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

Firda Aminy Ma’ruf¹, Adiwijaya², Untari Novia Wisesty³
¹ School of Computing, Telkom University
² firdamaruf@student.telkomuniversity.ac.id, adiwijaya@telkomuniversity.ac.id, untari.n.wisesty@gmail.com

Abstract. Microarray technology provides a way to monitor thousands of gene expressions at the same time. However, microarray data has a high dimensionality. This high dimensionality will affect the classification performance. In order to solve this issue, this research proposed the use of Minimum Redundancy Maximum Relevance (MRMR) as the dimension reduction method and Support Vector Machine (SVM) as the classifier. Principal Component Analysis (PCA) method was also used as a comparison to MRMR. Tests on lung cancer and ovarian cancer data with MRMR and SVM linear kernel classifier as well as polynomial kernel resulted in an F1-score of 1, with the number of features used for the classification was 20% of the original feature dataset. This means that the accuracy of the classification was 100% and the system built has an excellent performance. As for the colon cancer classification, the F1-score result using MRMR and SVM polynomial kernel classifier was greater than the classification without the dimension reduction method, which is 0.84. It is the same with the classification of leukemia cancer, where the MRMR and SVM polynomial kernel classifier obtained greater result than the result of leukemia classification without dimension reduction method, which the F1-score was 0.9657.

1. Introduction
Cancer is a disease caused by excessive and uncontrolled cell division inside the body. A total of 8.8 million people died from cancer in 2015 [1]. As the matter of fact, death by cancer can be avoided as much as 30%-50% by early identification of cancer type that affects a person, which will lead to an increase in successful treatment. However, it is a challenge for the researchers to diagnose cancer based on their morphological structure because of the very thin morphological differences for different types of cancer.

This difficulty encourages the use of data mining algorithms, specifically the use of gene expression to determine the types of cancer cells. The level of gene expression can indicate the activity of a gene in a body cell based on the number of mRNAs. It is known to contain information about the disease that may be in the gene sample, which may help experts to treat or even prevent the disease. DNA microarray technology has the ability to determine the level of thousands of genes simultaneously in an experiment [2], which pushes the development of cancer classification using gene expression data. Variety of methods have been implemented in the microarray gene classification process. For examples, statistical approaches such as Genetic Algorithm (GA), K-Nearest Neighbour (KNN) and Partial Least Square (PLS), were used to construct a classifier model for microarray gene expression. Fort et al. [3] created a classifier model using PLS to reduce the microarray data dimension and regularizing method of Ridge Penalize Logistic Regression. This model created was called RPLS and it was able to provide a low level of classification error. However, the problem with statistical approach is its inflexible
classification system, therefore if there’s a slight difference between the gene expression of the sample and the gene feature that has been determined, it will make it difficult to classify the sample [4].

Support Vector Machine (SVM) has received a lot of attention in machine learning and bioinformatics, where SVM has become the state of the art in pattern recognition because of its ease of use and the high performance it produces. Furey et al. proved that SVM with kernel function addition yielded 100% accuracy performance in classifying colon cancer data [5]. Furey used the dataset of ovarian cancer, colon cancer, and leukemia. Lisnawati E [6] proved that the classification system she built with SVM method using a linear kernel and Radial Basis Function (RBF) kernel were able to provide better accuracy than SVM without kernel function.

The objective of this research is to build a system which can reduce the microarray data dimension and accurately classify cancer based on microarray data. The microarray data has a large number of genes ($p$), which can reach tens of thousands (large dimension), while its number of samples ($n$) is relatively small. Therefore, it requires feature selection or feature extraction on microarray data to solve this curse of dimensionality and avoid overfitting of the classifier. In feature selection, the important thing to consider is choosing a feature subset that is smaller than the original feature set. This feature subset is essential to improve the classification efficiency and performance, and to reduce noise. Besides reducing the data dimension, the characteristic features also help to identify and observe the particular disease target.

Microarray data gives information about a gene expression, so there is a possibility that two genes or more carry the same information. If, for example, $g$, gene has a high relevancy to its label class is being selected into a feature subset, then another gene which is highly related to $g$ will also be selected. This condition is called feature redundancy. Redundancy will disrupt the classification efficiency and reduce the selected number of diverse features [10]. This is the limitation of some feature selection methods. Therefore, to solve the redundancy issue in microarray data, this research will use Minimum Redundancy Maximum Relevance (MRMR) as the feature selection method to reduce data dimension and select the subset of features. MRMR will maximize the relevancy of features with the class label while it minimizes the redundancy in each class. MRMR is probably the least popular feature selection method to reduce the microarray data dimension. However, the research done by Chris Deng and Hanchuan Peng [10] proved that the classification system built using SVM classifier and MRMR produced a small error rate for continuous cancer data, compared to the baseline classifier without dimension reduction method.

The system that will be built in this research will implement the SVM classifier with kernel functions (linear kernel, Radial Basis Function (RBF) kernel, and polynomial kernel) with Minimum Redundancy Maximum Relevance (MRMR) method as the feature selection and Principal Component Analysis (PCA) method as the feature extraction. PCA was used in this research as a comparison to MRMR, to analyze which dimension reduction method yields the best classification performance when combined with SVM method as the classifier. PCA proves itself as one of the most used feature extraction methods in cancer classification based on microarray data by often yielding good performance results [6][9]. PCA has advantages compared to other feature extraction methods, which are PCA is able to reduce the data dimension without losing much of its information, PCA is easy to implement, and has a relatively fast computing time compared to other feature extraction methods [9]. This method reduces the dimension of the system and at the same time preserves as much of the randomness (variance) in the high-dimensional space as possible [18].

SVM was chosen as the classifier because this research objective is to classify the cancer data into two classes and SVM is the suitable method since it works by finding the optimal hyperplane to separate the data into two classes. Moreover, SVM is suitable for DNA microarrays where the number of samples is much smaller than the number of features [18]. Based on the research by Ding et al [10], the combination of MRMR as the feature selection method and SVM as the classifier, produced the smallest error rate in lung cancer data classification, which was 5%. This was the smallest error rate yielded compared to the other 3 classifiers (Naïve Bayes, Linear Discriminant Analysis, Logistic Regression) when combined with MRMR to classify continuous cancer data. Therefore, this research will use the SVM method as the classifier with additional kernel functions. The kernel function allows SVM to process high dimensional data.
This paper is organized as follows: Section 2 gives the overview about microarray technology. Section 3 presents the design of the proposed classification system. Then, the test and the results will be presented in Section 4. Finally, Section 5 summarizes the paper.

2. Microarray technology overview
Cancer can be diagnosed using gene expression because the cause of cancer is gene abnormality inside the patient's body. Gene expression is a process of converting the information stored in a gene, into instructions for making proteins and non-protein coding RNAs. The constructs of gene expression, such as cDNA and ORFs, can be utilized for over-expression studies or to analyze the effect of a protein on a cellular phenotype.

Patrick O. Brown, Joseph DeRisi, and David Botstein developed DNA microarray in the mid-1990s, making it possible to observe thousands of gene expressions simultaneously. A microarray is a DNA arrangement whose diameter size is less than 250 microns. The microarray technology allows the observation of a cell's status and is able to identify a cell species from its gene expression profile. Microarray technology has the potential to revolutionize cancer diagnosis and therapy [8]. Microarray technology works by measuring the amount of mRNA hybridization in complementary single-stranded DNA (cDNA) strands. Each sample sequence (target) hybridizes to the complementary strand on the array (probe) allowing to confirm the presence of a particular gene. A microarray experiment usually compares gene expression levels in two different physiological conditions, for instance, one can compare the normal breast cell and the tumorous breast cell. Nucleic acids of both samples are labeled with distinct fluorescent dyes, Cy-3 (green) for the normal cell and Cy-5 (red) for the tumor cell. If an area in the chip is red, it is possible that the gene is attacked by a tumor. However, if the area is green, it means that the gene can be declared as normal. If the area is yellow, then the gene may be between normal or tumorous [7].

The microarray gene expression data can be represented in a table form, where every row represents a particular gene, each column represents a sample, and each input from the matrix is the measurement of the expression level of a particular gene in a sample, respectively. The main characteristic of the microarray gene expression data is that it has a very large number of genes ($p$), which can reach tens of thousands (large dimension), whereas the number of samples ($n$) is small, only about tens or hundreds. Therefore, dimension reduction is required to select or extract gene features that are able to differentiate the samples well according to their respective characteristics.

3. The proposed classification system design
The high dimension of microarray data is like a two-edged sword, where this high dimension data is most likely can be separated by a linear line (linearly separable). On the other hand, this high dimension will be a challenge when it comes to analyzing the data. This issue can be solved with the help of dimension reduction method. There are two types of dimension reduction method, namely, feature selection and feature extraction. This research proposed the use of one the feature selection methods which is Minimum Redundancy Maximum Relevance (MRMR) and the feature extraction method of Principal Component Analysis (PCA). The results of these dimension reduction processes are informative genes that will be used in the classification process by Support Vector Machine (SVM) to determine whether or not the data carry a cancerous cell. The reason as to why this research use 2 different dimension reduction methods is to compare which dimension reduction method will reduce the complexity of microarray data and yield the smallest classification error with the highest performance result.

The processes in the classification system built are divided into 4 main processes, which are:
1. Using the min-max normalization method, the data will be normalized first into a range of values between 0 and 1 before the dimension reduction process. This method will be explained in subsection 3.1 in this chapter.
2. The genes resulted from min-max normalization will be used in the dimension reduction process by MRMR and PCA. Subsection 3.2 will explain further about these methods, which covers how it works and the functions for these two methods.
3. Classify the chosen genes resulted from the dimension reduction process into 2 different classes. The general process of SVM will be presented in subsection 3.3 of this chapter.

4. Calculate the performance of the classification process using the F1-score measure. The processes above are illustrated in figure 1:

![Diagram](image)

**Figure 1.** The classification system with dimension reduction methods of PCA and MRMR and the classifier of SVM with a kernel function.

### 3.1 The data used and pre-processing method

The microarray data used for cancer classification in this research were colon cancer data, lung cancer data, ovarian cancer data, and leukemia data which were acquired from Kent Ridge Biomedical Data Repository [11]. The specifications of those data can be examined in table 1.

| Data          | Number of Class | Samples                                      | Features |
|---------------|-----------------|----------------------------------------------|----------|
| Colon Cancer  | 2               | 62 (22 Positives, 40 Negatives)              | 2,000    |
| Lung Cancer   | 2               | 181 (31 Mesothelioma, 150 ADCA)              | 12,533   |
| Ovarian Cancer| 2               | 253 (91 Normal, 162 Cancer)                  | 15,154   |
Leukemia 2 38 (27 ALL, 28 AML) 7,129

Here is an example of the microarray data features and samples, shown in table 2 of colon cancer microarray data. It can be seen that colon cancer data has 2,000 features (attributes) in total, which is larger than the number of its samples.

**Table 2.** The example of colon cancer microarray dataset used in this research.

| Object Data | 1st feature | 2nd feature | 3rd feature | ... | 2,000th feature | Class |
|-------------|-------------|-------------|-------------|-----|----------------|-------|
| 1           | 8,589.416   | 5,468.241   | 4,263.408   | ... | 28.70125       | Positive |
| 2           | 9,164.254   | 6,719.53    | 4,883.449   | ... | 16.77375       | Negative |
| 3           | 3,825.705   | 6,970.361   | 5,369.969   | ... | 15.15625       | Positive |
| 4           | 6,246.449   | 7,823.534   | 5,955.835   | ... | 16.085         | Negative |
| 5           | 3,230.329   | 3,694.45    | 3,400.74    | ... | 31.8125        | Positive |
| ...         | ...         | ...         | ...         | ... | ...            | ...    |
| 62          | 7,472.01    | 3,653.934   | 2,728.216   | ... | 39.63125       | Negative |

The current data has a considerable difference in value, therefore it has to be normalized into the range of value from 0 to 1 so that it can be easily processed. Min-max normalization is used to change the data and the function is as follow:

$$X_{ni} = \left( \frac{(X_i - \min(X))}{(\max(X) - \min(X))} \right)$$  \hspace{1cm} (1)

Where:

- $X_{ni}$ = the $i$-th actual data normalization
- $X_i$ = the $i$-th actual data with the range of the original data
- $X$ = the actual data with the range of the original data

From the min-max normalization process, the value of each feature changed into the range of 0-1. Here is the example of min-max normalization process on colon cancer data presented in table 3:

**Table 3.** The colon cancer data after min-max normalization.

| Object Data | 1st feature | 2nd feature | 3rd feature | ... | 2,000th feature | Class |
|-------------|-------------|-------------|-------------|-----|----------------|-------|
| 1           | 0.515       | 0.4658      | 0.4081      | ... | 0.1891         | 0     |
| 2           | 0.5593      | 0.6085      | 0.4927      | ... | 0.0906         | 1     |
| 3           | 0.1474      | 0.6371      | 0.559       | ... | 0.0772         | 0     |
| 4           | 0.3342      | 0.7344      | 0.6388      | ... | 0.0849         | 1     |
| 5           | 0.1015      | 0.2635      | 0.2905      | ... | 0.2148         | 0     |
| ...         | ...         | ...         | ...         | ... | ...            | ...   |
| 62          | 0.4287      | 0.2589      | 0.1988      | ... | 0.2794         | 1     |
After the min-max normalization step, the data was divided into training data and testing data using K-fold cross-validation method. The cross-validation process was repeated as many as 10. From \( k = 1 \) to 10, a dimension reduction process occurred to determine how many features used in the classification process. The advantage of K-fold cross-validation compared to random sampling is that all the data is used for both training and testing.

3.2 The dimension reduction process
The reason this research used MRMR is that this method has the advantage of choosing the features subset based on the minimum value of redundancy between features and the maximum relevance between features to the target class. F-test based MRMR calculates the relevancy value of \( i \)-th data to its class and the correlation value to determine redundancy between the \( i \)-th data and the other data. As for PCA, it’s known as the feature extraction method that is highly capable to reduce the dimensional complexity of microarray data because it will be extracted based on the eigenvectors and eigenvalues obtained. The steps of MRMR and PCA to reduce the dimensions of the microarray data will be explained in subsections 3.2.1 and 3.2.2, respectively.

3.2.1 MRMR process
The MRMR method that is used in this research is MRMR for a continuous variable because the microarray data used is a continuous microarray data. There are 2 types of MRMR function for a continuous variable, which are MRMR FCD (F-test Correlation Difference) and MRMR FCQ (F-test Correlation Quotient) [10]. Here are the functions:
1. MRMR FCD function: combine F-test with correlation using difference.
   \[
   \max_{i \in \Omega_S} \left| F(i, h) - \frac{1}{|S|} \sum_{j \in S} c(i, j) \right| \tag{2}
   \]
2. MRMR FCQ function: combine F-test with correlation using quotient.
   \[
   \max_{i \in \Omega_S} \left| \frac{F(i, h)}{\frac{1}{|S|} \sum_{j \in S} c(i, j)} \right| \tag{3}
   \]

3.2.2 PCA process
The following 6 simple steps explain the PCA process:
1. Collect \( n \) sample from dimension-\( m \) of data vector \( \vec{x}_1, \ldots, \vec{x}_n \) in \( \mathbb{R}^m \).
2. Reduce the mean of each data dimension with the following function:
   \[
   \vec{\mu} = \frac{1}{n} (\vec{x}_1 + \cdots + \vec{x}_n) \tag{4}
   \]
   Where: \( n \) = number of sample or number of observation data.
3. Form data matrix (\( B \)) and covariance matrix (\( S \))
   \( B \) is an \( m \times n \) sized matrix, with its \( i \)-th column \( \vec{x}_i - \vec{\mu} \). To form the matrix data, use the function below:
   \[
   B = [\vec{x}_1 - \vec{\mu}] \ldots [\vec{x}_n - \vec{\mu}] \tag{5}
   \]
   To form covariance matrix (with the size of \( m \times m \)), use the following function:
   \[
   S = \frac{1}{n-1} BB^T \tag{6}
   \]
   Where: \( n \) = number of sample or number of observation data.
4. Calculate the eigenvectors \( \vec{\mu}_1, \ldots, \vec{\mu}_m \) and the eigenvalues \( \lambda_1, \ldots, \lambda_m \) from \( S \) covariance matrix.
5. Sort the eigenvectors based on eigenvalues with descending order and choose the \( k \) eigenvectors with the biggest eigenvalues to form \( W \) matrix with the dimension of \( m \times k \) (in which each column represents eigenvector).
6. Form new data set
   This new data set is obtained from:
   \[
   Y = V_{row} \ast B \tag{7}
   \]
Where $V_{row}$ is acquired from Feature Vector$^T$ and $Y$ is the final data set.

To determine how many eigenvectors are chosen, this research used the cumulative proportion of variance (eigenvalues) to the total variance (eigenvalues). The percentage of variance (POV) shows the magnitude of the information percentage of the original variable contained in each feature vector (eigenvector) based on eigenvalues and provides an interpretation of how much data can be represented in the reduced dimension [12]. The POV for each major component (eigenvector) can be calculated using the following formula:

$$POV = \left( \frac{\lambda_i}{\sum \lambda_i} \times 100\% \right) + \left( \frac{\lambda_{i-1}}{\sum \lambda_i} \times 100\% \right)$$  \hspace{1cm} (8)

Where $\lambda_i$ = the eigenvalue and $\lambda_{i-1}$ = the previous eigenvalue.

The process of selecting eigenvectors based on the cumulative variance proportion is done after POV calculation by first determining the POV threshold (the selection criterion of eigenvectors or principal component).

Selection of eigenvectors $= \frac{\lambda_i}{\sum \lambda_i} \times 100\% >$ threshold  \hspace{1cm} (9)

The percentage of the threshold value is very influential on the results of the eigenvectors selection that will be used during data transformation in the PCA process. This transformed data will then be used in the classification process. The threshold literature on the cumulative proportion of variance on the choice of eigenvectors is a minimum threshold of 80% [13]. Research [6] also proved that a POV threshold of 80% for PCA generated the same accuracy number, even higher than PCA with a 95% POV threshold, for the classification of cancer data using the SVM classifier.

### 3.3 The classification process and performance calculation

The last step of this research is the cancer classification process and its performance calculation. The SVM classifier will determine the class of the microarray data, whether it is class 0 or class 1. SVM works by finding the most optimal hyperplane in separating classes in input space. Usually, SVM is used to classify linearly separable data. However, the problem that is mostly found in real life is non-linear. To overcome this issue, kernel function is used in data training which functions to map each data in the input space into a new vector space with a higher dimension. The kernel changes the dot product of $\vec{x}$ in each equation by adding the functions below. So, the kernel will transform from input space into feature space. Here are the kernel functions that will be used in this model:

1. Linear: $K(\vec{x}_i, \vec{x}_j) = \phi(\vec{x}_i) \phi(\vec{x}_j)$  \hspace{1cm} (10)

2. Radial Basis Function (Gaussian): $K(\vec{x}_i, \vec{x}_j) = \exp \left( -\frac{||\vec{x}_i-\vec{x}_j||^2}{2\sigma^2} \right)$  \hspace{1cm} (11)

3. Polynomial: $K(\vec{x}_i, \vec{x}_j) = (\vec{x}_i \cdot \vec{x}_j + 1)^P$, where $P$ is the degree of the polynomial  \hspace{1cm} (12)

To classify the testing data using the model that is acquired from the SVM classification using the training data, the steps taken in the training process are:

1. Represent the dataset

The dataset consists of input and target ($X_i$ dan $Y_i$)

$$D = \{(X_i, Y_i)| X_i \in \mathbb{R}^p, Y_i \in \{-1,1\}\}_{i=1}^{n}$$  \hspace{1cm} (13)

2. Choose two hyperplanes that separate the data without any data itself between the two. The selected hyperplanes must meet the following limitations:

   a. The data fulfill the equation below:

   The $x_i$ data that is included in the class of 1 (positive sample) is defined by:

   $$\vec{w} \cdot \vec{x}_i + b \geq +1$$  \hspace{1cm} (14)
The $x_i$ data that is included in the class of 0 (negative sample) is defined by:

$$\vec{w} \cdot \vec{x}_i + b \leq -1$$  \hfill (15)

b. The 2 selected hyperplanes must fulfill the following equation:

$$y_i(\vec{w} \cdot \vec{x}_i + b) \geq 1 \text{ for all } 1 \leq i \leq i$$  \hfill (16)

3. Maximize the distance between the two hyperplanes.

The distance between the two hyperplanes can be maximized by using the following function:

$$m = \frac{2}{\|w\|}$$  \hfill (17)

By minimizing the value of $\|w\|$, the maximum margin value will be obtained. After the value of $w$, $b$, and the hyperplane are obtained from the training data, then this model is used to classify test data. The classification performance was assessed by F1-score measurement.

4 Analysis of the results

The experiment scenario in this research is divided into five scenarios, which are:

1. Cancer data classification without dimension reduction method using SVM kernel (linear, RBF, polynomial) classifier.
2. Cancer data classification using MRMR FCD and SVM kernel (linear, RBF, polynomial) classifier.
3. Cancer data classification using MRMR FCQ and SVM kernel (linear, RBF, polynomial) classifier.
4. Cancer data classification using PCA and SVM kernel (linear, RBF, polynomial) classifier.
5. Cancer data classification using the combination of MRMR + PCA dimension reduction and SVM kernel (linear, RBF, polynomial) classifier.

The main observations from those scenarios above are the effect of dimension reduction on classifier performance, the number of features selected by MRMR, the number of features extracted by PCA, and the effect of using different kernels in SVM classifier.

4.1 Analysis of the effect of the dimension reduction process

The objective of this test is to analyze which dimension reduction method, feature selection or feature extraction, that yields better performance for the classifier. The analysis of the dimension reduction method effect on cancer data classification will be divided into four subsubsections, according to the four different types of cancer data this research used, which are observations on colon cancer data, lung cancer data, ovarian cancer data, and leukemia data.

The first testing scenario directly classified the cancer data using the SVM classifier with three different kernels, which are linear, RBF, and polynomial. To divide the test data and training data, K-fold cross-validation method was used with $K = 10$. Therefore, the results of accuracy, precision, recall, and F1-score were the average of 10 different results obtained from different test data and training data.

As for the second, third, fourth, and fifth testing scenario, it only used the linear kernel for the classifier. The number of features selected from MRMR process was determined manually, where the features selected were 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% of all the features that had been sorted in descending order based on its MRMR FCD or MRMR FCQ values. Then the selected features were compared, to decide how many numbers of features that provided the best classification performance and which method provided better performance. The percentage number of features that generally provided good performance results in classifying the four cancers data, was used in the fifth test scenario.

For the fifth test scenario, the POV threshold used was >80% based on the research of [6][13].

The fifth test scenario combined MRMR and PCA to select the features and then extract it. This last test scenario was divided into two, which were:

1. Classification of cancer data by combining MRMR FCD + PCA dimension reduction and classifier SVM kernel linear.
2. Classification of cancer data by combining MRMR FCQ + PCA dimension reduction and classifier SVM kernel linear.

Here is the analysis of dimension reduction method on cancer classification, divided into four subsubsections.

4.1.1 Analysis of dimension reduction process on colon cancer classification

The first testing scenario for colon cancer classification was to directly classify the data using SVM kernel (linear, RBF, polynomial) classifier. The highest F1-score acquired was 0.72 or the accuracy of 82.1429%, which is lower compared to the results acquired from the classification using the MRMR dimension reduction method. Figure 2 presents the performance results from classification using MRMR FCD and MRMR FCQ with SVM linear kernel classifier.

The F1-score obtained from the classification with MRMR FCQ using 50% features, is the highest F1-score of all classification using MRMR dimension reduction, which was 0.8167. In addition, the average computation time of this testing scenario was also faster than the MRMR FCD test scenario, which was 2.9 seconds. From the chart above, it can be inferred that by selecting features as much 50%, 60%, or 70% in MRMR FCD and MRMR FCQ generated good classification performances and high F1-scores for colon cancer data. When the feature percentage was 50%, classification of colon cancer data with MRMR FCQ dimension reduction and SVM linear kernel classifier generated the highest F1-score. Moreover, when the feature percentage was 60%, the MRMR FCD method was able to select features that generated the highest classification performance compared to when using other percentages of features.

For classification with PCA dimension reduction method, the range of POV threshold values that can be used was minimum POV threshold of 17.9771% with the number of eigenvectors (main components) selected as many as 1. The maximum threshold was 100% with the number of eigenvectors chosen was 55. The highest F1-score obtained from this testing scenario was 0.8233, which was acquired with the POV threshold of >70% and the feature used was 11 features.

With this small number of features used, the classification system consisted of dimension reduction process by PCA and SVM linear kernel classifier was able to provide better performance result than MRMR and SVM linear kernel classifier. However, the average computational time classification of colon cancer data with PCA dimension reduction and SVM linear kernel classifier was longer than the 2 previous test scenarios, which was 11.6 seconds.

Figure 3 is a chart of colon cancer data classification results from the fifth test scenario.
The highest F1-score obtained from the testing scenario was 0.82, but not as high as the F1-score obtained from the classification with PCA. So, it can be concluded that:

1. Classification of colon cancer data by dimension reduction method PCA was able to provide better classification performance compared to using MRMR dimension reduction and MRMR + PCA combined reduction method.
2. From the results of the 2, 3 and 5 test scenarios, it can be inferred that the difference in the number of the selected features of colon cancer by MRMR did not necessarily improve the performance of data classification.

4.1.2 Analysis of dimension reduction process on lung cancer classification

The classification of lung cancer data using the SVM classifier without dimension reduction method produced great results in general. Among the three kernel functions used, the linear kernel and polynomial kernel generated the best results, with F1-scores equal to 1. Lung cancer data classification using MRMR FCD and SVM linear kernel classifier, on average generated F1-score of 1, which means that the model built can classify all testing data perfectly. In addition, the average computation time for this test scenario was 37.1 seconds. The test scenario with the MRMR FCQ dimension reduction method provided the same good performance as lung cancer data classification using MRMR FCD, with F1-scores of 1. Hence for lung cancer data, even though the features were reduced by half using MRMR FCD or MRMR FCQ method, the classification performance results were still as good as the direct classification without the dimension reduction.

From the POV calculation carried out by PCA on lung cancer data, it was concluded that the minimum threshold was 18.6498% with the number of eigenvectors (main components) selected was 1 and the maximum threshold was 100% with the number of the eigenvectors selected was 28. Classification with dimension reduction method of PCA generated an F1-score of 1 when the features used were based on the POV threshold of >30% to >90%. The drawback of this test scenario was that the average computation time is longer, which was 2,364.5 seconds.

The results of the F1-score acquired from the previous testing scenarios showed that lung cancer classification using the dimension reduction method of MRMR or PCA was able to provide performance as good as the classification without dimension reduction process. However, the classification with the PCA dimension reduction method only produced F1-score of 1 when the features extracted were based on the POV threshold >30%.
The classification results from the last test scenario also obtained F1-scores = 1, even though the percentage number the features selected by MRMR were different (50%, 60%, or 70%). From the results of the tests above, it can be concluded that:

1. The initial lung cancer data has features that have high relevance to the class and low redundancy between features. Therefore, when MRMR calculated its relevance and redundancy value, not many informative features removed, resulted in a good classification performance.

2. Classification of lung cancer data using the dimension reduction method of PCA should extract features based on the POV threshold of >30% because when the feature extracted was based on POV >10% and >20%, there were only 1 and 2 features used, which were too small. If the number of features used is too small, the data training process will be affected, in the sense that the classifier model does not get enough features to "learn" its characteristics to determine the class of the sample in that feature.

4. Classification of lung cancer data with dimension reduction method produced results as good as classification without dimension reduction.

4.1.3 Analysis of dimension reduction process on ovarian cancer classification

The highest F1-score from ovarian cancer classification without dimension reduction method was obtained from the use of SVM linear kernel classifier, which was 1. The results of the second test scenario and the third test scenario are presented in figure 4.

The highest F1-score obtained from the classification with MRMR FCD and SVM linear kernel classifier was 1, that was when the number of features used was 90% and the second largest F1-score was 0.9947, which was obtained when classification used features of 60%, 70%, and 80%. The average computation time was 55.32 seconds. With the MRMR FCQ method, ovarian cancer data classification resulted in a better performance compared to MRMR FCD and also faster computation time at 54.6 seconds. Looking at Figure 4, the highest F1-score of this test is also 1. However, it is different from the MRMR FCD test scenario where the highest F1-score was obtained when the number of features used was 90%. MRMR FCQ and SVM linear kernel classifier generated F1-score = 1 when it used features as many as 20% to 90% of the total features. This means that the reduction method of MRMR FCQ is better than MRMR FCD methods in terms of selecting ovarian cancer features that have high relevance and low redundancies between features.

Figure 4. Charts of the comparison of F1-Score results from the classification of ovarian cancer data using MRMR FCD and MRMR FCQ.
The results of ovarian cancer classification with PCA dimension reduction method indicated that the selection of eigenvectors based on POV threshold >80% yielded the highest F1-score of 0.9577 compared to the POV threshold <90%. The F1-score results acquired from this test scenario were smaller than the results of classification using MRMR FCD or MRMR FCQ. Besides that, the average computation time of this test scenario was also slower, which was 4,092.23 seconds. Classification with eigenvector selection based on POV threshold 10%, 20%, 30%, and 40% obtained F1-score = 0. This was because the POV minimum threshold for ovarian cancer data was 40.1206%, which means when choosing features based of a threshold < 40.1206% there was no eigenvalues and eigenvectors selected as the principal components (PC).

Therefore, the final test scenario used features as many as 50%, 60%, and 70% selected by MRMR, because based on the second and the third test, these percentages of threshold provided F1-scores that was close to 1 but with fewer features selected compared to POV 90%. The POV threshold used to select the features was >80% because, from the fourth test results, classification using features based on that threshold produced the highest F1-score compared to other thresholds. Figure 5 is a chart of the performance results of the ovarian cancer data classification with the last test scenario.

![Figure 5. Charts of F1-score from the results of ovarian cancer data classification using a combination of MRMR FCD + PCA dimension reduction and MRMR FCQ + PCA.](image)

The highest F1-score from this classification was 0.99, which was obtained from the use of the MRMR FCQ + PCA dimension reduction method with 50% of features used. The conclusions that can be drawn from the classification of ovarian cancer data using these different scenarios are:

1. The dimensions reduction method of MRMR FCQ functioned better in selecting features of ovarian cancer which provided a high classification performance compared to MRMR FCD dimensions. The division of the relevance value to the redundancy value was able to choose features with little redundancy in ovarian cancer data. This is evidenced by the F1-score = 1, which was obtained from the classification system of MRMR FCQ and SVM linear kernel classifier with the percentage of features used as many as 20% - 90% of the total 15,154 features.

2. Increasing the selected number of ovarian cancer features by MRMR for classification did not have an effect on improving classification performance.

3. The classification performance will not necessarily improve when the POV threshold value is larger.

4.1.4 Analysis of dimension reduction process on leukemia cancer classification

The first test scenario of leukemia classification generated different results than the three previous cancer data classifications. The highest F1-score from leukemia cancer classification was obtained from the use of SVM polynomial kernel classifier, which was 0.9514. The next testing scenarios still used SVM linear
kernel classifier because the linear kernel is the most basic kernel function for finding hyperplanes. Figure 6 presents the results from the second test scenario and the third test scenario.

![Figure 6](image)

Figure 6. Charts of the comparison of F1-Score results from the classification of leukemia data using MRMR FCD and MRMR FCQ.

The second test scenario yielded the highest F1-score which was 0.9514, with the number of features used was 90% and the average computation time of 14.4 seconds. This score is greater than the F1-score obtained from the first test scenario and the third test scenario. With the MRMR FCQ dimension reduction, the highest F1-score obtained was 0.9429 and longer computation time at 14.76 seconds.

The results of POV calculation on leukemia cancer data indicated that the range of threshold values that can be used for the eigenvectors selection with PCA was at the minimum threshold of 17.0309% with 1 eigenvector selected and the maximum threshold was 100% with 34 eigenvectors selected. From the fourth test scenario, the selection of eigenvectors based on the POV threshold of >30% produced the best performance result, where the F1-score obtained was 0.951. However, this is smaller compared to the F1-score result from the second test scenario. Moreover, the average

![Figure 7](image)

Figure 7. Charts of F1-score from the results of leukemia data classification using a combination of MRMR FCD + PCA dimension reduction and MRMR FCQ + PCA.
computational time of the fourth test scenario was longer, which was 630.83 seconds. Figure 7 is a chart of the performance results of the leukemia cancer data classification with the fifth test scenario.

For the last test scenario, the percentages of features selected by MRMR were 50%, 60%, 70%. The reason is that the results from the second and third test scenario showed that with these small number of features selected compared to POV threshold >90%, it could generated F1-score that was close to 1.

The highest F1-score obtained from the last test scenario, which was 0.91, is not as high as the highest F1-score from the two previous tests. The conclusions that can be drawn from the results of the tests above are:

1. For leukemia cancer data, the reduction method of MRMR FCD provided better performance than the reduction method of FCQ and PCA MRMR dimensions. The highest F1-score obtained from the MRMR FCD classification with SVM linear kernel classifier, which was 0.9514, was higher than leukemia cancer data classification without dimension reduction method with SVM linear kernel classifier, which was 0.9257.

2. The classification of leukemia cancer data using a combined reduction method of MRMR and PCA did not provide a better F1-score compared to classification with only one method of dimension reduction. This is because there is a possibility that some discriminative and informative features were lost when selected and extracted afterward.

3. Increasing the POV threshold to select features did not improve the performance of leukemia cancer data classification.

4.2 Analysis of the effect of kernel usage on Support Vector Machine classifier

To analyze the effect of different kernel functions usage on the SVM classifier to the results of classification performance, this research used three different functions of the kernel. Those kernel functions were linear, RBF (Gaussian) and polynomial. There were 4 test scenarios implemented to see whether the SVM kernel classifier works best with the help of a dimension reduction method or not. These scenarios were:

1. Classification with SVM kernel (linear, RBF, polynomial) classifier without dimension reduction process.
2. Classification with MRMR FCD dimension reduction and SVM kernel (linear, RBF, polynomial) classifier.
3. Classification with MRMR FCQ dimension reduction and SVM kernel (linear, RBF, polynomial) classifier.
4. Classification with PCA dimension reduction and SVM kernel (linear, RBF, polynomial) classifier.

From the results of the previous test scenarios, it was concluded that the selection of features as many as 50%, 60%, and 70% by MRMR resulted in a better classifications performance. Therefore, the test scenario using MRMR dimension reduction used features that were selected as many as 50%, 60%, and 70%. As for the PCA test scenario, the selection of eigenvectors was based on the POV threshold value of >80%, because the results from [6][13] stated that it is better to select eigenvectors based on the POV threshold >80%.

The final results of all cancer data classifications using the SVM classifier with 3 different kernel functions can be seen in figure 8.
Classification of colon cancer data, lung cancer data, and ovarian cancer data using SVM linear kernel classifier with and without dimension reduction method (MRMR FCD, MRMR FCQ, and PCA) on average generated F1-scores that were greater than the F1-score results from the classification with the two other kernels. This is because the data from those three cancers can still be easily separated by a linear hyperplane despite its large dimensions. Their initial data were well-separated into its own class.

SVM polynomial kernel classifier outperformed SVM linear kernel classifier and SVM RBF kernel classifier on the performance of leukemia classification, where the highest F1-score was 0.9657 or the accuracy of 95%, which was obtained from the classification using MRMR FCQ dimension reduction method. This result proves that after the dimension reduction process, the data variance of leukemia data changed and required the use of polynomial kernel in determining the optimal non-linear hyperplane that could properly classify data with a high dimension such as leukemia data.

5. Conclusion and future work
Based on the results of several test scenarios that had been done in this research, the effect of dimension reduction process and different kernel functions usage on SVM classifier performance can be concluded. The number of features selected by MRMR method, the number of eigenvectors selected according to POV threshold by the PCA method, and the kernel functions in SVM, have a significant effect on the performance result of the classification and the computational time in each process of cancer classification. The performance of classification will not always increase when the number of features selected by MRMR is increased. This is proven by the results of colon, ovarian, and leukemia classification with MRMR dimension reduction and SVM linear kernel classifier, where the F1-score
results fluctuate when more features were added. It also applies for the classification with the dimension reduction method of PCA, where the greater the POV threshold value used in selecting features will not always improve the classification results.

Moreover, the results from this research able to prove that the performance of the cancer classification with dimension reduction method of MRMR is as good or even better than the performance of classification without the dimension reduction method. One proof is the results of colon cancer classification, where MRMR FCQ dimension reduction and SVM linear kernel classifier yielded F1-score of 0.8167 using 50% features selected from the total of 2000 features. This is greater than the F1-score result of the classification without dimension reduction process, which was only 0.72. This proves that although the selected features are smaller compared to the original features set, they have better generalization properties and effectively cover the whole characteristics of the class. The classification for continuous microarray data by MRMR FCQ dimension reduction and SVM kernel (linear, RBF, polynomial) classifier generated a better performance than the classification by MRMR FCD dimension reduction and SVM kernel classifier. This is proven by the chart in Figure 8. The combination of relevance value division to redundancy value is able to select features with the lowest redundancy and the highest relevance.

Comparing the use of feature selection and feature extraction method in cancer classification, it can be inferred that both of them produced equally good performance, even better than the performance of the classification without dimension reduction method. This research showed that the use of MRMR dimension reduction method was able to produce the highest F1-score in the classification of lung cancer, ovarian cancer, and leukemia. Whereas the PCA and SVM classifier produced the highest F1-score in the classification of colon cancer with 11 features used, which was 0.8233. In conclusion, the type of data must be considered when choosing which dimension reduction method will be applied in the classification system to achieve good performance. Each method has its own advantages and disadvantages. For example, in this research, PCA generally extracted features into smaller number compared to features selected by MRMR. However, MRMR computation time was a lot faster.

From the observation on the effect of different kernels usage on the performance of the classification, it can be concluded that the use of SVM linear kernel classifiers provided the best classification results on colon cancer data, lung cancer data, and ovarian cancer data. On the other hand, the F1-score result of leukemia classification was the highest by using the SVM polynomial kernel classification. The polynomial kernel with the 3rd order projected the leukemia data into a higher vector space and determined the optimal non-linear hyperplane that classified the leukemia data properly.

Some suggestions that can be applied for further research is by adding the Sigmoid kernel function to the SVM classifier, comparing the calculation results of the redundancy value as well as the relevance value of the feature set before and after the dimension reduction process by MRMR, and test all of the Kent Ridge Biomedical dataset using the classification model built in this research.

REFERENCES
[1] World Health Organization, Cancer Fact Sheet http://www.who.int/mediacentre/factsheets/fs297/en/ accessed on the date of February 11th 2017.
[2] Zhang LJ, Li ZJ, Hu X 2007 A hybrid gene select. method for cancer classification Austria VLDB Workshop on Data Mining in Bioinformatics (Vienna)
[3] Fort G, Lambert-Lacroix S 2004 Classification using partial least squares with penalized logistic regression Bioinformatics 21(7) pp 1104-1111
[4] Vanitha CD, Devaraj D, Venkatesulu M 2015 Gene expression data classification using support vector machine and mutual information-based gene selection Procedia Computer Science 47 pp 13-21
[5] Furey TS, Cristianini N, Duffy N, Bednarski DW, Schummer M, Haussler D 2000 Support vector machine classification and validation of cancer tissue samples using microarray expression data Bioinformatics 16(10) pp.906-914

[6] E Lisnawati, Adiwijaya, D T Murdiansyah, Untari N Wisesty Deteksi Kanker Berdasarkan Klasifikasi Microarray Data Menggunakan Principal Component Analysis dan Support Vector Machine

[7] Singh RK, Sivabalakrishnan M 2015 Feature selection of gene expression data for cancer classification: a review Procedia Computer Science 50 pp 52-7

[8] Schmidt U and Begley CG 2003 Cancer diagnosis and microarrays The international journal of biochemistry & cell biology 35(2) pp119-124

[9] NCBI Principal components analysis to summarize microarray experiments: application to sporulation time series https://www.ncbi.nlm.nih.gov/pubmed/10902193 accessed on the date of April 13th 2017

[10] Ding C, Peng H 2005 Minimum redundancy feature selection from microarray gene expression data Journal of bioinformatics and computational biology 3 02 pp 185-205

[11] Kent Ridge Biomedical Data Set Repository http://leo.ugr.es/elvira/DBCRepository/ accessed on the date of April 10th 2017

[12] M Mariana 2013 Analisis Komponen Utama. MATEMATIKA DAN PEMBELAJARAN 1 2 pp 189-204

[13] Nelson Spencer Machine Learning: Supervised Learning pt. 2 http://www.nelsonspencer.com/blog/2015/2/15/machine-learning-supervised-learning-pt-2 accessed on the date of February 11th 2017

[14] Nurfalah A, Adiwijaya, Suryani AA 2016 Cancer Detection Based On Microarray Data Classification Using PCA And Modified Back Propagation Far East Journal of Electronics and Communications 16(2) p 269

[15] Marianah L, Nhita F, Adiwijya A 2015 Klasterisasi Data Microarray Menggunakan Metode Clique Partitioning Indonesian Symposium on Computing 2014

[16] H Aydadenta, Adiwijaya 2018 A Clustering Approach for Feature Selection in Microarray Data Classification using Random Forest Journal of Information Processing System 14(5)

[17] Adiwijaya, U N Wisesty, E Lisnawati, A Aditsania, D S Kusumo 2018 Dimensionality Reduction using Principal Component Analysis for Cancer Detection based on Microarray Data Classification Journal of Computer Science 14(10)

[18] Varghese N, Verghese V, Gayathri P, Jaisankar N 2012 A survey of dimensionality reduction and classification methods International Journal of Computer Science and Engineering Survey 3(3) p 45.