Indonesian Journal of Electrical Engineering and Computer Science
Vol. 20, No. 2, November 2020, pp. 712~719
ISSN: 2502-4752, DOI: 10.11591/ijeecs.v20.i2.pp712-719

Comparison of microarray breast cancer classification using support vector machine and logistic regression with LASSO and boruta feature selection

Nursabillilah Mohd Ali[1], Nor Azlina Ab Aziz[2], Rosli Besar[3]
[1,2,3]Centre for Engineering Computational Intelligence, Faculty of Engineering and Technology, Multimedia University, Malaysia
[1]Rehabilitation Engineering & Assistive Technology Research Group, Faculty of Electrical Engineering, Universiti Teknikal Malaysia Melaka, Malaysia

Article Info

Article history:

Received Feb 17, 2020
Revised Apr 24, 2020
Accepted Apr 14, 2020

Keywords:

Boruta
Breast cancer
LASSO
LR
Micrarray data
SVM

ABSTRACT

Breast cancer is the most frequent cancer diagnosis amongst women worldwide. Despite the advancement of medical diagnostic and prognostic tools for early detection and treatment of breast cancer patients, research on development of better and more reliable tools is still actively conducted globally. The breast cancer classification is significantly important in ensuring reliable diagnostic system. Preliminary research on the usage of machine learning classifier and feature selection method for breast cancer classification is conducted here. Two feature selection methods namely Boruta and LASSO and SVM and LR classifier are studied. A breast cancer dataset from GEO web is adopted in this study. The findings show that LASSO with LR gives the best accuracy using this dataset.

Corresponding Author:

Nursabillilah Mohd Ali,
Rehabilitation Engineering and Assistive Technology Research Group,
Faculty of Electrical Engineering,
Universiti Teknikal Malaysia Melaka,
Jalan Hang Tuah Jaya, 76100 Durian Tunggal, Melaka, Malaysia.
Email: nursabillilah@utem.edu.my

1. INTRODUCTION

According to the Global Cancer Statistics Report [1-2], it is estimated that more than 2 million patients were newly diagnosed with cancer and almost 627,000 breast cancer mortalities were reported. The increasing number of patients being diagnose with cancer shows a tremendous alarming rate with high mortality and morbidity in developed and developing countries. It is also stated that breast cancer is the second leading cause of cancer death, it is also the most common cancer diagnosed amongst women worldwide [3]. In fact, according to the Malaysian Cancer Registry Report 2007 to 2011 it is the most prevalent cancer among Malaysian female as shown in Figure 1.

Breast cancer survival rate is greatly influenced by malignancy's stage during diagnosis [4]. Lack of early detection disgnostic system and delayed treatments contributed to high death among breast cancer patients. Hence, early cancer diagnosis and treatment is needed to reduce the risk of cancerous and abnormal tissue from spreading to other organs [5].

Traditionally, human intervention by medical doctors and physicians are needed to detect, identify and validate the existance of breast cancer. However, this method is subject to human error, inaccuracy, time-consuming and cost. Computer aided technology can help to overcome the disadvantages of the traditional method. The diagnosis are normally delivered using result of mammography, ultra-sound, Contrasted-Enhanced (CE), x-ray, CT scan and magnetic resonance imaging (MRI) [6-8]. Among all of these

imaging tools, mammogram is the most frequent and widely used in breast cancer screening. However, it is not efficient for patients with age of under 40 years old due to heavy and dense breast tissues [9]. CE digital mammography could provide more accurate result than mammogram and ultrasound in heavy and dense breast patients. However, it is not extensively used because of expensive and could cause enormous level of radiation [10]. Similar to CE, MRI also able to detect small injury that cannot be detected via mammogram. It also very costly and could result in the over-diagnoses [11].

Microarray technology is high-throughtput technology used to produce the huge amount of gene expression profiles in cancerous and non-cancerous cells. It provides alternative ways to the imaging technologies. Microarray genes profiles carries important information which can be used in ensuring efficient drugs in targeted theraphy, disease monitoring and also in identifying new potential cancerous cells marker. Typically, microarray genes expression data comes with huge (up to thousands) number of genes with small number of sample/class; e.g. number of genes (p) > number of sample (n) [12]. This condition is known as 'curse of dimensionality' in which there are imbalance number of number of genes/features (p) with respect to number of sample (n) [12]. The large amount of information carried by microarray data has opened up research on application of intelligent system and machine learning in classifying this data to help the medical doctor and histopathologist in analysing the result. Figure 2 shows the transformation of the microarray from image to gene expression profiles.

This work focus on our preliminary findings on application of Support Vector Machine and Logistic Regression in classification of breast cancer microarray data. To tackle the dimensionality issue, two feature selection methods are considered, LASSO and Boruta. Even though, several works had been reported on the application of machine learning for microarray data, not many focus solely on its application towards breast cancer. Microarray data identifies breast cancer using distinct subtypes of genes tumors profiles. The finding shows that LR with LASSO has the best performance.

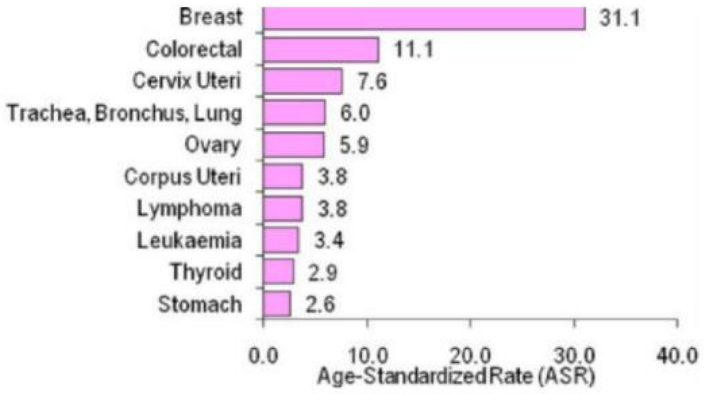


Figure 1. Ten most frequent cancers in Malaysian females 2007-2011 [3]

	samples	1007_s_at	1053_at	117_at	121_at	1255_g_at	1294_at	13
2	relapse (1	0	1	0	0	0	0	
3	ER Status	ER+	ER+	ER+	ER-	ER+	ER+	ER
4	GSM3677	1.9509	0.1262	0.2569	0.4178	-0.4378	0.4715	
5	GSM3677	1.9371	-0.0503	0.0801	0.6076	-0.3644	0.3133	
6	GSM3677	2.2929	0.047	0.2308	0.6326	-0.4288	0.3789	
7	GSM3678	1.8104	0.2041	0.1208	0.4302	-0.3678	0.1935	
8	GSM3678	1.9095	0.2324	0.176	0.3928	-0.4003	0.1794	
9	GSM3678	1.6403	0.0562	0.0743	0.4557	-0.3918	0.2966	
10	GSM3678	2.1778	0.202	0.0965	0.6071	-0.3657	0.3624	
11	GSM3678	1.7961	-0.0307	0.0187	0.4759	-0.3907	0.1766	
12	GSM3678	1.6642	0.2271	0.2325	0.408	-0.4855	0.1862	
13	GSM3678	1.9228	0.0672	0.106	0.4792	-0.3848	0.3577	
14	GSM3678	2.0022	0.0479	0.2232	0.4597	-0.4554	0.1292	
15	GSM3678	1.6571	0.2583	0.1135	0.649	-0.4199	0.6532	
16	GSM3678	1.9211	-0.0113	0.2078	0.4548	-0.3715	0.4922	
17	GSM3679	1.6322	0.0639	0.1833	0.6341	-0.4187	0.6876	

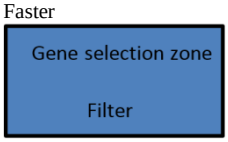
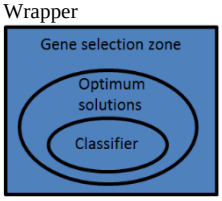
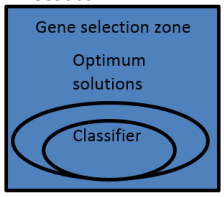
(a) (b) (c)

Figure 2. Physical sample from DNA human body(a) Gene expression (in image expression) using high density DNA technology(b) to microarray gene expression profiles in dataset form (c)

2. RELATED WORK

Gene selection method can be categorized to three methods namely filter, wrapper, embedded and hybrid representation. Table 1 shows the summarized of gene selection technique for microarray profiles with their advantages and disadvantages.

Table 1. Comparison on Feature Selection type and the element representation [13-16]

Method Representation	Element *Benefits*	*Limitation*	*Instances*
Filter High rank Low score Faster 	-Very fast and simple computation -Low time complexity -Only the highest ranking features selected while remove others -Independent features	-Ignore interaction with the classifier -Feature not dependent on any classifier -Generate redundancy -Only determined genes based on scored and rank features and ignore others -Difficult to determine starting value to rank the features	-Information gain -Chi-square method -Variance threshold -Fisher criterion -Correlation coefficient -t-Test
Wrapper 	-Usually optimize the classifier performance -Start calculate by using randomness value -Feature dependent -Optimum solutions -Interaction between feature selection and subset -Have interaction with classifier -Feature selection process not sensitive to dataset	-Computationally more expensive than filter method due to iterations step and cross validation -Randomness and iterations search will takes enough times to eliminate redundancy	-Genetic algorithm -Ant Colony -Particle swarm optimization -Sequential feature selection -Recursive feature elimination
Embedded 	-Taking benefits from filter and wrapper models -Quite similar to wrapper -Optimize objective function/learning algorithm -Feature dependencies -Computational time better than wrapper -Interact with classifier	-Always tend to over fitting -Dependent classifier	-Decision tree -LASSO regression -SVM-RFE
Hybrid Combination between filter//wrapper or other FS	-Taking advantages from numerous methods (filter and wrapper and others) -Can combine between other approaches to taking its advantages	-Complex -computational time	-Mutual information -Multi-objective optimization -kNN -SVM -Random forest

In the past research, classification of cancer had been involved with machine learning method and many advanced methods have been emerged with fast and accurate result [17]. Machine learning (ML) algortihms have been intensively used in model prediction, targeting and classification in various applications such as in medical data predictions. ML is defined as a model that 'study', 'acquire' and 'memorize' the prior or previous data to forecast the coming and future data. There are a lot of study focuses on various kind of techniques such as statistical, probabilistic and optimization that can be applied as the 'study'or learning models. These learning models such as artificial neural network, K-nearest neighbor, support vector machine (SVM), logistic regression (LR) and random forest (RF) are amongst the classifier that have been widely used in many research.

ML learning models come up with two types namely supervised and unsupervised learning [18]. The supervised learning constructs from known class (labeled training data) whereas unsupervised learning form the features from unknown source/class data (unlabeled training data). In this work, we focused on supervised learning models to construct and classify the breast cancer microarray data.

SVM is one of the most popular ML approach. SVM is prevailing in recognizing linear and non-linear models in large and complex data. The objective is to create the hyperplane/boundary based on its orientation and positions so that the data points will be far from nearby data points from every class. This nearby points are known as support vector. In a work done by [18] using binary cancerous microarray data such as colon, ovarian, central nervous and lung, it is shown that RF outperform SVM. However, it is stated that the result of classification accuracy either high or low can be influenced by dimensional reduction method used in the feature selection task [18]. Thus, it is important to choose the best feature selection method prior to the classification.

LR is generated from the basic concept of linear model that has intercept and coefficient value. It is coming form statistics appracoh that is used for data prediction by finding the area under the receiver operating characeristc (ROC). It is noted that ROC graph can be determined by a simple and intuitive method [19]. In a recent research, Almugren and Alshamlan [20] presented cancer classification technique to classify different type of cancer. The approach uses different type of cancer based on wrapper method using metaheuristic application. On the basis of the comprehensive literature review, the used of wrapper based using Boruta feature selection had been rarely studied. It is interesting to look at the aforementioned approaches for cancer classification using standard dataset available for cancer to have fair comparison. This paper proposed a novel approach to classifiy a diagnose breast cancer using human RNA microarray dataset based on the used of feature selection and without applied feature selection using LR and SVM classifier.

3. METHODOLOGY

The system studied here starts with feature selection where its input is the raw microarray data consisting of all features and the output is a reduced data with only the selected features. The reduced data is then fed to classifier to be classified. The flowchart in Figure 3 presented this system.

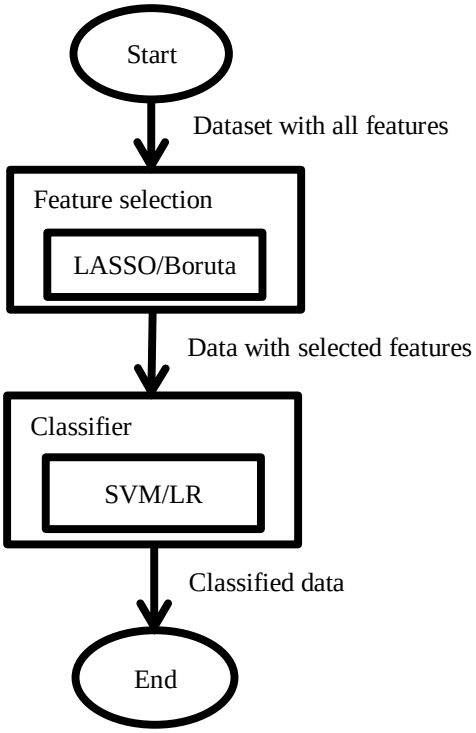


Figure 3. Flowchart of the classification system

3.1. Dataset

A human recurrence breast cancer microarray dataset is retrieved from Gene Expression Omnibus (GEO) database for this study. The data is labelled as estrogen receptor active (ER+) or estrogen receptor inactive (ER-) or in other words either the gene features are belonging to normal or abnormal cells. The distribution of the samples is illustrated in Table 2. The classification of such data is important for potential treatment therapy.

Table 2. Representation of breast cancer microarray dataset GSE2034 (recurrence dataset)

	ER+	ER-	Sample	Gene	Class
GSE2034	180	106	286	22283	Binary

In this work the raw data retrieved from the GEO web is splitted into two separate file namely features and target. No preprocessing is done prior to this. Next, the data is divided into training and test set with the ratio of 20% test and 80% training.

3.2. Feature selection

Feature selection is used to reduce high dimensionality of microarray features so that only the informative features are used for classification. Since, microarray medical data consists of huge number of input features, feature selection or gene selection is required in this study before it is feed into classification task. In this work, the effectiveness of Boruta feature selection method which introduced in 2010 by Kursa and Rudnicki [21] and LASSO [22] is studied.

It is a wrapper based method that make used of random forest classification algorithm implementation. Boruta is also known as ensemble method where the selector task can be implemented by selecting multiple unbiased weak classifiers and separately classify informative and noninformative features/attributes. It makes use of calculating average and standard deviation of the loss accuracy. The RF classifier that built in the Boruta algorithm will decide the important parameters/element by randomly reducing the misleading fluctuations and correlations. Research had shown that Boruta is computationally intensive method. However, when it is used together with random forest classifier, it is able to efficiently optimize dataset and contribute to good classification accuracy [23].

LASSO however is different from Boruta, as it constructs a linear model and generate regression coefficient using *Lasso or L1* distance. In linear model method, it has added penalty to the gene features to avoid overfitting. From the penalty that is applied to the coefficients, the *L1* has the attribute to reduce and shrink less or more coefficients to zero. Hence, based on these concept, the feature can be removed from the model. It is actually generated from the basic concept of linear model that has intercept and coefficient value. In short, LASSO will select the features with the coefficient are non-zero.

3.3. Classification

Support Vector Machine (SVM) is a supervised learning classification method that is introduced by Vapnik in 1995 [24]. It is a discriminate classifier that could be used in finding optimal hyperplane based on margin maximization. It searches the best hyperplane to separate the binary classes by maximizing the margin or distance between data points and the separable data. Normally, it is used for linear and non-linearly separable data. Figure 4 shows conventional separable margin for data splitting. It is frequently chosen for microarray data classification as seen in [25].

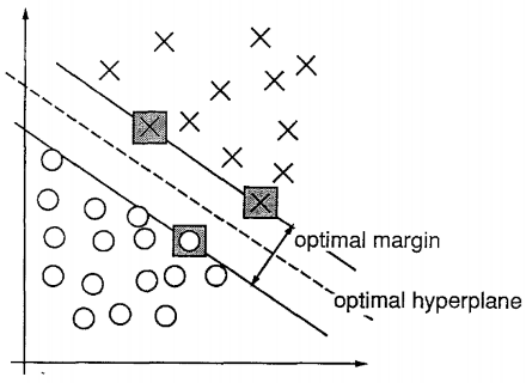


Figure 4. Conventional SVM hyperplane margin for separable data

Logistic regression also known as Longit is a statistic model introduced by Berkson in 1944 [26]. LR is widely used for binary classification. The aim of the LR is to find the probability of *p* outcome that occurs to one or two targeting data. In the contexts of LR, it makes used of linear function to finding the maximum likelihood of the predictor value. Maximum likelihood means the estimation of LR model in

generating exponential data curve. Figure 5 shows the linear model between linear regression and logistic regression in terms of the line and curve model. It is based on exponential distribution function in which is formulated with vectorization of parameters using known function of data. It is preferable to be used since it is simple and easy to be implemented.

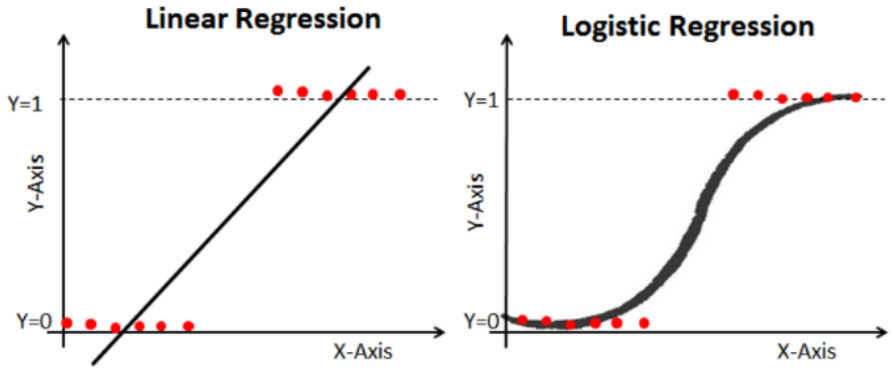


Figure 5. Linear model versus LR

4. RESULTS AND ANALYSIS

The effectiveness of the feature selection and classifier algorithms are studied here, (1) the performance of SVM and LR without any feature selection algorithm are compared followed by (2) the performance with the implementation of Boruta and LASSO feature selection algorithm with SVM and LR classifier. This is to observe the effectiveness of classification of the proposed result without and with the feature selection algorithm.

The experiments were carried out using a desktop computer equipped with Intel® core ™ i5-7200U CPU @ 2.50GHz 2.71 GHz and 8GB RAM. The algorithm was written and analysed using Python programme in Spyder web environment. For criteria (1), using LR classifier, 72.22% accuracy is achieved whereas the SVM classifier obtained 59.72% accuracy. This shows that without the feature selection, LR performs better than SVM for the selected dataset.

For criteria (2), when feature selection is implemented together with the classification algorithms, LR maintains to perform better than SVM when combined with both Boruta and LASSO feature selection algorithm. Boruta feature selection shows improvement of classification result to 69.44% and 73.61% for Boruta+SVM and Boruta+LR respectively. Between SVM and LR, Boruta gives significant improvement to SVM. Combination of LASSO+LR shows a very significant improvement of almost 37% in comparison to using LR algorithm on itw own. Unfortunately, there is no changes observed when LASSO is used with SVM. The result is similar with using SVM classifier only. Table 3 shows the preliminary result using the proposed method namely Boruta and LASSO feature selection with and without classifier respectively SVM and LR.

Table 3. Comparison result of Microarray data based on two criteteria mentioned in (1) and (2)

Criteria	SVM	LR
Without Feature Selection (FS)	59.72%	72.22%
Boruta	69.44%	73.61%
LASSO	59.72%	98.61%

5. CONCLUSION

Application of SVM and LR together with Boruta and LASSO feature selection techniques on the microarray breast cancer classification process is explained in this work. There is no improvement observed when using LASSO in SVM. However, using Boruta feature selection, the classification accuracy increase by almost 10% when Boruta is used prior to classification using SVM. Whereas, combination of LASSO+LR show substantial improvement when the accuracy achieved shoot to almost 99% which is better than the classification with LR without any feature selection. However, Boruta+LR only increase by lesser than 2% from classification result of LR. Future work will include more dataset, other feature selection and classifier algorithms, tuning of parameters in the classifiers and feature selection algorithms.

ACKNOWLEDGEMENTS

This work is supported by Universiti Teknikal Malaysia Melaka, the Ministry of Education Malaysia, under Funding Number: KPT(BS)850320045568 through SLAB Sponsorship Awards and Multimedia Univerity, Malaysia. Also, thanks to the referred reviewers for their valuable comments and suggestions that made considerable improvement to our work.

REFERENCES

[1] GLOBOCAN, "Estimated Cancer Incidence, Mortality and Prevalence Worldwide in 2012," *Section of Cancer Surveillance*. pp. 1–7, 2012, doi: 10.1111/j.1440-1754.2007.01232.x.
[2] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal, "Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries," *CA. Cancer J. Clin.*, 2018.
[3] Ministry of Health Malaysia, "Early Detection of Common Cancers And Referral Pathways: Module for Health Care Providers," *MOH, Module 2017*, 2017. [Online]. Available: http://www.moh.gov.my/resources/index/Penerbitan/Rujukan/NCD/Early_Detection_Of_Common_Cancers_And_Referal_Pathways_Module_For_Health_Care_Providers.pdf. [Accessed: 16-Nov-2018].
[4] D. R. Youlden, S. M. Cramb, N. A. M. Dunn, J. M. Muller, C. M. Pyke, and P. D. Baade, "The descriptive epidemiology of female breast cancer: an international comparison of screening, incidence, survival and mortality," *Cancer Epidemiol.*, vol. 36, no. 3, pp. 237–248, 2012.
[5] F. Ahmad, N. A. M. Isa, M. H. M. Noor, and Z. Hussain, "Intelligent breast cancer diagnosis using hybrid GA-ANN," in *2013 Fifth International Conference on Computational Intelligence, Communication Systems and Networks*, 2013, pp. 9–12.
[6] S. Bauer, R. Wiest, L.-P. Nolte, and M. Reyes, "A survey of MRI-based medical image analysis for brain tumor studies," *Phys. Med. Biol.*, vol. 58, no. 13, p. R97, 2013.
[7] S. H. Heywang-Köbrunner, A. Hacker, and S. Sedlacek, "Advantages and disadvantages of mammography screening," *Breast care*, vol. 6, no. 3, pp. 199–207, 2011.
[8] J. Nahar, K. S. Tickle, A. B. M. S. Ali, and Y.-P. P. Chen, "Early Breast Cancer Identification: Which Way to Go? Microarray or Image Based Computer Aided Diagnosis!," in *Network and System Security, 2009. NSS'09. Third International Conference on*, 2009, pp. 456–461.
[9] T. Onega *et al.*, "Facility mammography volume in relation to breast cancer screening outcomes," *J. Med. Screen.*, vol. 23, no. 1, pp. 31–37, 2016.
[10] F. Han, W. Sun, and Q.-H. Ling, "A novel strategy for gene selection of microarray data based on gene-to-class sensitivity information," *PLoS One*, vol. 9, no. 5, p. e97530, 2014.
[11] O. A. Alomari, A. T. Khader, M. A. Al-Betar, and Z. A. A. Alyasseri, "A Hybrid Filter-Wrapper Gene Selection Method for Cancer Classification," in *2018 2nd International Conference on BioSignal Analysis, Processing and Systems (ICBAPS)*, 2018, pp. 113–118.
[12] S. Turgut, M. Dağtekin, and T. Ensari, "Microarray breast cancer data classification using machine learning methods," in *2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT)*, 2018, pp. 1–3.
[13] M. Dashtban and M. Balafar, "Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts," *Genomics*, vol. 109, no. 2, pp. 91–107, 2017.
[14] S. Kar, K. Das Sharma, and M. Maitra, "Gene selection from microarray gene expression data for classification of cancer subgroups employing PSO and adaptive K-nearest neighborhood technique," *Expert Syst. Appl.*, vol. 42, pp. 612–627, 2015.
[15] M. S. Mohamad, S. Omatu, S. Deris, and M. Yoshioka, "A modified binary particle swarm optimization for selecting the small subset of informative genes from gene expression data," *IEEE Trans. Inf. Technol. Biomed.*, vol. 15, no. 6, pp. 813–822, 2011.
[16] S. Shahbeig, M. S. Helfroush, and A. Rahideh, "A fuzzy multi-objective hybrid TLBO–PSO approach to select the associated genes with breast cancer," *Signal Processing*, vol. 131, pp. 58–65, 2017.
[17] T. P. Vital, M. M. Krishna, G. V. L. Narayana, P. Suneel, and P. Ramarao, "Empirical Analysis on Cancer Dataset with Machine Learning Algorithms," in *Soft Computing in Data Analytics*, Springer, 2019, pp. 789–801.
[18] H. Aydadenta, "On the classification techniques in data mining for microarray data classification," in *Journal of Physics: Conference Series*, 2018, vol. 971, no. 1, p. 12004.
[19] D. W. Mount *et al.*, "Using logistic regression to improve the prognostic value of microarray gene expression data sets: application to early-stage squamous cell carcinoma of the lung and triple negative breast carcinoma," *BMC Med. Genomics*, vol. 7, no. 1, p. 33, 2014.
[20] N. Almugren and H. Alshamlan, "FF-SVM: New FireFly-based gene selection algorithm for microarray cancer classification," in *2019 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*, 2019, pp. 1–6.
[21] M. B. Kursa and W. R. Rudnicki, "Feature selection with the Boruta package," *J Stat Softw*, vol. 36, no. 11, pp. 1–13, 2010.
[22] R. Tibshirani, "Regression shrinkage and selection via the lasso," *J. R. Stat. Soc. Ser. B*, vol. 58, no. 1, pp. 267–288, 1996.
[23] M. B. Kursa, "Robustness of Random Forest-based gene selection methods," *BMC Bioinformatics*, vol. 15, no. 1, p. 8, 2014.
[24] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, 1995.
[25] A. Statnikov, L. Wang, and C. F. Aliferis, "A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification," *BMC Bioinformatics*, vol. 9, no. 1, p. 319, 2008.
[26] J. Berkson, "Application of the logistic function to bio-assay," *J. Am. Stat. Assoc.*, vol. 39, no. 227, pp. 357–365, 1944.

BIOGRAPHIES OF AUTHORS

Nursabillilah Mohd Ali was born in Melaka, Malaysia, in 1985. She received the B.Eng. (Hons.) and M.Sc. degrees from Universiti Teknikal Malaysia Melaka and International Islamic University, Malaysia in 2009 and 2014, respectively, all in mechatronic engineering. She has been an academic staff since 2009, where now she is a Senior Lecturer of Universiti Teknikal Malaysia Melaka. She is a Chartered Engineer of the Engineering Council UK and a Graduate Engineer of the Board of Engineers Malaysia. Currently, she is working toward the Ph.D. degree at the Multimedia University. Her research interests include bioinformatics system, DNA gene expression, optimization algorithm and machine learning.

Nor Azlina Ab Aziz received her Ph.D. degree from University of Malaya, Malaysia. She is currently a senior lecturer with the Faculty of Engineering and Technology, Multimedia University, Melaka, Malaysia. She is also the chairperson of the Center for Engineering Computational Intelligence, Multimedia University. She has published and presented numerous scientific papers in international journals and conferences, and lead multiple research projects. Her research interests include the fundamental aspects and applications of computational intelligence in wireless communication, bioinformatics, operational research and affective computing.

Rosli Besar is currently Associate Professor of Multimedia University. He received the B.Eng. (Hons) and M.Sc. degrees from the University of Science Malaysia (USM), Malaysia, in 1990 and 1993, respectively and the Ph.D. degree from the Multimedia University, Malaysia, in 2004. His current interests include Signal and Image Processing and Medical Imaging.