ICIC Express Letters
Volume 15, Number 5, May 2021

ICIC International ©2021 ISSN 1881-803X
pp. 429–437

MICROARRAY DATA CLASSIFICATION USING MINIMUM REDUNDANCY MAXIMUM RELEVANCE AND MODIFIED LOGISTIC REGRESSION FOR HIGH ACCURACY CANCER DETECTION

Azka Khoirunnisa and Adiwijaya

School of Computing
Telkom University
Jl. Telekomunikasi No. 1, Bandung 40257, Indonesia
{ azkhoi; adiwijaya }@telkomuniversity.ac.id

Received September 2020; accepted December 2020

Abstract. *Microarray data technology is one of the popular techniques for cancer detection, where thousands of gene expressions can be read at the same time. However, one of the well-known constraints specifically related to microarray data is the large number of genes that cause high dimensional data in comparison with the small number of available experiments or cases. This will give lower classification accuracy. Some popular approaches have been introduced, and one of them is a combination of Principal Component Analysis (PCA) and Logistic Regression to overcome the issues. However, the accuracy is yet to be improved. This paper presents a new technique which is a combination of minimum Redundancy Maximum Relevance (mRMR) and Modified Logistic Regression for data reduction and data classifier, respectively. The result shows an improvement of accuracy, with the average accuracy up to 93.33%.*

Keywords: Microarray data, Cancer, Classification, mRMR, Logistic Regression

1. **Introduction.** Cancer is known as the second largest cause of death worldwide. This deadly disease could attack everyone, either children or adults. Based on the data of IARC in 2018, the number of cancer sufferers has raised to 18.1 million sufferers with 9.6 million mortalities where it is around 1 in 6 deaths is caused by cancer [1]. One way to eradicate the cancer is to do early detection, so that the patient could get the right treatment and subsequently reduce the mortality rate due to cancer disease. There are some classical methods to detect cancer disease, such as using USG (Ultrasonography), blood test, pap smear, endoscope and gene expression [2]. However, with new technology available, researchers are now attempting to use DNA microarray where it can be used to detect cancer via the analysis of gene expressions attacked by cancer.

A microarray holds hundreds to thousands of genes in the form of DNA where a DNA microarray could be used as a cancer detector. This analysis of gene expression uses 2 types of DNA: the normal one and the attacked one. These DNAs are isolated and given a different color fluorescence. Furthermore, each DNA will also emit different colors based on the color it is assigned [3].

DNA microarrays are microscope slides that are printed with thousands of tiny spots in defined positions, with each spot containing a known DNA sequence or gene. Often, these slides are referred to as gene chips or DNA chips. The DNA molecules attached to each slide act as probes to detect gene expression, which is also known as the transcriptome or the set of messenger RNA (mRNA) transcripts expressed by a group of genes.

Every microarray experiment generates a large amount of data and requires certain computational techniques to interpret expression profiles. Available techniques/algorithms

DOI: 10.24507/icicel.15.05.429

reduce the number of genes to a non-redundant information set and subdivide a panel of cancer types into groups that share common features, for example, expression of known cancer markers or the response to a specific therapy. The characteristic profiles of the groups can then be used to identify unknown specimens [4].

Most techniques to analyze cancer expression profiles can be divided into "supervised" classifications or "unsupervised" clustering methods. Supervised methods require additional information regarding the genes whose expression is being examined and which was obtained independently of the microarray experiment, e.g., gene function or tissue origin. This information is then used to find patterns that classify the samples into the given categories. Unsupervised clustering does not require additional information, and is aimed at the discovery of novel, unbiased patterns in the data [5]. However, the dimension is still large that affects the performance and computational time of the system.

Therefore, dimensional reduction process must be conducted prior to the classification process. This process will reduce computational time and subsequently increase the performance of the system [6]. There are several methods that can be used to perform dimensional reduction. One of the commonly used is based on Principal Component Analysis (PCA) [7]. The combination of PCA and Logistic Regression has been used to improve the accuracy of classification [8]. However, the accuracy is yet to be improved.

This paper proposes a new technique which is a combination of minimum Redundancy Maximum Relevance (mRMR) and Modified Logistic Regression for data reduction and data classifier, respectively. Logistic Regression model is known to have the risk of overfitting problem. Moreover, overfitting is associated to the curse of dimensionality and small sample size problem. Microarray is known to have the curse of dimensionality and the small sample size problem. This is the reason why the model has the risk of overfitting.

Overfitting models will have a low prediction accuracy. Therefore, the classifier was being modified to overcome the problem, by adding a regularization term to the cost function. Therefore, the model was able to obtain the optimal set of thetas for the model.

The structure of this research is organized as follows. In Section 2, some of the previous research which is related to this research is discussed. The system design of the proposed method is presented in Section 3. Furthermore, the experiment result and the discussion are discussed in Section 4. Finally, some conclusions are stated and discussed in Section 5.

2. **Related Works.**

Turgut et al. [9] used 8 different classification methods to find the most suitable method. The datasets used in this research were 2 breast cancer datasets from different sources. This research concludes that Support Vector Machine (SVM) and Logistic Regression are the best classifier among the others.

In 2018, Ma'ruf et al. [10] used the mRMR method as feature selection method with SVM as the classifier. The system performance is good, with F1-score of 0.81667. This research also compared the result of system with dimension reduction and without it. Furthermore, the system without dimension reduction gave F1-score of 0.72. This research concludes that the features that are selected from the dimension reduction process have a better generalization and represent the characteristic of the class well.

Aydadenta and Adiwijaya [11] used random forest as the dimension reduction method to decrease the redundancy of microarray data. The result is compared with the result of system using random forest without redundancy reduction. Furthermore, the performance of system using random forest gave the average accuracy of 72.63%. Meanwhile, random forest with redundancy reduction gave 91.24% of accuracy. Finally, this research concludes that redundancy reduction could be used to increase the performance of microarray data classification.

In our previous research [8], we used the Principal Component Analysis (PCA) as the feature extraction method and Logistic Regression as the classifier. The performance of the system gave the average accuracy of 72.58%. This low accuracy of the system is due to the fact that Logistic Regression has the risk of overfitting, that leads to a low accuracy of the system. To address this issue, this study will implement a modification to the classifier, by adding a regularization term to the cost function of Logistic Regression.

3. **Proposed Method.** Datasets used are 4 different cancer datasets from Kent-Ridge [4]. These datasets consist of Lung Cancer, Leukemia, Colon Cancer and Ovarian Cancer. Furthermore, details of these datasets are shown in Table 1.

Table 1. Details of cancer datasets

Dataset	Records	Features	Class
Colon Cancer	62	2.000	2
Lung Cancer	181	7.129	2
Leukemia	72	15.154	2
Ovarian Cancer	253	12.533	2

In Table 1, "Records" represents the quantity of patients that is being observed. While "Features" represents the information of the gene expressions of each record. Since microarray data had much smaller number of records compared to the features, the dimensional reduction process using feature extraction or feature selection was necessary. Furthermore, the 'cancer' class is denoted by 1 while the 'not cancer' class is denoted by 0.

The overview of system in this study consists of preprocessing the datasets, feature selection, splitting dataset, classifier building, classification and system evaluation. Furthermore, the system design can be seen in Figure 1.

The features in the datasets had a wide range of value, which could affect the system performance. Therefore, the datasets are being preprocessed using normalization, so that the value of the features on each dataset would be around 0 and 1.

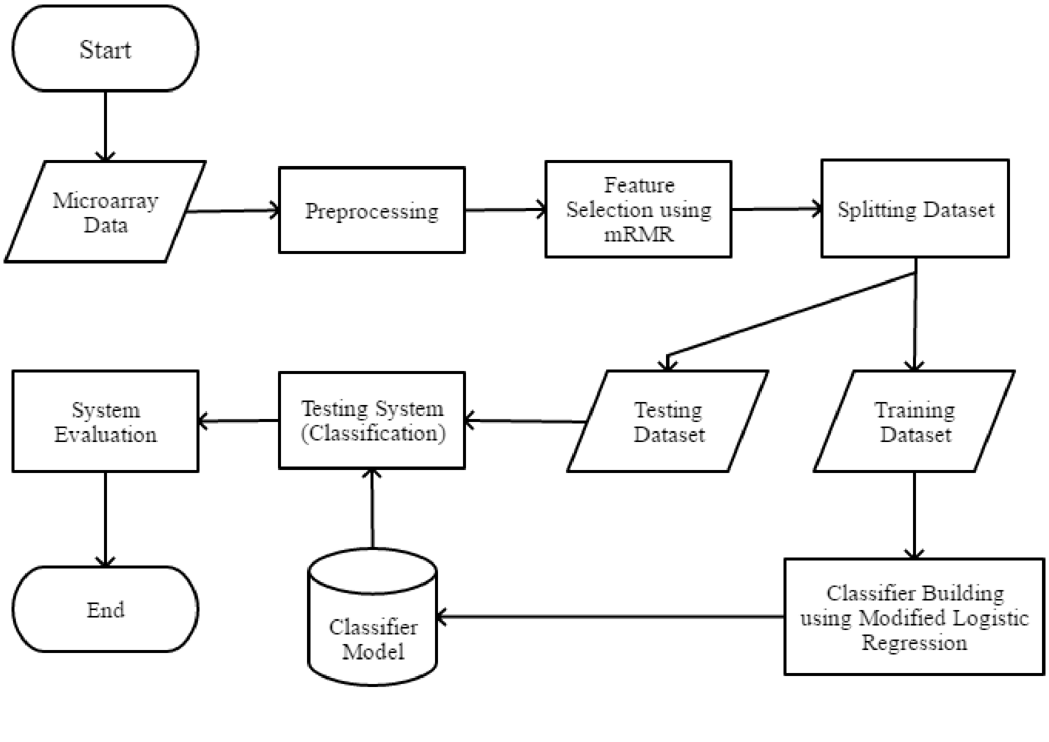

Figure 1. System design

3.1. Minimum Redundancy Maximum Relevance (mRMR).

mRMR is a part of feature selection method, which aims to select the features that have a high correlation with the class (relevance) and a low correlation between each feature (redundancy).

Since microarray has continuous data attributes, the value of maximum relevance is calculated using F-statistic between the attributes (genes) and the classification variable h. Furthermore, value of F-test of gene variable g_i in K classes represented as h has the form:

$$F(g_i, h) = \left[\sum_k \frac{n_k \left(\bar{g}_k - \bar{g} \right)}{K - 1} \right] \Big/ \sigma^2 \tag{1}$$

where $\bar{g}$ is the mean value of g_i in all sample of genes, $\bar{g}_k$ is the average value of g_i within the kth class, and σ^2 is the pooled variance (where n_k and σ_k are the size and the variance of the kth class). The t-test of 2-class classification will be reduced using F-test, with the correlation $F = t^2$. Thus, the maximum relevance of feature set S could be written as:

$$\max V_f, V_f = \frac{1}{|S|} \sum_i F(i, h) \tag{2}$$

The minimum redundancy condition may be specified in several ways. On this research, we use Pearson correlation coefficient $c(g_j, g_i) = c(i, j)$, and the condition is

$$\min W_c, W_c = \frac{1}{|S|^2} \sum_{i,j} |c(i, j)| \tag{3}$$

The set of MRMR feature is obtained by simultaneously optimizing the condition in Equations (2) and (3).

3.2. Logistic Regression.

Logistic Regression is known as an approximation of a mathematical model with an aim to analyze the correlation between some independent variables and a dependent variable. Logistic Regression is different from linear regression on the dependent variable type. Linear regression is applied for numerical variables, while Logistic Regression is applied for dichotomous variables [13]. Since the cancer data consist of 2 classes, which are 'cancer' and 'not cancer', Logistic Regression was selected as the classifier.

The formal model of Logistic Regression can be seen in Equation (4):

$$\log \frac{p(x)}{1 - p(x)} = \theta_0 + x \cdot \theta_j \tag{4}$$

Here $p(x)$ is a linear function of $\vec{x}$ and θ is the parameter of Logistic Regression. The quantity of θ is affected by the quantity of data that is being processed. Furthermore, the probability value of $p(x)$ is obtained by using an exponential function that simplifies the logarithm in Equation (4) to yield Equation (5):

$$p(x; b, w) = \frac{e^{\theta_0 + x \cdot \theta_j}}{1 + e^{\theta_0 + x \cdot \theta_j}} = \frac{1}{1 + e^{-(\theta_0 + x \cdot \theta_j)}} \tag{5}$$

The mis-classification rate is minimized by predicting the value of p as 0, if the value of $p < 0.5$. Meanwhile, the value of p is predicted as 1 if the value of $p \geq 0.5$ [8].

The key problem of Logistic Regression is to find the value of θ, so the algorithm will revolve to optimize parameter θ. The first step of θ optimization is to define the cost function $J(\theta)$. Cost function represents optimization objective, which measures how badly models are performing. The cost function needs to be minimized so that an accurate model with minimum error could be developed. The lowest cost function is denoted by

$\arg\min J(\theta)$ [14]. Furthermore, the cost function of Logistic Regression could be written as:

$$J(\theta) = -\frac{1}{N}\left[\sum_{j=1}^{N} y^{(j)} \cdot \log\left(h_\theta\left(x^{(j)}\right) + \left(1 - y^{(j)}\right)\log\left(1 - h\left(x^{(j)}\right)\right)\right)\right] \tag{6}$$

where h_θ is the prediction class of the data and $y^{(j)}$ is the actual class of the data.

Modified Logistic Regression. Logistic Regression attempts to predict a result based on the independent variables. However, the model appears to have more predictive power than it actually does. It is caused by the sampling bias. It will lead to the overstated accuracy of its prediction or known as "overfit".

Overfitting is associated to the curse of dimensionality and small sample size problem. Overfitting occurs when the input has a high number of dimensions and a small number of sample size [14]. Microarray is known to have the curse of dimensionality and the small sample size problem. This is the reason why the model has the risk of overfitting.

Overfitting models will have a low prediction accuracy. Therefore, we need to overcome the overfitting problem. There are some methods to overcome the problem, and one of them is by using regularization [3]. Regularization is a technique used to reduce the error of a model by fitting a function appropriately on the given training set.

There are several techniques of regularization. On this research, we will apply Lp regularization and find the effect on logistic classification process. Lp regularization will add an extra term to the cost function of Logistic Regression. The original cost function of Logistic Regression is defined by Equation (6). Furthermore, the regularized loss function is given by

$$E(\theta, D) = J(\theta, D) + \lambda R(\theta) \tag{7}$$

The general Lp regularization is defined as:

$$R'(\theta) = \lambda \|\theta_j\|_p^p = \lambda\left(\sum_{j=0}^{|\theta|} |\theta_j|^p\right) \tag{8}$$

where $\|\theta\|$ is the Lp norm of θ and λ is the parameter of regularization.

4. **Experiment Results and Discussion.** The results of testing and analysis of this research consist of four parts: result and analysis of previous research, the application of PCA + Modified Logistic Regression, mRMR + Logistic Regression and mRMR + Modified Logistic Regression. Furthermore, the comparison of the results and the analysis were conducted.

4.1. **Principal component analysis with Logistic Regression.** This research aimed to improve the accuracy of cancer detection system on the previous research. The method used on the previous research was Principal Component Analysis (PCA) with Logistic Regression. PCA extracts features using eigenvalues and eigenvectors. Furthermore, the eigenvectors were selected based on the Proportion of Variance (PPV).

PPV represented the percentage of selected eigenvalues to the total number of eigenvalues [14]. The PPV used on the research was 60%, 70%, 80%, 90% and 95%. The result of previous research is shown in Table 2. The table shows that the value of PPV is not correlated to the system accuracy of each dataset. PPV value represented the number of selected features of PCA process, and the higher value of PPV indicates the higher number of selected features.

Table 2. Experiment result of previous research

PPV	Accuracy (%)				
	Colon	**Leukemia**	**Lung**	**Ovarian**	**Avg**
60%	50	60	57.22	90	**64.31**
70%	66.67	80	61.11	92	**74.95**
80%	58.33	70	47.22	93.6	**67.29**
90%	66.67	80	58.33	100	**76.25**
95%	66.67	90	63.89	100	**80.14**
Average	**61.668**	**76**	**57.554**	**95.12**	**72.58**

4.2. **Principal component analysis with Modified Logistic Regression.** As the comparison to the previous research, the system using PCA as the dimension reduction with Modified Logistic Regression as the classifier was implemented. The number of assigned PPVs was equivalent to the previous research. Dimension reduction process using PCA produced a subset feature with smaller dimension, called Principal Component (PC). Furthermore, the number of PCs for each assigned PPV can be seen in Table 3.

Table 3. Number of PCs for each cancer dataset

PPV	Number of PCs			
	Colon	**Leukemia**	**Lung**	**Ovarian**
60%	3	11	31	4
70%	6	16	54	5
80%	10	22	68	12
90%	22	29	93	35
95%	36	34	99	85

Based on Table 3, the value of assigned PPV is correlated to the number of PC's obtained. When the assigned value of PPV is larger, then the number of obtained PCs would be greater. Meanwhile, the number of features of each cancer data is not correlated to the number of PC's obtained. Despite Ovarian Cancer has twice larger features number than Lung Cancer, Lung Cancer dataset obtains a greater number of PCs for all assigned PPV values. This means that the Ovarian Cancer dataset has a high redundancy.

Furthermore, the accuracy of each cancer data is as follows.

Table 4. Experiment result using PCA + Modified Logistic Regression

PPV	Accuracy (%)				
	Colon	**Leukemia**	**Lung**	**Ovarian**	**Avg**
60%	50	70	55.55	94	**67.39**
70%	66.67	80	55.55	96	**74.55**
80%	75	70	61.11	98	**76.03**
90%	66.67	90	50	98	**76.17**
95%	66.67	100	63.89	100	**82.64**
Average	**65**	**82**	**57.22**	**97.2**	**75.35**

Table 4 shows that the value of PPV is also not correlated to the system accuracy of each cancer data. Besides, the average accuracy of each cancer data shows an improvement on Colon and Leukemia Cancer dataset compared to the previous research. However, the average of the system is higher than the previous research, i.e., 75.35%.

4.3. **mRMR with Logistic Regression.** On this research, the testing of 4 different cancer datasets is implemented using mRMR as the feature selection method and Modified Logistic Regression as the classifier. Furthermore, features proportion indicated the ratio of selected features to the total number of features. As the comparison to the modified method, the test was also conducted using mRMR and Logistic Regression.

The parameter of mRMR was feature proportion, representing the number of selected features to the total number of features. On this research, the value of assigned feature proportion was 1%, 2%, 3%, 4% and 5%. Furthermore, the number of selected features for each cancer data is presented in Table 5.

Table 5. Number of selected features for each cancer dataset

Features proportion	Number of selected features			
	Colon	**Leukemia**	**Lung**	**Ovarian**
1%	20	71	125	151
2%	40	142	251	303
3%	60	213	376	454
4%	80	285	502	606
5%	100	355	627	756

Based on Table 5, the number of selected features of each cancer data is directly proportional to the assigned value of features proportion. This was happened because mRMR worked by ranking the features based on the F-statistic and Pearson correlation, then selected the features based on assigned features proportion. Colon Cancer dataset has 2000 features, with 5% number of features proportion, and the number of selected features is 100 features. Meanwhile, Leukemia Cancer dataset has 7129 features. As a result, the number of selected features using 5% features proportion is 355 features. This means that the number of selected features is the result of multiplying the assigned features proportion to the total number of features.

The accuracy of the system is shown in Table 6.

Table 6. Experiment result using mRMR + Logistic Regression

Features proportion	Accuracy (%)				
	Colon	**Leukemia**	**Lung**	**Ovarian**	**Average**
1%	59.76	68.75	82.83	96	**76.84**
2%	66.27	78.125	85.1	98	**81.87**
3%	85.79	59.38	63	96	**76.04**
4%	79.2	59.38	82.83	96	**79.35**
5%	79.2	78.125	84.61	88	**82.48**
Average	**74.04**	**68.75**	**79.67**	**94.8**	**79.31**

Table 6 shows that the value of features proportion is not directly correlated to the system accuracy of each cancer dataset. The highest testing accuracy of Colon Tumor dataset is obtained at proportion value of 3%. Meanwhile, the highest testing accuracy of Ovarian Cancer dataset was obtained at proportion value of 2%. This means that for Colon Tumor dataset, the number of informative features is 60. Informative features are the features with high relevance with the class and low redundancy each other. Meanwhile, the number of informative features on Ovarian Cancer dataset is 375 features. Furthermore, the highest accuracy was obtained at the proportion value of 5%. This shows that the highest accuracy is obtained when the data could be represented well. The average accuracy of the system is higher than the system using PCA as the dimension reduction, i.e., 79.31%.

4.4. **mRMR with Modified Logistic Regression.**

The obtained accuracy of system using mRMR + Modified Logistic Regression is shown in Table 7. Based on Table 7, value of features proportion is also not correlated to the obtained accuracy. The highest accuracy of Lung Cancer dataset is obtained with features proportion value of 2%. This means that for Lung Cancer dataset, the number of informative features is 142 features. Meanwhile, the highest accuracy of Ovarian Cancer dataset obtained with features proportion of 4%. This shows that for Lung Cancer dataset, the original dataset is represented well with only 2% of the data. Meanwhile, the Ovarian Cancer dataset needs 4% number of the original data to represent well.

TABLE 7. Experiment result using mRMR + Modified Logistic Regression

Features proportion	Accuracy (%)				
	Colon	**Leukemia**	**Lung**	**Ovarian**	**Average**
1%	92.31	85.71	97.29	90	**91.33**
2%	92.31	100	100	96	**97.07**
3%	85.79	100	89.19	98	**93.25**
4%	79.2	100	89.19	100	**92.09**
5%	79.2	100	94.59	98	**92.95**
Average	**85.76**	**97.14**	**94.05**	**96.4**	**93.33**

Furthermore, the highest average accuracy is obtained with features proportion value of 2%. This means that with only 2% number of data, the best accuracy could be obtained. The average accuracy of the system is higher than the system using Logistic Regression as the classifier, i.e., 93.33%. Moreover, a discussion with medical doctor concludes that the proposed system is valuable for validating their diagnosis result of cancer patients.

5. **Conclusions.**

In this research, mRMR is applied as the dimension reduction method with Modified Logistic Regression as the classifier to classify 4 different cancer datasets. The result of the experiment is compared with 3 other scenarios, which are PCA + Logistic Regression, PCA + Modified Logistic Regression and mRMR + Logistic Regression.

The result of classification process using mRMR as the dimension reduction and Logistic Regression classifier gives a higher accuracy than the PCA with the same classifier, for Colon and Lung Cancer dataset. This implies that mRMR could produce an optimal set of features by removing redundancy and keeping only the relevant features.

The assigned features proportion for mRMR is not directly correlated to the obtained accuracy of the system. For Colon Cancer dataset using Modified Logistic Regression classifier, highest accuracy is obtained with 1% and 2% features proportion. Meanwhile, the Ovarian Cancer dataset needs 4% features proportion to obtain the highest accuracy.

Modified Logistic Regression method gave the higher accuracy compared to the Logistic Regression method for all of the datasets. This implies that Modified Logistic Regression could produce a set of optimal thetas for the system. On previous research, using PCA + Logistic Regression system gave the average accuracy of 72.58%. Furthermore, the system using mRMR + Logistic Regression gave the average accuracy of 79.31%. Meanwhile, PCA + Modified Logistic Regression and mRMR + Modified Logistic Regression gave the average accuracy of 75.35% and 93.33% respectively.

Acknowledgment. This work is partially supported by Telkom University. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] M. Kumari, *Cancer Notes*, Lecture Notes, 2020.

[2] R. Sharma et al., Smart living for elderly: Design and human-computer interaction considerations, in *Human Aspects of IT for the Aged Population. Healthy and Active Aging. ITAP 2016. Lecture Notes in Computer Science*, J. Zhou and G. Salvendy (eds.), Cham, Springer, 2016.

[3] G. Manikandan and S. Abirami, A survey on feature selection and extraction techniques for high-dimensional microarray datasets, *Knowledge Computing and Its Applications*, pp.311-333, DOI: 10.1007/978-981-10-8258-0_14, 2018.

[4] J. Tian, H. Zhang, D. Lu, X. Zhou, A. Dong and X. Zheng, Stochastic access scheme for delay sensitive applications in wireless Ad Hoc networks, *Proc. of the 10th EAI International Conference on Mobile Multimedia Communications (MOBIMEDIA'17)*, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), Brussels, BEL, pp.314-320, DOI: 10.4108/eai.13-7-2017.2270206, 2017.

[5] C. Liu, X. Chen, Y. Li and M. Peng, Coexistence study between NB-IoT and cdma2000 systems, *Proc. of the 10th EAI International Conference on Mobile Multimedia Communications (MOBIMEDIA'17)*, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), Brussels, BEL, pp.159-163, 2017.

[6] A. Mortazavi and M. H. Moattar, Robust feature selection from microarray data based on cooperative game theory and qualitative mutual information, *Advances in Bioinformatics*, vol.2016, DOI: 10.1155/2016/1058305, 2016.

[7] S. Yeom, I. Giacomelli, M. Fredrikson and S. Jha, Privacy risk in machine learning: Analyzing the connection to overfitting, *IEEE the 31st Computer Security Foundations Symposium (CSF)*, Oxford, pp.268-282, DOI: 10.1109/CSF.2018.00027, 2018.

[8] A. Khoirunnisa, Adiwijaya and A. A. Rohmawati, Implementing principal component analysis and multinomial logit for cancer detection based on microarray data classification, *The 7th International Conference on Information and Communication Technology (ICoICT)*, Kuala Lumpur, Malaysia, pp.1-6, DOI: 10.1109/ICoICT.2019.8835320, 2019.

[9] S. Turgut, M. Dağekin and T. Ensari, Microarray breast cancer data classification using machine learning methods, *Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, Istanbul, pp.1-3, 2018.

[10] F. A. Ma'ruf, Adiwijaya and U. N. Wisesty, Analysis of the influence of minimum redundancy maximum relevance as dimensionality reduction method on cancer based on microarray data using support vector machine classifier, *The 2nd International Conference on Data and Information Science*, Bandung, Indonesia, 2018.

[11] H. Aydadenta and Adiwijaya, A clustering approach for feature selection in microarray data classification using random forest, *Journal of Information Processing Systems (JIPS)*, vol.14, no.5, pp.1167-1175, 2018.

[12] W. Maharani, Adiwijaya and A. A. Gozali, Degree centrality and eigenvector centrality in Twitter, *The 8th International Conference on Telecommunication Systems Services and Applications (TSSA)*, Kuta, Bali, Indonesia, pp.1-5, 2014.

[13] T. Gałecki, *The Environment of Support of a Massive Distributed Computing*, Ph.D. Thesis, Instytut Informatyki, Indonesia, 2019.

[14] X. Wan, The influence of polynomial order in logistic regression on decision boundary, *IOP Conference Series: Earth and Environmental Science*, vol.267, no.4, 2019.

[15] R. Liu and D. F. Gillies, Overfitting in linear feature extraction for classification of high-dimensional image data, *Pattern Recognition*, vol.53, pp.73-86, DOI: 10.1016/j.patcog.2015.11.015, 2016.

[16] R. K. Patel and V. K. Giri, Development of feature extraction and classification for bearing fault analysis of induction motor, *The 5th IEEE Uttar Pradesh Section International Conference on Electrical, Electronics and Computer Engineering (UPCON)*, DOI: 10.1109/upcon.2018.8596763, 2018.

[17] Q. Zou, K. Qu, Y. Luo, D. Yin, Y. Ju and H. Tang, Predicting diabetes mellitus with machine learning techniques, *Frontiers in Genetics*, vol.9, DOI: 10.3389/fgene.2018.00515, 2018.

[18] R. Ocampo-Vega, G. Sanchez-Ante, M. A. de Luna, R. Vega, L. E. Falcón-Morales and H. Sossa, Improving pattern classification of DNA microarray data by using PCA and Logistic Regression, *Intelligent Data Analysis*, vol.20, no.S1, pp.S53-S67, DOI: 10.3233/ida-160845, 2016.

[19] M. Al-Rajab, J. Lu and Q. Xu, Examining applying high performance genetic data feature selection and classification algorithms for colon cancer diagnosis, *Computer Methods and Programs in Biomedicine*, vol.146, pp.11-24, DOI: 10.1016/j.cmpb.2017.05.001, 2017.

[20] B. Ye and P. Liu, Classification of high-dimensional data: A random-matrix regularized discriminant analysis approach, *International Journal of Innovative Computing, Information and Control*, vol.15, no.3, pp.955-967, 2019.