuracy. In this paper, we compare Minimum Redundancy Maximum Relevance (MRMR) and Least Absolute Shrinkage and Selection Operator (LASSO) to reduce the dimension of microarray data. Moreover, by using Random Forest (RF) Classifier, the performance of classification (cancer detection) is compared. Based on the simulation, it can be concluded that LASSO is better than MRMR because it can produce an evaluation of 100% in lung and ovarian cancer, 92% colon cancer, 93% prostate tumor, and 83% central nervous system.

Keywords – Cancer, Microarray, Minimum Redundancy Maximum Relevance (MRMR), Least Absolute Shrinkage and Selection Operator (LASSO), Random Forest



## I. INTRODUCTION

Cancer is a disease that can affect all organs of humans. The formation of abnormal cells that grow beyond the limits quickly and spread to other organs is one hallmark of cancer. Based on data from the World Health Organization (WHO) factsheets in 2018, death is experienced because cancer has already reached 9.6 million people [1]. Based on data from the American Cancer Society in 2018, it provides an estimate of the number of new cancer cases and deaths in 2019. It is estimated that 1.76245 million new cancer cases and 606.880 deaths from cancer [2] the known way to detect cancer by a classification based on microarray data of gene expressions.

Microarray is currently used as an alternative to diagnose or detect cancer. Microarray gene information expresses in cell samples that can be used to examine thousands of genes simultaneously [16][17]. Microarray data is big data or data with high dimensions because microarray data has more features than samples. Therefore, there must be done of dimension reduction process [3][5]. Based on Adiwijaya's research in 2018 [4]. His research proves that the classification of microarray data uses. The Support Vector Machine (SVM) is a classification method with linear kernel and kernel polynomial functions in Lung Cancer. It has proposed a system that can provide an F1-score of 1 in the use of the Minimum Redundancy dimension reduction method Maximum Relevance (MRMR) with the number of features used for classification is 10% of the original number of features. That means that the accuracy obtained from the classification is 100%, and the system performance built is excellent. Ding Research. C and Hanchuan. P [9] proves that the MRMR method can help improve classification performance in good performance accuracy. Meanwhile, In 2018, research by H. Aydadenta and Adiwijaya [3] showed that microarray data classification using the Random Forest classification method and dimension reduction as a selection feature on the Relief Method obtained 85.87% accuracy for Colon Cancer, 98.9% for Lung Cancer, and 89% for prostate tumor.

In 2007, Somnath. D and Susmita. D [6] proved that the LASSO (Least Absolute Shrinkage and Selection Operator) method is more effective than the PLS (Partial Least Squares) method because lung cancer data have a higher percentage. The approach used

between LASSO and PLS is average imputation to predict the survival time of someone who has cancer. In 2008, Adiwijaya et al. [17] conducted research on the Analysis and Implementation of the Minimum-Redundancy-Maximum-Relevance (MRMR) Feature Selection on the Naïve Bayes classification method. The selected selection feature aims to reduce the size of the data but still produce a good accuracy value. It is not too large if there is a decrease in the accuracy of the data classification [18,19].

In 2010, Li et al. BMC Bioinformatics [8] researched on "classification of G-protein coupled receptors based on support vector machines with maximum relevance of minimum redundancy and genetic algorithms." This study uses the Support Vector Machine (SVM) classification method and dimension reduction as a selection feature using the Minimum Redundancy Maximum Relevance (MRMR) method to predict and classify GPCR directly from amino acid sequence data. A Genetic Algorithm (GA) is used to find the optimal feature subset. The accuracy of three-layer predictors: the superfamily, family, and subfamily levels. Those are obtained from the Cross-Validation test on two non-redundant datasets. The yield can be around 0.5% to 16% higher than GPCR-CA and GPCRPred.

This research differs from previous research studies for the research [4] feature selection of Minimum Redundancy Maximum Relevance only improves the accuracy of lung cancer data. Therefore, this research will use MRMR to enhance the accuracy of the other datasets. In a research [6], LASSO has higher than PLS presentation in the lung cancer data. In research [3], Random Forest has not sufficiently increased if with the Relief Method selection feature. Therefore, this research will use the Random Forest classification with MRMR and LASSO selection features to improve accuracy in other datasets.

## II. RESEARCH METHOD

### *A. Microarray Data*

Microarray datasets used in this research were five data, among others are data Central Nervous System, Colon Cancer, Lung Cancer, Prostate Tumor, and Ovarian Cancer. The dataset was obtained from (http://leo.ugr.es/elvira/DBCRepository/).

In Table 1, there are data names, features, notes (amount of data), and classes that have been provided are positive classes and negative classes in microarray data.

### *B. Proposed Scheme*

In this research, there are five steps, according to Fig.1. First, preprocessing. At the stage of preprocessing, data is split and normalized. Second, dimension reduction using feature selection. The selected feature selection method in this research is the Minimum Redundancy Maximum Relevance and the Least Absolute Shrinkage and Selection Operator to get optimal features. Third, the process of finding the best parameters with hyperparameter tuning and cross-validation. Fourth, the classification process using the Random Forest. Finally, evaluate the results of cancer detection on microarray data.

a) Preprocessing

At this stage, the data is processed to be ready for the reduction or even classified. The step that has been done in this process is splitting data into training and testing data. Normalization data is used to change the range of data values with a range from 0 to 1. Data were normalized using the min-max method following Equation (1) [4][7]:

$$\text{normalization} = \frac{(a(\text{i}) - \min(a))}{(\max(a) - \min(a))} \quad (1)$$

Where:

$a(i)$ = the i-th original data
$a$ = the original data
$min$ = for the minimum value in the data
$max$ = for the maximum value in the data

Table 1. Microarray Data

| No. | Name Data | Feature | Record | Positive Class | Negative Class |
|---|---|---|---|---|---|
| 1 | Central Nervous System | 7129 | 60 | Class 1 | Class 0 |
| 2 | Colon Cancer | 2000 | 62 | Positive | Negative |
| 3 | Lung Cancer | 12533 | 181 | Mesothelioma | ADCA |
| 4 | Prostate Tumor | 12600 | 136 | Relapse | Non - relapse |
| 5 | Ovarian Cancer | 15154 | 253 | Cancer | Normal |

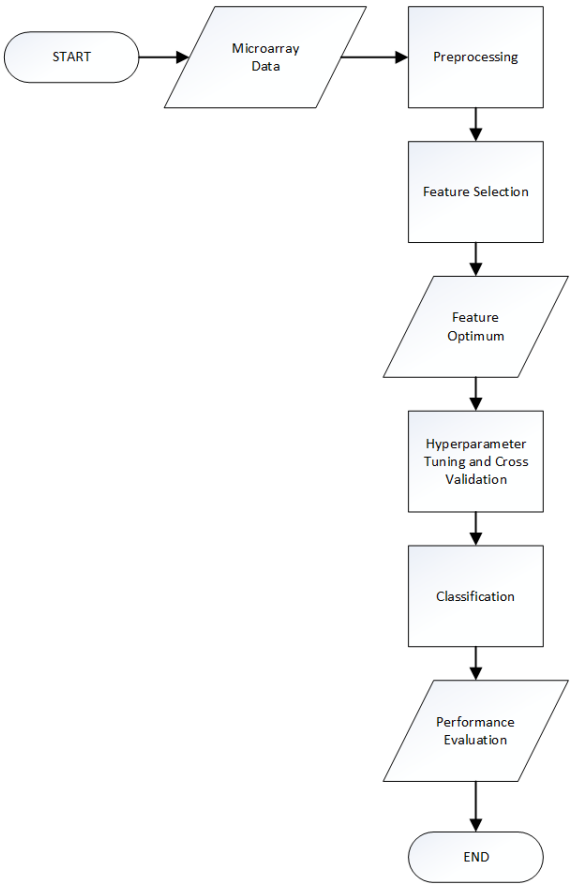


Fig.1. Proposed Scheme

b) Feature Selection

Microarray data has enormous dimensions. The microarray data feature is not comparable to the number of data records. That resulted in suboptimal accuracy of the results of the classification process. Feature selection used is the Minimum Redundancy Maximum Relevance (MRMR) and Least Absolute Shrinkage and Selection Operator (LASSO). Both selection features can select the best features that will be used in the classification process.

- Minimum Redundancy Maximum Relevance (MRMR)

  This method is a method for selecting features that have maximum relevance to the target class and minimum redundancy with other features [8]. The minimum redundancy process is calculated using the correlation function. The calculating process the value of maximum relevance of a feature of the class using the F-test function.

  After that, the calculation process is performed on the F-test Correlation Difference (FCD) function. If the value is greater than redundancy, there is the same information, and that should be done is to remove the feature [4].

  The function used in this research was the F-test Correlation Difference [8][9]:

$$max_{\mathrm{i}\in\Omega\mathrm{S}}\,[F(\mathrm{i},\mathrm{h}) - \frac{1}{|n|}\sum_{i\,\epsilon\,n}|c(\mathrm{i},\mathrm{j})|] \qquad (2)$$

  Where:

  $F(i,h)$ = number of features selected relevance
  $n$ = number of features selected
  $c(i,j)$ = number of features selected redundancy

- Least Absolute Shrinkage and Selection Operator (LASSO)

  Microarray data is data that has many variables. LASSO is a statistical method to overcome dimensional microarray, overfitting the data and multicollinearity. LASSO is one of regression techniques to shrink the independent variable [12]. LASSO build models to feature coefficient set to zero or close to zero, when the feature is not zero that features selected. In 2015 [5], explained that the LASSO can model and linear supervised learning. Function of LASSO [6]:

$$\hat{Y} = \hat{\alpha} + \sum_{j=1}^{p}\hat{\beta}j\,Xj \qquad (3)$$

  Where:
  $\alpha$ = control the LASSO coefficient that is set with restrictions.

  Constraint,

  $\sum_{j=1}^{p}|\widehat{\beta}j| \leq s$, where s is the parameter of depreciation a number of coefficients to zero; thus, can be used for the selection of variables as well.

  β = shrink the LASSO coefficient which correlates with zero or near to zero

c) Hyperparameter Tuning and Cross Validation

Hyperparameter tuning is one method for finding the best parameters on multiple methods of classification. This research used two search techniques for tuning hyperparameter is random search and grid search at random forests classification. Random search is a technique combining random parameters to find the best solution on a model. Grid search is technique combines all the possibilities of which have been determined by hyperparameter to obtain optimal values of these parameters [14]. For grid search, not all random search parameters are used again because the n_estimators parameter is searched again.

Early experiments were performed with random search techniques. There are some parameters in random search is n_estimators (number of trees), max_features (maximum number of features), min_samples_split (minimum number of data points placed in a node before the node is split), min_samples_leaf (minimum number of data points allowed in a leaf node) and bootstrap (method for sampling data points).

Each parameter is given a value as in Table 2. After the search, the best parameter used again for a search grid search technique with the appropriate values of each parameter that it finds more optimal parameters. The classification also using cross validation to re-validate the data to achieve a good accuracy. This research uses a 3-fold cross validation.

Table 2. Hyperparameter Tuning with Values

| Hyperparameter Tuning | Hyperparameter Values |
|---|---|
| n_estimators | start = 200, stop = 1000 |
| max_features | 'auto', 'log2' |
| min_samples_split | 2, 5, 10 |
| min_samples_leaf | 1, 2, 4 |
| bootstrap | 'True', 'False' |

d) Classification Based on Random Forests

Scientists are aware of any changes in the DNA of genes in each specific disease. Therefore, it would be very difficult in the development of tests to detect these changes, particularly in cancer. Microarray is learning to know the extent to which genes turn on or off certain genes in cells and networks [10]. Classification is the process of finding information [12], one of which is on microarray data. One is used for the classification of microarray data is Random Forest. Random Forest is one method of classification which consists of a decision tree. This method is constructed at random as its name and is assisted by a decision tree. This method can help eliminate the correlation between the decision tree [3]. Random Forest is one of the methods that produces very accurate predictions and can handle a very large number of input variables without overfitting [13].

According Aydadenta. H, Adiwijaya (2018) [3] in the referred paper According Breiman (2001) [13] each tree is formed beforehand at random. All features are placed at each node. Once divided equally, calculate the "Best Split" feature based on the features of the training data [13]. When selecting a random value means no need to look at other attributes, as seen just randomly chosen value [20]. Parameters are most needed in this method is the election of the random value and the number of trees that will be built on a random forest. [3].

e) Performance Evaluation

Performance calculation of a model is needed to find out whether the model is correct or not in the classification process. At the time of using a performance calculation accuracy, precision, and recall, it uses the variables listed in the confusion matrix. There are variables True Positive, True Negative, False Positive dan False Negative. Positive labels are categorized as cancerous and negative labels are categorized as normal or not cancerous. Each of these variables in Table 3 have an explanation, namely:

- True Positive (TP) is the value when the predicted data shows positive cancer (affected by cancer) and the actual data shows positive cancer (affected by cancer).
- True Negative (TN) is a value when predictive data shows cancer negative (normal or not cancerous) and actual data shows cancer negative (normal or not cancer).
- False Positive (FP) is the value when the predicted data shows cancer positive (affected by cancer) and the actual data shows negative cancer (normal or not cancer).
- False Negative (FN) is a value when predictive data shows cancer negative (normal or not cancer) and actual data shows cancer positive (affected by cancer).

Table 3. Confusion Matrix

| | Predicted Positive | Predicted Negative |
|---|---|---|
| **Actual Positive** | True Positive (TP) | False Negative (FN) |
| **Actual Negative** | False Positive (FP) | True Negative (TN) |

The formula for accuracy, precision, and recall as follows:

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+FP+TN} \text{ x } 100\% \quad (4)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (5)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (6)$$

## III. RESULT

In this research, we provide three scenarios experiment with five types of cancer data is Colon Cancer, Lung Cancer, Ovarian, Prostate Tumor and Central Nervous System. This research applies the LASSO + Random Forest method, MRMR FCD +

Random Forest method and the Random Forest method without dimension reduction in python applications. Each scenario has been done ± 10 times to achieve the best accuracy. The results of the first experiment to search the number of features the best selection of two-dimensional reduction method below,

Table 4. The Number of Features with Different Alpha

| Data | α | | |
|---|---|---|---|
| | 0.01 | 0.001 | 0.0001 |
| Colon Cancer | 22 | 44 | 50 |
| Lung Cancer | 20 | 95 | 140 |
| Ovarian Cancer | 8 | 64 | 188 |
| Prostate Tumor | 14 | 87 | 109 |
| Central Nervous System | 35 | 46 | 52 |

Table 4 presents the number of features with a range of alpha 0-1 obtained from each dataset with a difference of three alpha value is 0.01, 0.001 and 0.0001. The smaller the alpha value, the greater the features obtained. vice versa.

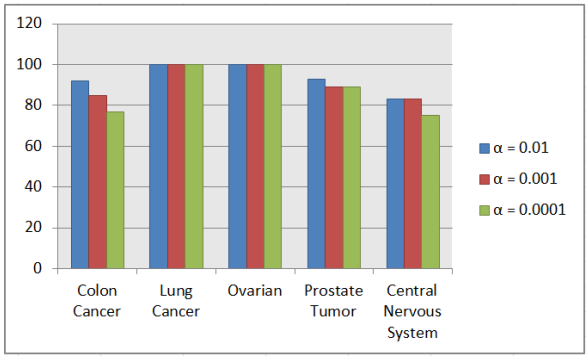


Fig.2. The Graph Looks for The Best Alpha Parameter Values for The Feature Selected in The LASSO Method

Graph Fig.2 presents the accuracy of each data to determine the best alpha will be used on the final outcome of LASSO + Random Forest method. On the stem blue diagram is an alpha with a value of 0.01 i.e by 92% datasets Colon Cancer, 100% Lung Cancer, 100% Ovarian, 93% Prostate Tumor, and 83% Central Nervous System. On the stem red diagram is an alpha with a value of 0.001 i.e by 85% datasets Colon Cancer, 100% Lung Cancer, 100% Ovarian, 89% Prostate Tumor, and 83% Central Nervous System. Last, on the stem green diagram is an alpha with a value of 0.0001 which is the 77% datasets Colon Cancer, 100% Lung Cancer, 100% Ovarian, 89% Prostate Tumor and 75%. Once analyzed, the alpha with a value of 0.01 have an accuracy better than the other then the alpha value alpha 0.01 which will be selected for dimension reduction process on LASSO method.

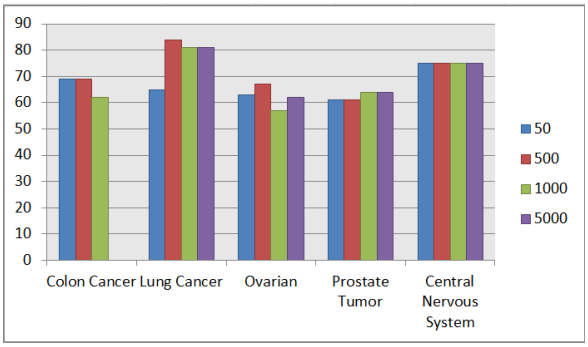


Fig.3. The Graph Looks for The Best Number of Features in The MRMR Method

Graph Fig.3 presents the accuracy of any data to determine the best n-features or the number of features that will be used at the end of MRMR + Random Forest method. This research tried several features of 50, 500, 1000 and 5000 randomly to determine the reduction of data redundation by trying to take the number of features ± to the mid-limit of the actual number of features. In colon cancer only try the number of features 50, 500, and 1000 because the actual number of features is only 2000. Meanwhile, other data has several features over 5000. Seen that colon cancer, lung cancer, ovarian and central nervous has the best accuracy when the number of features = 500. While prostate tumor best accuracy at 1000 features.

Table 5. Before and After of Feature Selection

| Data | The Best Number of Feature Selection | | | |
|---|---|---|---|---|
| | LASSO | | MRMR | |
| | Before | After | Before | After |
| Colon Cancer | 2000 | 22 | 2000 | 500 |
| Lung Cancer | 12533 | 20 | 12533 | 500 |
| Ovarian Cancer | 15154 | 8 | 15154 | 500 |
| Prostate Tumor | 12600 | 14 | 12600 | 1000 |
| Central Nervous System | 7129 | 35 | 7129 | 500 |

Table 5 presents provide data comparison before and after the number of feature selection for LASSO and MRMR method. For LASSO with the best alpha parameters is 0.01 and with the number of features listed in the table. The best number of features with MRMR method for data colon cancer, lung cancer, ovarian cancer, and central nervous system is 500 features, and the prostate tumor is 100 features.

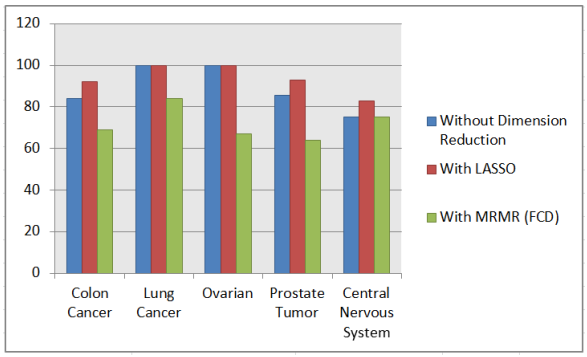


Fig.4. Best Accuracy Graph Results from All Scenarios Experiment With Classifier Random Forests Without Dimension Reduction, With Dimension Reduction (Comparison between LASSO and MRMR)

Graph Fig.4 presents the best accuracy of the three scenarios cancer trial data classification without dimension reduction and dimension reduction method (comparing LASSO and MRMR). In the blue bar shows the classification accuracy of the Random Forest without dimension reduction 84% of Colon Cancer, 100% Lung Cancer, 100% Ovarian, 85.71% Prostate Tumor, and 75% Central Nervous System. The red bar shows the accuracy of the Random Forest classification with dimensional reduction (LASSO) of 92% Colon Cancer, 100% Lung Cancer, 100% Ovarian, 93% Prostate Tumor, and 83% Central Nervous System. The green bar shows the accuracy of the Random Forest classification with dimensional reduction (MRMR) of Colon Cancer 69%, Lung Cancer 84%, Ovarian 67%, Prostate Tumor 64%, and Central Nervous System 75%.

Lung Cancer and Ovarian constant gain 100% accuracy without reducing the dimensions and dimension reduction (LASSO). Colon Cancer, Prostate Tumor and Central Nervous System can increase accuracy when using dimension reduction (LASSO) with an accuracy of 92%, 93% and 83%. While the dimension reduction using MRMR, there is no improvement. Just for Central Nervous System have a constant gain is 75%. It is seen that the best results of the comparison of the accuracy of the dimension reduction is LASSO, because LASSO is able to improve the best accuracy of the five data with the Random Forest classification.

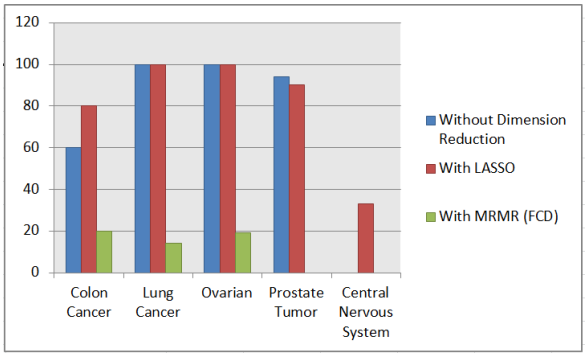


Fig.5. Precision of The Three Scenarios Experiment

Figure 5 presents a graph precision values of the three scenarios experiments. On the stem diagram of blue showed precision without dimension reduction to mean that the right of developing cancer was 60% in the data Colon Cancer, 100% Lung Cancer, 100% Ovarian, 94% Prostate Tumor and 0% Central Nervous System of the overall predicted cancer. On the stem diagram of red indicate the precision with dimension reduction (LASSO) in the sense that the right of developing cancer was 80% in the data Colon Cancer, 100% Lung Cancer, 100% Ovarian, 90% Prostate Tumor and 33% Central Nervous System of the overall predicted cancer. On the stem diagram green color indicates the precision with dimension reduction (MRMR) in the sense that the right of developing cancer was 20% in the data Colon Cancer, 14% Lung Cancer, 19% Ovarian, 0% Prostate Tumor and 0% Central Nervous System of the overall predicted cancer.

For a percentage of 100% precision shows that 100 people who tested positive for cancer from overall positive prediction were large. 94% precision shows that 94 people predicted positive for cancer out of the overall positive prediction were large. 90% precision shows that 90 people predicted positive for cancer out of the overall positive prediction were large. 80% precision shows that 80 people who tested positive for cancer out of all positive predictions were large. 60% precision shows that 60 people predicted positive for cancer out of all those positive predicted. 33% precision shows that 33 people who tested positive for cancer out of all positive predictions were large. 20% precision shows that the 20 people who predicted positive cancers from the overall positive predicted were large. 19% precision shows that 19 people who tested positive for cancer out of all positive predictions were large. 14% precision shows that 14 people predicted positive for cancer than overall positive predicted. 0% precision indicates that no one is positive for cancer or predicted positive for cancer.

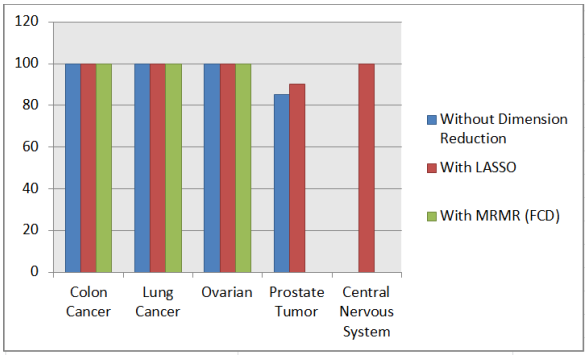


Fig.6. Recall of the three scenarios experiment

Graph Fig.6 presents the recall value of the three scenarios experiments. The blue bar diagram shows recall without dimension reduction, meaning that what is predicted to be cancer is 100% in the Colon Cancer,

100% Lung Cancer, 100% Ovarian, 85% Prostate Tumor and 0% Central Nervous System compared to the whole who actually has cancer. The red bar diagram shows the dimensional reduction (LASSO) meaning that what is predicted to be cancer is 100% in the Colon Cancer, 100% Lung Cancer, 100% Ovarian, 90% Prostate Tumor and 100% Central Nervous System compared to the whole who actually was affected cancer. The green diagram shows the dimensional reduction (MRMR) meaning that what is predicted to be cancer is 100% in the Colon Cancer, 100% Lung Cancer, 100% Ovarian, 0% Prostate Tumor and 0% Central Nervous System compared to the whole who actually was affected cancer.

The 100% recall presentation showed that 100 people were positive for cancer compared to all people who were cancer positive. 90% recall showed that 90 people were positive for cancer compared to all people who were cancer positive. 85% recall showed that 85 people were positive for cancer compared to overall people who were cancer positive. 0% recall shows that no one is predicted to be cancer positive compared to all people who are cancer positive.

The results of the analysis of Fig. 5 and Fig. 6 are for the precision value showing what percentage of a person's cancer from the overall predicted cancer. To remember the value indicates what percentage of a person estimates overall cancer from actual cancer. There are several conditions between precision and recall, condition 1 when precision is lower than recall because the greater the amount of data provided, the possibility of TP and FP values (the amount of irrelevant data) is also greater. Condition 2 when a recall is lower than precision because the greater the amount of data provided, the greater the possibility of TP and FN values and condition 3 when precision and recall have the same value because the amount of relevant data is greater or balanced when compared to irrelevant there is no error.

Table 6. Confusion Matrix for LASSO

| Data | Confusion Matrix for LASSO | | | | |
|---|---|---|---|---|---|
| | TP | FP | FN | TN | Accuracy |
| Colon Cancer | 4 | 1 | 0 | 8 | 92% |
| Lung Cancer | 7 | 0 | 0 | 30 | 100% |
| Ovarian Cancer | 21 | 0 | 0 | 30 | 100% |
| Prostate Tumor | 9 | 1 | 1 | 17 | 93% |
| Central Nervous System | 1 | 2 | 0 | 9 | 83% |

Table 6 and 7 presents confusion matrix as accuracy measurement for LASSO and MRMR method. For example colon cancer data in LASSO, the value TP is 4, which means there are have 4 true positive when the predicted data shows positive cancer (affected by cancer), and the actual data shows positive cancer (affected by cancer).

Table 7. Confusion Matrix for MRMR

| Data | Confusion Matrix for MRMR | | | | |
|---|---|---|---|---|---|
| | TP | FP | FN | TN | Accuracy |
| Colon Cancer | 1 | 4 | 0 | 8 | 69% |
| Lung Cancer | 1 | 6 | 0 | 30 | 84% |
| Ovarian Cancer | 4 | 17 | 0 | 30 | 67% |
| Prostate Tumor | 0 | 10 | 1 | 17 | 64% |
| Central Nervous System | 0 | 3 | 0 | 9 | 75% |

The value FP is 1, which means there are have 1 false positive when the predicted data shows cancer positive (affected by cancer), and the actual data shows negative cancer (normal or not cancer). The value FN is 0, which means there is no false negative when predictive data shows cancer negative (normal or not cancer) and actual data shows cancer positive (affected by cancer). The value TN is 8, which means there are have 8 true negative when predictive data shows cancer negative (normal or not cancerous) and actual data shows cancer negative (normal or not cancer).

## IV. DISCUSSION

Based on the simulation, we gained some knowledge from the results of this research. The first scenario with LASSO + Random Forest can improve accuracy higher than other scenarios. As in Colon Cancer data, if without reduction dimension is 84%, after being reduced by LASSO it becomes 92%. Then Prostate from 85.71% to 93% and Central Nervous System from 75% to 83%.

The second scenario with MRMR FCD + Random Forest is not able to improve accuracy like other scenarios. One example is Prostate Tumor data from 85.71% to 64%. In fact, there is a decrease because MRMR just has a random feature. The third scenario is Random Forest without dimension reduction, all data are almost able to improve accuracy for the better. So, in this research, only the first scenario with LASSO + Random Forest is able to improve accuracy for the better because of the Alpha parameters that are in the LASSO method, and the Random Forest classification also has hyperparameter tuning to help improve better accuracy.

## V. CONCLUSSION

The conclusion of the process and results of the research are as follow. Selection of the number of features of LASSO dimension reduction method is very influential because of their alpha parameter that support to improve the accuracy of the best against each dataset, the smaller the value of alpha hence the lower the accuracy. Likewise, the greater the accuracy of the alpha value is higher, but not very influential

MRMR for parameter *n*-features done at random, the large or small did not specify the value *n*-feature also good accuracy results. Based on the results of the classification random forest method dimension reduction (LASSO) value obtained is greatly improved compared to when no dimension reduction, evidenced in the data Colon Cancer from 84% to 92%, Prostate Tumor from 85.71% to 93% and the Central Nervous System of 75% to 83%. While the data Lung Cancer and Ovarian are equally good.

Based on Random forest classification result, the comparison of two-dimension reduction between LASSO methods and MRMR show that LASSO is able to improve the accuracy of classification. Selection hyperparameter tuning on random forest classification was also influential in finding the optimum value of each parameter to be used in the classification as the number of trees to be built and the number of variables.

## ACKNOWLEDGMENT

The authors would like to thank Telkom University have supported this research and publish the paper.

## REFERENCES

[1] World Health Organization, Cancer Factsheets, 2018.

[2] Rebecca L. Siegel, MPH; Kimberly D. Miller, MPH; Ahmedin Jemal, DVM, PhD; CA Cancer J Clin, American Cancer Society, 69:7-34;2019.

[3] Aydadenta, Husna, and Adiwijaya Adiwijaya. "A Clustering Approach for Feature Selection in Microarray Data Classification Using Random Forest." Journal of Information Processing Systems 14.5, pp. 1167-1175, 2018

[4] Ma'ruf, Firda Aminy, and Untari Novia Wisesty. "Analysis of the influence of Minimum Redundancy Maximum Relevance as dimensionality reduction method on cancer classification based on microarray data using Support Vector Machine classifier." In Journal of Physics: Conference Series, vol. 1192, no. 1, p. 012011. IOP Publishing, 2019

[5] Z. M. Hira and D. F. Gillies, "A review of feature selection and feature extraction methods applied on microarray data", Advances in Bioinformatics, vol. 2015, article ID, 198363, 2015.

[6] Somnath, D and Susmita, D., "Predicting Patient Survival from Microarray Data by Accelerated Failure Time Modeling Using Partial Least Squares and LASSO". Journal of Biometrics, Maret vol 63 No.1, pp.259-271. USA. 2007

[7] Zhu. C, Gao. D, "Influence of Data Preprocessing", Journal of Computing Science and Engineering, vol.10, No.2, pp. 51-57, June, 2016.

[8] Li. Z, Zhou. X, Dai. Z, Zou. X, "Classification of G-protein coupled receptors based on support vector machine with maximum relevance minimum redundancy and genetic algorithm", BMC Bioinformatics, 2010.

[9] Ding. C, Hanchuan Peng, "Minimum Redundancy Feature Selection from Microarray Gene Expression Data". *Journal of bioinformatics and computational biology* 3 (02) pp.185-205. 2005

[10] National Human Genome Research Institute. [Online] https://www.genome.gov/about-genomics/fact-sheets/DNA-Microarray-Technology [Accesed 24 Oktober 2019]

[11] Adiwijaya, Aulia MN, Mubarok MS, and Novia WU and Nhita, F. A, " comparative study of MFCC-KNN and LPC-KNN for hijaiyyah letters pronounciation classification system. Information and Communication Technology (ICoIC7)." In 5th International Conference on pp, pp. 1-5. 2017.

[12] Gerard Biau, "Analysis of a Random Forests Model", Journal of Machine Learning Research 13 (2012) 1063-1095.

[13] Farmani, D.K, Kencana. N, Sukarsa.G, "Perbandingan Analisis Least Absolute Shrinkage and Selection Operator dan Partial Least Squares", e-Jurnal Matematika, Vol. 1, No. 1, Agustus 2012, 75-80.

[14] Breiman. L, "Random Forest", Machine Learning, Kluwer Academic Publishers, Manufactured in The Netherlands, 45, 5–32, 2001.

[15] Muhammad Murtadha ramadhan,. Imas Sukaesih Sitanggang,. Fahrendi Rizky Nasution,. Abdullah Ghifari. "Parameter Tuning in Random Forest Based on Grid Search Method for Gender Classification Based on Voice Frequency". International Conference on Computer, Electronics and Communication Engineering. 2017

[16] Adiwijaya, Wisesty UN, E. Lisnawati, A. Aditsania, and Dana S. Kusumo. "Dimensionality Reduction using Principal Component Analysis for Cancer Detection based on Microarray Data Classification." Journal of Computer Science 14, no. 10, 2018.

[17] M.D. Purbolaksono, K. C. Widiastuti, Adiwijaya, M. S. Mubarok, and F. A. Ma'ruf. "Implementation of mutual information and bayes theorem for classification microarray data." In Journal of Physics: Conference Series, vol. 971, no. 1, p. 012011. IOP Publishing, 2018.

[18] Mabarti, I., Aditsania, A., "Implementation of Minimum Redundancy Maximum Relevance (MRMR) and Genetic Algorithm (GA) for Microarray Data Classification with C4.5 Decision Tree". Journal of Data Science and Its Applications, 3(1), 2020.

[19] Daeli, N.O.F, Adiwijaya. Sentiment analysis on movie reviews using Information gain and K-nearest neighbor. Journal of Data Science and Its Applications, 3(1), 2020

[20] Manuel, Bram, and Dodie Tricahyono. "Classifying electronic word of mouth and competitive position in online game industry." *Journal of Data Science and Its Applications* 1(1) pp. 20-27. 2018.