International Journal on Advanced Science Engineering Information Technology

Vol.7 (2017) No. 4-2
ISSN: 2088-5334

Improved Support Vector Machine Using Multiple SVM-RFE for Cancer Classification

Nurul Nadzirah Mohd Hasri[#], Nies Hui Wen[#], Chan Weng Howe[#], Mohd Saberi Mohamad[*,$,^], Safaai Deris[*,$,^], Shahreen Kasim[+],

[#]*Artificial Intelligence and Bioinformatics Research Group, Universiti Teknologi Malaysia, 81310 Johor Bahru, Johor, Malaysia*

[*]*Center for Computing and Informatics, Universiti Malaysia Kelantan, City Campus, Pengkalan Chepa, 16100, Kota Bharu, Kelantan, Malaysia.*
E-mail: saberi@umk.edu.my

[$]*Artificial Intelligence and Big Data Institute, Universiti Malaysia Kelantan, City Campus, Pengkalan Chepa, 16100, Kota Bharu, Kelantan, Malaysia*

[^]*Faculty of Creative Technology and Heritage, Universiti Malaysia Kelantan, Karung Berkunci 01, 16300, Bachok, Kelantan, Malaysia*

[+]*Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, 86400, Batu Pahat, Johor, Malaysia*

***Abstract*— Support Vector Machine (SVM) is a machine learning method and widely used in the area of cancer studies especially in microarray data. A common problem related to the microarray data is that the size of genes is essentially larger than the number of samples. Although SVM is capable of handling a large number of genes, better accuracy of classification can be obtained using a small number of gene subset. This research proposed Multiple Support Vector Machine- Recursive Feature Elimination (MSVM-RFE) as a gene selection to identify the small number of informative genes. This method is implemented in order to improve the performance of SVM during classification. The effectiveness of the proposed method has been tested on two different datasets of gene expression which are leukemia and lung cancer. In order to see the effectiveness of the proposed method, some methods such as Random Forest and C4.5 Decision Tree are compared in this paper. The result shows that this MSVM-RFE is effective in reducing the number of genes in both datasets thus providing a better accuracy for SVM in cancer classification.**

***Keywords*— support vector machine (SVM); multiple support vector machine- recursive feature elimination (MSVM-RFE); leukemia; lung cancer**

I. INTRODUCTION

Cancer or called as malignancy is a group of disease involving uncontrolled and abnormal cell growth [1]. Furthermore, cancer is one of the main cause of death in this world. However, not all of uncontrolled and abnormal cell growths are cancerous, it all depends on the number of active and inactive cell in it. There are more than one hundred types of cancer such as skin cancer, breast cancer, leukemia, prostate cancer and other. Presently, in the field of medical, the major research is in the area of cancer analysis where there is a demand in developing a powerful method for the purpose of a cancer diagnosis.

Customarily, physical analyses of tissues are performed for cancer or tumour prognosis and diagnosis utilizing Computed Tomography (CT) scan, Chest X-ray, and Magnetic Resonance Imaging (MRI) [2], [3]. However, they can only identify the malignant cells in the late stage of cancer and would bring about low survival rates [4]. Thus, the studying of cancer to identify the formation of tumour at the earlier stage propels in molecular biologies such as DNA, protein, and RNA was proposed. However, as that review was investigated and examined utilizing low-throughput information, prior knowledge of disease is required for securing the information of candidate markers. Moreover, there is a limitation in finding the novel markers which is caused by the presence of an only small number of markers [5].

Therefore, microarray method is presented. In Gordon *et al*., [6] has investigated the cancer diagnosis method based on gene expression and showed the best performance in the accuracy of classification. The usefulness of the microarray data has motivated many researchers to perform large-scale

area of this study. The era of microarray technologies for measuring genome-wide expression profiles has prompted the development and improvement of various methods and techniques to distinguish between different classes of a complex disease like cancer through transcriptome analysis [7]–[10]. Besides, many of classification methods such as Support Vector Machine (SVM) [11], Directed Random Walk (DRW) [12] and Artificial Neural Network (ANN) [13] have been developed to help this era. Fig. 1 shows the visualization of the process in the microarray analysis.

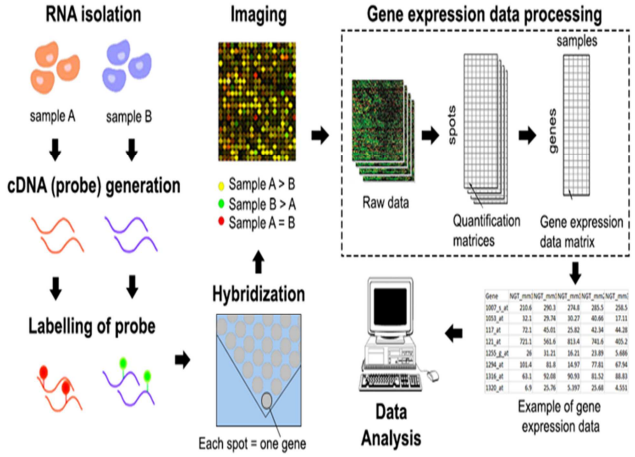

Fig. 1 Visualization of the process in microarray analysis

However, the preparation of microarray data is a critical step in biological function analyses, especially in cancer classification. The microarray data will produce a huge dataset with high dimensionality that contains informative genes, redundant genes, irrelevant genes, and noisy genes. Gene selection is a method to alleviate the problems of the irrelevant or redundant gene.

The pure Support Vector Machines (SVM) is also known as support vector network. It is one of the machine learning models that introduced by Vapnik [14]. Furthermore, it is a machine learning model with a related algorithm that analyzes the data and identifies the pattern of the data. It is used for the classification and regression analysis [11]. An SVM model is a representation data as a point in the space. It will construct the hyperplane on the map to classify the data and predict the group based on which side of the gap it's fall on such as in Fig. 2. Using the kernel trick, it can handle and effectively perform non-linear classification. It has four types of kernel, i.e., linear, polynomial, radial basis function (RBF) and sigmoid.

Besides, it is robust to a high variable-to-test degree and a huge number of variables, able to learn productively complex classification function, and manage to utilize intense regularization standards to abstain from overfitting [14]-[17]. It is broadly utilized as a part of the area of cancer studies and commonly in microarray data [24]–[26], [18]-[20].

Regrettably, the size of features or genes in microarray data is essentially bigger than the number of samples. In any case, the scantiness of a microarrays gene expression is so compelling that even an SVM classifier is unable to accomplish a palatable performance. Thus, the preprocessing step of gene selection or feature selection before undergoing the classification is vital for more trustworthy cancer classification [27], [21].

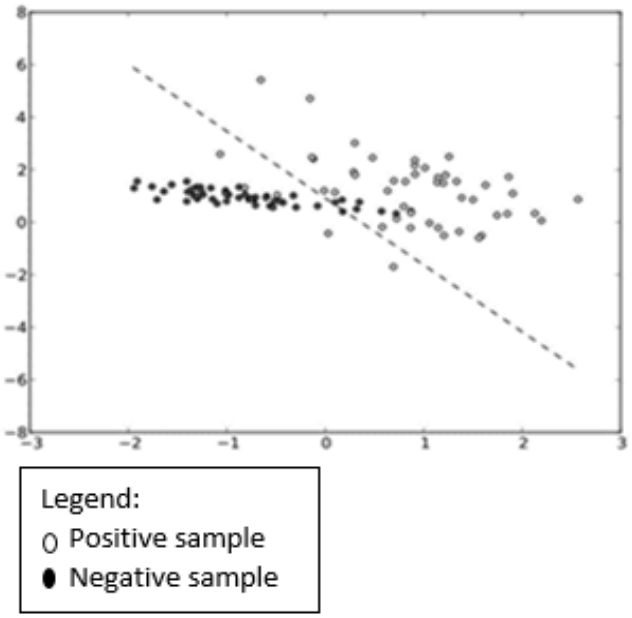

Fig. 2 The example of Linear-SVM-scatterplot

II. Material and Method

A. *Multiple Support Vector Machine Recursive Feature Elimination (MSVM-RFE)*

Currently, most of the microarray data for the cancer classification generated bewildering amounts of raw data, and the number of genes is larger than the number of samples. To secure against spurious results, gene selection is a better solution to solve the vital machine learning problem. Identifying a small number of informative genes is the objective of gene selection. Many evolutions had been made in the Support Vector Machine- Recursive Elimination (SVM-RFE), from the basic of SVM-RFE to two-stage SVM-RFE and multiple SVM-RFE.

Reducing the dimensionality of the dataset will yield a good analysis [24], [18]. Multiple SVM-RFE (MSVM-RFE) [28], [22] is an upgraded version of the original SVM-RFE. MSVM stand for multiple SVMs that use a backward elimination procedure to eliminate the lowest weight of the gene, similar to the SVM-RFE. However, at each step, the computation of feature ranking score is based on the statistical analysis of weight vector of multiple linear SVMs that being trained on a subset of the training data. This approach makes the result of MSVM-RFE to be better and more accurate compared to the SVM-RFE.

Furthermore, repeating the selection procedure on a few subsamples from bootstrap resampling on the training data is one of the ways to stabilize the gene selection method. This idea is applied to every step of recursive MSVM-RFE, rather than apply this idea on SVM-RFE all in all. Moreover, MSVM-RFE also used cross-validation instead of bootstrap sampling as the resampling method explores the higher possibility of choosing and determining a better subset of the gene in the recursive procedure. Hence, MSVM-RFE is a

meaningful and powerful approach in gene selection to select the informative genes for cancer classification. Based on these reasons, the MSVM-RFE has been chosen for the purpose of gene selection in this research to enhance SVM.

In order to train on different subsamples of original training data, we have t linear SVMs. The w_j is a weight vector of the jth linear SVMs, w_{ji} is a corresponding weight value associated with the ith feature and let $v_{ji} = (w_{ji})^2$. The score of feature ranking is computed with the following formula:

$$c_i = \frac{\bar{v}_i}{\sigma_{v_i}} \quad (1)$$

$$\bar{v}_i = \frac{1}{t}\sum_{j=1}^{t} v_{ji} \quad (2)$$

$$\sigma_{v_i} = \sqrt{\frac{\sum_{j=1}^{t}(v_{ji} - \bar{v}_i)^2}{t-1}} \quad (3)$$

where $\bar{v}_i$ is mean and σ_{v_i} is a standard deviation for the $\bar{v}_i$, However it is important to normalize the weight vectors before computed the ranking score for each gene.

$$w_j = \frac{w_j}{\|w_j\|} \quad (4)$$

The procedure of MSVM-RFE start with ranking the gene set, R = [ ]. From a selected gene subset of S = [1,…,d], the following step is repeated until all the features or genes are ranked. Firstly, the multiple linear SVMs are trained on subsamples of the original training data, with genes in set S as the input variable. Secondly, compute and normalize the weight vectors. By using the first equation, compute the ranking scores c, for genes in S. Next, find the gene with the smallest ranking score and eliminate that gene from the subset S. Lastly, update the list in gene set R. Fig. 3 shows the recursive procedure of MSVM-RFE.

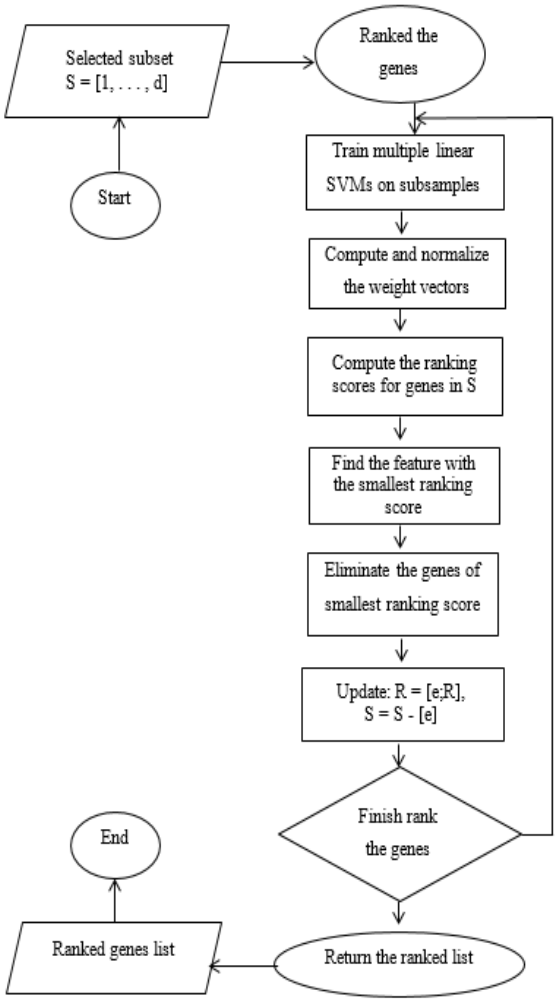

Fig. 3 The recursive procedure of MSVM-RFE

B. Improvement of SVM Using MSVM-RFE as Gene Selection

The improvement of accuracy in SVM has been accomplished through the enhancement made in this work using MSVM-RFE as gene selection. The classifier package that implements SVM has been used in this research. This package is a compilation of function for the application and creation of highly optimized, robustly evaluated ensembles of SVM. It creates a highly optimized ensemble of Radial Basis Function (RBF) SVM classifiers. The flow of MSVM-RFE is shown in Fig. 4 while Fig. 5 shows the enhancement of SVM using MSVM-RFE as gene selection.

Firstly, preprocessing step needs to be performed to sort the dataset following the standard input dataset that should be formatted as a data frame. They consist of the row for each observation and column for each gene or feature. In MSVM-RFE, the first column should be the true class label whereas, for the SVM classifier, the data should be separated between the class and factor. Afterwards, the data will be transposed and sorted according to input setting.

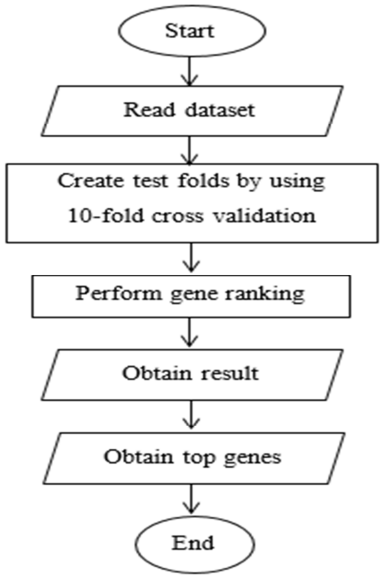

Fig. 4 The flow chart of MSVM-RFE

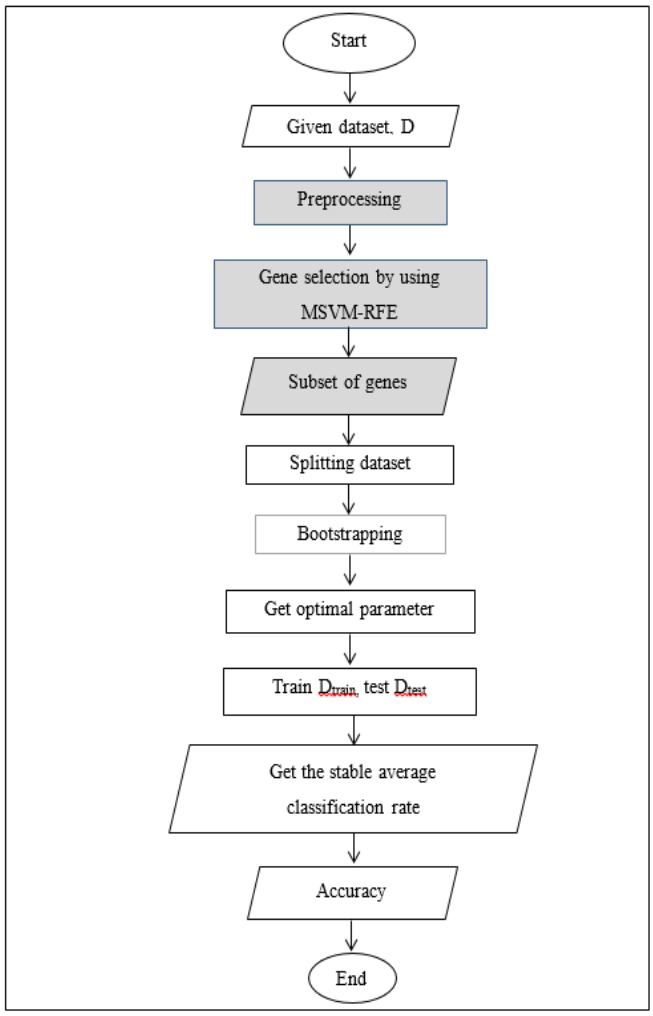

Fig. 5 The flow work of enhancement. The shaded figure is the part of enhancement

Then the MSVM-RFE is performed, where the first step is to set up the fold. This step has been repeated using 10-fold cross-validation on the training set for defining which remark are in which fold. Then, reformat the folds into a list that contains the test set indices for the fold. Then, by using this fold, perform gene ranking for all 10 training set. Indicate k=10 for the k-fold cross validation as the "multiple" part of MSVM-RFE, where the standard SVM-RFE is k=1. In this function, there are also has halved above parameter that allows us to cut the features or genes. This parameter will cut the features or genes in half each round, rather than one by one. In this work, this parameter has set halve.above=100 same as prior work. Hence, the genes will be cut in half each round until the number of genes fewer than 100 remains. The output from this step is a vector of genes or features, currently sorted from most to least useful. Lastly, the output of top feature for this step is the list of genes that are ordered by average rank across the 10 folds, where the lower the numbers in average rank, the better the result.

In the SVM step, the dataset needs to be split into separate training and testing subsets using a bootstrapping approach combined with a heuristic optimization algorithm and parallel processing to minimize and reduce the computation time. Test set has been kept aside during the training process of the SVM model. Generally, the test set comprises of one-third of the original samples. Then, bootstrapping is repeated until reasonable winning parameter combination is produced. The optimal parameters from the bootstrap step are utilized to train another classifier with the full train dataset and test the test dataset. Because the data in the testing set already contains known values for the attribute for predict, it is easy to determine whether the model's guesses are correct and obtain the better accuracy.

In this research, firstly the classifier that implements SVM classification has been run without gene selection of MSVM-RFE. The random genes in a dataset of varying numbers such as 10, 20, 30 and 40 until 100 genes have been tested. Then, SVM has been run with MSVM-RFE. By using the output of top features from MSVM-RFE result with varying number of genes such as 10, 20, 30 and 40 until 100 have been undergoing the classifier package to perform SVM. The best subset of the gene is repeated the classification process for 20 times to obtain average of the accuracy.

In this work, two types of datasets have been used, which are leukemia and lung datasets. The basic information of the dataset including the number of total genes, samples, and sizes of the class is shown in Table 1. The size of the class for leukemia dataset consists of 47 patients with acute lymphoblastic leukemia (ALL) and 25 patients with acute myeloid leukemia (AML). Meanwhile, the class size of lung dataset consists of 150 patients with adenocarcinoma (ADCA) and 31 patients with malignant pleural mesothelioma (MPM).

TABLE I
THE LIST OF THE DATASET

Dataset	Genes	Samples	Source
Leukemia	7129	72	[7]
Lung	12533	181	[6], [29], [23]

III. RESULT AND DISCUSSION

To evaluate the performance, the accuracy of the result is calculated according to [12].

$$\text{Accuracy} = \frac{\text{correctly predicted data}}{\text{total testing data}} \times 100\% \qquad (5)$$

The results of leukemia dataset are compared with several methods which are standard SVM, enhanced SVM with MSVM-RFE as gene selection, random forest by Moorthy and Mohamad [30], [24], Random forest with MSVM-RFE [31], [25] and varSeIFE [32], [26]. The result based on accuracy and computational time. The result is tabulated in Table 2. The shaded row in the table indicates the best method based on highest accuracy and shortest time.

The overall comparison of accuracy and computational time is presented to demonstrate the enhancement accomplished. Based on the result, the enhanced SVM with combine MSVM-RFE show an improvement in terms of better accuracy and lower computational time which can lead to lower computational cost.

The random forest has many advantages such as good predictive performance even though most predictive genes

are noisy and can handle large input genes without gene deletion [32], [26]. However, the result of accuracy for enhanced SVM is higher than standard Random Forest and Random Forest with MSVM-RFE. The enhanced SVM achieve the highest result of accuracy with 0.986 and the shortest time with only 0.013 hours. The result of enhanced SVM followed by Random Forest with MSVM-RFE with 0.949 for accuracy and 0.060 hour, standard Random Forest with 0.925 for accuracy and 0.013 hour, varSeIFR with 0.911 for accuracy and 1.280 hour and lastly standard SVM with 0.881 for accuracy and 1.830 hour. Thus, based on this result it proved that MSVM-RFE is a power gene selection by improving classification method with a better result.

TABLE II
THE RESULT OF ACCURACY AND COMPUTATIONAL TIME FOR LEUKEMIA DATASET

Method	Accuracy	Computational Time (Hours)
SVM	0.881	1.830
SVM + MSVM-RFE	0.986	0.013
Random Forest	0.925	0.330
Random Forest + MSVM-RFE	0.949	0.060
varSeIFR	0.911	1.280

Meanwhile, the result of lung dataset had been compared with some methods which are standard SVM, enhanced SVM with MSVM-RFE as gene selection and C4.5 Decision Tree [33], [27] methods. The result based on accuracy is tabulated in Table 3. The shaded row in the table indicates the best method based on highest accuracy. All the findings show that the enhanced SVM outperforms the SVM without gene selection and C4.5 Decision Tree in terms of higher classification accuracy with 0.989. The result of enhancing SVM follow by C4.5 Decision Tree with 0.926 of accuracy and standard SVM with 0.920. Thus, it is proven that better accuracy can be gained by reducing the number of genes as the result of implementing gene selection.

TABLE III
THE RESULT OF ACCURACY FOR LUNG DATASET

Method	Accuracy
SVM	0.920
SVM + MSVM-RFE	0.986
C4.5 Decision Tree	0.926

IV. CONCLUSIONS

In this study, MSVM-RFE is implemented in the standard SVM as gene selection to handle a large number of genes in microarray data for identifying the small informative gene. Multiple SVM-RFE (MSVM-RFE) is an upgraded version of the original SVM-RFE. This method using a backward elimination procedure that eliminates the lowest weight of the gene, same like SVM-RFE. However, at each step, the computation of feature ranking score is based on the statistical analysis of weight vector of multiple linear SVMs that being trained on a subset of the training data. The implementation of MSVM-RFE yields better and more accurate result compared to the SVM-RFE. MSVM-RFE is implemented to enhance the performance of SVM in terms of accuracy and computational time. The performance of the enhanced method has been compared with several methods such as Random Forest, Random Forest with MSVM-RFE, varSeIFE and C4.5 Decision Tree by using two different datasets of gene expression which are leukemia and lung cancer. All the findings show that the enhance SVM outperform the SVM without gene selection in terms of higher classification accuracy in both datasets and lower computational time in leukemia dataset. However, this research still has some limitations such as larger dataset are taking longer computational time, and all the datasets used require preprocessing before undergo the gene selection and classification processes. Therefore, there are many works that can be done in future to improve the results of the used method. Firstly, the result of this research can be compared with more performance measurement such as error rate, specificity, and sensitivity. Secondly, to implement, test and analyze the strength of MSVM-RFE with other classifier and compare the result with this research. Lastly, to add more type of dataset other than leukemia and lung data.

ACKNOWLEDGMENT

This research is supported by Universiti Teknologi Malaysia through the Tier 1 research grants (Grant numbers: Q.J130000.2528.11H11).

REFERENCES

[1] J. Lieberman, A. Lal, and D. Chowdhury, "Therapeutic and diagnostic strategies," U.S. Patent Application No. 15/162,337, 2016.
[2] C. F. Mountain and C. M. Dresler, "Regional lymph node classification for lung cancer staging," *Chest*, 111(6), 1718-1723, 1997.
[3] C. F. Mountain, "Revisions in the international system for staging lung cancer," *Chest*, 111(6), 1710-1717, 1997.
[4] J. A. Tsou, J. A. Hagen, C. L. Carpenter, and I. A Laird-Offringa, "DNA methylation analysis: a powerful new tool for lung cancer diagnosis," *Oncogene*, 21(35), 5450, 2002.
[5] W. Engchuan, A. Meechai, S. Tongsima, and J. H. Chan, *Cross-Platform Pathway Activity Transformation and Classification of Microarray Data*, in Computational Intelligence in Information Systems, pp. 139-148: Springer International Publishing, 2015.
[6] G. J. Gordon, R. V. Jensen, L. L. Hsiao, S. R. Gullans, J. E. Blumenstock, S. Ramaswamy and R. Bueno, "Translation of microarray data into clinically relevant cancer diagnostic tests using gene expression ratios in lung cancer and mesothelioma," *Cancer research*, 62(17), 4963-4967, 2002.
[7] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov and C. D. Bloomfield, "Molecular classification of cancer: class discovery and class prediction by gene expression monitoring," *science*, 286(5439), 531-537, 1999.
[8] A. A. Alizadeh, M. B. Eisen, R. E. Davis, C. Ma, I. S. Lossos, A. Rosenwald, and J. I. Powell, "Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling," *Nature*, 403(6769), 503-511, 2000.
[9] S. Ramaswamy, K. N. Ross, E. S. Lander, and T. R. Golub, "A molecular signature of metastasis in primary solid tumors," *Nature genetics*, 33(1), 49-54, 2003.
[10] A. Perez-Diez, A. Morgun, and N. Shulzhenko, *Microarrays for cancer diagnosis and classification*, in Microarray Technology and Cancer Gene Profiling, pp. 74-85: Springer New York, 2007.
[11] C. C. Chang, and C. J. Lin, "LIBSVM: a library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, 2(3), 27, 2011.

[12] C. S. Seah, S. Kasim, and M. S. Mohamad, "Specific Tuning Parameter for Directed Random Walk Algorithm Cancer Classification," *International Journal on Advanced Science, Engineering and Information Technology*, 7(1), 176-182, 2017.

[13] N. M. Nawi, A. S. Hussein, N. A. Samsudin, N. A. Hamid, M. A. M. Yunus, and M. F. Ab Aziz, "The Effect of Pre-Processing Techniques and Optimal Parameters selection on Back Propagation Neural Networks," *International Journal on Advanced Science, Engineering and Information Technology*, 7(3), 2017.

[14] V. N. Vapnik and V. Vapnik, *Statistical learning theory*. New York: Wiley, 1998, vol. 1.

[15] M. P. S. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares and D. Haussler, "Knowledge-based analysis of microarray gene expression data by using support vector machines," *Proceedings of the National Academy of Sciences*, 97(1), 262-267, 2000.

[16] A. Statnikov, C. F. Aliferis, L. Tsamardinos, D. Hardin and S. Levy, "A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis," *Bioinformatics*, 21(5), 631-643, 2005.

[17] U. K. Hassan, N. M. Nawi, and S. Kasim, "Classify a protein domain using sigmoid support vector machine," in *Information Science and Applications (ICISA), 2014 International Conference*, *IEEE*, May. 2014. pp. 1-4.

[18] U. K. Hassan, N. M. Nawi, and S. Kasim, A. A, Ramli, M. F. M., Fudzee, and M. A. Salamat,"Classify a Protein Domain Using SVM Sigmoid Kernel". In *Recent Advances on Soft Computing and Data Mining*, pp. 143-151, Springer, Cham.

[19] R., Hassan, R. M., Othman, P., Saad, and S., Kasim, "A compact hybrid feature vector for an accurate secondary structure prediction" *Information Sciences*, *181*(23), 5267-5277.

[20] S., Ismail, R. M., Othman, S., Kasim, R., Hassan, H., Asmuni, and J., Taliba, "Pairwise protein substring alignment with latent semantic analysis and support vector machines to detect remote protein homology", *International Journal of Bio-Science and Bio-Technology*, Volume 3, Issue 3, 2011, Pages 17-34.

[21] F. M., Abdullah, R. M., Othman, S., Kasim, R., Hashim, R., Hassan, H., Asmuni, and J., Taliba, "An Optimal Mesh Algorithm for Remote Protein Homology Detection", *Communications in Computer and Information Science*, Volume 151 CCIS, Issue PART 2, 2011, Pages 471-497.

[22] S., Ismail, R. M., Othman, S., Kasim, R., Hassan, H., Asmuni, and J., Taliba, "Pairwise protein substring alignment with latent semantic analysis and support vector machines to detect remote protein homology" *Communications in Computer and Information Science*,Volume 151 CCIS, Issue PART 2, 2011, Pages 526-546.

[23] F. M., Abdullah, R. M., Othman, S., Kasim, R., Hashim, R., Hassan, H., Asmuni, and J., Taliba, "An Optimal Mesh Algorithm for Remote Protein Homology Detection", *International Journal of Bio-Science and Bio-Technology*, Volume 3, Issue 2, June 2011, Pages 13-38.

[24] I. Guyon, J. Weston, S. Barnhill and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine learning*, 46(1-3), 389-422, 2002.

[25] S. Mukherjee, "Classifying microarray data using support vector machines," *A practical approach to microarray data analysis*, 1, 166-185, 2003.

[26] E. B. Huerta, B. Duval, and J. K. Hao, *A hybrid GA/SVM approach for gene selection and classification of microarray data*, in Applications of Evolutionary Computing, pp. 34-44 Berlin, Heidelberg: Springer, 2006.

[27] Y. Tang, Y. Q. Zhang and Z. Huang, "Development of two-stage SVM-RFE gene selection strategy for microarray expression data analysis," *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 4(3), 365-381, 2007.

[28] K. B. Duan, J. C. Rajapakse, H. Wang and F. Azuaje, "Multiple SVM-RFE for gene selection in cancer classification with expression data," *IEEE transactions on nanobioscience*, 4(3), 228-234, 2005.

[29] G. Roffo, S. Melzi, and M. Cristani, "Infinite feature selection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 4202-4210.

[30] K. Moorthy and M. S. Mohamad, "Random forest for gene selection and microarray data classification," *Bioinformation*, 7(3), 142, 2011.

[31] K. Moorthy, "Improved Random Forest with Multiple Support Vector Machine for Gene Selection and Classification of Microarray Data," Doctoral dissertation, Universiti Teknologi Malaysia, 2015.

[32] R. Díaz-Uriarte and S. A. De Andres, "Gene selection and classification of microarray data using random forest," *BMC bioinformatics*, 7(1), 3, 2006.

[33] A. C. Tan and D. Gilbert, "Ensemble machine learning on gene expression data for cancer classification," 2003.