SEMARAK ILMU PUBLISHING

Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Classification of Breast Cancer Subtypes using Microarray RNA Expression Data

Muhammad Shazwan Suhiman[1], Sayang Mohd Deni[1,*], Ahamd Zia Ul-Saufie Mohamad Japeri[2], Aszila Asmat[2], Lirong Wang[3]

[1] Mathematical Sciences Studies, College of Computing, Informatics, and Mathematics, Universiti Teknologi MARA (UiTM), 40450 Shah Alam, Selangor, Malaysia
[2] D Mathematical Sciences Studies, College of Computing, Informatics, and Media, Universiti Teknologi MARA (UiTM), Pahang Campus, 27600, Raub, Pahang, Malaysia
[3] School of Mathematics and Finance, Science and Technology, hunan University of Humanities, Science and Technology, Loudi, 417000, P.R.China

ARTICLE INFO

Article history:
Received 2 June 2023
Received in revised form 23 November 2023
Accepted 9 April 2024
Available online 25 May 2024

Keywords:

Breast cancer classification; Feature selections; Machine learning

ABSTRACT

Breast cancer is a heterogeneous disease that involves molecular alteration, cellular alterations, and clinical outcome for which the classification of Breast cancer remains a challenge to diagnose. Current practice uses immunohistochemistry markers and clinical variables to classify Breast cancer, but this approach has limitations due to the inclusion of other tumour subtypes and healthy individuals. Machine learning approaches based on mRNA expression data offer new possibilities for researchers to investigate the potential of molecular biomarkers as one of the diagnostic characteristics. The purpose of this study is to evaluate features (genes) rank through feature selection method for Breast cancer diagnostic test. Three feature selection methods of IG, relief and mRMR were applied and subsets of top 100, 50, 25, 10, 5 and 3 were created. Each subset was tested with SVM, LR and RF classifiers and its performance was assessed using confusion matrix. The result of this study found that the feature selection of IG, reliefF and mRMR was able to achieve highest accuracy with SVM, LR and RF classifier. mRMR with RF classifier achieved highest accuracy with the least number of top rank genes with 25 genes. Hybrid feature selection approached (mRMR + SVM) improved accuracy of top 3 highest rank genes using SVM, LR and RF classifier. Future work should aim to use other feature selection methods and classifiers to explore the classification accuracy with the least features subset in multiclass cancer dataset.

1. Introduction

Breast cancer is notable as one of the leading causes of death among women with 14% of cancer deaths worldwide [1]. Breast cancer, like most other types of cancer is known as a heterogeneous disease, a disease which caused by different genes and alleles. Current practice uses classical

* *Corresponding author.*
E-mail address: sayan929@uitm.edu.my

https://doi.org/10.37934/araset.46.1.7585

immunohistochemistry markers (IHC) and traditional clinicopathological variables such as 2 tumour size and grade, Estrogenic-receptor (ER), Progesterone receptor (PR), Human Epidermal growth factor receptor 2 (HER2) to subdivide this disease classification. This practice helps to explain some of the complexities of Breast cancer for a better prognosis [2]. However, this conventional approach has significant limitation as it also identified in other tumour subtypes and even within healthy person who are under stress [3].

The advancement of global gene expression profiling improves Breast cancer classification by unveiling the distinct intrinsic molecular characteristics apart from commonly used clinical-pathological variables. There is an urgency in using genomic stratification in Breast cancer, though the molecular heterogeneity is a well-recognised characteristic in Breast cancer, but it is poorly considered in current clinical setting. Some biomarkers may have shown potential to detect certain specific types of Breast cancer, thus finding new biomarkers to different types of Breast cancer to predict prognosis should be done on a larger scale. Ribonucleic acid (RNA) has been one of the reliable diagnostic indicators for cancer and numerous studies have mentioned the association of messenger RNA (mRNA) with Breast cancer. A recent study by Gera *et al.,* [4] found an association between Cyclin-dependent kinase 2-associated protein 1 (CDK2AP1) with Breast cancer cell where its expression was 38 folds lower than of adjacent non-cancerous cells. In contrast, CDK2AP1 levels were three times higher in disease-free patients at 10 years than in patients who died of Breast cancer.

The advent of microarray data from public database allows us to monitor thousands of different gene expression levels simultaneously that help in disease diagnostic, particularly on Breast cancer. There are challenges to handle all features in a microarray dataset as it requires a lot of time in terms of rendering and analysing. This was when Machine learning application started in biomedical research since it enables analysis to be done in a short period. In recent years, Support Vector Machine (SVM) has been among the top machine learning used in predicting mRNA expression to classified Breast cancer. SVM already gained it attention as one of the reliable machine learning classifiers and also for feature selection in gene expression dataset [5]. In a study where SVM classifier were chose together with feature selections of mRMR on RNA seq cancer data resulted an accuracy of 0.751 in comparing between people with cancer and healthy person [6]. SVM was the only classifier chose for a hybrid method feature selections where an accuracy of 100% was recorded [7]. Another study conducted by Kim *et al.,* [8] using logistic regression found that this classifier was able to perform well in the classification of multiclass gene expression data. On the other hand, Random Forest was mentioned as the alternative reliable machine learning in related to microarray data [5]. Study on feature selection Lasso or Relief on RF found an accuracy of 85.6% [9]. A recent finding of using hybrid anova and lasso methods for feature selection of microarray data and testing on spark environment denoted that RF perform best with 100% in two dataset and 96% in one dataset in a less time consumed [10]. Other applications of machine learning approach particularly on predictive models could be found in various field of studies including intervention and prevention of diabetic retinopathy [11], diagnosis of Alzheimer's disease [12], prediction of dengue outbreaks [13] and heart failure [14], prediction and classification on future PM_{10} concentrations [15], forecasting reservoir water level [16], forecasting daily sales data [17] and rainfall prediction in flood prone areas [18] and brain MRI image classification for Alzheimer's disease [19].

However, microarray data have apparent characteristics of high-dimensional features with respect to small sample size lead classifier to overfit, therefore, one of the effective methods to solve this situation is by using feature selection methods [20]. Feature selections help to filter out the features that redundant and does not give value in improving prediction and presenting only important distinct features that carry high weightage that will increase accuracy in classification [21]. Previous study showed that Binary dataset have shown a tremendous accuracy with Relief, Lasso [9].

Further research explored to find the very least features but still having high accuracy, and a study reported three features enough to attained high accuracy (100%) using feature selection-machine learning (FS-ML) [7]. A multiclass dataset would give a different challenge, especially if the selected dataset has features higher than 50000 with 6 subtypes. Past study had reported the using of PCA which end up with low accuracy due to loss of information [22]. To solve this problem, potential feature selections method like mRMR and reliefF and a hybrid method were proposed together with classifier that can handle with multiclass dataset. SVM, Naïve Bayes (NB), k-NN and Decision Tree (DT) machine learning classifier was previously reported in other study involving multiclass dataset with much lower features [23].

This study focuses entirely on RNA as a dependent variable in Breast cancer since it has been recognized as a promising biomarker in cancer treatment. The sophisticated approach of machine learning can shed light on how RNA relates to Breast cancer that can be discovered reliably and can aid medical practitioners in cancer management. In addition, this study may benefit the pharmaceutical industry in developing miRNA-based therapy, which has the potential to be the next game changer for cancer treatment. Thus, the goal of this study is to develop and evaluate predictive models that can classify Breast cancer subtypes using RNA-microarray data, which can enhance existing diagnosis and aid in therapy options.

2. Methodology

2.1 Dataset

Data on Breast cancer gene expression was downloaded from Curated Microarray Database (CuMiDa) which consists of 151 samples [24]. The target variable is identified as the type of Breast cancer in group classifier which means the desired prediction would be either group classification of Basal, HER, Luminal A, Luminal B, HER, Normal or Cell_line. Of all the samples, 33 were HER subtype, 41 were basal subtype, 14 were cell line subtype, 29 were luminal A, 30 luminal D and 7 were normal subtype. The input variable is distinct mRNA which consists of 54677 genes or features level. The approach to achieve the objective of the study is shown in Figure 1.

At the first stage, three feature selection methods of Information Gain (IG), reliefF and Maximum Relevance Minimum Redundancy (mRMR) together with ranker evaluator were applied to make a feature selection through individual feature score. A few subsets from each feature selection methods' top ranked features were created as followed, top 100, top 50, top 25, top 10, top 5 and top 3. After that, each subset will be evaluated with three different classifier which were Support Vector Machine (SVM), Logistic Regression (LR) and Random Forest (RF) with 10 folds cross validation. The performance of classifier will be evaluated using classification confusion matrix. The best feature selection methods will then continue with another set of experiment on testing feature selection using SVM hybrid method for stage 2. A subset of top 10, top 5 and top 3 were created and then validated with the same three classifier with 10 folds cross validation. Finally, the performance of each classifier will be evaluated using classification confusion matrix.

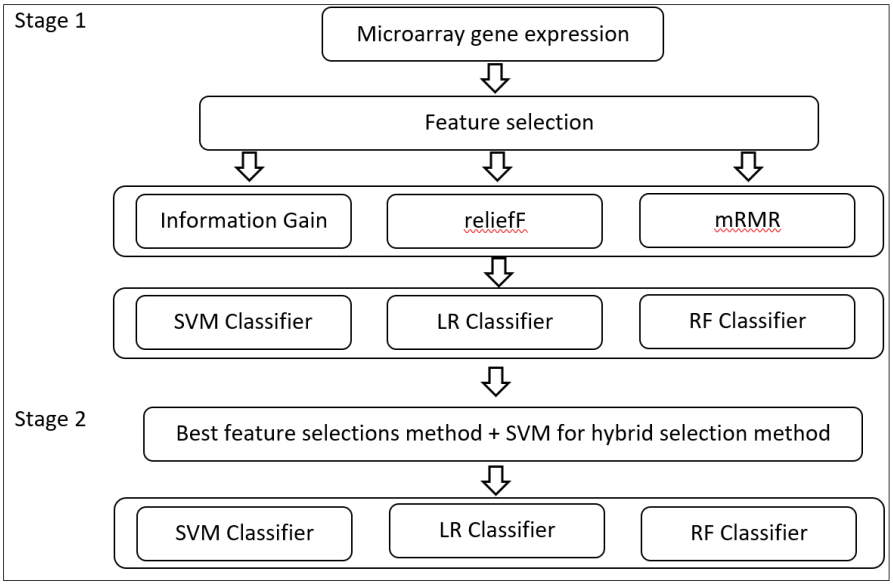


Fig. 1. Breast cancer Classification Procedure

2.2 Feature Selection

Feature selection is an important method in removing features that obviously irrelevant and redundant. Since microarray dataset have high dimension of features, feature selection become mandatory in this study. There are four feature selection methods applied to the dataset namely Information Gain (IG), reliefF and maximum Relevance Minimum Redundancy (mRMR).

Multiple studies reported using IG in their study [20,23,25]. IG can be identified as a classical popular feature selection in microarray data as it is reliable and less time consuming. The differences between entropy and conditional entropy, which show the reduction of uncertainty, can be used to assess the relevance of genes in a certain category as indicated in the Eq. (1).

$$g(Y,X) = H(y) - H(Y|X) \tag{1}$$

where $H(Y|X)$ signifies the conditional entropy, which depicts the uncertainty based on the known variable, and $H(y)$ means the entropy of dataset Y, which quantifies the uncertainty associated in predicting the value of a random variable.

Maximum Relevance Minimum Redundancy (mRMR) method investigates the correlations between features and target according to mutual information. Max Relevance and Min-Redundancy are two criteria that used in the mRMR method strategy. This method will evaluate based on its relevance to target and its redundancy to other features, this will then generate two feature lists which are MaxRel feature list and mRMR feature list. The feature lists are ranked according to their importance to target. This can be read by the Mutual Information measure (MI) value, a higher MI value indicates a strong correlation [6]. MI is defined by:

$$I(x,y) = \int\int p(x,y)\log\frac{p(x,y)}{p(x)p(y)}\,dx\,dy \tag{2}$$

where $p(x,y)$ represents the joint probabilistic density, and $p(x)$ and $p(y)$ represent the marginal probabilistic densities. mRMR has been reported increase accuracy with both binary and multiclass data when using hybrid method of mRMR and Artificial Bee Colony (ABC) [26].

On the other hand, ReliefF was designed to deal with multiclass dataset other than incomplete and noisy dataset, in feature selecting of the high dimensional dataset [23]. ReliefF focusing on optimal gene subsets that can differentiate instances that close with each other. This is done by calculating the quality of features that have weights greater than the threshold using the distinction of a feature value between a given instance and the two nearest instances, namely Hit and Miss which later will update the quality estimation on W_i.

$$W_i = W_i + \frac{\sum_{k=1}^{k} D_H}{n_k} + \sum_{c=1}^{c-1} P_c \frac{\sum_{k=1}^{k} D_{MC}}{n_k} \quad (3)$$

where n_k is the number of instances in class k, D_H (or D_{MC} is the sum of distance between the selected instance and each) H (or Mc), P_c is the probability of class c [27].

2.3 Predictive Modelling

All the subsets stated at 2.1 were tested with three machine learnings model which were Support Vector Machine (LibSVM), Logistic regression (LR) and Random Forest (RF). All the models will be validated in WEKA. A 10 folds cross validation was set in each of the run [28].

SVM is a supervised machine learning that categories binary class problem into multiple classes of labelled training data by binarization technique [29]. SVMs used hyperplane boundary to separate between classes, with first will be assumed as linear models. The separation of when determining the hyperplanes depend on the notion of margin. A clean separation between two classes indicates by maximum margin hyperplane that marked by none training data points in it and large margin does not exist on each side. In the case of miRNA, the solution in maximizing margin size was by support vectors in both classes which with b is the bias on the equation will dictates the shift in the hyperplane boundary. Feature vectors (x_i, y_i), that consist in miRNA dataset is shown in Eq. (4) where yi ∈ {+1, −1}.

$$y_i(wx_i + b) - 1 \geq 0 \quad (4)$$

Logistic regression technique was used when the dependent variable is categorical. Mainly used for binary classification where in microarray data, let y would be an array of binary disease status (1 for cancer and 0 for normal). Let $x = x_{j1}, x_{j2}, \ldots, x_{jn-1}$ expression vector, where x_j is the expression level of the j^{th} gene. Building a logistic prediction model with microarray data, on the other hand, is fundamentally different from normal logistic modelling since the number of genes (predictors) p can be thousands while the number of arrays (subjects) n is often less than 100 [30]. Multiclass classification was implemented using the same principles as binary classification.

On the other hand, Random Forest classifier consists of a large number of decision trees where the new sample classification was made depending on the voting by each of the individual decision trees which known also as ensemble learning method. The Random Forest was constructed by the following manner. Each decision trees will be selected n samples at random, certain features also been selected at random, where the best split of features 26 based on, for example Information Gain is used as the binary split on that node which repeated until the predefined minimum node size is reached. Later, the new data classification was done by aggregating all the decision trees predictions on look on the majority vote [20].

2.4 Performance Measure

Results from each classifier was evaluated using confusion matrix as shown in Table 1. True positive (TP) is the number of the correctly predicted positive samples, False positives (FP) is the number of the rest of the positive samples which incorrectly predicted. On the other hand, True False (TF) defines as the number of the correctly predicted negative samples and False negative (FN) is the number of the rest of the negative samples which incorrectly predicted.

Table 1
Confusion Matrix

Predicted	Actual	
	Y=0 (Negative)	Y=1 (Positive)
Y=0 (Negative)	True Negative (TN)	False Positive (FP)
Y=1 (Positive)	False Negative (FN)	True Positive (TP)

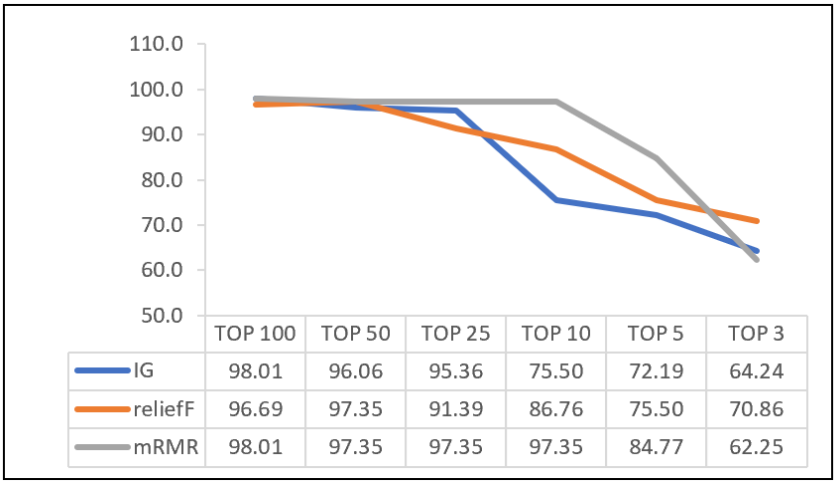
Then, classification performance was measured using classification accuracy (CA), which is defined as the ability of each classifier to make the correct decision divided by the total number of the data [31]. This measure is suitable for multiclassification performance, and it is computed according to Eq. (5).

$$\mathrm{CA} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \tag{5}$$

3. Results

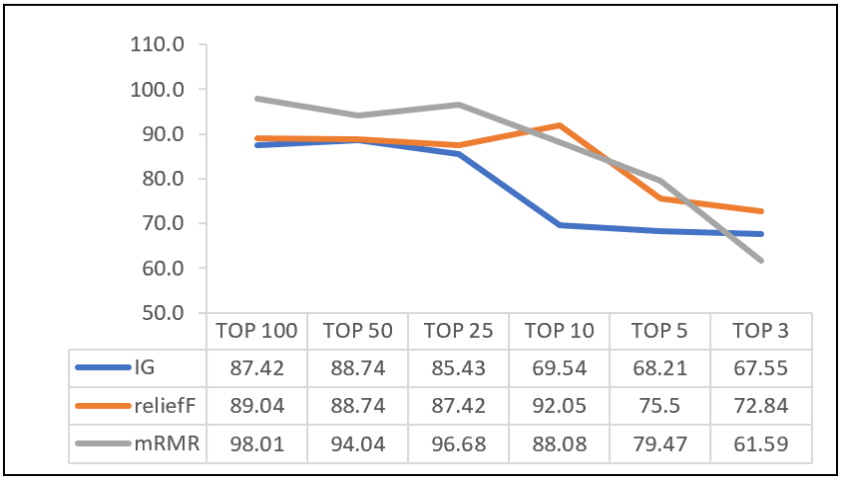
At the first stage, three feature selection methods were applied to dataset which included IG, ReliefF and mRMR. Features were ranked based on significance value, where in this study subsets of top 100, 50, 25, 10, 5 and 3 were used. All these subsets will then be tested with three machine learning classifiers which were Support Vector Machine (SVM), Logistic regression (LR) and Random Forest (RF) and its performance are shown in Figure 2 to Figure 4 respectively.

The performance of feature selection methods in three different classifiers showed best at around top 100, top 50, and 25. The performance later drops clearly on top 10 and 5 and perform worst at top 3. mRMR feature selection performance was consistently high for the top 100, 50, 25, and 10 for all three classifiers given the accuracy is within 88.08% to 98.01%. ReliefF performance for all three algorithms showing the same pattern, of maintaining accuracy for top 100 and 50 before decreasing on top 25 and 10 but perform the best for the top 3 for the SVM and LR algorithms. IG performance varied between three algorithms. It performs the best in RF with the accuracy within 95% for the top 100, 50 and 25, and 10. However, with SVM validation, it only recorded accuracy around 95 % with its top 100 and top 50 subsets. IG performs worst in LR with its highest accuracy only 88.74%. Thus, from this section, mRMR was identified as the best feature selections due to its high accuracy in all the classifiers and it recorded the least number of features for the highest accuracy achieved. A study by Dashtban and Balafar [32] on small round blue cell tumour (SRBCT) which was the same type of multiclass microarray data found a 100% accuracy achieved with Naïve Bayes classifier using top 50 genes subset with feature selection methods of fisher-score and Laplacian-score. However, the dataset was lacked due to a relatively small number of original features with only 2308 genes. A microarray dataset with less than 10000 was questioned on its stability as criticized in a review paper conducted by Hambali *et al.*, [21].



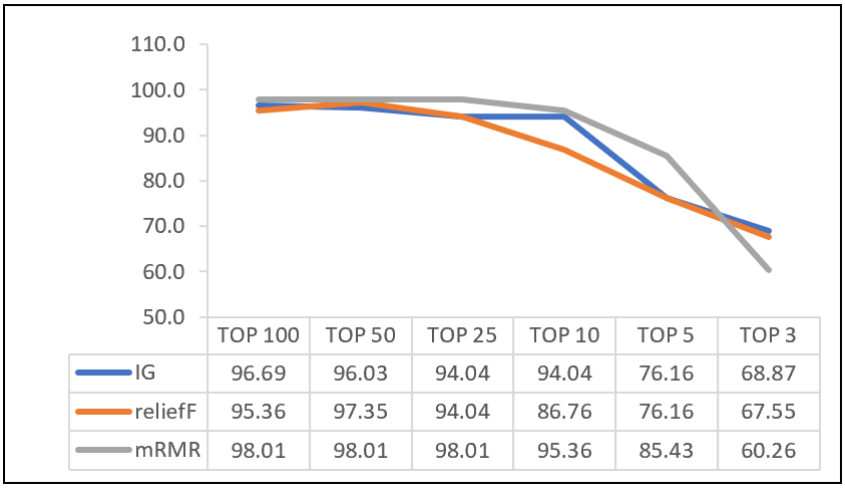
	TOP 100	TOP 50	TOP 25	TOP 10	TOP 5	TOP 3
IG	98.01	96.06	95.36	75.50	72.19	64.24
reliefF	96.69	97.35	91.39	86.76	75.50	70.86
mRMR	98.01	97.35	97.35	97.35	84.77	62.25

Fig. 2. Performance of top genes with SVM



	TOP 100	TOP 50	TOP 25	TOP 10	TOP 5	TOP 3
IG	87.42	88.74	85.43	69.54	68.21	67.55
reliefF	89.04	88.74	87.42	92.05	75.5	72.84
mRMR	98.01	94.04	96.68	88.08	79.47	61.59

Fig. 3. Performance of top genes with Logistic Regression



	TOP 100	TOP 50	TOP 25	TOP 10	TOP 5	TOP 3
IG	96.69	96.03	94.04	94.04	76.16	68.87
reliefF	95.36	97.35	94.04	86.76	76.16	67.55
mRMR	98.01	98.01	98.01	95.36	85.43	60.26

Fig. 4. Performance of top genes with Random Forest

The analysis continued with testing feature selection using SVM hybrid method for stage 2. The best feature selection method of mRMR will be combined with SVM feature evaluator method before being validated with the same three classifiers. In this section, only subset with top 3 genes of hybrid (mRMR+SVM) was tested, as we aimed to get high accuracy in lowest number possible of features

used. The feature selected were 226961_at (SPARC like 1(SPARCL1), 210930_s_at (erb-b2 receptor tyrosine kinase 2(ERBB2) and 200795_at (proline rich 15(PRR15)). Top 3 features were used according to the report, by using this method of FS+SVM, they will be able to get high accuracy using only 3 features in binary dataset [7]. The confusion matrix of each classifier was presented in Table 2.

Table 2
Confusion Matrix for SVM, Logistic Regression and Random Forest

Classifier	Confusion Matrix							Classification
		a	b	c	d	e	f	
SVM	a	38	1	0	0	0	2	Correctly classified = 138/151
	b	0	27	0	0	0	3	Incorrectly classified = 13/151
	c	0	0	14	0	0	0	Percentage = 91.39%
	d	0	0	0	7	0	0	
	e	0	0	0	0	28	1	
	f	1	3	0	1	1	24	
Logistic Regression	a	38	0	0	1	0	2	Correctly classified = 130/151
	b	1	23	0	0	0	6	Incorrectly classified = 21/151
	c	1	0	13	0	0	0	Percentage = 86.09%
	d	0	0	0	7	0	0	
	e	0	0	0	0	27	2	
	f	2	4	0	0	2	22	
Random Forest	a	39	0	0	1	0	1	Correctly classified = 133/151
	b	1	25	0	0	0	4	Incorrectly classified = 18/151
	c	0	0	13	0	0	1	Percentage = 88.08%
	d	0	0	0	7	0	0	
	e	0	0	0	0	27	2	
	f	2	4	0	1	1	22	

Note: The letters a, b, c, d, e and f represent basal, HER, cell_line, normal, luminal_A and luminal B respectively

The performance of each classifier using hybrid method of mRMR and SVM (mRMR+SVM) then were compared with performance when using feature selection method of mRMR only and the results were shown in Table 3.

Table 3
Performance of top genes hybrid method vs filter-only method

	mRMR+SVM	mRMR
SVM	91.39	62.25
Logistic Regression	86.09	61.59
Random Forest	88.07	60.26

According to Table 3, overall hybrid method of mRMR+SVM improved its top 3 accuracies in SVM classifier to 91.39% from 62.25%. The hybrid method also improved top 3 accuracies to 86.09% from 62.25% when using LR classifier, and the same occurred when validated with Random Forest classifier where top 3 accuracies improved from 88.07% to 60.26%. In summary, hybrid method of mRMR+SVM abled to increase top 3 features subset accuracy significantly, with SVM classifier was the highest accuracy with 91.39%. By checking on DAVID repository platform by Huang *et al.,* [33], a powerful free public online bioinformatics resources that provide functional interpretation of large list of genes derived from studies including microarray, found out that three of the top ten feature

genes from ReliefF (erb-b2 receptor tyrosine kinase 2(ERBB2), estrogen receptor 1(ESR1) and forkhead box A1(FOXA1)), and four from mRMR (CD93 molecule(CD93), erb-b2 receptor tyrosine kinase 2(ERBB2), estrogen receptor 1(ESR1), migration and invasion enhancer 1(MIEN1)) were directly correlated with the Breast cancer disease. It suggests the feature rank of feature selections has a meaningful correlation to the real biological function which has higher chances to act as the biomarker for the disease.

4. Conclusions

Finding from this study showed all three features selection methods of IG, ReliefF and mRMR has abled to help classifiers achieved a high cross validation accuracy with much lower features. This solved the problem on to handling high-dimensional features of this dataset. Observation from the result clearly indicates that ReliefF and mRMR performance were better than IG where IG scored lowest in all subsets tested except top 3 in all classifiers as compared to ReliefF and mRMR. Unlike mRMR and ReliefF, IG classic method ignore the features dependencies that lead to its bad performance. The characteristic of the mRMR algorithm that chose features based on mutual information of maximum-relevance to the target and minimum redundancy has performed the best in term of consistent highest accuracy with lowest number of features as compared to IG and ReliefF. This signifies that adequate information of the whole dataset features was greatly represent by the top rank features although it performed lowest on its top 3 accuracy. However, hybrid method chosen (mRMR+SVM) has helped to boost accuracy on the top 3 rank indicates that the hybrid method works in a sense that after filter method of mRMR removed irrelevant features, the latter feature selection method of SVM managed to find the best features subset from the remaining features pool. SVM has been agreed by many literatures works incredibly well with small datasets and features and abled to separate different multiclass through its hyperplanes.

There are multiple feature selection techniques that can be used for further study which include wrapper, embedded and other hybrid methods that have yet to receive much attention to be tested in multiclass microarray type dataset. In addition, one could use other classifiers to explore the classification accuracy with the least number of features in multiclass cancer dataset. Through accurate classification, a reliable mRNA biomarker can be found and will benefit medical practitioner in cancer management, where quick, accurate and cost-effective diagnostic can be done. Thus, the end goal is that, hopefully this study may help to increase the likelihood of Breast cancer patient surviving chance.

Acknowledgement

The authors wish to thank to Universiti Teknologi Mara (UiTM) for giving the opportunity to complete part of this research project under Master of Data Science program. They also acknowledge their sincere appreciation to the reviewers for their valuable suggestion and remarks to improve the manuscript.

References

[1] Yerukala Sathipati, Srinivasulu, and Shinn-Ying Ho. "Identifying a miRNA signature for predicting the stage of breast cancer." *Scientific reports* 8, no. 1 (2018): 16138. https://doi.org/10.1038/s41598-018-34604-3

[2] Sherafatian, Masih. "Tree-based machine learning algorithms identified minimal set of miRNA biomarkers for breast cancer diagnosis and molecular subtyping." *Gene* 677 (2018): 111-118. https://doi.org/10.1016/j.gene.2018.07.057

[3] Chen, Lei, Tao Zeng, Xiaoyong Pan, Yu-Hang Zhang, Tao Huang, and Yu-Dong Cai. "Identifying methylation pattern and genes associated with breast cancer subtypes." *International journal of molecular sciences* 20, no. 17 (2019): 4269. https://doi.org/10.3390/ijms20174269

[4] Gera, Ritika, Leon Mokbel, Wen G. Jiang, and Kefah Mokbel. "mRNA expression of CDK2AP1 in human breast cancer: correlation with clinical and pathological parameters." *Cancer genomics & proteomics* 15, no. 6 (2018): 447-452. https://doi.org/10.21873/cgp.20103

[5] George, G., and V. Cyril Raj. "Review on feature selection techniques and the impact of SVM for cancer classification using gene expression profile." *arXiv preprint arXiv:1109.1062* (2011).

[6] Zhang, Yu-Hang, Tao Huang, Lei Chen, YaoChen Xu, Yu Hu, Lan-Dian Hu, Yudong Cai, and Xiangyin Kong. "Identifying and analyzing different cancer subtypes using RNA-seq data of blood platelets." *Oncotarget* 8, no. 50 (2017): 87494. https://doi.org/10.18632/oncotarget.20903

[7] Gao, Lingyun, Mingquan Ye, Xiaojie Lu, and Daobin Huang. "Hybrid method based on information gain and support vector machine for gene selection in cancer classification." *Genomics, Proteomics and Bioinformatics* 15, no. 6 (2017): 389-395. https://doi.org/10.1016/j.gpb.2017.08.002

[8] Kim, Yongdai, Sunghoon Kwon, and Seuck Heun Song. "Multiclass sparse logistic regression for classification of multiple cancer types using gene expression data." *Computational Statistics & Data Analysis* 51, no. 3 (2006): 1643-1655. https://doi.org/10.1016/j.csda.2006.06.007

[9] Güçkiran, Kıvanç, İsmail Cantürk, And Lale Özyilmaz. "DNA microarray gene expression data classification using SVM, MLP, and RF with feature selection methods relief and LASSO." *Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi* 23, no. 1 (2019): 126-132. https://doi.org/10.19113/sdufenbed.453462

[10] Albaldawi, Wafaa S., and Rafah M. Almuttairi. "Hybrid ANOVA and LASSO methods for feature selection and linear support vector, multilayer perceptron and random forest classifiers based on spark environment for microarray data classification." In *IOP Conference Series: Materials Science and Engineering*, vol. 1094, no. 1, p. 012107. IOP Publishing, 2021. https://doi.org/10.1088/1757-899X/1094/1/012107

[11] Khairudin, Zuraida, Nurfatin Adila Abdul Razak, Hezlin Aryani Abd Rahman, Norbaizura Kamarudin, and Nor Azimah Abd Aziz. "Prediction of diabetic retinopathy among type II diabetic patients using data mining techniques." *Malaysian Journal of Computing (MJoC)* 5, no. 2 (2020): 572-586. https://doi.org/10.24191/mjoc.v5i2.10554

[12] Abdullah, Mohammad Nasir, Yap Bee Wah, AB Abdul Majeed, Yuslina Zakaria, and Norshahida Shaadan. "Identification of blood-based multi-omics biomarkers for alzheimer's disease using firth's logistic regression." *vol* 30 (2022): 1197-1218. https://doi.org/10.47836/pjst.30.2.19

[13] Salim, Nurul Azam Mohd, Yap Bee Wah, Caitlynn Reeves, Madison Smith, Wan Fairos Wan Yaacob, Rose Nani Mudin, Rahmat Dapari, Nik Nur Fatin Fatihah Sapri, and Ubydul Haque. "Prediction of dengue outbreak in Selangor Malaysia using machine learning techniques." *Scientific reports* 11, no. 1 (2021): 939. https://doi.org/10.1038/s41598-020-79193-2

[14] Mansur Huang, Nur Shahellin, Zaidah Ibrahim, and Norizan Mat Diah. "Machine learning techniques for early heart failure prediction." *Malaysian Journal of Computing (MJoC)* 6, no. 2 (2021): 872-884. https://doi.org/10.24191/mjoc.v6i2.13708

[15] Shaziayani, Wan Nur, Ahmad Zia Ul-Saufie, Sofianita Mutalib, Norazian Mohamad Noor, and Nazatul Syadia Zainordin. "Classification prediction of PM10 concentration using a tree-based machine learning approach." *Atmosphere* 13, no. 4 (2022): 538. https://doi.org/10.3390/atmos13040538

[16] Aquil, Mohammad Amimul Ihsan, and Wan Hussain Wan Ishak. "Comparison of Machine Learning Models in Forecasting Reservoir Water Level." *Journal of Advanced Research in Applied Sciences and Engineering Technology* 31, no. 3 (2023): 137-144. https://doi.org/10.37934/araset.31.3.137144

[17] Amir, Wan Khairul Hazim Wan Khairul, Afiqah Bazlla Md Soom, Aisyah Mat Jasin, Juhaida Ismail, and Aszila Asmat. "Sales Forecasting Using Convolution Neural Network." *Journal of Advanced Research in Applied Sciences and Engineering Technology* 30, no. 3 (2023): 290-301. https://doi.org/10.37934/araset.30.3.290301

[18] Ramlan, Siti Zuhairah, and Sayang Mohd Deni. "Rainfall prediction in flood prone area using deep learning approach." In *Soft Computing in Data Science: 6th International Conference, SCDS 2021, Virtual Event, November 2–3, 2021, Proceedings 6*, pp. 71-88. Springer Singapore, 2021. https://doi.org/10.1007/978-981-16-7334-4_6

[19] Khaw Li Wen, Shahrum Shah Abdullah. "Mri Brain Image Classification Using Convolutional Neural Networks and Transfer Learning." *Journal of Advanced Research in Computing and Applications* 31, no. 1(2023): 20-26. https://www.akademiabaru.com/submit/index.php/arca/article/view/5210

[20] Rehman, Oneeb, Hanqi Zhuang, Ali Muhamed Ali, Ali Ibrahim, and Zhongwei Li. "Validation of miRNAs as breast cancer biomarkers with a machine learning approach." *Cancers* 11, no. 3 (2019): 431. https://doi.org/10.3390/cancers11030431

[21] Hambali, Moshood A., Tinuke O. Oladele, and Kayode S. Adewole. "Microarray cancer feature selection: Review, challenges and research directions." *International Journal of Cognitive Computing in Engineering* 1 (2020): 78-97. https://doi.org/10.1016/j.ijcce.2020.11.001

[22] Dang, Thuy Hang, Dung Pham Trung, Hoai Linh Tran, and Quang Le Van. "Using dimension reduction with feature selection to enhance accuracy of tumor classification." In *2016 International Conference on Biomedical Engineering (BME-HUST)*, pp. 14-17. IEEE, 2016. https://doi.org/10.1109/BME-HUST.2016.7782082

[23] Shukla, Alok Kumar, and Diwakar Tripathi. "Identification of potential biomarkers on microarray data using distributed gene selection approach." *Mathematical biosciences* 315 (2019): 108230. https://doi.org/10.1016/j.mbs.2019.108230

[24] Feltes, Bruno Cesar, Eduardo Bassani Chandelier, Bruno Iochins Grisci, and Márcio Dorn. "Cumida: An extensively curated microarray database for benchmarking and testing of machine learning approaches in cancer research." *Journal of Computational Biology* 26, no. 4 (2019): 376-386. https://doi.org/10.1089/cmb.2018.0238

[25] Mazlan, Umi Hanim, and Puteh Saad. "Classification of breast cancer microarray data using Radial Basis Function Network." In *2012 International Conference on Statistics in Science, Business and Engineering (ICSSBE)*, pp. 1-4. IEEE, 2012. https://doi.org/10.1109/ICSSBE.2012.6396523

[26] Alshamlan, Hala, Ghada Badr, and Yousef Alohali. "mRMR-ABC: a hybrid gene selection algorithm for cancer classification using microarray gene expression profiling." *Biomed research international* 2015 (2015). https://doi.org/10.1155/2015/604910

[27] Shreem, Salam Salameh, Salwani Abdullah, Mohd Zakree Ahmad Nazri, and M. A. L. E. K. Alzaqebah. "Hybridizing ReliefF, MRMR filters and GA wrapper approaches for gene selection." *J. Theor. Appl. Inf. Technol* 46, no. 2 (2012): 1034-1039.

[28] Sivagami, P. "Supervised learning approach for breast cancer classification." *Int. J. Emerg. Trends Technol. Comput. Sci* 1, no. 4 (2012): 115-129.

[29] Deepika, J., P. Selvaraju, Mahesh Kumar Thota, Mohit Tiwari, Dunde Venu, K. Manjulaadevi, and N. Geetha Lakshmi. "Efficient classification of kidney disease detection using Heterogeneous Modified Artificial Neural Network and Fruit Fly Optimization Algorithm." *Journal of Advanced Research in Applied Sciences and Engineering Technology* 31, no. 3 (2023): 1-12. https://doi.org/10.37934/araset.31.3.112

[30] Liao, J. G., and Khew-Voon Chin. "Logistic regression for disease classification using microarray data: model selection in a large p and small n case." *Bioinformatics* 23, no. 15 (2007): 1945-1951. https://doi.org/10.1093/bioinformatics/btm287

[31] Arias-Duart, Anna, Ettore Mariotti, Dario Garcia-Gasulla, and Jose Maria Alonso-Moral. "A confusion matrix for evaluating feature attribution methods." In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3708-3713. 2023. https://doi.org/10.1109/CVPRW59228.2023.00380

[32] Dashtban, M., and Mohammadali Balafar. "Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts." *Genomics* 109, no. 2 (2017): 91-107. https://doi.org/10.1016/j.ygeno.2017.01.004

[33] Huang, Da Wei, Brad T. Sherman, and Richard A. Lempicki. "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources." *Nature protocols* 4, no. 1 (2009): 44-57. https://doi.org/10.1038/nprot.2008.211

Name of Author	Email
Muhammad Shazwan Suhiman	shazwansuhiman@gmail.com
Sayang Mohd Deni	sayan929@uitm.edu.my
Ahmad Zia Ul-Saufie Mohamad Japeri	Ahmadzia101@uitm.edu.my
Aszila Asmat	aszila@uitm.edu.my
Lirong Wang	584978631@qq.com